A QUALITATIVE COMPARISON OF MPEG VIEW SYNTHESIS AND LIGHT FIELD RENDERING

Lode Jorissen, *Patrik Goorts*, *Bram Bex*, *Nick Michiels*, *Sammy Rogmans*, *Philippe Bekaert*, *Gauthier Lafruit*

Hasselt University - tUL - iMinds
Expertise Centre for Digital Media - Wetenschapspark 2 - 3590 Diepenbeek, Belgium

ABSTRACT

Free Viewpoint Television (FTV) is a new modality in next generation television, which provides the viewer free navigation through the scene, using image-based view synthesis from a couple of camera view inputs. The recently developed MPEG reference software technology is, however, restricted to narrow baselines and linear camera arrangements. Its reference software currently implements stereo matching and interpolation techniques, designed mainly to support three camera inputs (middle-left and middle-right stereo). Especially in view of future use case scenarios in multi-scopic 3D displays, where hundreds of output views are generated from a limited number (tens) of wide baseline input views, it becomes mandatory to fully exploit all input camera information to its maximal potential. We therefore revisit existing view interpolation techniques to support dozens of camera inputs for better view synthesis performance. In particular, we show that Light Fields yield average PSNR gains of approximately 5 dB over MPEG's existing depth-based multiview video technology, even in the presence of large baselines.

Index Terms **—** Light Fields, MPEG, Stereo, View Synthesis

1. INTRODUCTION

In this paper, we study view interpolation, which synthesizes novel viewpoints from a limited set of real input camera views. Such technology opens opportunities to innovative applications, such as free viewpoint navigation in future television broadcasting and enhanced depth perception in 3D movie creation. View synthesis may use many different techniques, e.g. stereo vision, plane sweeping, light field creation, etc. To select the most optimal method for future applications and standards, we propose to compare the MPEG reference software, which uses disparity estimation on stereo image pairs, with light field methods, which use a large number of images to estimate the apparent movement of objects in the scene. Both methods are compared using MPEG's reference datasets, as described in section 2.

The experiments show that the light field method provides better results in most cases, both qualitatively and quantitatively, especially in complex scenes. This demonstrates the potential of the light field method, especially when a large number of cameras are available.

2. EXPERIMENTAL SETUP

In our experiments we tested both view synthesis algorithms on a set of rectified test image sequences, i.e. the Dog sequence by Fujii Laboratory [1], the Bee sequence by NICT [2], and the San Miguel sequence [1]. The Dog sequence consists of images from a real scene, while the San Miguel and Bee sequences consist of ray-traced images. In all scenes the cameras were placed with evenly spacing on a single line. The setup is shown in Figure 1.

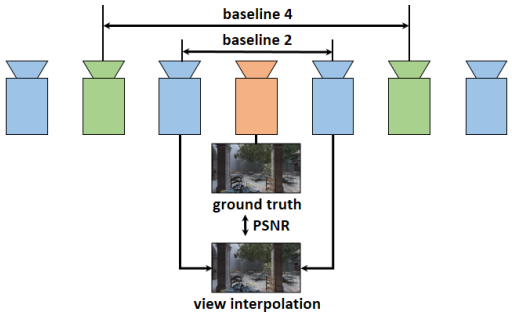


Figure 1: The camera setup of the datasets. The cameras are placed on a line, equally spread. We skip a number of cameras to simulate a larger baseline. The middle camera is always the ground truth camera, and is never used for view synthesis. This ground truth image is compared with the synthesized image using the PSNR metric.

We also tested the light field method on the Xmas sequence by Fujii Laboratory which consists of images captured from a real scene. The setup of the cameras is similar to that of the Dog and San Miguel sequences. Since there was no camera calibration information available we were unable to test the MPEG reference software on this sequence. We did, however, compare these results with state-of-the-art light field methods.

For each experiment we increased the baseline between the cameras to study the influence on the quality. This was achieved by removing cameras from the set. The frames from which the color information was used for the image synthesis are given in Table 1. The baseline units used in the experiments represent the distance between two adjacent cameras.

Performance figures are obtained through benchmarking with the MPEG reference software, which uses disparity estimation on stereo image pairs [2], warping and inpainting methods for view synthesis [3]. These view synthesis results are compared to the light field method (see section 2.1), which uses a large number of images to estimate the apparent movement of objects in the scene for depth extraction.

As quality metric we calculate the Peak Signal-to-Noise Ratio (PSNR) between the synthesized view and the ground truth image.

[1] http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/
[2] http://www.fujii.nuee.nagoya-u.ac.jp/NICT/NICT.htm

978-1-4799-4758-4/14/$31.00

	Dog		Bee		San Miguel	
Baseline	Left	Right	Left	Right	Left	Right
2	39	41	141	139	91	93
4	38	42	142	138	90	94
8	36	44	144	136	88	96
16	32	48	148	132	84	100
32	24	56	156	124	76	108
Synth. View	40		140		92	

Table 1: Cameras used for view interpolation. For each baseline and sequence the left and right cameras from the reconstructed view are given.

This metric is widely used and allows the comparison with state-of-the-art techniques.

2.1. Light Fields

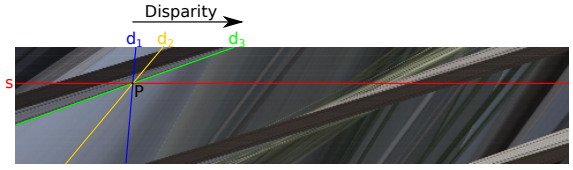
The light field interpolation algorithm is based on the work of Kim et al. [4] who present a scene reconstruction method that utilizes the properties of their light field setup to estimate highly detailed depth maps. The algorithm uses a 3D representation of light fields where one dimension s represents the starting position of each ray in the light field, and the other two, u and v, represent the direction of the rays. The pixel (u, v) in camera s represents the radiance of the ray passing through the camera center and the world position of pixel (u, v). The 3D light field can be represented as a cube by stacking all images on top of each other as shown in Figure 2a.

Each (u, s)-slice out of this cube delivers an epipolar image (EPI) [5], such as shown in Figure 2b. An EPI contains a set of lines, that corresponds to points in the scene. Different slopes (the positive angle between the x-axis and the line) are the result of the disparity of objects between adjacent frames. In particular, objects that are closer to the cameras have a smaller slope than the objects that are farther away. The depth of each pixel can hence be estimated through the slope of its corresponding EPI line. Kim et al. [4] exploit this property to generate a sparse representation of the light field, i.e. each EPI v, intersecting the cube at height v, is represented as a set of tuples $(d, u, s, \mathbf{r})$ where $\mathbf{r}$ is the average radiance of the scene point and d is its disparity.

To determine the slopes of the EPI lines one selects a scanline s in the EPI. For each pixel on the scanline a score is computed for all possible disparities and their corresponding slopes. This score considers the equivalence of all pixels along the slope, i.e. slopes for which more pixels along the corresponding line are similar to the selected pixel receive a higher score than slopes with only a very few similar pixels. The disparity with the highest score is then assigned to the pixels similar to the selected pixel on s. For each pixel on the EPI scanline s a tuple $(d, u, s, \mathbf{r})$ is created, where d represents the disparity with the highest depth score for pixel (u, s). Figure 2b shows an example of this slope estimation step. These steps result in a set of tuples for each EPI v' that can be used to generate a detailed depth map of the scene. In our experiments we used the scanlines corresponding to the images listed in Table 1.

A view synthesis for a camera at position s' can be obtained by intersecting the lines of the tuples in each EPI v' with a virtual scanline at row s' in that EPI. For each line the color of its tuple is assigned to the pixel at the intersection. The scanline, and thus also the camera, does not have to be at a discrete position. Note that the tuples need to be traversed in a back to front manner to make sure that occluding objects are in front of the occluded objects, which can be easily obtained by sorting them by their disparity value. The synthesized view is obtained by moving the filled scanline s' in each EPI v' to the scanline at row v' in the image. Holes are filled by interpolating the values of the closest pixels on the scanline.

(a) A light field cube, generated by stacking the images on top of each other. (u, v) represents image coordinates while s determines the center of the corresponding camera of each image.



(b) The EPI for the (u, s)-slice marked by the red lines in Figure 2a. Lines of scene points closer to the camera, e.g. the pillars, have a smaller slope than the objects further away, e.g. the tree in the background. The slope of the line through a pixel P along the scanline s can be determined by looping over all possible disparities and selecting the slope with the best depth score. In this example the slope for disparity d_3 has the best depth score.

Figure 2: The light field method.

2.2. MPEG Reference Software

We used the MPEG depth estimation reference software (DERS)[3] and the view synthesis reference software (VSRS) v6.0[4] for benchmarking, which have been originally developed by Nogaya University and Poznan University of Technology, updated with improvements throughout MPEG's standardization process [6, 7].

DERS uses stereo matching based on aggregation blocks and a refinement step based on graph cuts using the approach of Boykov et al. [2]. The method uses 2 or 3 input images, and can only generate depth maps corresponding to input viewpoints.

The two depth maps calculated for the given input viewpoints are used to synthesize the novel viewpoint using VSRS. The input color images are warped and blended, based on the previously calculated depth maps [3]. Holes are filled using the closest pixel values where data is available. Texture patches, however, are used when the holes are relatively big [8, 9].

3. RESULTS

We compare the MPEG reference software and light field methods in function of baseline distance, i.e. intercamera distance, to study how the quality of the synthesized view behaves for larger

[3] http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/mpeg2/DE.htm
[4] http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/mpeg2/VS.htm

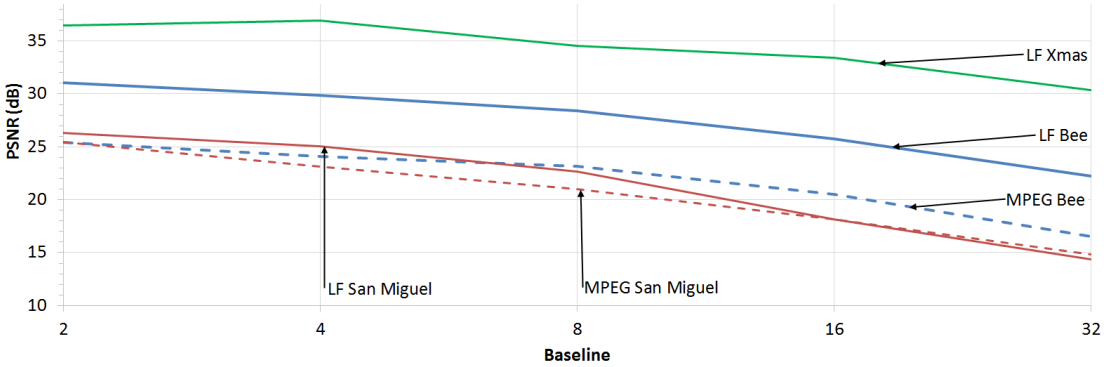


Figure 3: Comparison of the PSNR values for the light field method (LF) and the MPEG stereo method (MPEG), for the Bee, San Miguel, and XMas datasets, using varying baselines. The baseline is represented as baseline units. The PSNR values declines when the baseline increases. In most cases, the light field method is better than the MPEG method, demonstrating the usefulness of the light field method.

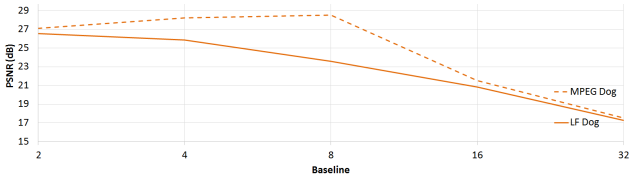


Figure 4: Comparison of the PSNR values for the light field method (LF) and the MPEG stereo method (MPEG), for the Dog dataset, using varying baselines. The baseline is represented as baseline units. In this case, the values are higher for the MPEG dataset. The visual results, however, are better for the light field method, as shown in Figure 5.

baselines. The quality metric we use is the PSNR values of the interpolated views compared to their ground truth data. The results are shown in Figures 3 and 4. Furthermore we provide a visual comparison between close-ups of the Bee and Dog scenes in Figures 5.

Figure 3 shows the PSNR of the view synthesis for the Xmas, Bee and San Miguel sequences. The quality of the Bee Light Field view synthesis is consistently 5 dB better than the results obtained with VSRS. The complex scene of San Miguel still provides a gain of up to 2 dB with Light Fields compared to VSRS. Visually, it is clear from Figure 5 that our method keeps the details of the bee intact for a much larger baseline than the MPEG reference software, which shows ghosting artifacts and missing scene portions.

With the Dog sequence, at first sight, the situation seems somewhat reversed w.r.t. the PSNR quality metric, as shown in Figure 4. However, visual inspection from Figure 5 nuances these findings: while the background in this scene starts to fade for larger baselines, other parts of the scene such as the face of the woman are almost completely kept intact even for large baselines where the MPEG reference software clearly fails.

This is due to inconsistent lighting in the images. The light field method tries to match pixels that have a radiance consistent to the radiance of the selected pixel. This incurs a large PSNR penalty, especially since the background covers a large portion of the scene.

For the Xmas sequence there was no camera calibration information available that allowed us to test the sequence for the MPEG reference software. Since our light field algorithm just assumes that the cameras are equally spread, we were still able to do a light field view synthesis for this sequence. With reference to Figure 3, light fields on the Xmas sequence yielded favorable results, compared to the more challenging Bee and San Miguel sequences, which are recent and novel. Xu et al. [10] and Jiang et al. [11] did a comparison of other light field based interpolation algorithms. In order to compare our algorithm with their results we extended our method to be able to determine the minimum and maximum depths. This was achieved with the help of the Hough Transform which detects the lines in the EPI's. A comparison of their results extended with the results of our light field method can be found in Table 2. These results show that the Pixel Matching based Interpolation (PMI), Block Matching based Interpolation (BMI) and Multi-Epipolar Lines based ray-space Interpolation (MELI) methods perform worse than our implementation in situations where there is no prior depth knowledge, while our method is comparable with the Directionality based Linear Interpolation (DLI) method. On the other hand, the Multi-Stage Search (MSS), Full Search (FS) and Radon Transform based ray-space interpolation (RTI) algorithms do perform better in this particular scene.

BL	PMI	BMI	MELI	DLI	**LF**	MSS/FS	RTI
4	24.5	27.1	30.9	37.1	**36.4**	41.3	42.2
10	23.7	26.5	30.5	34	**35.5**	41.5	41.9
20	23.9	26.5	30.4	31	**31.3**	41	41.5

Table 2: The results of Xu et al. [10] (PMI, BMI, MELI and RTI) for the Xmas sequence and the results of Jiang et al. [11] (DLI, MSS and FS) compared to the results of our light field implementation. The values represent the PSNR in dB. The baseline (BL) is represented as baseline units.

4. CONCLUSION

In this paper, we compared light field image synthesis with the MPEG depth estimation and view synthesis reference software. The former consistently exploits the information from dozens of cameras, while the latter has been designed for only three input camera views. We varied the baseline to determine the effect on quality, measured with the PSNR metric. The results show that the light field method is better for most of the scenes (approximately 5 dB), and always visually better.

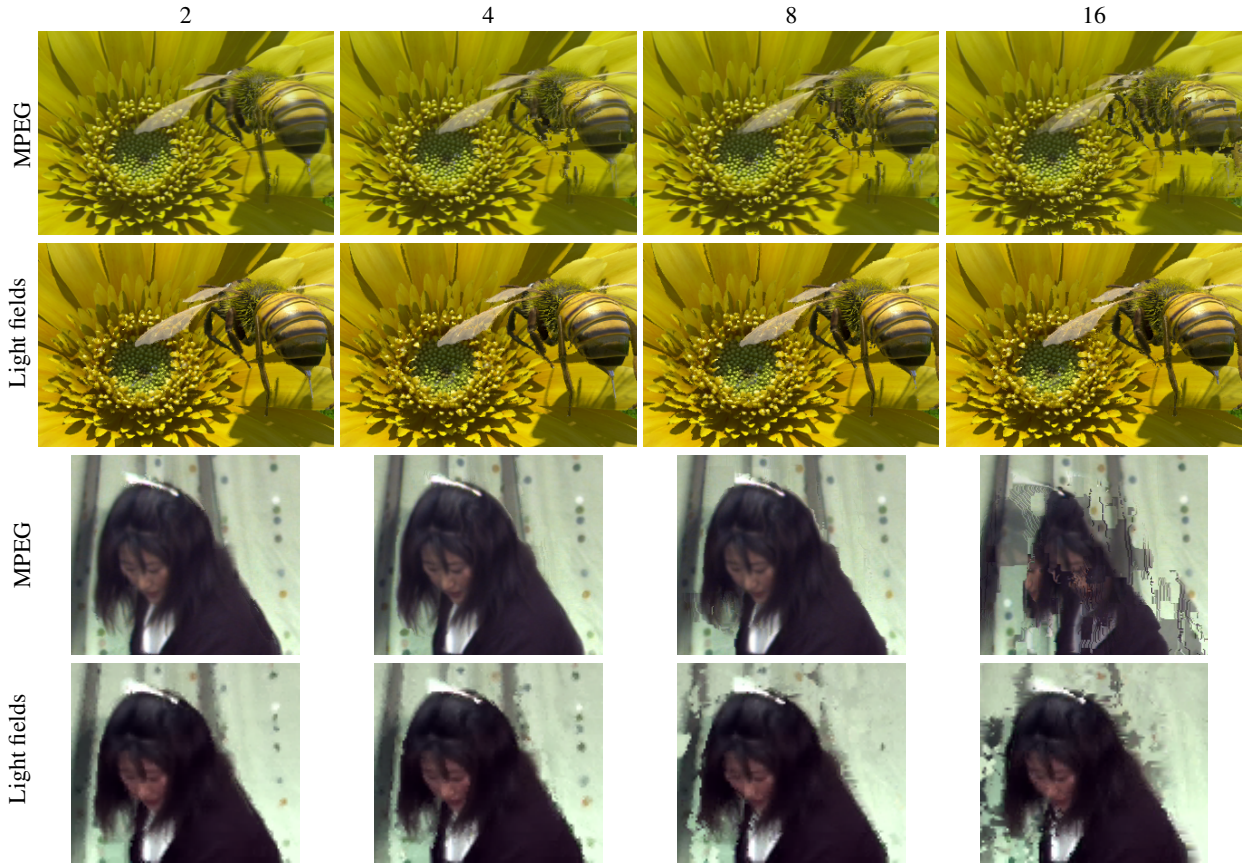


Figure 5: A visual comparison of the Bee (top two rows) and the Dog sequence (bottom two rows) for baselines 2, 4, 8 and 16 (from left to right). The details of the bee are preserved much longer when the light field method is used. The MPEG method also suffers from ghosting artifacts. In the case of the dog sequence, the overall performance of the MPEG method is better than the light field method, but the foreground objects are preserved significantly better with the light fields.

5. REFERENCES

[1] Patrik Goorts, Mohammad Javadi, Sammy Rogmans, and Gauthier Lafruit, "San miguel test images with depth ground truth, ISO/IEC JTC1/SC29/WG11/M33163," Tech. Rep., MPEG, Valencia, March 2014.

[2] Yuri Boykov and Vladimir Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 26, no. 9, pp. 1124–1137, 2004.

[3] Masayuki Tanimoto, Toshiaki Fujii, and Kazuyoshi Suzuki, "View synthesis algorithm in view synthesis reference software 2.0 (vsrs2.0)," *ISO/IEC JTC1/SC29/WG11/M16090*, vol. 16090, 2009.

[4] Changil Kim, Henning Zimmer, Yael Pritch, Alexander Sorkine-Hornung, and Markus Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. on Graphics (Proc. of ACM SIGGRAPH)*, vol. 32, no. 4, pp. 73:1–73:12, 2013.

[5] Robert C. Bolles, H. Harlyn Baker, David, and H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *Intl. Journal of Computer Vision*, vol. 1, no. 1, pp. 7–55, 1987.

[6] Olgierd Stankiewicz, Krzysztof Wegner, Masayuki Tanimoto, and Marek Domanski, "Enhanced depth estimation reference software (ders) for free-viewpoint television, ISO/IEC JTC1/SC29/WG11/M31518," Tech. Rep., MPEG, Geneva, October 2013.

[7] Krzysztof Wegner, Olgierd Stankiewicz, Masayuki Tanimoto, and Marek Domanski, "Enhanced view synthesis reference software (vsrs) for free-viewpoint television, ISO/IEC JTC1/SC29/WG11/M31520," Tech. Rep., MPEG, Geneva, October 2013.

[8] Martin Koppel, Xi Wang, Dimitar Doshkov, Thomas Wiegand, and Patrick Ndjiki-Nya, "Depth image-based rendering with spatio-temporally consistent texture synthesis for 3-d video with global motion," in *Image Processing (ICIP), 2012 19th IEEE Intl. Conf. on*. IEEE, 2012, pp. 2713–2716.

[9] Patrick Ndjiki-Nya, Martin Koppel, Dimitar Doshkov, Haricharan Lakshman, Philipp Merkle, K Muller, and Thomas Wiegand, "Depth image-based rendering with advanced texture synthesis for 3-d video," *Multimedia, IEEE Trans. on*, vol. 13, no. 3, pp. 453–465, 2011.

[10] Lingfeng Xu, Ling Hou, Oscar C. Au, Wenxiu Sun, Xingyu Zhang, and Yuanfang Guo, "A novel ray-space based view generation algorithm via radon transform," *3D Research*, vol. 4, no. 2, pp. 1–15, 2013.

[11] Gangyi Jiang, Mei Yu, Xien Ye, Liangzhong Fan, and Randi Fu, "New method of ray-space interpolation for free viewpoint video," in *Image Processing, 2005. ICIP 2005. IEEE Intl. Conf. on*, Sept 2005, vol. 2, pp. II–1138–41.