Light field scale-depth space transform for dense depth estimation

Ivana Tošić, Kathrin Berkner

Ricoh Innovations, Corp.
Menlo Park, USA
Email: {ivana, berkner}@ric.ricoh.com

Abstract

Recent development of hand-held plenoptic cameras has brought light field acquisition into many practical and low-cost imaging applications. We address a crucial challenge in light field data processing: dense depth estimation of 3D scenes captured by camera arrays or plenoptic cameras. We first propose a method for construction of light field scale-depth spaces, by convolving a given light field with a special kernel adapted to the light field structure. We detect local extrema in such scale-depth spaces, which indicate the regions of constant depth, and convert them to dense depth maps after solving occlusion conflicts in a consistent way across all views. Due to the multi-scale characterization of objects in proposed representations, our method provides depth estimates for both uniform and textured regions, where uniform regions with large spatial extent are captured at coarser scales and textured regions are found at finer scales. Experimental results on the HCI (Heidelberg Collaboratory for Image Processing) light field benchmark show that our method gives state of the art depth accuracy. We also show results on plenoptic images from the RAYTRIX® camera and our plenoptic camera prototype.

I. Introduction

Compared to traditional imaging systems, plenoptic systems provide additional capabilities and functionalities such as single-snapshot multi-spectral imaging [1], refocusing [2] and 3D imaging [3]. This is achieved by inserting a micro-lens array in front of the imaging sensor. After calibration, plenoptic data can be demultiplexed to a set of multi-view images that form a 4-dimensional (4D) data structure called the light field (LF) [4]. Prior to development of plenoptic cameras, LFs have been acquired by camera arrays or by a moving camera rig, capturing images from regularly spaces viewpoints. Such acquisition leads to 3D, 4D or higher dimensional pixel arrays, which represent specific samplings of the 7D plenoptic function [5].

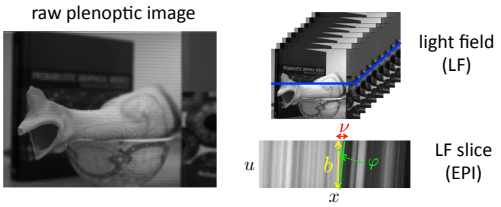


Fig. 1. Example of a LF extracted from a plenoptic image. The blue line indicates the pixels extracted in the horizontal $x-u$ slice (EPI) of LF, displayed below. Angle φ of ray at pixel x uniquely determines the disparity ν of x.

Even though LFs contain multi-view data of a 3D scene and, therefore, depth information of objects, extracting dense depth maps from LFs still represents a challenging problem because of the high dimensionality of LFs. Estimation of globally consistent dense depth maps from multiple views or LFs typically requires global optimization [6], [7], which is of prohibitive complexity for such high-dimensional data processing. Therefore, there is an essential need for local LF processing algorithms that efficiently and robustly extract depth information, while simultaneously handling occlusions in a given 3D scene. To address this problem, we exploit the particular geometry of LFs obtained by plenoptic sensors or planar camera arrays, where viewpoints are regularly spaced on a planar surface. A parametrization of such 4D LFs is usually given in coordinates (x, y, u, v), where (x, y) are pixel coordinates for an image taken from a viewpoint (u, v). An example of a LF obtained from a plenoptic image is shown in Figure 1. We can see that a 2D $x-u$ slice of the LF, obtained by cutting the LF across views, has a "linear" or "ray" structure, where the angle of a ray corresponds to a different depth value of that point in a 3D scene. This

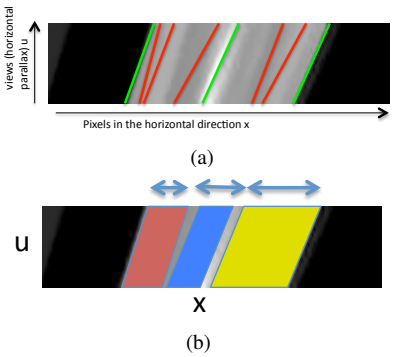


Fig. 2. Detection of ray edges and rays. a) While ray edges (green lines) can be easily detected and their angle estimated, there are problems with estimating angles for pixels within uniform regions (red lines). b) Our approach: we detect whole rays and estimate their position, angle and width.

structure has been observed by Bolles et al. [8] who named these slices *epipolar plane images* (EPIs).

For LFs with such EPI structure, the task of dense depth estimation can be formulated as a problem of estimating the angle of rays for each pixel x in a given EPI. In Figure 2 we show an example of an EPI. We can see that a typical EPI consists of discontinuities that we call *ray edges*, and uniform regions (stripes) bounded by ray edges that we will simply call *rays* throughout the rest of the paper. Each ray is parametrized by its position on x-axis, its width and its angle with the u-axis. While depth at ray edges can be reliably estimated via for example line fitting (green lines in Figure 2a), estimating depth within rays is challenging as there is an ambiguity in angle estimation (red lines in Figure 2a) due to the same value of pixels within the ray. Similar ambiguities exist in stereo and multi-view approaches, where most solutions impose smoothness constraints in a global optimization [6] or in a variational framework [9]. Here, we take a different approach by formulating the problem as a ray detection problem where the goal is to *simultaneously* detect rays and determine their positions, angles and widths, as illustrated in Figure 2b. Once we have detected rays and estimated their widths, we can assign the same depth value to all points within the same ray. This way, we can obtain a dense depth map without performing global optimization.

The proposed solution to ray detection and estimation is based on multi-scale LF analysis using scale-space theory. We exploit the construction of *light field scale-depth spaces* that take into account the specific properties of LFs, as presented in our prior work [10]. We name them **Lisad** spaces, short for **Li**ght field **s**cale **a**nd **d**epth spaces. Lisad spaces are parametrized both in terms of scale of objects recorded by the LF and in terms of objects' depth. In [10], we have introduced a "Ray Gaussian" kernel for construction of Lisad spaces that satisfy the scale-invariance property, and shown an application to 3D key-point detection. In this paper, we propose a construction of Lisad spaces based on the normalized second derivative of the Ray Gaussian and formulate ray detection as extrema detection is such Lisad spaces. We prove theoretically that normalized second-derivative Lisad spaces (referred to from now on as *Lisad-2*) satisfy the scale-invariance property and do not exhibit any angle bias. Such bias would result in inaccurate depth assignment to foreground vs. background objects. Detected extrema in the Lisad-2 spaces provide scale and depth estimates for rays of different sizes, where the scale parameter is proportional to the width of the ray. Rays are further converted into per-pixel depth values after solving occlusion conflicts. Obtaining the width of rays via their scale thus represents the crucial benefit of using Lisad spaces for dense depth estimation. Moreover, this approach includes only local processing and does not require any iterative estimation.

Our contribution with respect to prior art is a new local method for ray detection and dense depth estimation by multi-scale 3D analysis of LFs. Analysis of Lisad-2 spaces allows for a joint ray detection and ray angle estimation, which has previously always been done separately [8], [11]. We evaluate the depth estimation accuracy on the HCI (Heidelberg Collaboratory for Image Processing) LF benchmark database [12] and show that it outperforms all algorithms presented in the benchmark, including the state of the art approach of Wanner and Golduecke [9]. We also show that estimated depth maps from plenoptic images obtained with the RAYTRIX® plenoptic camera [3] and our own plenoptic camera prototype provide information about the uniform regions while maintaining sharp edges.

II. Prior art

We first review prior art on LF dense depth estimation, then briefly give background on scale-spaces and finally describe our recent work on LF scale-depth spaces [10].

A. Depth estimation from light fields

Unlike dense depth estimation from stereo or multiple views, dense depth estimation from LFs has not been much investigated until the last few years. The reason is that acquiring LFs using plenoptic cameras has only recently become an excellent alternative to difficult and expensive acquisition with camera arrays. Hand-held plenoptic sensors [2], [3] and multi-aperture systems [13] offer a possibility to record LFs using hardware with a small form-factor. In recent years, we have seen a myriad of papers addressing different challenges of plenoptic imaging, ranging from hardware design, super-resolution methods to 3D geometry estimation. What differentiates most prior work in LF depth estimation from the more general approaches

of depth from multiple views [6], [14] is that they exploit the particular EPI structure described in Sec. I.

Bolles et al. analyzed the light fields by detecting edges, peaks and troughs in the EPI images and then fitting straight lines to those edges in order to estimate depth of image features [8]. Criminisi et al. went one step ahead to group lines in EPIs into 2D EPI stripes and 3D EPI tubes [11]. Gelman et al. used active contours and the level-set method to segment the LF into depth layers [15]. A common characteristic of these approaches is that they use existing image processing tools for detection or segmentation of rays in EPIs, followed by a separate step of their angle (depth) estimation. Our approach differs in that sense: by finding extrema in Lisad spaces we **jointly** detect rays of different scales and estimate their angles.

Wanner and Goldluecke [7], on the other hand, do not detect rays, but calculate the structure tensor for each pixel in each view, which gives local estimates of angles of rays in the EPIs. Local estimates are then integrated into an objective based on variational regularization and solved using global optimization [7]. Their global approach gives high quality dense depth maps, but also requires more computation time. They have recently introduced a more efficient algorithm that performs smoothing of their local depth estimates [9]. What differentiates our work from theirs is that our depth estimates are obtained by processing the whole rays, thereby operating on all views at the same time, without the need to impose smoothness on views after depth estimation. We also differ from their global approach [7] by having only local computations within a spatial neighborhood of a given pixel (in x-axis), while still preserving computation over all views (in u-axis). Spatially local computation is of large importance for applications of plenoptic sensors that have limited computational capabilities. Another prior work that proposes local depth estimation from light fields has been presented by Kim et al. [16]. Their method requires, however, LFs with high spatio-angular resolution, which are hard to obtain with plenoptic sensors.

Finally, the dataterm used in our depth estimation method is guaranteed not to introduce an angle bias, a problem first time identified by Criminisi et al. [11].

B. Gaussian scale spaces

Research on scale-space theory has a long history, dating back to the seminal paper by Witkin [17]. A plethora of prior works introduce the theory and construction of scale-spaces for representing signals by filtering them with kernels of different scales. Scale spaces have found many applications in analysis of images, videos, tomography data, medical images (see [18] for a review of literature on scale-spaces). One of most well known applications is the Scale-Invariant-Feature-Transform (SIFT), where feature detection is based on finding extrema in the scale-spaces built upon the Difference of Gaussian (DoG) kernel [19].

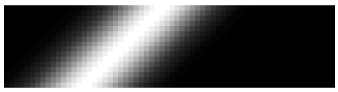


Fig. 3. Example of a Ray Gaussian kernel with $\varphi = \pi/4$ and $\sigma = 6$.

The most commonly used kernel for constructing scale spaces is the Gaussian kernel: $G_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$. Its associated scale space in 1D case is defined as $I(x,\sigma) = I(x) * G_\sigma(x)$, where $*$ denotes convolution. An important property of Gaussian scale-spaces is the scale invariance property: $(J * G_\sigma)(x) = (I * G_{s\sigma})(sx)$, where $J(x) = I(sx), s \in \mathbb{R}$. This property says that a feature at scale σ elicits the same response as that feature at a larger scale $s\sigma$, which allows for scale-invariant processing of signals. This property is necessary for dealing with the object size variations in image processing [18]. Examples include edge detection by finding extrema in the scale-space built upon the normalized first derivative of the Gaussian $\sigma \frac{dG_\sigma}{dx}$ and blob detection by finding extrema in the scale-space built upon the normalized second derivative of the Gaussian $\sigma^2 \frac{d^2 G_\sigma}{dx^2}$ [20]. Our approach bears similarities to blob detection using normalized second derivative Gaussian scale-spaces [20], but is specifically constructed for analysis of LFs, as explained in the next section.

C. Light field scale and depth (Lisad) spaces

In [10] we have presented a method for constructing Lisad spaces based on a *Ray-Gaussian* kernel, which was the first time scale-depth spaces for LFs were introduced. We have defined the Ray-Gaussian (RG) function as:

$$\mathcal{R}_{\sigma,\varphi}(x,u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x+u\tan\varphi)^2}{2\sigma^2}}, \qquad (1)$$

where x and u are coordinates of pixels in a 2D EPI, φ is the angle that the RG forms with the u-axis and σ is the width parameter of the kernel[1]. An example of a RG function is given in Figure 3. We can see that it is Gaussian in x-direction and a ridge in u-direction. The slant of the ridge is equal to $\tan\varphi$, which multiplies u in the shift of x in the exponent. We have further used the RG to construct the LF scale-depth space $\mathcal{L}(x;\sigma,\varphi)$ in the following way:

$$\mathcal{L}(x;\sigma,\varphi) = (I * \mathcal{R}_{\sigma,\varphi})(x,u)|_{u=0}, \qquad (2)$$

where $u = 0$ is chosen because we need to evaluate convolution only over x (pixel domain) and not over views u. This is because the LF features will be present in all

[1]Note that the RG is not just a special case of an affine transformation of a 2D Gaussian, where σ in u-direction is ∞. Such affine Gaussian is $1/\sigma\sqrt{2\pi}\exp\left(-(x\cos\varphi + u\sin\varphi)^2/(2\sigma^2)\right)$.

views (except in rare cases of occlusion), so we do not need to localize them within the views. That is,

$$(f * g)(x,u)|_{u=0} = \iint_{x'\,u'} f(x-x',-u')g(x',u')dx'du'.$$

Note here that $\mathcal{L}(x;\sigma,\varphi)$ does not depend on u since the convolution is only over x, and that is has both scale σ and angle φ parameters. Note also that we refer to the constructed space as scale-depth space, since the ray angle is uniquely related to depth. In [10] we have shown the scale-invariance property of the RG kernel, which is important for building its associated scale-depth spaces. We review the results from [10] here for completeness. Unlike in standard image scale-spaces, we analyze the scale-invariance by downsampling only in x since downsampling in u (dropping views) is usually undesirable.

Lemma 2.1: [Scale-invariance of RG] [10] The following equality holds: $\mathcal{R}_{\sigma,\varphi}(x,u) = s\mathcal{R}_{s\sigma,\varphi'}(sx,u)$, where $\varphi' = \arctan(s\tan\varphi)$, $\varphi \in (-\pi/2,\pi/2)$ and $s>0$.

For proof of this lemma and proofs of the following two propositions, please see [10]. Lemma 2.1 shows that a RG with scale σ and angle φ is equal to its downsampled version at scale $s\sigma$ and angle $\varphi' = \arctan(s\tan\varphi)$, with values multiplied by s and for downsampling in x by factor s. Using Lemma 2.1, we have also shown the scale-invariance property of the Lisad space, as given by the following proposition.

Proposition 2.2: [Scale-invariance of RG scale-depth space [10].] If we have a LF slice (i.e., EPI) $J(x,u)$ such that $J(x,u) = I(sx,u)$ (i.e., I is a downsampled version of J over x), then it holds:

$$(J * \mathcal{R}_{\sigma,\varphi})(x,u)|_{u=0} = (I * \mathcal{R}_{s\sigma,\varphi'})(sx,u)|_{u=0}, \quad (3)$$

where $\varphi' = \arctan(s\tan\varphi)$, $\varphi \in (-\pi/2,\pi/2)$ and $s>0$.

Finally, in [10] we have shown another property of the RG, which relates to the angle invariance of its inner product with an EPI.

Proposition 2.3: [10] If we have a real function $f_\varphi(x,u) = h(x+u\tan\varphi)$, where $x,u \in \mathbb{R}$, $\varphi \in (-\pi/2,\pi/2)$ and h is a 1D embedding of f_φ, then $\forall\varphi$ it holds: $\langle f_\varphi, \mathcal{R}_{\sigma,\varphi}\rangle = \langle f_0, \mathcal{R}_{\sigma,0}\rangle$.

This property means that if an EPI can be embedded in a 1D space for a given φ, i.e., $I(x,u) = h(x+u\tan\varphi)$, $\forall x \in \mathbb{R}$, then its inner product with the RG of angle φ will always have the same value, independent of the value of φ. Note that only LFs without occlusions satisfy this assumption, which is not always the case in real LFs. However, we can assume that this requirement is satisfied locally. This is an important property of Lisad spaces because it assures that there is no angle (depth) bias. Finally, note that the proven invariance to scale and angle differentiates the Ray Gaussian from other kernels or matching fitters that do not exhibit such properties.

III. Depth estimation by ray detection

In order to estimate depth of objects in a given LF, we need to estimate angles of all rays in all EPIs. First, we need a way to detect rays along with their positions in the slice, their widths and their angles. We propose to detect rays in EPIs by finding extrema (local minima and maxima) of the normalized second derivative Ray Gaussian Lisad space. We present the framework for horizontal ($x-u$) EPIs, but the same holds for vertical ($y-v$) EPIs. Those will be combined later.

A. Lisad space of the RG second derivative

We first show that scale-invariance holds for scale-depth spaces built upon the "normalized" Ray Gaussian second derivative $\sigma^2\mathcal{R}''_{\sigma,\varphi} = \sigma^2\frac{d^2}{dx^2}\mathcal{R}_{\sigma,\varphi}$. We define the normalized second derivative RG Lisad space as: $\mathcal{L}''_n(x;\sigma,\varphi) = (I * \sigma^2\mathcal{R}''_{\sigma,\varphi})(x,u)|_{u=0} = (I * \sigma^2\frac{d^2}{dx^2}\mathcal{R}_{\sigma,\varphi})(x,u)|_{u=0}$ and refer to it as Lisad-2 space.

Proposition 3.1: [Scale-invariance of Lisad-2 space.] If we have an EPI $J(x,u)$ such that $J(x,u) = I(sx,u)$ (i.e., I is a downsampled version of J over x), then it holds:

$$(J * \sigma^2\mathcal{R}''_{\sigma,\varphi})(x,u)|_{u=0} = (I * s^2\sigma^2\mathcal{R}''_{s\sigma,\varphi'})(sx,u)|_{u=0}, \quad (4)$$

where $\varphi' = \arctan(s\tan\varphi)$, $\varphi \in (-\pi/2,\pi/2)$ and $s>0$.

Proof

$$(J * \sigma^2\frac{d^2}{dx^2}\mathcal{R}_{\sigma,\varphi})(x,u)|_{u=0} =$$

$$= \iint_{x'\,u'} \sigma^2\frac{d^2}{dx^2}\mathcal{R}_{\sigma,\varphi}(x-x',-u')J(x',u')dx'du'$$

$$\overset{(L\,2.1)}{=} \iint_{x'\,u'} s\sigma^2\frac{d^2\mathcal{R}_{s\sigma,\varphi'}(sx-sx',-u')}{dx^2}I(sx',u')dx'du'$$

$$\overset{(w=sx')}{=} \iint_{w\,u'} s\sigma^2\frac{d^2\mathcal{R}_{s\sigma,\varphi'}(sx-w,-u')}{dx^2}I(w,u')\frac{dw}{s}du'$$

$$= \iint_{w\,u'} \sigma^2\frac{d^2\mathcal{R}_{s\sigma,\varphi'}(sx-w,-u')}{d(sx-w)^2}\frac{d(sx-w)^2}{dx^2}I(w,u')dwdu'$$

$$= (I * s^2\sigma^2\frac{d^2}{dx^2}\mathcal{R}_{s\sigma,\varphi'})(sx,u)|_{u=0}.$$

Besides scale invariance, note that Proposition 2.3 relating to the depth invariance of the inner product with a LF slice holds also for $\sigma^2\mathcal{R}''_{\sigma,\varphi}$.

Figure 4 shows an example of generating a Lisad-2 space for a given 2D EPI. The $x-u$ EPI (top left panel) is convolved over dimension u with the normalized second derivative of a Ray Gaussian kernel (top middle panel), giving a 3D Lisad-2 space with coordinates (x,σ,φ) (top right panel). In the bottom panels, we show examples of 2D slices through the Lisad-2 volume when fixing a location x_0, a scale σ_0, and an angle φ_0. From left to right panels, we show slices (x_0,σ,φ), (x,σ_0,φ) and (x,σ,φ_0).

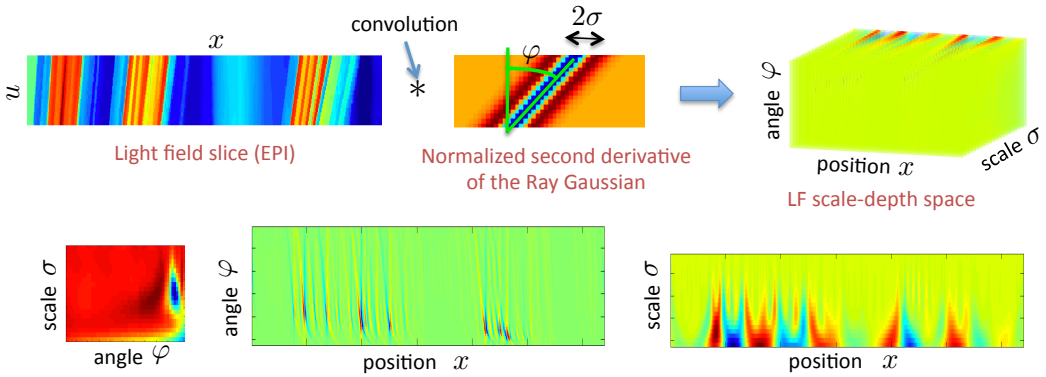


Fig. 4. Illustration of the scale-depth space of a light field slice, built upon the second derivative of a Ray Gaussian. Top panels: Convolution of an EPI with a normalized second derivative of the Ray Gaussian gives a 3D light field scale-depth space (Lisad-2). Bottom panels show Lisad-2 slices; from left to right: angle-scale, position-angle and position-scale.

Extrema in those slices are located in the blue and red regions. The exact coordinates of the extrema are found through search over the entire 3D volume. Whereas the (x, σ, φ_0) looks like a typical scale-space visualization with extrema located at the bottom of figure along the x-axis, the slices (x_0, σ, φ), (x, σ, φ_0) exhibit a different structure, with extrema being located inside the volume.

B. Ray estimation by extrema detection

Similar to using second derivative Gaussian scale-spaces for blob detection, we use the normalized second derivative Ray-Gaussian Lisad spaces to find rays in the EPIs. Namely, it can be easily shown that an extremum in Lisad space will be located exactly in the middle of the ray, where the width of the ray is exactly 2σ of that extremum.

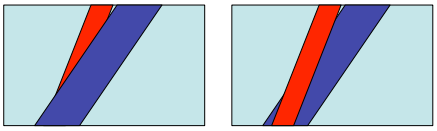
Parameters of P extrema points $\{(x_p, \sigma_p, \varphi_p)\}_{p=1,\dots,P}$ give us the following information about each ray p:

- position of the center of the ray x_p;
- width of the ray $2\sigma_p$;
- angle of the ray φ_p.

From the angle φ_p we get depth of that ray by using the camera calibration parameters as $d_p = fb/\tan(\varphi_p)$, where f is camera focal length and b is the distance between neighboring cameras.

C. Occlusion detection

After we have detected rays and found their parameters, we need to resolve occlusion conflicts between overlapping rays. Since we have the position and width of each ray, we can easily find pairs that overlap. Once we have found overlapping rays, we need to decide on their ordering from foreground to background. Because larger angle of rays indicates smaller depth (closer objects, larger parallax), rays with larger angles should always be in the foreground



(a) (b)

Fig. 5. Ray ordering under occlusion. a) Possible. b) Impossible.

as shown in Figure 5a. Due to noise in images, detected rays sometimes conform to the situation presented in Figure 5b, which is an impossible ordering. When we find such cases of overlapping rays, we remove the "occluded" ray (blue in Figure 5b) from the rays set. For situations conforming to Figure 5a, we keep the rays and we can additionally record the information about the occlusion.

To decide the ordering we do not use as metric the value of the Lisad-2 space $\mathcal{L}''(x; \sigma, \varphi)$ because the invariance to angle values does not hold in case of occlusion. That means that this metric would be sensitive to occlusion. Instead we evaluate the variance of the ray along its direction, within the ray width equal to 2σ. The ray with the smaller variance is considered to be in front, because it has less variation as it does not cross the occlusion boundary.

D. Ray elimination, combination and post-processing

Beside rays eliminated due to occlusion conflicts, we also remove rays that have one weak edge. Those rays are sometimes detected next to object boundaries, meaning that one side of the ray is an object edge while the other side is within a uniform background. To solve this problem, we use the normalized first derivative Ray-Gaussian Lisad space (Lisad-1) to detect ray edges (see [10]) and then keep the rays that have ray edges on both sides. Moreover, we impose a condition that those ray edges have an angle

value within a small threshold from the angle of the ray.

Lisad space construction (for first and second derivatives), ray detection and occlusion detection are performed separately on horizontal and vertical EPIs. After we have eliminated the weak rays, we convert information about each ray (its position, scale, angle and value of the scale-depth space) into a dataterm for depth estimation. Namely, for each ray, we record its angle value for pixels within that ray (within $\pm\sigma$ from the center of the ray). Therefore, each ray p assigns an angle value φ_p to pixels in $(x_p - \sigma_p, x_p + \sigma_p)$. After we have done that for all rays, we might have situations where a pixel has multiple assignments originating from multiple rays. For each pixel and each angle value, we then assign a dataterm value equal to the value of the Lisad-2 space of the ray which gave that angle value. Therefore, we obtain a dataterm $M_h(x, y, \varphi)$ from horizontal EPIs and a dataterm $M_v(x, y, \varphi)$ from vertical EPIs. The final dataterm is then equal to $M = M_h + M_v$. To assign an angle value φ_i to each pixel (x_i, y_i), we take the $\varphi_i = \arg\max_\varphi M(x_i, y_i, \varphi)$. Angles are then converted to depth using $d = fb/\tan(\varphi)$. Therefore, we obtain a depth map from the detected rays (using the Lisad-2 space) and from the ray edges (using the Lisad-1 space). We can further combine these two depth maps by taking only the confident estimates from both maps. We say that an estimate is confident if its dataterm value exceeds a certain percentage of the maximal dataterm value. In a similar way, we combine depth estimates from different color channels of LFs.

Finally, we should note that after this depth assignment there might be pixels with no depth value assigned, due to no rays for that pixel or if the depth estimate is not confident enough. We inpaint these regions by median filtering with masking of the missing regions during filtering. Initial depth map in the non-missing regions is then combined with the inpainted depth map values in the mission regions. Finally, we perform total-variation denoising [21] using the ℓ_1 noise model to remove the outliers.

The flow-chart of the depth estimation method is shown in Figure 6. Dashed lines denote the detection of 3D edges in the EPIs and their angles to help with depth assignment. Ray elimination, combination and post-processing (inpainting of missing regions and denoising) are all grouped within the depth assignment block.

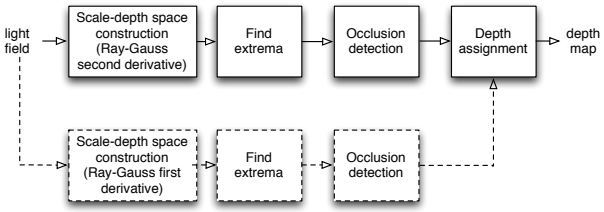


Fig. 6. Flow-chart of depth estimation method using Lisad spaces.

IV. Experimental results

We have first evaluated our method on the LF benchmark database hosted by the Heidelberg Collaboratory for Image Processing (HCI) [22] and compared the depth estimation accuracy to the best reported method of the benchmark. We have chosen that database because the LFs are obtained with BLENDER™ rendering and thus contain the ground truth depth. Note that the method presented in [16] is not evaluated on the HCI benchmark and that the source code is not available online. Therefore, we cannot provide comparisons to [16] for the HCI database.

For our method implementation, we have used 64 samples of the angle parameter to form the angle space, distributed uniformly in depth within the depth range of the datasets. The conversion from depth to disparity for the HCI datasets is provided in [22], where the disparity is represented as the slant of the ray in the EPI. In our notation, the disparity is equal to $\tan\varphi$. For the scale parameters, we have used 3 octaves with 4 samples per octave. Prior to depth estimation, we have converted the RGB light fields into the YCbCr colorspace and then performed depth estimation per color channel in that space. Local extrema in Lisad spaces are found by comparing to nearest neighbors (total of 26 neighbors for a 3D volume). For Lisad-1, we have used only the first scale since most edges in these datasets are very sharp and are captured by small scale derivatives of Gaussian. For occlusion detection, we consider pairs of rays with angle difference larger than six steps of the angle parameter. When their angle difference is smaller, we keep both rays. For ray elimination, the ray needs to be within one step of angle difference from its bounding 3D edges. All parameters are the same for HCI datasets. Finally, note that our horizontal EPIs are extracted for the middle vertical views and vertical EPIs for the middle horizontal views. That means that we are using only the cross-hair views.

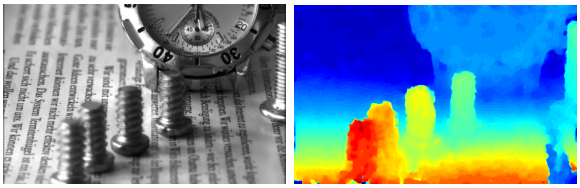
To compare with the benchmark datasets, we use two metrics: the mean squared error (MSE) of disparity estimates and the percentage of pixels with depth error smaller than 1%. The table with disparity MSE for LF datasets, for a range of depth estimation algorithms, is given in Figure 7 in [22]. We compare our algorithm to the best reported value in [22], [9], for Buddha and Mona datasets. These two datasets have 9 views in each horizontal and vertical directions. Middle views are shown in Figure 7. We have chosen these two datasets because our second metric (depth accuracy) is reported only for these two datasets in Figure 7 in [9]. For other datasets, [9] reports only average values. Among all prior art algorithms (therefore excluding ours), the algorithm that gives the smallest disparity MSE value

(a) Buddha (b) Mona

Fig. 7. Middle views for the Buddha and Mona datasets.

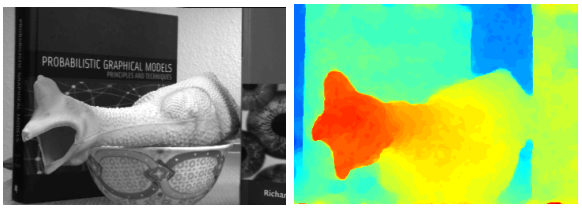
Dataset	Metric	Ours	Best prior work
Buddha	disparity MSE	**0.0048**	0.0055 [23]
	$< 1\%$ depth error	**98.8%**	96.5% [9]
Mona	disparity MSE	**0.0061**	0.0082 [23]
	$< 1\%$ depth error	**97.6%**	95.4% [9]

TABLE I. Results for Buddha and Mona datasets in terms of the mean squared error (MSE) of disparity values and the percentage of pixels with depth error less than 1%.



(a) Middle view (b) Disparity map

Fig. 9. Disparity estimation for the RAYTRIX® "watch" image.



(a) Middle view (b) Disparity map

Fig. 10. Disparity estimation for the "chicken" image from our plenoptic camera prototype.

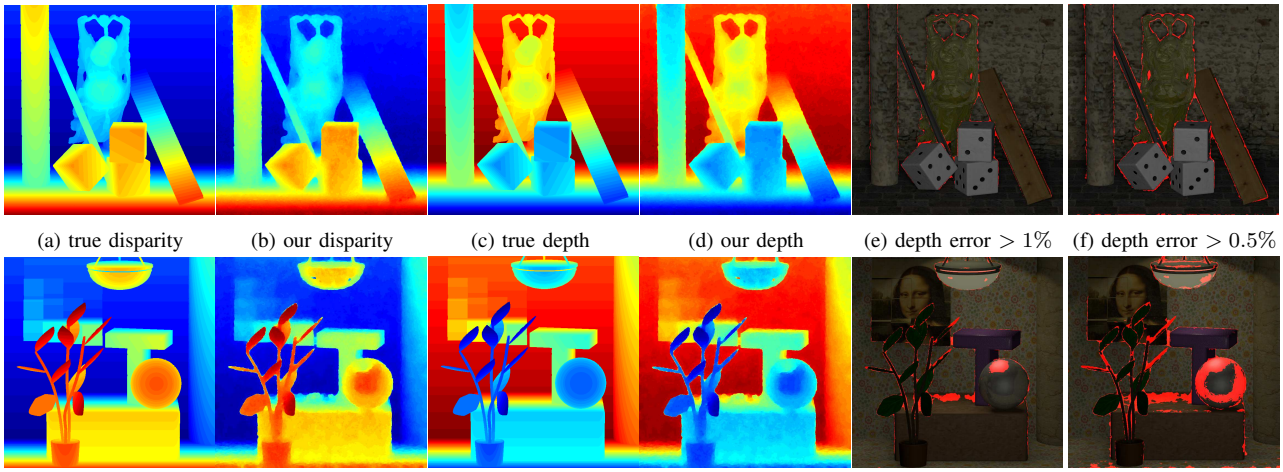
on the HCI benchmark is the algorithm of [23] (performs constrained denoising on each epipolar plane image and takes into account occlusion ordering constraints), while the algorithm that is the best with respect to the second metric is [9] (structure tensor local estimate + TV-L1 denoising). We thus compare our algorithm to these two algorithms using corresponding metrics.

Table I shows the obtained disparity and depth accuracy for Buddha and Mona, for the two above mentioned metrics, compared to previously best reported results on the benchmark. We can see that our method outperforms the best prior art, both in terms of disparity MSE and percentage of pixels with depth error less than 1%. Also note here that the prior methods use all views (81 for these datasets), while our method uses only the cross-hair (17 views). We expect that adding other views will increase the accuracy of our method, making it even more advantageous to prior work. We show visually the ground truth and obtained disparity and depth maps for Buddha in Figure 8(a-d) and Mona in Figure 8(g-j). To see the pixels whose depth errors are larger than a certain threshold (1% and 0.5%) we display them in red overlaid on the original images in Figures: 8(e-f) and (k-l), for Buddha and Mona respectively. We can see that most errors are located around edges and specular reflections.

Finally, we show disparity estimation results for the "watch" image from the RAYTRIX® camera, obtained from the HCI database, and for the "chicken" image from our own plenoptic camera prototype, in Figures 9 and 10, respectively. The angle here is sampled with 64 steps uniformly in the angle space, since we do not have conversion to depth. For the RAYTRIX® dataset, we use a larger set of octaves: six for the Lisad-2 space and two for the Lisad-1 space. One reason for this sampling is that the "watch" LF contains larger uniform regions. Another reason is that, due to the processing applied when converting a plenoptic image to the LF [24] during the creation of the benchmark set, the LF contains blurry edges. We can see that for both watch and chicken images we get a dense and smooth disparity map even for the uniform regions.

V. Conclusions

We have presented a novel method for dense depth estimation from light fields, based on extrema detection in continuous light field scale-depth spaces built upon the normalized second derivative of the Ray Gaussian kernel. We have proven theoretically the scale-invariance of such scale-depth spaces and that their values do not exhibit depth bias. Furthermore, we have formulated the depth estimation problem as a ray detection problem, which is solved by finding extrema in the formed scale-depth spaces. Detected rays further allow for efficient occlusion detection by finding overlapping rays. We then propose a set of ray combination and post-processing steps to enhance the quality of depth maps. We show that our method outperforms the best values previously reported on the HCI benchmark, for Buddha and Mona datasets. Our method is purely local and has potential for efficient implementation. We are working on building datasets with objects containing more uniform regions, where our scale-depth space formulation and ray-bundle detection can offer more benefits than for highly textured datasets presented in the HCI benchmark.



(a) true disparity (b) our disparity (c) true depth (d) our depth (e) depth error $> 1\%$ (f) depth error $> 0.5\%$

(g) true disparity (h) our disparity (i) true depth (j) our depth (k) depth error $> 1\%$ (l) depth error $> 0.5\%$

Fig. 8. Visual comparison of estimated depth/disparity maps to the ground truth depth/disparity maps for Buddha (a-f), Mona (g-l).

VI. Acknowledgements

Raytrix is a registered trademark of the Raytrix GmbH. Blender may be a trademark of NaN Holding B.V. in the U.S. and/or other countries.

References

[1] R. Horstmeyer, G. Euliss, R. Athale, and M. Levoy, "Flexible multimodal camera using a light field architecture," in *IEEE International Conference on Computational Photography*, 2009. 1

[2] R. Ng, M. Levoy, M. Brédif, and G. Duval, "Light field photography with a hand-held plenoptic camera," *Technical Report CSTR*, 2005. 1, 2

[3] C. Perwass and L. Wietzke, "Single lens 3d-camera with extended depth-of-field," in *Proceedings of SPIE Electronic Imaging*, 2012. 1, 2

[4] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 31–42. 1

[5] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," *Computational models of visual processing*, vol. 1, no. 2, pp. 3–20, 1991. 1

[6] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proceedings of the European Conference on Computer Vision*, 2002. 1, 2, 3

[7] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4d light fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1, 3

[8] 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*, vol. 1, no. 1, pp. 7–55, 1987. 2, 3

[9] S. Wanner and B. Goldluecke, "Variational Light Field Analysis for Disparity Estimation and Super-Resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606–619, 2014. 2, 3, 6, 7

[10] I. Tošić and K. Berkner, "3d keypoint detection by light field scale-depth space analysis," *Submitted to the IEEE International Conference on Image Processing (ICIP)*, 2014. 2, 3, 4, 5

[11] A. Criminisi, S. B. Kang, R. Swaminathan, R. Szeliski, and P. Anandan, "Extracting layers and analyzing their specular properties using epipolar-plane-image analysis," *Computer Vision and Image Understanding*, vol. 97, no. 1, pp. 51–85, 2005. 2, 3

[12] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4d light fields," in *Vision, Modelling and Visualization (VMV)*, 2013. 2

[13] 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)*, vol. 32, no. 6, p. 166, 2013. 2

[14] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 3

[15] A. Gelman, J. Berent, and P. Dragotti, "Layer-based sparse representation of multiview images," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–15, 2012. 3

[16] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 73, 2013. 3, 6

[17] A. Witkin, "Scale-space filtering: A new approach to multi-scale description," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1984. 3

[18] T. Lindeberg, *Scale-Space*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2007. 3

[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. 3

[20] T. Lindeberg, "Feature detection with automatic scale selection," *International journal of computer vision*, vol. 30, no. 2, pp. 79–116, 1998. 3

[21] A. Chambolle, "An Algorithm for Total Variation Minimization and Applications," *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1-2, pp. 89–97, 2004. 6

[22] S. Wanner, C. Straehle, and B. Goldluecke, "Globally consistent multi-label assignment on the ray space of 4d light fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 6

[23] B. Goldluecke and S. Wanner, "The variational structure of disparity and regularization of 4d light fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 7

[24] S. Wanner and B. Goldluecke, "Spatial and angular variational super-resolution of 4d light fields," in *Proceedings of the European Conference on Computer Vision*, 2012. 7