A light field FDL-HSIFT feature in scale-disparity space

Zhaolin Xiao, M Eng Zhang, Haiyan Jin, Christine Guillemot

A LIGHT FIELD FDL-HSIFT FEATURE IN SCALE-DISPARITY SPACE

Zhaolin Xiao[1,2], *Meng Zhang*[1], *Haiyan Jin*[1,2], *Christine Guillemot*[3]

1. Xi'an University of Technology, Xi'an, China, 710048
2. Shaanxi Key Laboratory for Network Computing and Security Technology, Xi'an, China, 710048
3. Institut National de Recherche en Informatique et en Automatique, Rennes, France

ABSTRACT

Many computer vision applications rely on feature matching, hence the need for computationally efficient and robust 4D light field (LF) feature detectors and descriptors for applications using this imaging modality. In this paper, we propose a novel LF feature extraction method in the scale-disparity space, based on a Fourier disparity layer representation. The proposed feature extraction takes advantage of both the Harris feature detector and SIFT descriptor, and is shown to yield more accurate feature matching, compared with the LiFF light field feature with low computational complexity. In order to evaluate the feature matching performance with the proposed descriptor, we generated synthetic LF datasets with ground truth matching points. Experimental results with synthetic and real datasets show that, our solution outperforms existing methods in terms of both feature detection robustness and feature matching accuracy.

Index Terms— Feature detection, feature descriptor, feature matching, light fields, Fourier disparity layer.

1. INTRODUCTION

In the past decades, 2D feature detectors such as SIFT [1], SURF [2], FAST [3], and ORB [4] has been instrumental in many computer vision applications with 2D images, such as structure-from-motion, 3D reconstruction, object tracking, or scene recognition. However, 2D image feature detection and matching can be significantly affected by occlusions, scale and illumination changes, and non-Lambertian reflections. It has been shown in [5] that 3D feature detection from RGB-D images can be more robust than 2D feature detection.

Light fields, unlike 2D imaging and RGB-D images, by recording the flow of rays emitted by the scene along different directions, yield a 4D spatio-angular representation of the scene, from which one can extract information about the parallax and depth of the scene. 4D LF features therefore hold promises to solve limitations of 2D image features in presence of occlusions, illumination changes and non Lambertian scenes. This is investigated in [6] and [7] where the authors exploit depth information in the LF to build scale-depth descriptors. Another category of approaches builds upon 2D descriptors, by computing 2D detectors on the different sub-aperture images, and then imposing angular consistency using epipolar geometry [8], [9] or using optical flows [10], [11]. The authors in [12] instead simultaneously consider all sub-aperture images and extend the SIFT descriptor to 4D LF by searching for features in a joint 4D scale-slope space, i.e. in the scale space as SIFT but at different depths, or slopes of structures in epipolar plane images.

Despite the above work, defining robust and computationally efficient 4D LF feature extractors and descriptors is still a widely open problem. One question inherent to the LiFF descriptor is the discretization of the depth space, which has obvious implications on computational complexity. The depth space discretization corresponds to a list of slopes, a higher number of slopes giving a better performance, but a higher computational complexity. The optimal list of slopes is not easy to determine. The authors recommend using as many slopes as there are samples in the LF angular dimension.

In this paper, we propose a novel 4D LF feature, called FDL-HSIFT feature, based on the Harris detector and SIFT descriptor computed on the Fourier disparity layer representation [13]. The FDL is a compact representation which samples the LF in the depth (or equivalently the disparity) dimension by decomposing the scene as a discrete sum of layers. The layers, and their corresponding disparity values, are automatically found, from a subset of LF views, using regularized least square regression performed in the Fourier domain, independently at each spatial frequency. The proposed feature is therefore defined in the 4D LF scale-disparity space, the disparity being discretized thanks the FDL construction. The representation being compact, it leads to a reduced computational complexity without loosing in terms of performance.

In summary, the contributions of the paper are as follows

- We introduce a FDL based scale-disparity space (FDL-SDIS), which benefits potential LF feature extraction.
- By combining the Harris detector and the SIFT descriptor, we propose a novel FDL-HSIFT blob feature in the FDL based scale-disparity space.
- We also created a synthetic dataset using blender which gives ground truth matches, in order to evaluate the proposed feature descriptor.

Corresponding author: Haiyan Jin. This work has been funded in part by the NSFC (No. 61871319 and No. 62031023), in part by the Natural Science Basic Research Plan of Shaanxi Province(No.2019JM-221), in part by the EU H2020 Research and Innovation Programme under grant agreement No 694122 (ERC advanced grant CLIM).

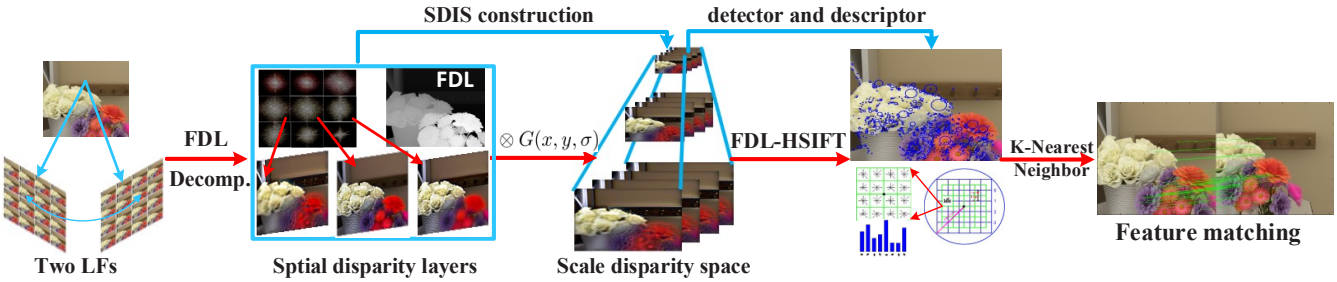

Fig. 1. Overview of FDL-HSIFT feature matching.

2. RELATED WORK

2.1. Image and LF based features extraction

Image feature detection refers to the problem of identifying and localizing interest points, blobs and regions. Classical 2D image feature detectors such as [14], or SUSAN [15], SIFT[1], FAST [3], SURF [2] and ORB [4] feature detectors have been widely employed in many computer vision application. These feature detectors are mainly based on specific image gradient distributions, which have local or global invariance to possible image translation, rotation, or to scale or affine transformation.

In parallel, in the past two decades, many acquisition devices have been designed to capture LF, ranging from camera arrays, [16], to single cameras mounted on moving gantries, and plenoptic cameras [17]. Overviews of these devices can be found in [18] and [19]. Several LF feature detectors and descriptors have been introduced in the literature [7, 8, 20, 12]. Tosic and Berkner [7] proposed to detect edge keypoints in the LF scale-depth (Lisad) space, constructed using a modified Gaussian kernel, and parameterized both in the object scale and depth. The authors in [8] use epipolar geometry to impose angular consistency between 2D descriptors computed in the different sub-aperture images, while the authors in [10], [11] instead use optical flows. These methods based on 2D features computed on the different sub-aperture images suffer from the limitations of 2D feature descriptors. Tsai *et al.*[20] instead propose a method to distinguish between refracted and Lambertian image features using a LF camera, based on textural cross-correlation to characterise apparent feature motion across the LF. The authors in [12] extend the SIFT descriptor to 4D LF to detect blob features in a joint 4D scale-slope space, i.e. in the scale space as SIFT but at different depths corresponding to slopes of structures in epipolar plane images. The method yields more precise feature matching in the context of structure from motion (SfM).

2.2. Light field representation

Let us consider the 4D LF representation proposed in [21] and [22] describing the radiance along rays by a function $LF(x, y, u, v)$, and based on a parameterization of orientations of light rays with two parallel planes. The pairs (x, y) and (u, v) represent the spatial and angular coordinates of light rays respectively.

By applying spatial or frequency refocusing [23], a LF can be represented by a sequence of images refocused at different depth planes, called focal stack. The 3D focal stack can also be inversely converted to a 4D representation (Sub-aperture images, SAIs), *e.g.* using the focal stack based deconvolution [24] [25], or recovering from the Fourier domain [26]. The LF can be instead represented by a set of layers, each one corresponding to a different disparity value [13], and computed using a fast disparity regularized least square regression in frequency domain.

For simplicity of notation, let us consider only one 2D slice of the light field with only one spatial and one angular dimension. The Fourier transform of the light field can be computed as [13]

$$\hat{L}(w_x, w_u) = \sum_k \delta(w_u - d_k w_x)\hat{L}^k(w_x) \qquad (1)$$

where w_x and w_u are spatial and angular frequency terms. $\delta(w_u - d_k w_x)$ is a dirac function, which simulates the aperture function with infinitely small aperture size. Each function $\hat{L}^k$ can be derived as,

$$\hat{L}^k(w_x) = \int_{\Omega_k} e^{-2i\pi x w_x} L(x, 0)dx \qquad (2)$$

and interpreted as the Fourier transform of the central view $L(x, 0)$ only considering a spatial region Ω_k of disparity d_k, hence the name Fourier Disparity Layers (FDL). More generally, the Fourier Transform $\hat{L}_{u_0}$ of L_{u_0} (a LF view at angular coordinate u_0 defined by $L_{u_0}(x) = L(x, u_0)$), given a set of n disparity values $\{d_k\}_{k\in[1,n]}$, can be decomposed as [13]

$$\hat{L}_{u_0}(\omega_x) = \sum_k e^{+2i\pi u_0 d_k \omega_x} \hat{L}^k(\omega_x). \qquad (3)$$

The FDL representation is therefore composed of the set of layers $\{L^k(x)\}$ (for a 2D slice) which can be derived from the inverse Fourier transform of $\hat{L}^k(w_x)$.

The FDL construction is done using linear regression which automatically finds the correct discretization in the depth or disparity space, leading to a more compact representation, compared to a focal stack, that has been shown efficient for various processing applications, e.g. rendering, view synthesis or varying aperture size and shape.

3. THE PROPOSED FDL-HSIFT FEATURE

3.1. FDL based scale-disparity space

We consider the FDL representation with the 4D notation, i.e. the set of layers $\{L^k(x,y)\}$ derived from the inverse Fourier transform of $\hat{L}^k(w_x, w_y)$, for feature extraction. The different layers define a discretization of the disparity space. To make the proposed feature robust to scale variance, by using a Gaussian kernel based scale transformation, the scale-disparity space (SDIS) is thus defined by the set of layers

$$\Psi^{k,\sigma}(x,y) = L^k(x,y,\sigma) = L^k(x,y) \otimes G(x,y,d_k,\sigma) \quad (4)$$

where $G(\sigma)$ is a Gaussian kernel associated with disparity d_k. We construct the representation $\Psi^{k,\sigma}(x,y)$ in the SDIS for each given input LF, number k of disparity layers and scale factor σ.

3.2. FDL-HSIFT Feature detection and matching

To detect a FDL-HSIFT feature, we first use a Harris detector in the SDIS representation. A displacement $(\Delta x, \Delta y)$ in the spatial dimension of the SDIS can be represented as,

$$\Psi^{k,\sigma}(\Delta x, \Delta y) = \sum_{x,y} \eta(x,y)[\Psi^{k,\sigma}(x+\Delta x, y+\Delta x) - \Psi^{k,\sigma}(x,y)]^2 \quad (5)$$

where $\eta(x,y)$ is a window function. By applying the Taylor expansion to Equation 5, the displacement is derived as,

$$\Psi^{k,\sigma}(\Delta x, \Delta y) = \sum_{x,y} [\Psi_x \Delta x, \Psi_y \Delta y]^2 = [\Delta x, \Delta y] M \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \quad (6)$$

where Ψ_x and Ψ_y are 1st-order partial derivatives of Ψ in the x and y directions respectively. The matrix M is the Harris matrix or structure tensor defined as

$$M = \sum_{x,y} \eta(x,y) \begin{bmatrix} \Psi_x^2 & \Psi_x \Psi_y \\ \Psi_x \Psi_y & \Psi_y^2 \end{bmatrix} \quad (7)$$

To detect corner points, edges or flat areas, we compute the response $R(x,y)$ of the Harris detector at each pixel of coordinates (x,y) as in [14]

$$R(x,y) = det(M) - \lambda \cdot (trace(M))^2 \quad (8)$$

where λ is the empirical coefficient within $[0.04, 0.06]$, and then simply select the top percentages of maximum response.

To represent the detected feature blobs, we employ a SIFT-like description in the SDIS space. For a given image blob with center (x,y), we compute a histogram of gradient orientations for every $\Psi^{(k,\sigma)}(\cdot)$. The histogram has 8 bins covering the 360 degree. Then, the highest peak is selected as the major direction, which, as in SIFT, preserves invariance to rotation. The gradient magnitude $m(x,y)$ and orientation $\theta(x,y)$ are computed as

$$m(x,y) = \sqrt{(Dx)^2 + (Dy)^2}$$
$$\theta(x,y) = arctan\frac{Dy}{Dx} \quad (9)$$

in which,

$$Dx = \Psi^{k,\sigma}(x+1,y) - \Psi^{k,\sigma}(x-1,y)$$
$$Dy = \Psi^{k,\sigma}(x,y+1) - \Psi^{k,\sigma}(x,y-1) \quad (10)$$

For a single $\Psi^{(k,\sigma)}$, the 128D descriptor $f(\Psi^{(k,\sigma)})$ consists of 16 blocks (4×4) with 8 directions in each. In this way, we extract the FDL-HSIFT descriptor $F(x,y)$ from a 16×16 neighborhood. $F(x,y)$ can be represented by

$$F(x,y) = \{f | f(\Psi^{(k,\sigma)}), (k,\sigma) \in SDIS\} \quad (11)$$

Algorithm 1 FDL-HSIFT matching algorithm

Input: LF_1, LF_2.
Output: Matching point set $(p_1,p_2)|p_1 \in Q_1, P_2 \in Q_2$.

1: Construct the FDL of each LF (LF_1 and LF_2) by inverse Fourier transform of Equation (2);
2: Construct the SDIS $\Psi_n^{(k,\sigma)}(x,y)$ for each LF indexed by n ($n = 1, 2$) by Equation(4);
3: **for** each pixel $Q_n = \{(x,y)|(x,y) \in \Psi_n^{(k,\sigma)}(x,y)\}$ **do**
4: apply the $Harris$ corner detector using Equation 8;
5: descriptor $f_n(\Psi_n^{(k,\sigma)}(x,y))$ is computed as $f_n(\Psi_n^{(k,\sigma)}(x,y)) \overset{SIFT}{\longleftarrow} \Psi_n^{(k,\sigma)}(x,y)$ as in [1];
6: compute the curvature ratio r by calculating Equation (12) between the features f_1 and f_2 at feature points $p \in Q_1$ and $p_2 \in Q_2$ in the two LFs
7: **if** $r < 0.6$ **then**
8: output $[p_1, p_2]$ as matching between LF_1 and LF_2.
9: **end if**
10: **end for**

When applying FDL-HSIFT feature matching, we use a cosine based metric for measuring the distance $< f_1(p), f_2(q) >$ between two feature vectors at two pixel positions in LF_1 and LF_2 respectively. The variables p and q denote the candidate matching coordinates in the two different LFs.

$$dist(p,q) = \max_{(k_1,\sigma_1),(k_2,\sigma_2)} (cos(f_1^{k_1,\sigma_1}(p), f_2^{k_2,\sigma_2}(q))),$$
$$\text{where} \quad f_1 \in F_1, f_2 \in F_2 \quad (12)$$

In this paper, we take the final matching decision by using a principal curvature ratio $r = dist(p, q_{max})/dist(p, q_{2nd})$. It is a positive matching only when $r < 0.6$, which means the distance p to the nearest point q_{max} is obvious less than the distance p to the second nearest point q_{2nd}. We summarize the FDL-HSIFT feature detection and matching algorithm in Algorithm 1. In the experiments, we set the number of FDL layers $k = 9$ and use 4 scale levels ($\sigma = 1.6$). These parameter settings may need to be adjusted for complex scene with large disparities.

4. EXPERIMENTS

Given that no LF dataset is available with feature matching ground truth, the authors in [12] evaluate their LiFF feature

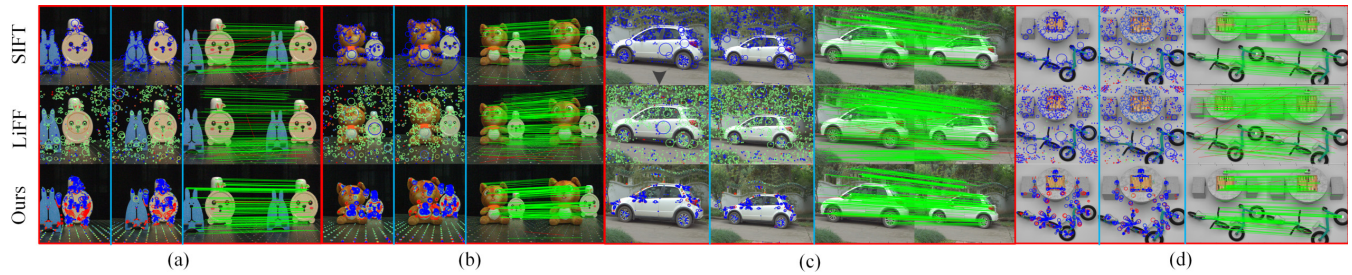

Fig. 2. Feature detection and matching results with different datasets. (a), (b) and (c) are results on real LFs; (d) is the result on synthetic LFs. For each dataset, the first two columns show detection results, the third columns is matching results. The circle size and color denote the scale and disparity of detected features.

in the context of a SfM algorithm. We instead created LF datasets with ground truth matching points. For each test data, we generate a pair of LFs with a known rotation, translation and camera settings. The LF includes 9×9 views, each view is 512×512 in spatial resolution, within the disparity range $[-2, 2]$ pixels. The central views of a pair of LFs are shown in figure 3(a). Using Blender, we can do a pixel-wise cross-check of matching (figure 3(c)) and feature matching (figure 3(d)) using ground truth depth (figure 3(b)).

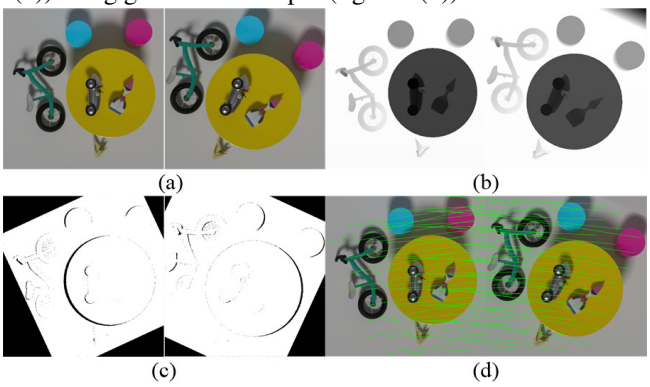

Fig. 3. Example of Blender LF matching dataset. (a) Central views of two LFs, with translation and rotation between the two; (b) Corresponding depth maps of (a); (c) Matching binary masks (black means that a matching point does not exist); (d) Pixel-wise matching ground truth of two LFs.

4.1. Feature detection and matching

First, we evaluate the detecting and matching performance of the proposed FDL-HSIFT, in comparison with the classical SIFT [1] and LiFF [12] descriptors. Since pixel-wise matching ground truth is available, we calculate both the precision and recall on the created synthetic LF datasets. The proposed FDL-HSIFT detection and matching outperforms the state-of-the-art LiFF feature (see Tab.1). Then, we conducted the same comparative assessment with real-world LF datasets. The results can be seen in Fig.2. Given that, in this case, the matching ground truth is not known, we check correct matches as in [12]. But depth estimation may not be very precise for every pixel. For this reason, we set the matching error tolerance to 20 pixels, which is the distance between the computed and the theoretical positions. As shown in Tab.1, our algorithm finds more feature blobs, and is also better in terms of precision.

Table 1. Comparison of feature matching on synthetic LFs (total matches, precision, recall in each grid)

method	chess	sofa	cabinets	office
SIFT	38	68	247	9
	0.868	0.956	**0.984**	0.556
	0.022	0.025	0.140	0.003
LiFF	121	169	178	51
	0.876	0.959	0.803	0.490
	0.070	**0.062**	0.082	**0.014**
Ours	171	121	393	29
	0.953	**0.992**	0.97	**0.724**
	0.108	0.046	**0.211**	0.012

Table 2. Runtime comparison

method	chess	sofa	cabinets	office
SIFT	421.95s	220.87s	585.80s	227.35s
LiFF	86.29s	55.52s	97.01s	**26.40s**
Ours	**62.93s**	**32.33s**	**82.72s**	34.18s

4.2. Computational complexity

Tab. 2 gives a runtime comparison. The runtime of the proposed feature matching pipeline includes three parts: time of SDIS construction, feature detection and description and feature matching. Assuming a 4D $LF(x, y, u, v)$ is decomposed into SDIS as $\Psi(x, y, k, \sigma)$, except for the time of SDIS construction, the computational complexity of FDL-HSIFT will be theoretically decreased $(u \times v)/(k \times \sigma)$ times than repeated SIFT. The computational complexity of FDL-HSIFT is lower than LiFF in most of cases. One reason is the compact representation of FDL with a lower number of layers compared to images in a focal stack, as used by the LiFF feature detector, making our algorithm more suitable for high angular resolution LFs.

5. CONCLUSION

In this paper, we propose a FDL-HSIFT feature detection and matching algorithm in LF scale-disparity spaces. To perform a quantitative analysis, we have created with Blender pairs of synthetic LF datasets, with ground truth matches. Experimental results show that the proposed algorithm has better precision and lower computational complexity compared to the state-of-the-art LiFF feature detector. Future work will be dedicated to analysis and optimization of FDL-HSIFT on real-world LFs, *e.g.* in presence of noise and distortion.

6. REFERENCES

[1] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *ECCV*, 2006.

[3] Edward Rosten and Tom Drummond, "Machine learning for high-speed corner detection," in *ECCV*. Springer, 2006, pp. 430–443.

[4] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*, 2011, pp. 2564–2571.

[5] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *CVPR*, 2013, pp. 564–571.

[6] Alireza Ghasemi and Martin Vetterli, "Scale-invariant representation of light field images for object recognition and tracking," in *Computational Imaging XII*. International Society for Optics and Photonics, 2014, vol. 9020, p. 902015.

[7] Ivana Tošić and Kathrin Berkner, "3d keypoint detection by light field scale-depth space analysis," in *ICIP*, 2014, pp. 1927–1931.

[8] José Abecasis Teixeira, Catarina Brites, Fernando Pereira, and João Ascenso, "Epipolar based light field key-location detector," in *International Workshop on Multimedia Signal Processing*, 2017, pp. 1–6.

[9] Ole Johannsen, Antonin Sulc, and Bastian Goldluecke, "On linear structure from motion for light field cameras," in *ICCV*, 2015, pp. 720–728.

[10] Kazuki Maeno, Hajime Nagahara, Atsushi Shimada, and Rin-ichiro Taniguchi, "Light field distortion feature for transparent object recognition," in *CVPR*, 2013, pp. 2786–2793.

[11] Yichao Xu, Hajime Nagahara, Atsushi Shimada, and Rin-ichiro Taniguchi, "Transcut: Transparent object segmentation from a light-field image," in *ICCV*, 2015, pp. 3442–3450.

[12] Donald G Dansereau, Bernd Girod, and Gordon Wetzstein, "Liff: Light field features in scale and depth," in *CVPR*, 2019, pp. 8042–8051.

[13] M. Le Pendu, C. Guillemot, and A. Smolic, "A fourier disparity layer representation for light fields," *IEEE TIP*, vol. 28, no. 11, pp. 5740–5753, 2019.

[14] Christopher G Harris, Mike Stephens, et al., "A combined corner and edge detector.," in *Alvey vision conf.* Citeseer, 1988, vol. 15, pp. 10–5244.

[15] Stephen M Smith and J Michael Brady, "Susan: new approach to low level image processing," *International journal of computer vision*, vol. 23, no. 1, pp. 45–78, 1997.

[16] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy, "High performance imaging using large camera arrays," in *ACM SIGGRAPH*, pp. 765–776. 2005.

[17] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan, "Light field photography with a hand-held plenoptic camera," in *ACM SIGGRAPH*, 2005, pp. 735–744.

[18] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 926–954, 2017.

[19] I. Ihrke, J. Restrepo, and L. Mignard-Debise, "Principles of light field imaging: Briefly revisiting 25 years of research," *IEEE Signal Processing Magazine*, vol. 33, no. 5, pp. 59–69, 2016.

[20] Dorian Tsai, Donald G Dansereau, Thierry Peynot, and Peter Corke, "Distinguishing refracted features using light field cameras with application to structure from motion," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 177–184, 2018.

[21] Marc Levoy and Pat Hanrahan, "Light field rendering," in *Annual conf. on Computer graphics and interactive techniques*, 1996, pp. 31–42.

[22] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen, "The lumigraph," in *Proceedings of the 23rd annual conf. on Computer graphics and interactive techniques*, 1996, pp. 43–54.

[23] Ren Ng, *Fourier slice photography*, Ph.D. thesis, Stanford University, 2006.

[24] A. Levin and F. Durand, "Linear view synthesis using a dimensionality gap light field prior," in *CVPR*, 2010, pp. 1831–1838.

[25] K. Kodama and A. Kubota, "Efficient reconstruction of all-in-focus images through shifted pinholes from multi-focus images for dense light field synthesis and rendering," *IEEE TIP*, vol. 22, no. 11, pp. 4407–4421, 2013.

[26] F. Perez, A. Perez, M. Rodriguez, and E. Magdaleno, "Light field recovery from its focal stack," *Journal of Mathematical Imaging and Vision*, vol. 56, no. 3, pp. 1–18, 2016.