Rapid Light Field Depth Estimation with Semi-Global Matching

Yuriy Anisimov, Oliver Wasenmüller, Didier Stricker
Department Augmented Vision, German Research Center for Artificial Intelligence (DFKI)
Department of Computer Science, University of Kaiserslautern
Kaiserslautern, Germany
{yuriy.anisimov, oliver.wasenmueller, didier.stricker}@dfki.de

***Abstract*—Running time of the light field depth estimation algorithms is typically high. This assessment is based on the computational complexity of existing methods and the large amounts of data involved. The aim of our work is to develop a simple and fast algorithm for accurate depth computation. In this context, we propose an approach, which involves Semi-Global Matching for the processing of light field images. It forms on comparison of pixels' correspondences with different metrics in the substantially bounded light field space. We show that our method is suitable for the fast production of a proper result in a variety of light field configurations.**

I. Introduction

In terms of computer vision a light field, originally described by Gershun [1], can be interpreted as a set of 2-dimensional images, projected from a scene with an equal physical distance (baseline) between the adjacent viewpoints in the horizontal or vertical direction.

Real-world capturing of such a set can be done in different ways. The simplest one involves an ordinary camera on the moving stage, which is shifted on constant length for every viewpoint, producing a 3-dimensional light field. Light fields can be captured by so-called plenoptic camera [2], which has a micro-lens array placed in front of the sensor image plane, providing a 4-dimensional light field. Such cameras were presented by Ng *et al.* [3] and Perwass *et al.* [4]. Multi-camera arrays with individual lenses can be also used for light field capturing. A large-scale version of this camera type was proposed by Wiburn *et al.* [5], whereas small-scale versions were presented by Venkataraman *et al.* [6] and Anisimov *et al.* [7]

Light field cameras are attracting considerable interest due to the possibility of their utilization in different industrial (optical inspection), biological (three-dimensional microscopy) and cinematic applications. This demand can be explained by the features of light field images, *e.g.* digital refocusing, which is an ability to change the focus on already acquired images [3]. Another useful feature of light field images is a possibility of accurate depth estimation. In contrast to classical multi-view systems, the matching correspondence search can be simplified by preserving the constant baseline between views in light field images.

An overview of the actual light field depth estimation algorithms is provided in Section II-A. We can state a major challenge in this area as a trade-off between runtime and the depth quality. Existing approaches are computationally demanding and involve large amounts of data for processing. It limits rapid depth map calculation. Most studies have tended to focus on depth map accuracy rather than execution time. We mainly address the runtime issue and aim towards real-time processing.

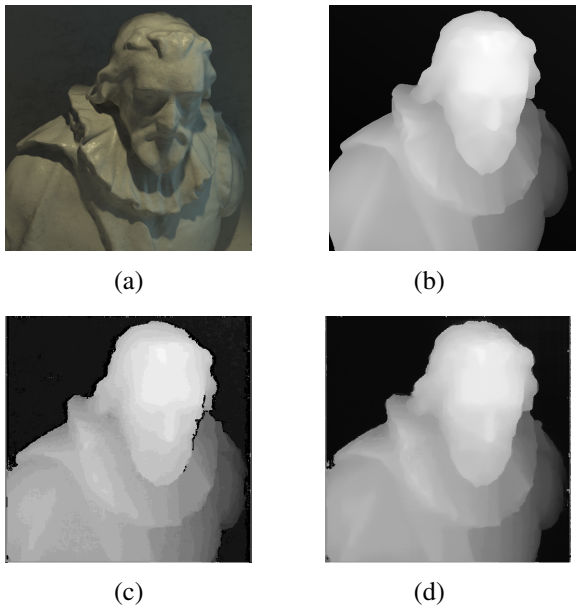

Fig. 1. Result of our algorithm for a synthetic scene from 4D Light Field Benchmark [8], [9]: (a) Center image from a light field, (b) ground truth, (c) initial disparity map, (d) final disparity map.

The proposed algorithm is based on pixel matching in light field space with different similarity measurements for estimation of a dense depth map. We choose Census and ℓ_2-norm for generation of matching cost, which is explained in sections III-B and III-C. Although this principle gives good results in terms of depth map quality, computational demand and consequential running time of this method do not allow its direct use for fast or even real-time estimation. In order to reduce the number of calculations for this step, we generate the initial disparity map from some of the views in light field image and use this information as "boundaries" for high-intensive correspondence search among all views, which is outlined in Section III-D. For disparity quality improvement a semi-global matching (SGM) method [10] is applied for

978-1-7281-4914-1/19/$31.00

previously generated costs. This method recommended itself as one of the most developed and optimized for providing accurate real-time depth results, which is fully justified by works reviewed in Section II-B.

The presented approach follows the core idea of Anisimov *et al.* [7], [11], but introduces the following features. We consider our work to be the first attempt to apply the SGM method directly to matching costs from light field images. For initial disparity map calculation we use a fusion of disparity maps for different views w.r.t. reference view and do not apply refinement to it, unlike in [11]. Together with that, we propose the utilization of Census transform for RGB images, in contrast to classical grayscaled images. As a result of all these contributions, we overcome limitations of previous algorithm [11] such as objects' boundaries sharpness and noise in discontinuity areas. The workflow of our method is presented in Section III.

This approach achieves almost the best running time among other methods together with acceptable depth quality level. We provide a verification of our results by 4-dimensional Light Field Benchmark [8], [9] in Section IV. Together with that, we present real-world qualitative results. On top of the benchmark estimators, we use a M-metric, proposed in [11], to evaluate algorithms in terms of efficiently estimated pixels per unit of time. Also, we show the usage of different image similarity measurements methods for matching cost estimation dependently of various light field configuration properties.

II. Related work

A. Light field processing

Several light field depth estimation approaches utilize Epipolar Plane Image (EPI) structure, proposed by Bolles *et al.* [12] and defined as a "slice" of the light field, in which the slope of the line with reference to a certain point in an image is proportional to the depth value. Wanner and Goldluecke present a reformulation of stereo matching to a constrained labeling problem on EPIs with further variational regularization [13]. Kim *et al.* [14] present an approach for depth reconstruction from high-resolution images using a slope analysis for the lines in EPIs. Wang *et al.* [15] extend this approach for the occlusions-handling case. Johannsen *et al.* [16] present a method based on a sparse decomposition of a light field with further depth-orientation dependency retrieving. Sheng *et al.* [17] extract multi-orientation EPIs and aggregate local depth maps from them with the preservation of occlusions. Several methods do not consider EPI for depth estimation. Neri *et al.* [18] adaptively combine data term and multi-view stereo with a multi-resolution approach for reducing the complexity of the algorithm. Sabater *et al.* [19] propose a method for real-time depth estimation, using a zero-normalized cross-correlation as an image similarity measurement together with pyramid strategy for coarse-to-fine reconstruction. An approach from Jeon *et al.* [20] utilizes the cost volume from shifted sub-aperture images with different similarity measurements and further result refinement via Graph Cuts (GC) [21]. Huang [22] presents a novel framework derived from Markov random fields [23]. Based on this framework, an empirical Bayesian algorithm for depth estimation is developed. Anisimov and Stricker [11] propose an efficient approach for depth estimation by initializing the line fitting hypothesis search with a result of SGM. In this work we follow this core idea, but introduce some unique features as mentioned in Section I.

In recent years interest in the usage of neural networks for the light field processing has been growing. These methods can produce a relatively fast depth result on high-performance graphics processing units (GPU). Heber and Pock [24] utilize a Convolutional Neural Network (CNN) for predicting the orientation of a 2-dimensional hyperplane, which represent the depth information, in the 4-dimensional light field space. Sun *et al.* [25] augment the EPI with Hough and Radon transform, providing the modified result to CNN. Jeon *et al.* [26] present an extension of the algorithm [20] with learning-based matching costs. A paper by Shin *et al.* [27] explains an architecture of the multi-stream network, which currently achieves the top result in 4D Light Field Benchmark [8], [9].

B. Semi-global matching

Original SGM description was proposed by Hirschmuller [10]. Because of the high-quality output and relatively small runtime, this algorithm found its application in stereo reconstruction-related fields. Our method utilizes the SGM routine described by Haller *et al.* [28] with the cost aggregation principles from [10]. Recent work of Hernandez-Juarez *et al.* [29] shows outstanding real-time results of SGM on embedded GPU. They present a parallel version of this approach with Center-Symmetric Census transform as a cost generation method. One approach with the SGM extension for an arbitrary number of images is presented by Bethmann and Luhmann [30]. Matching cost values in their method are computed as voxels in the object space, in which the minimization process is performed.

III. Proposed approach

This section provides an overview of the steps of the proposed algorithm. We describe light field parametrization in Section III-A and provide a comparison of image similarity metrics in Section III-B. Estimation of matching cost among the views in light field images is described in Section III-C. Forming of initial disparity map is presented in Section III-D and the final result generation is outlined in Section III-E. Result post-processing is described in Section III-F.

A. Light field parametrization

Levoy and Hanrahan [31] describe a light field in the form of two-plane parametrization: a plane of spatial coordinates (u, v) for a 2-dimensional view in the light field image, and angular plane (s, t) for the viewpoint representation. The whole light field space can be denoted as $L(u, v, s, t)$. The way of finding the matching pixel for a specified disparity hypothesis in light field views can be formulated using this definition. For a given reference light field view $(\hat{s}, \hat{t})$ and a disparity hypothesis

d, matching pixel position (u, v) in the view (s, t) can be determined as [14]:

$$\hat{p}(u, v, s, t, d) = L(u + (\hat{s} - s)d, v + (\hat{t} - t)d, s, t). \quad (1)$$

B. Image similarity measurement

Generation of a matching cost, based on the comparison of image elements such as pixels, patches or windows, is an important step for disparity estimation algorithms. Commonly used pixel-wise functions are "city block" and Euclidean distance [32] (ℓ_1- and ℓ_2-norm respectively). The result of these functions can be used directly for the cost forming or can be further improved in a form of Earth Mover Distance [33] or Kernel Density Estimation (KDE) [34]. Algorithms from [14] and [11] are using KDE in a form of Epanechnikov kernel [35].

In window-based methods, widely used measures are the sum of absolute differences, the sum of squared differences and normalized cross-correlation [36]. These estimators can provide more accurate results in contrast to pixel-based methods, but the computation time increases since for each pixel in an image more pixels around are involved, which can limit usage on window-based approaches in rapid estimation algorithms.

One of the efficiently used method for the matching cost generation is Census transform as detailed by Zabih *et al.* [37]. Radiance value of a pixel in an image I is compared to pixels nearby within a set of pixels coordinates D, lying in a window. It results in a bit string for a pixel in Census-transformed image I_c:

$$I_c(u, v) = \bigotimes_{[i,j] \in D} \xi(I(u, v),\, I(u + i, v + j)), \quad (2)$$

where $\otimes$ stands for bit-wise concatenation. Pixel relations are defined as:

$$\xi(p_1, p_2) = \begin{cases} 0, & p_1 \leqslant p_2 \\ 1, & p_1 > p_2 \end{cases}. \quad (3)$$

Although this method is based on window around a pixel, the number of calculations can be reduced by using a sparse window for census transform.

C. Matching cost generation

Two image similarity measurements are used in this work for the generation of matching cost. A Census-based matching cost function can be defined through a Hamming distance between corresponding pixels from Census-transformed images. For two images in Census-transformed light field L_c with coordinates $(\hat{s}, \hat{t})$ and (s, t):

$$C_c(u, v, d) = HD(L_c(u, v, \hat{s}, \hat{t}), \hat{p}_c(u, v, s, t, d)), \quad (4)$$

where $\hat{p}_c$ stands for matching pixel position in L_c (Eq. 1). HD is the Hamming distance function. For two vectors x_i and x_j ($|x_i| = |x_j| = n$, here and further $|\dots|$ denotes cardinality) it can be determined as a number of elements with different values.

$$HD(x_i, x_j) = \sum_{k=1}^{n} x_{ik} \oplus x_{jk}. \quad (5)$$

In our approach Census transformation is extended to be applied for RGB images. It is performed for each view channel separately, and the result is composed by a sum of Hamming distances of pixels in every channel. Cost, generated by this method, is used in calculations of initial disparity map, explained in Section III-D.

Radiance comparison for correspondence matching in light field space is performed by $\ell 2$-norm. For two light field views with coordinates $(\hat{s}, \hat{t})$ and (s, t) is determined as:

$$C_{\ell 2}(u, v, d) = \|L(u, v, \hat{s}, \hat{t}) - \hat{p}(u, v, s, t, d)\|_2. \quad (6)$$

Cost from this method is used during the final disparity map estimation, described in Section III-E.

D. Initial disparity map

In order to create boundary values for correspondence search in whole light field space by more computationally-intensive algorithm the initial disparity map is calculated by using lower computationally-intensive algorithm. According to this concept, we first estimate the disparity maps for the set of four cross-lying views in the light field image $V = \{(\hat{s}, 0), (\hat{s}, t_{max}), (0, \hat{t}), (s_{max}, \hat{t})\}$, where s_{max} and t_{max} correspond to horizontal and vertical angular dimensions of light field image. Disparity maps are estimated relative to reference view $(\hat{s}, \hat{t})$. Calculations in this step are simplified compared to whole light field space correspondence matching by reducing the number of processed views and preserving changes only in one angular direction. Corresponding Census-transformed views from a light field image are used for this estimation. For every view in a set V the cost is achieved as

$$C_{c_i}(u, v, d) = HD(L_c(u, v, \hat{s}, \hat{t}), \hat{p_c}(u, v, V_i, d)), \quad (7)$$

where $i = [1, |V|]$, and disparity hypothesis d lies in predefined range $d \in T = [d_{min}, d_{max}]$, which covers all possible shifts of the pixels within two opposing views on one angular axis *e.g.* $(\hat{s}, 0)$ and $(\hat{s}, t_{max})$.

Results of the cost matching are then individually aggregated with original SGM method. For each pixel $p = (u, v)$ and $d \in T$, after traversing in direction r, formulated as a 2-dimensional vector $r = \{\Delta u, \Delta v\}$, aggregated cost L_r is

$$\begin{aligned} L_r(p, d) = {} & C(p, d) + \\ & \min\big(L_r(p - r, d), \\ & L_r(p - r, d - 1) + P1, \\ & L_r(p - r, d + 1) + P1, \\ & \min_t L_r(p - r, t) + P2\big), \end{aligned} \quad (8)$$

where $P1$ and $P2$ are penalty parameters for neighbourhood disparities, $P2 \geqslant P1$ and $C(p, d) = C_c(u, v, d)$ for the case

	$BadPix$		MSE		$Runtime$, $\log_{10}$		M, %/s	
	Median	Average	Median	Average	Median	Average	Median	Average
BSL [11]	13.41	12.74	5.43	7.28	0.71	0.78	18.39	22.25
EPI1 [16]	22.89	24.32	3.93	5.98	1.93	1.95	0.91	0.87
EPI2 [13]	22.94	22.65	5.72	8.24	0.94	0.92	9.05	9.31
EPINET [27]	**3.38**	**4.93**	**1.21**	**2.48**	0.29	0.30	**48.91**	48.12
LF [20]	16.15	16.19	7.96	9.13	3.00	3.00	0.08	0.08
LF_OCC [15]	17.82	15.07	2.70	6.76	2.69	2.72	0.17	0.17
OFSY [38]	11.33	12.04	5.43	7.03	2.30	2.30	0.46	0.47
RM3DE [18]	7.99	10.22	1.46	3.92	1.65	1.68	1.96	1.95
RPRF [22]	9.89	10.02	3.76	5.68	1.55	1.54	2.53	2.64
FSL (ours)	11.92	12.95	3.97	6.64	**0.25**	**0.23**	48.33	**56.09**

TABLE I
EVALUATION OF DIFFERENT ALGORITHMS WITH GENERAL METRICS ON 4D LIGHT FIELD BENCHMARK [8], [9]

described in this section. Costs are then summarized among all directions:

$$C_s(p,d) = \sum_r L_r(p,d). \tag{9}$$

We compute disparity map separately using the winner-takes-all (WTA) strategy on the summarized cost:

$$D_s(p) = \arg\min_d C_s(p,d). \tag{10}$$

As a result we obtain four intermediate disparity maps $D_i, i = 1...|V|$, which already correspond to the reference view of light field image, unlike in the algorithm of Anisimov and Stricker [11], where disparity maps had to be projected on the coordinates of the reference view. For obtaining the initial disparity map D_{init} the intermediate maps are fused using a confidence threshold φ. As an initialization step $D_{init} = D_1$, for every pixel (u,v) we verify if inequality $|D_{init}(u,v) - Di(u,v)| < \varphi$ for the rest of intermediate maps is true. If it is so – new value in initial disparity map is defined as an average between $D_{init}(u,v)$ and $D_i(u,v)$, if not – the pixel is discarded as uncertain. To partially cover "holes" without a valid disparity value, which appear after the described fusion, we apply a one- or two-pass median-based filling of particularly this holes with nearby values within a window. An example of initial disparity map can be found in Fig. 1 (c).

D_{init} is used for generation of boundaries for the further estimation. These boundaries will limit matching cost generation in the whole light field space. Two structures named high and low borders (D_H and D_L respectively) are generated by using the border threshold λ in such a manner:

$$D_H(u,v) = D_{init}(u,v) + \lambda; D_L(u,v) = D_{init}(u,v) - \lambda. \tag{11}$$

The values, which lies outside on predefined disparity range ($D_H > d_{max}$, $D_L < d_{min}$) are saturated accordingly. Invalid values from D_{init} are marked in the corresponding borders for re-computation on the whole disparity range T. Also, we exclude edges from the borders, so the objects boundaries are reconstructed better by involving all information from light field space. For that, the technique with Sobel operator [39] for gradient extraction from reference view is applied.

E. Final disparity map

Within the border for disparity values, we perform a correspondence search across all views in light field image for computing matching cost. Matching cost S for a pixel $p = (u,v)$ with respect to the reference view $(\hat{s},\hat{t})$ for each possible hypothesis $d \in [D_L(p), D_H(p)]$ can be computed using the ℓ_2-norm as

$$S(u,v,d) = \sum_{s=1}^{n}\sum_{t=1}^{m} \|L(u,v,\hat{s},\hat{t}) - \hat{p}(u,v,s,t,d)\|_2, \tag{12}$$

We do not use generated cost values directly for disparity estimation. Instead, for the result improvement, SGM method is used to aggregate the generated matching cost. Eq. 8 is applied to the $S(u,v,d)$ instead of $C_s(p,d)$, result for every pixel traversing direction is provided to $C_{s_\ell 2}$. SGM in this step performed only for the disparity values, which are in the predefined bounded range $d \in [D_L(p), D_H(p)]$, so values outside the boundaries do not affect the aggregation process. We determine the final disparity map D_f according to the lowest value of the aggregated costs $C_{s_\ell 2}$ for each pixel by applying the WTA strategy for Eq. 10.

F. Post-processing

Usage of the Census transformation or ℓ_2-norm gives us an opportunity for sub-pixel disparity value estimation. A popular technique for it is based on parabolic interpolation of cost values. For each pixel p in interpolated disparity map D_n procedure for calculation of the interpolated value can be expressed as follows:

$$D_n(p) = D_f(p) + \frac{C_s(p,d-1) - C_s(p,d+1)}{2\,(2\,C_s(p,d) - C_s(p,d-1) - C_s(p,d+1))} \tag{13}$$

Interpolation can only be performed on pixels, in which disparity value $D_f(p) \in [D_L(p)+1, D_H(p)-1]$ and $|[D_L(p)+$

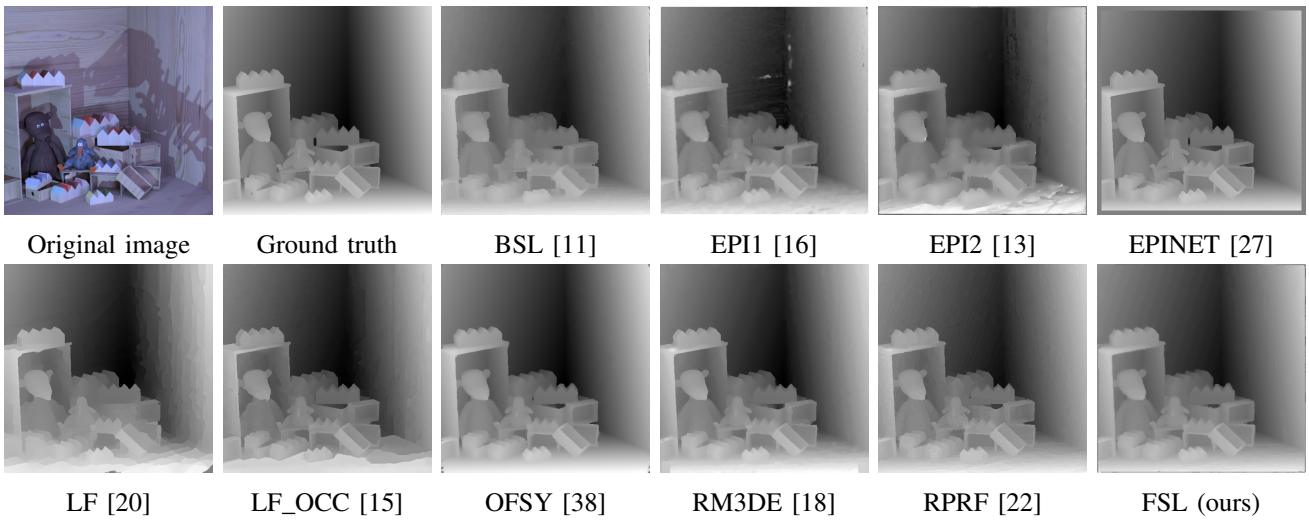

Fig. 2. Qualitative result for "dino" scene from 4D Light Field Benchmark [8], [9].

$1, D_H(p)-1]| \geqslant 3$. If the disparity value does not satisfy these conditions, then $D_n(p) = D_f(p)$ After this step we apply a median filter to D_n in order to remove the impulse noise.

IV. Experiments

Our method is evaluated with the 4-dimensional Light Field Benchmark [8], [9]. We provide a comparison of the proposed algorithm with the state-of-the-art methods presented in Section II-A: BSL [11], EPI1 [16], EPI2 [13], EPINET [27], LF [20], LF_OCC [15], OFSY [38], RM3DE [18], RPRF [22]. Together with that, qualitative results for real-world EPFL [40] and Middlebury datasets [41], [42] are presented.

A. Synthetic dataset

A dataset with synthetic scenes is provided by Honauer *et al.* [9] via 4D Light Field Benchmark (4DLFB) [8]. Twelve synthetic scenes are used for the comparison; each of them is represented by the 9x9 light field, collected from 8-bit RGB images with 512x512 pixel resolution. Camera settings and disparity ranges are provided for every scene; high-resolution disparity and depth maps are provided only for categories for training.

B. Evaluation measures

These are several general metrics given by a benchmark. Fundamental criteria for evaluation of our approach are the estimation of the percentage of pixels, in which absolute difference of the result and ground truth larger than the specified threshold, which is set to 0.07 in our comparison, formulated as the $BadPix$ metric in the mentioned benchmark, together with the runtime of the algorithm. Algorithms are also compared by the Mean Squared Error (MSE) over image pixels. Corresponding formulas and descriptions are presented by Honauer *et al.* [9], the result for different photo-consistency metrics, which are not covered in this paper, can be found online in the 4D Light Field Benchmark [8].

Also, we compare algorithms with M-metric [11], which is based on the runtime and $BadPix$ metric and formulated as a percentage of correctly computed pixels per second:

$$M = \frac{100\% - BadPix}{Runtime}\Big(\frac{\%}{sec.}\Big). \tag{14}$$

C. Algorithm parameters

Configuration of our algorithm is adjusted for an optimal result for $BadPix$ metric. Ranges for depth hypotheses are set accordingly to configuration files of each scene, all other parameters remain same for all scenes. For the SGM from Section III-D penalty parameters $P1$ and $P2$ are empirically set to 21 and 45 respectively, whereas the penalties for the extended SGM from Section III-E are equal to 17 and 35. 16 traversing directions for SGM are used. We use a sparse 7x7 pattern for Census transformation in a configuration from [11]. Confidence threshold φ is set to 3, and the border penalty λ – to 2. Window size for median-based filters is equal to 3.

D. Results

Results of the comparison with metrics from Section IV-B are presented in Table I. Visualization of disparity maps for a scene "dino" is presented in Fig. 2; per-scene evaluation together with disparity map results for other scenes can be found online in the 4D Light Field Benchmark [8] under the FSL acronym. Comments for the result are provided in Section IV-F.

E. Real-world scenes

Real-world tests were performed with two image similarity measurements for final disparity map estimation. We used an approach, described in Section III-E with $\ell 2$-norm matching

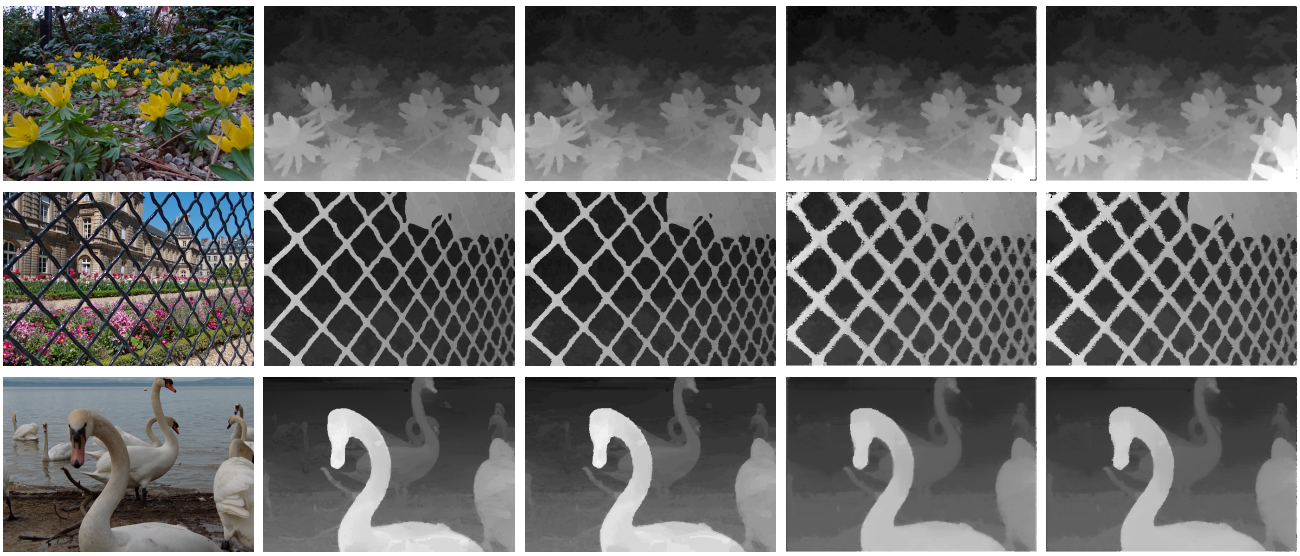

Fig. 3. Qualitative results for scenes from EPFL dataset [40]

cost, and we also modified in to use with Census cost by modifying Eq. 12 to:

$$S(u,v,d) = \sum_{s=1}^{n}\sum_{t=1}^{m} HD(L_c(u,v,\hat{s},\hat{t}), \hat{p}_c(u,v,s,t,d)). \tag{15}$$

We provide qualitative results of our algorithm for light field images from EPFL dataset [40], edited and uploaded by C.-T. Huang for the publication [22]. Light fields consist of 3x3 RGB image with resolution 541x376. A visualization of the result is presented in Fig. 3. On the testing machine in average it took 21 s. for RPRF [22] to generate a result when our approach provides the output in 1.5 s.

Also, for additional tests we use 3-dimensional light fields provided by Middlebury 2006 dataset [41], [42]. Each scene is represented by 7 images in a row with a large baseline between them. Original resolution is 1240-1396x1110 pixels, we used half-sized images for our experiments. For these images correspondence search is performed in boundary-less manner. Qualitative result for some scenes with different image similarity measurements is visualized in Fig. 4. We do not provide ground truth images for this comparison, because they are available only for two bordering views on the light fields, and we compute our disparity map with respect to the central light field image. Processing of these scenes took 8 s. on average. For both datasets, penalty parameters for SGM are set to 30 and 120, other algorithm parameters remain same and described in IV-C.

F. Discussion

The results of our algorithm are expectedly consistent with results from BSL [11]. Quantitatively, the average values of $BadPix$ and MSE are improved, which can be explained as an effect of applying SGM to the generated costs. In general, our algorithm produces average quality results within other algorithms. Qualitatively we can see an improvement of the subjective sharpness level compared to BSL. Comparison based on photo-realistic metrics proposed by Honauer *et al.* [9] also shows a slight improvement of the result. However, a visual issue with step effect on the disparity map is recognized. Further experiments are required to determine exactly how this problem can be eliminated. We believe that this effect can be partially compensated by proper adjustment of penalty parameters for the extended SGM or with more sophisticated post-processing filtration.

Currently, we achieve the best result for runtime and the second best for M-metric (14). The approach EPINET [27] can be considered as top-of-the-line, providing good results in terms of depth quality together with the satisfying runtime. However, their algorithm is performed on the high-end GPU. Such heavy computational resources are not required by our approach, which utilizes a central processing unit (CPU) without specifically employed thread parallelism. The exact configuration is explained in Section IV-G. There is a field of improvement for this algorithm with SIMD-instructions and exploitation of parallelism *e.g.* for independent of each other traversing directions.

Borders from the initial depth map help to reduce the number of sampled hypotheses by further processing in a range from 50% (real-world scenes) up to 97% (synthetic scenes). Runtime of the correspondence search in the light field space is affected proportionally.

A possible gain of the number of bordered pixels can be achieved by the improvement of the matching algorithm for initial disparity map estimation. For instance, it can be done with different configurations of Census window *e.g.* by using adaptive Census window based on gradients. Real-world tests show that utilization of ℓ_2-norm for global correspondence

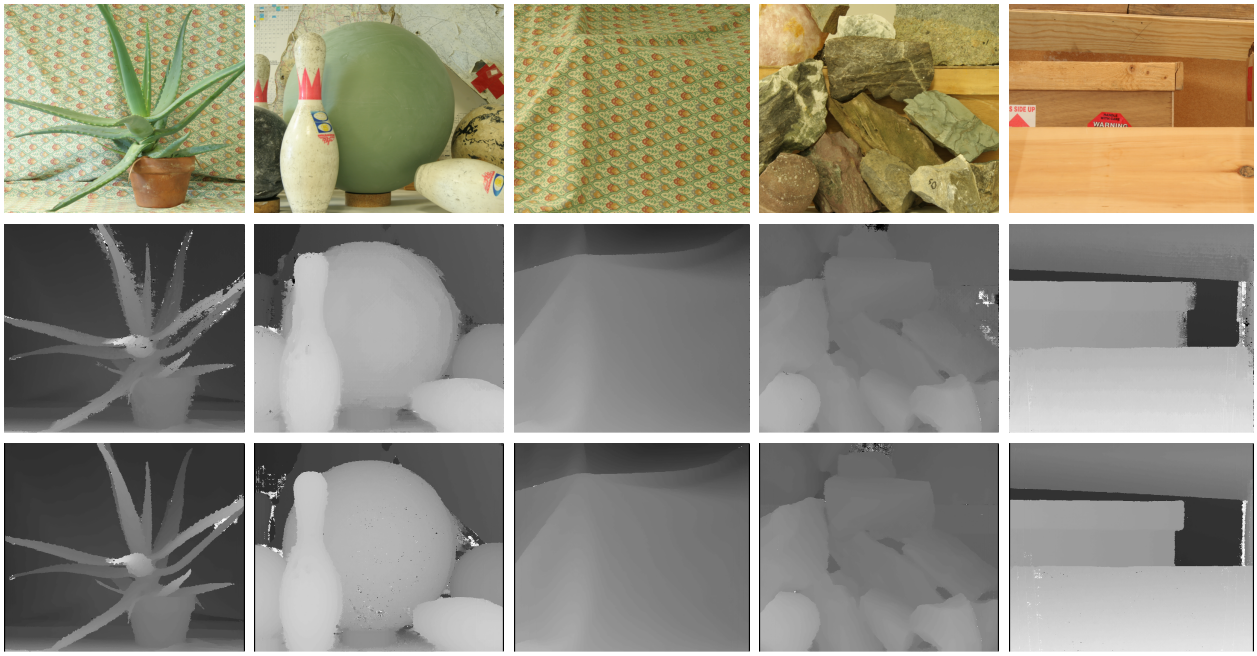

Fig. 4. Qualitative results of presented method for Middlebury dataset [41], [42]: ℓ_2-norm in the middle and Census-based in the bottom row

search cannot be considered as effective in some cases. ℓ_2-norm works better for the case of synthetic images and real-world scenes with narrow baseline in light field images, whereas Census-based estimation provides a better result for other real-world cases. Configuration with Census-based matching cost tested with Middlebury dataset images (Fig. 4) revealed a fact that computation of the initial disparity map and utilization of it for correspondence search with the mentioned similarity measurement shows worse runtime compare to the configuration with initialization. In our opinion, it happens mainly because of a large number of depth hypotheses in this dataset. The initial map was computed for EPFL dataset for both similarity configurations, and the time difference was insignificant.

We assume that decision for the selection of proper matching cost and the boundaries configuration should depend on the image size, the maximum possible displacement between two boundary light field views and quality of the image by itself (*e.g.* it should not be noisy). Without boundaries, Census-based configuration has somewhat similar to the method of Tomioka *et al.* [43] in terms of aggregation principles.

During the tests with real-world images, we came up with an idea of adjustment for border threshold λ dependent on the initial depth value, so that for small disparities this threshold is higher rather than for higher values. Such a strategy can help to determine the more accurate values for the farther objects, where the pixel discontinuity can be a crucial factor for the right depth estimation.

Additional limitations have been observed during these tests. On Fig. 3 the visual problem with grids in the middle scene can be noticed. Also, we were not able to generate proper interpolation for a bottom scene in Fig. 3, it could be a problem related to the use of a small number of images for reconstruction. Although bordering of depth values with the initial map helps to reduce depth mismatching noise, some of the wrongly calculated pixels still can ”survive” this filtering, which is visible in images on Fig. 4. These mistakes appear either in the areas marked previously as non-consistent or on the object edges. Our algorithm fails with depth estimation on image boundaries for scenes with a relatively large distance between images. The explanation of this problem is related to our selection of the central light field image as a reference view. In this case search for a matching pixel from image boundary fails since there is no match in most of the images and therefore our algorithm can not aggregate enough depth score for the correct value. A solution for this problem was proposed by Kim *et al.* [14], where the position of reference images changes over time and cost aggregation is performed from the new position. However, for the presented method such an improvement can greatly increase the running time, which is in contradiction with our main objective.

G. Environment

Hardware configuration for depth processing includes CPU Intel Xeon E3-1245 V2 @ 3.40 GHz, forced to work in single-thread mode. Our algorithm is implemented in C and compiled using GCC v.7.3.1 with /O3 option.

V. Conclusion

In this paper, we proposed a fast depth estimation method from light fields by extending an efficient stereo matching algorithm with a methodology of enlargement to multi-view sampling by the correspondence search in light field space. The evaluation against state-of-the-art methods showed that our algorithm produces comparable depth map results, which

have been proven by different quantitative metrics. In contrast to many state-of-the-art approaches, our proposed method produces depth maps in a relatively small amount of time. As a result, our method provides almost the best result in terms of computed pixels per unit of time. In addition, our approach works for different configurations of 3-dimensional and 4-dimensional synthetic and real-world light field images. Further work will investigate additional real-time improvements, *e.g.* by utilizing a pyramidal scheme with downscaled images together with hardware-specified optimizations, such as parallelism on multi-core CPU or GPU.

ACKNOWLEDGMENTS

This work was partially funded by the Federal Ministry of Education and Research (Germany) in the context of the project DAKARA (13N14318). The authors are grateful to Kiran Varanasi and Jonathan Wray for the provided support.

REFERENCES

[1] A. Gershun, "The light field," *Studies in Applied Mathematics*, 1939.
[2] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *International Conference on Computational Photography (ICCP)*, 2009.
[3] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report CSTR*, 2005.
[4] C. Perwass and L. Wietzke, "Single lens 3d-camera with extended depth-of-field," in *Human Vision and Electronic Imaging XVII*. International Society for Optics and Photonics, 2012.
[5] B. Wiburn, "High performance imaging using arrays of inexpensive cameras," Ph.D. dissertation, Stanford University, 2004.
[6] K. Venkataraman, D. Lelescu, J. Duparré, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar, "Picam: An ultra-thin high performance monolithic camera array," *ACM Transactions on Graphics (TOG)*, 2013.
[7] Y. Anisimov, O. Wasenmüller, and D. Stricker, "A compact light field camera for real-time depth estimation," *International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2019.
[8] "4d light field benchmark," http://hci-lightfield.iwr.uni-heidelberg.de, accessed: 30.07.2019.
[9] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Asian Conference on Computer Vision (ACCV)*, 2016.
[10] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005.
[11] Y. Anisimov and D. Stricker, "Fast and efficient depth map estimation from light fields," *International Conference on 3D Vision (3DV)*, 2017.
[12] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International Journal of Computer Vision*, 1987.
[13] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4d light fields," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.
[14] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, "Scene reconstruction from high spatio-angular resolution light fields." *ACM Trans. Graph.*, 2013.
[15] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *International Conference on Computer Vision (ICCV)*. IEEE, 2015.
[16] O. Johannsen, A. Sulc, and B. Goldluecke, "What sparse light field coding reveals about scene structure," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.
[17] H. Sheng, P. Zhao, S. Zhang, J. Zhang, and D. Yang, "Occlusion-aware depth estimation for light field using multi-orientation epis," *Pattern Recognition*, 2018.
[18] A. Neri, M. Carli, and F. Battisti, "A multi-resolution approach to depth field estimation in dense image arrays," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015.
[19] N. Sabater, G. Boisson, B. Vandame, P. Kerbiriou, F. Babon, M. Hog, R. Gendrot, T. Langlois, O. Bureller, A. Schubert *et al.*, "Dataset and pipeline for multi-view lightfield video," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017.
[20] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. So Kweon, "Accurate depth map estimation from a lenslet light field camera," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.
[21] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *European Conference on Computer Vision (ECCV)*. Springer, 2002.
[22] C.-T. Huang, "Robust pseudo random fields for light-field stereo matching," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.
[23] A. Blake, P. Kohli, and C. Rother, *Markov random fields for vision and image processing*. Mit Press, 2011.
[24] S. Heber and T. Pock, "Convolutional networks for shape from light field," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.
[25] X. Sun, Z. Xu, N. Meng, E. Y. Lam, and H. K.-H. So, "Data-driven light field depth estimation using deep convolutional neural networks," in *International Joint Conference Neural Networks (IJCNN)*, 2016.
[26] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. W. Tai, and I. S. Kweon, "Depth from a light field image with learning-based matching costs," *Transactions on Pattern Analysis and Machine Intelligence*, 2018.
[27] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
[28] I. Haller, C. Pantilie, F. Oniga, and S. Nedevschi, "Real-time semi-global dense stereo solution with improved sub-pixel accuracy," in *Intelligent Vehicles Symposium (IV)*. IEEE, 2010.
[29] D. Hernandez-Juarez, A. Chacón, A. Espinosa, D. Vázquez, J. C. Moure, and A. M. López, "Embedded real-time stereo estimation via semi-global matching on the GPU," in *International Conference on Computational Science 2016 (ICCS)*, 2016.
[30] F. Bethmann and T. Luhmann, "Semi-global matching in object space," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2015.
[31] M. Levoy and P. Hanrahan, "Light field rendering," in *Computer Graphics and Interactive Techniques (SIGGRAPH)*. ACM, 1996.
[32] C.-C. Chen and H.-T. Chu, "Similarity measurement between images," in *Computer Software and Applications Conference (COMPSAC)*, 2005.
[33] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. Journal of Computer Vision*, 2000.
[34] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, 1962.
[35] V. A. Epanechnikov, "Non-parametric estimation of a multivariate probability density," *Theory of Probability & Its Applications*, 1969.
[36] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.
[37] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European Conference on Computer Vision (ECCV)*. Springer, 1994.
[38] M. Strecke, A. Alperovich, and B. Goldluecke, "Accurate depth and normal maps from occlusion-aware focal stack symmetry," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.
[39] I. Sobel and G. Feldman, "A 3x3 isotropic gradient operator for image processing," *Talk at the Stanford Artificial Project*, 1968.
[40] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Quality of Multimedia Experience (QoMEX)*, 2016.
[41] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007.
[42] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Computer Vision and Pattern Recognition (CVPR)*, 2007.
[43] T. Tomioka, K. Mishiba, Y. Oyamada, and K. Kondo, "Depth map estimation using census transform for light field cameras," *IEICE TRANSACTIONS on Information and Systems*, 2017.