Benchmarking of several disparity estimation algorithms for light field processing

Faezeh Sadat Zakeri, Michel Bätz, Tobias Jaschke, Joachim Keinert, Alexandra Chuchvara

Faezeh Sadat Zakeri, Michel Bätz, Tobias Jaschke, Joachim Keinert, Alexandra Chuchvara, "Benchmarking of several disparity estimation algorithms for light field processing," Proc. SPIE 11172, Fourteenth International Conference on Quality Control by Artificial Vision, 111721C (16 July 2019); doi: 10.1117/12.2521747

SPIE.

Event: Fourteenth International Conference on Quality Control by Artificial Vision, 2019, Mulhouse, France

Benchmarking of Several Disparity Estimation Algorithms for Light Field Processing

Faezeh Sadat Zakeri*[a], Michel Bätz[a], Tobias Jaschke[a]
, Joachim Keinert[a], Alexandra Chuchvara*[b]
[a]Fraunhofer IIS, Moving Picture Technologies, Am Wolfsmantel 33, Erlangen, Germany; [b]Tampere University of Technology; Department of Signal Processing, Korkeakoulunkatu 10, Tampere, Finland

ABSTRACT

A number of high-quality depth imaged-based rendering (DIBR) pipelines have been developed to reconstruct a 3D scene from several images taken from known camera viewpoints. Due to the specific limitations of each technique, their output is prone to artifacts. Therefore the quality cannot be ensured. To improve the quality of the most critical and challenging image areas, an exhaustive comparison is required. In this paper, we consider three questions of benchmarking the quality performance of eight DIBR techniques on light fields: First, how does the density of original input views affect the quality of the rendered novel views? Second, how does disparity range between adjacent input views impact the quality? Third, how does each technique behave for different object properties? We compared and evaluated the results visually as well as quantitatively (PSNR, SSIM, AD, and VDP2). The results show some techniques outperform others in different disparity ranges. The results also indicate using more views not necessarily results in visually higher quality for all critical image areas. Finally we have shown a comparison for different scene's complexity such as non-Lambertian objects.

Keywords: Depth image-based rendering, disparity estimation, quality evaluation.

1. INTRODUCTION

Depth Image-Based Rendering (DIBR) [1] is a key technology for the processing and distribution of three-dimensional content. Given a two-dimensional image from a scene that was taken from a specific viewpoint and a corresponding depth or stereo-derived disparity map, DIBR enables the generation of novel views that shows synthesized viewpoints of the scene. The ability of DIBR to synthesize novel views enables the generation of enhanced additional views for stereoscopic and multi-view displays and gives control over the 3D depth impression [2]. In DIBR, the quality of an underlying depth map is correlated to the quality of the novel views generated from it. For example, mismatches, misalignments of depth and color edges, or over-smoothed depth edges can lead to visible artifacts in the novel views. Accordingly, we decided to investigate the effects of the density of original views, the disparity range within two adjacent input views on the perceived quality of the rendered novel views. We also have analyzed the limitations of our eight chosen DIBR techniques for different object properties like edges, homogeneous regions, periodic structures, fine edges, and non-Lambertian properties. The main focus of this work is to understand the behavior and robustness of each technique for different use cases and applications.

*{faezeh.zakeri, michel.baetz, tobias.jaschke, joachim.keinert}@iis.fraunhofer.de

*aleksandra.chuchvara@tut.fi

Fourteenth International Conference on Quality Control by Artificial Vision, edited by Christophe Cudel, Stéphane Bazeille, Nicolas Verrier, Proc. of SPIE Vol. 11172, 111721C · © 2019 SPIE · CCC code: 0277-786X/19/$21 · doi: 10.1117/12.2521747

There are a number of benchmarks analyzing DIBR techniques already published [3] and [4]. However, due to rapid growth in the number of DIBR techniques also remarkable improvement of the-state-of-the-art, a new study of recent DIBR techniques that put the improved state-of-the-art together with the newest ones is required. Furthermore, a benchmark that concentrates the evaluation on light field features such as density of views together with capabilities of DIBR techniques like handling different disparity ranges is missing [5]. This works contributes by analyzing the behavior of the different DIBR techniques for the purpose of light field processing.
The remainder of this paper is structured as follows: A brief review of the literature related to the selected disparity estimation algorithms and their expected behavior will be given in Section 2. Then, the experimental setup will be discussed by illustrating three aspects in more details in Section 3. Afterward, we will continue by presenting and discussing our results in Section 4. Eventually, multiple conclusive remarks will be made in Section 5.

2. LITERATURE REVIEW

We have selected the stereo disparity estimation algorithms such that they each represents majority of the most used approaches from classic ones to the state-of-the-art and put them against each other to highlight their benefits and limitations relatively. In this chapter the chosen methods are introduced:

- Block Matching Plus (BMPlus),
- Absolute Differences Census (ADC) [6],
- Census with Cross Aggregation (CCA),
- Facebook Surround 360 (OFFB) [7],
- Semi Global Matching (SGM) [8],
- Mesh Stereo (MS) [9],
- Shearlet [10],
- Superpixels-based Depth Estimation [11].

One very basic disparity estimation method is block matching [12] which involves several steps. The first step is to divide current view into several blocks. A corresponding block then is being compared with the other blocks as well as neighboring blocks in the adjacent views. The result of the comparison ends with a motion vector that simulates the movement of the corresponding block from one location to another. This step continues by comparing all the blocks within the current view. The metric to calculate the vector can be any cost function like absolute differences, (AD). The method called BMPlus, is our implementation of the block matching method that uses Census costs [13]. Block matching plus in addition applies cost volume filtering [14] on the costs are computed for a chosen pixel p. The filtering approach constructs a 3D cost volume which stores the calculated costs for a choosing label l at image coordinate x and y.Then it obtains a solution for a labeling problem by choosing the label of the lowest cost at each pixel. This smooths the cost volume such that color edges are better preserved which outperforms the common filtering methods by improving the quality of the results [14].
Absolute difference census and cross-based aggregation are both alternative block matching methods. Absolute difference census is a block matcher with cross-based cost aggregation of AD and census costs [6]. This can be viewed as a joint over the cost volumes that are calculated by AD and census cost functions. Cross-based aggregation costs are found to produce more robust cost volumes that outcome to more accurate disparities [6]. The only difference between ADC and CCA is that in cross-based aggregation only census costs are utilized.
The fourth included method is Facebook Surround 360 that we called optical flow Facebook (OFFB) which computes a dense flow map. In general, optical flow describes a sparse or dense vector field, where a displacement vector is assigned to a certain pixel position that points to where that pixel can be found in another image. In the context of scene flow estimation which is performed on images with additional depth values, every pixel is assigned a depth displacement as well [15]. The workflow is provided by Facebook is based on dense optical flow that can be applied to generate disparity maps directly.
We also considered SGM since it is a robust method that has been used compared and evaluated for years [8]. In contrast to block matching approaches, SGM is a pixel-wise approach that computes 1D minimum costs at each pixel coordinate from all directions and it solves an energy minimization problem which considers a small penalty for all pixels in the neighborhood of N pixels of pixel p for which disparity change is small [8].
Mesh stereo is another technique we used. The workflow of the proposed integrated approach is to first partition an input stereo pair into 2D triangles with shared vertices according to edge distribution and local visual consistency. Then to

compute proper disparity values for the vertices using a region based stereo model imposing photo consistency and normal smoothness [9].
Shearlet dense reconstruction method is one of the newest algorithms for 3D reconstruction of scenes with non-Lambertian properties [10]. Shearlet reconstruction utilizes sparse representation of epipolar-plane images (EPI) [16] in shearlet transform domain [10]. In order to handle straight lines, shearlet transform is justified for EPI. The method is implemented iteratively as a regularization problem with adaptive thresholding [10]. Equation 1 demonstrates number of reconstructed samples N from sparse number of given views K.

$$N = (K-1) * drange \tag{1}$$

Whereas, $drange$ indicates the range of disparities in the original views that is defined as:

$$\max disp - \min disp \tag{2}$$

Superpixels-based depth estimation that we called it superpixels approach only for the sake of simplicity is our last compared method [11]. Superpixels approach is an image segmentation technique that effectively propagates depth information from textured to ambiguous textureless areas. It estimates initial disparities via standard window-based stereo and groups them per image segment. By lifting the domain from the pixel to the superpixels level, both the computational complexity and the susceptibility to noise are reduced [11].

3. EXPERIMENTAL SETUP

Figure 1 summarizes our experimental setup in two general steps. We import the original views into the chain of algorithms in order to generate refined disparity maps first, using selected disparity estimation algorithms. Further, we feed view rendering pipeline by importing the original views and their estimated disparity maps. Eventually the rendered novel views are produced as the output of view rendering pipeline that are main results to be used for evaluation and comparison. In this section, we outline the three mentioned questions and address them one after the other. We explore effects of disparity range, density of the original views and complexity of the scene in sections 3.1, 3.2 and 3.3 respectively. After estimation of disparity maps per each pair of images by all chosen techniques, we merged disparity maps using their neighboring camera pairs.

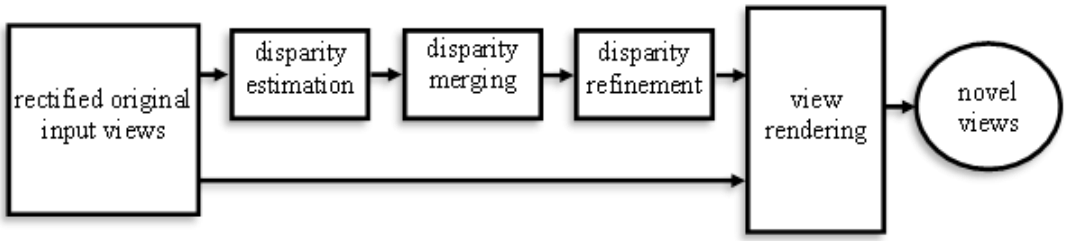

Figure 1. Block diagram showing the stages of experimental setup

As it is demonstrated in Figure 2, disparity maps are estimated using each two adjacent cameras in two directions.

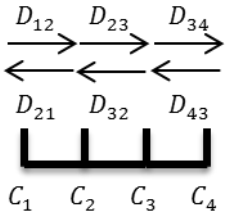

Figure 2. Example of merging approach between camera pairs.

Then for the cameras with more than one neighbor like C_2 and C_3 a function merges the neighboring disparity maps such that:

$$D_2 = merge(D_{21}, D_{23}) \tag{3}$$

$$D_3 = merge(D_{32}, D_{34}) \tag{4}$$

Once disparity maps are merged, each of them is refined by simply applying a sparse percentile filter in which each pixel in the disparity map is replaced by the disparity value of that pixel in the histogram of its neighbor pixels that corresponds with a preset percentile. This permits removal of outliers without degrading the image quality [17]. Finally the filtered maps simply are inpainted in a linewise manner such that disparities with zero value are filled with their 1D-surrounding smallest disparity value.

We eventually imported refined disparity maps and input views into a view interpolation pipeline [18]. In view interpolation first, disparity maps are forward projected [19] to a predefined desired virtual camera position in between original known camera views. This is done using the projection matrix is interpolated from the matrices of the two original views involved which includes the virtual view position. After forward warping step, projected disparities are backward warped [19] to the original camera views in order to be assigned with their intensity values. Consequently, the intensity value of the original views get blended in a linear manner and allocated to backward warped disparities.

The steps explained above are required to generate the final results for all chosen techniques except for the Shearlet approach. The Shearlet approach is an end-to-end reconstruction method and the current public binary that is available online [10] does not output the disparity maps. Thus, we had to compare the rendering results of all other techniques with the same view rendering pipeline as it is described before in a way that results are comparable with of the ones of Shearlet. Equation 1 shows that the density of reconstruction is dependent on disparity range saying that Shearlet generates more views in between original input views in compare with other techniques. Therefore the only way to include Shearlet in this comparison is to find a rendered view in a specific position based on Equation 1 such that its match from other techniques is available. This is the only way to ensure the rendered view used for comparison is at the same render position for all the methods.

3.1 Disparity Range

In this subsection, it is elaborated how the disparity range between two adjacent views are considered in the setup to be analyzed for its effects on the quality of the rendered novel views.

Taking into account that the provided images in the JPEG Pleno data set [20] are pre-rectified, meaning that images are aligned such that they all belong to a common image plane, we shrunk JPEG Pleno into a subset of five input views.

We further downsampled the views with the factor of $1-5$ into five different resolutions. Each of the downsampled subsets has different range of disparity between two adjacent views. We assigned labels from “narrow” to “very large” to each subset shown in Table 1 to categorize subsets based on their disparity range.

Table 1. Subsets, their characteristics with assigned labels

subsets with five input views	disparity range in pixels
$factor = 1$ very large	[52 100)
$factor = 2$ large	[36 52)
$factor = 3$ medium	[24 36)
$factor = 4$ small	[16 24)
$factor = 5$ narrow	[0 16)

3.2 Density of the Original Views

This subsection illustrates the manipulation of density of the original views in the experimental setup.
Paying attention to the fact that DIBR techniques are dealing with stereo situation, where initial disparities are estimated using stereo input views. In cases where the provided views are more than two images, the estimated disparity maps are fused with other camera pairs.
In order to explore the impact of number of input views, we shrunk JPEG Pleno into sparse subsets with four, five and ten views by simply skipping the views. We rendered novel views between input images such that we reconstruct dense JPEG Plano data set. We compared then the results with the corresponding ground truth images at the same position.

3.3 Complexity of the Scene

Moreover, by choosing JPEG Pleno as our ground truth which describes a complex scene we could compare selected techniques for various object properties. JPEG Pleno includes transparent objects, fine edges, periodic patterns, reflective and specular objects, homogenous colors, rich textures, shadows and occlusion zones. Most of the mentioned properties are challenging for all the disparity estimators, and their rendered results are prone to perceivable artifacts. Some visual examples are shown in Figure 3-Figure 4.

4. RESULTS AND EVALUATION

For an accurate evaluation, we compare the rendered images to the ground truth images. Taking into consideration of the different levels of quality reduction during several stages of processing, having ground-truth data to conclude weather the artifacts or noise are from the processing steps or the data itself is crucial.
As our initial aim for comparison of different techniques is to understand which techniques achieves higher quality in terms of visual perception, we observed sequences of rendered results and compared them by ground truth visually as it is shown in Figure 3- Figure 4.
In order to support our observations, we produced probability of detection using VDP2 quality metric [21] is shown map in Figure 3-Figure 4, second row. VDP2 is a visual metric that compares a pair of images (a reference and a test image) and predicts visibility that the differences between the images are visible for an average human observer. VDP2 metric also predicts quality such that quality degradation with respect to the reference image, expressed as a mean-opinion-score. The probability of detection map foresees how likely a human observer notices a difference between the two images [21].

Figure 3. Zoomed crop of a novel rendered view at a same rendered position for all techniques and corresponding ground truth (first row). The corresponding crop of probability of the detection map by VDP 2 (second row). The selected crop is part of an object with sharp edges. The results show different quality levels of disparity estimators.

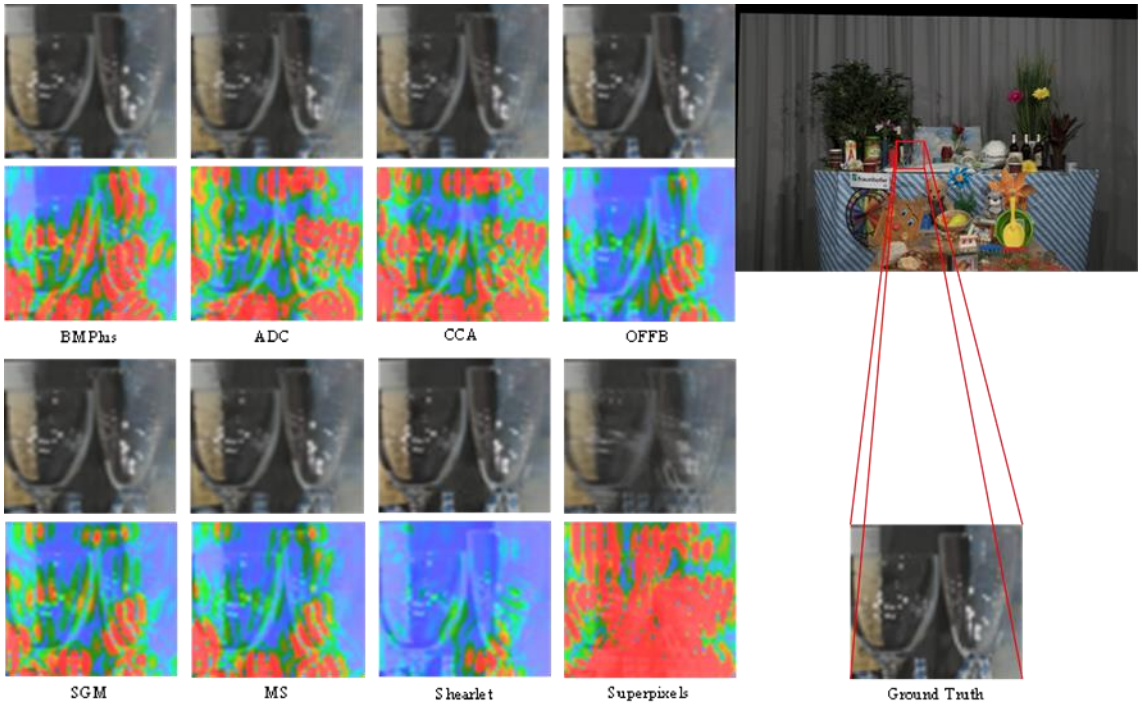

Figure 4. Zoomed crop of a novel rendered view at a same rendered position for all techniques and corresponding ground truth (first row). The corresponding crop of probability of the detection map by VDP 2 (second row). The selected crop is part of a transparent object. The results show different quality levels of disparity estimators.

To assist our evaluation with more metrics, we computed PSNR and SSIM as is shown Figure 5 for subsets introduced in Table 1.

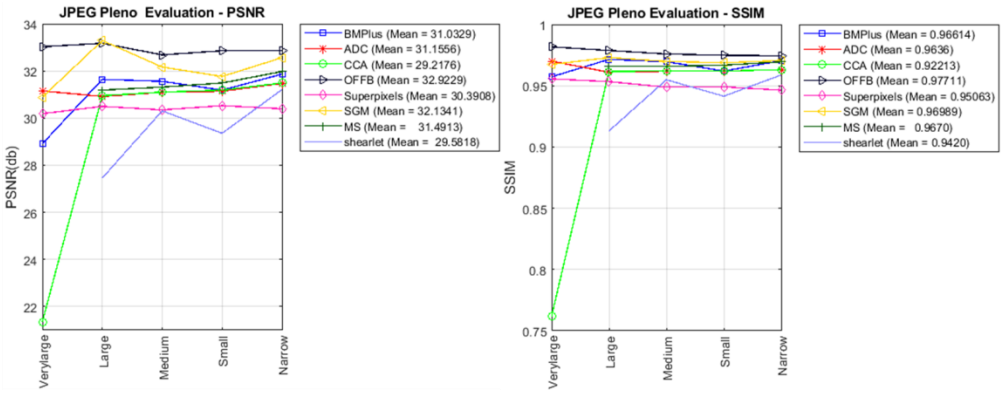

Figure 5. PSNR and SSIM results for subset of JPEG Pleno with five views. The horizontal axis is labelled based on Table 1. In both plots, for very large range the results are truncated for Shearlet and Mesh Stereo due to technical difficulties.

PSNR and SSIM plots show a significant increase for all the methods when disparity range decreases from very large to large. There is an exception for ADC in which PSNR and SSIM values both drop. The decrement of PSNR and SSIM values for ADC is not as significant as the increment of the values for all other methods. This basically refers to incapability of most of the methods for large disparity range.

From large to narrow range most of the methods show increment in their PSNR and SSIM values and finally a convergence. Except for Shearlet that shows noticeable drops from small to narrow range. Based on what is explained in [10], the shearlet reconstruction only allows four levels scaling in shearlet domain. Therefore with disparity range below sixteen pixels (narrow), the lowest possible scaling number used for initialization of scale parameter is probably higher than it should be. We think this might be the reason why PSNR and SSIM values are dropping in this range. In addition to that, due to the limitations in implementation of Shearlet approach and Mesh Stereo, rendering for very large range is currently not possible and only for these two methods plots are truncated in "very large" region.

To represent the influence of the density of the original input views on the quality of rendered novel views, we plot PSNR and SSIM values in Figure 6 per each subset as it explained in section 3.2.

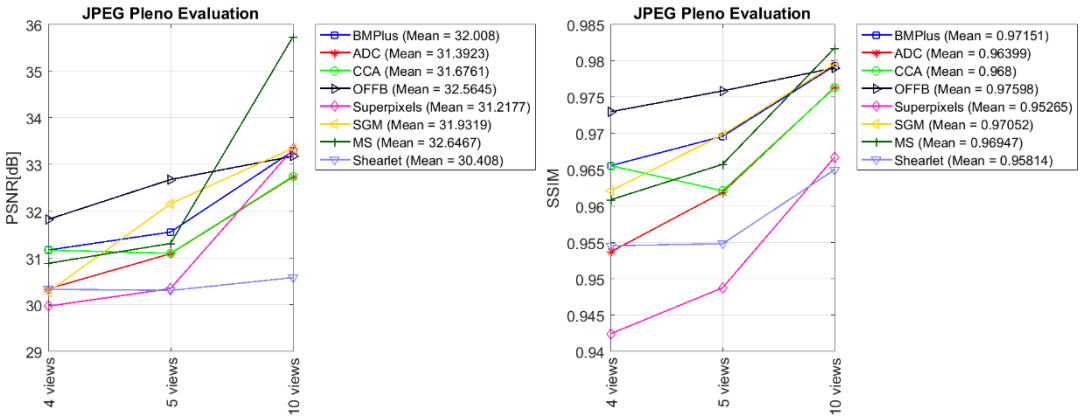

Figure 6. PSNR and SSIM results for subsets of JPEG Pleno with four, five and ten input views. The horizontal axis shows the density of the subsets.

In this case, both PSNR and SSIM values shows a general increasing manner by the increase of density of the input views for all techniques.
Note that CCA and Shearlet behave in opposite to all other methods where the input views increases from 4 to 5. PSNR plot shows that, increasing the density of input views from 5 to 10 has the most impact on MS and the least on Shearlet in compare with other techniques.
Similarly to PSNR plot, absolute difference maps (AD) in Figure 7.a (third row) confirm the manner of PSNR plot. AD maps indicate decrement of difference by using additional views in general. They also demonstrate that changing the density of input views from 4 to 5 does not influence the absolute differences between the rendered novel views and ground truth.
Aiding our evaluation, we computed probability of detection map and the difference of ground truth with the rendered novel view for one Figure 7.a shows occlusion zones where the objects are located in front of each other.
The detection map of the novel view using four input views does not show big difference with the map corresponds to the novel view is rendered using five input views. Considering that the difference is perceivable by comparing the maps of novel view using four and five input views and the map of novel view rendered with ten input views. Another example of areas with sharp edges is shown in Figure 7.b. By increasing the density of original input views the probability that an average observer perceives the difference between the rendered novel view and ground truth is increasing.
This manner can be an explanation for saying that not necessarily using more input views brings more visual quality into the last stage of DIBR pipeline. The addition of views can improve the quality of the rendered views in some areas such as occlusion zones. There other image areas which additional use of input views introduces other types of artifacts. Foreground fattening where the objects in front are being assigned the disparity of their nearby pixels of the background is an example.

5. CONCLUSIONS

In this paper, a benchmark of quality performance of eight chosen DIBR techniques is done. We studied the correlation between the quality of the rendered novel views and the kind of disparity estimators. Moreover we did evaluation of the methods in order to understand the performance of the methods in various setups such that the influence of density of the subsets used for rendering and the disparity range within adjacent views are investigated. We realized using denser light fields not necessarily results in higher visual quality. More than that, the PSNR and SSIM cannot predict quality on many areas within the image since the quality of rendered views is very dependent on the material properties of the object as well its geometrical position in the scene. One reason for that could be the fact that increasing density solves for occlusions which increases PSNR values but it introduces other types of artifacts like foreground bleeding or background fattening. This increases the probability that an average human observer perceives the degradation on image quality.
Analysis of the results allows us to know about the advantages and drawbacks of the techniques such that for scenes with distinct complexity we can decide which methods we should apply. As an example, OFFB outperforms superpixels approach in most of the image areas even though its corresponding disparity map looks noisier than the map of superpixels approach. Also, OFFB and Shearlet outperform other methods on non-Lambertian surfaces. Moreover, analysis show that even though superpixels-based disparity maps are showing smooth disparity maps in compare with other methods, they transfer more artifacts to rendered view in almost all the challenging object properties.

Acknowledgements

The work in this paper was funded from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging.

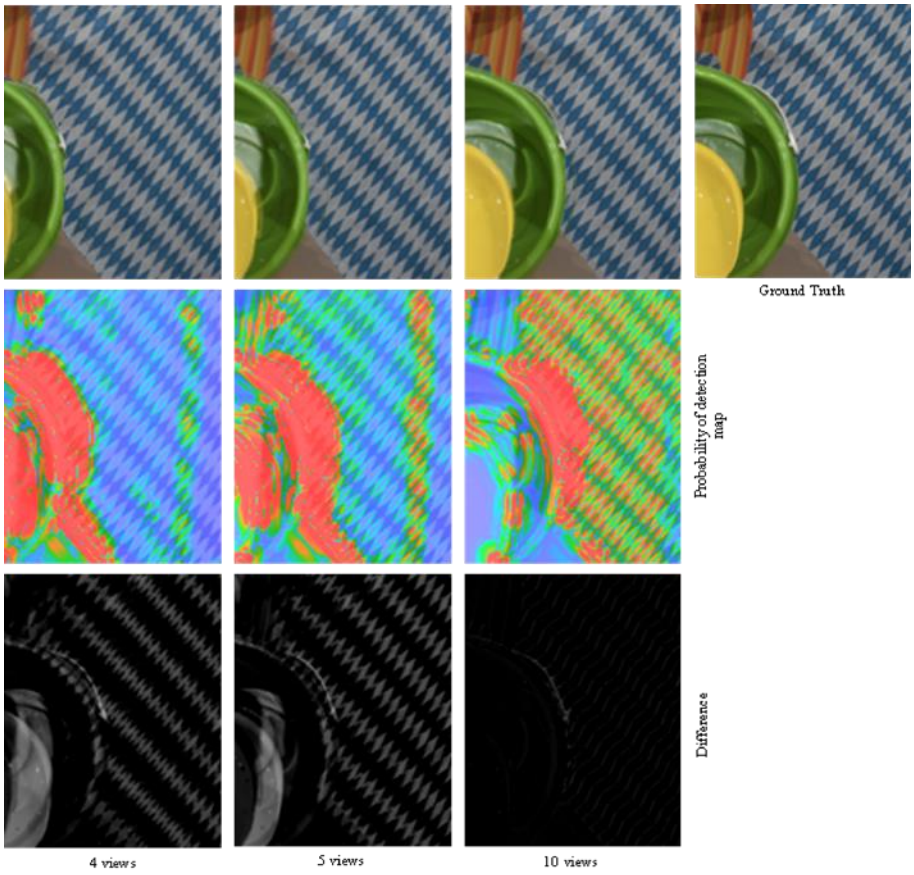

Figure 7.a. Visual results in areas with occlusion

Figure 7.b. Visual results in areas with sharp edges

Figure 7. Zoomed crop of rendered view for subsets of JPEG Pleno with different densities. The first row shows rendered view at the same position and the corresponding ground truth (top right). Second row shows the probability of detection map. Third row shows the absolute difference between rendered view (first row) and the corresponding ground truth (top right).

REFERENCES

[1] Shum, H., and Kang, S.B, (2000). "A Review of Image-based Rendering Techniques", In Visual Communications and Image Processing, The International Society for Optical Engineering, Volume 4067, pp.(2-14).

[2] Nezveda, M., (2014). PhD thesis, "Evaluation of Depth Map Post-processing Techniques for Novel View Generation", Fakultät für Informatik der Technischen Universität Wien.

[3] Slabaugh, G., Culbertson, B., Malzbender, T., and Shafer, R., (2001). "A survey of methods for volumetric scene reconstruction from photographs". In Mueller K., Kaufman A.E. (eds) Volume Graphics 2001, Eurographics, Springer, Vienna, pp.81-100.

[4] Dyer, C., (2001), "Volumetric scene reconstruction from multiple views", In L. S. Davis, editor, Foundations of Image Understanding, Kluwer, pp.469-489.

[5] Seitz, S., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R., (2006). "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms", In Proceeding of Conference on Computer Vision and Pattern Recognition, DC, USA, IEEE, Volume 1, pp.519-528.

[6] Mei, X., Sun, X., Zhou, M., Jiao, S., Wang, H., and Zhang, X., (2011). "On Building an Accurate Stereo Matching System on Graphics Hardware", In Proceeding International Conference on Computer Vision Workshop, Barcelona, Spain, IEEE, pp.467-474.

[7] Surround 360, (2016). Facebook.

[8] Hirschmuller, H., (2005). "Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information", Proceeding Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, IEEE, Volume 2, pp.807-814.

[9] Zhang, C., Li, Z., Cheng, Y., Cai, R., Chao, H., and Rui, Y., (2015). "MeshStereo: A Global Stereo Model with Mesh Alignment Regularization for View Interpolation", In International Conference on Computer Vision (ICCV), Santiago, Chile, IEEE.

[10] Vagharshakyan, S., Bregovic, R., and Gotchev, A., (2018). "Light Field Reconstruction Using Shearlet Transform", Transactions on Pattern Analysis and Machine Intelligence. IEEE, Volume 40, Issue 1, pp.133-147.

[11] Bodis-Szomoru, A., Riemenschneider, H., and V.Gool, L., (2014). "Fast approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels", In Proceeding Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, IEEE, pp.469-476.

[12] Lu, J., and Liou, M., (1997). "A Simple and Efficient Search Algorithm for Block-Matching Motion Estimation, Transactions on Circuits and Systems for Video Technology", IEEE, Volume 7, Issue 2, pp.429-433.

[13] Zabih, R., and Woodfill, J., (1994). "Non-parametric local transforms for computing visual correspondence", In Proceeding European Conference on Computer Vision, London, UK, Springer , pp.151-158.

[14] Hosni, A., Rhemann, C., Bleyer, M., Rother, C., and Gelautz, M, (2013). "Fast Cost-Volume Filtering for Visual Correspondence and Beyond, Transactions on Pattern Analysis and Machine Intelligence", IEEE, Volume 35, Issue 2, pp.504-511.

[15] Brox, T., Bruhn, A., Papenberg, N., and Weickert, J., (2004). "High Accuracy Optical Flow Estimation Based on a Theory for Warping", In Proceeding European Conference on Computer Vision, Berlin, Springer, pp.25-36.

[16] Bolles, R., Baker, H., and Marimont, D., (1987). "Epipolar-plane image analysis: An approach to determining structure from motion", International Journal of Computer Vision, Volume 1, Issue 1, pp-7-55.

[17] Duin, R.P.W., Haringa, H., and Zeelen, R., (1986) "Fast percentile filtering, Pattern Recognition Letters", North-Holland, Elsevier Science Inc., New York, NY, USA, Volume 4, Issue 4, pp.269-272.

[18] Merkle, P., Kauff, P., and Wiegand, T., (2008). "Intermediate View Interpolation based on Multiview Video plus Depth for Advance 3D Video Systems", International Conference on Image Processing, San Diego, CA, USA, IEEE.

[19] Martin, N., and Roy, S., (2018). "Fast View Interpolation from Stereo: Simpler can be Better, Fourth International Symposium on 3D Data", Processing, Visualization and Transmission, Atlanta, Georgia, USA, Georgia Institute of Technology.

[20] Ziegler, M.,Veld, R., Keinert, J., and Zilly, F., (2017). "Acquisition System for Dense Lightfield of Large Scenes", In preceedings 3DTV-CON, Copenhagen, Denmark, IEEE.

[21] Mantiuk, R., Kim, K., Rempel, A., and Heidrich, W., (2011). "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions", Transactions on Graphics, New York, NY, USA, ACM, Volume 30, Issue 4.