Open access • Proceedings Article • DOI:10.1109/ICIP.2018.8451466

Depth Estimation with Occlusion Handling from a Sparse Set of Light Field Views

Xiaoran Jiang, Mikael Le Pendu, Christine Guillemot

Institutions: French Institute for Research in Computer Science and Automation

Published on: 07 Oct 2018 - International Conference on Image Processing

Topics: Depth map, View synthesis and Light field

Depth estimation with occlusion handling from a sparse set of light field views

Xiaoran Jiang, Mikaël Le Pendu, Christine Guillemot

DEPTH ESTIMATION WITH OCCLUSION HANDLING FROM A SPARSE SET OF LIGHT FIELD VIEWS

Xiaoran Jiang, Mikaël Le Pendu, Christine Guillemot*

INRIA, Rennes, France

ABSTRACT

This paper addresses the problem of depth estimation for every viewpoint of a dense light field, exploiting information from only a sparse set of views. This problem is particularly relevant for applications such as light field reconstruction from a subset of views, for view synthesis and for compression. Unlike most existing methods for scene depth estimation from light fields, the proposed algorithm computes disparity (or equivalently depth) for every viewpoint taking into account occlusions. In addition, it preserves the continuity of the depth space and does not require prior knowledge on the depth range. The experiments show that, both for synthetic and real light fields, our algorithm achieves competitive performance to state-of-the-art algorithms which exploit the entire light field and usually generate the depth map for the center viewpoint only.

***Index Terms*—** depth estimation, light field, stereo matching, optical flow, low rank approximation

1. INTRODUCTION

Light fields, by capturing light rays emitted by a 3D scene along different orientations, give a very rich description of the scene enabling a variety of computer vision applications. The recorded 4D light field can also be regarded as an arrays of views giving information about the parallax and depth of the scene. Existing methods for depth estimation from light fields can be classified into several main categories: methods based on sub-aperture images (SAI), on epipolar plane images (EPI) or on refocused images. The methods based on SAI compute matches between the extracted views, assuming that they are well rectified with a constant baseline [1, 2]. The authors in [1] use robust PCA to estimate the disparity which will minimize the rank of the matrix containing all the views warped on the center one. A method is described in [2] which instead computes a cost volume based on the similarity between sub-aperture images and the center view shifted at sub-pixel locations to evaluate the matching cost of different disparity labels. The authors in [3] apply an optical flow estimator on a sequence of light field views along an angular dimension to estimate several disparity maps which are then aggregated to create a single disparity map which is then converted into a depth map.

Another type of methods for plenoptic depth estimation uses EPIs [4, 5]. Indeed, the slope of the line composed of the corresponding pixel in an EPI is proportional to the depth of the pixel [6]. The authors in [4] use structure tensors to estimate the local slopes which are then regularized using a variational labeling framework for global consistency, while a spinning parallelogram operator is proposed in [5] to estimate the slopes of these structures. A third category of methods uses images in a focal stack and use defocus cues, possibly combined with other measures, to estimate depth [7, 8]. This relies on the assumption that in-focus points are projected at the same spatial position in the different views.

However, most of these methods require that the light field is densely sampled. Furthermore, the methods using cost volumes or those that consider depth estimation as a multi-labelling problem require discretizing the depth space. In order to keep the computational cost within a manageable limit, the number of discretization levels in these methods should be kept low at the expense of accuracy. Finally, most methods only compute the depth map of the center view, hence do not give geometry information on pixels which are not visible from this viewpoint, yet this information may be required for problems such as light field reconstruction, or view synthesis.

In this paper, we propose a novel depth estimation algorithm exploiting only a sparse subset of light field views. The method computes a disparity (or equivalently depth) map for any light field viewpoint, hence is particularly interesting for applications such as light field reconstruction or view synthesis from a subset of views and for compression [9]. Unlike [2,5,7,8], the proposed method does not demand discretization of the disparity space, nor prior knowledge about the disparity range. According to the metrics defined in [10, 11], the experiments show that our approach achieves competitive performance compared to state-of-the-art methods that make use of the whole set of light field views.

2. ALGORITHM OVERVIEW

We consider the 4D representation of light fields defined by a function $L(x, y, u, v)$ of 4 parameters at the intersection of the light rays with 2 parallel planes. The light field can be seen as capturing an array of viewpoints (called sub-aperture images and denoted by $L_{u,v}$) of the scene with varying angular coordinates $u \in [\![1 \,..\, U]\!]$ and $v \in [\![1 \,..\, V]\!]$. In this paper, we propose to estimate one consistent and accurate disparity map per position

*Mikaël Le Pendu is now with Trinity College Dublin.

This work has been funded by the EU H2020 Research and Innovation Programme under grant agreement No 694122 (ERC advanced grant CLIM).

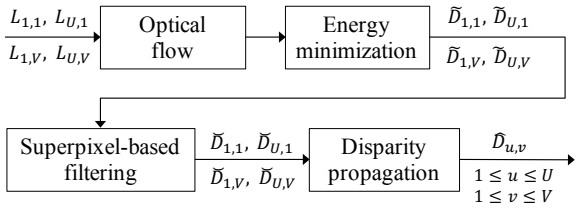


Fig. 1: Algorithm pipeline.

(u, v), given only a sparse subset of light field views. This disparity map can then be easily converted into a depth map.

The algorithm proceeds as follows (*cf.* Fig. 1). The four corner views $L_{1,1}, L_{U,1}, L_{1,V}$ and $L_{U,V}$ are taken as input, since these views contain the geometry of the whole scene, accounting for disocclusions. Multiple rough disparity estimates for these input views are first computed using an optical flow estimator. These candidates are then aggregated by using an energy minimization and the resulting map per input view is further enhanced by using a novel superpixel-based bilateral filtering. These refined disparity maps are warped to novel viewpoints and a global inpainting using low rank approximation is performed to fill the holes. Finally, at each novel viewpoint, warped candidates are merged with respect to their disparity uncertainty via a winner-take-all process.

3. DISPARITY ON REFERENCE VIEWS

3.1. Optical flow-based initialization

Let $\mathcal{R} = \{\mathbf{r}\}$ denote the set of input viewpoints, with $\mathbf{r} = (u_r, v_r)$. Horizontal and vertical disparity maps are computed between each input view $L_{\mathbf{r}}$ and each of the other views $L_{\mathbf{r}'}$, with $\mathbf{r}' \in \mathcal{R} \setminus \mathbf{r}$, using an optical flow estimator,

$$d_{\mathbf{r}\to\mathbf{r}'} = \begin{pmatrix} d^X_{\mathbf{r}\to\mathbf{r}'} \\ d^Y_{\mathbf{r}\to\mathbf{r}'} \end{pmatrix} = OpticalFlow(L_{\mathbf{r}}, L_{\mathbf{r}'}). \tag{1}$$

$d^X_{\mathbf{r}\to\mathbf{r}'}$ and $d^Y_{\mathbf{r}\to\mathbf{r}'}$ are respectively the horizontal and vertical disparity from the view $L_{\mathbf{r}}$ to $L_{\mathbf{r}'}$. In the experiments, we used Epicflow [12], however, any state-of-the-art optical flow estimator can be used.

Assuming that the light field views are well rectified, a scene point moves only horizontally from position $(1,1)$ to $(U,1)$, and only vertically from position $(1,1)$ to $(1,V)$: $d^X_{(1,1)\to(1,V)} = 0$ and $d^Y_{(1,1)\to(U,1)} = 0$. For simplicity and without loss of generality, we also suppose the 4D light field is a square grid of regularly spaced views, i.e. $U = V$ and the horizontal and vertical baseline between views is uniform. Therefore, a set of 4 estimates that reveal the same disparity information is obtained at view $L_{1,1}$: $\mathcal{C}_{1,1} = \{d^X_{(1,1)\to(U,1)}, d^Y_{(1,1)\to(1,V)}, d^X_{(1,1)\to(U,V)}, d^Y_{(1,1)\to(U,V)}\}$. These maps are normalized by dividing them by $(U-1)$ (where U is the number of views in each dimension) such that they represent disparity between adjacent views, assuming a constant baseline. The sets of estimated maps $\mathcal{C}_{U,1}$, $\mathcal{C}_{1,V}$, $\mathcal{C}_{U,V}$ for the other input views are computed in a similar manner. In the sequel, we let $d^i_{\mathbf{r}}$ denote the i^{th} estimated disparity map in the set $\mathcal{C}_{\mathbf{r}}$.

3.2. Energy minimization

As optical flow estimators only work on pairs of images, which do not fully exploit angular diversity of all input views, each pair gives a rough disparity estimate. To have a more reliable estimate $\tilde{D}_{\mathbf{r}}, \mathbf{r} \in \mathcal{R}$, at each pixel $\mathbf{p}$, one disparity value is selected among all candidates $d^i_{\mathbf{r}}(\mathbf{p})$ by minimizing the energy E

$$\tilde{D}_{\mathbf{r}}(\mathbf{p}) = \underset{d^i_{\mathbf{r}}(\mathbf{p})}{\operatorname{argmin}}\, E(d^i_{\mathbf{r}}, \mathbf{p}), \tag{2}$$

where $E(d^i_{\mathbf{r}}, \mathbf{p})$ denotes the energy value at pixel $\mathbf{p}$ computed using $d^i_{\mathbf{r}}$. These energy values for all pixels form a so-called energy map $E(d^i_{\mathbf{r}})$ expressed as

$$E(d^i_{\mathbf{r}}) = E_c(d^i_{\mathbf{r}}) + \lambda_1 E_g(d^i_{\mathbf{r}}) + \lambda_2 E_s(d^i_{\mathbf{r}}), \tag{3}$$

The energy term E_c is a color consistency term computed between $L_{\mathbf{r}}$ and the projected views as

$$E_c(d^i_{\mathbf{r}}) = \sum_{\mathbf{r}'\in\mathcal{R}\setminus\mathbf{r}} \left(M^{\mathbf{r}',i}_{\mathbf{r}} \odot \mathcal{E}_1(L_{\mathbf{r}}, L^{\mathbf{r}',i}_{\mathbf{r}}) \right) \oslash \sum_{\mathbf{r}'\in\mathcal{R}\setminus\mathbf{r}} M^{\mathbf{r}',i}_{\mathbf{r}}, \tag{4}$$

where $\mathcal{E}_1(I, I')$ is defined as the pixel-wise sum of square errors for the three color components: $\mathcal{E}_1(I, I') = \sum_{C\in\{R,G,B\}} (I_C - I'_C)^2$. The symbols $\oslash$ and $\odot$ denote respectively pixel-wise division and Hadamard product. $L^{\mathbf{r}',i}_{\mathbf{r}}$ stands for the warped image from position $\mathbf{r}'$ to $\mathbf{r}$ by using the i^{th} disparity candidate $d^i_{\mathbf{r}}$. Here, backward warping is applied, and horizontal and vertical disparities between $\mathbf{r}$ and $\mathbf{r}'$ are obtained by multiplying the normalized disparity $d^i_{\mathbf{r}}$ by the angular position offset $\mathbf{r}' - \mathbf{r}$. $M^{\mathbf{r}',i}_{\mathbf{r}}$ is the corresponding binary mask discarding the disoccluded pixels from the energy summation. If one pixel $\mathbf{p}$ falls into holes of all the warped images using one candidate disparity map $d^i_{\mathbf{r}}$, (i.e. $\sum_{\mathbf{r}'\in\mathcal{R}\setminus\mathbf{r}} M^{\mathbf{r}',i}_{\mathbf{r}}(\mathbf{p}) = 0$) the candidate is discarded for that pixel by setting $E_c(d^i_{\mathbf{r}}, \mathbf{p})$ to infinity.

The second term E_g is a gradient consistency term computed between the reference view and the warped views:

$$E_g(d^i_{\mathbf{r}}) = \sum_{\mathbf{r}'\in\mathcal{R}\setminus\mathbf{r}} \left(M^{\mathbf{r}',i}_{\mathbf{r}} \odot \mathcal{E}_2(L_{\mathbf{r}}, L^{\mathbf{r}',i}_{\mathbf{r}}) \right) \oslash \sum_{\mathbf{r}'\in\mathcal{R}\setminus\mathbf{r}} M^{\mathbf{r}',i}_{\mathbf{r}}, \tag{5}$$

where $\mathcal{E}_2$ denotes the pixel-wise square error on the gradients: $\mathcal{E}_2(I, I') = (\nabla_x I - \nabla_x I')^2 + (\nabla_y I - \nabla_y I')^2$. Finally, the smoothness term E_s

$$E_s(d^i_{\mathbf{r}}) = \sum_{\mathbf{r}'\in\mathcal{R}\setminus\mathbf{r}} \sqrt{\mathcal{E}_2(d^i_{\mathbf{r}}, L_{\mathbf{r}}) \odot \left(|\nabla_x d^i_{\mathbf{r}}|^2 + |\nabla_y d^i_{\mathbf{r}}|^2 \right)} \tag{6}$$

helps preserving edges. In fact, if an edge (large gradient) in the disparity map is misaligned with the corresponding one in the color image, the term $\mathcal{E}_2(d^i_{\mathbf{r}}, L_{\mathbf{r}})$ should be high. Otherwise, when the two edges are aligned, E_s does not penalize the energy in spite of the large gradient of the disparity. In practice, to bring the disparity map and the color image to the same scale, their values are both normalized between 0 and 1.

A confidence measure of the selected disparity values can be deduced based on the minimum energy:

$$\tilde{F}_{\mathbf{r}} = \exp\left(-\frac{\min_i E(d^i_{\mathbf{r}})}{2\sigma_e^2}\right) \quad (7)$$

where σ_e controls the "width" of the distribution.

3.3. Superpixel-based edge-preserving filtering

The previous energy-based voting approach is fast and efficient to select the most likely disparity value among all candidates. Nevertheless, the resulting map may benefit from further enhancement due to the fact that: 1/- none of the candidate values may be correct; 2/- the disocclusion masks used in the energy computation (Eq. 4-5) may be erroneously estimated due to disparity inaccuracy.

To enhance the disparity estimate for each input view, we propose a novel *superpixel-based edge-preserving filtering*. We first identify the pixels with the lowest 5% confidence measure as the set of pixels $\Omega_{\mathbf{r}}$ for which the disparity is potentially wrong. A filtering of these unreliable disparities is performed by computing a weighted average of reliable nearby values:

$$\forall \mathbf{p} \in \Omega_{\mathbf{r}}, \breve{D}_{\mathbf{r}}(\mathbf{p}) = \frac{1}{Z_{\mathbf{p}}} \sum_{\mathbf{q} \in \mathcal{N}_{\mathbf{p}} \cap \overline{\Omega}_{\mathbf{r}}} w_{\mathbf{p},\mathbf{q}} \tilde{D}_{\mathbf{r}}(\mathbf{q}), \quad (8)$$

where $Z_{\mathbf{p}}$ is a normalization factor

$$Z_{\mathbf{p}} = \sum_{\mathbf{q} \in \mathcal{N}_{\mathbf{p}} \cap \overline{\Omega}_{\mathbf{r}}} w_{\mathbf{p},\mathbf{q}}. \quad (9)$$

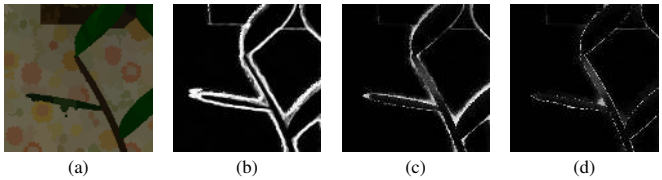
As for the bilateral filter, the weights $w_{\mathbf{p},\mathbf{q}}$ are defined as a function of a spatial G_{σ_s} and photometric G_{σ_c} kernel as:

$$\begin{aligned} w_{\mathbf{p},\mathbf{q}} &= G_{\sigma_s}(\|\mathbf{p}-\mathbf{q}\|).G_{\sigma_c}(\|L(\mathbf{p})-L(\mathbf{q})\|) \\ &= \exp(-\frac{\|\mathbf{p}-\mathbf{q}\|}{2\sigma_s^2} - \frac{\|L(\mathbf{p})-L(\mathbf{q})\|}{2\sigma_c^2}). \end{aligned} \quad (10)$$

However, unlike classical image-guided bilateral filtering for which the pixel neighborhood $\mathcal{N}_{\mathbf{p}}$ is usually a square window centered in $\mathbf{p}$, in the proposed filter, the neighborhood is defined by superpixels assuming that pixels inside a superpixel are likely to have close depth. In our experiments, SLIC [13] implementation is used. To best adapt the size of the neighborhood to the reliability of disparity values, a fine over-segmentation in superpixels is first performed and if a superpixel s_i contains less than 50% of reliable disparity values, then it is merged to the most similar neighbor superpixel $s_s \in N_{s_i}$, N_{s_i} being the neighborhood of s_i. The most similar superpixel is chosen by the following minimization

$$s_s = \underset{s_j \in N_{s_i}}{\operatorname{argmin}} \left\|\mu(s_i) - \mu(s_j)\right\| + \left\|var(s_i) - var(s_j)\right\|, \quad (11)$$

where the mean color μ and the variance var are both calculated in the CIELAB color space. An example showing the effectiveness of our superpixel-based filtering is given in Fig. 2.



(a) (b) (c) (d)

Fig. 2: An example of superpixel-based edge-preserving filtering. (a) Superpixel over-segmentation; (b) Uncertainty ($1-\tilde{F}$) measured on the unfiltered map; (c) Error on the unfiltered disparity map with respect to the ground truth; (d) Error after filtering with respect to the ground truth.

4. DISPARITY PROPAGATION

The refined disparity map $\breve{D}_{\mathbf{r}}$ ($\mathbf{r} \in \mathcal{R}$) thus obtained for each input view $L_{\mathbf{r}}$ is projected (forward warping) to the novel position $\mathbf{s} \in [\![1 \;..\; U]\!] \times [\![1 \;..\; V]\!]$ by using the disparity information itself. Thus, at each position $\mathbf{s}$, there are four pairs of warped maps and corresponding inpainting masks $(\breve{D}^{\mathbf{r}}_{\mathbf{s}}, \breve{M}^{\mathbf{r}}_{\mathbf{s}})$. We thus construct the matrix $\breve{H}$ of $4 \times U \times V$ columns, each column being a vectorized warped disparity map with holes: $\breve{H} = \left[\text{vec}(\breve{D}^{\mathbf{r_1}}_{\mathbf{s_1}}) \,|..|\, \text{vec}(\breve{D}^{\mathbf{r_4}}_{\mathbf{s_1}}) \,|..|\, \text{vec}(\breve{D}^{\mathbf{r_1}}_{\mathbf{s_N}}) \,|..|\, \text{vec}(\breve{D}^{\mathbf{r_4}}_{\mathbf{s_N}})\right]$ ($N = U \times V$). The mask matrix is defined in the similar way: $\breve{M} = \left[\text{vec}(\breve{M}^{\mathbf{r_1}}_{\mathbf{s_1}}) \,|..|\, \text{vec}(\breve{M}^{\mathbf{r_4}}_{\mathbf{s_1}}) \,|..|\, \text{vec}(\breve{M}^{\mathbf{r_1}}_{\mathbf{s_N}}) \,|..|\, \text{vec}(\breve{M}^{\mathbf{r_4}}_{\mathbf{s_N}})\right]$. Given that the warped disparity maps are highly correlated, $\breve{H}$ can be efficiently inpainted using a matrix completion method which formalizes the problem as

$$\begin{aligned} &\min_{\hat{H}} \text{rank}(\hat{H}) \\ &\text{s.t. } P_{\breve{M}}(\hat{H}) = P_{\breve{M}}(\breve{H}), \end{aligned} \quad (12)$$

where $P_{\breve{M}}$ is the sampling operator such that $P_{\breve{M}}(H)_{i,j}$ is equal to $H_{i,j}$ if $\breve{M}_{i,j} = 1$, and zero otherwise.

The low rank matrix completion is solved using the Inexact ALM (IALM) method [14]. The inpainting works well in practice because the disocclusions in the different warped views are unlikely to overlap. The inpainting is globally performed by processing all the view positions at the same time. While the superpixel-based filtering enhances spatial coherence in the disparity of the input views, angular correlation is exploited here.

After inpainting, four disparity maps $(\hat{D}^{\mathbf{r}}_{\mathbf{s}}, \mathbf{r} \in \mathcal{R})$ per view are extracted from the matrix $\hat{H}$. In order to obtain a unique disparity map $\hat{D}_{\mathbf{s}}$ per view, a pixel-wise winner-take-all selection is performed based on the confidence values $\hat{F}^{\mathbf{r}}_{\mathbf{s}}$:

$$\begin{aligned} \forall \mathbf{s}, \forall \mathbf{p}, \mathbf{r}_{\text{win}} &= \underset{\mathbf{r} \in \mathcal{R}}{\operatorname{argmax}}\ \hat{F}^{\mathbf{r}}_{\mathbf{s}}(\mathbf{p}), \\ \hat{D}_{\mathbf{s}}(\mathbf{p}) &= \hat{D}^{\mathbf{r}_{\text{win}}}_{\mathbf{s}}(\mathbf{p}). \end{aligned} \quad (13)$$

Contrary to the reference view where the color information can be exploited to measure the confidence (*cf.* Section 3.2), here $\hat{F}^{\mathbf{r}}_{\mathbf{s}}$ is inferred by projecting (forward warping) the confidence $\tilde{F}_{\mathbf{r}}$ from the reference view $\mathbf{r}$ to the target view $\mathbf{s}$.

Finally, a step of total variation regularization (TV-L1) using the primal-dual algorithm [15] is applied on the different epipolar slice images of the resulting disparity maps in order to enforce view consistency.

Table 1: Quality evaluation of the estimated disparity maps on center view. Our method (4 corner views input) is compared against two state-of-the-art methods (4D full LF input), namely *LF [2] and SPO [5]. The best results are marked in **bold**.

	MSE*100			BadPix(0.01)			BadPix(0.03)			BadPix(0.07)			Q25		
Light fields	[2]	[5]	Ours	[2]	[5]	Ours	[2]	[5]	Ours	[2]	[5]	Ours	[2]	[5]	Ours
StillLife	2.02	**1.72**	2.56	81.2	76.2	**71.3**	51.0	32.1	**25.0**	20.9	**6.8**	9.2	1.36	1.02	**0.87**
Buddha	1.13	0.97	**0.82**	57.7	41.2	**34.9**	24.4	14.8	**12.3**	10.1	6.7	**5.4**	0.51	0.34	**0.31**
MonasRoom	0.76	0.58	**0.53**	46.0	42.5	**38.6**	22.1	**17.8**	18.6	11.7	**7.8**	8.2	0.38	0.34	**0.33**
Butterfly	4.79	**0.74**	1.84	82.5	78.9	**70.8**	49.1	48.5	**36.0**	15.4	14.1	**6.7**	1.47	1.22	**0.85**
Boxes	14.15	**8.23**	12.71	72.7	**62.3**	65.8	45.5	**28.1**	37.7	26.4	**15.8**	23.9	0.89	**0.62**	0.68
Cotton	9.98	1.44	**1.18**	60.5	**41.7**	42.6	23.3	11.1	**10.7**	8.9	**2.7**	4.1	0.59	**0.36**	0.42
Dino	1.23	**0.29**	0.88	76.6	57.5	**49.1**	48.4	**17.9**	20.0	20.9	**3.4**	9.5	1.08	0.55	**0.43**
Sideboard	4.16	**0.92**	10.31	67.8	64.3	**61.7**	39.3	**31.0**	37.5	23.0	**10.4**	19.6	0.73	0.66	**0.51**
Average	4.78	**1.86**	3.85	68.1	58.1	**54.4**	37.9	25.2	**24.7**	17.2	**8.5**	10.8	0.88	0.64	**0.55**

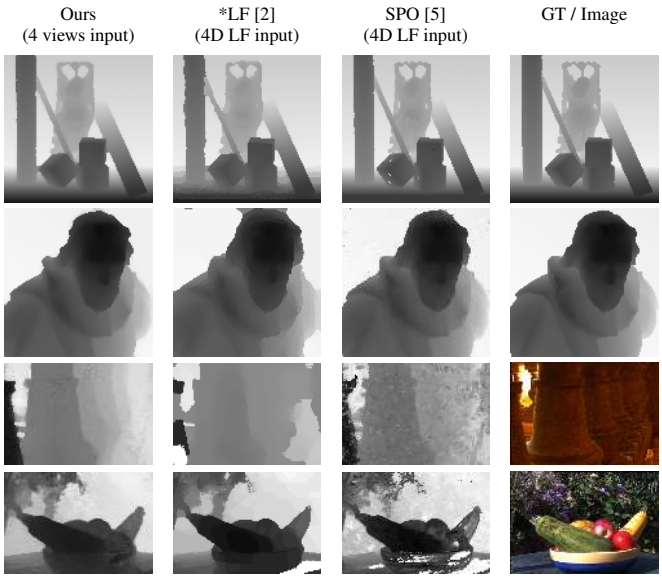


Fig. 3: Visual comparison of the estimated disparity maps on center view. The first 2 rows are synthetic LFs "Buddha" and "Cotton", the last 2 rows are Illum real LFs "Stone_Pillars_Outside" and "Fruits".

5. PERFORMANCE ASSESSMENT

The performance of our method has been assessed using different datasets: both the synthetic LFs from the HCI datasets [10, 16] and the real ones captured by Lytro Illum cameras [17, 18] are considered. In our experiments, the center 7×7 sub-aperture views are considered for all test LFs. We have set the parameters as follows: $\lambda_1 = 2$, $\lambda_2 = 2$, $\sigma_e = 0.1$, $\sigma_s = 2$ and $\sigma_c = 1$.

Synthetic data. Ground truth disparity on center view is available for HCI synthetic light fields. Using the evaluation metrics defined in [10,11], we compare in Table 1 our scheme against two state-of-the-art methods, namely *LF [2] and SPO [5]. In our experiments, the number of discretized disparity levels is kept the same as in [11], i.e. 100 for *LF and 256 for SPO. In order to achieve satisfying results, both methods require a high sampling rate on the angular dimension. Despite the fact that our method is disadvantaged as we only exploit the corner views and infer the disparity for other views, our method achieves significantly better performance than *LF, and comparable results to SPO. Especially, our method is the best among the three evaluated methods for BadPix(0.01), BadPix(0.03) and Q25 which indicates "the best case accuracy" of a given algorithm. One of the main reasons for this gain is that our algorithm preserves continuity of the disparity space. Nevertheless, for scenes with very fine details, such as "Sideboard", the sampling rate (4 input views out of 49 of the light field) turns out to be too low.

Real data. Limited by current technology, real light fields captured by plenoptic imaging devices are often prone to noise and distortion, which significantly compromises the accuracy of many depth estimation algorithms. In Fig. 3, we observe that SPO performs best on object boundary, but fails to provide consistent estimates on noisy homogenous zones. Our algorithm achieves a good balancing between accuracy and robustness. Readers are invited to view more simulation results and animations (estimation on every viewpoint) on the web page `https://www.irisa.fr/temics/demos/lightField/DepthEstim/DepthEstimXR.html`. Our method has been also used to compute disparity maps from video light fields captured with a Raytrix camera [19] (`http://clim.inria.fr/Datasets/RaytrixR8Dataset-5x5/index.html`).

Runtime. Thanks to the low sampling rate, our algorithm is comparatively efficient. Simulations have been carried out on a Macbook Pro with a 2.8GHz Intel Core i7 processor and 16G RAM. For a light field of resolution $7 \times 7 \times 512 \times 512$, it takes approximately 13 mins to estimate scene depth using our method, against 63 mins for SPO and 16 mins for *LF. Moreover, our method generates 49 disparity maps at a time. Note that the code is written in Matlab and could be further optimized in the future.

6. CONCLUSIONS

In this paper, we have proposed a light field depth estimation algorithm using only a small subset of light field views and inferring the depth for other views without exploiting the color information of all the views. The experiments show that, both for synthetic and real light fields, our algorithm achieves competitive performance compared with state-of-the-art algorithms, despite the fact that the algorithm uses much less light field information.

7. REFERENCES

[1] Stefan Heber and Thomas Pock, "Shape from light field meets robust PCA," in *European Conference on Computer Vision (ECCV)*, 2014.

[2] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon, "Accurate depth map estimation from a lenslet light field camera," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] Yang Chen, Martin Alain, and Aljosa Smolic, "Fast and accurate optical flow based depth map estimation from light fields," in *Irish Machine Vision and Image Processing Conference (IMVIP)*, 2017.

[4] Sven Wanner and Bastian Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Transactions of Pattern analysis and machine intelligence*, vol. 36, no. 3, 2013.

[5] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Journal Computer Vision and Image Understanding*, vol. 145, pp. 148–159, 2016.

[6] Robert C. Bolles, H. Harlyn Baker, and David H. Marimont, "Epipolarplane image analysis: An approach to determining structure from motion," *International Journal of Computer Vision*, pp. 1–7, 1987.

[7] Michael W. Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *International Conference on Computer Vision (ICCV)*, 2013.

[8] Ting-Chun Wang, Alexei Efros, and Ravi Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *International Conference on Computer Vision (ICCV)*, 2015.

[9] Xiaoran Jiang, Mikael Le Pendu, and Christine Guillemot, "Light field compression using depth image based view synthesis," in *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2017.

[10] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Asian Conference on Computer Vision (ACCV)*, 2016.

[11] Ole Johannsen, Katrin Honauer, Bastian Goldluecke, Anna Alperovich, Federica Battisti, Yunsu Bok, Michele Brizzi, Marco Carli, Gyeongmin Choe, Maximilian Diebold, Marcel Gutsche, Hae-Gon Jeon, In So Kweon, Alessandro Neri, Jaesik Park, Jinsun Park, Hendrik Schilling, Hao Sheng, Lipeng Si, Michael Strecke, Antonin Sulc, Yu-Wing Tai, Qing Wang, Ting-Chun Wang, Sven Wanner, Zhang Xiong, Jingyi Yu, Shuo Zhang, and Hao Zhu, "A taxonomy and evaluation of dense light field depth estimation algorithms," in *Conference on Computer Vision and Pattern Recognition - LF4CV Workshop*, 2017.

[12] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid, "EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk, "SLIC Superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 2274–2282, 2012.

[14] Zhouchen Lin, Risheng Liu, and Zhixun Su, "Linearized alternating direction method with adaptive penalty for low rank representation," in *Neural Information Processing Systems (NIPS)*, 2011.

[15] Antonin Chambolle, Vicent Caselles, Matteo Novaga, Daniel Cremers, and Thomas Pock, "An introduction to Total Variation for Image Analysis," *hal-00437581*, 2009.

[16] Sven Wanner, Stephan Meister, and Bastian Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *VMV Workshop*, 2013.

[17] "INRIA Lytro image dataset," `https://www.irisa.fr/temics/demos/lightField/LowRank2/datasets/datasets.html`.

[18] "ICME 2016 Grand Challenge dataset," `http://mmspg.epfl.ch/EPFL-light-field-image-dataset`.

[19] Laurent Guillo, Xiaoran Jiang, Gauthier Lafruit, and Christine Guillemot, "Light field video dataset captured by a R8 Raytrix camera (with disparity maps)," *MPEG and JPEG contributions, ISO/IEC JTC1/SC29/WG11 MPEG2018/m42468, ISO/IEC JTC1/SC29/WG1 JPEG2018/m79046*, 2018.