PARALLAX VIEW GENERATION FOR STATIC SCENES USING PARALLAX-INTERPOLATION ADAPTIVE SEPARABLE CONVOLUTION

Yuan Gao and Reinhard Koch

Christian-Albrechts-University of Kiel, 24118 Kiel, Germany

{yga, rk}@informatik.uni-kiel.de

ABSTRACT

Reconstructing a Densely-Sampled Light Field (DSLF) from a Sparsely-Sampled Light Field (SSLF) is a challenging problem, for which various kinds of algorithms have been proposed. However, very few of them treat the angular information in a light field as the temporal information of a video from a virtual camera, *i.e.* the parallax views of a SSLF for a static scene can be turned into the key frames of a video captured by a virtual camera moving along the parallax axis. To this end, in this paper, a novel parallax view generation method, Parallax-Interpolation Adaptive Separable Convolution (PIASC), is proposed. The presented PIASC method takes full advantage of the motion coherence of static objects captured by a SSLF device to enhance the motion-sensitive convolution kernels of a state-of-the-art video frame interpolation method, *i.e.* Adaptive Separable Convolution (AdaSepConv). Experimental results on three development datasets of the grand challenge demonstrate the superior performance of PIASC for DSLF reconstruction of static scenes.

Index Terms— Parallax View Generation, View Synthesis, Densely-Sampled Light Field Reconstruction, Sparsely-Sampled Light Field Capture, Parallax-Interpolation Adaptive Separable Convolution

1. INTRODUCTION

Due to the rapidly growing market demand for the Virtual Reality (VR) [1] and Free Viewpoint Video (FVV) [2] contents, how to acquire a Densely-Sampled Light Field (DSLF) from a real-world static or dynamic scene for VR or FVV rendering is currently becoming a hot research topic. Moreover, DSLFs are capable of facilitating several computer vision applications that rely on light field data, *e.g.*, depth estimation, super-resolution, synthetic aperture imaging, visual odometry, segmentation and compression [3]. However, building a DSLF capture system with a large number of densely-positioned cameras would be prohibitively expensive in data processing, camera synchronization and calibration. Therefore, sparse light field capture systems [4, 5, 6, 7, 8] have been designed for real-time light field video capture using a coarse set of cameras. How to reconstruct DSLFs from the sparse light fields captured by such systems is a challenging problem, which constitutes the main topic of the grand challenge on DSLF reconstruction.

To solve the DSLF reconstruction problem for a Sparsely-Sampled Light Field (SSLF) with parallax views for a static scene, a novel parallax view synthesis method, which is based on Adaptive Separable Convolution (AdaSepConv) [9], is proposed in this paper. Specifically, the proposed parallax view generation approach, Parallax-Interpolation Adaptive Separable Convolution (PIASC), applies a fine-tuning strategy to enhancing the convolution kernels of AdaSepConv with the consideration of the motion coherence of static objects in a parallax-view capture system. The PIASC method is evaluated on all the three development datasets of the grand challenge and further compared with AdaSepConv. Experimental results demonstrate the effectiveness of the proposed PIASC method and its superiority over the AdaSepConv approach for DSLF reconstruction of static scenes.

2. RELATED WORK

The DSLF reconstruction problem has been attempted to be solved by a lot of methods that can be generally categorized into several types, including light fields [10, 11], image-based rendering [12, 13], depth-image-based rendering [14, 15], optical flow [16, 17], light field angular super-resolution and learning-based view synthesis. Since the last two method categories are more related to the work in this paper, they are described in more detail below.

Light Field Angular Super-Resolution: Kalantari *et al.* propose a deep learning-based approach comprising both the color and disparity estimator components for view synthesis using a consumer light field camera [18]. More recently, light field angular super-resolution is explored in the area of angular detail restoration on Epipolar Plane Images (EPIs) of a light field with sparse parallax images. Vagharshakyan *et al.* present a DSLF reconstruction solution by dealing with the inpainting problem on EPIs using sparse regularization in shearlet transform domain, which is demonstrated to be effective in reconstructing non-Lambertian scenes containing semi-transparent objects [19, 20]. Wu *et al.* propose a blur-restoration-deblur framework for processing EPIs of a SSLF and restore the angular detail of the blurred and up-sampled EPIs with a Convolutional Neural Network (CNN) [21].

Learning-Based View Synthesis: Flynn *et al.* apply a deep architecture to addressing the view interpolation problem for

978-1-5386-1737-3/18/$31.00

synthesizing novel natural imagery between wide baseline views in the real-world environments [22]. Niklaus *et al.* leverage a deep neural network to estimate spatially-adaptive 2D convolution kernels, which capture both the motion and interpolation information for pixel-wise video frame synthesis [23]. However, due to the large memory requirement for storing the convolution kernel information for all the image pixels, the whole desired virtual frame may not be synthesized at once by this method. To overcome this limitation, Niklaus *et al.* propose a spatially-adaptive separable convolution method by approximating each 2D convolution kernel with a pair of 1D kernels [9]. Liu *et al.* employ an end-to-end fully-convolutional deep network, Deep Voxel Flow (DVF), for video frame interpolation and extrapolation with sharp and realistic results [24].

3. METHODOLOGY

3.1. Preliminary

The parallax view generation of a SSLF for a static scene can be simplified as a novel view interpolation problem by utilizing the color information of only two RGB images from a pair of adjacent parallax views in this SSLF. Suppose the two RGB images are denoted by $\mathcal{I}_1$ and $\mathcal{I}_2$, and the intermediate view to be reconstructed is represented by $\tilde{\mathcal{I}}$. All these three images have the same resolution, $m \times n$ pixels. The AdaSepConv approach proposed in [9] is essentially to estimate two 2D convolution kernels, $\mathbf{K}_1(x, y)$ and $\mathbf{K}_2(x, y)$, for each 2D point (x, y) on $\tilde{\mathcal{I}}$. Specifically, each 2D kernel $\mathbf{K}_\mu(x, y)$ is approximated by a pair of 1D vectors $(\boldsymbol{v}^\mu_{x,y}, \boldsymbol{h}^\mu_{x,y})$:

$$\mathbf{K}_\mu(x, y) = \boldsymbol{v}^\mu_{x,y}(\boldsymbol{h}^\mu_{x,y})^{\mathrm{T}}. \tag{1}$$

The final color information for (x, y) on $\tilde{\mathcal{I}}$ is recovered by

$$\tilde{\mathcal{I}}(x, y, c) = \sum_{\mu=1}^{2} \big(\mathbf{K}_\mu(x, y) * \mathbf{P}_\mu(x, y, c)\big). \tag{2}$$

Here, '$*$' is the covolution operation symbol and c stands for the color channel, *i.e.* $c \in \{r, g, b\}$. Besides, $\mathbf{P}_\mu(x, y, c)$ represents the image patch centered at (x, y) in the c channel of $\mathcal{I}_\mu$, which has the same size as $\mathbf{K}_\mu(x, y)$, *i.e.* $k \times k$. Compared with the Adaptive Convolution (AdaConv) method proposed in [23], the AdaSepConv approach reduces the number of unknown kernel parameters from $2m \times n \times k^2$ to $2m \times n \times 2k$, thereby enabling a high-resolution synthesized view to be generated at once efficiently.

3.2. Parallax-Interpolation AdaSepConv (PIASC)

The DSLF reconstruction for a SSLF of a static scene can be treated as the frame interpolation of a standard video that is captured by a virtual camera moving along the parallax axis of a SSLF capture system. The AdaSepConv method is originally designed for novel frame synthesis for videos containing objects moving in different directions at varying speeds. Additionally, constructing a dedicated fully convolutional network based on AdaSepConv for the purpose of DSLF reconstruction is not always easy, considering that public high-resolution and high-quality real-world light field datasets are not as common as public high-definition and high-fidelity real-world videos and the training process would be enormously time- and effort-consuming. In order to take full advantage of the state-of-the-art video frame interpolation method, AdaSepConv, for tackling the DSLF reconstruction problem and to avoid its cumbersome re-training process, a novel DSLF reconstruction method, PIASC, is proposed. More details about it are introduced as below.

Fig. 1. A flow chart of reconstructing a DSLF from a sparse set of parallax views. The novel views are reconstructed recursively in three steps. The circles with solid lines represent duplicates of the under-sampled ground-truth camera views for the sub-challenge category $\mathcal{C}_1$ on a development dataset as described in Section 4.1. The circles with dash lines are unknown parallax views to be reconstructed.

A 2D convolution kernel $\mathbf{K}_\mu(x, y)$ generated by the deep neural network of AdaSepConv contains both motion and resampling information for any object moving in any direction. However, in the grand challenge on DSLF reconstruction, a SSLF dataset is composed of a sparse set of parallax images for a static scene; in other words, static objects in these parallax images have only one motion direction that coincides with the parallax axis of the SSLF dataset. Intuitively, performing a fine-tuning strategy that enhances the motion-sensible convolution kernels of AdaSepConv should be beneficial to the parallax view synthesis for a SSLF. Accordingly, the proposed PIASC method implements this fine-tuning process by adjusting all the coefficient values in the convolution kernels with an elaborately designed weight matrix $\mathbf{W}$, *i.e.*

$$\hat{\mathbf{K}}_\mu(x, y) = \frac{k^2}{\sum_{i=1}^{k}\sum_{j=1}^{k} \mathbf{W}(i, j; \sigma)} \mathbf{W} \circ \mathbf{K}_\mu(x, y), \tag{3}$$

where

$$\mathbf{W}(i, j; \sigma) = \exp\left(-\frac{1}{2}\left(\frac{|j - \bar{k}|}{\sigma}\right)^2\right), \quad i, j \in \mathbb{Z} \cap [1, k], \;\; \bar{k} = \frac{k+1}{2}. \tag{4}$$

Here, '$\circ$' denotes the element-wise (Hadamard) product and $\hat{\mathbf{K}}_\mu(x, y)$ represents the horizontal-motion-enhanced convolution kernel generated by PIASC for the DSLF reconstruction along the horizontal parallax axis of a SSLF. The weight matrix $\mathbf{W}$ is similar to a Gaussian kernel; however, only the coordinate information along the vertical axis of $\mathbf{W}$ is taken into account by PIASC, which is because of the horizontal-parallax feature of the datasets in the grand challenge.

3.3. DSLF Reconstruction for SSLFs

After the above introduction about the proposed PIASC method, this section is dedicated to investigating how to lever-

Input: Dataset size t $(= 193)$;
Camera view interval $\tau \in \{8, 16, 32\}$;
Camera view $\mathcal{I}_\mu$, $\mu \in \{1, 1+\tau, 1+2\tau, \ldots, t\}$.
Output: Reconstructed view $\tilde{\mathcal{I}}_\omega$, $\omega \in \mathbb{Z}^+$ and $\omega \leqslant t$.

/* range $(1,\ t,\ \tau) = \{1, 1+\tau, 1+2\tau, \ldots, t\}$ */
for ω **in** range $(1,\ t,\ \tau)$ **do**
  $\tilde{\mathcal{I}}_\omega \leftarrow \mathcal{I}_\omega$;
end
while $\tau > 1$ **do**
  $\dot{\tau} \leftarrow \frac{\tau}{2}$;
  for ω **in** range $(1,\ t-\tau,\ \tau)$ **do**
    $\tilde{\mathcal{I}}_{\omega+\dot{\tau}} \leftarrow$ PIASC $(\tilde{\mathcal{I}}_\omega,\ \tilde{\mathcal{I}}_{\omega+\tau})$;
  end
  $\tau \leftarrow \dot{\tau}$;
end

Algorithm 1: A parallax view generation algorithm for a SSLF dataset, which is created from a DSLF dataset.

age this approach to reconstruct a DSLF from a SSLF. The overall process of DSLF reconstruction for a SSLF is depicted in Algorithm 1. The camera view interval τ denotes the sampling interval on a DSLF dataset comprising ground-truth parallax views. The under-sampled parallax views in this DSLF dataset form a SSLF dataset, which is firstly used to recover a portion of parallax views of a desired unknown DSLF. The orange circles with solid lines illustrated in Fig. 1 stand for these reconstructed views, which are essentially duplicates of all the views in the SSLF dataset. The unknown views in the middle of adjacent reconstructed views are then synthesized by utilizing the PIASC method, corresponding to the Step 1 and yellow dash-line circle '5' in Fig. 1. Finally, this operation is repeated recursively until all the parallax views of the desired unknown DSLF are reconstructed, *i.e.* the Step 2 and 3 in Fig. 1.

4. EXPERIMENTS

4.1. Experimental Settings

Grand Challenge Introduction: The grand challenge on DSLF reconstruction has three sub-challenges, *i.e.* three categories of decimated-parallax imagery for DSLF reconstruction, which are denoted by $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$. In particular, these three categories have different numbers of camera views along parallax axis, such that the adjacent images in $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ have varying disparity ranges, *i.e.* narrow ($\approx$ 8 pixels), moderate ($\approx 15-16$ pixels), and wide ($\approx 30-32$ pixels), corresponding to $\tau = 8, 16, 32$ in Algorithm 1 separately.

Development Datasets: To evaluate the performance of different DSLF reconstruction algorithms, three Development Datasets (DDs) of varing 3D scenes are provided by the grand challenge. The development datasets are composed of pre-rectified horizontal-parallax multi-perspective RGB images with the same resolution, *i.e.* $m = 1,280$ and $n = 720$ in Section 3.1. Two of the datasets, Lambertian DD and Complex DD, are captured by a high-quality low-noise camera mounted on a highly-precise gantry for two different real 3D scenes. The third one, Synthetic DD, is rendered by Blender for a photorealistic 3D scene.

Table I. The lowest per-view PSNR results (in dB, explained in Section 4.1) for the performance evaluation of different methods on three development datasets for three sub-challenge categories of the grand challenge.

Lambertian DD				
Cat.	AdaSepConv ($\mathcal{L}_1$)	AdaSepConv ($\mathcal{L}_F$)	PIASC ($\mathcal{L}_1$)	PIASC ($\mathcal{L}_F$)
$\mathcal{C}_1$	43.001	43.162	**44.253**	43.657
$\mathcal{C}_2$	41.619	41.708	**43.091**	42.160
$\mathcal{C}_3$	38.857	38.760	**38.988**	38.436
Synthetic DD				
Cat.	AdaSepConv ($\mathcal{L}_1$)	AdaSepConv ($\mathcal{L}_F$)	PIASC ($\mathcal{L}_1$)	PIASC ($\mathcal{L}_F$)
$\mathcal{C}_1$	36.329	36.156	**36.451**	36.186
$\mathcal{C}_2$	35.256	35.143	**35.491**	35.271
$\mathcal{C}_3$	32.539	32.312	**32.666**	32.333
Complex DD				
Cat.	AdaSepConv ($\mathcal{L}_1$)	AdaSepConv ($\mathcal{L}_F$)	PIASC ($\mathcal{L}_1$)	PIASC ($\mathcal{L}_F$)
$\mathcal{C}_1$	34.620	34.682	**34.736**	34.645
$\mathcal{C}_2$	30.884	30.897	**30.974**	30.866
$\mathcal{C}_3$	27.500	26.922	**27.538**	26.896
Average performance for each sub-challenge category across all the DDs				
Cat.	AdaSepConv ($\mathcal{L}_1$)	AdaSepConv ($\mathcal{L}_F$)	PIASC ($\mathcal{L}_1$)	PIASC ($\mathcal{L}_F$)
$\mathcal{C}_1$	37.983	38.000	**38.480**	38.163
$\mathcal{C}_2$	35.920	35.916	**36.519**	36.099
$\mathcal{C}_3$	32.965	32.665	**33.064**	32.555

Evaluation Criteria: The reconstructed parallax views for each sub-challenge category on a develop dataset are compared against ground-truth horizontal-parallax images in this develop dataset. The per-view PSNR is exploited to perform the quality evaluation for a reconstructed view $\tilde{\mathcal{I}}$ with using its corresponding ground-truth view $\mathcal{I}$, *i.e.*

$$\mathrm{MSE} = \frac{1}{m \times n \times 3} \sum_{x=1}^{m} \sum_{y=1}^{n} \left\| \tilde{\mathcal{I}}(x,y) - \mathcal{I}(x,y) \right\|_2^2, \qquad \mathrm{PSNR} = 10 \log_{10} \left(\frac{255^2}{\mathrm{MSE}} \right). \tag{5}$$

The lowest per-view PSNR for a sub-challenge category on a development dataset is selected as the single quality measure for this category on this development dataset.

Implementation Details: The pre-trained fully convolutional neural networks of AdaSepConv with $\mathcal{L}_1$ loss and perceptual loss $\mathcal{L}_F$ are from [9]. Regarding the parameters for the weight matrix $\mathbf{W}$ in (4), $k = 51$ and $\sigma = 200$.

4.2. Results and Analysis

The quantitative evaluation results of the proposed methods and baseline approaches are exhibited in Table I. As can be seen from this table, for the same sub-challenge category, all the methods achieve the best performance on the Lambertian DD, but the worst performance on the Complex DD. This is because the real 3D scene of Lambertian DD consists of objects with Lambertian reflectance only; however, the photo-realistic 3D scene of Synthetic DD has predominantly semi-transparent objects and the real 3D scene of Complex DD includes depth variations, occlusions and reflective objects. In other words, the scene-complexity order for the development datasets is Lambertian DD $<$ Synthetic DD $<$ Complex DD.

Besides, the proposed PIASC method with $\mathcal{L}_1$ loss outperforms the other three approaches on all the three development datasets for each sub-challenge category, which proves the effectiveness of PIASC $(\mathcal{L}_1)$ method for DSLF reconstruction of static scenes. Moreover, the average performance for all the methods on each sub-challenge category across all the three development datasets is shown at the bottom of Table I. It can be found that, for $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$, the PIASC $(\mathcal{L}_1)$ method achieves average-PSNR improvements of 1.26%, 1.67% and 0.30% compared to the maximal average-PSNR values of AdaSepConv $(\mathcal{L}_1)$ and AdaSepConv $(\mathcal{L}_F)$ on these three sub-challenge categories. It implies that the proposed PIASC $(\mathcal{L}_1)$ approach is more effective for DSLF reconstruction of static scenes with moderate occlusions and specular reflections.

5. CONCLUSION

This paper presents a novel parallax view generation algorithm based on PIASC for the grand challenge on DSLF reconstruction for decimated-parallax imagery of static scenes. The proposed PIASC method fully leverages the object-motion coherence of a horizontal-parallax SSLF to enhance the motion-sensitive convolution kernels, which are generated by one of the state-of-the-art learning-based video frame synthesis approaches, *i.e.* AdaSepConv. Experimental results on three development datasets with varying level of scene complexity show that PIASC achieves the better DSLF reconstruction performance than the AdaSepConv approach.

6. ACKNOWLEDGMENTS

The work in this paper was funded from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676401, European Training Network on Full Parallax Imaging, the German Research Foundation (DFG) No. K02044/8-1. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

7. REFERENCES

[1] J. Yu, "A light-field journey to virtual reality," *IEEE MultiMedia*, vol. 24, no. 2, pp. 104–112, 2017. 1

[2] A. Smolic, "3D video and free viewpoint video - From capture to display," *Pattern Recognition*, vol. 44, no. 9, pp. 1958–1968, 2011. 1

[3] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," *IEEE J-STSP*, vol. 11, no. 7, pp. 926–954, 2017. 1

[4] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," *ACM TOG*, vol. 24, no. 3, pp. 765–776, 2005. 1

[5] L. Dąbała, M. Ziegler, P. Didyk, F. Zilly, J. Keinert, K. Myszkowski, H.-P. Seidel, P. Rokita, and T. Ritschel, "Efficient multi-image correspondences for on-line light field video processing," *Computer Graphics Forum*, vol. 35, no. 7, pp. 401–410, 2016. 1

[6] S. Esquivel, Y. Gao, T. Michels, L. Palmieri, and R. Koch, "Synchronized data capture and calibration of a large-field-of-view moving multi-camera light field rig," in *3DTV-CON Workshops*, 2016. 1

[7] Y. Gao, S. Esquivel, R. Koch, M. Ziegler, F. Zilly, and J. Keinert, "A novel Kinect V2 registration method for large-displacement environments using camera and scene constraints," in *ICIP*, 2017, pp. 997–1001. 1

[8] Y. Gao, S. Esquivel, R. Koch, and J. Keinert, "A novel self-calibration method for a stereo-ToF system using a Kinect V2 and two 4K GoPro cameras," in *3DV*, 2017. 1

[9] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *ICCV*, 2017, pp. 261–270. 1, 2, 3

[10] M. Levoy and P. Hanrahan, "Light field rendering," in *SIGGRAPH*, 1996, pp. 31–42. 1

[11] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *SIGGRAPH*, 1996, pp. 43–54. 1

[12] H.-Y. Shum, S.-C. Chan, and S.-B. Kang, *Image-based rendering*, Springer Science+Business Media, 2007. 1

[13] H.-Y. Shum, S.-B. Kang, and S.-C. Chan, "Survey of image-based representations and compression techniques," *IEEE TCSVT*, vol. 13, no. 11, pp. 1020–1037, 2003. 1

[14] C. Zhu, Y. Zhao, L. Yu, and M. Tanimoto, *3D-TV system with depth-image-based rendering*, Springer Science+Business Media, 2013. 1

[15] C. Fehn, R. De La Barré, and S. Pastoor, "Interactive 3-DTV-concepts and key technologies," *Proceedings of the IEEE*, vol. 94, no. 3, pp. 524–538, 2006. 1

[16] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *IEEE TPAMI*, vol. 34, no. 9, pp. 1744–1757, 2012. 1

[17] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *IJCV*, vol. 92, no. 1, pp. 1–31, 2011. 1

[18] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM TOG*, vol. 35, no. 6, pp. 193:1–193:10, 2016. 1

[19] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform," *IEEE TPAMI*, vol. 40, no. 1, pp. 133–147, 2018. 1

[20] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Accelerated shearlet-domain light field reconstruction," *IEEE J-STSP*, vol. 11, no. 7, pp. 1082–1091, 2017. 1

[21] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on EPI," in *CVPR*, 2017, pp. 1638–1646. 1

[22] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "DeepStereo: Learning to predict new views from the world's imagery," in *CVPR*, 2016, pp. 5515–5524. 2

[23] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *CVPR*, 2017, pp. 2270–2279. 2

[24] Z. Liu, R. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *ICCV*, 2017, pp. 4473–4481. 2