LIGHT FIELD SYNTHESIS USING INEXPENSIVE SURVEILLANCE CAMERA SYSTEMS

Frederike Dümbgen[†*] *Christopher Schroers*[‡] *Kenny Mitchell**

† School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland
‡ Disney Research Studios *Disney Research Los Angeles and Edinburgh Napier University

ABSTRACT

We present a light field synthesis technique that achieves accurate reconstruction given a low-cost, wide-baseline camera rig. Our system integrates optical flow with methods for rectification, disparity estimation, and feature extraction, which we then feed to a neural network view synthesis solver with wide-baseline capability. We propose two novel warping methods that improve the accuracy of disparity estimation and view synthesis. The methods enable the use of off-the-shelf surveillance camera hardware in a simplified and expedited capture workflow. A thorough analysis of the process and resulting view synthesis accuracy over state of the art is provided.

***Index Terms*—** Light field reconstruction, view synthesis, camera arrays, surveillance cameras, disparity estimation

1. INTRODUCTION

The task of image-based view synthesis is to reconstruct novel viewpoints given one or more sample views of a scene. Light field *video* synthesis heightens this challenge to reproduce novel views of dynamic scenes for high-quality view-dependent appearance. Such viewpoint synthesis from real-world captured light fields finds uses in television, video games and virtual reality, as well as architectural visualization [1].

A variety of devices have been employed to capture light fields, all showing a trade-off between equipment cost, assembly and calibration effort, reconstruction performance and workflow simplicity. Micro-lenslet array cameras such as Lytro [2] and Raytrix [3], for out-of-the-box light field capture have limited baseline and are primarily appropriate for light field focus manipulation uses. Single scan swept paths using *e.g.* GoPros [4] also provide a low cost and simple usage, but require custom constructions and only deal with static scenes. More involved multi stereo pairs [5] and large camera arrays [6], can be highly complex and costly requiring special purpose ingest hardware assembly and calibration, with custom synchronization and machine vision components. Meanwhile, surveillance camera systems have been developed for consumer ease of use of multi-camera ingest with low-cost packaged storage, and with ever increasing resolution and frame rate. Since such systems are not designed for accurate synchronization, homogeneous image capture or calibration, we develop new methods to enable these widely accessible systems for the purpose of light field synthesis.

We propose a view synthesis algorithm adapted to low-cost camera rigs, with particular attention for human motion capture. We develop an integrated process to solve camera rectification, whilst simultaneously refining disparity estimations, by employing optical flow. Given refined disparities we then adapt the neural network view synthesis method of Kalantari *et al.* [7], overcoming their limitation of narrow baseline cameras by selective application of a disparity prior. The contributions of this paper are three-fold:

- A multi-camera rectification method making use of optical flow methods.
- Two novel warping strategies to reduce artefacts in novel view and disparity estimates.
- A depth-guided feature extraction method for learning-based view synthesis.

Our proposed method allows for the creation of a detailed representation of scenes, such as shown in Figure 1, at arbitrary sampling density. Such representations can be fed into light field displays or used for Virtual Reality experiences [8].

2. RELATED WORK

In most light field capture setups, cameras are arranged with uniform spacing on lines or grids, thus facilitating the sampling of the 4D light field. This leads to a particular geometric property in the epipolar plane images (EPIs): the slopes of lines formed by corresponding pixels correspond to their depth [9]. Various methods making use of this property have been proposed for 3D reconstruction from light fields [10]. For instance, Kim *et al.* [11] recover high-resolution depth maps of urban scenes from a linearly moving camera using a sparse EPI representation and a fine-to-coarse estimation approach. More recently, learning-based approaches [12] have been considered because of their ability to account for nonlinear phenomena induced by view-dependent effects and miscalibration.

EPI-based methods require very accurate calibration in

*The work was performed while at Disney Research Los Angeles.

Fig. 1: View synthesis results from the proposed methods. We use images 1, 4 and 7 to synthesize intermediate views (2,3,5,6).

terms of both lighting and camera poses. The method proposed by Kalantari *et al.* [7] does not rely on EPIs and is therefore more suitable for inexpensive camera rigs such as ours.

Recently, there has been a surge in view synthesis solutions based on the so-called Layered Depth Images (LDI), originally introduced by Shade [13]. Both Tulsani *et al.* [14] and Dhamo *et al.* [15] proposed a method to learn a 2-layered depth plane representation, allowing to synthesize both depth and textures in (visible) foreground and (occluded) background. While the goal of the above methods is to learn the content of non-visible pixels, our method is not restricted to one or few views and uses information from neighboring views to fill in missing information in the case of disocclusions.

Penner *et al.* [16] and Flynn *et al.* [17] used a similar multi-layer representation on multiple input images, constructing the color image at the target view from a probabilistic depth volume, and probabilistic disparities, respectively. Zhou *et al.* [18], introduce multiplane images, which enable a single global scene representation, encompassing colors and visibility factors at a predefined set of depth planes.

While LDI-based approaches outperform the work of Kalantari *et al.* in terms of occlusion handling, they have a limited capability in resolving view-dependent effects, which is of high importance in human motion capture. By introducing a prior-guided feature extraction and improved warping schemes for Kalantari's pipeline, we create a system that can model view-dependent effects and correctly resolve disocclusions.

3. REFINED RECTIFICATION AND DISPARITY ESTIMATION

3.1. Multi-Camera Rectification

While a solution for rectifying camera pairs facing the same direction always exists [19], [20], rectification of more than two cameras can only be approximated. Our multi-camera rectification method is based on the solution proposed by Nozick [21], but tailored to motion capture setups like ours. Firstly, since all cameras share a large field of view, one single camera is enough as reference (no iterative updates between views are required). Secondly, we set the orientation to the average of all cameras, which ensures a good match between rectified and original poses. We model the focal lengths of our N cameras, to be the minimizers of the following cost function:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \sum_{m=1}^{M} \left| \left(f(\theta_i, x_m) - u_{1m} \right)_y \right|, \qquad (1)$$

where $\theta_i \in \mathbb{R}^2$ contains the unknown focal lengths for camera i, M is the number of calibration points and u_{1m} is the image coordinate of point m in the left-most camera. For mapping the calibration point $x_m \in \mathbb{R}^3$ to the rectified image of camera i, we use the standard pinhole model [20], denoted by $f(\theta_i, x_m)$.

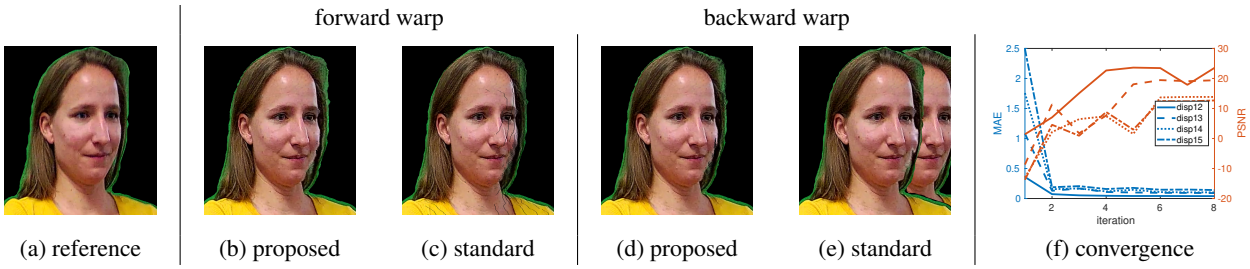
We solve (1) with the Nelder-Mead minimization scheme provided in SciPy.[1] The optimization converges in only a few iterations and the final cost per calibration point is on average less than 0.5 pixels for all camera pairs. The performance was experimentally found to be independent of the choice of the reference camera.

The rectification accuracy decreases with increasing distance from the location of the calibration pattern in the camera frames. However, since optical flow is an important component of our view synthesis pipeline, we can favorably use it to refine the rectification too: by forward-warping the input images with the vertical component of the optical flow field, we obtain accurately rectified images in the region of interest, without the need for re-calibration. As shown later, this significantly improves reconstruction accuracy.

3.2. Disparity Estimation

Given the input images from a set of cameras, we obtain the optical flow field between each camera pair using the PWC-

[1] `www.scipy.org`, version 1.1.0.



(a) reference (b) proposed (c) standard (d) proposed (e) standard (f) convergence

Fig. 2: Results of the proposed warping strategies. Figures (b) to (e) show that the cracking and ghosting artefacts of the naive approaches are correctly resolved using the proposed method. Figure (f) shows the disparity refinement convergence results. dispij denotes the error between the warped input disparity at position i and the target disparity at position j. Both MAE and PSNR are depicted.

Net [22] method. This choice is motivated by its recent success and software availability. We use the PyTorch implementation by Niklaus [2], and perform a coarse foreground segmentation before feeding the input images into the pipeline.

To obtain a disparity map from the optical flow field, we first warp the flow images using the vertical flow component, like in the rectification refinement. This estimate is then refined by exploiting the high redundancy between pairs of disparity maps: indeed, warping one disparity map with itself, one should obtain the neighboring disparity map and vice-versa. This leads to the following iterative pairwise horizontal refinement algorithm:

$$\widetilde{D}_{ij} = g(D_{ji}^{(k)}, -D_{ji}^{(k)}), \quad D_{ij}^{(k+i)} = \max\left(\widetilde{D}_{ij}, D_{ij}^{(k)}\right), \tag{2}$$

where D_{ij} is the disparity map from camera i to j, max is the pixel-wise maximum, and $g(D, I)$ is our forward-warping operator applied to input disparity D and image I. Choosing the maximum of warped and reference disparity map ensures that we do not introduce contents from the background to the foreground. This iterative process converges in a few steps, as shown in Figure 2 (f).

3.3. Improved Warping Strategies

We propose two warping strategies which reduce artefacts typically arising from standard approaches.

Forward Warping Our forward warping method is based on the strategy proposed by Jantet [23]. The method uses the insight that in pixel-wise warping, sequentially warped points are usually adjacent or overlapping. When the intensities are overlapping, we keep the more recent value, which is reasonable as long as we adapt the correct scanning order [24]. If there is a gap between two sequentially projected pixels (denoted by Δp) there are two plausible scenarios. For small gaps ($\Delta p \leq p_f \in \mathbb{N}$), there is a crack; an artefact naturally arising when there are changes in smooth disparity regions which are bigger than one pixel, shown in Figure 2 (c). Beyond the threshold ($\Delta p > p_f$), there is a disocclusion. We interpolate cracks as reported previously [23] but leave disocclusions empty, to fill them with information from neighboring views in the view synthesis process.

Backward Warping In the same spirit as the forward warping scheme, we introduce a latency threshold p_b. We mask pixels from the input image during warping as soon as we moved away by p_b or more pixels. This eliminates the ghosting effect, which typically occurs in standard backward warping, shown in Figure 2 (e). The latency makes sure pixels can be used more than once, which is important in particular in smooth disparity regions.

The obtained warping schemes could be extended to non-rectified image pairs by considering pixel neighborhoods rather than lines [25]. We empirically find that setting both thresholds p_t and p_b to 1 or 2 pixels consistently yields good results.

[2] https://github.com/sniklaus/pytorch-pwc

Fig. 3: Comparison of crop of reference image (left), with the synthesized view from the neighboring camera, with rectification refinement (middle). Without refinement (right), the details are blurry.

4. DEPTH-GUIDED VIEW SYNTHESIS

We propose an adaptation of the learning-based framework provided by Kalantari *et al.* [7], and equip it to work for a wider range of baselines and subjects than originally trained for, without the need for retraining. The framework consists of two convolutional neural networks: one for estimating the target disparity from input images, and one for estimating the target color image from input images and estimated target disparity. Our first contribution addresses the warping strategy. The authors use standard backward warping throughout the process because of its differentiability. Since the networks are trained end-to-end and optimized for view synthesis, the system learns to account for artefacts by predicting false disparity values around boundaries. We found that by using our warping strategies, both the target disparity map and color

images gain in quality.

Even with the above precautions, reconstructing the geometry of faces remains challenging because of symmetries, homogeneous regions and complicated view-dependent effects. To overcome this, we propose a novel prior-guided feature extraction scheme. Denote by $p_i \in \mathbb{R}^2$ the position of camera i, lying in a common 2D plane after rectification. Our goal is to generate the view L_q at position $q \in \mathbb{R}^2$, given the input images L_{p_i} and the input disparities D_{p_i}. The disparities are obtained from (2) and normalized with respect to the baseline such that $D_{p_i q} = (p_i - q)\, D_{p_i}$. Instead of warping the input images to $L = 100$ uniform disparity values in a fixed interval, we propose a prior-guided warping scheme which yields a higher resolution in the disparity ranges of interest and thus leads to better view synthesis performance. We warp the input images to L disparity values d_l, sampled uniformly from a window of size K around the prior. Introducing $D_{p_i q}(d_l) = D_{p_i q} + d_l$, the contrast features are given by

$$L_q^l = \frac{1}{N} \sum_{i=1}^{N} g\left(D_{p_i q}(d_l), L_{p_i}\right), \text{for } l = 1 \dots L \qquad (3)$$

where N is the number of input cameras. The variance features are deduced in the same way.

Since the disparity and color networks were trained for disparities between -21 and 21 pixels, their output and input, respectively, are linearly mapped to the correct range.

5. EXPERIMENTAL RESULTS

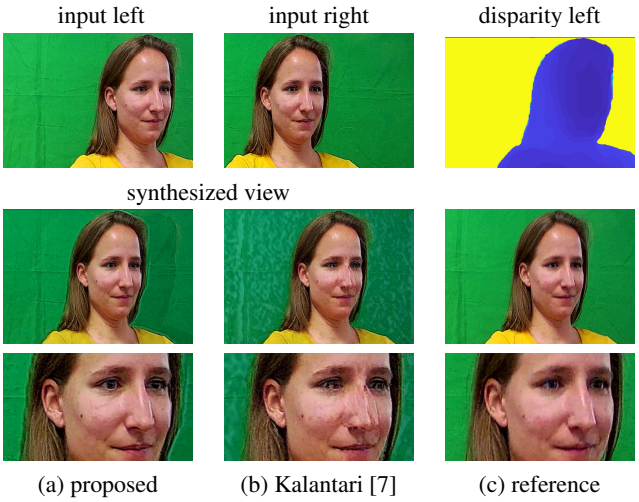
We perform experiments with a linear camera rig made of 16 surveillance cameras and a single ingest 4 TB storage system. The 16 channel networked video recorder (NVR) sustains up to 4 K resolution at 30 fps. However, the particular cameras used are 5 MP varifocal zoom 2.8 mm-12 mm. An optical synchronization LED signal is used to align frames temporally before processing by the synthesis pipeline. Two adjacent cameras are on average 7 cm apart.

In Table 1, we quantitatively evaluate the disparity estimation with and without bidirectional refinement for two different baselines: narrow (adjacent cameras) and wide (2 cameras apart). We use our improved forward and backward warping schemes and also compare to the standard artefact-prone schemes. The error measure is the mean absolute error (MAE) between warped and target image. Our method clearly improves on classical warping schemes, and refinement is always beneficial; less so for narrow-baseline than for wide-baseline. We also compare the view synthesis qualitatively in Figure 3. When no refinement is performed, the synthesized image looses in detail, visible in particular in textured regions, such as the lines in the sweater.

In a second experiment, we sample the light field at arbitrary linear density. Two sample results are shown in Figure 1. The only visible artefacts are in the background, for which the disparity prior was set to zero. We observe a slight tone change, which is however consistent across the generated views and thus not harmful for real-world applications.

	forward		backward		
	narrow	wide	narrow	wide	
Ours w. R	**3.65**	**7.10**	**3.52**	**6.92**	× e-02
Ours w/o R	3.66	7.27	3.62	7.14	
Standard	4.12	8.34	7.73	17.1	

Table 1: Mean absolute error of refinement (R) performance, for forward and backward warping and two different baselines.



(a) proposed (b) Kalantari [7] (c) reference

Fig. 4: Results of our view synthesis method compared to Kalantari [7]. Given the wide-baseline rectified input images, we synthesize the middle image using the disparities as guides.

Finally, we compare the synthesized image with a reference image at the same location. One such result is shown in Figure 4. We compare our synthesis to the method by Kalantari [7], where we adapt their disparity range to reasonable values for fairness. While the algorithm performs well at the subject boundaries, it fails to correctly reconstruct the facial details. Our method shows no visible artefacts at the boundaries or in the face: even though the disparity prior is only coarse, the result is sharp and even fine details such as the eyes are reconstructed realistically.

6. CONCLUSION

We have proposed a system for light-field reconstruction from inexpensive camera arrays. While we focus on prior information from optical flow, the method can be adapted to depth information provided by monocular depth estimation or depth sensors. In future work we plan to address the temporal consistency of the reconstruction, and to exploit redundancy of consecutive frames to speed up the reconstruction process.

7. REFERENCES

[1] Jean Yves Guillemaut and Adrian Hilton, "Joint multi-layer segmentation and reconstruction for free-viewpoint video applications," *International Journal of Computer Vision*, vol. 93, no. 1, pp. 73–100, 2011.

[2] "Lytro," `https://en.wikipedia.org/wiki/Lytro`, Accessed: 2019-02-05.

[3] "Raytrix - 3d light field camera technology," `www.raytrix.de`, Accessed: 2019-02-05.

[4] Ryan S. Overbeck, Daniel Erickson, Daniel Evangelakos, and Paul Debevec, "The making of welcome to light fields vr," in *ACM SIGGRAPH 2018 Talks*, New York, NY, USA, 2018, pp. 63:1–63:2.

[5] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross, "High-quality single-shot capture of facial geometry," *ACM Transactions on Graphics*, vol. 29, no. 4, pp. 1, 2010.

[6] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy, "High performance imaging using large camera arrays," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 765–776, 2005.

[7] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi, "Learning-Based View Synthesis for Light Field Cameras," *ACM Transactions on Graphics*, vol. 35, no. 6, 2016.

[8] Charalampos Koniaris, Maggie Kosek, David Sinclair, and Kenny Mitchell, "Compressed animated light fields with real-time view-dependent reconstruction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 4, pp. 1166–1680, 2018.

[9] Robert C Bolles, H Harlyn Baker, and David H Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International journal of computer vision*, vol. 1, no. 1, pp. 7–55, 1987.

[10] Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu, "Light Field Image Processing: An Overview," *IEEE Journal on Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 926–954, 2017.

[11] Changil Kim, Henning Zimmer, Yael Pritch, Alexander Sorkine-Hornung, and Markus Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Transactions on Graphics*, vol. 32, no. 4, 2013.

[12] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim, "EPINET: A Fully-Convolutional Neural Network Using Epipolar Geometry for Depth from Light Field Images," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018.

[13] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski, "Layered Depth Images," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, New York, 2009, pp. 231–242, ACM New York.

[14] Shubham Tulsiani, Richard Tucker, and Noah Snavely, "Layer-structured 3D Scene Inference via View Synthesis," *arXiv preprint arXiv:1807.10264*, 2018.

[15] Helisa Dhamo, Keisuke Tateno, Iro Laina, Nassir Navab, and Federico Tombari, "Peeking Behind Objects: Layered Depth Prediction from a Single Image," *arXiv preprint arXiv:1807.08776*, 2018.

[16] Eric Penner and Li Zhang, "Soft 3D reconstruction for view synthesis," *ACM Transactions on Graphics*, vol. 36, no. 6, 2017.

[17] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely, "DeepStereo: Learning to Predict New Views from the World's Imagery," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5515–5524.

[18] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely, "Stereo magnification: learning view synthesis using multiplane images," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 65, 2018.

[19] Andrea Fusiello, Emanuele Trucco, and Alessandro Verri, "A compact algorithm for rectification of stereo pairs," *Machine Vision and Applications*, vol. 12, no. 1, pp. 16–22, 2000.

[20] Richard Hartley and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[21] Vincent Nozick, "Camera array image rectification and calibration for stereoscopic and autostereoscopic displays," *Annals of Telecommunications, Springer*, vol. 68, no. 11, pp. 581–596, 2013.

[22] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz, "PWC-Net: CNNs for Optical Flow using Pyramid, Warping, and Cost Volume," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943.

[23] Vincent Jantet, *Layered depth images for multi-view coding*, Ph.D. thesis, Université Rennes 1, 2012.

[24] Leonard McMillan, "A list-priority rendering algorithm for redisplaying projected surfaces," Tech. Rep., Chapel Hill, NC, USA, 1995.

[25] Beerend Ceulemans, Shao-Ping Lu, Gauthier Lafruit, and Adrian Munteanu, "Robust multiview synthesis for wide-baseline camera arrays," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2235–2248, 2018.