Tampere University

SUREN VAGHARSHAKYAN

Densely Sampled Light Field Reconstruction

Tampere University Dissertations 273

Tampere University Dissertations 273

SUREN VAGHARSHAKYAN

Densely Sampled Light Field Reconstruction

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Information Technology and Communication Sciences
of Tampere University,
for public discussion at Tampere University
on 26 June 2020, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Information Technology and Communication Sciences
Finland

Responsible supervisor and Custos	Professor Atanas Gotchev Tampere University Finland	
Supervisor	Dr Robert Bregovic Tampere University Finland	
Pre-examiners	Dr Christine Guillemot INRIA Rennes France	Professor Gordon Wetstein Stanford University USA
Opponents	Professor Sebastian Knorr Ernst Abbe University of Applied Sciences Jena Germany	Dr Martin Alain Trinity College Dublin Ireland

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-1614-3 (print)
ISBN 978-952-03-1615-0 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1615-0

PunaMusta Oy – Yliopistopaino
Tampere 2020

PREFACE

This thesis is the result of six years spent as a doctoral student with the Faculty of Information Technology and Communication Sciences of Tampere University, formerly Tampere University of Technology. This long journey started with moving to Finland in early 2012 having applied for a Ph.D. position at the suggestion of Aram Danielyan, so he is the main culprit for my being in Finland.

I would like to thank my thesis pre-examiners Dr. Christine Guillemot and Prof. Gordon Wetzstein for their positive reviews and recommending publishing the draft.

I would like to express my sincere gratitude to Prof. Atanas Gotchev for all his guidance during my studies, for his help in reviewing my work, his continuous support, and the inspiration he gives.

I thank Dr. Robert Bregovic for his valuable comments and fruitful discussions. Sharing his expertise in signal processing helped me achieve the results presented in this dissertation. I also want to express my gratitude to my co-author, postdoc. Erdem Sahin whose skills in holography and wave optics were extremely valuable.

Additionally, I would like to thank Ahmed Durmush, Evgeniy Belyaev, Jani Mäkinen, Mihail Georgiev, Olli J. Suominen, Pavlo Molchanov, Sergio Moreschini and Ugur Akpinar, and to express my appreciation to all the many members of the 3D Media Research Group that I've had the pleasure of working with over the years.

ABSTRACT

The emerging light-field and holographic displays aim at providing an immersive visual experience, which in turn requires processing a substantial amount of visual information. In this endeavour, the concept of plenoptic or light-field function plays a very important role as it quantifies the light coming from a visual scene through the multitude of rays going in any direction, at any intensity and at any instant in time. Such a comprehensive function is multi-dimensional and highly redundant at the same time, which raises the problem of its accurate sampling and reconstruction.

In this thesis, we develop a novel method for light field reconstruction from a limited number of multi-perspective images (views).

First, we formalize the light field function in the epipolar image domain in terms of a directional frame representation. We construct a frame (i.e. a dictionary) based on the previously developed shearlet system. The constructed dictionary efficiently represents the structural properties of the continuous light field function. This allows us to formulate the light field reconstruction problem as a variational optimization problem with a sparsity constraint.

Second, we develop an iterative optimization procedure by adapting the variational in-painting method originally developed for 2D image reconstruction. The designed algorithm employs an iterative thresholding and yields an accurate reconstruction using a relatively sparse set of samples in the angular domain.

Finally, we extended the method using various acceleration approaches. More specifically, we improve its robustness by an additional overrelaxation step and make use of the redundancy between different color channels and between epipolar images through colorization and wavelet decomposition techniques.

Extensive experiments have demonstrated that these methods constitute the state of the art for light field reconstruction. The resulting densely-sampled light fields have high visual quality which is beneficial in applications such as holographic stereograms, super-multiview displays, and light field compression.

CONTENTS

ABBREVIATIONS

BP	basis pursuits
CNN	convolutional neural network
dB	decibel
DCT	discrete cosine transform
DSEPI	densely sampled epipolar-plane image
DSLF	densely sampled light field
EPI	epipolar-plane image
essBW	essential bandwidth
FFoV	finite field of view
FSW	finite scene width
GCD	greatest common divisor
LF	light field
PSNR	peak signal-to-noise ratio
RIP	restricted isometry property

ORIGINAL PUBLICATIONS

Publication I S. Vagharshakyan, R. Bregovic and A. Gotchev. Image Based Rendering Technique via Sparse Representation in Shearlet Domain. *2015 IEEE International Conference on Image Processing (ICIP)*. Sept. 2015, 1379–1383.

Publication II E. Sahin, S. Vagharshakyan, J. Mäkinen, R. Bregovic and A. Gotchev. Shearlet-Domain Light Field Reconstruction for Holographic Stereogram Generation. *2016 IEEE International Conference on Image Processing (ICIP)*. Sept. 2016, 1479–1483.

Publication III S. Vagharshakyan, R. Bregovic and A. Gotchev. Accelerated Shearlet-Domain Light Field Reconstruction. *IEEE Journal of Selected Topics in Signal Processing* 11.7 (Oct. 2017), 1082–1091.

Publication IV S. Vagharshakyan, R. Bregovic and A. Gotchev. Light Field Reconstruction Using Shearlet Transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.1 (Jan. 2018), 133–147.

1 INTRODUCTION

Modern display systems aim at creating visual content that can be perceived by humans as a natural scene. A number of display prototypes that specifically address the issue of higher realism by recreating 3D visual cues have been presented in recent years. These include super multi-view„ integral imaging, light field, and holographic displays [97], [96], [87] . In general, these systems require a high amount of information to accurately depict the desired 3D scene. In practice, visual information acquisition systems can vary from a gantry moving a single conventional camera, through arrays of cameras, to the recently developed plenoptic cameras and complex systems composed of multiple sensors. All these are technology-limited. Hence, robust and efficient methods are needed to close the gap between the amount of data that can be feasibly sensed and the visual content required by the new generation of display systems. To formalize such methods, both the sensed and required information should be presented in a common framework.

A common approach used in computer graphics is to model and reconstruct the scene using geometrical primitives [30], [51], [42], [79]. More specifically, sensed visual information is used to model the underlying 3D scene in terms of geometric structure and corresponding material information. The latter represents the complex light-scattering distribution properties of the scene's surface. If such accurate decomposition is achieved, the required visual information can be subsequently rendered from the model. Various methods have been used to address particular difficulties in finding such a decomposition or model.

An alternative concept of image-based rendering, or light field, has been introduced in [38]. This concept provides a generalized framework for modeling the visual information of a given scene in terms of a multidimensional function with corresponding properties. In contrast to the scene reconstruction approach, image-based rendering does not explicitly estimate and reconstruct geometrical information. Instead, multiperspective images are used to model a continuous light field function.

This helps to efficiently describe both the required and acquired visual information as differently-sampled versions of the same continuous light field function. This raises the fundamental problem of its reconstruction from a given set of measurements, which in turn encompasses the related problems of parametrization, acquisition, processing, and analysis.

1.1 Scope of the thesis

The main goal of this thesis is to develop an effective and efficient method for continuous light field reconstruction from a limited number of multi-perspective images (views) representing a 3D visual scene. In contrast with conventional methods, which employ scene geometry information in terms of depth maps to synthesise missing views, we consider the light field reconstruction as a multi-dimensional signal sampling and reconstruction problem and employ modern sparsification approaches to solve it.

The solution is sought by exploiting the light field function's structural redundancy. In our concept, the acquisition system consists of a rectified camera array which can efficiently describe the correspondences between the captured images. These correspondences are formalized by the respective epipolar-plane images (EPI) and the distinct directed-line structure therein. We favor the use of a discrete representation, referred to as densely-sampled light field (DSLF), which serves as an intermediary between the sensed images and the visual data needed to drive any display. With respect to DSLF, the sensed images form EPIs with holes. Therefore, from a practical point of view, the reconstruction problem is reformulated as an inpainting problem of filling in these holes.

Based on a detailed analysis of the light field's structure and properties, we formulate a regularized EPI reconstruction method utilising the sparsity in the shearlet transform domain. This method is generalized for the case of continuous light field function reconstruction. The redundancies in the color, spatial and angular domains are properly incorporated in the corresponding variational optimization approaches. The method's performance is demonstrated by integrating it into important applications.

1.2 Structure of the thesis

Chapter 2 covers the required theoretical preliminaries. It presents definitions of the light field, and essential results about its parameterization, sampling and reconstruction. Sufficient sampling conditions for various basic 3D scenes are derived using Fourier analysis. The current state of the art in light field reconstruction is reviewed.

The main contribution as a novel light field reconstruction method is presented in Chapter 3.It includes the problem formulation and the required theoretical considerations underpinning the proposed method in sufficient detail.

Chapter 4 presents various approaches aimed at accelerating the main method, whereas Chapter 5 presents several important applications which are improved by using our proposed reconstruction method. Chapter 6 summarises the work and offers some concluding remarks.

1.3 Links to the publications

The dissertation encompasses four publications. An early work on intermediate view synthesis formulated in terms of in-painting is presented in Publication I. It contains the first version of the main method, where the regularization is formulated in terms of sparsity in the shearlet transform domain (Section 3.2). Further theoretical and practical extensions of the proposed method along with its extensive evaluation are presented in Publication IV. Particularly, it develops the modified shearlet frame for efficient interpolation which is presented in Section 3.3. Publication III develops various accelerated versions of the main method, achieved by utilizing colorization techniques and inter-EPI decorrelation methods (Section 4). Publication II presents the importance and efficiency of our reconstruction method for the case of holographic stereogram generation, as presented in Section 5.

The author of the thesis is the first author for publications P.I, P.III, and P.IV and was in charge of the analysis, implementation, experimental evaluation and the scientific write-up of the presented methods. The author is the second listed author in P.II. He contributed to developing the reconstruction method and generating experimental results (Sections 3 and 4 in P.II). The first author, postdoctoral researcher Erdem Sahin, had a leading role in formulating the theoretical background connecting the holographic stereogram formation with the DSLF representation (Section 2 in P.II).

2 PRELIMINARIES

This chapter presents the general theoretical framework which forms the basis for developing the novel contributions presented later in the thesis. It summarises the mathematical formalization of the light field, its modeling and parametrization, and the sampling and reconstruction in the spatial and frequency domains. It also overviews the most recent research studies on light field reconstruction based on depth estimation, machine learning and sparse representation.

2.1 Light field modeling and parametrization

Light is the medium which conveys information about the visual world. Its formalization can be done following the plenoptic function concept, as proposed by Adelson and Bergen [1]. In this concept, light from a scene is modeled as a dense field of rays, where each ray is parameterized by its location in the 3D space (V_x, V_y, V_z), its direction (θ, ϕ), wavelength λ and time t. Thus, the resulting plenoptic $7D$ function, representing the field, is a function of seven variables

$$P = P(\theta, \phi, \lambda, t, V_x, V_y, V_z).$$

The plenoptic function can be simplified by considering fixed time instants and wavelength (e.g. for only the primary colours). Essentially, this reduces the number of corresponding parameters,yielding a 5D light field formed by a set of panoramic images at different 3D locations , expressed either in polar or Cartesian coordinates [66].

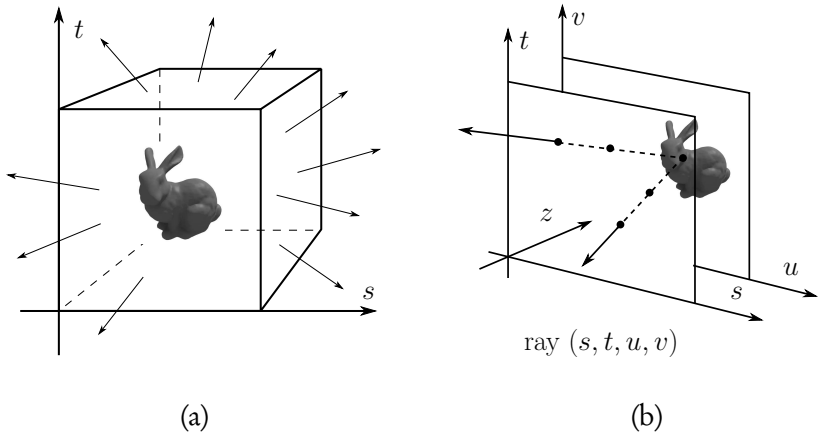


Figure 2.1 (a) Light radiance information can be described by considering radiance over the rays intersecting the cube sides. (b) Two-plane parameterization, where an arbitrary ray is parameterized using intersection points with two parallel planes.

2.1.1 The Lumigraph and 4D Light Field

In an attempt to further simplify the light field representation and make it practical for a finite scene, Gortler et al. have proposed the Lumigraph [38]. This representation considers a 3D scene placed inside a virtual cube. Each of the cube's six faces is parametrized by a pair of orthogonal axes s and t, defining a plane, as shown in Fig. 2.1 (a). The direction of any ray coming from the scene is then parametrized by the intersection with the (s,t) plane and a second plane parallel to it, as illustrated in Fig. 2.1 (b). Any 4D point (s,t,u,v) maps to a ray, thus giving rise to a 4D light field representation, similar to the plenoptic function for finite-size scenes.

A two-plane light field parametrization, leading to a 4D light field, as in Fig. 2.1 (b), has been independently proposed by Levoy and Hanrahan [59], emphasizing its usability for scenes in an occlusion-free region of space.

Both the Lumigraph and the two-plane parametrization offer simple yet effective ways of describing light fields, and are highly applicable for scene analysis and view synthesis. In particular, they allow a continuous light field L to be represented as a weighted sum of its discrete samples, in generic signal processing fashion. Given the coefficients $x_{i,j,p,q}$, associated with samples at locations (i,j,p,q), and introducing localized 4D reconstruction kernels $B_{i,j,p,q}(s,t,u,v)$, the continuous light field $\tilde{L}$ is

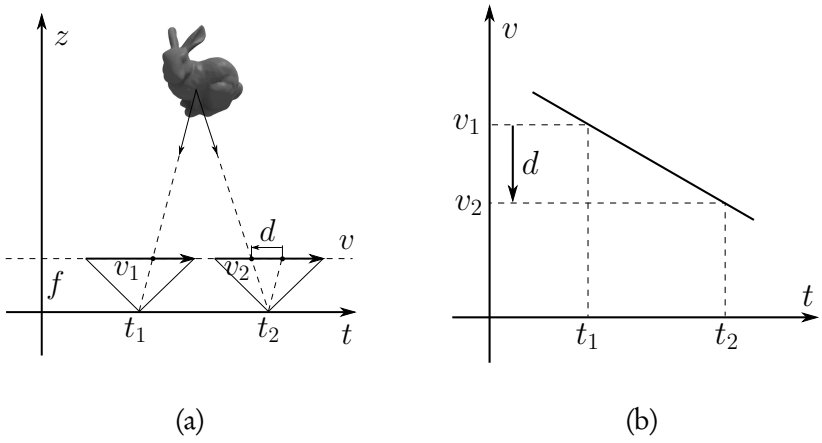


Figure 2.2 Epipolar-plane image formation using a pinhole camera moving over a fixed line. (a) Captured data can be interpreted as a two plane parameterization of a light field for fixed values of the s and u axes. (b) Corresponding epipolar-plane image where the radiance of the same 3D point is observed from different camera positions distributed over the straight line.

obtained by

$$\tilde{L}(s,t,u,v) = \sum_i \sum_j \sum_p \sum_q x_{i,j,p,q} B_{i,j,p,q}(s,t,u,v).$$

For computational efficiency, the use of quadrilinear interpolation has been proposed in [38]. This choice of reconstruction kernels tolerates a certain level of artifacts caused by the lack of any band-limited property of the light field.

2.1.2 Epipolar-plane image

The concept of an epipolar-plane image (EPI) originates from the stereo vision geometry referred to as *epipolar geometry* [42]. Consider a stereo pair of images formed under the pinhole camera model with optical centres c_L and c_R and epipoles e_L, e_R. A world point P, its projections on the left and right images, p_L and p_R, the optical centres, and the epipoles all lie on an *epipolar plane*. Correspondingly, given the projection point p_L, its corresponding point p_R lies on the epipolar line $p_R - e_R$, according to the *epipolar constraint* [42]. This constraint allows the search for matching features to be simplified from a two-dimensional problem to a one-dimensional problem (e.g. by rectifying the left and right images so that they lie on a single plane,

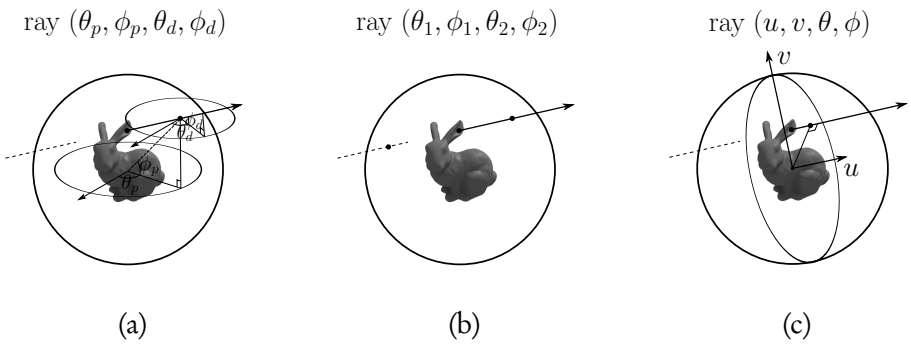


Figure 2.3 Identical ray in different spherical light field parameterizations.

which makes all the epipolar lines parallel to the baseline).

The epipolar geometry has been generalized for the case of multiple images by Bolles et al. in [10]. They attempted to describe a static 3D scene by a dense sequence of images, which are interpreted as a solid block of data with equal temporal and spatial continuity. The authors have demonstrated that, in the case of fixed camera locations, each slice of the solid data block has a distinct structure directly related to the scene's geometry [10].

The Bolles et al. approach is directly applicable to the two-plane parameterized LF. An EPI can be formed by fixing the parameters s and u and varying the parameters t and v. In this way, the slice $E_{(s,u)}(t,v) = l(u,v,s,t)$ is conventionally called a horizontal EPI and the slice $E_{(v,u)}(t,v) = l(u,v,s,t)$ is called a vertical EPI. EPIs implicitly characterize the scene's structure, as each object point is transformed into a line with a slope dependent on the point's depth. A simple example of an EPI is presented in Fig. 2.2.

2.1.3 Alternative parametrizations

Concrete visual acquisition systems usually motivate corresponding particular light field parametrizations. For example, spherical and cylindrical parametrizations have been introduced for the efficient representation of multiple images captured from the same location. Efficient methods for stitching multiple images acquired with conventional cameras into compound 360-degree cylindrical [22] or spherical panoramic images [78] have been proposed. An arbitrary virtual view can be synthesized by warping the corresponding panoramic image to simulate panning and zooming

effects. This approach can be interpreted as a light field interpolation from a fixed point using cylindrical or spherical parameterizations.

Concentric mosaics is a parametrization model where the plenoptic function is reduced to three dimensions [76]. Each ray is described by its radius, rotation angle, and vertical elevation. The corresponding acquisition system consists of a camera moving over planar concentric circles. Novel views are synthesized by combining the appropriate rays so that the required rays are obtained by composing slits of the captured images. This view synthesis only works when the corresponding viewpoint is located inside the planar circular region.

Spherical parametrization assumes finite-size scenes confined within a unit sphere [46]. The light ray is parametrized using an intersection point on the *positional* sphere (θ_p, ϕ_p) used as a convex hull of the scene, as illustrated in Fig. 2.3 (a). The direction of the ray is identified by the intersection point with the *directional* sphere (θ_d, ϕ_d). Two-sphere or spherical 4D light field parameterization is defined as a function $l^{\text{sphere}}(\theta_p, \phi_p, \theta_d, \phi_d)$, or as a superposition of two functions,

$$l^{\text{sphere}} = [l^p(\theta_p, \phi_p)](\theta_d, \phi_d) = l^d(\theta_d, \phi_d),$$

where each of them is defined on a sphere such as $l^p : V \to (V \to C), l^d : V \to C$ and $V = \{(\theta, \phi) | 0 \leq \theta < 2\pi, -\pi/2 \leq \phi \leq \pi/2\}$.

Alternative sphere-sphere (2SP) and sphere-plane (SPP) parameterizations have been proposed in [14]. Each ray is parameterized by its intersection points with the same sphere (2SP) Fig. 2.3 (b), or by its angle and the 2D coordinate of the intersection point of the ray and the orthogonal plane (SPP) Fig. 2.3 (c).

2.1.4 Acquisition systems

The developed parametrizations have stimulated the development of various light field acquisition systems. For example, spherical parametrization is utilized in the Stanford spherical gantry [59]. Another spherical LF capturing system has been introduced in [69]. It can be considered as an advance on the concentric mosaic and consists of two cameras rotated over a sphere surface. The captured spherical LF data allows an efficient synthesis of views located within the recorded spherical volume.

The two-plane parametrization is directly associated with an array of cameras (e.g. having a camera plane and an image plane). An example of a corresponding

acquisition system has been proposed in [91]. Identical data can be acquired using a conventional camera moved by a gantry as demonstrated in [92]. The so-called light field cameras, or plenoptic cameras, have been introduced in [34, 68]. In contrast to conventional cameras, plenoptic cameras accommodate an additional micro-lens array located between the main lens and the sensor. The sensed data is subsequently interpreted as a narrow-baseline, two-plane paremeterized and uniformly-sampled LF. Such aqcuisition systems have proved themselves to be usable, particularly in microscopy [60].

Later in this dissertation, the two-plane parameterization is adopted as the main way to describe light fields. Acquisition systems with a relatively wide baseline, e.g. using camera array or gantry, are considered.

2.2 Light field sampling and reconstruction

The aim of light field rendering is to synthesize any arbitrary light ray or slice given a set of rays or images representing a sampled version of the continuous light field. In a classical signal sampling and reconstruction approach, the task requires specifying the critical sampling rate and designing the corresponding anti-aliasing filter. This, in turn, requires an analysis of the light field bandwidth, i.e. its support in a Fourier domain.

This section overviews the classic works on LF sampling and reconstruction, mainly based on the seminal paper by Chai et al. [20]. The two-plane parameterization is used as a starting formalization and most of the results are derived for scenes with Lambertian reflectance.

2.2.1 Light field representation in a Fourier domain

Consider (s,t) to be the camera plane and (u,v) to be the image plane, as illustrated in Fig. 2.2 (a). An arbitrary ray intersecting both planes uniquely determines a quadruple $q = (u,v,s,t)$. Note, that in [20], the plane coordinates of (u,v) are defined relative to the (s,t) coordinates, in contrast to the global coordinate systems considered in the Lumigraph [38].

A sampled LF l_s is obtained with the use of a sampling function (pattern) $p(q)$

$$l_s(q) = l(q)p(q) \tag{2.1}$$

i.e. $p(q) = 1$ when q is from the sampling grid and $p(q) = 0$ otherwise. A continuous LF reconstruction is obtained with the use of a suitable anti-aliasing filter r

$$l_r(q) = [r * l_s](q).$$

To find the Fourier spectra of both l and l_s, the concept of epipolar-plane images (EPIs), as presented in Section 2.1.2, is further utilized, along with a few simplifications. First, an occlusion-free scene is considered. This implies that the same 3D point can be observed from any location on the camera plane. Second, the Lambertian reflectance of the scene surface guarantees constant radiance in different directions for a fixed point on the surface. Under these assumptions, every EPI is formed as a union of distinct lines, where each line corresponds to a particular scene point. The line intensity corresponds to the light radiance from the point in different directions and, for the Lambertian case, this is a constant. The disparity d of the observed 3D point between two images located at (s,t_1) and (s,t_2) can be calculated as $d = v_2 - v_1 = (t_1 - t_2)f/z$, where $z = z(q)$ represents the scene depth, i.e. the distance of the surface point corresponding to the q^{th}-ray from the camera plane, and f is the distance between the two planes.

Without loss of generality, $t_1 = 0$ is assumed to be the origin of the axis t. Thus, for the light field l, the following equation is valid

$$l(q) = l\left(u + \frac{f}{z(q)}s, v + \frac{f}{z(q)}t, 0, 0\right).$$

This shows considerable redundancy in the 4D light field. For a constant-depth plane $z(q) = z_0$, as shown in [20], its Fourier transform takes the form:

$$L(\Omega_u, \Omega_v, \Omega_s, \Omega_t) = 4\pi^2 L'(\Omega_u, \Omega_v)\delta\left(\Omega_s - f\Omega_u/z_0\right)\delta\left(\Omega_t - f\Omega_v/z_0\right), \tag{2.2}$$

where $L'(\Omega_u, \Omega_v)$ is the Fourier transform of $l'(u,v) = l(u,v,0,0)$ and δ is the Dirac delta function. In this case, the support of the 4D function L on the 2D plane (Ω_v, Ω_t) is bounded by the line $\Omega_t = \Omega_v f/z_0$, as shown in Fig. 2.4 (a). Identically, on the plane

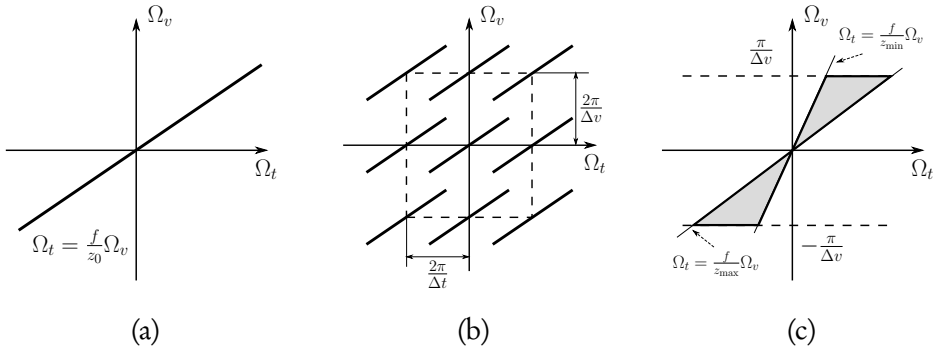


Figure 2.4 The Fourier transform support on (Ω_t, Ω_v) plane, (a) continuous light field with a constant depth, (b) sampled light field with a constant depth, (c) depth varies between $z_{\min}$ and $z_{\max}$.

(Ω_u, Ω_s), the corresponding line is $\Omega_s = \Omega_u f / z_0$.

Assume a uniform lattice sampling pattern defined by the sampling intervals Δu, Δv, Δs, and Δt. The corresponding sampling function is $p(q) = \text{Ш}_{\Delta q}(q)$, where $\text{Ш}_T(t)$ is the Dirac comb function and $\Delta q = (\Delta u, \Delta v, \Delta s, \Delta t)$. Thus, the Fourier transform of the sampled LF L_s at frequency $\Omega_q = (\Omega_u, \Omega_v, \Omega_s, \Omega_t)$ is given by the convolution

$$L_s(\Omega_q) = L(\Omega_q) * \text{Ш}_{2\pi/\Delta q}(\Omega_q).$$

This effectively creates replicas of the LF baseband. For the baseband of the constant-depth scene at z_0, illustrated in Fig. 2.4 (a), the spectrum of the sampled light field gets the form shown in Fig. 2.4 (b). Periodic replicas appear at intervals $\frac{2\pi}{\Delta t}$ along the Ω_t axis and $\frac{2\pi}{\Delta v}$ along the Ω_v axis.

A scene with spatially varying depth with depth values within the range of $[z_{\min}, z_{\max}]$ has support in the Fourier domain bounded by the lines $\Omega_t = f\Omega_v / z_{\max}, \Omega_t = f\Omega_v / z_{\min}$ as shown in Fig. 2.4 (c) [20].

2.2.2 Optimal sampling and minimum samplig curve

In terms of frequency-domain support, a reconstruction filter should retain the baseband and suppress its replicas. A direct approach is to apply a low-pass filter assuming a constant depth at infinity (c.f. Fig. 2.5 (a)). A better approach is to consider a directional filter at a constant-depth plane z_{opt}, where $z_{\text{opt}}^{-1} = (z_{\min}^{-1} + z_{\max}^{-1})/2$ (Fig. 2.5 (b)).).To avoid replica overlapping, the sampling interval between adjacent

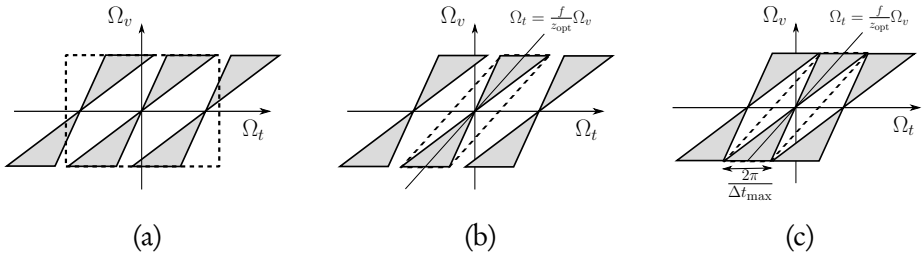


Figure 2.5 (a) Direct reconstruction filter based assuming with implicit assumption of infinite depth. (b) Filtering using z_{opt}. (c) Optimal packing in frequency domain is achieved in case of critical camera spacing distance $\Delta t_{\max}$.

cameras has to be sufficiently small, which in turn creates practical problems for the capture setting. It has been shown that the maximum camera spacing distance which ensures non-overlapping bands is defined by [20]:

$$\Delta t_{\max} = \frac{1}{K_{f_v} f \left(z_{\min}^{-1} - z_{\max}^{-1} \right)},$$

where $K_{f_v} = \min(B_v^s, 1/(2\Delta v), 1/(2\delta v))$ is the maximum frequency in the axis Ω_v. K_{f_v} depends on the complexity of the texture information, represented by the highest scene texture frequency B_v^s and on the rendering camera resolution δv. If the textural complexity is ignored and full-resolution images are rendered, then the maximal frequency is $K_{f_v} = 1/(2\Delta v)$.

The location of spectral replicas for the case of $\Delta t_{\max}$ is illustrated in Fig. 2.5 (c).

As shown in [20], the error in the case of maximum camera spacing $\Delta t_{\max}$ is $e = d_{z_{\min}} - d_{z_{\text{opt}}} = d_{z_{\text{opt}}} - d_{z_{\max}} = 1/(2K_{f_v})$. Thus, in the worst case, the error is $e = \Delta v$, i.e. one pixel.

Another result relating the scene geometry and sampling rate has been reported in [62]. The authors again consider occlusion-free scenes with Lambertian reflectance and they approximate the scene depth with a fixed depth plane, Z. Reconstruction through bilinear interpolation in ray space is considered. For a point with an actual depth of z_0, the error is defined as $|z_0 - Z|$. It has been demonstrated that in order to avoid ghosting artifacts due to aliasing in the novel view interpolation, the sampling distance between adjacent camera locations d should satisfy $\left| z_0^{-1} - Z^{-1} \right| d \leq \delta$, where δ is the camera's spatial resolution. This result implies that the sampling distance should be taken so as to ensure that the disparities between adjacent views do

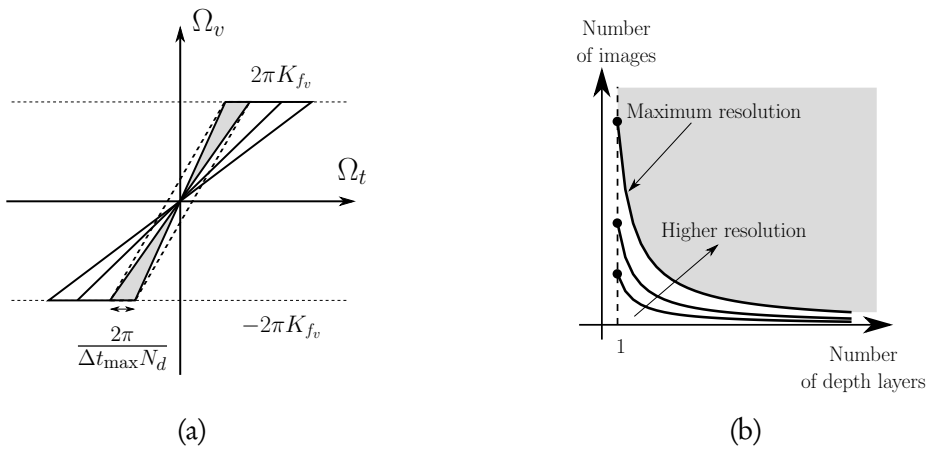


Figure 2.6 (a) Uniform multi-layer depth decomposition represented in the frequencies domain. (b) Minimum sampling curve for different rendering resolutions. Any point in the highlighted region represents redundancy for rendering in joint image and geometric space.

not exceed one pixel. An artifact-free LF rendering for a scene within a depth range $[z_{\min}, z_{\max}]$ can be achieved by specifying a constant-depth plane at

$$z_{\text{opt}} = \frac{2 z_{\min} z_{\max}}{z_{\min} + z_{\max}}.$$

This again suggests that the sampling rate on the camera plane is inversely proportional to both the uncertainty of the geometry information and the camera resolution.

The analysis presented so far can be generalised for the case of multiple depth layers. The idea is to represent the scene depth in terms of multi-layer decomposition, where each layer can be processed independently using the corresponding optimal filter for a constant-depth plane z_i:

$$z_i^{-1} = \lambda_i z_{\min}^{-1} + (1-\lambda_i) z_{\max}^{-1}, \quad \lambda_i = (i-0.5)/N_d, \quad i = 0, \ldots, N_d - 1.$$

In the Fourier domain, a uniform depth layering implies a division of the original support, as illustrated in Fig. 2.6 (a). Given N_d number of layers, the required minimum sampling rate is reduced and the achieved maximum camera spacing is related with the layer bandwidth $\Delta t_{\max,N_d} = \Delta t_{\max} N_d$.

In [20], an *optimal sampling curve* in the joint image and geometric space has been

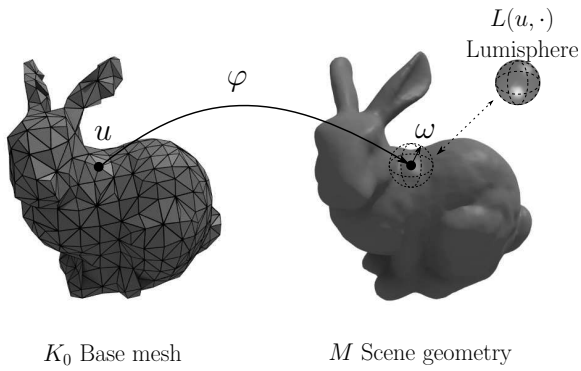


Figure 2.7 An illustration of the relation between base mesh and scene geometry in surface light field parameterization as described in [93].

derived, as shown in Fig. 2.6 (b). For the full parallax LF interpolation, it illustrates the optimal relation between the number of necessary images and the number of uniform depth layers as $N_d\sqrt{N_i} = K_{f_v}$.

Filtering based LF reconstruction using the optimal sampling curve still implies capturing of a significant number of images. Therefore more advanced methods are required to achieve similar reconstruction quality using a lower number of images.

2.3 Surface light field

A direct parameterization of the scene surface is an alternative to the depth layering approach, and this allows more complex scenes to be analysed [93].

The geometrical scene surface M is modelled by a projection $\varphi : K_0 \to M \in \mathbb{R}^3$ using the simplified base mesh K_0. The radiance of a light ray in direction ω at a point u on K_0 is defined as the surface light field (SLF) $L(u,\omega)$. It is simplified to a piece-wise linear function depending on the ω parameter, and for a fixed u it is referred to as a *lumisphere* [55]. Each *lumisphere* is recovered from a given set of images using the least-squares approximation. The quality of the reconstruction using the SLF parameterization greatly depends on the accuracy of the geometry of the approximated scene. However, it allows non-Lambertian scenes to be represented efficiently. The relation between the two-plane parameterized light field and the surface light field together with the related spectral analysis has been developed in [19], [25].

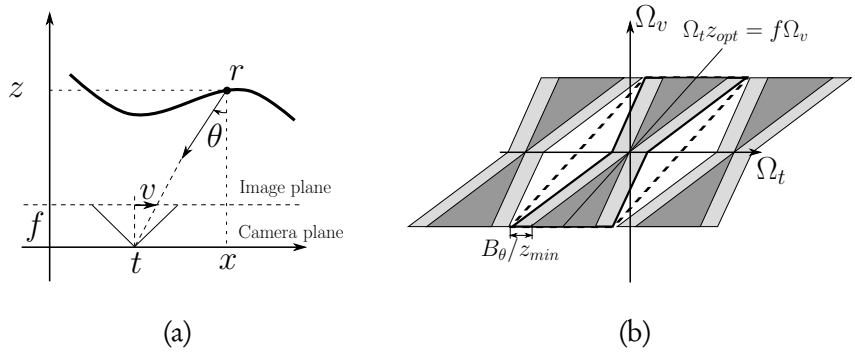


Figure 2.8 (a) The relation between the two-plane and SLF parameterizations. (b) The spectral support of the LF of a non-Lambertian surface.

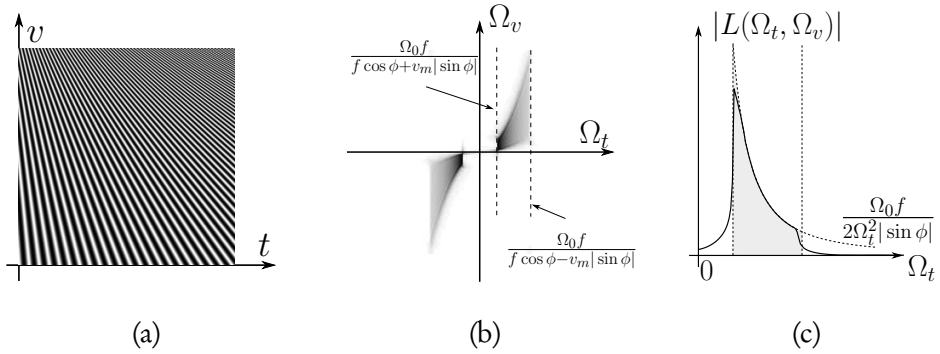


Figure 2.9 (a) Epipolar-plane image for a Lambertian scene of slanted plane cosine texture. (b) Corresponding magnitude in the frequency plane with theoretical limits (dashed lines). (c) The decay rate of the magnitude of the LF Fourier transform coefficients along Ω_t parameter.

Without loss of generality, the frequency analysis can be developed for 2D epipolar-plane image. Similarly, the SLF $l_s(r,\theta)$ is considered only on the plane (x,z), where r is an arc length and θ is the direction of an emitted light ray. The two-plane parameterization is given by $l(t,v)$ as shown in Fig. 2.8 (a). As seen in the figure, for the occlusion-free case, $(x-t)/z = v/f$. An approximation $\theta \approx v/f$ is valid for pinhole cameras with relatively narrow fields of view. For the scene with constant-depth, $z = z_0$, the two parameterizations are related as $l(t,v) = l_s(vz_0/f + t, v/f)$ and the corresponding Fourier transforms are related as

$$L(\Omega_v, \Omega_t) = f L_s(\Omega_t, f\Omega_v - z_0\Omega_t).$$

For a non-Lambertian scene, the SLF is not a band-limited function. Nevertheless,

as shown in [19], the following approximation of its Fourier transform L_s can be considered

$$L_s(\Omega_r, \Omega_\theta) \approx L_s(\Omega_r, \Omega_\theta) I_{B_\theta}(\Omega_\theta),$$

where I_{B_θ} is the indicator function with bandwidth B_θ. This approximation suggests that the surface's BRDF is band-limited, or for a fixed point on the surface, the light radiance slowly changes depending on the angle θ. As a consequence

$$L(\Omega_v, \Omega_t) = f L_s(\Omega_t, f\Omega_v - z_0\Omega_t) I_{B_\theta}(f\Omega_v - z_0\Omega_t)$$

and the spectrum has additional finite width $\frac{2B_\theta}{\sqrt{z_0^2+f^2}}$ perpendicular to its tilt. An illustration of this result is presented in Fig. 2.8 (b).

Depth that is modelled as a tilted plane is another case considered in [19]. For this, the relation between the two parameterizations is

$$l(v,t) = l_s\left(\frac{vz_0 + ft - fx_0}{f\cos\phi - v\sin\phi}, \frac{v}{f}\right)$$

where ϕ is the angle of the tilted plane starting at point (x_0, z_0).

To give a more illustrative example, a sinusoidal texture $l_s(r,\theta) = \cos(\Omega_0 r)$ is considered. For a limited field of view $|\theta| < v_m/f$, it has been shown that the magnitude of the Fourier transform L in the first quadrant ($\Omega_t > 0, \Omega_s > 0$) is [19]

$$|L(\Omega_t, \Omega_v)| = \begin{cases} \dfrac{\Omega_0 f}{2\Omega_t^2 |\sin\phi|}, \dfrac{\Omega_0 f}{f\cos\phi + v_m|\sin\phi|} \le \Omega_t \le \dfrac{\Omega_0 f}{f\cos\phi - v_m|\sin\phi|} \\ 0, \text{otherwise.} \end{cases} \quad (2.3)$$

This result is illustrated in Fig. 2.9 (a) for a slant $\phi = \pi/6$. One can clearly see the bounds in Eq. 2.3 along the Ω_t axis and the function decay (Fig. 2.9 (b) (c)).

In the case of a scene with occlusions, an approach using silhouettes has been proposed in [19]. The scene is initially divided into independent objects without self-occlusions. The corresponding LF is represented as a composition of LFs generated by each object separately, and the silhouettes are defined as masking functions representing occlusions. The spectrum of an occluded object is its unoccluded spectrum modulated by all the silhouettes of the occluding objects. The magnitude of the Fourier transform for a simple scene with three constant depth planes occluding

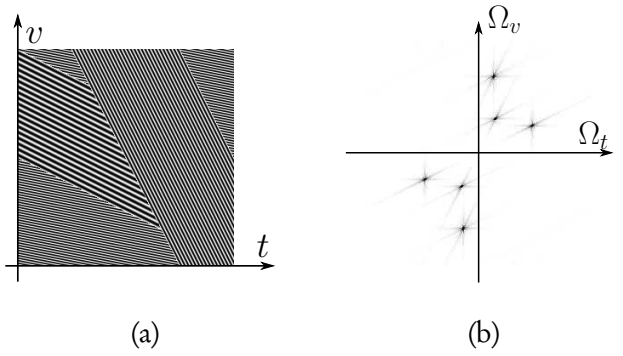


Figure 2.10 (a) Epipolar-plane image for a Lambertian scene consisting of three constant depth planes occluding each other. (b) Corresponding magnitude in the frequency plane.

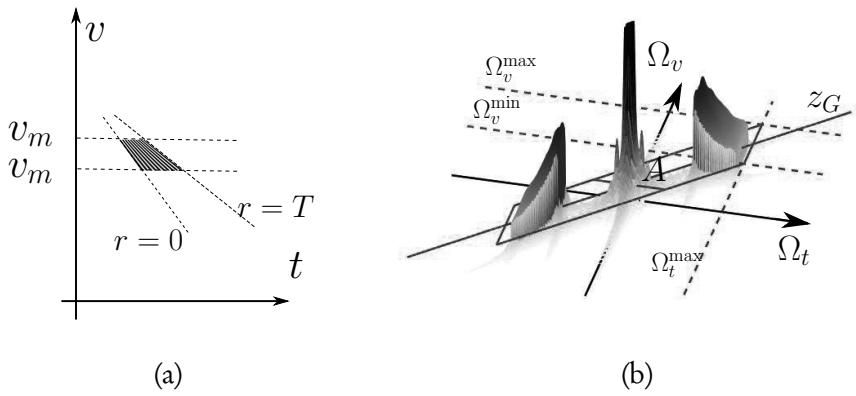
each other is shown in Fig. 2.10.

The general case for arbitrary depth has been studied in [25]. The bandwidth of the LF is, in general, non band-limited, and the method proposes only analysing the boundaries for an essential bandwidth. The main assumption is of an occlusion-free scene given on the plane (t,v) by depth function $z(x)$ and corresponding texture function $g(r)$ described in the curvilinear coordinates $r = r(x)$. The essential bandwidths (essBW) are bounded by

$$\text{essBW}_{\Omega_t}\{l\} = \frac{\sqrt{1+\max|z'|^2}}{1-v_m \max|z'|}\text{BW}\{g\}, \quad \text{essBW}_{\Omega_v}\{l\} \leq \frac{z_{\max}\sqrt{1+\max|z'|^2}}{1-v_m \max|z'|}\text{BW}\{g\},$$

where $\text{BW}\{g\}$ is the bandwidth of the texture function g. As remarked in [37], the worst case for the bandwidth expansion along the Ω_t axis is determined by the steepest slope on the surface, whereas the worst case along the Ω_v axis is when the surface is at the steepest point and furthest away from the camera, and the pixel is at the boundary of the field-of-view.

The studies in [36] and [37] have proposed generalizations in terms of finite field of view (FFoV) and finite scene width (FSW). A scene on a plane slanted by an angle ϕ with finite length given by $(x(r), z(x)), r \in [0, T], x \in [x_1, x_2]$ is considered, c.f. Fig. 2.11 (a). The texture function h is defined as $h(r) = g(r)$ for $r \in [0, T]$, and $h(r) = 0$, otherwise, where g is a band-limited function. The FFoV condition implies that the LF is 0, for $|v| > v_m$. The final, rather complex form of the LF Fourier transform is given in [37]. A generalization of Eq. 2.3 is obtained for



(a) (b)

Figure 2.11 (a) Light field parameterization for a slanted plane surface. (b) An epipolar-plane image for a slanted plane scene with $\cos(\Omega_0 r)$ texture function and FFoV, FSW constraints. (b) Corresponding to an epipolar-plane frequency plane representation with highlighted essential bandwidth obtained in [37].

the essential part of the bandwidth along the Ω_t axis in the first quadrant as follows

$$\Omega_t^{\max} = \frac{\mathrm{BW}\{g\} + 2\pi/T}{\cos\phi - \bar{v}_m |\sin\phi|}.$$

The same along the Ω_v axis is

$$\Omega_v^{\max} = \Omega_t^{\max}\frac{z_{\max}}{f} + n(\phi, \bar{v}_m)\frac{\pi}{v_m} \quad \Omega_v^{\min} = \Omega_t^{\max}\frac{z_{\min}}{f} - n(\phi, \bar{v}_m)\frac{\pi}{v_m}$$

for $\bar{v}_m = v_m/f$ and

$$n(\phi, \bar{v}_m) = \frac{3\cos^2\phi + 3.5(\bar{v}_m \sin\phi)^2}{3\cos^2\phi + (\bar{v}_m \sin\phi)^2}.$$

Under the no-occlusion constraint, $n(\phi, \bar{v}_m) \in [1, 1.625]$. Moreover, the essential bandwidth for a slanted plane is bounded by a parallelogram as illustrated in Fig. 2.11 (c), where

$$z_G = \frac{z_{\max} + z_{\min}}{2}$$

and the bandwidth is

$$A = \frac{z_{\max} - z_{\min}}{z_G}\Omega_t^{\max} + n(\phi, \bar{v}_m)\frac{2\pi f}{v_m z_G}.$$

Based on these results, the optimal sampling distance on the camera plane for a slanted plane case is

$$\Delta t_{\max} = \frac{2\pi}{A} = \frac{2\pi z_G v_m}{v_m \Omega_t (z_{\max} - z_{\min}) + 2\pi n(\phi, \bar{v}_m) f}.$$

The EPI under FFoV and FSW constraints and for a cosine texture function $g(r) = \cos(\Omega_0 r)$ is shown in Fig. 2.11 (a). The corresponding magnitude in the frequency domain is shown in Fig. 2.11 (b).

2.4 General methods for light field reconstruction

This section presents an overview of recent approaches to light field reconstruction. The majority of these approaches are based on an accurate depth estimation, aimed at subsequent scene reconstruction [89], [45], [48], [50]. These are are reviewed in Subsection 2.4.1. Subsection 2.4.2 is devoted to reviewing methods for depth or disparity estimation with consecutive post-processing using modern machine learning approaches [49], [95], [94], [40], [41], [75], [7]. An alternative reconstruction method working in a continuous Fourier domain [74] is reviewed in Subsection 2.4.3. Approaches to the related spatial super-resolution problem are presented in Subsection 2.4.4 [29], [6].

2.4.1 Depth based methods

Wanner et al. have proposed a method for estimating disparity directly from the light field data [89]. First, a fast estimate is made of the local disparity using a structure tensor working on epipolar plane images. Then, globally consistent depth maps are obtained from the local estimates using convex regularization. The accurately estimated sub-pixel precision disparity maps are then used for spatial and angular super-resolution of the light field function, formulated as variation inverse problems.

This method has been developed for processing data from plenoptic cameras, whose disparity range is relatively small.

Conventional disparity estimation methods, employ a three-step framework: cost volume construction (evaluating different depth hypotheses), cost volume filtering (regularization using aggregation in the spatial domain), and label selection (selection

of most probable depth hypothesis, typically by winner-takes-all) [45]. A similar framework has been adopted in [48] for accurate estimation of the disparity from light fields acquired with plenoptic cameras, i.e. images with relatively small disparity ranges. Therefore, the approach focuses on developing an accurate sub-pixel displacement algorithm, required for the cost volume construction. The latter consists of two components to accommodate both the color differences between views and the differences in the gradient images. The sub-pixel precision disparity map associated with the central view can be further used for LF angular and spatial interpolation.

Alternative methods for disparity estimation from plenoptic data have been developed in [80], [100].

As opposed to using plenoptic data with a small disparity range, the method described in [50] processes data captured by a gantry providing horizontal parallax images. More specifically, the constructed system accurately moves a DSLR camera horizontally to capture high-resolution multi-perspective images. For such a high amount of data, a direct estimation of disparity maps using conventional methods is inefficient. Therefore, the proposed technique utilizes a fine-to-coarse refinement technique to obtain accurate disparity maps from sufficiently-dense sampled light fields and avoids explicit global regularization. A novel sparse representation for a set of adjacent EPIs is proposed , comprising a set of distinct lines, achieved by considering a densely sampled LF. This representation is obtained first at the edges of the high-resolution image and then proceeds to successively coarser EPI resolutions with the aim of obtaining disparity estimation on the smooth spatial areas where the edges are not well-defined. The technique effectively utilizes the EPI constraints and is especially efficient for processing high spatio-angular LF datasets.

2.4.2 Machine learning methods

In recent years, there have been several attempts to solve the light field reconstruction problem by machine learning means. Kalantari at. al. [49] have proposed a novel learning-based approach aimed at intermediate view synthesis. The training dataset is formed of plenoptic images captured with a Lytro Illum camera. Using only 4 of the sub-aperture views, the aim is to reconstruct all the intermediate sub-aperture views. This approach includes two convolutional neural networks (CNN): one for disparity estimation and another for final view synthesis using the estimated disparity.

Both networks are trained simultaneously by minimizing the error between the synthesized and ground truth views. The disparity map at a location q on the camera plane (u,v) is formalized as $D_q = g_d(K)$, where g_d is a transform modeled by a CNN, working on a cost volume K computed by warping all the available views to the novel view position q using predefined disparity levels $d_1,\ldots,d_L$. For view synthesis, another CNN denoted by g_c is used: $L_q = g_c(H)$, where H represents a set of all the back-warped input views using the disparity map D_q. The role of the color prediction CNN g_c is to model the complex relationship between the final image and the warped images around any occlusions. The two neural networks are trained together using a dataset containing various patches of 60×60 spatial resolution. This method has demonstrated superior results in comparison to [89], [48], especially around occlusion boundaries.

The LF angular super-resolution problem has been cast as EPI high-frequency details reconstruction in [95]. The insufficient sampling in the angular domain is modeled in the EPI domain as

$$E_L = E_H\downarrow,$$

where $\downarrow$ denotes a down-sampling operation applied to the original E_H densely sampled EPI, where the disparity does not exceed 1 pixel. The inverse problem is solved by minimizing the following function

$$\min_f ||E_H - D_k f((E_L * k)\uparrow)||_2^2,$$

where $\uparrow$ denotes bicubic interpolation, k denotes a predefined 1D kernel in the spatial domain, D_k denotes a non-blind deblurring operator, and f represents a high-frequency reconstruction operator in the angular domain, modeled as a CNN.

The convolution kernel k (usually a Gaussian function) extracts low-frequency features from E_L. The CNN f is designed as a residual neural network with three convolution layers with decreasing kernel sizes together with a rectified linear unit. It is used to predict angular domain detail information from blurred and upsampled EPI. Further, the spatial detail of the EPI is recovered through a non-blind deblurring operation [52]. The whole densely sampled light field is reconstructed by applying the proposed "blurring - restoration - deblurring" framework for each EPI in both the horizontal and vertical directions.

Applying this method directly on an LF with a wide disparity range yields poor

reconstruction since the operator D_k does not perform well for large blur kernels k. The results reported in [95] are for a maximum disparity of 5 pixels. In a subsequent work [94], this narrow disparity limitation has been overcome using a disparity-assisted rendering technique, wherein it is sufficient that the disparities are roughly discretized [45]. For each discrete disparity region, appropriate shearing is applied to the corresponding EPI region so that the disparity range becomes small enough to be processed by the original "blurring - restoration - deblurring" method. The final result is obtained by blending together multiple super-resolved EPI regions.

Heber and Pock have proposed a two-step method for disparity estimation by analyzing all the EPIs [40]. For a given 4D LF, hyperplane orientations are predicted for the central image using a CNN applied to both horizontal and vertical EPIs. The next step is to refine the predicted orientation (disparities) using a generalized total variation regularization procedure [11]. In a follow-up publication, the authors have improved the previous results by designing a neural network that is directly applied to 3D subsets of the 4D LF using one angular and two spatial dimensions [41]. This approach allowed the artifacts in the spatial domain caused by the independent processing of each EPI slice to be significantly suppressed.

Shin at. al. have proposed an end-to-end neural network architecture for disparity estimation from a 4D light field [75]. The input consists of a stack of horizontal, vertical and two diagonal views containing the central view. This network exhibits a multi-stream structure, such that each 1D image stack subset is processed through three convolution layers to obtain sets of features describing it. The feature sets are concatenated and processed together by additional convolution layers followed by RELU. This work has also emphasized the importance of using multiple augmentation techniques in order to avoid overfitting. The method has demonstrated top performance for the HCI 4D Light Field Benchmark [44].

Extracting intrinsic information from LF data has been attempted in [7]. An encoder-decoder network has been designed to decompose a non-Lambertian scene LF into disparity, diffuse and specular components. The encoder part works on the EPI volume by processing each EPI independently. Using a number of residual blocks, it gradually reduces the EPI volume in the spatial domain and increases its feature domain. The encoded features of the paired EPI volumes are then concatenated and further processed by multiple decode pathways. The auto-encoder path reconstructs the original input data, while three other decoders get the corresponding

disparity, diffuse and specular components. All the decoders are constructed with residual blocks with transpose convolution layers. Before the last layer, the diffuse and specular decoders concatenate the features to assist in computing the final radiance as the sum of the diffusion and specular information following the dichromatic reflection model.

2.4.3 LF reconstruction through sparsification in a continuous Fourier domain

All the LF representations and sampling conditions discussed so far have specified stringent requirements for the sampling intervals in order to retain the properties of the underlying continuous function in its sampled version so that it can subsequently be reconstructed correctly. Applying these requirements to LF sensing would generate a high amount of data, which would be difficult to handle. At the same time, this data is highly redundant. This raises the idea of using sparsification approaches, that is, to find a suitable domain where the LF signal is sparse and only a small amount of data (in the form of transform coefficients) is needed for its reconstruction.

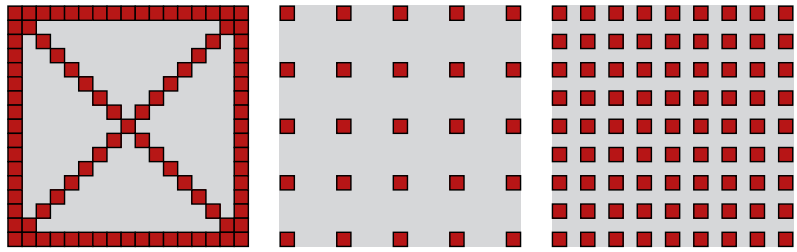
Shi et al. have proposed applying the sparsity concept in the continuous Fourier domain [74]. The authors argue that LF sparsity in a spectral domain mainly holds true in the continuous Fourier case, but does not necessarily hold in the discrete Fourier case, mainly due to the required windowing.

A signal $x = x[n]$ of length N is k-sparse in the continuous Fourier domain if it can be represented as a combination of $k < N$ frequencies at arbitrary and non-integer locations:

$$x[n] = \frac{1}{N} \sum_{l=0}^{k} a_l \exp\left(2\pi i\, \omega_l \frac{n}{N}\right).$$

The problem of reconstructing a signal from given measurements can be formulated as an estimation of the frequencies $\{\omega_l\}_{l=0}^{k}$ and the corresponding coefficients $\{a_l\}_{l=0}^{k}$. Thus, to recover two-dimensional signal samples $\{x[u,v], \forall u,v = 0,\ldots,N-1\}$ from a set of measurements $x_S = \{x[u,v], \forall (u,v) \in S\}$, one can assume k-sparsity in the continuous Fourier domain and solve the following minimization problem

$$\underset{a_l,\omega_{u_l},\omega_{v_l}}{\arg\min} \sum_{(u,v)\in S} \left\| x(u,v) - \frac{1}{N} \sum_{l=0}^{k} a_l \exp\left(2\pi i\, \frac{u\omega_{u_l} + v\omega_{v_l}}{N}\right) \right\|_2^2 .$$



(a) (b) (c)

Figure 2.12 Sampling pattern where every rectangle represents one view from a LF consisting of 17×17 views. (a) box and two diagonals pattern consisting of 93 views used for method [74]. (b), (c) uniformly decimated setup consisting of 5×5 and 9×9 views respectively.

In matrix notations $a = \{a_l\}_{l=0}^{k}, \omega = \left\{(\omega_{u_l}, \omega_{v_l})\right\}_{l=0}^{k}$, the problem becomes

$$\arg\min_{a,\omega} ||x_S - A_\omega a||_2^2.$$

The problem is solved by alternating minimization. For fixed k frequency locations ω, the corresponding optimal coefficients a are estimated directly using pseudo inverse, $a = A_\omega^\dagger x_S$. The optimal solution of the frequency locations ω is obtained by minimizing

$$\omega^* = \arg\min_{\omega} \left\| x_s - A_\omega A_\omega^\dagger x_S \right\|_2^2,$$

using gradient descent. In [74], the gradient is approximated by evaluating an error function over 8 directions around every frequency position and updating it in the most descending direction.

The 4D light field reconstruction problem is formulated as finding the values $L(x,y,u,v)$ at all angular locations (u,v), given a sampling set S. Technically, 2D slices $\hat{L}_{\omega_x,\omega_y}(u,v)$ for fixed spatial frequencies are reconstructed independently.

The sampling set S is composed of a set of 1D discrete sampling lines on the camera plane, as illustrated in Fig. 2.12. A 1D discrete line on a 2D discrete grid is defined by the parameters (α_u, α_v) and (τ_u, τ_v) when $\mathrm{GCD}(\alpha_u, \alpha_v) = 1, 0 \leq \alpha_u, \alpha_v, \tau_u, \tau_v < N$ so that the sampling positions are

$$\{(\alpha_u t + \tau_u \mod N, \alpha_v t + \tau_v \mod N), t = 0, \ldots, N-1\}.$$

To obtain reliable initial estimates for the frequency locations, the authors of [74] have proposed a voting scheme based on the Fourier projection slice theorem, which effectively utilizes the available 1D discrete sampling lines.

This method have shown prominent reconstruction quality especially for light fields generated by non-Lambertian scenes.

2.4.4 Spatial super-resolution methods

So far, light field reconstruction has been addressed in terms of reconstructing missing angular views, a problem also referred to as angular super-resolution. A related problem of increasing the resolution can be formulated in the spatial domain. Spatial super-resolution is particularly important for plenoptic cameras where the capture of angular views comes at the expense of reducing their spatial resolution. Still, plenoptic imagery exhibits significant redundancy between images, which has motivated a number of studies on LF spatial super-resolution [29], [99], [88], [71], [6]. More specifically, in [29], the LF spatial super-resolution has been formulated as a low-rank matrix approximation problem employing a convolutional neural network. By rearranging the discrete 4D light field data $l(s,t,u,v)$, a 2D matrix is formed from columns of vectorized sub-aperture images $I_{s,t}(u,v)$, such that $L = [\text{vec}(I_{s,t}), ...] \in \mathbb{R}^{m \times n}$. The acquisition of the low-resolution LF L^L from the high-resolution LF L^H is modeled as

$$L^L = \downarrow_\alpha B L^H + \eta,$$

where $\downarrow_\alpha$ is a downsampling operator by a factor α, B is a blurring kernel and η is some additive noise. The inverse problem is ill-posed and requires additional regularization. In [29], a dimensionality reduction is considered for the LF data in order to significantly eliminate redundant information. Using a precalculated optical flow, a forward warping operator $\Gamma(L^L) = L^L_\Gamma$ is applied to the low-resolution LF. The reduction is achieved by approximating $L^L_\Gamma \approx A^L$ by rank-k matrix using singular value decomposition. The A^L rank-k matrix allows the data to be decomposed into linearly independent and dependent groups of vectors (columns), such that $A^L_I \in \mathbb{R}^{m \times k}$ represents the set of linearly independent columns and $A^L_D \in \mathbb{R}^{m \times n-k} \approx W_L A^L_I$ represents the linearly dependent columns with W_L coefficients.

Instead of direct estimation of the high-resolution LF L^H, its forward warped version $\Gamma(L_H)$ is estimated first. The problem of restoration or prediction of the

linearly independent columns A_I^H of $\Gamma(L^H)$ from A_I^L is solved by minimizing $\arg\min_{A_I^H} \|f(A_I^H) - A_I^L\|^2$ using a 10 layer CNN represented by f. Linear dependent columns of $\Gamma(L^H)$ are obtained using the same W coefficients found from the low-resolution LF decomposition, $A_D^H = W_L A_I^H$. The obtained A_I^H and A_D^H together form the A^H matrix which is the rank-k estimate of $\Gamma(L^H)$.

Applying back warping (inverse of the forward warping transform Γ) on A^H generates holes in the restored images at the occlusion locations. Occlusion filling is accomplished by inpainting using the structure tensor from a low-resolution EPI for guidance [29]. The obtained $\overline{L^H}$ estimate of the unknown L^H still contains artifacts due to imperfect back warping. These are tackled by an iterative back-projection refinement

$$\overline{L_{k+1}^H} = \overline{L_k^H} + \left(L^L - \uparrow_\alpha \left(\downarrow_\alpha B\overline{L_k^H}\right)\right).$$

The spatial super-resolution problem has also been addressed in a graph-based regularization framework [71]. The forward model is formalized by blurring and sub-sampling

$$I^L = SBI^H + \eta,$$

where S is a downsampling matrix and B is a matrix implementing the blurring operation. A two-term regularization is applied, given that $F_k^{k'}$ is a warping transform from image (view) k to image k' and $H_k^{k'}$ is a binary matrix marking the corresponding occlusions. The first regularization term is formed as $H_k^{k'} I_{k'}^L \approx H_k^{k'} SBF_k^{k'} I_k^H$, i.e. assuming that any low-resolution image $I_{k'}^L$ can be approximated by a high-resolution image I_k^H, excluding occlusions. The second graph-based regularization term enforces the LF structure between views. The proposed final minimization problem is

$$\underset{I^H=\{I_k^H\}}{\arg\min} \sum_k \left\| SBI_k^H - I_k^L \right\|_2^2 + \sum_k \sum_{k' \neq k} \left\| H_k^{k'} \left(SBF_k^{k'} I_k^H - I_k^L \right) \right\|_2^2 + I^H L I^H,$$

where L represents the Laplacian matrix of the graph attached to enforce the LF structure between high-resolution views.

The spatial super-resolution problem is solved using methods developed for image denoising. A similar approach is considered in [6]. The new super-resolution algorithm is based on the denoising method presented in [5]. The super-resolution

problem is cast as a minimization problem

$$\arg\min_{\omega} \frac{1}{2} \left\| L^L - D_\alpha L^H \right\|_2^2 + \lambda_\omega \left\| \omega \right\|_0, \text{subject to} \quad L^H = \Psi\omega,$$

where D_α is a combined blurring and downsampling operator. The parameter ω denotes the coefficients in the transform domain defined by the corresponding analysis and synthesis transforms Φ and Ψ, where collaborative filtering is employed. The solution giving a high-resolution LF is found through an iterative procedure:

$$\begin{aligned} L_i^H &= \Psi T_\tau \left(\Phi L_i^H \right) \\ L_{i+1}^H &= L_i^H + \beta U_\alpha \left(L^L - D_\alpha L_i^H \right), \end{aligned} \tag{2.4}$$

where T_τ is an element-wise hard thresholding operator with a threshold value $\tau = \sqrt{2\lambda_\omega}$, and U_α is an upsampling matrix employing bicubic interpolation with an additional guided filter if necessary. The compound transform $\Psi T_\tau \Phi$ represents the denoising operator, which is formed by utilizing the original BM3D denoising method [24].

Compressive sensing techniques have been successfully used for light field capture with the corresponding dictionary learning. A new camera model, employing a coded attenuation (amplitude) mask placed between the sensor and lenses has been proposed in [65]. The sensed discrete signal is formed as

$$\mathrm{y} = \sum_i \Phi_i \mathrm{l}_i,$$

where y represents the compressive measurements, and $\mathrm{l_i}$ represents the LF angular views formed on the camera plane. Diagonal matrices $\{\Phi_i\}$ are formed by differently shearing the same coded mask. By concatenating together the variables $\mathrm{l} = [\mathrm{l}_1, \mathrm{l}_2, \ldots]$ and $\Phi = [\Phi_1, \Phi_2, \ldots]$, a compact notation is given by

$$\mathrm{y} = \Phi \mathrm{l}.$$

This problem is ill-posed since the number of measurements (sensor size in pixels) is significantly smaller than the actual 4D LF resolution. Therefore, this issue has been addressed by utilizing sparse coding techniques [65]. An overcomplete dictionary D is constructed to effectively represent the natural light field l data, giving rise to

the (small number of non-zero) representation coefficients α. This reformulates the acquisition model as

$$y = \Phi D \alpha.$$

The dictionary is learned from a big set of 4D light field patches $\{p_i\}$ employing the corresponding minimization problem

$$\min_{D,\{\alpha_i\}} \sum_i ||p_i - D\alpha_i||_2 \text{, subject to } \forall i, ||\alpha_i|| \leq k,$$

where k is the sparsity level enforced on the corresponding coefficient α_i.

A typical choice for the measurements matrix Φ is a random matrix generating incoherent measurements with respect to the dictionary. However, in the case of the coded mask, it has a strictly diagonal structure. Therefore, inspired by the results presented in [28], the optimal mask code $f = [f_1, f_2, \ldots, f_m]$ is obtained by solving the following optimization problem for a given dictionary D,

$$\min_f ||\mathrm{I} - (\Phi D)^\intercal \Phi D||_F \text{, subject to } \forall i = 1, \ldots m, 0 \leq f_i \leq 1, \sum_i f_i/m > \tau.$$

Additional constraints on f provide control over the optical light efficiency τ (formulated in terms of mean light transmission) of the camera system.

The constructed dictionary and the corresponding optimal coded mask have been evaluated for LFs with different challenging 3D scenes, containing non-Lambertian reflectancies and partial occlusions. It has been demonstrated that the achieved light field reconstruction quality is higher than it is for conventional plenoptic cameras at the cost of substantial computational resources.

3 LIGHT FIELD RECONSTRUCTION

3.1 Problem formulation

As discussed in Chapter 2, a continuous 4D light field carries the information fully describing 3D visual scenes. Recreation of the continuous light field can provide people with a highly realistic visual experience.

Applications, such as super multi-view displays and digital holography, require a dense set of LF samples. The corresponding replica of the true continuous light field is formed in the subsequent optical system. Such applications can be considered as optical reconstructors of the continuous LF from a dense, yet discrete set of light rays (samples). Other applications might require getting rays at a desired (i.e. arbitrary) location and direction from a given dense set of rays.

Hereafter, the LF reconstruction is addressed in the context of two-plane parametrization. The considered LF samples are acquired by a rectified array of cameras uniformly placed over a fixed plane in space. Assuming a Lambertian scene and sufficient spatial resolution of the captured views, the frequency analysis of an LF function presented in Chapter 2 and the results presented in [62] suggest that an arbitrary ray can be computed using a local interpolation method, given that the disparities between neighboring samples in the angular domain are less than $1px$. Correspondingly, the densely sampled LF (DSLF) is defined as a two-plane parameterized function, where the disparity between adjacent views is $1px$ at most (c.f. Fig. 3.1). This DSLF can serve as the sought-after intermediary between given LF samples and continuous LF restoration. The required sampling density of a wide baseline DSLF is not achievable by direct capture for common 3D scenes. Hence, its reconstruction has to be performed from a coarser sampling. Thus, the light field reconstruction problem, or equivalently, the view interpolation problem is reformulated as the problem of obtaining DSLF from a coarse set of views or a coarse set of samples. Hereafter, it will be referred to as a DSLF reconstruction problem. In this formulation, DSLF

reconstruction mainly implies LF reconstruction in the angular domain. However, as shown in Section 5.1, the same idea can be generalized for spatial super-resolution or reconstruction.

In the simpler case of horizontal parallax, DSLF is three-dimensional data, which is still computationally expensive to handle. This problem can be split into stages, where one first looks at the LF 2D slices, referred to as epipolar images (EPI) (Section 2.1). Each EPI can be considered independently since the correlated information in the angular domain is essentially contained within the epipolar planes. In the EPI slices, an acquired uniform set of rectified views is interpreted as a set of rows located at the proper positions of the densely sampled grid. This is referred to as densely-sampled EPI (DSEPI) (c.f. Fig. 3.1). The problem of DSEPI reconstruction can be considered independently for each particular EPI, so that only coarse sets of rows are available. In this approach, the DSLF problem is decomposed into solving multiple independent 2D problems for every DSEPI.

For any desired DSEPI, only several distinct rows are available, while the rest has to be reconstructed. One can conclude that DSEPI reconstruction is similar to the problem of inpainting. Inpainting is the process of filling in (rather isolated) holes or gaps in images, using information from the neighboring pixels and structure. It has been extensively studied in conventional 2D image processing [39]. In contrast to conventional inpainting, however, the empty regions in DSEPI are bigger than the available informative image samples. This motivates us to seek an inpainting solution in the context of a sparsifying dictionary [73], [12], relying on the hypothesis that the DSEPI can be conveniently represented with some proper dictionary.

Based on the structural properties of the DSLF, a specific predesigned dictionary can be considered. Specifically, continuous EPIs have a very distinct structure of directed lines and stripes. Our hypothesis is that a dictionary composed of directed and multi-scale atoms should be suitable for the above-defined inpainting task. As discussed further in this chapter, we adopt the shearlet frames as suitable candidates for this task. A detailed formulation of the proposed suitable dictionary is presented in Section 3.3. The proposed reconstruction method represents a modification of the iterative thresholding algorithm described in Section 3.2. The proposed DSLF reconstruction uses an iterative procedure, wherein the sparsity of coefficients from the predesigned dictionary is employed.

Further extensions of the reconstruction method using various acceleration tech-

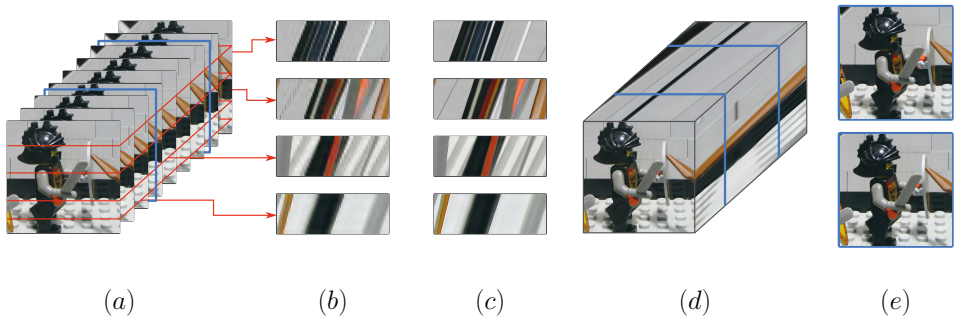


Figure 3.1 An illustration of the problem of a densely sampled LF reconstruction in terms of DSEPI reconstruction. (a) Coarse set of input views. (b) Corresponding EPIs for different rows scaled for explicit illustration of the discontinuous sampling in the angular domain. (c) Reconstructed DSEPIs illustrating the continuous structure of epipolar lines. (d) DSLF formed from the reconstructed DSEPIs. (e) Two intermediate views.

niques, including efficient inter-EPI processing and colorization are presented in Chapter 4. Certain applications where the proposed reconstruction method can be efficiently used are presented in Chapter 5.

3.2 Sparse representation and regularization

The problem of DSLF reconstruction from coarse samples can be viewed from the perspective of the light field frequency-domain analysis presented in Chapter 2. It was demonstrated that the required sampling rate for getting a faithful reconstruction was too high even for the case of Lambertian scenes with no occlusions. To overcome this limitation, a reconstruction method should use other approaches, for example based on sparsification. Sparse signal representation is usually defined with respect to a dictionary or a coordinate system, where the signal of interest is sparse. Finding the sparse representation usually requires going through an analysis/synthesis procedure, requiring an additional regularization term, which has to be designed with the specific application in mind. This way, reliable (consistent, stable) reconstruction from coarser sampling can be achieved. In this section, we briefly overview the basics of sparse representation and regularization techniques. Building on this theory, we further utilise the main reconstruction method using the sparse representation framework presented in [12], [73].

3.2.1 Sparse regularization

Various image processing problems, such as denoising, deblurring, inpainting, super-resolution, acquisition process can be formalized by a system of linear equations, referred to as a forward model

$$y = Mx + e, \tag{3.1}$$

where x is the unknown signal of interest, M is the image-formation transform represented by a measurement matrix and y is the vector of acquired measurements. For example, the measurement matrix M for the deblurring problem represents a convolution matrix and for the inpainting problem, it represents a binary mask. In many cases, perturbation of the acquired observations y is also considered. It is modeled by adding a noise term e, such that e is a vector of independent and identically distributed Gaussian random variables with zero mean and standard deviation of σ representing noise model ($e \sim \mathcal{N}(0, \sigma^2 I)$).

Obtaining the unknown signal x for the given set of measurements y would be referred to as the inverse problem of the forward model (Eq. 3.1). The inverse problem is said to be well-posed if for each given y the problem has a unique solution and the solution depends continuously on the data y [81]. Otherwise, the problem is considered to be ill-posed. Typical image processing problems are ill-posed problems.

The inverse problem can be addressed in terms of ordinary *least squares* (LS)

$$\arg\min_x ||y - Mx||_2^2.$$

For a full rank square matrix M, the solution is given simply as $M^{-1}x$ achieving zero in minimization.

In most image processing problems, the matrix M is underdetermined, i.e. the number of measurements is smaller than the signal size. Typically this implies an infinite set of x minimizing the l_2 norm. Hence, the problem is ill-posed and additional assumptions are required. Usually, assumptions are given in the form of regularization for x to determine the desirable unique solution.

For example, in the case of a full rank matrix M, the solution for the system of linear equations can be obtained using the Moore-Penrose pseudoinverse $M^\dagger y = M^\intercal (MM^\intercal)^{-1} y$. However, the regularization term, in this case, implicitly implies a

minimum l_2 norm for the desirable solution.

In some problems, it might not be easy or possible to formalize the regularization term as a simple quantity such as a minimum l_2 norm. Therefore, the solution for an ill-posed problem is formulated in terms of minimizing the cost function, composed of a fidelity term f and a penalty (regularization) term R weighted by some scalar $\lambda > 0$:

$$\arg\min_x f(x) + \lambda R(x). \tag{3.2}$$

The fidelity term $f(x)$ ensures the consistency between the solution and the measurements and the penalty term or regularizer $R(x)$ enforces a prior model of the signal.

If the penalty term is defined as $R(x) = 1/2||\Gamma x||_2^2$ for some suitable matrix Γ and the fidelity term as $1/2||y - Mx||_2^2$, a generalized solution for the *least squares* problem is obtained as

$$\arg\min_x \frac{1}{2}||y - Mx||_2^2 + \frac{\lambda}{2}||\Gamma x||_2^2 = A^\intercal(AA^\intercal + \lambda\Gamma^\intercal\Gamma)^{-1}y.$$

This is known in the literature as Tikhonov regularization or *ridge regression* in the case of $\Gamma = I$ [81].

For a set of problems, the desirable solution is sought in the form of sparse representation in a predefined dictionary. For a fixed dictionary given by a matrix D, the sparse representation of a signal x refers to the set of coefficients ω, which is sparsest. That is

$$\arg\min ||\omega||_0, \text{ subject to } x = D\omega, \tag{3.3}$$

where $||\omega||_0 = \#(\omega_k \neq 0)$ is l_0 pseudo-norm, giving the number of non-zero coefficients. The matrix D represents the analysis operator of the corresponding dictionary $D = \{\phi_n\}_{n\in\Gamma}$, such that $(Dx)[n] = \langle\phi_n, x\rangle$. Finding sparse representation is an NP-Hard problem. Nevertheless, for sufficiently sparse α, a solution can be found with a convex relaxation by replacing the l_0 pseudo-norm with the l_1 norm:

$$\min_\omega ||\omega||_1, \text{ subject to } x = D\omega. \tag{3.4}$$

This approach is referred to as Basis Pursuits (BP) [15].

It has been demonstrated that the problems 3.3 and 3.4 are equivalent in the case

of *sufficient sparsity* defined as

$$||x||_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(D)}\right),$$

where $\mu(D) = \max_{k \neq l} \frac{\langle \phi_k \phi_l \rangle}{||\phi_k||_2 ||\phi_l||_2}$ is the so-called *mutual coherence* of the dictionary. A condition for determining the uniqueness of a sparse solution has been formulated using the so-called *restricted isometry property* (RIP), [16].

The design or choice of the dictionary used in the sparse representation (3.3) plays a crucial role in addressing an inverse problem. The choice is usually motivated by the set of considered signals. A number of dictionaries have been studied for the sparse representation of natural images captured by conventional digital cameras. Dictionaries using discrete cosine transform (DCT) or wavelet transform have been designed and successfully applied to various problems [64].

An alternative to the fixed dictionary approach is to learn a dictionary from a set of training data representing the considered class of signals. One of the common techniques for dictionary learning is referred to as sparse coding [56]. For a given set of signals $Y = \{y_i\}$, sparse coding aims at finding (learning) a dictionary D, such that each signal $y \in Y$ can be sparsely represented or approximated by a linear combination of dictionary elements, typically formulated as the following optimization problem

$$\min_{D,\{\omega_i\}} \sum_i \left(\frac{1}{2}||y_i - D\omega_i||_2^2 + \lambda ||\omega_i||_0\right), \quad \text{subject to } ||\phi_k||_2 = 1, \forall k. \tag{3.5}$$

The normalization condition for all the dictionary elements $||\phi_k||_2 = 1$ is required to avoid arbitrarily small coefficients α_i compensated by the elements with high magnitude.

Typically, the dictionary learning problem of Eq. 3.5 is solved using an alternating iteration scheme consisting of sparse approximation and dictionary refinement steps. The sparse approximation step aims to obtain the α_i coefficients implying a fixed dictionary D, using greedy strategies or replacing the l_0 norm with the l_1 norm. The dictionary refinement step updates the dictionary elements with fixed coefficients α_i obtained at the previous step. The explicit form of the final dictionary also depends on setting additional desirable constraints, such as an upper bound on mutual coherence. An alternative method *K-SVD* for joint updating dictionary elements with

corresponding coefficients is presented in [2].

For some considered inverse problems, the penalty term in Eq. (3.2) might have a nontrivial representation and a direct solution minimizing Eq. (3.2) might not be feasible since it might not be possible to differentiate the penalty term. Nevertheless, several methods are proposed in the literature to resolve this problem.

For the main inverse problem in Eq. (3.1), the corresponding regularized minimization problem (3.2) is considered. The fidelity term is represented as $f(x) = \frac{1}{2\sigma^2}||y - Mx||_2^2$. For a general regularizer $R(x)$, an alternating direction method of multipliers (ADMM) has been proposed [90]. The solution is obtained through an iterative minimization procedure

$$\begin{aligned} a) \quad & x_{k+1} = \arg\min_x \frac{1}{2\sigma^2}||y - Mx||_2^2 + \frac{\gamma}{2}||x - (v_k - u_k)||_2^2 \\ b) \quad & v_{k+1} = \arg\min_v \frac{\gamma}{2}||(x_{k+1} + u_k) - v||_2^2 + \lambda R(v) \\ c) \quad & u_{k+1} = u_k + (x_{k+1} - v_{k+1}) \end{aligned} \tag{3.6}$$

To avoid the explicit formulation of the regularizer $R(x)$, a Plug-and-Play (P&P ADMM) method is presented in [21]. In this method, the minimization step Eq. 3.6 (b) is replaced with

$$b)\ v_{k+1} = \mathscr{D}(x_{k+1} + u_k, \sqrt{\lambda/\gamma}). \tag{3.7}$$

where $\mathscr{D}(\cdot;\sigma)$ is a denoising operator. Furthermore, an adaptive update rule for the parameter $\gamma = \gamma_k$ is presented for robust and accelerated convergence. The denoising operator can be defined as thresholding in a certain transform domain $\mathscr{D}(x,\sigma) = \Psi\mathscr{T}_{t(\sigma)}(\Phi x)$, where $\mathscr{T}_t(x)$ is a thresholding operator [64].

Another method is presented in [12], seeking a sparse representation with respect to a certain frame, defined by analysis and synthesis transforms Φ and Ψ, correspondingly. The penalty function is defined for the transform coefficients as $R(\omega) = ||\omega||_p$, $p = 0, 1$. For the inpainting problem, the *framelet* method [12] aims to find a solution applying the following frequency-spatial alternate minimization

$$\begin{aligned} a) \quad & \omega_k = \arg\min_\omega \frac{1}{2}||\Phi x_n - \omega||_2^2 + \lambda||\omega||_1 \\ b) \quad & x_{k+1} = \arg\min_x \frac{1}{2}||\Psi\omega_k - x||_2^2 + \chi_C(x), \end{aligned} \tag{3.8}$$

where $\chi_C(x)$ is an indicator function for the subset of admissible solutions $C = \{x; Mx = y\}$. It has been shown that the minimization problems 3.8 can be solved with a simple iterative algorithm

$$\omega_k = \mathscr{T}_\lambda \Phi(My + (I - M)\Psi\omega_k).$$

where $\mathscr{T}_\lambda(\omega)$ is a soft thresholding operator representing the proximal mapping, which corresponds to the penalty term in terms of l_1 norm as in Eq. 3.8 (a). The solution in the signal domain is given by $x_k = \Psi\omega_k$. The same iterative algorithm can be performed entirely in the signal domain as

$$x_{k+1} = \Psi\mathscr{T}_\lambda \Phi(x + M(y - x)) \tag{3.9}$$

Given $\omega = \Phi x$, it has been shown that the convergence of the algorithm 3.9 is equivalent to the minimization problem

$$\arg\min_{\omega} \frac{1}{2} ||M\Psi\omega - y||_2^2 + \frac{\kappa}{2} ||(I - \Phi\Psi)\omega||_2^2 + \lambda ||\omega||_1 . \tag{3.10}$$

with $\kappa = 1$. This minimization problem is referred as a *balanced* approach [12], [73].

For comparison, the commonly used *synthesis* approach, which corresponds to $\kappa = 0$ in Eq. 3.10:

$$\text{(Synthesis)} \quad \arg\min_{\omega \in \text{Range}(\Phi)} \left\{ \frac{1}{2} ||y - M\Psi\omega||_2^2 + \lambda ||\omega||_1 \right\}.$$

A corresponding solution can be obtained using a proximal gradient iterative algorithm given as

$$\omega_{k+1} = \mathscr{T}_\lambda(\omega_k + \Phi M(y - \Psi\omega_k)), \tag{3.11}$$

performed entirely in the coefficients domain.

Another formulation known in the literature as the analysis model corresponds to $\kappa \to \infty$ in Eq. 3.10 :

$$\text{(Analysis)} \quad \arg\min_{x} \left\{ \frac{1}{2} ||y - Mx||_2^2 + \lambda ||\Phi x||_1 \right\}.$$

All three approaches are identical for the case of an orthonormal basis Φ, $\Phi\Phi^\intercal = I$. However, if the transform is defined by a frame, the corresponding *synthesis*, *analysis*,

and *balanced* approaches cannot be derived from one another.

In order to derive a solution for the *balanced* approach 3.10, a general minimization problem is considered

$$\arg\min_x F_1(x) + F_2(x),$$

where F_1 is a semi-continuous convex function and F_2 is a differentiable convex function satisfying the condition $\beta \|\nabla F_2(x) - \nabla F_2(y)\|_2 \leq \|x - y\|_2$. In [23], the solution is presented by a proximal forward-backward splitting algorithm that guarantees weak convergence of a series $\{x_k\}$ to the minimum, given by

$$x_{k+1} = \text{prox}_{\gamma F_1}(x_k - \gamma \nabla F_2(x_k)),$$

when $0 < \gamma < 2\beta$.

As shown in [13], the proximal forward-backward splitting algorithm can be directly applied to the problem (3.10), if $F_1(\omega) = \lambda \|\omega\|_1$ and $F_2(\omega) = \frac{1}{2}\|M\Psi\omega - y\|_2^2 + \frac{\kappa}{2}\|(I - \Phi\Psi)\omega\|_2^2$. The proximal mapping of the F_1 is given by the soft thresholding operator

$$\text{prox}_{\gamma F_1}(\omega) = \mathcal{T}_{\gamma\lambda}(\omega).$$

The gradient of F_2 can be calculated as

$$\nabla F_2(\omega) = \kappa(I - \Phi\Psi)\omega + \Phi(M\Psi\omega - y).$$

Taking into account that for the case of inpainting the measurement matrix M is diagonal and $\Psi\Phi = I$, for some β:

$$\beta \|(\kappa(I - \Phi\Psi) + \Phi M\Psi)(\omega_1 - \omega_2)\|_2 \leq \|\omega_1 - \omega_2\|_2 .$$

Therefore, the iterative algorithm for solving the problem (3.10) can be summarized as follows

$$\begin{aligned} a) \quad & x_k = \Psi\omega_k \\ b) \quad & \eta_k = \Phi\left(x_k + \frac{1}{\kappa}(y - Mx_k)\right) \\ c) \quad & \omega_{k+1} = \mathcal{T}_{\gamma\lambda}(\omega_k + \gamma\kappa(\eta_k - \omega_k)) \end{aligned} \tag{3.12}$$

3.3 Shearlet frame

Among other properties, natural images are characterized by smooth regions delineated by anisotropic edges. For their efficient representation, dictionaries, which optimally approximate anisotropic features, are required. The developments of such dictionaries and systems are described in the literature in the context of the so-called *cartoon-like* functions. *Cartoon-like* functions are defined as having second continuous derivatives (C^2) on the unit square, except for a closed C^2 discontinuity curve. The optimal approximation, in this case, is defined using the decay rate of l_2 norm error of the best N-term approximation. Conventional Fourier or 2D wavelet transforms are not suitable for this task since their separable structure does not efficiently represent singularities over a discontinuity curve. It has been shown that the best N-term approximation rate using 2D wavelet transform is $\mathcal{O}(N^{-1})$ [53]. Using an approach such as adaptive triangulation with N triangles, referred to as *Wedgelet*, an approximation rate of $\mathcal{O}(N^{-2})$ can be achieved [27].

Obtaining a similar approximation rate but using a non-adaptive system is desirable for the efficient representation of discontinuity regions. For this task, *steerable pyramid* transform has been proposed in [77]. However, this system does not provide an optimal approximation rate for anisotropic features.

An optimal approximating rate has been attempted with the *Curvelets* transform [18] by maintaining the orientation property at various scales and locations. The *Curvelets* system achieves a significantly better approximation rate than the 2D wavelet systems *cartoon-like* functions discussed above. It has been developed further in the form of *Contourlets*, which act as discrete filter-banks of the *Curvelets* system, thus providing an efficient implementation [26] .

In all these representation systems, the goal has been to achieve a non-adaptive representation optimally approximating anisotropic features.

The *Shearlets* system as an alternative approach has been introduced in [53]. The directional property, in this case, is achieved using shear transform, in contrast to the rotation transform used in the methods presented before. This property makes the shearlet frame particularly interesting for the epipolar-plane image representation since its structure is formed by a shearing operator rather than rotation. The shearlet system has an elaborate theoretical background and has demonstrated an optimal sparse approximation of $\mathcal{O}(N^{-2}(\log(N))^3)$ for *cartoon-like* functions [53]. A further

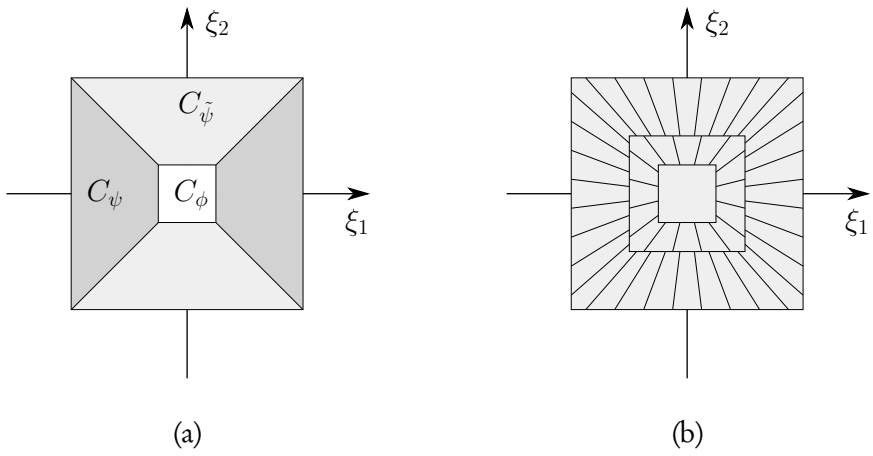


Figure 3.2 (a) The outlined regions correspond to frequency plane separation for shearlet transform design. Two cone-adapted regions correspond to a $C_\psi, C_{\tilde{\psi}}$ set of filters and the central rectangle region corresponds to a C_ϕ low pass filter. (b) Frequency plane tilling obtained by whole shearlet transform using two scales of decomposition $J = 2$.

modification in the form of compactly supported shearlet systems has been presented in [54]. While this system does not form a Parseval frame, it is still applicable for the task of efficient approximation of *cartoon-like functions*. The theoretical framework of the universal shearlet system has been studied in [33]. It generalizes various systems as members of the $alpha$-parameterized family so that for $\alpha = 2$ it represents the wavelet system, for $\alpha = 1$ it is the parabolic shearlet system and for $\alpha \longrightarrow 0$ the system forms *ridglets* [17].

The shearlet transform proposed in this thesis is an adaptation of the non-separable shearlet transform presented in [61]. It has been specifically modified for efficient representation of functions with singularities distributed over straight lines rather than parabolic curves, similarly to the *Ridglet* transform.

The cone-adapted discrete shearlet system SH is a set of 2D functions formed by shearing S, parabolic scaling A, and translation transforms applied on the generator functions: scaling function ϕ and shearlets $\psi, \widetilde{\psi} \in L^2(\mathbb{R}^2)$. For $c = (c_1, c_2) \in \mathbb{R}^2_+$, the system is defined as follows

$$\mathrm{SH}(\phi, \psi, \widetilde{\psi}; c) = \Phi(\phi; c_1) \cup \Psi(\psi; c) \cup \widetilde{\Psi}(\widetilde{\psi}; c). \tag{3.13}$$

The role of the three subsets is illustrated in Fig. 3.2 (a). The subset $\Psi(\psi; c)$

corresponds to the cone-shaped region C_ψ, the subset $\tilde{\Psi}(\tilde{\psi};c)$ corresponds to the region $C_{\tilde{\psi}}$, and the subset $\Phi(\phi;c_1)$ - to the central part C_ϕ. This division of the frequency plane is achieved using the following definitions

$$\Phi(\phi;c_1)=\{\phi_m=\phi(\cdot-c_1 m);m\in\mathbb{Z}^2\}$$
$$\Psi(\psi;c)=\{\psi_{j,k,m}=2^{3/4j}\psi(S_k A_{2^j}\cdot-M_c m);j\geq 0,|k|\leq\lceil 2^{j/2}\rceil,m\in\mathbb{Z}^2\}$$
$$\tilde{\Psi}(\tilde{\psi};c)=\{\tilde{\psi}_{j,k,m}=2^{3/4j}\tilde{\psi}(S_k^{\intercal}\tilde{A}_{2^j}\cdot-\tilde{M}_c m);j\geq 0,|k|\leq\lceil 2^{j/2}\rceil,m\in\mathbb{Z}^2\}$$

where A and $\tilde{A}$ are scaling matrices, S_k is the shearing matrix, $M_c=\mathrm{diag}(c_1,c_2)$ and $\tilde{M}_c=\mathrm{diag}(c_2,c_1)$ are translation sampling matrices,

$$A=\begin{pmatrix}2^j & 0\\ 0 & 2^{j/2}\end{pmatrix},\ \tilde{A}=\begin{pmatrix}2^{j/2} & 0\\ 0 & 2^j\end{pmatrix},\ S_k=\begin{pmatrix}1 & k\\ 0 & 1\end{pmatrix}.$$

Following the formalization in [61], the factor $j/2$ has to be integer, otherwise $\lceil j/2\rceil$ is taken. Usually, a discrete image or function f^d is given as input. To maintain the discretization of the continuous transform, it is assumed that a continuous 2D function f can be represented by the discrete signal f^d and the scaling function ϕ for a sufficiently large $J>0$, i.e.

$$f(x_1,x_2)=\sum_{(k_1,k_2)\in\mathbb{Z}^2}2^J f^d[k_1,k_2]\phi(2^J x_1-k_1,2^J x_2-k_2),$$

where the 2D scaling function is $\phi(x_1,x_2)=\phi^1(x_1)\phi^1(x_2)$ and 1D scaling ϕ^1 and wavelet ψ^1 functions satisfy the following two-scale equations

$$\phi^1(x)=\sum_{k\in\mathbb{Z}}h[k]\sqrt{2}\phi^1(2x-k)\quad\text{and}\quad\psi^1(x)=\sum_{k\in\mathbb{Z}}g[k]\sqrt{2}\phi^1(2x-k).$$

The Fourier coefficients of the trigonometric polynomial H_j and G_j

$$H_0\equiv 1,\ H_j(\xi)=\prod_{i=0}^{j-1}H(2^i\xi),\ G_j(\xi)=G(2^{j-1}\xi)H_{j-1}(\xi),\ j=0,\ldots,J \tag{3.14}$$

are denoted by g_j and h_j.

The counterpart of the 2D scaling function $\phi(x_1,x_2)$ is the 2D wavelet function

$\psi(x_1, x_2)$, which has been formed in a non-separable manner as proposed in [61]

$$\hat{\psi}(\xi_1, \xi_2) = P(\xi_1/2, \xi_2)\hat{\psi}^1(\xi_1)\hat{\phi}^1(\xi_2),$$

where $P(\xi_1, \xi_2)$ is a trigonometric polynomial representing a 2D fan filter with wedge-shaped essential support. The choice of this filter leads to significant improvements in the numerical results compared to the design of the filter as a Kronecker product of 1D filters, as shown in [61]. Consequently, it can be shown that by appropriate selection of the sampling grid M_c, the coefficients of the shearlet transform corresponding to the system elements $\{\psi_{j,0,m}\}_{m\in\mathbb{Z}^2}$ can be calculated by applying digital filter $p_j * \left(g_{J-j} \otimes h_{J-j/2}\right)$ on discrete signal f^d, where p_j are the Fourier coefficients of a scaled 2D fan filter $P(2^{J-j-1}\xi_1, 2^{J-j/2}\xi_2)$.

It is easy to see that

$$\psi_{j,k,m}(\cdot) = \psi_{j,0,m}(S_{k2^{-j/2}}\cdot).$$

This motivates the need for the discretization of the shearing operator. However, this is not straightforward since the shearing operator does not preserve the regular grid $\mathbb{Z}^2$. Nevertheless, in [61] the discretization has been achieved in the following way. First, an upsampling by the factor of $2^{j/2}$ (denoted by $\uparrow 2^{j/2}$) is applied. This is followed by filtering with the corresponding low-pass filter $h_{j/2}$, and resampling over a new integer grid by applying the S_k shearing operator. Finally, low-pass filtering $\overline{h}_{j/2}$ followed by downsampling with a factor on $2^{j/2}$ (denoted by $\downarrow 2^{j/2}$) is applied. Thus, the discretized shearing operator $S^d_{k2^{-j/2}}$ is defined as follows

$$S^d_{k2^{-j/2}}(r) = \left(\left(r_{\uparrow 2^{j/2}} *_1 h_{j/2}\right)(S_k\cdot) *_1 \overline{h_{j/2}}\right)_{\downarrow 2^{j/2}}$$

The coefficients corresponding to the set of elements $\{\psi_{j,k,m}\}_{m\in\mathbb{Z}}$ in the shearlet system can be calculated using convolution of the discrete signal with the following digital filter

$$\psi^d_{j,k} = S^d_{k2^{-j/2}}\left(p_j * \left(g_{J-j} \otimes h_{J-j/2}\right)\right).$$

This discrete shearlet frame is designed for the efficient representation of signals with parabolic singularities. By replacing the scaling matrix with $A = \mathrm{diag}(2^j, 2^{-1})$, a modified shearlet system is obtained. It maintains shears along straight lines and determines the required number of shears on each scale of the frequency plane tilling.

The corresponding digital filters become

$$\Psi^d_{j,k} = S^d_{k2^{-(j+1)}}\left(p_j * \left(g_{J-j} \otimes h_{J+1}\right)\right), j = 0,\ldots,J-1, |k| \leq 2^j + 1.$$

This set of transform filters corresponds to only one cone-shaped region C_ψ in the frequency plane highlighted in Fig. 3.2 (a). The region $C_{\tilde{\psi}}$ is covered by the filters $\tilde{\psi}^d_{j,k}$ directly obtained as $\hat{\tilde{\psi}}^d_{j,k}(\xi_1,\xi_2) = \hat{\psi}^d_{j,k}(\xi_2,\xi_1)$. The central region C_Φ corresponds to one filter $\phi^d = h_J \otimes h_J$.

The constructed discrete shearlet system is not orthogonal, therefore the dual frame elements are required for synthesis transform. Using auxiliary notation

$$\hat{\Psi}^d = |\hat{\phi}^d|^2 + \sum_{j=0}^{J-1} \sum_{|k|\leq 2^j+1} \left(|\hat{\psi}^d_{j,k}|^2 + |\hat{\tilde{\psi}}^d_{j,k}|^2\right),$$

the dual elements are defined as follows

$$\hat{\varphi}^d = \frac{\hat{\phi}^d}{\hat{\Psi}^d}, \quad \hat{\gamma}^d_{j,k} = \frac{\hat{\psi}^d}{\hat{\Psi}^d}, \quad \hat{\tilde{\gamma}}^d_{j,k} = \frac{\hat{\tilde{\psi}}^d}{\hat{\Psi}^d}.$$

Finally, the analysis operator corresponding to the construction shearlet frame is given by

$$S(f^d_J) = \left\{ s_{j,k} = f^d_J * \bar{\psi}^d_{j,k}, \tilde{s}_{j,k} = f^d_J * \bar{\tilde{\psi}}^d_{j,k}, s_0 = f^d_J * \bar{\phi}^d \right\} \tag{3.15}$$

and the inverse (synthesis) operator can be calculated using the dual elements

$$S^*\left(\{s_{j,k}, s_0\}\right) = \sum_{j=0}^{J-1} \sum_{|k|\leq 2^j+1} \left(s_{j,k} * \gamma^d_{j,k} + \tilde{s}_{j,k} * \tilde{\gamma}^d_{j,k}\right) + s_0 * \phi^d.$$

3.4 Epipolar-plane image reconstruction

The proposed inpainting techniques for DSEPI reconstruction utilize the sparse regularization as presented in Section 3.2 and the shearlet frame as presented in Section 3.3.

The input is given by subsampled EPI in angular dimension so that the disparities are in the range of $[d_{\min}, d_{\max}]$ in pixels. With no loss of generality, a pre-shearing

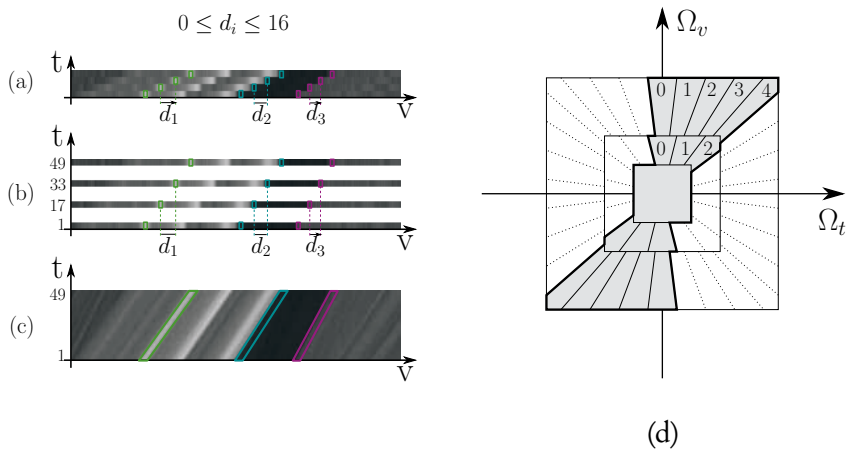


Figure 3.3 (a) Subsampled densely sampled epipolar-plane image, assuming that disparities between consecutive rows are no more than 16px. (b) Subsampled data can be interpreted as every 16-th row if a densely sampled light field is desired. (c) Corresponding densely sampled light field with disparities don't exceed 1px. (d) Highlighted shearlet transform atoms used in the EPI reconstruction algorithm. The selected atoms correspond to the EPI anisotropic structure. Each disparity layer ($k = 0, 1, ...4$) is represented with one transform atom in each scale ($j = 0, 1$).

transform can be applied to guarantee positive disparities with the disparity $[0, d_{\text{range}}]$ ($d_{\text{range}} = d_{\text{max}} - d_{\text{min}}$). Such an input subsampled EPI is illustrated in Fig. 3.3 (a) for the case $d_{\text{range}} = 16$. All rows of the given subsampled EPI are interpreted as every d_{range}-th row of the corresponding DSEPI, as illustrated in Fig. 3.3 (b). This interpretation guarantees that in the final reconstructed EPI, the densely sampled condition will be satisfied.

One can identify continuous epipolar lines formed by the given rows in an EPI like in Fig. 3.3 (b), however, broken by missing rows. These missing rows are the regions to be inpainted. This leads to the following formulation of the problem: reconstruct DSEPI, as illustrated in Fig. 3.3 (c), from the given d_{range}-th rows by filling in the missing rows (i.e. inpainting empty regions). Hence, the forward model of subsampled EPI obtained from DSEPI can be expressed in the form

$$y = Mx,$$

where x is the desirable DSEPI, y is an EPI where only every d_{range}-th rows are given and M is a binary masking matrix. The inpainting problem of obtaining x for a

given y and M is ill-posed and requires additional regularization. In our case, this regularization is defined by imposing sparsity in the shearlet transform domain.

The disparities are confined within the range of $[0,1]$ px for DSEPI and the corresponding support in the frequency plan will be within a certain region. Thus, not all elements of the shearlet transform are required. In the regularization transform only elements corresponding to disparities within the range $[0,1]$ are used, as illustrated in Fig. 3.3 (d). The analysis of the regularization transform built from the shearlet transform (Eq. 3.15) is such that for discrete signal x the analysis transform is defined as

$$S(x) = \left\{ x * \overline{\psi^d_{j,k}}, x * \overline{\phi^d}, j = 0,\dots J-1, k = 0,\dots,2^{j+1} \right\} \tag{3.16}$$

The algorithm for DSEPI reconstruction proposed in [P I], [P IV] employs iterative hard thresholding in the following form

$$x_{k+1} = S^* \left(\mathscr{T}_{\lambda_k} \left(S(x_k + \alpha_k (y - M x_k)) \right) \right). \tag{3.17}$$

As depicted in Fig. 3.4, the analysis and the synthesis transforms are implemented via Fourier-domain multiplication with pre-calculated directional filters corresponding to the transform elements.

The hard thresholding operator

$$(\mathscr{T}_\lambda x)(k) = \begin{cases} x(k), & |x(k)| \geq \lambda \\ 0, & |x(k)| < \lambda \end{cases}$$

is used with a linearly decaying thresholding value λ_n in the range $[\lambda_{\max}, \lambda_{\min}]$. The algorithm is a simplified version of Eq. 3.12 in Section 3.2, where $\alpha_k = 1/\kappa$, $\gamma = 1/\kappa$ and $\lambda_k = \gamma\lambda$. A high value of the parameter α_k provides additional acceleration to the convergence rate. This effect is partially related to the sparsity of the measurement matrix M. Typically, the number of available samples is significantly smaller than the number of reconstructed samples. Therefore, significant amplification is required to increase the influence of the available samples at every thresholding iteration. However, an unlimited increase of the parameter α_k would lead to divergence of the series x_k. Using the method proposed in [9], the parameter is made adaptable as

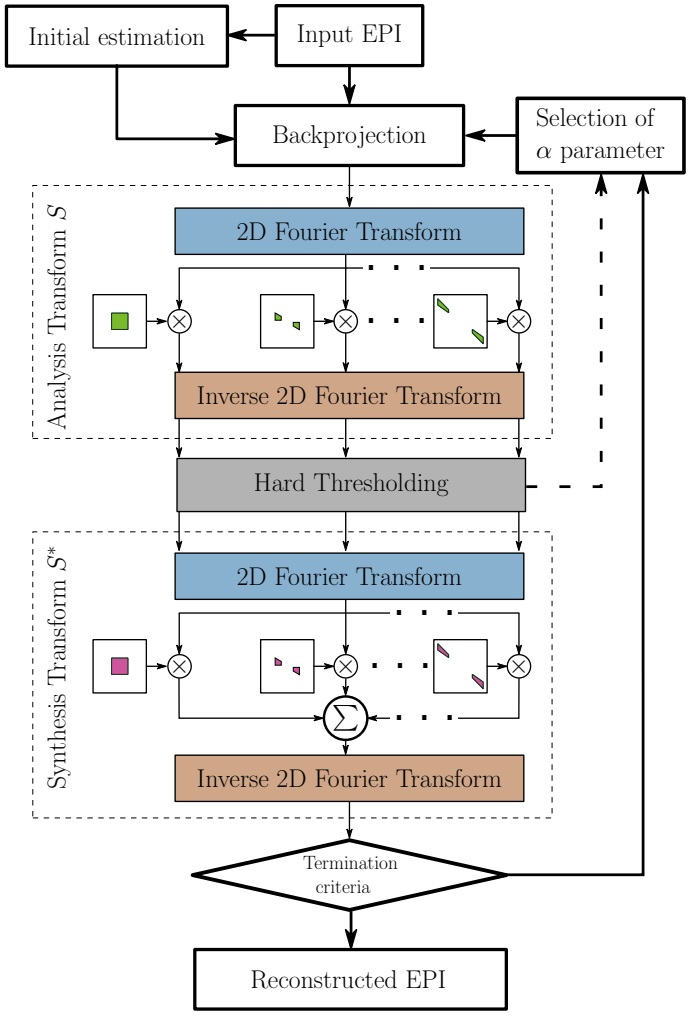


Figure 3.4 Diagram of the proposed reconstruction method.

follows

$$\alpha_k = \frac{||\beta_k||_2^2}{||MS^*(\beta_k)||_2^2}, \tag{3.18}$$

where $\beta_k = S_{\Gamma_k}(y - Mx_k)$ and S_{Γ_k} includes the shearlet transform coefficients only from the set Γ_k, the support of $S(x_k)$.

The number of shearlet decomposition scales has to be determined in relation to the disparity range: $J = \lceil \log_2(d_{\text{range}}) \rceil$. This choice determines a small enough bandwidth for the central low-pass filter, which guarantees that the retained frequencies

(a) (b) (c)

Figure 3.5 DSLF reconstruction method performance for *Lego Knights* dataset. (a) Ground-truth intermediate view. (b) Same view obtained by reconstructing every EPI. (c) Absolute difference between the two.

are alias-free. A smaller value of J would directly influence the reconstruction quality. A small value will not guarantee enough directional elements of the transform, which leads to poor reconstruction quality, thus the choice of the appropriate number of scales for the transform with the corresponding number of shears is crucial for accurate reconstruction. A comparison of the reconstruction quality for various scaling numbers is presented in [P IV].

3.5 View interpolation

While the problem of DSLF reconstruction has been equivalently formulated in the form of DSEPI reconstruction, the final cumulative result of reconstructing all DSEPIs can be quantified over the new angular views which have been generated. This allows comparing the method in Section 3.4 with methods aimed at interpolating virtual views from a given set of multi-perspective views, most of them employing information about the scene geometry in the form of a depth map [P IV]. An example of intermediate view interpolation for the dataset *Lego Knights* from [86] is presented in Fig. 3.5. The depicted intermediate view is reconstructed by applying an iterative reconstruction algorithm for every EPI, using only 5 input views.

The DSLF reconstruction method has been developed for uniform sampling along the angular dimension. However, a non-uniform sampling in the camera plane can also be considered and can be easily incorporated in the presented reconstruction

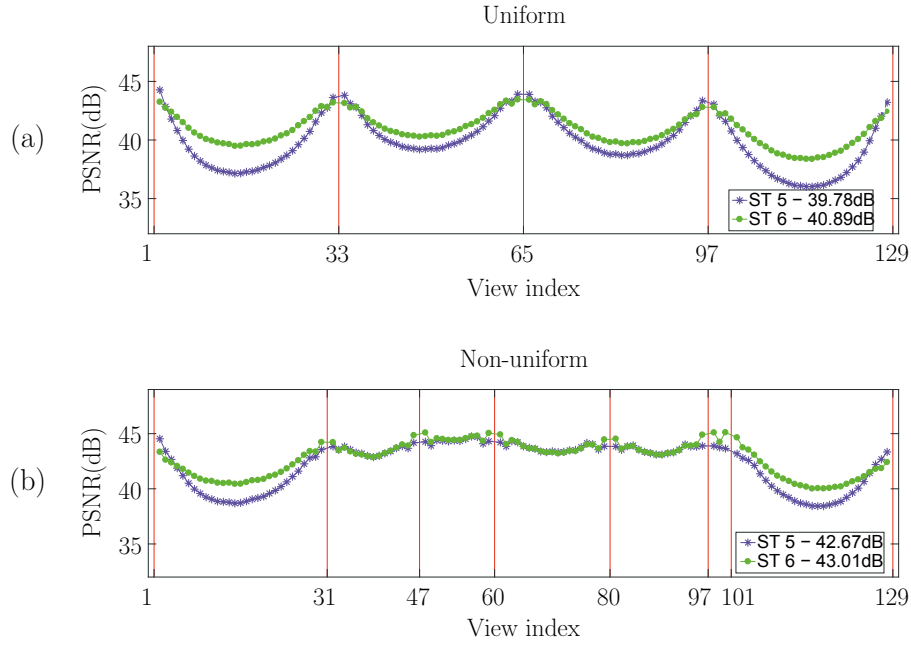


Figure 3.6 Comparison of the DSLF reconstruction performance in terms of PSNR for uniform (a) and non-uniform (b) sampling. Methods referred to as *ST 5* and *ST 6* correspond to the used number of scales $J = 5$ and $J = 6$ in the shearlet transform.

procedure. For a given set of non-uniformly spaced cameras, the assumption is again that they are samples of some densely sampled LF and are placed accordingly. The case is illustrated in Fig. 3.6. As seen in the figure, the performance of the reconstruction method is highly dependent on the distance between available samples (c.f. Fig. 3.6 (b)). It is important to emphasize that the necessary parameter d_{range} is not determined straightforwardly as in the case of uniform sampling. Instead, it is recommended to choose it as the maximum disparity range corresponding to the adjacent samples being furthest from each other. This allows for minimizing the reconstruction error in coarsely sampled regions.

The proposed reconstruction method is applied directly to EPIs and thus avoids a direct geometry estimation in the form of depth maps. This has the advantage of processing scenes where depth estimation would be ambiguous, which is the case of non-Lambertian scenes. Consider, for example, a 3D scene containing a semi-transparent surface. Direct depth estimation is quite demanding using the input captured views. In contrast, the proposed method demonstrates identical performance in terms of

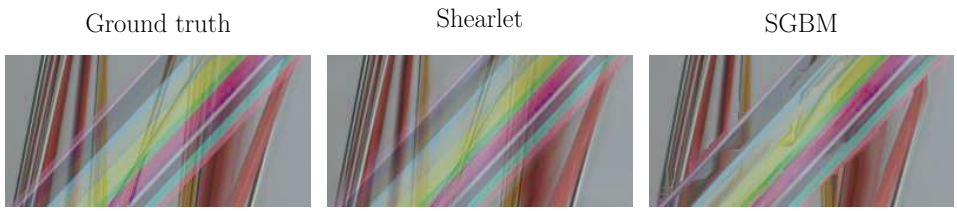


Figure 3.7 Semi-transparent DSEPI reconstruction using the proposed method and SGBM [43].

reconstruction error for Lambertian and semi-transparent regions of the scene. This aspect of the proposed method has been evaluated using a synthetic dataset, where the ground truth DSLF is available and only a coarse set of samples is used as an input for the reconstruction. The performance of the proposed method has been compared with the disparity estimation algorithm SGBM [43]. An illustration of the reconstruction results for an EPI with the semi-transparent region is given in Fig. 3.7.

4 ACCELERATION METHODS

The proposed main reconstruction method, as summarized by Eq. 4.1 is applied independently on each EPI of a given LF. Since typical LF reconstruction requires the reconstruction of multiple EPIs, this is a good start for developing parallelized computing approaches for its implementation. However, it is still a computationally demanding method. This chapter discusses approaches for its acceleration. Two approaches have been considered. First, we have aimed at improving the convergence in the iterative procedure. Second, we have exploited the correlations between adjacent EPIs in different domains (i.e. employing correlation in the spatial domain and between color channels). Finally, we have also proposed a particular way of handling full-parallax imagery. All the acceleration methods presented in this section have been published in [P III].

4.1 Overrelaxation

The speed of convergence for the iterative algorithm in Eq. (3.17) can be changed with the parameter α_k. Its adaptive calculation, as in Eq. (3.18), is computationally demanding. This one step in the algorithm costs almost as much as computing as the rest of the iteration in Eq. (3.17). A preferable solution would be to avoid the adaptive selection of the parameter α. However, fixing α has to be done optimally: its value has to be high enough to provide fast convergence and at the same time small enough to avoid divergence. The proposed solution aims at introducing additional overrelaxation steps to provide stable convergence even for high values of α [P III].

The proposed solution is similar to the double overrelaxation steps presented in [70]. Consider the current iteration step where x^{new} is the result obtained by Eq. (3.17). It can be updated using the result from the previous iteration x^{old} by

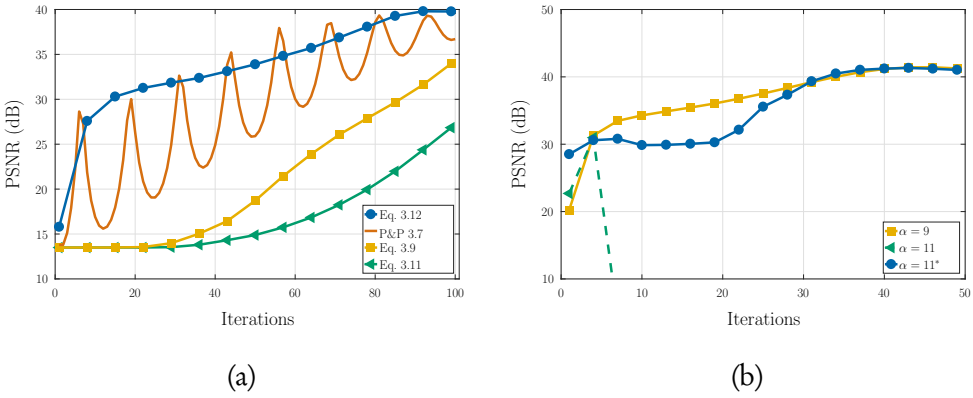


Figure 4.1 (a) Performance comparison of different EPI reconstruction algorithms. (b) Convergence performance for different values of the parameter α for the algorithm 4.1.

minimizing a quadratic error norm against the ground-truth y

$$x^{\text{new}} = \min_{\beta} \left\| y - M(x^{\text{new}} + \beta(x^{\text{new}} - x^{\text{old}})) \right\|_2^2.$$

The solution for the optimization parameter β in the case of diagonal matrix M can be obtained as follows

$$\beta^* = \frac{(y - x^{\text{new}})^\intercal M(x^{\text{new}} - x^{\text{old}})}{(x^{\text{new}} - x^{\text{old}})^\intercal M(x^{\text{new}} - x^{\text{old}})}.$$

The proposed full accelerated algorithm is

$$\begin{aligned} p_k &= \Psi \mathcal{T}_{\lambda_k}(\Phi(x_k + \alpha(y - Mx_k))) \\ q_k &= p_k + \beta(p_k - x_{k-1}),\ \beta = \frac{(y - p_k)^\intercal M(p_k - x_{k-1})}{(p_k - x_{k-1})^\intercal M(p_k - x_{k-1})} \\ x_{k+1} &= q_k + \gamma(q_k - x_{k-2}),\ \gamma = \frac{(y - q_k)^\intercal M(q_k - x_{k-2})}{(q_k - x_{k-2})^\intercal M(q_k - x_{k-2})}. \end{aligned} \tag{4.1}$$

The additional steps provide stable convergence which allows the choice of a fixed and relatively high parameter α, which is also common for all processed EPIs. This might be rendered suboptimal for some EPIs, but it establishes a common procedure for all EPIs, which significantly simplifies the proposed algorithm.

A comparison between reconstruction methods with differently-controlled con-

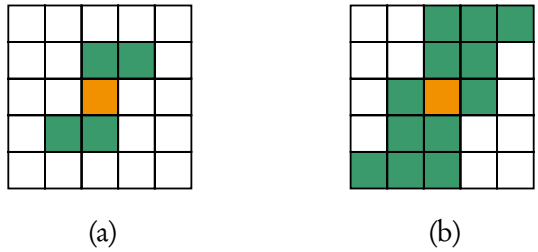


Figure 4.2 (a) Proposed w window (green) for modeling guidance map. (b) Neighborhood (green) for forming matting Laplacian matrix entry for reference pixel (orange).

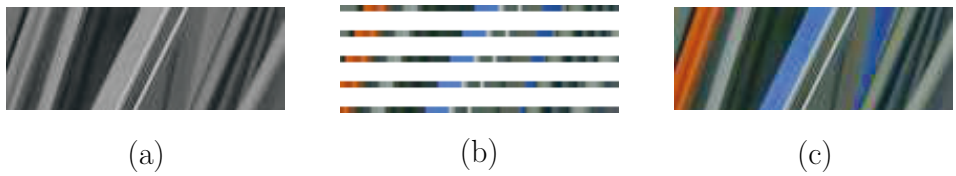


Figure 4.3 (a) Grayscale DSEPI obtained by luminance channel reconstruction using shearlet transform used as a guide for colorization. (b) Color information of the DSEPI is available only from input coarse set of views. (c) Colorization result obtained by solving problem Eq. 4.2.

vergence is presented in Fig. 4.1 (a). As seen in the figure, the algorithm in Eq. (4.1) has to be favored as it provides fast convergence within a fixed parameter α and a relatively small number of iterations Fig. 4.1 (b).

4.2 Colorization

Guided colorization refers to the problem of recovering the color of an image when the color is only available in isolated regions, using a structure (guidance) from a grayscale image [58], [57]. In an approach proposed in [57], the local properties of a grayscale image have been exploited and used for determining the correct propagation of the available color information. The colorization problem is closely related to the alpha matting problem, which assumes a guided foreground/background separation [57].

As a computationally inexpensive alternative to the proposed DSEPI reconstruction method, every color channel of the DSEPI can be reconstructed using some colorization technique as long as a suitable guidance map is available. The hypoth-

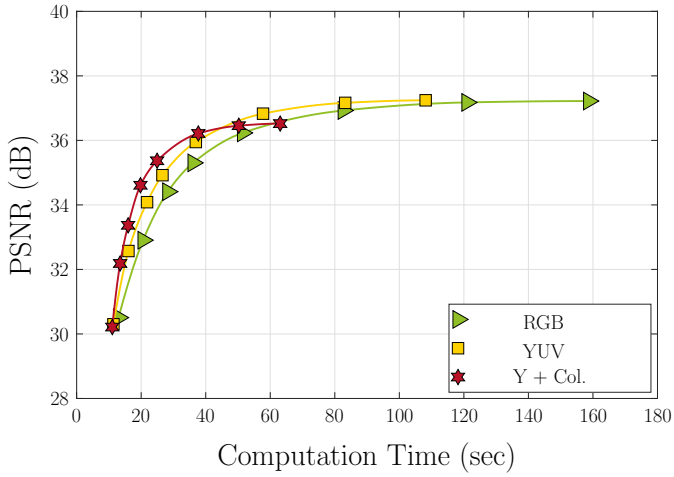


Figure 4.4 The average performance over multiple datasets of the colorization technique (Y+Col.) compared with the reference reconstruction method applied on every color channel independently (RGB) and with efficient reconstruction in YUV color space (YUV).

esis has been formulated as follows [P III]: Let the luminance channel of DSEPI be reconstructed by the method in Eq. 4.1. Using this reconstructed luminance channel as guidance map E, apply guided colorization to reconstruct any color component of the DSEPI x where the color is only available at the locations of the given input views.

Following the method in [57], the unknown color image x is modeled as a linear function of the given guidance map E at each pixel within a small spatial window w

$$x[i] \approx aE[i] + b, \quad \forall i \in w.$$

The solution can be formulated in terms of cost function minimization

$$\min_{x,a,b} J(x,a,b) = \sum_j \left(\sum_{i \in w_j} (x[i] - a_j E[i] - b_j)^2 + \varepsilon a_j^2 \right),$$

where ε is a regularization coefficient to provide numerical stability.

As shown in [57], the minimization problem above can be reformulated in terms of *matting Laplacian* matrix Λ such that

$$\min_{a,b} J(x,a,b) = x^\intercal \Lambda x,$$

where the symmetric matrix Λ depends only on E and w. The (i,j)—th element of the matrix Λ is defined as

$$\Lambda_{ij} = \sum_{k|(i,j)\in w_k} \left(\delta_{ij} - \frac{1}{|w_k|} \left(1 + \frac{(E[i]-\mu_k)(E[j]-\mu_k)}{\frac{\varepsilon}{|w_k|} + \sigma_k} \right) \right),$$

where μ_k, σ_k correspond to the mean and the variance of E within the window w_k; $|w_k|$ denotes the cardinality of the window w_k, and $\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$.

Due to the limitation of 1px maximum disparity in DSEPI, the anisotropic structure in a small window is quite restricted. This property allows us to define a better shape of the window w for the DSEPI colorization problem as illustrated in Fig. 4.2 (a). It is important to mention that the summation in the definition of the matrix elements $\Lambda_{i,j}$ takes all windows w_k, which contain the pair i, j. Therefore for a fixed pixel i, the corresponding pixels j which are involved in the summation are contained in the window presented in Fig. 4.2 (b).

Thus, taking into account the available color information, the guided colorization problem can be formulated as a constrained quadratic minimization,

$$\min_x x^\intercal \Lambda x, \quad \text{subject to} \quad Mx = y, \tag{4.2}$$

where M represents the diagonal measuring matrix and y represents the available color information identical to the one in Eq. 4.1. An approximate solution can be obtained using the conjugate gradient method for the system of linear equations $(\Lambda + \lambda M)x = \lambda My$ with significantly high λ.

An example of the colorization result is given in Fig. 4.3. A detailed evaluation of the colorization approach has been performed in [P III]. It has been observed that the quality of the colorization depends mainly on the accuracy of the guidance map. Therefore, to provide an overall accelerated reconstruction, it is required to efficiently distribute the processing time between the reconstruction of the luminance channel Y by shearlet regularization and the subsequent RGB channel reconstruction by colorization. As seen in Fig. 4.4, some acceleration in terms of better quality reconstruction for the same amount of time has been achieved. The comparison has been performed against the full RGB reconstruction and reconstruction in a YUV color space, where priority has been given to the Y channel which has been processed

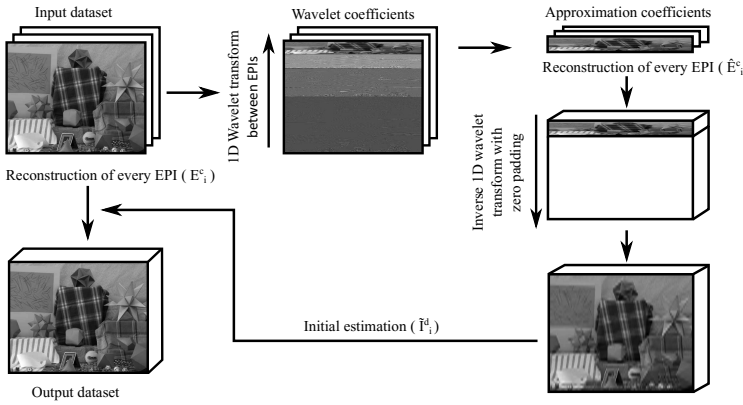


Figure 4.5 Reconstruction flowchart using wavelet transform approximation (lowpass) coefficients as an initial estimation for original set reconstruction.

for twice as long as the U and V channels, due to its higher significance.

4.3 Decorrelation Transform

Another acceleration can be approached by exploiting the spatial correlation between neighboring EPIs. Initially, this idea was pursued by distributing the whole set of EPIs in a processing tree based on the similarity between neighboring EPIs. In this approach, the reconstruction is applied over a tree such that each leaf takes the parent reconstruction result as its initial estimate in order to decrease the needed number of iterations [85]. This idea has been further developed by using wavelet transform [P III]. An initial decorrelating wavelet transform is applied columnwise, followed by DSEPI reconstruction using the approximation wavelet coefficients only. This reconstruction is then used as an initial estimate for the DSEPI reconstruction using shearlet transform. The flowchart of this method is presented in Fig. 4.5. An evaluation of the method has shown a significant dependence on the spatial structure of the input images. For scenes consisting of objects with a simple vertical structure, the wavelet pre-processing significantly decreases the computation time, while for scenes with more complex structures, this method brings only marginal improvement in terms of time versus reconstruction quality.

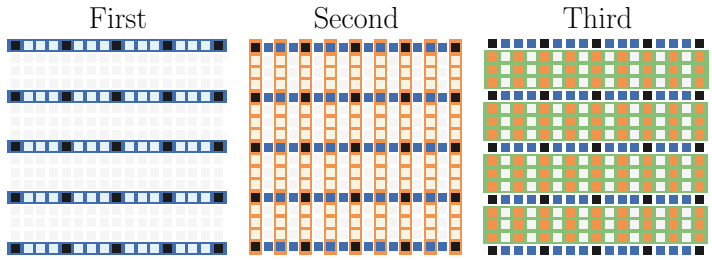


Figure 4.6 Proposed fast processing order illustrated for 17×17 array of images. Reconstruction is divided into three steps (blue, orange, green) to decrease the disparity range in the successive steps.

4.4 Full parallax processing

The core DSEPI reconstruction algorithms have been developed for the case of 1D parallax, that is the pinholes of the cameras providing the input multiview images are considered to be located in a line. The resulting stack of images is three dimensional and the reconstruction is performed on the corresponding stack of 2D EPIs.

In general, the light field is a four-dimensional function, i.e. full parallax. This case can be handled in a separable manner. First, the horizontal parallax is considered and DSEPIs are reconstructed in the corresponding direction. Then, the same reconstruction procedure is applied for the vertical parallax. This approach can be referred to as *direct processing order*. It implies that the same disparity range is considered in both the horizontal and vertical parallax directions. However, the disparity range is the factor which determines the number of shearlet decomposition levels and thus strongly influences the overall reconstruction time. For this reason, smaller disparity ranges are always preferable if it is possible to accommodate them. A *hierarchical reconstruction (HR) order* has been introduced in [P IV] aimed at reducing the computational time by reducing the number of shearlet levels. The HR approach can be illustrated by an example given in Fig. 4.6. The considered setting contains 5×5 input images of a given LF and the goal is to reconstruct the corresponding DSLF which contains 17×17 images. The reconstruction is carried out in three steps, as presented in the figure. The first step is the ordinary horizontal-parallax reconstruction step. During the second step, only selected lines of images along the vertical parallax are fully reconstructed. In this way, the third step contains 12 horizontal parallax sets

still to be reconstructed. These are notable as their disparity range is half the size of the original disparity range, which means a lower number of shearlet levels have to be used. For some large disparity ranges, the number of steps can be increased identically in such a way that at each step, the disparity range of the 1D parallax lines of images to be processed is twice as low as the disparity range in the previous step.

5 APPLICATIONS

The reconstruction method proposed in this thesis can be efficiently used in various applications. As well as the direct application of providing dense LF information, this chapter presents several valuable applications, one of which is super-resolution in the spatial domain. The original angular reconstruction algorithm is extended for spatial super-resolution by considering simultaneous super-resolution in the spatial domain and dense sampling in the angular domain. Another application is obtaining a holographic stereogram and our method has been successfully applied for the accurate calculation of fringe patterns. The last application is light field compression. As shown in the chapter, the intra-view prediction scheme can utilize this reconstruction method, which improves compression, especially in a low bitrates scenario.

5.1 Spatial super-resolution

A new camera design called a light field (plenoptic) camera is presented in [68]. It allows a 4D light field to be captured with a single photographic exposure shot. 4D light field acquisition is achieved using an addiction microlens array between the main lens and sensor. The idea has been developed further in [63] by introducing the focused plenoptic camera. The raw data acquired from the plenoptic camera can be rearranged to form a discrete 4D LF. Theoretically, given certain conditions, identical LF data could be obtained using pinhole cameras uniformly placed within a small baseline. However, in practice this is not achievable due to the physical size of the camera. The LF obtained by the plenoptic camera allows dynamic refocusing using relatively simple post-processing algorithms [47], [67], something which is not achievable with a single shot of a conventional camera. One drawback is that the created refocused images have a smaller spatial resolution than the full sensor size image. Also, due to the small baseline between the elemental images, the disparities change within a small range. Therefore, spatial super-resolution of the captured LF ,

rather than angular super-resolution, is considered in multiple publications [35] [8], [98] [88], .

For the case of super-resolution by a factor of n of the given LF with the range of disparities d_{range} the decimation factor of $n \times d_{\text{range}}$ in both the horizontal and vertical angular dimensions of the unknown high-resolution DSLF should be considered.

Section 3.4 presented a reconstruction method aimed at angular super-resolution. To generalize this to the case of spatial super-resolution for an LF obtained with a plenoptic camera, angular and spatial reconstruction (super-resolution) are considered together. Thus, the following two linked forward models are considered:

$$y = M^{\text{spatial}} x^{\text{sr}}, \quad x^{\text{sr}} = M^{\text{angular}} x^{\text{ds}},$$

where y is a given low-resolution LF formed from x^{sr} the spatial high-resolution LF using $n \times n$ spatial downsampling represented with a matrix M^{spatial}. In turn, the final x^{ds} spatial high-resolution DSLF is related with x^{sr} through the M^{angular} angular domain decimation matrix (decimation factor is $n \times d_{\text{range}}$).

The proposed reconstruction method for the super-resolution problem is represented by sequentially solving two inverse problems using the algorithm from Eq. 3.17 and gradient descent, i.e.

$$\begin{aligned} x^{\text{ds}}_{k+1} &= S^* \left(\mathscr{T}_{\lambda_k} \left(S(x^{\text{ds}}_k + \alpha_k (x^{\text{sr}}_k - M^{\text{angular}} x^{\text{ds}}_k)) \right) \right), \\ x^{\text{sr}}_{k+1} &= x^{\text{sr}}_k + \tau \left(A^{\text{spatial}} (y - M^{\text{spatial}} x^{\text{sr}}_k) + \gamma M^{\text{angular}} x^{\text{ds}}_{k+1} \right) \end{aligned} \tag{5.1}$$

where the A^{spatial} transform is an approximation of the inverse M^{spatial} represented with an interpolation filter and guided filtering similar to those chosen in Eq. (2.4) from [6]. The second step in the iterative procedure represents gradient descent for the minimization problem $\arg\min_{x^{\text{sr}}} \left\| y - M^{\text{spatial}} x^{\text{sr}} \right\|_2^2 + \gamma \left\| x^{\text{sr}} - M^{\text{angular}} x^{\text{ds}}_{k+1} \right\|_2^2$.

Two cases of spatial downsampling are considered to evaluate the proposed algorithm. The downsampling operator is formed from the filtering operator and decimation by a certain factor. The examined filtering operators are Gaussian filtering and an averaging filter. The considered decimation factors are $\times 2$, $\times 3$, $\times 4$. Identical assumptions have been analyzed in recent publications [6], [71]. The super-resolution results for the Stanford Light field dataset [86] are presented in Tables 5.1 and 5.2. To reduce computation time only 512×512 central parts of the images are used. Along the angular dimensions, only the 5×5 central subset is used. The error

Gaussian filtering	Upsampling			LFBM5D			Shearlet+GF		
	×2	×3	×4	×2	×3	×4	×2	×3	×4
Amethyst	30.13	29.37	28.36	**34.05**	31.87	29.15	33.49	**32.08**	**29.88**
Bracelet	27.66	26.71	25.35	**34.15**	30.73	26.24	33.98	**30.91**	**27.23**
Chess	32.26	31.43	30.25	**37.65**	**35.43**	31.60	37.00	34.75	**32.07**
Eucalyptus Flowers	23.83	23.26	22.58	**26.33**	24.63	22.83	26.13	**24.74**	**23.24**
Jelly Beans	42.95	42.06	40.69	**47.47**	**46.41**	**44.12**	46.92	46.23	43.34
Lego Bulldozer	29.25	28.42	27.22	**34.40**	**32.38**	**29.47**	34.15	32.33	29.25
Lego Knights	29.79	29.03	27.91	34.51	**32.90**	**30.11**	**34.56**	32.61	29.72
Lego Truck	28.75	28.03	26.98	**32.46**	**30.74**	28.01	32.29	30.63	**28.41**
Tarot Cards and Crystal Ball (Large Angle)	28.04	27.20	25.94	**33.48**	30.19	27.18	33.28	**30.34**	**27.41**
Tarot Cards and Crystal Ball (Small Angle)	28.19	27.51	26.34	**32.60**	**30.57**	27.21	32.11	30.24	**27.71**
The Stanford Bunny	37.07	36.23	34.87	**41.71**	**40.88**	**38.46**	41.37	40.47	37.48
Treasure Chest	24.58	23.90	22.96	**28.35**	25.96	23.84	28.28	**26.36**	**24.12**
Average	30.21	29.43	28.29	**34.76**	**32.72**	29.85	34.46	32.64	**29.99**

Table 5.1 Reconstruction results in PSNRs for the case of downsampling using the Gaussian filter [6]. Upsampling represents basic interpolation using a low-pass filter designed for the corresponding decimation factor.

Averaging filter	Upsampling			GB+DR			Shearlet+GF		
	×2	×3	×4	×2	×3	×4	×2	×3	×4
Amethyst	35.46	31.18	28.95	**37.96**	**33.35**	**30.56**	35.31	31.78	29.75
Bracelet	36.24	29.42	26.06	**37.54**	**34.28**	**28.57**	36.52	30.60	27.05
Chess	38.32	33.54	30.81	**41.36**	**36.16**	**32.73**	38.72	34.43	31.76
Eucalyptus Flowers	27.43	24.21	22.84	**31.07**	**26.38**	**24.09**	27.49	24.45	23.11
Jelly Beans	**49.84**	45.13	41.53	48.58	**46.85**	**44.04**	46.68	46.38	43.74
Lego Bulldozer	35.40	30.75	27.95	33.90	**32.68**	**30.52**	**35.54**	31.89	29.12
Lego Knights	34.93	30.88	28.53	**38.06**	**33.83**	**30.90**	35.24	32.14	29.70
Lego Truck	32.87	29.59	27.42	**36.38**	**31.57**	**29.02**	32.90	30.17	28.18
The Stanford Bunny	43.06	38.92	35.84	**44.75**	**41.16**	**38.19**	41.43	40.33	37.72
Treasure Chest	29.59	25.35	23.38	**34.73**	**27.93**	**25.37**	30.05	25.94	23.96
Average	36.31	31.90	29.33	**38.43**	**34.42**	**31.40**	35.99	32.81	30.41

Table 5.2 Reconstruction results in the case of downsampling using averaging of $n \times n$ block of pixels.

is calculated in terms of average PSNR(dB) using 17 px cropping from the borders in order to remove border artifacts. As can be seen in the tables, the shearlet-based spatial super-resolution is comparable with the two state of the art methods, while being with lower computational complexity. Visual examples of reconstructed images are given in Fig. 5.1.

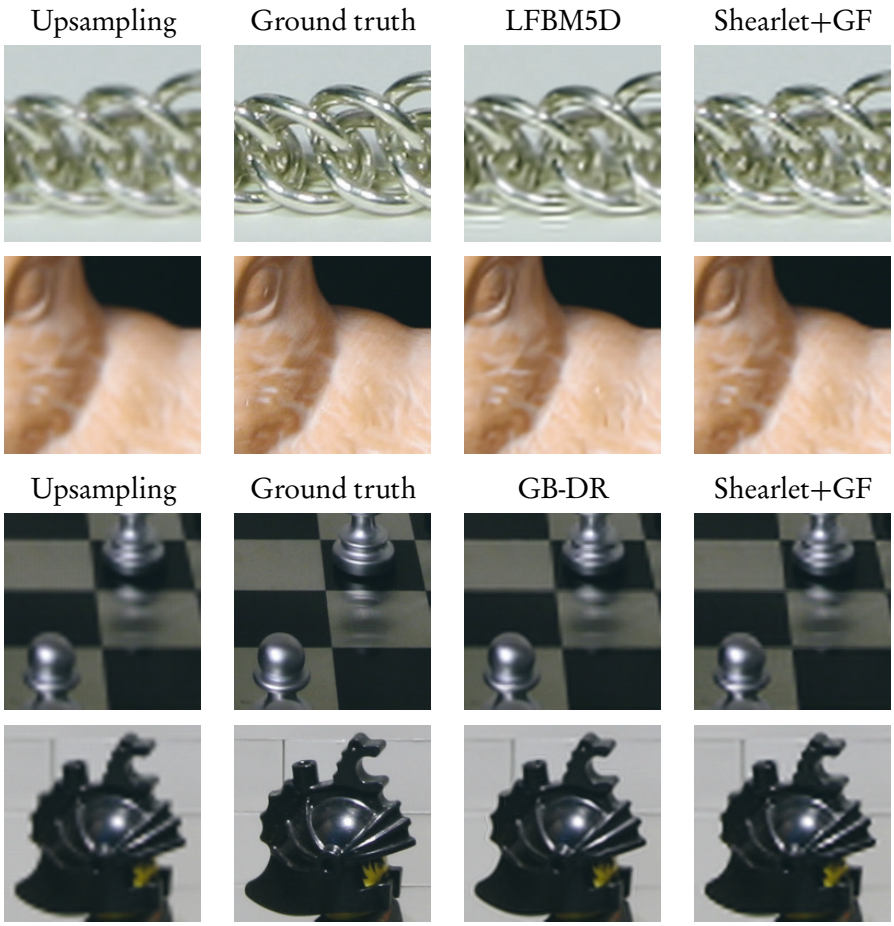


Figure 5.1 Examples of reconstructions for $\times 4$ super-resolution using LFBM5D [6], GB-DR [71], and proposed Shearlet+GF

.

5.2 Holographic stereogram

Our proposed view reconstruction method can be used in various applications requiring a dense set of views. Calculating a holographic stereogram is one example of such an application, which is considered in [P II]. Each ray of the LF is represented as a windowed plane wave with its corresponding amplitude. Thus, the whole LF forms a superposition of plane waves. For convenience, the two planes of the LF parameterization are located on the camera plane and the hologram plane. Thus, all the rays intersecting a point on the hologram plane form a so-called *hogel*. A fringe

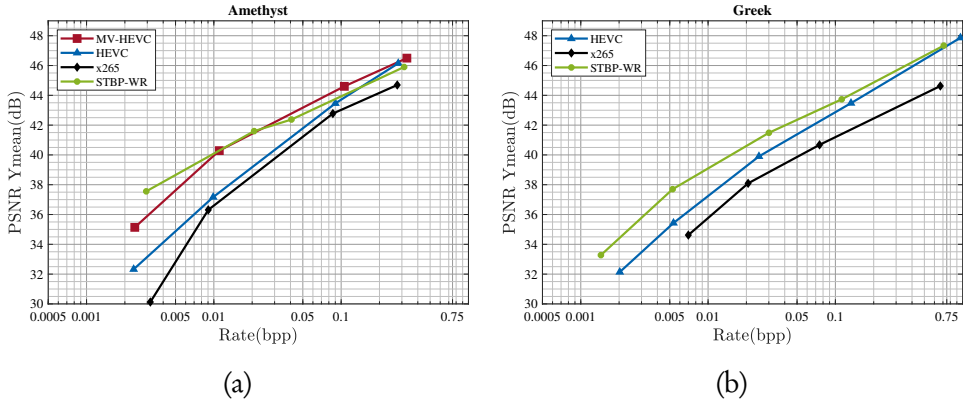


Figure 5.2 Illustration of the efficiency of the method proposed in [4] (STBP-WR).

pattern corresponding to a hogel on the holographic stereogram is calculated using a superposition of the plane waves that correspond to the rays in the hogel. The resolution of each hogel is directly related to the angular resolution of the given LF. Thus, high angular resolution provides an accurate calculation of the fringe pattern corresponding to each hogel. All the fringe patterns together form a full holographic stereogram. The experimental results described in [P II] used synthetic data so that the required ground-truth LF data is available. The results reported in the publication show the value of using DSLF for the holographic stereogram calculation and the efficiency of the proposed shearlet-based algorithm for obtaining DSLF. Reconstruction is employed on the 7×7 LF so that the upsampled DSLF has 49×49 angular resolution. The reconstructed DSLF is further converted to a holographic stereogram using bilinear resampling in order to obtain continuous spatial frequencies. The proposed method for the calculation of a holographic stereogram performed better than the depth-based approach. More details about these experiments can be found in [P II].

5.3 Light field compression

In [3], [4] Waqas et al. successfully used DSLF reconstruction for a general LF compression task. Typically, the LF compression problem is interpreted as compression of corresponding sub-aperture images. A significant improvement in compression can be achieved by using an enhanced inter-view prediction scheme. Nevertheless,

providing such an inter-view prediction scheme is not an easy task. An alternative approach is proposed in [3], [4]. If the input LF has first been uniformly decimated in the angular domain to form a set of key views, this significantly decreases the number of images that need to be compressed. The key views are converted into a pseudo video sequence and compressed (encoded) using high-efficiency video coding (HEVC). On the decoding side, the full LF is reconstructed using only the decoded key views. The optimal parameters for decimation and encoding have been considered in [3]. As an anchor method, the direct encoding of the full set of images from the LF as a pseudo video sequence is considered. The obtained results demonstrate the efficiency of the proposed compression scheme in terms of lower bit-rates than for the anchor method. Since the reconstruction method based on the shearlet transform relies only on key views, in a low bit-rate scenario, the bit budget allows the achievement of high quality key views and, as a consequence, high quality for the reconstructed DSLF. On the other hand, the anchor achieves efficient compression at high bit-rates high quality by efficiently encoding the residual information. This property is illustrated in Fig. 5.2 by comparing different compression methods for two full parallax light field datasets. More results can be found in [4].

6 DISCUSSION AND CONCLUSIONS

The general approach for intermediate view interpolation from multi-view imagery consists of two steps: depth estimation using the available views and the subsequent view synthesis. Inaccuracies in the depth estimation result in significant artifacts in the final synthesized views which leads to poor reconstruction quality. Alternative approaches cast this problem within the light field framework. By considering two-plane LF parameterization, the intermediate view synthesis problem can be reformulated as reconstructing the continuous 4D function (light field) which carries all the visual information about the 3D scene. In early research, the light field reconstruction problem was addressed from a classical sampling perspective, i.e. analysing the LF spectrum and suggesting sampling and reconstruction under specified limitations for the 3D scene. This thesis has proposed a novel method for reconstructing a light field from a given set of views. Our approach addresses the problem using modern signal processing techniques, namely frame decomposition and sparsification in a suitable transform domain. At first, the general light field reconstruction problem is formalized as a reconstruction of a densely sampled light field from a given insufficient number of samples (coarse set of samples) along the angular dimension. Further, the inverse problem of the densely sampled light field reconstruction is addressed by considering a global optimization of a minimization problem with a regularization term formed as the sparsity in a certain transform domain. Based on previous studies about the spectral properties of light fields, we have proposed selecting the shearlet transform as a convenient transform. The resulting optimization problem is solved with an iterative algorithm initially presented in Publication I, and further developed in Publication IV. The presented reconstruction scheme allows us to avoid any direct depth estimation and the only required information is the disparity range of the whole visible scene. The method has demonstrated state of the art performance, especially for scenes containing semi-transparent objects.

To address the problem of speeding up the proposed method, several acceleration

techniques have been proposed in Publication III. We have demonstrated that a multi-channel light field can be reconstructed using a colorization technique that can be efficiently used to avoid calculating the computationally demanding overcomplete transform. Additionally, acceleration is achieved by considering a decorrelation transform to provide an adequate initial estimate for the iterative reconstruction method. For an efficient reconstruction of a full parallax light field, a hierarchical order of 1D reconstructions is proposed.

Obtaining the densely sampled light field allows the consideration of several applications where dense visual information of the scene is required. As presented in Publication II, calculating holographic stereograms is one such application. The presented results show the efficiency of the densely sampled light field reconstruction for calculating holographic stereograms, something which cannot be acquired otherwise. The proposed reconstruction method can also be applied for LF spatial super-resolution, as reported in this thesis.

The shearlet approach to LF reconstruction is based on rigorous mathematical formalism, which couples modern signal sparsification theory with ray optics based light propagation modelling. In recent years, machine learning has emerged as a powerful tool to solve problems related with image classification and reconstruction. Methods for light field reconstruction based on convolutional neural networks have claimed state of the art performance. The method proposed in the thesis compares favorably against those more recent methods, especially for wide camera baselines and corresponding high disparities [32], [31]. In fact, due to its accurate modeling of directional properties of the plenoptic function, the method is well suited for machine learning based extensions. In [32], it has been extended by a deep CNN to predict the residuals of the shearlet coefficients in shearlet domain, thus improving its performance for moderate disparity ranges and increasing the processing speed 2.4 times. In [31], the iterative regularization procedure has been replaced by a self-supervised learned regularizer, trained directly on sparsely sampled light field data with small disparity ranges (≤ 8 pixels). This has led to an improved performance for datasets with large disparity ranges (16 - 32 pixels) and 9 times speedup over the original approach.

REFERENCES

[1] E. H. Adelson and J. R. Bergen. The Plenoptic Function and the Elements of Early Vision. *Computational Models of Visual Processing*. 1991, 3–20.

[2] M. Aharon, M. Elad and A. Bruckstein. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing* 54.11 (Dec. 2006), 4311–4322.

[3] W. Ahmad, S. Vagharshakyan, M. Sjöström, A. Gotchev, R. Bregovic and R. Olsson. Shearlet Transform Based Prediction Scheme for Light Field Compression. *2018 Data Compression Conference*. Mar. 2018, 396–396.

[4] W. Ahmad, S. Vagharshakyan, M. Sjöström, A. Gotchev, R. Bregovic and R. Olsson. Shearlet Transform based Light Field Compression Under Low Bitrates. *IEEE Transactions on Image Processing* (Jan. 2020). accepted for publication.

[5] M. Alain and A. Smolic. Light Field Denoising by Sparse 5D Transform Domain Collaborative Filtering. *19th International Workshop on Multimedia Signal Processing (MMSP)*. Oct. 2017, 1–6.

[6] M. Alain and A. Smolic. Light Field Super-Resolution via LFBM5D Sparse Coding. *IEEE International Conference on Image Processing (ICIP)*. Oct. 2018, 2501–2505.

[7] A. Alperovich, O. Johannsen, M. Strecke and B. Goldluecke. Light Field Intrinsics With a Deep Encoder-Decoder Network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[8] T. E. Bishop, S. Zanetti and P. Favaro. Light Field Superresolution. *IEEE International Conference on Computational Photography (ICCP)*. Apr. 2009, 1–9.

[9] T. Blumensath and M. E. Davies. Normalized Iterative Hard Thresholding: Guaranteed Stability and Performance. *IEEE Journal of Selected Topics in Signal Processing* 4.2 (Apr. 2010), 298–309.

[10] 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* 1.1 (Mar. 1987), 7–55.

[11] K. Bredies, K. Kunisch and T. Pock. Total Generalized Variation. *SIAM Journal on Imaging Sciences* 3.3 (Jan. 2010), 492–526.

[12] J.-F. Cai, R. H. Chan and Z. Shen. A Framelet-Based Image Inpainting Algorithm. *Applied and Computational Harmonic Analysis* 24.2 (Mar. 2008), 131–149.

[13] J.-F. Cai and Z. Shen. Framelet Based Deconvolution. *Journal of Computational Mathematics* 28.3 (2010), 289–308.

[14] E. Camahort, A. Lerios and D. Fussell. Uniformly Sampled Light Fields. *Rendering Techniques '98*. 1998, 117–130.

[15] E. J. Candès, J. Romberg and T. Tao. Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information. *IEEE Transactions on Information Theory* 52.2 (Feb. 2006), 489–509.

[16] E. J. Candes and T. Tao. Decoding by Linear Programming. *IEEE Transactions on Information Theory* 51.12 (Dec. 2005), 4203–4215.

[17] E. J. Candès and D. L. Donoho. Ridgelets: a Key to Higher-dimensional Intermittency?: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 357.1760 (1999), 2495–2509.

[18] E. J. Candès and D. L. Donoho. New Tight Frames of Curvelets and Optimal Representations of Objects With Piecewise C^2 Singularities. *Communications on Pure and Applied Mathematics* 57.2 (2004), 219–266.

[19] Cha Zhang and Tsuhan Chen. Spectral Analysis for Sampling Image-based Rendering Data. *IEEE Transactions on Circuits and Systems for Video Technology* 13.11 (Nov. 2003), 1038–1050.

[20] J.-X. Chai, X. Tong, S.-C. Chan and H.-Y. Shum. Plenoptic Sampling. *27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '00. 2000, 307–318.

[21] S. H. Chan, X. Wang and O. A. Elgendy. Plug-and-Play ADMM for Image Restoration: Fixed-Point Convergence and Applications. *IEEE Transactions on Computational Imaging* 3.1 (Mar. 2017), 84–98.

[22] S. E. Chen. QuickTime VR: An Image-based Approach to Virtual Environment Navigation. *22nd Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '95. 1995, 29–38.

[23] P. L. Combettes and J.-C. Pesquet. Proximal Splitting Methods in Signal Processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Ed. by H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke and H. Wolkowicz. New York, NY, 2011, 185–212.

[24] A. Danielyan, V. Katkovnik and K. Egiazarian. BM3D Frames and Variational Image Deblurring. *IEEE Transactions on Image Processing* 21.4 (Apr. 2012), 1715–1728.

[25] M. N. Do, D. Marchand-Maillet and M. Vetterli. On the Bandwidth of the Plenoptic Function. *IEEE Transactions on Image Processing* 21.2 (Feb. 2012), 708–717.

[26] M. N. Do and M. Vetterli. The Contourlet Transform: an Efficient Directional Multiresolution Image Representation. *IEEE Transactions on Image Processing* 14.12 (Dec. 2005), 2091–2106.

[27] D. L. Donoho. Sparse Components of Images and Optimal Atomic Decompositions. *Constructive Approximation* 17.3 (Jan. 2001), 353–382.

[28] J. M. Duarte-Carvajalino and G. Sapiro. Learning to Sense Sparse Signals: Simultaneous Sensing Matrix and Sparsifying Dictionary Optimization. *IEEE Transactions on Image Processing* 18.7 (July 2009), 1395–1408.

[29] R. Farrugia and C. Guillemot. Light Field Super-Resolution using a Low-Rank Prior and Deep Convolutional Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[30] C. Fehn. Depth-image-based Rendering (DIBR), Compression, and Transmission for a New Approach on 3D-TV. *Stereoscopic Displays and Virtual Reality Systems XI*. Vol. 5291. 2004, 93–105.

[31] Y. Gao, R. Bregovic and A. Gotchev. Self-Supervised Light Field Reconstruction Using Shearlet Transform and Cycle Consistency. *arXiv e-prints*, arXiv:2003.09294 (Mar. 2020).

[32] Y. Gao, R. Bregovic, R. Koch and A. Gotchev. DRST: Deep Residual Shearlet Transform for Densely Sampled Light Field Reconstruction. *arXiv e-prints*, arXiv:2003.08865 (Mar. 2020).

[33] M. Genzel and G. Kutyniok. Asymptotic Analysis of Inpainting via Universal Shearlet Systems. *SIAM Journal on Imaging Sciences* 7.4 (2014), 2301–2339.

[34] T. Georgiev and C. Intwala. Light Field Camera Design for Integral View Photography. *Adobe Technical Report* (2006).

[35] T. Georgiev and A. Lumsdaine. Superresolution with Plenoptic Camera 2.0. *Technical Report, Adobe Systems Incorporated* (2009).

[36] C. Gilliam, P. L. Dragotti and M. Brookes. A Closed-form Expression for the Bandwidth of the Plenoptic Function Under Finite Field of View Constraints. *IEEE International Conference on Image Processing (ICIP)*. Sept. 2010, 3965–3968.

[37] C. Gilliam, P. Dragotti and M. Brookes. On the Spectrum of the Plenoptic Function. *IEEE Transactions on Image Processing* 23.2 (Feb. 2014), 502–516.

[38] S. J. Gortler, R. Grzeszczuk, R. Szeliski and M. F. Cohen. The Lumigraph. *23rd Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '96. 1996, 43–54.

[39] C. Guillemot and O. Le Meur. Image Inpainting: Overview and Recent Advances. *IEEE signal processing magazine* 31.1 (2013), 127–144.

[40] S. Heber and T. Pock. Convolutional Networks for Shape from Light Field. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016, 3746–3754.

[41] S. Heber, W. Yu and T. Pock. Neural EPI-Volume Networks for Shape from Light Field. *IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, 2271–2279.

[42] A. Heyden and M. Pollefeys. Multiple View Geometry. *Emerging topics in computer vision* (2005), 45–107.

[43] H. Hirschmuller. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.2 (Feb. 2008), 328–341.

[44] K. Honauer, O. Johannsen, D. Kondermann and B. Goldluecke. A Dataset and Evaluation Methodology for Depth Estimation on 4D Light Fields. *Asian Conference on Computer Vision*. 2016.

[45] A. Hosni, C. Rhemann, M. Bleyer, C. Rother and M. Gelautz. Fast Cost-Volume Filtering for Visual Correspondence and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.2 (Feb. 2013), 504–511.

[46] I. Ihm, S. Park and R. K. Lee. Rendering of Spherical Light Fields. *Fifth Pacific Conference on Computer Graphics and Applications*. Oct. 1997, 59–68.

[47] A. Isaksen, L. McMillan and S. J. Gortler. Dynamically Reparameterized Light Fields. *27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '00. 2000, 297–306.

[48] H. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. Tai and I. S. Kweon. Accurate Depth Map Estimation From a Lenslet Light Field Camera. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, 1547–1555.

[49] N. K. Kalantari, T.-C. Wang and R. Ramamoorthi. Learning-Based View Synthesis for Light Field Cameras. *ACM Transactions on Graphics (TOG)* 35.6 (Nov. 2016), 1–10.

[50] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung and M. Gross. Scene Reconstruction From High Spatio-angular Resolution Light Fields. *ACM Transactions on Graphics (TOG)* 32.4 (July 2013), 1–12.

[51] V. Kolmogorov and R. Zabih. Multi-Camera Scene Reconstruction via Graph Cuts. *7th European Conference on Computer Vision-Part III*. ECCV'02. 2002, 82–96.

[52] D. Krishnan and R. Fergus. Fast Image Deconvolution using Hyper-Laplacian Priors. *Advances in Neural Information Processing Systems 22*. 2009, 1033–1041.

[53] G. Kutyniok, J. Lemvig and W.-Q. Lim. *Shearlets: Multiscale Analysis for Multivariate Data*. Ed. by G. Kutyniok and D. Labate. 2012.

[54] G. Kutyniok and W.-Q. Lim. Compactly Supported Shearlets Are Optimally Sparse. *Journal of Approximation Theory* 163.11 (2011), 1564–1589.

[55] A. W. F. Lee, W. Sweldens, P. Schröder, L. Cowsar and D. Dobkin. MAPS: Multiresolution Adaptive Parameterization of Surfaces. *25th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '98. 1998, 95–104.

[56] H. Lee, A. Battle, R. Raina and A. Y. Ng. Efficient Sparse Coding Algorithms. *Proceedings of the 19th International Conference on Neural Information Processing Systems*. NIPS'06. Canada: MIT Press, 2006, 801–808.

[57] A. Levin, D. Lischinski and Y. Weiss. A Closed-Form Solution to Natural Image Matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.2 (Feb. 2008), 228–242.

[58] A. Levin, D. Lischinski and Y. Weiss. Colorization Using Optimization. *ACM Transactions on Graphics (TOG)* 23.3 (Aug. 2004), 689–694.

[59] M. Levoy and P. Hanrahan. Light Field Rendering. *23rd Annual Conference on Computer Graphics and Interactive Techniques*. 1996, 31–42.

[60] M. Levoy, R. Ng, A. Adams, M. Footer and M. Horowitz. Light Field Microscopy. *ACM Transactions on Graphics (TOG)* 25.3 (July 2006), 924.

[61] W.-Q. Lim. Nonseparable Shearlet Transform. *IEEE Transactions on Image Processing* 22.5 (May 2013), 2056–2065.

[62] Z. Lin and H.-Y. Shum. A Geometric Analysis of Light Field Rendering. *International Journal of Computer Vision* 58.2 (July 2004), 121–138.

[63] A. Lumsdaine and T. Georgiev. The Focused Plenoptic Camera. *IEEE International Conference on Computational Photography (ICCP)*. Apr. 2009, 1–8.

[64] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 2008.

[65] K. Marwah, G. Wetzstein, Y. Bando and R. Raskar. Compressive Light Field Photography Using Overcomplete Dictionaries and Optimized Projections. *ACM Transactions on Graphics (TOG)* 32.4 (July 2013), 1–12.

[66] L. McMillan and G. Bishop. Plenoptic Modeling: An Image-based Rendering System. *22nd Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '95. 1995, 39–46.

[67] R. Ng. Fourier Slice Photography. *ACM Transactions on Graphics (TOG)* 24.3 (July 2005), 735–744.

[68] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan et al. Light Field Photography With a Hand-held Plenoptic Camera. *Computer Science Technical Report (CSTR)* 2.11 (2005), 1–11.

[69] R. S. Overbeck, D. Erickson, D. Evangelakos, M. Pharr and P. Debevec. A System for Acquiring, Processing, and Rendering Panoramic Light Field Stills for Virtual Reality. *ACM Transactions on Graphics (TOG)* 37 (Dec. 2018), 1–15.

[70] K. Qiu and A. Dogandžić. Double Overrelaxation Thresholding Methods for Sparse Signal Reconstruction. *44th Annual Conference on Information Sciences and Systems (CISS)*. Mar. 2010, 1–6.

[71] M. Rossi and P. Frossard. Geometry-Consistent Light Field Super-Resolution via Graph-Based Regularization. *IEEE Transactions on Image Processing* 27.9 (Sept. 2018), 4207–4218.

[72] E. Sahin, S. Vagharshakyan, J. Mäkinen, R. Bregovic and A. Gotchev. Shearlet-Domain Light Field Reconstruction for Holographic Stereogram Generation. *2016 IEEE International Conference on Image Processing (ICIP)*. Sept. 2016, 1479–1483.

[73] Z. Shen, K. Toh and S. Yun. An Accelerated Proximal Gradient Algorithm for Frame-Based Image Restoration via the Balanced Approach. *SIAM Journal on Imaging Sciences* 4.2 (Jan. 2011), 573–596.

[74] L. Shi, H. Hassanieh, A. Davis, D. Katabi and F. Durand. Light Field Reconstruction Using Sparsity in the Continuous Fourier Domain. *ACM Transactions on Graphics (TOG)* 34.1 (Dec. 2014), 1–13.

[75] C. Shin, H. 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. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 2018, 4748–4757.

[76] H.-Y. Shum and L.-W. He. Rendering with Concentric Mosaics. *26th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '99. 1999, 299–306.

[77] E. P. Simoncelli, W. T. Freeman, E. H. Adelson and D. J. Heeger. Shiftable Multiscale Transforms. *IEEE Transactions on Information Theory* 38.2 (Mar. 1992), 587–607.

[78] R. Szeliski, H.-Y. Shum, H.-Y. Shum and H.-Y. Shum. Creating Full View Panoramic Image Mosaics and Environment Maps. *24th Annual Conference on Computer Graphics and Interactive Techniques*. 1997, 251–258.

[79] M. Tanimoto. Overview of FTV (Free-Viewpoint Television). *IEEE International Conference on Multimedia and Expo*. June 2009, 1552–1553.

[80] M. W. Tao, S. Hadap, J. Malik and R. Ramamoorthi. Depth From Combining Defocus and Correspondence Using Light-field Cameras. *IEEE International Conference on Computer Vision (ICCV)*. Dec. 2013, 673–680.

[81] A. N. Tikhonov, A. Goncharsky, V. V. Stepanov and A. G. Yagola. *Numerical Methods for the Solution of Ill-posed Problems*. Vol. 328. 2013.

[82] S. Vagharshakyan, R. Bregovic and A. Gotchev. Image Based Rendering Technique via Sparse Representation in Shearlet Domain. *2015 IEEE International Conference on Image Processing (ICIP)*. Sept. 2015, 1379–1383.

[83] S. Vagharshakyan, R. Bregovic and A. Gotchev. Accelerated Shearlet-Domain Light Field Reconstruction. *IEEE Journal of Selected Topics in Signal Processing* 11.7 (Oct. 2017), 1082–1091.

[84] S. Vagharshakyan, R. Bregovic and A. Gotchev. Light Field Reconstruction Using Shearlet Transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.1 (Jan. 2018), 133–147.

[85] S. Vagharshakyan, R. Bregovic and A. Gotchev. Tree-Structured Algorithm for Efficient Shearlet-domain Light Field Reconstruction. *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. 2015, 478–482.

[86] V. Vaish and A. Adams. *The (New) Stanford Light Field Archive*. `http://lightfield.stanford.edu`. 2008.

[87] K. Wakunami, M. Yamaguchi and B. Javidi. High-Resolution Three-Dimensional Holographic Display Using Dense Ray Sampling From Integral Imaging. *Optics letters* 37.24 (2012), 5103–5105.

[88] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun and T. Tan. LFNet: A Novel Bidirectional Recurrent Convolutional Neural Network for Light-Field Image Super-Resolution. *IEEE Transactions on Image Processing* 27.9 (Sept. 2018), 4274–4286.

[89] S. Wanner and B. Goldluecke. Variational Light Field Analysis for Disparity Estimation and Super-Resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.3 (Mar. 2014), 606–619.

[90] Z. Wen, D. Goldfarb and W. Yin. Alternating Direction Augmented Lagrangian Methods for Semidefinite Programming. *Mathematical Programming Computation* 2.3-4 (2010), 203–230.

[91] B. S. Wilburn, M. Smulski, H.-H. K. Lee and M. A. Horowitz. Light Field Video Camera. *Media Processors 2002*. Vol. 4674. 2001, 29–37.

[92] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz and M. Levoy. High Performance Imaging Using Large Camera Arrays. *ACM Transactions on Graphics (TOG)* 24.3 (July 2005), 765–776.

[93] D. N. Wood, D. I. Azuma, K. Aldinger, B. Curless, T. Duchamp, D. H. Salesin and W. Stuetzle. Surface Light Fields for 3D Photography. *27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '00. 2000, 287–296.

[94] G. Wu, Y. Liu, L. Fang, Q. Dai and T. Chai. Light Field Reconstruction Using Convolutional Network on EPI and Extended Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[95] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai and Y. Liu. Light Field Reconstruction Using Deep Convolutional Network on EPI. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017, 1638–1646.

[96] X. Xiao, B. Javidi, M. Martinez-Corral and A. Stern. Advances in Three-Dimensional Integral Imaging: Sensing, Display, and Applications. *Applied optics* 52.4 (2013), 546–560.

[97] T. Yasuhiro, T. Kosuke and N. Junya. Super Multi-view Display With a Lower Resolution Flat-panel Display. *Optics Express* 19.5 (Feb. 2011), 4129–4139.

[98] Y. Yoon, H. Jeon, D. Yoo, J. Lee and I. S. Kweon. Learning a Deep Convolutional Network for Light-Field Image Super-Resolution. *IEEE International Conference on Computer Vision Workshop (ICCVW)*. Dec. 2015, 57–65.

[99] Y. Yoon, H. Jeon, D. Yoo, J. Lee and I. S. Kweon. Light-Field Image Super-Resolution Using Convolutional Neural Network. *IEEE Signal Processing Letters* 24.6 (June 2017), 848–852.

[100] Z. Yu, X. Guo, H. Ling, A. Lumsdaine and J. Yu. Line Assisted Light Field Triangulation and Stereo Matching. *IEEE International Conference on Computer Vision (ICCV)*. Dec. 2013, 2792–2799.

PUBLICATIONS

PUBLICATION

I

Image Based Rendering Technique via Sparse Representation in Shearlet Domain

S. Vagharshakyan, R. Bregovic and A. Gotchev

*2015 IEEE International Conference on Image Processing (ICIP)*2015, 1379–1383

IMAGE BASED RENDERING TECHNIQUE VIA SPARSE REPRESENTATION IN SHEARLET DOMAIN

Suren Vagharshakyan, Robert Bregovic, Atanas Gotchev

Department of Signal Processing, Tampere University of Technology, Tampere, Finland

ABSTRACT

In this paper we propose a method for reconstructing a densely sampled light field from a given sparse set of perspective views from rectified cameras without an explicit estimation of the scene depth. The desired intermediate views are synthesized by inpainting of epipolar-plane images, utilizing their sparsity in the shearlet domain. For the purpose of shearlet-domain representation, compactly supported shearlets have been constructed using different directional filters for different scales in an attempt to provide better directional selectivity at lower scales. The reconstruction procedure with shearlet-domain sparsity condition is implemented through an iterative thresholding algorithm. The performance of the method is quantified by tests on synthetic and real visual data and compared favorably against depth-image based rendering.

Index Terms— Light field, sparse reconstruction, shearlet, image based rendering

1. INTRODUCTION

Modern image based rendering (IBR) methods are based on two, fundamentally different, approaches. First approach is based on estimating the scene geometry, e.g. in the form of depth map(s), from a given set of images (views) [1], [2], [3] and synthesizing the desired views using the estimated depth maps and the given images [4], [5]. Second approach is based on the light field (LF) concept as introduced by Levoy and Hanrahan [6]. This concept considers each pixel of the given views as a sample of a multidimensional LF function, therefore the view synthesis problem transforms to the problem of continuous LF reconstruction and subsequent interpolation at the desired points, performed with no use of explicit depth estimation. In [7], different kernels for interpolation with the usage of available geometrical information are considered. However, this interpolation technique requires a substantial number of samples (images), as discussed in [8] where Lin and Shum derive precise bounds of the LF sampling.

In order to synthesize novel views without ghosting artefacts based only on linear interpolation one needs to sample the LF such that the disparity between nearby views is less than *one* pixel. Hereafter, we refer to this kind of sampled LF as *densely sampled*. Densely sampled LF provides sufficient information about scene's visual content for all practical image-based applications such as refocused image generation [9], depth estimation [10], [11], novel view generation for free viewpoint television [12] and holographic stereogram [13].

In order to capture a densely sampled LF, the required distance between nearby camera positions can be estimated based on the lower bound of the depth of the scene and the camera resolution. Furthermore, camera resolution should provide enough samples to properly capture highest spatial texture frequency in a scene [14].

In [15], it has been shown that seismic data from limited number of measurements can be efficiently reconstructed by using an inpainting technique based on shearlet-domain representation. We employ this idea and present a method for reconstruction of a densely sampled LF from a given sparse set of views, which requires no explicit depth information. The proposed method is based on a sparse representation in shearlet domain of every decimated epipolar-plane image (EPI) slice of the densely sampled LF. Available data (captured views) can be interpreted as known rows in the EPI's. By applying inpainting technique on every EPI, we can reconstruct all unknown samples of the densely sampled LF. The proposed method enables one to capture the scene with a smaller number of cameras and still be able to reconstruct the densely sampled LF.

2. EPIPOLAR-PLANE IMAGES

Epipolar-plane image was first introduced by Bolles *et al*. in [16]. In comparison with regular photo images, an EPI has a specific and distinct structure, see Fig. 1(b). Any captured point of the scene is revealed in one of the EPIs as a line whose slope relates to disparity and directly depends on the distance of the point from the capturing plane (depth). The intensity over the line is related with the intensity of emanated light from that scene point. Within the pinhole camera model assumption, the disparity is defined as $\Delta d = \frac{f}{z}\Delta t$, where f is the focal distance in pixel size, z is the depth of the point, and Δt is the distance between nearby camera positions (see [14] for more details). The corresponding line slope in the EPI is f/z.

The Lambertian reflectance model (any point in the scene emanates light in every direction with the same intensity) drives the distinct structure of EPI formed by lines with constant intensity distribution. Chai *et al*. presented a spectral analysis of the EPI slices of a LF depending on the scene depth and LF sampling rates in different dimensions [14]. It is interesting to point out that the spectrum of the EPI has a bow-tie type shape. Densely sampled LF guaranties that the spectrum of each EPI is always contained in a region similar to the one highlighted in Fig. 1(d). As shown in [14], the visual information of each depth slice is contained in a line passing through DC component in the frequency domain representation of the EPI. In order to obtain space of functions where EPI data will be presented sparsely, we need to provide an analysis tool for identification and separation of the lines in the frequency domain corresponding to different depth slices. While in spatial domain analysis atoms should be similar to lines with different slopes, their spectrum should have bow-tie type shape, as shown with different colors in Fig. 1(d).

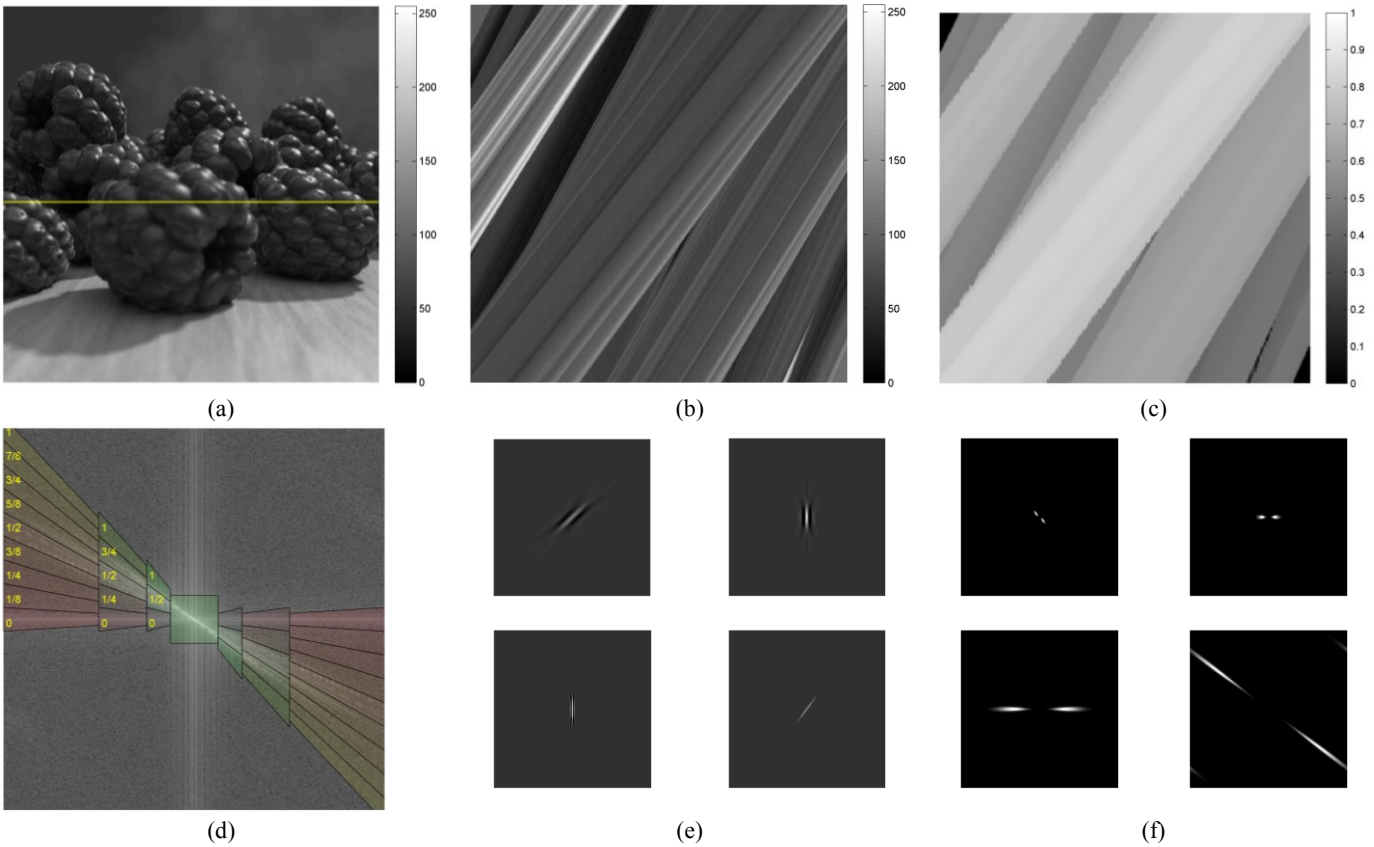


Figure 1. (a) Example of scene image.(b) Example of densly sampled light field EPI corresponding to row highlighted in yellow in (a). (c) EPI of disparity map. (d) Frequnecy domain characteristics of EPI with desirable frequency domain truncation, presented in 3 scales and central low pass filter with disparity values of corresponding shears. (e, f) Example of several constructed sheralet atoms in spatial and frequency domains.

3. SPARSE REPRESENTATION IN SHEARLET DOMAIN

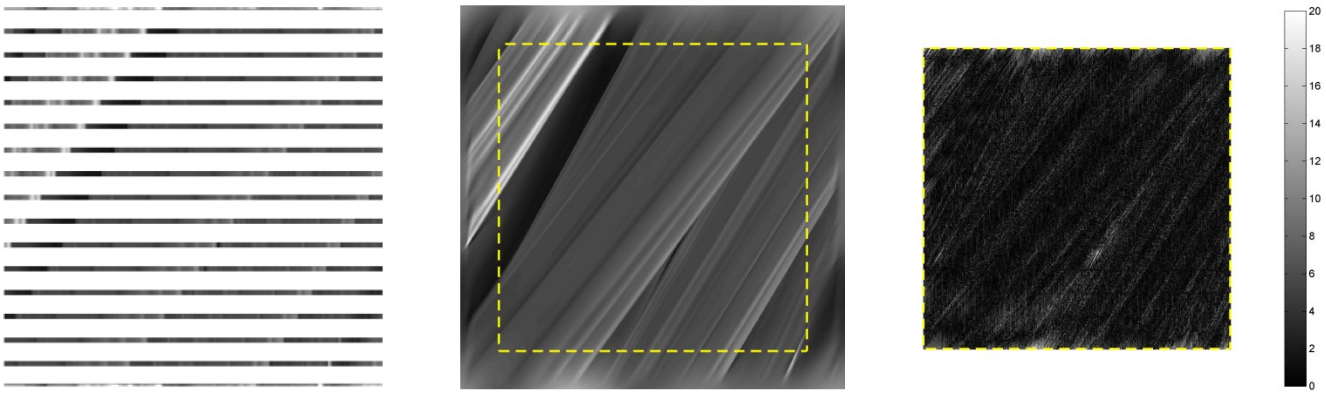
Shearlet frames, as developed in [17], [18], [19], are a perfect tool for the aforementioned sparse representation of the EPI. The elements of shearlet frames are translation-invariant functions whose spectrum covers a region similar to the one presented in Fig 1(f). Shearlet frame is described by number of scales and number of shears (directions) in each scale. An example is the Fast Finite Shearlet Transform (FFST) presented in [17]. FFST consists of a set of atoms that build a tight frame. Those atoms give almost perfect behavior in the frequency domain. However, in the spatial domain non-compact support of the atoms leads to ringing type artifacts. As a result, the approximation quality around the edges, where EPI does not comply with the band limited function condition, is drastically reduced. Another example of basis elements are the so-called compactly supported shearlets, as presented in [18]. Compactly supported shearlets are constructed in spatial domain using scaling and shearing operators. The compact support of the atoms was achieved by slightly changing the behavior in the frequency domain in comparison to atoms of the FFST.

In order to provide good directional properties at lower scales in frequency domain we propose to use different directional filters for different scales in the process of constructing a frame of compactly supported shearlet. Our construction follows the method proposed in [18], [19]. Fig. 1(e, f) presents examples of several constructed frame elements for different scales and shears.

4. RECONSTRUCTION ALGORITHM

We can interpret the set of captured views as given measurements of the unknown densely sampled EPI, as illustrated in Fig. 2(a). The problem tackled in this paper is to find (reconstruct) all missing data in the EPI. In order to simplify the notations, in this paper we assume rectangular size of EPI (in most case the horizontal resolution of the camera is higher than number of cameras, however, the corresponding EPI can be partially processed using overlapping rectangle windows with the size of the number of cameras).

Let $f \in \mathbb{R}^{N\times N}$ be the unknown complete EPI matrix, where each row represents corresponding image row and $g \in \mathbb{R}^{N\times N}$ be incomplete EPI where only rows with available views are presented, while everywhere else is 0. Further, f and g are used in their column-wise reshaped $\mathbb{R}^{N^2}$ vector version with keeping same notations for f and g. Let the mask matrix (measuring matrix) $H \in \mathbb{R}^{N^2\times N^2}$ be $H(i,i) = 1$ if $g(i) \neq 0$ and 0 otherwise. Analysis and synthesis matrix of the shearlet frame will be denoted as $S \in \mathbb{R}^{M\times N^2}$ and $S^* \in \mathbb{R}^{N^2\times M}$, respectively, where $M = \eta N^2$ and η is the number of all shears in all scales of the shearlet.



(a) (b) (c)

Figure 2. (a) Example of the input data for the proposed algorithm, where the original data was decimated by factor 32. (b) Reconstructed EPI, yellow square representing the region which was used for reconstruction quality estimation. (c) Absolute difference between the ground truth and reconstructed EPI.

Reconstruction of missing rows of g can be formulated as an inpainting problem, with prior condition to have sparse solution in the shearlet domain, i.e.

$$\min_{f\in\mathbb{R}^{N^2}} \|Sf\|_0, \text{ subject to } g = Hf \tag{1}$$

It was shown in [20] that the problem (1) can be efficiently solved through the following iterative thresholding algorithm

$$f_{n+1} = S^*\left(H_{\lambda_n}\left(S(f_n + \alpha(g - Hf_n))\right)\right) \tag{2}$$

where $H_\lambda(x) = \begin{cases} x, |x| \geq \lambda \\ 0, |x| < \lambda \end{cases}$, is a hard thresholding operator and α is a chosen relaxation parameter. The thresholding parameter λ_n decreases with the iteration number. Initial value of f_0 can be chosen as 0 everywhere. After sufficient iterations, f_n reaches a satisfying solution for the problem (1). More details can be found in [20], [21], [22], [23].

5. EXPERIMENTAL RESULTS

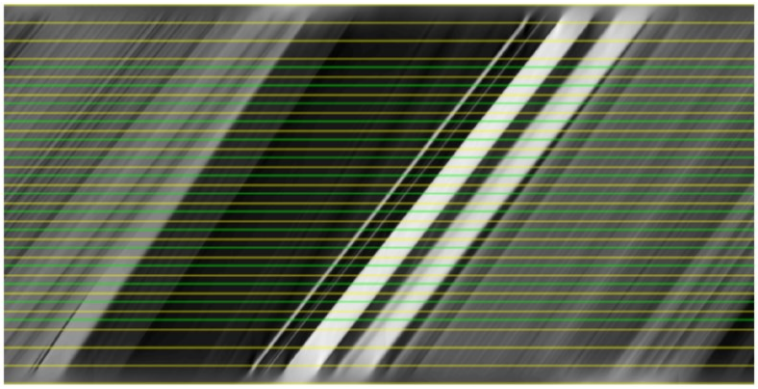
We will illustrate the proposed method on synthetic data as well as on a real-world dataset captured by cameras.

5.1. Synthetic Data

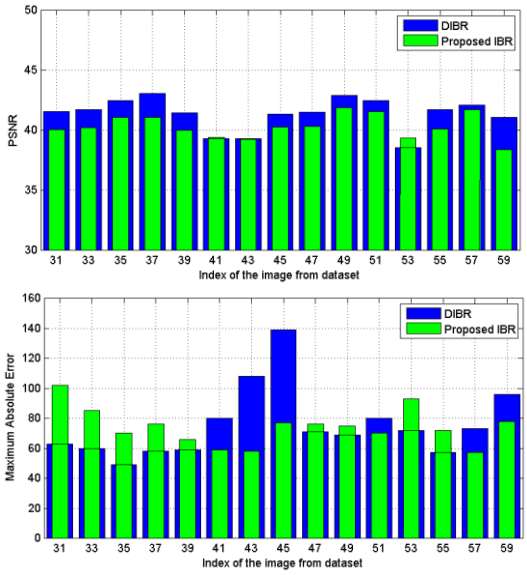
To construct synthetic data we used Blender (open source shareware, www.blender.org). It enables simulating a desired parallel positioned camera capturing system. Our generated synthetic data consists of 511 images with 511×511 resolution. Captured views provide horizontal parallax with disparity values in the range of $[0, 1]$ pixels between views. One of the EPIs generated from the rendered images is shown in Fig. 1(b) with the corresponding frequency domain characteristic in Fig 1(d) and the corresponding ground truth disparity map in Fig 1(c). As an input data for the reconstruction algorithm we use every 32nd view, thus 17 views. An example of the input data for the proposed algorithm is shown in Fig 2(a). In that case the input dataset consist of images with disparity values in the range $[0, 32]$ pixels between two consecutive images. Shearlet frame is constructed using 6 scales and a central low pass filter. In each scale from low to high we have $[2, 3, 5, 9, 17, 33]$ shears respectively. Each set of shears for fixed scale uniformly covers the $[0,1]$ range of disparities. Example of a similar separation (fewer scales) is illustrated in Fig 1(d). Shearlet is a translation-invariant frame thus its synthesis and analysis transforms are easy to implement using convolution operator. Convolution implemented through Fourier transform implicitly assumes circular replication of the signal. This increases the undesirable border effects and decreases the algorithm performance around image borders. In this paper, a Kaiser window is used to reduce these border effects. In Fig. 2(a) example of sparse EPI (input data) is presented, Fig. 2(b) shows the corresponding reconstructed result and Fig. 2(c) shows the residual calculated only over the region within the yellow rectangle. In the presented case, the mean-square-error (MSE) is 8 and the mean-absolute-error (MAE) is 25. Both are calculated with respect to the ground truth data. This example shows that by using the proposed method, a densely sampled LF can be reconstructed by using only a small number of captured views.

5.2. Real Data

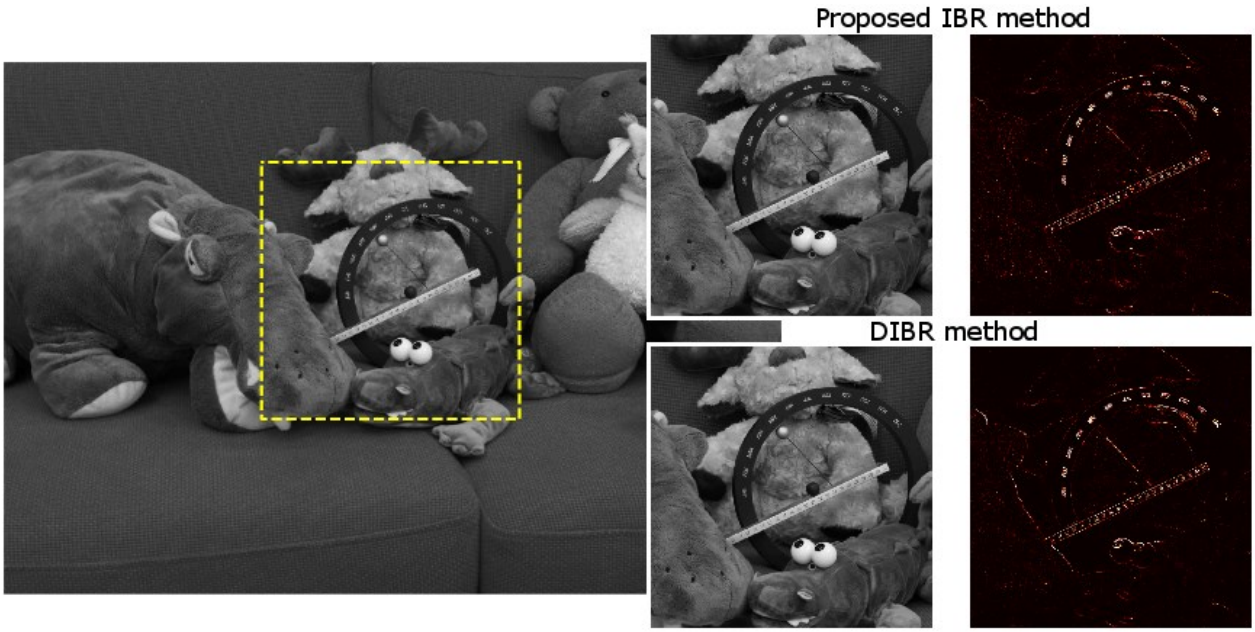
As a real captured dataset we use the "Couch" dataset used in [3]. It consists of 101 images with 2679x4020 resolution as well as 51 estimated disparity maps for the central views obtained by the algorithm proposed in [3] using the whole set of images. Given disparity estimation shows that maximal disparity between consecutive images is about 11px. We applied the presented algorithm to the grayscale images. 15 views were reconstructed using the odd number indexed views from the dataset. An example of a reconstructed EPI is presented in Fig. 3(a), where input (selected) rows for the reconstruction algorithm are highlighted in yellow and rows in green represent views used for assessing the algorithm performance. Same input data is used for depth image based rendering algorithm based on 3D warping and blending implemented as described in [5]. The reconstruction quality of the two algorithms is presented in Fig. 3(b, c). As seen in the figure, both approaches result in reconstructed images with good PSNR with respect to reference captured images whereas the proposed algorithm has in average a lower maximum absolute error.



(a)



(b)



(c)

Figure 3. (a) Reconstructed EPI for the real dataset. Rows which are highlighted with yellow color represent odd indexed views from original data set which were used as an input data for the algorithm and green color represents even indexed views from original dataset which are used for algorithm quality estimation. (b) Evaluation of intermediate views reconstruction in PSNR(top) and MAE (bottom) for regular depth image based rendering (DIBR) algorithm and prosposed algorithm. (c) Ground truth image from dataset(left), reconstruction results for highlighted part of the image and absolute differences between the ground truth and reconstructed images for DIBR (bottom) and proposed algorithm (top).

6. CONCLUSION

In this paper we presented a method for reconstructing densely sampled LF from a given sparse set of views by processing the corresponding EPI images in shearlet domain. We have shown, by using synthetic and real data examples, that the proposed method is very effective in reconstructing densely sampled LFs out of small number of given views. The strength of the proposed method lies in its ability to reconstruct the complete dataset (whole LF) at ones in comparison with classical IBR techniques where each view has to be reconstructed individually. The proposed method establishes a new approach for LF interpolation.

7. REFERENCES

[1] A. Gelman, P. L. Dragotti and V. Velisavljevic, "Multiview image compression using a layer-based representation," in *17th IEEE International Conference on Image Processing (ICIP),* 2010, pp. 493-496.

[2] J. Berent, P. L. Dragotti and M. Brookes, "Adaptive layer extraction for image based rendering," in *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on,* 2009, pp. 1-6.

[3] 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.,* vol. 32, pp. 73, 2013.

[4] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in *Electronic Imaging 2004,* 2004, pp. 93-104.

[5] M. Li, H. Chen, R. Li and X. Chang, "An improved virtual view rendering method based on depth image," in *Communication Technology (ICCT), 2011 IEEE 13th International Conference on,* 2011, pp. 381-384.

[6] 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.

[7] S. J. Gortler, R. Grzeszczuk, R. Szeliski and M. F. Cohen, "The lumigraph," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, 1996, pp. 43-54.

[8] Z. Lin and H. Shum, "A geometric analysis of light field rendering," *International Journal of Computer Vision,* vol. 58, pp. 121-138, 2004.

[9] R. Ng, "Fourier slice photography," in *ACM Transactions on Graphics (TOG)*, 2005, pp. 735-744.

[10] I. Tosic and K. Berkner, "Light field scale-depth space transform for dense depth estimation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014, pp. 441-448.

[11] 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, pp. 606-619, 2014.

[12] M. Tanimoto, "Overview of free viewpoint television," *Signal Process Image Commun,* vol. 21, pp. 454-461, 2006.

[13] J. Jurik, T. Burnett, M. Klug and P. Debevec, "Geometry-corrected light field rendering for creating a holographic stereogram," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* 2012, pp. 9-13.

[14] J. Chai, X. Tong, S. Chan and H. Shum, "Plenoptic sampling," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques,* 2000, pp. 307-318.

[15] S. Hauser and J. Ma, *Seismic Data Reconstruction Via Shearlet-Regularized Directional Inpainting*, http://www.mathematik.uni-kl.de/uploads/tx_sibibtex/seismic.pdf, accessed 15 May 2012.

[16] 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, pp. 7-55, 1987.

[17] S. Häuser, "Fast finite shearlet transform," *ArXiv Preprint arXiv:1202.1773,* 2012.

[18] P. Kittipoom, G. Kutyniok and W. Lim, "Construction of compactly supported shearlet frames," *Constructive Approximation,* vol. 35, pp. 21-72, 2012.

[19] G. Kutyniok, W. Lim and R. Reisenhofer, "Shearlab 3D: Faithful digital shearlet transforms based on compactly supported shearlets," *ArXiv Preprint arXiv:1402.5670,* 2014.

[20] M. Elad, J. Starck, P. Querre and D. L. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *Applied and Computational Harmonic Analysis,* vol. 19, pp. 340-358, 2005.

[21] J. Starck and J. M. Fadili, "An overview of inverse problem regularization using sparsity," in *Ieee Icip,* 2009, pp. 1453-1456.

[22] M. Fadili, J. Starck and F. Murtagh, "Inpainting and zooming using sparse representations," *The Computer Journal,* vol. 52, pp. 64-79, 2009.

[23] M. Kowalski, "Thresholding rules and iterative shrinkage/ thresholding algorithm: A convergence study," in *International Conference on Image Processing (ICIP) 2014,* 2014.

PUBLICATION

II

Shearlet-Domain Light Field Reconstruction for Holographic Stereogram Generation

E. Sahin, S. Vagharshakyan, J. Mäkinen, R. Bregovic and A. Gotchev

*2016 IEEE International Conference on Image Processing (ICIP)*2016, 1479–1483

SHEARLET-DOMAIN LIGHT FIELD RECONSTRUCTION FOR HOLOGRAPHIC STEREOGRAM GENERATION

Erdem Sahin, Suren Vagharshakyan, Jani Mäkinen, Robert Bregovic, Atanas Gotchev

Department of Signal Processing, Tampere University of Technology, Tampere, Finland

ABSTRACT

Holographic stereograms (HSs) constitute one of the most widely used types of computer-generated holograms. The scene information required to calculate the HSs can be acquired by conventional digital cameras. It is, however, usually required that the scene should be captured from dense set of view points. Therefore, relieving this requirement is critical in the sense of easing the capture process. In this paper, in the capture stage of holographic stereograms, we employ our previously presented light field reconstruction algorithm [1], where we utilize sparse representation of light fields in the shearlet domain and reconstruct dense light fields from their highly under-sampled versions. The simulation results demonstrate that we can relieve the dense view sampling requirement of HSs, e.g. by as high as 8×8 sub-sampling factor, and still keep the perceived image quality of holographic reconstructions at satisfactory levels. This enables, for example, replacing the scanning camera setups with the more convenient multi-camera arrangements.

Index Terms— Holographic stereogram, light field, sparse reconstruction, shearlet

1. INTRODUCTION

An end-to-end holographic capture and display is usually regarded as the ultimate way of 3D scene replication. As it is well-known, however, holographic capture relies on the wave interference principle which requires illumination of the scene with a coherent light source. This in turn makes the recording of real-life scenes difficult for various practical reasons. The endeavor towards solving this problem goes back to 1966 when the first computer-generated hologram was proposed by Brown *et al.* [2]. Computer-generated holography (CGH), in a sense, simulates the optical recording process carried out in holographic capture. In other words, the interference of reference and object waves is calculated numerically. Therefore, it enables obtaining holographic information from a synthetically generated scene or a real-life scene which is illuminated by (incoherent) white light.

Unlike other model-based approaches (which require scene depth information), such as Fresnel hologram [3] and phase-added stereogram [4], holographic stereogram (HS) [5] is an image-based CGH technique which relies only on a set of captured images of the scene. Due to this relieved scene capture requirement as well as their efficient way of calculation, HSs have found various applications, especially, with the recently developed holographic print techniques [6]. The set of captured images required for HSs are usually described using the light field (LF) formalism [7]. The sampling requirements of the LF, which is to be used in HS calculation, is determined based on the human visual system (HVS) as well as the properties of the scene. These requirements mostly result in densely sampled LFs which then set challenging constraints on the practical capture setups. As a consequence of these constraints, the scene is usually captured by a scanning camera which should be moved by an accurate camera positioning system with step sizes in millimeter or sub-millimeter ranges [8]. Relieving the sampling requirements of LF is crucial for HSs, not just to ease the tedious work to be accomplished by such capture systems but also to enable new capture setups. For example, relaxing the camera movement step size to centimeter ranges could make a multi-camera setup useful for capturing, which then removes the static scene assumption of the scanning camera systems and enables dynamic scene capture.

Cao *et al.* [9] and Rivenson *et al.* [10] have used the recently emerged compressive sensing techniques to reconstruct 3D scenes from sub-sampled versions of Fresnel holograms and HSs, respectively. In these approaches, the relation between the scene and the hologram plane is modeled via superposition of 2D-to-2D propagations defined between a set of parallel planes representing the scene and the hologram plane. Thus, independent treatment of different sections of the scene leaves the applicability of such methods questionable for scenes with occlusions. As the HSs actually rely on the captured LF, we aim to solve the sparse capturing problem within the context of LF reconstruction from its sub-sampled versions. By this way, we surpass the above-mentioned modeling problem, which then makes our approach applicable to scenes with occlusions. The novel view synthesis approaches are directly applicable to the problem considered in this paper. Indeed, in [11, 12, 13], depth-image based rendering techniques have been used to reduce the number of captured images required for CGH. The performance of such approaches is dictated by the quality of the depth estimation which is

very much scene dependent. Our LF reconstruction algorithm, which has been previously presented in [1], relies on an image based rendering technique. Thus, it does not require explicit depth information. Furthermore, it is particularly useful for the problem we consider here with its competence in LF reconstruction, i.e. its ability of acceptable quality dense LF reconstructions from highly sub-sampled LFs. The sub-sampling rates we consider in this paper are significantly higher (which results in sparser set of cameras), for example, than those reported in the recently presented LF reconstruction technique [14].

2. HOLOGRAHIC STEREOGRAMS

A holographic stereogram contains the sampled LF information on a certain plane (hologram plane). The LF required to calculate HS can be described using the hologram plane and the capture plane which defines camera locations. Thus, the required LF information is obtained from cameras capturing perspective views as illustrated in Fig. 1. The light rays crossing these two planes can be equivalently parametrized using the hologram plane and two angular coordinates which correspond to positions on the camera plane. Please note that in Fig. 1 and in the following derivations we consider the 2D cross-section of the 3D space for simplicity, the extension to 3D case is straightforward.

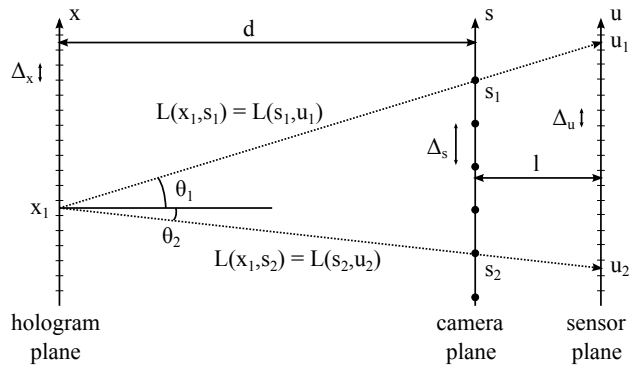


Fig. 1. Parametrization of light field capture for holographic stereograms.

Let us denote the hologram, capture and camera sensor planes by x, s and u, respectively. We define the LF parametrized by the x and s planes as $L_1(x, s)$, similarly another LF is defined using the s and u planes as $L_2(s, u)$. Denoting the distance between the camera and hologram planes as d, and the distance between the sensor and the center of projection of the camera as l, the relation between L_1 and L_2 is given by

$$L_1(x, s) = L_2(s, u_x), \tag{1}$$

where $u_x = s + l(s - x)/d$. The hologram, capture and sensor planes are discretized by the sampling steps Δ_x, Δ_s, Δ_u, which represent the holographic element (hogel) size, distance between adjacent cameras and the pixel size of the camera sensor, respectively. If the magnification equation given by $\Delta_x = \Delta_u d/l$ is satisfied and Δ_s is chosen to provide integer pixels of disparity D for adjacent views, i.e. $D = \Delta_s l/d \in \mathbb{Z}$, for those points on the hologram plane, then there is a one-to-one correspondence between the discrete LFs $L_1[m, i]$ and $L_2[i, k]$. Thus, the LF $L_1[m, i]$ can be directly read from the captured images.

HSs encode the LF information in the form of holographic fringes. The object field can be expressed as a superposition of windowed plane waves emitted from different hogels to different directions, and the amplitudes of the plane waves are specified by the corresponding LF (intensity) samples. Thus,

$$O_{HS}(x) = \sum_m \mathrm{rect}\left(\frac{x - m\Delta_x}{\Delta_x}\right) \times \sum_i \sqrt{L_1[m, i]} \exp(j2\pi f_x^{mi} x), \tag{2}$$

where f_x^{mi} are the spatial frequencies of the plane waves on the hologram plane. They are determined by the propagation directions of the corresponding rays as $f_x^{mi} = (1/\lambda) \sin \theta_x^{mi}$, where λ is the wavelength of the monochromatic light and θ_x^{mi} represent the incidence angles of the rays along the $x-$axis. The inner sum in Eq. 2 produces the spatial pattern to be written inside each hogel and it can be found via applying an inverse Fourier transform operation to the reordered images according to Eq. 1. The interference pattern of a reference wave and the object wave found by Eq. 2 creates the intensity fringe patterns of the HS which then reconstruct the object wave when illuminated with the same reference wave. In this paper, we consider the complex object wave as the HS to avoid the reconstruction noise that would be introduced by the conjugate object wave so as to evaluate our approach in a more reliable way, as will be demonstrated in Sec. 4.

The diffractive properties of HSs are mainly determined by the hogel size. The hogel size of the HS is usually chosen based on the properties of HVS and an average intended observation distance. Assuming that HVS is a diffraction-limited imaging system, the minimum distance between two points at distance d that can be resolved by the HVS is given by the Rayleigh criterion as [15]

$$\Delta_x^{HVS} = \frac{1.22\lambda d}{T}, \tag{3}$$

where T is the pupil size of the human eye (which is typically in $2mm$-$8mm$). Thus, having a hogel size smaller than or equal to this minimum resolvable distance will ensure maximized perceived image resolution. On the other hand the pupil size sets an upper limit for the angular resolution of HS which needs to be satisfied for smooth experience of view-dependent image properties such as motion parallax [16]. These two criteria usually impose a dense LF sampling

for the capture setup which complicates the capture process. In the following section, we propose a method to relieve the sampling requirements regarding the capture setup.

3. LIGHT FIELD RECONSTRUCTION

One widely used way of LF analysis is to utilize the epipolar-plane image (EPI) representation. An epipolar-plane image can be formed by taking the slices of the LF, i.e. for a 4D LF $L(s,t;u,v)$ it is obtained as $E(s,u)$ or $E(t,v)$, depending on the direction for which the analysis will be carried out. The problem of densely sampled LF reconstruction can be formulated as reconstruction of each densely sampled EPI slice from only a sparse (decimated) set of samples as illustrated in Fig. 2(a) and Fig. 2(b). Here, we refer to the LF sampling case as the densely sampled LF, when the disparity range of the scene between adjacent views is within $-1:1$ pixels with respect to the disparity of points on the focused scene plane. It has been shown that the continuous LF function can be obtained from such a densely sampled LF using linear interpolation [17].

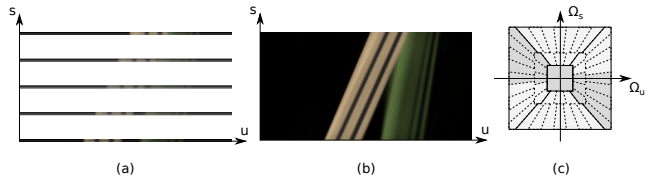


Fig. 2. Reconstruction of the densely sampled EPI. (a) Decimated EPI. (b) Corresponding reconstruction result. (c) Tiling of the frequency domain of EPI by the shearlet atoms.

The LF reconstruction problem can be efficiently solved using regularization in the shearlet domain, since the LFs exhibit sparse representations in this domain [1]. Fig. 2(c) illustrates how the frequency domain of EPI is tiled by the shearlet atoms. The shearlet atoms are distributed such that each disparity value in EPI is revealed as direction in tiling. The reconstruction of unknown samples in EPI can be modeled as estimation of $\boldsymbol{a}$ given that $\boldsymbol{b} = \boldsymbol{H}\boldsymbol{a}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the vectorized versions of the densely sampled and decimated EPI, respectively, $\boldsymbol{H}$ is the masking matrix representing the known samples positions. Reconstruction is obtained by iterative hard thresholding procedure with decreasing threshold [1]. That is,

$$\boldsymbol{a}_{n+1} = \boldsymbol{S}^* \left\{ \boldsymbol{T}_{\lambda_n} \left\{ \boldsymbol{S} \left[\boldsymbol{a}_n + \alpha(\boldsymbol{b} - \boldsymbol{H}\boldsymbol{a}_n) \right] \right\} \right\},$$

where $\boldsymbol{S}, \boldsymbol{S}^*$ are the shearlet analysis and synthesis transform matrices, respectively, α is acceleration coefficient and $\boldsymbol{T}_{\lambda_n}$ is the hard thresholding operator with threshold λ_n. After sufficient number of iteration, $\boldsymbol{a}_n$ is obtained as the solution with the corresponding sparse representation of $\boldsymbol{S}\boldsymbol{a}_n$.

In this paper, we consider full parallax viewing of HSs. Therefore, the full parallax LF reconstruction is achieved by consecutively reconstructing each horizontal parallax set and then repeating the same procedure for each vertical parallax set. For a more detailed discussion of the LF reconstruction algorithm, we refer the reader to [1].

4. SIMULATION RESULTS

We consider the simulation setup illustrated in Fig. 3, where we use the 3D modeling software Blender [18] for designing the scene and rendering the perspective images. The holographic stereogram is intended to be viewed by an observer at distance $d = 200mm$. The pupil size of the observer is assumed to be $2mm$. The hologram parameters are then set (according to Sec. 2) as: pixel size $X_x = 2\mu m$, hogel size $\Delta_x = 64\mu m$ and total number of pixels $N = 8092$.

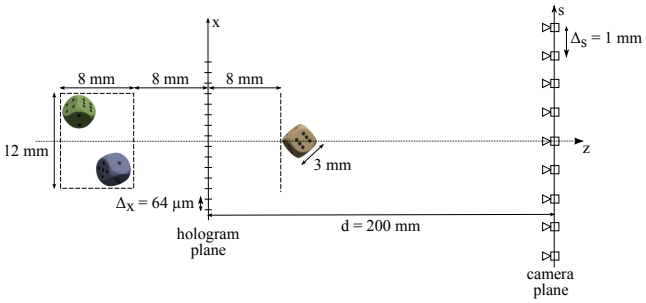


Fig. 3. The simulated scene.

The camera plane is also chosen to be $d = 200mm$ away from the hologram plane. We first implement the dense LF sampling case for which the LF samples required for HS calculation can be simply obtained from the captured LF data by linear interpolation, as pointed out in Sec. 3. For the scene shown in Fig. 3, it is required that the camera spacing should be at most $1mm$ to be able to capture the dense LF. Therefore, we put 49×49 cameras at $s,t \in \{-24\Delta_s, -23\Delta_s, ..., 24\Delta_s\}$ with spacing of $\Delta_s = 1mm$ on the camera plane. In the second setup, we employ the sparse set of 7×7 cameras at $s,t \in \{-24\Delta_s, -16\Delta_s, ..., 24\Delta_s\}$ to demonstrate how our LF reconstruction algorithm relieves the LF capture stage of HSs. Given this sparse set of cameras we estimate all 49×49 dense set of views using two different approaches: one is our shearlet-based LF reconstruction method discussed in Sec. 3 and the other one is depth-based light field reconstruction which utilizes ground truth depth maps (provided by Blender) for each sparse viewpoints. In the depth-based approach, following the procedure in [13], we separate the depth in three dominant layers (corresponding to depths of three dice) that are consequently used in the rendering of the dense light field. Then, we calculate three holographic stereograms by using both the original and the estimated dense sets of views for the above-mentioned two approaches. In all HS calculations,

the scene is assumed to be illuminated by a monochromatic light with wavelength of $534nm$ and correspondingly only the green channels of the images are utilized. The HSs are calculated hogel by hogel using IFFT, as suggested by Eq. 2, where the discrete set of spatial frequencies are obtained by resampling the continuous set of spatial frequencies via bi-linear interpolation.

We compare the qualities of obtained HSs by simulating the viewing process. In particular, we employ wave field modeling and find the image perceived by the observer using the Fresnel diffraction model as

$$I(u,v) = |\mathcal{F}_l \{T(s,t)\mathcal{F}_d \{O_{HS}(x,y)\}\}|^2, \qquad (4)$$

where $\mathcal{F}_z$ is the Fresnel propagation operation by distance z, and $T(s,t)$ represents the lens transfer function of the human eye [15]. The human eye is modeled as a camera with a thin lens having a circular aperture of diameter $2mm$. The distance between the pupil and retina, l, is assumed to be $25mm$ and the eye is focused on the hologram plane. The pixel size on the retina is set to $\Delta_x l/d = 8\mu m$, i.e. it corresponds to one hogel size according to lens magnification. In order to reduce the speckle noise in the reconstructed images, we do the calculation given by Eq. 4 in a multiplexed manner. The HSs are expressed as a sum of their sub-sampled versions (sub-sampling is done over hogels), where the sub-sampling factor is 6×6. The corresponding 36 reconstructed images are then (incoherently) summed up to find the final reconstructed image. By this way, we eliminate the interference between a given hogel and the hogels within its 11×11 neighborhood.

In Fig. 4, we show reconstructed images from three HSs, corresponding to different LFs, at different observer positions on the camera plane. The HS reconstructions corresponding to original dense set of images, estimated sets of images via the depth-based reconstruction, and estimated sets of images via our LF reconstruction algorithm are given in columns (b), (c), and (d), respectively. At each observation point shown in different rows, we take the green channels of *ideal* images rendered in Blender as reference (shown in column (a)) and calculate PSNRs for the image regions within the bounding boxes of dice. From column (b) to column (d), the corresponding PSNRs are found as $22.83dB$, $22.63dB$, $22.55dB$ for the top row; $22.90dB$, $22.89dB$, $22.87dB$ for the middle row; and $23.13dB$, $22.37dB$, $23.12dB$ for the bottom row.

Regarding the comparison between the reconstructed and corresponding ideal images, the significant parts of degradations are caused by the common factors that are the errors introduced in HS generation (e.g. due to plane wave assumption) and the speckle noise inherent to subsequent holographic reconstruction process. Otherwise we can provide acceptable quality reconstructed (perceived) images which are similar to those obtained from the original dense set of views, and in doing this we relieve the view sampling requirement by a factor of 8×8 compared to the dense LF capture case. Furthermore, our LF reconstruction and the depth-based method also result in similar reconstructed images from HSs. This makes our method more preferable, since it does not require depth estimation which is usually susceptible to artifacts such as misregistration. There are significant implications of these achievements. For example, it would be possible to capture the required LF information for the setup considered in this section by using a 7×7 array of cameras which is placed at $800mm$ away from the hologram plane, where each camera has $50mm$ focal length and $4K \times 4K$ sensor with pixel size of $\sim 4.3\mu m$. The distance between the adjacent cameras of this multi-camera setup would be $8mm \times (800/200) = 32mm$ which is a feasible value for camera spacing (it should have been $1mm \times (800/200) = 4mm$ for direct capture of dense LF).

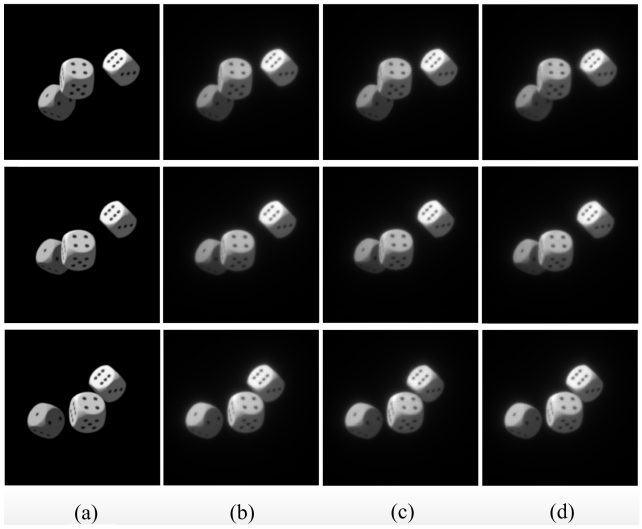


Fig. 4. HS reconstructions for views given in (a). HS is calculated from: (b) original dense set of images, (c) estimated dense set of images via depth-based approach (ground truth depth is utilized), (d) estimated dense set of images via our light field reconstruction algorithm. The observer is at: top row: $(6\Delta_s, 12\Delta_s)$, middle row: $(-12\Delta_s, 8\Delta_s)$, bottom row: $(-10\Delta_s, -10\Delta_s)$.

5. CONCLUSIONS

We have presented a way to relieve the light field sampling required by the holographic stereograms. In particular, we have utilized the sparse representation of light fields in the shearlet domain. The ability of acceptable quality dense light field reconstruction from its highly under-sampled versions (e.g. by a factor of 8×8) have led us to capture a much sparser sets of multi-view images than that is originally required for holographic stereogram calculation. As an important implication of this, we have demonstrated that the usually employed scanning camera setups can be replaced with the more convenient multi-camera arrangements.

6. REFERENCES

[1] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Image based rendering technique via sparse representation in shearlet domain," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1379–1383.

[2] B. R. Brown and A. W. Lohmann, "Complex spatial filtering with binary masks," *Appl. Opt.*, vol. 5, no. 6, pp. 967–969, 1966.

[3] B. P. Waters, "Holographic image synthesis utilizing theoretical method," *Appl. Phys. Lett.*, vol. 9, pp. 405–407, 1966.

[4] M. Yamaguchi, H. Hoshino, T. Honda, and N. Ohyama, "Phase-added stereogram: calculation of hologram using computer graphics technique," *Proc. SPIE*, vol. 1914, pp. 25–31, 1993.

[5] T. Yatagai, "Stereoscopic approach to 3-d display using computer-generated holograms," *Appl. Opt.*, vol. 15, no. 11, pp. 2722–2729, 1976.

[6] H. Kang, E. Stoykova, J. Park, S. Hong, and Y. Kim, "Holographic printing of white-light viewable holograms and stereograms," in *Holography - Basic Principles and Contemporary Applications*, E. Mihaylova, Ed. Intech, 2013.

[7] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. 1996, SIGGRAPH '96, pp. 31–42, ACM.

[8] S. J. Hart, "Methods and apparatus for making holograms," 1998, US Patent 5,796,500.

[9] X. Cao, X. Sang, Z. Chen, Y. Zhang, J. Leng, N. Guo, B. Yan, J. Yuan, K. Wang, and C. Yu, "Fresnel hologram reconstruction of complex three-dimensional object based on compressive sensing," *Chin. Opt. Lett.*, vol. 12, no. 8, pp. 080901, 2014.

[10] Y. Rivenson, A. Stern, and J. Rosen, "Compressive multiple view projection incoherent holography," *Opt. Express*, vol. 19, no. 7, pp. 6109–6118, 2011.

[11] B. Katz, N. T. Shaked, and J. Rosen, "Synthesizing computer generated holograms with reduced number of perspective projections," *Opt. Express*, vol. 15, no. 20, pp. 13250–13255, 2007.

[12] Y. Ohsawa, K. Yamaguchi, T. Ichikawa, and Y. Sakamoto, "Computer-generated holograms using multiview images captured by a small number of sparsely arranged cameras," *Appl. Opt.*, vol. 52, no. 1, pp. A167–A176, 2013.

[13] J. Jurik, T. Burnett, M. Klug, and P. E. Debevec, "Geometry-corrected light field rendering for creating a holographic stereogram," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, June 16-21, 2012*, 2012, pp. 9–13.

[14] X. Cao, Z. Geng, and T. Li, "Dictionary-based light field acquisition using sparse camera array," *Opt. Express*, vol. 22, no. 20, pp. 24081–24095, 2014.

[15] J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, 2nd edition, 1996.

[16] M. Lucente, *Diffraction-Specific Fringe Computation for Electro-Holography*, Ph.D. thesis, Massachusetts Institute of Technology, USA, 1994.

[17] Z. Lin and H.-Y. Shum, "A geometric analysis of light field rendering," *Int'l J. of Computer Vision*, vol. 58, no. 2, pp. 121–138, 2004.

[18] "Blender," http://www.blender.org/about/.

PUBLICATION III

Accelerated Shearlet-Domain Light Field Reconstruction

S. Vagharshakyan, R. Bregovic and A. Gotchev

IEEE Journal of Selected Topics in Signal Processing 11.7 (2017), 1082–1091

Accelerated Shearlet-Domain Light Field Reconstruction

Suren Vagharshakyan, *Member, IEEE,* Robert Bregovic, *Member, IEEE,* and Atanas Gotchev, *Member, IEEE*

***Abstract*—We consider the problem of reconstructing densely sampled light field (DSLF) from sparse camera views. In our previous work, the DSLF has been reconstructed by processing epipolar-plane images (EPI) employing sparse regularization in shearlet transform domain. With the aim to avoid redundant processing and reduce the overall reconstruction time, in this article we propose algorithm modifications in three directions. First, we modify the basic algorithm by offering a faster and more stable iterative procedure. Second, we elaborate on the proper use of color redundancy by studying the effect of reconstruction of an average intensity channel and its use as a guiding mode for colorizing the three color channels. Third, we explore similarities between EPIs by their grouping and joint processing or by effective decorrelation to get an initial estimate for the basic iterative procedure. We are specifically interested in GPU-based computations allowing an efficient implementation of the shearlet transform. We quantify our three main approaches to accelerated processing over a wide collection of horizontal- as well as full-parallax datasets.**

***Index Terms*—Light field reconstruction, Graphics processing units, Densely sampled light field**

I. Introduction

3D visual scenes are completely represented by the light field they emanate. Given that the light field is a continuous function, its capture and consequent reconstruction is an important task, especially for visualization applications, which require multiple perspective views (e.g. super multiview displays [1]) or dense parallax (e.g. digitally printed holograms [2]). Many other light field image processing applications, such as depth estimation, compression, synthetic aperture imaging would benefit from accurately reconstructed light field [3]. A typical way of capturing light fields from real world scenes is to use a set of identical parallel cameras which are uniformly positioned on a plane. In order to support continuous parallax, such capturing setup requires that the cameras are densely positioned [4]. To overcome the demand for synchronously controlled high amount of cameras, the approach is to use a coarse set of cameras and to devise a consecutive light field reconstruction method, which can deliver densely sampled views from the coarse set of captured ones.

The approaches for reconstructing dense intermediate views from a given sparse set of views can be categorized into two categories. First, those are methods aimed at extracting geometry information about the scene in the of form of high quality depth maps collocated with the given input images, which can be used for depth-based view rendering [5] or unstructured lumigraph rendering [6]. Such methods utilize correspondences between images, found by block matching [7], and employ some global optimization of cost functions usually formed by data and smoothness terms [8], [9]. Apart from having problems with occlusions, when using sparse views, these methods result in over-smoothed depth estimates, and for finding finest details aligned with object boundaries they still need relatively densely positioned cameras [10]. Second category includes methods operating directly on the light field and aimed at employing some sparsity priors for this data. For example, the work [11] exploits sparse representation of full parallax 4D light field in continuous Fourier domain using a small number of 1D viewpoint trajectories.

For both categories, reconstructing a dense set of images is a computationally demanding problem. Global optimization methods aimed at obtaining multiple high quality depth maps do not scale well with the number of images and their resolution. The method in [10] targeted processing of 50 views with overall of 21-megapixels and accounted of about 50 mins of processing time. The runtime of the method in [11] ranged from 2 to 3 hours using a cluster of machines.

Previously, we have proposed a method for light field reconstruction which utilizes sparsification in shearlet transform domain [12], [13]. The method reconstructs DSLF from an undersampled light field captured by a small number of wide-baseline cameras. It demonstrated superior performance while compared with Motion Picture Experts Group's depth estimation reference software (DERS) [14] and view synthesis reference software (VSRS) [15], and with the state of the art in depth-from-stereo scene geometry reconstruction [16]. The method handles both horizontal and full-parallax capture settings and is highly successful when reconstructing non-Lambertian scenes formed by semi-transparent objects [13]. In this article, we further develop the method by proposing computational acceleration approaches based on inherent similarities in the assumed data representation and further algorithm tuning. Our aim is to decrease the necessary computational time while keeping or even increasing the reconstruction quality for a large set of test data.

The article is structured as follows: the light field parameterization and a summary of light field reconstruction algorithm from [13] are presented in Section II. Different acceleration approaches are proposed in Section III. Computing and evaluation setup, algorithm implementation, experimental results and discussions are presented in Section IV.

II. Reconstruction of Densely Sampled Light Field

4D light field is parameterized by the so-called two-plane parameterization $L(u, v, s, t)$ (Fig. 1), where (s, t) and (u, v) cor-

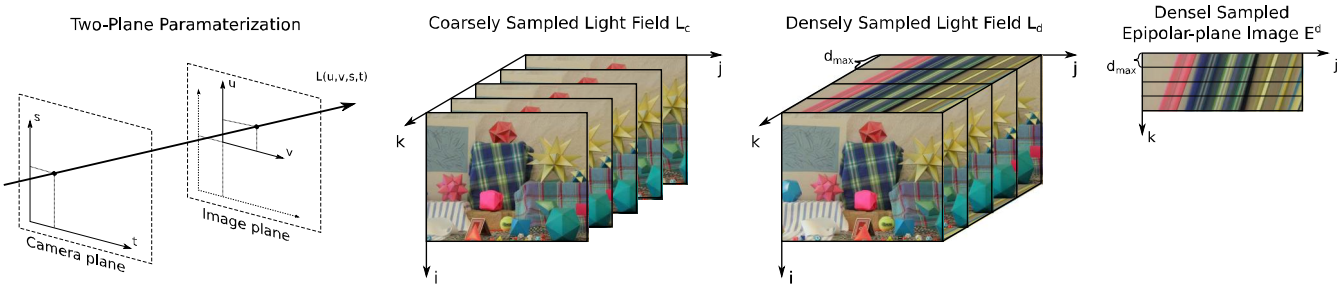


Fig. 1. Light field two-plane parameterization and corresponding discrete coarsely and densely sampled light field parameterizations.

respond to the camera plane and the image plane respectively [17]. This parameterization allows to conveniently describe and denote both the set of images (views) captured by a multi-camera setup and the required dense set of images fully representing the 3D scene of interest. For the sake of simpler notations and easier illustrations, hereafter we consider the case of horizontal parallax and explain the generalisations to full parallax when needed. Fixing the camera motion to horizontal direction only implies that the parameter $s = s_0$ corresponding to the vertical parallax can be omitted.

$$L(u, v, t) = L(u, v, s_0, t).$$

We aim at reconstructing continuous light field $L(u, s, t)$ from sampled views of 3D scenes. For such scenes and corresponding light fields, we define the densely sampled light field (DSLF), denoted by $L^d(i, j, k)$, as the light field having a maximal disparity between adjacent views less than 1 pixel. The k-th captured image $I^d_k(i, j) = L^d(i, j, k)$ corresponds to an image of the L continuous light field sampled at $t_k = k\Delta t_d$. The necessary parameter Δt_d for capturing DSLF can be calculated based on camera intrinsic parameters and specifying the minimum depth of the scene from the camera (capturing) plane. The continuous light field can be reconstructed from DSLF by linear interpolation [4]. Therefore, DSLF is a convenient representation and is in the core of many LF-based algorithms, such as refocusing, free view-point rendering and smooth-parallax visualization [3]. However, a direct capture of views providing one-pixel disparity is impractical.

In our previous work we have presented a method for DSLF reconstruction from coarsely sampled views [12]. A coarsely sampled light field is assumed to be a decimated version of DSLF, where the decimation factor is denoted by d_{max}

$$L^c(i, j, k) = L^d(i, j, d_{max}k).$$

Subsequently, the maximal disparity between adjacent coarsely sampled views $I^c_k(i, j) = L^c(i, j, k)$ is no more than d_{max}. In our previous work we presented an iterative algorithm which reconstructs L^d from L^c for $d_{max} \leq 32$. The algorithm works in EPI domain. More specifically, DSLF L^d is reconstructed by reconstructing every densely sampled epipolar-plane image (DSEPI) defined as

$$E^d_i(k, j) = L^d(i, j, k)$$

from the given decimated samples E^c_i such that $E^c_i(k, j) = E^d_i(d_{max}k, j)$. An example of coarsely-sampled and densely-sampled light fields is given in Fig. 1. Note how rows with step size d_{max} form E^c out of E^d. The specific value of $d_{max} \leq 32$ is related with image resolution and has been selected for practical reasons. The method can be applied for higher disparity ranges too, however this would impose processing images with higher resolution, which in turn would significantly increase the required amount of memory [13]. Therefore, in this article we consider the limit case of $d_{max} \leq 32$.

Below, we summarize the algorithm for DSEPI reconstruction [13]. To simplify the notations, we denote the unknown DSEPI matrix by $f \in \mathbb{R}^{N\times N}$. The decimated EPI $g \in \mathbb{R}^{N\times N}$ has the same dimension and contains sensed values at each d_{max}-th row while the other rows are set to 0. The relation between the two EPIs is formalized by setting a binary measurement matrix $M \in \mathbb{R}^{N\times N}$ which has zero values elsewhere than $M(kd_{max}, j) = 1$. Then, $g = M \odot f$, where $\odot$ is element-wise matrix multiplication. The direct and inverse shearlet transforms are denoted by $S : \mathbb{R}^{N\times N} \to \mathbb{R}^{\eta\times N\times N}$ and $S^* : \mathbb{R}^{\eta\times N\times N} \to \mathbb{R}^{N\times N}$, respectively, where η is the number of all shears in all scales of the shearlet transform. More details about the shearlet transform construction can be found in [13]. The reconstruction of unknown rows of the matrix g is formulated under the prior condition for having sparse solution in the shearlet domain, i.e.

$$\min_{f\in\mathbb{R}^{N\times N}} \|S(f)\|_0, \text{ subject to } g = M \odot f$$

which can be efficiently solved through the following iterative thresholding algorithm [13]:

$$f_{n+1} = S^*(T_{\lambda_n}(S(f_n + \alpha_n(g - M \odot f_n)))), \quad (1)$$

where the acceleration parameter α_n is chosen as follows

$$\alpha_n = \frac{\|\beta_n\|_2^2}{\|M \odot S^*(\beta_n)\|_2^2}, \quad (2)$$
$$\beta_n = S_{\Gamma_n}(y - M \odot f_n), \Gamma_n = \text{supp}(f_n),$$

and $(T_\lambda f)(k) = \begin{cases} f(k), & |x(k)| \geq \lambda \\ 0, & |x(k)| < \lambda \end{cases}$ is a hard thresholding operator. The initial value can be set to $f_0 = S^*_0(S_0(g))$, where S_0 and S^*_0 are direct and inverse transform using only low-pass element in the shearlet transform. The thresholding parameter λ_n is set to decrease with the iteration number n. In our case we apply a linear decrease from λ_{max} to λ_{min}, for L iterations such as $n = 0, \ldots, L$.

It is important to mention that one has to set a few parameters while running the algorithm. These are the number

of iterations L which directly influences the computational time; the initial estimation f_0 which can reduce the necessary number of iterations; and the threshold range $[\lambda_{max}, \lambda_{min}]$.

The generalization to full parallax is straightforward by successively implementing the basic algorithm along the horizontal and vertical camera axes. A more computationally-efficient alternative, referred to as hierarchical reconstruction [13], implements the reconstructions in a specific order, aimed at reducing the maximum disparity between input views after each iteration, thus reducing the required number of shearlet transform scales and the related processing time.

III. Accelerated Processing

The method in Section II is applicable for any EPI. A preferable solution would use the same set of parameters for every EPI and run the reconstructions independently. One can select optimal thresholding parameters for the reconstruction algorithm as $[\lambda_{max}^{opt}, \lambda_{min}^{opt}]$ and fix a common number of iterations N for all EPIs. In this case, the computation time linearly depends on the number of EPIs and the fixed number of iterations. By distributing the required computations equally between multiple GPUs, one can achieve the fastest computational time for this independent processing. Further acceleration can be achieved by speeding up the iterative algorithm itself and by utilizing similarities between EPIs.

A. Faster convergence by double overrelaxation

In this section we propose a modification of the main iterative algorithm aimed at its faster convergence. As presented in (2), the convergence is controlled by the parameter α_n, which was originally designed to provide stability for varying content for the price of increased computations [13]. As a computationally less expensive alternative, here we propose another update mechanism, based on the so-called double overrelaxation (DORE), similar to the one presented in [18]. Assume the light field EPI matrices are reordered in column vector form and assume the parameter $\alpha_n = \alpha$ is fixed. The thresholding operation

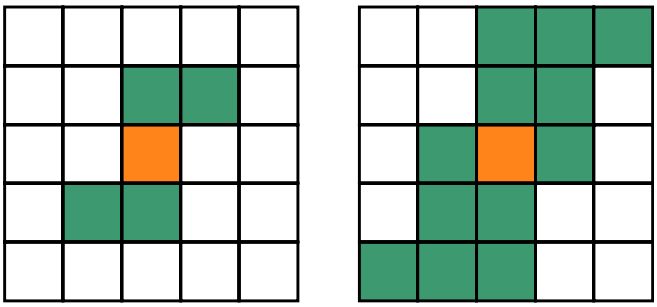
$$\hat{f}_n = S^*(T_{\lambda_n}(S(f_n + \alpha(g - M \odot f_n)))) \tag{3}$$

is followed by a two-step overrelaxation

$$\begin{aligned} \tilde{f}_n &= \hat{f}_n + \beta_1(\hat{f}_n - f_{n-1}) \\ \beta_1 &= \frac{(g - \hat{f}_n)^\top H(\hat{f}_n - f_{n-1})}{(\hat{f}_n - f_{n-1})^\top H(\hat{f}_n - f_{n-1})}, \end{aligned} \tag{4}$$

$$\begin{aligned} f_{n+1} &= \tilde{f}_n + \beta_2(\tilde{f}_n - f_{n-2}) \\ \beta_2 &= \frac{(g - \tilde{f}_n)^\top H(\tilde{f}_n - f_{n-2})}{(\tilde{f}_n - f_{n-2})^\top H(\tilde{f}_n - f_{n-2})}, \end{aligned} \tag{5}$$

where $H \in R^{N^2 \times N^2}$ is a diagonal matrix containing the elements of the measuring matrix M along its main diagonal. For additional stability we clamp the values of $\beta_1, \beta_2 \in [0, 1]$. The role of the double over-relaxation is to tackle potential instabilities by keeping the next iteration anchored to the previous iterations. Eq. (4) and (5) provide closed-form solutions for the respective line search problems $\beta_1 = \underset{\beta}{\operatorname{argmin}} ||H(g - (\hat{f}_n + \beta(\hat{f}_n - f_{n-1}))||^2$ and $\beta_2 = \underset{\beta}{\operatorname{argmin}} ||H(g - (\tilde{f}_n + \beta(\tilde{f}_n - f_{n-2}))||^2$. This leads to finding an optimal linear combination between consecutive solutions such that the error is minimized over the given samples defined by the matrix H.



(a) (b)

Fig. 2. (a) Proposed window (green) for modelling guidance with respect to reference pixel (orange). (b) Neighbourhood (green) for forming Laplacian matrix entry with respect to reference pixel (orange).

B. Color spaces and guided colorization

A trivial approach is to convert the RGB color channels into YUV colour space and process EPIs there, while expecting significantly less energy in the U and V colour channels. Specifically, we apply reversible color transform (RCT [19]) without any quantization of values, i.e.

$$\begin{cases} \text{Y} = (\text{R} + 2*\text{G} + \text{B})/4 \\ \text{U} = \text{B} - \text{G} \\ \text{V} = \text{R} - \text{G} \end{cases}.$$

Usually, the spatial information in U and V channels is highly redundant for natural images. Therefore, in the case of processing in YUV colour space with given N number of iterations, we reconstruct Y channel with N iterations and U, V channels with $N/2$ iterations. Compared with the reconstruction in RGB colour space, the overall number of iterations is reduced from $3N$ to $2N$.

Furthermore, we investigate the possibility of applying the fully reconstructed Y-channel EPI as a guide in reconstructing R, G and B color channels from their decimated EPI versions. This type of problem can be solved by methods previously developed for image colorization [20], [21]. It has been also shown that colorization can be considered as a particular case of the more general problem of alpha matting [22]. Specifically, we adopt the so called closed-form alpha matting algorithm proposed in [23] and modify it for the purpose of reconstructing color EPIs.

Following the notations in Section II, we denote the targeted EPI color channel and its decimated version by f and g respectively. Let us denote also the reconstructed Y-channel EPI by E. Then, the targeted color channel pixels f_i are modelled as a linear function of the known (i.e. guiding) image pixels E_i, within a small window w

$$f_i \approx aE_i + b, \forall i \in w.$$

For natural images, typically, the small window has been assumed to be a square window of 3×3 pixels around the reference pixel. For the case of EPI, we propose to use a different window to leverage the directional information presented in EPI. We show the proposed shape in Fig. 2(a).

The cost function minimization problem can be formulated as follows

$$J(f, a, b) = \sum_{j=1}^{N^2} \left(\sum_{i \in w_j} (f_i - a_j E_i - b_j)^2 + \varepsilon a_j^2 \right),$$

where the regularisation term εa_j^2 is added for numerical stability. In [23], it has bee shown that an equivalent minimization problem can be formulated using *matting Laplacian* matrix Λ, which removes the need to identify a and b

$$J(f) = \min_{a,b} J(f, a, b) \sim J(f) = f^T \Lambda f,$$

where the entries of the matrix $\Lambda \in R^{N^2 \times N^2}$ are calculated as follows

$$\Lambda(i, j) = \sum_{k|(i,j) \in w_k} \left(\delta_{ij} - \frac{1}{\aleph_k} \left(1 + \frac{1}{\frac{\varepsilon}{\aleph_k} + \sigma_k^2} (E_i - \mu_k)(E_j - \mu_k) \right) \right).$$

In the above equation, δ_{ij} denotes the Kronecker delta, and $\aleph_k$, μ_k, σ_k^2 denote the cardinality, the mean and the variance of the window w_k respectively. For the entry $L(i, j)$, the summation is done over all windows w_k which contain pixels with indices i and j. For a reference pixel at position i and for our choice of window shape, the pixel positions j are shown in Fig. 2(b). Given the true colors at the decimated EPI g, the problem is reformulated as

$$\text{minimize } f^T \Lambda f, \text{ s.t. } Hf = g,$$

where H is the diagonally-arranged measurement matrix M. The so-formulated problem is solved using the conjugated gradient method.

C. Group Processing of Similar EPIs

Previously, we have presented an attempt to accelerate the basic algorithm utilizing similarities between EPIs [24]. The method suggested constructing a tree, which defines the order of processing depending on similarity between EPIs. In the constructed tree, each node corresponds to an EPI. The tree is constructed by comparing EPIs for their similarity in terms of l^2 norm and consecutively connecting the most similar pairs of EPIs. Iterating over all EPIs, one obtains a connected graph. Then, the processing is performed from top to bottom, and the EPI being processed uses the reconstructed EPI at its parent node as an initial estimate. Our hypothesis was that the reconstruction over the graph would allow for adaptively choosing the number of iterations for each EPI depending on the similarity to its initial estimate. This approach heavily depends on the threshold defining similarity between EPIs and setting the same reconstruction parameters for different datasets is problematic. Therefore, in this article we adopt a more systematic approach toward exploring the EPI similarities, which would allow an easier tuning of algorithm parameters. First, we consider grouping of similar EPIs, done by comparing l^2 distances between EPIs against a predefined threshold t_s. Having the EPIs organized in groups, we fully reconstruct the average EPI over each group and use it as a guidance map to reconstruct the other EPIs in the group by the approach proposed in Subsection III-B.

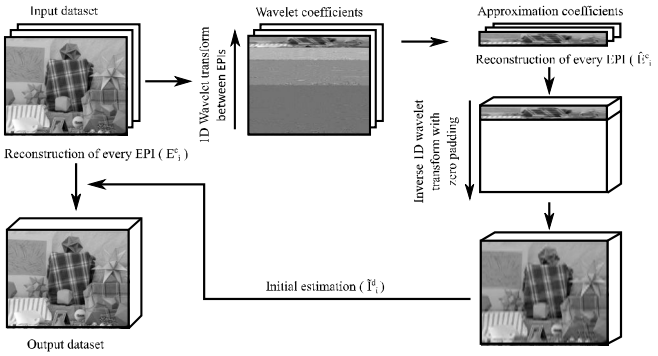


Fig. 3. Reconstruction flowchart using wavelet transform approximation coefficient as an initial estimation.

D. Initialization by Wavelet Transform

The redundancy between EPIs can be regarded as redundancy in the vertical direction in the given multi-perspective images. Instead of local grouping of similar EPIs as in Subsection III-C, we consider the alternative of decorrelating the vertical image lines by a fixed transform, e.g. a wavelet transform. Namely, a wavelet transform is performed on $E_i^c(\cdot, \cdot)$ along the i axis which is equivalent to performing 1D wavelet transform vertically on every input image $I_k^c(i, \cdot)$ along i axis. By performing L level of 1D wavelet transform between EPIs $E_i^c, i = 1, ..., p$, we expect to split them into EPIs with small-magnitude detail coefficients and $\tilde{E}_i^c, i = 1, ..., p/2^L$ EPIs with higher-magnitude approximation coefficients. The approximation coefficients gather most of the information of the original set of EPIs. The reconstruction is then applied directly on EPIs formed by wavelet transform approximation coefficients. The obtained set of densely sampled EPIs $\tilde{E}_i^d, i = 1, ..., p/2^L$ contain a good amount of directional structures (more global ones), however, they still require further processing to obtain desirable quality of reconstruction (add details from the original set of EPIs). Therefore, the inverse wavelet transform can be applied on the reconstructed EPIs of the approximation coefficients with an appropriate padding with zeros corresponding to detail coefficients. The obtained set of EPIs $\tilde{I}_i^d, i = 1, ..., p$ is used as an initial estimate for reconstruction of the original input E_i^c EPIs by performing additional processing by the modified basic algorithm. The processing times for the two steps can be set independently. The flowchart of the approach is shown in Fig. 3.

IV. Experimental Evaluation

A. Algorithm Implementation

We have implemented the core reconstruction algorithms, on a GPU using CUDA Toolkit [25]. Since the shearlet transform is a translation invariant transform, it can be efficiently

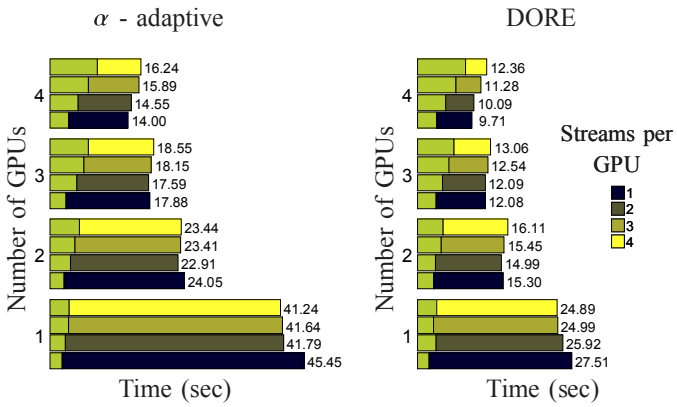


Fig. 4. Computational time required to perform 50 iterations of α-adaptive algorithm (left) and $DORE$ (right) on 100 EPIs using different parallelization between GPUs. Light green color represents the necessary time for initialization of the algorithm. For resolution 256×512 using more than one stream per GPU doesn't provide acceleration.

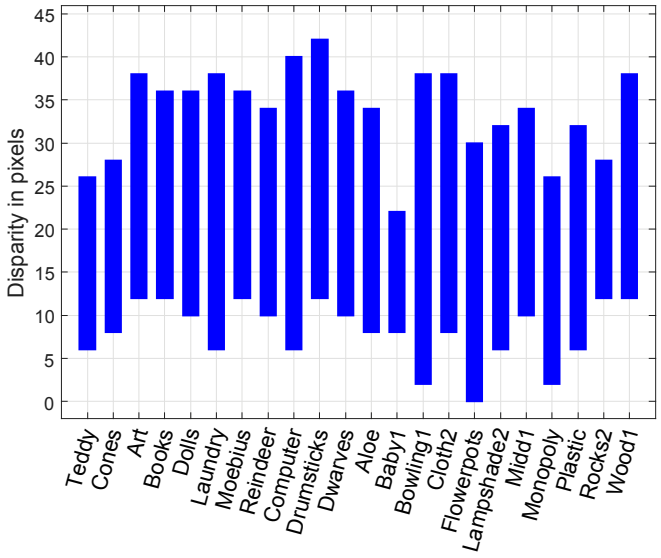


Fig. 5. Illustration of the disparity ranges $[d_{min}, d_{max}]$ between adjacent views for input dataset used in this paper (obtained from ground true disparity maps).

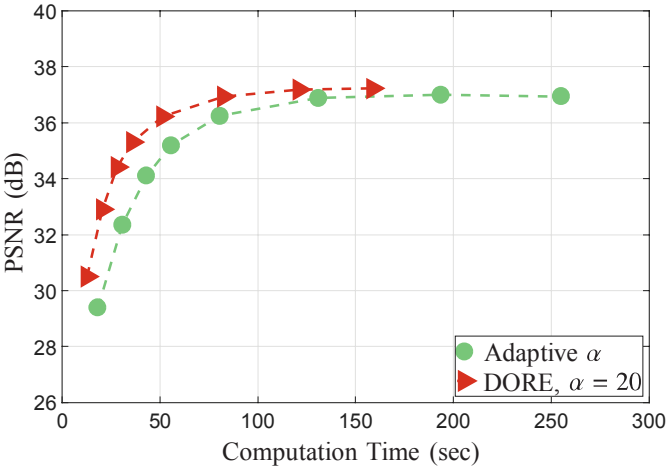


Fig. 6. Comparison of average performance between basic algorithm with adaptive selection of the parameter α and DORE with fixed α = 20 for the horizontal-parallax datasets.

computed using the Fast Fourier Transform (FFT). In our case, we used the cuFFT library to get an FFT implementation on the GPU [26]. For generating our experimental results we have used a system consisting of four Nvidia GeForce GTX Titan X GPUs. The computational time for reconstructing an EPI mainly depends on the number of overall iterations that have to be performed. In order to achieve fastest computation for a given set of input EPIs with a corresponding number of iterations, the set has been distributed between GPUs such that the overall number of iterations that has to be performed are approximately equal for each GPUs. On the level of one GPU, the whole iterative processing is performed independently from other GPUs. Depending on the size of the processed EPI, we get different occupation of GPU kernels at a time. We consider EPIs with the size of 256×512 processed with the shearlet transform at 5 scales using algorithm presented in Sections II, III-A. In both cases, 100 EPIs are processed. The computational times are presented in Fig. 4. Note that reconstructing in the case of 256×512 with S^5 transform, only one process per GPU is sufficient for both algorithms, while DORE is significantly faster.

B. Evaluation

In our comparative tests, we have used datasets presented in [27], [28] for horizontal parallax and in [29] for full parallax datasets. In overall, 22 horizontal-parallax datasets and two full-parallax datasets of various depth and spatial content have been used. In all experiments, the input data is formed by every second view of the test dataset. The other views form the reference. The algorithm performance has been evaluated by comparing the difference between the reconstructed and the reference views in terms of PSNR (dB). $J = 5$ scale levels have been considered for the shearlet transform (S^5) which corresponds to an intermediate view reconstruction, where the maximum disparity between adjacent views is in the range of $[0, 32]$ pixels. The exact disparity ranges of each scene as obtained from the ground truth disparity maps are shown in Fig. 5. In order to shift the available (existing) disparity ranges to $[0, d_{max} - d_{min}]$, for each input dataset we perform first a horizontal shearing by $-d_{min}$ on all EPIs. After reconstructing $d_{max} - d_{min} - 1$ intermediate views, a shearing by $d_{min}/(d_{max} - d_{min})$ is applied to return the imagery to the original disparity range. In general, we evaluate algorithms presented in Section III for different number of iterations in order to compare trend of convergence speed of different algorithms in average for all datasets.

In our first comparative test, we present the average reconstruction quality for the algorithm modification based on DORE with $\alpha = 20$ and the original α adaptive algorithm. The comparison is done for $5, 10, 15, 20, 30, 50, 75, 100$ iterations per EPI. The results for the horizontal parallax datasets are shown in Fig. 6, while the reconstruction results for the full-parallax datasets are shown in Fig. 7. The DORE algorithm provides faster convergence for all datasets and results in

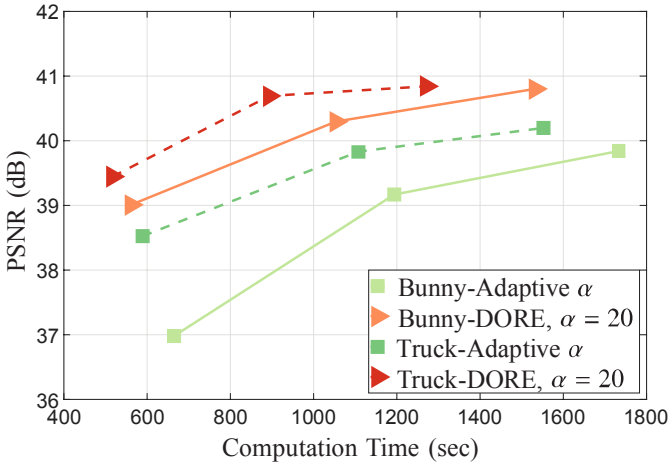


Fig. 7. Comparison of performance between basic algorithm with adaptive selection of the parameter α and DORE with fixed $\alpha = 20$ for the full-parallax datasets.

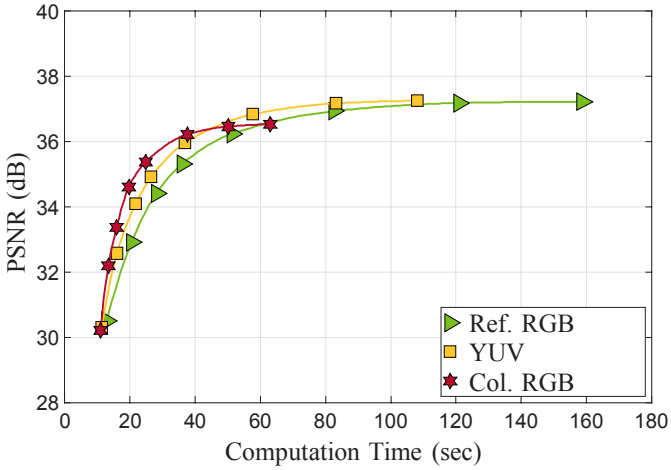


Fig. 8. Comparison between reference algorithm (RGB), YUV, and colorization of RGB for the horizontal-parallax datasets.

better quality enjoying also faster processing. In all subsequent experiments, we use the DORE algorithm with $\alpha = 20$, referring to it as the reference algorithm.

Next, we aim at quantifying the performance of the colorization algorithm. For the horizontal-parallax datasets, we present the trend in reconstruction quality for different number of iterations, see Fig. 8. The reference algorithm reconstructs the three color channels, R, G, and B in an equal number of iterations, while in YUV, priority is given to the Y channel, which is processed twice longer than the U and V channels. In the case of colorization, an average intensity channel is formed as $Y = (1/3) * (R + G + B)$ and fully reconstructed in varying number of iterations, then each of the color channels is reconstructed by colorization using the reconstructed Y channel as a guidance. As can be seen in the figure, the prioritized processing brings some improvement over the reference algorithm and the algorithm based on colorization is significantly faster as it processes a single channel only. All three algorithms saturate in performance, which means that after some number of iterations, no quality improvement is

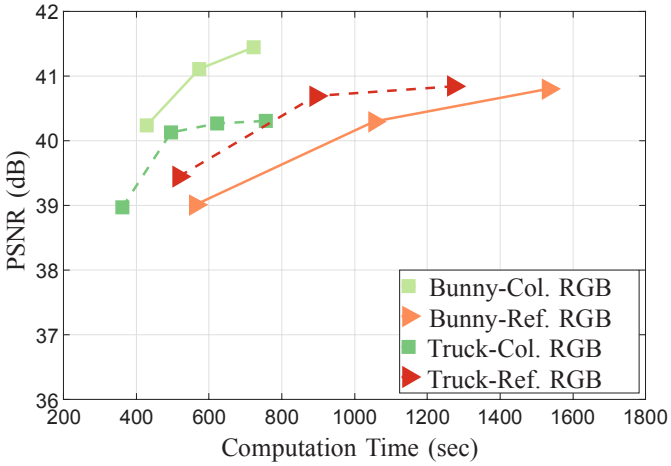


Fig. 9. Comparison between reference algorithm (RGB), and colorization of RGB for the full-parallax datasets.

achieved. The colorization algorithm saturates at lower level, which indicates that the structural differences in the three color channels have not been fully reconstructed in the averaged intensity channel. The reached values after 100 iterations for each of the three algorithms are the following: $RGB-37.22dB$, $YUV-37.24dB$, $Col.RGB-36.22dB$. These results suggest that the colorization algorithm is preferable in case of limited computation time, since it converges faster, while for the case of better computing resources, the best quality is achieved by the YUV color space processing.

Fig. 9 presents the results for the full-parallax datasets. For the *Bunny* dataset, the colorization shows a significant improvement both in terms of time and quality, while for the *Truck* dataset, the results are in agreement with the average result over the horizontal-parallax datasets.

Fig. 8 presents the average results over the whole group of test scenes. The result in terms of rate of convergence vary substantially for the individual test scenes. In order to further analyse the algorithm based on colorization, we look at the saturation points for the reference algorithm and the colorization algorithm for each individual dataset. The two algorithms are run for increasing number of iterations and saturation points are estimated at the iteration where further improvement is negligible. Denote by $E(k), T(k), k = 1, \ldots$ the quality level (e.g. PSNR), and the corresponding time in seconds for the running iteration k. We define $k_{\max} = \arg\max_k E(k)$ and define the saturation point as $k_{sat} = \arg\max_k \left(\frac{E(k+1)-E(k)}{E(k_{\max})} < 0.0015 \right)$. The idea is illustrated in Fig. 10 for the Teddy dataset, where the saturation points are given by the circles around the corresponding iterations.

Having found two saturation points per dataset, one for the reference and one for the colorization algorithm, one can compare them in terms of quality variation (ISNR) and time acceleration ratio. In other words, we compare the best achievable quality per dataset for the two algorithms versus the time acceleration ratio it brings. Fig. 11 presents this comparison over the horizontal-parallax datasets. In the figure, each dot represents one dataset, the $x-$ axis represents the relative time acceleration achieved by the colorization algorithm versus the

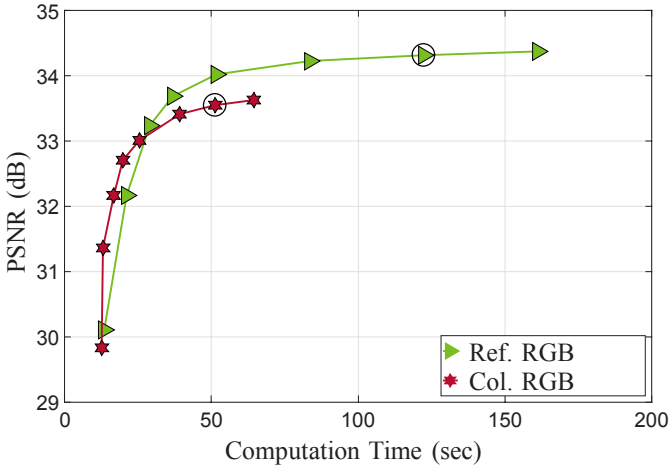


Fig. 10. Saturation points for the reference and colorization algorithms for the Teddy dataset. The obtained saturation points are illustrated by black circles.

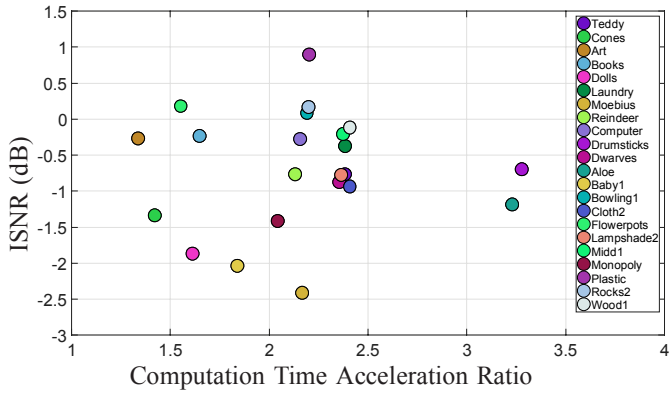


Fig. 11. Comparison of saturation points of the colorization and reference algorithms for the horizontal-parallax test datasets.

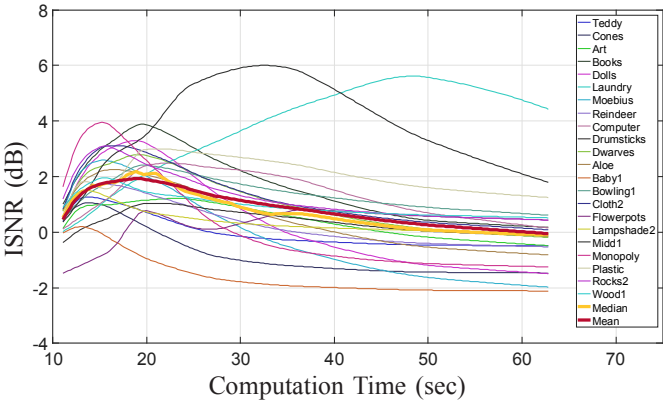


Fig. 12. ISNR of colorization versus reference algorithm for individual sequences.

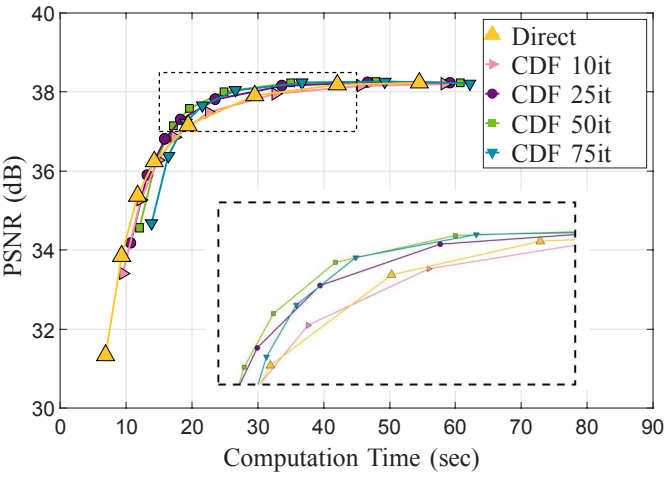


Fig. 13. Comparison of the reconstruction trends depending on number of iterations used for obtaining initial estimation in method utilizing wavelet transform. In the legend of the figure presented corresponding number of iterations used for processing initial estimations.

reference one, and the $y-$ axis represents the improvement in the signal-to-noise ratio ISNR (dB). As seen in the figure, there are a few sets, where the acceleration in time comes together with improved quality, while for the majority of datasets, the acceleration is achieved for the price of reduced quality.

Another way to illustrate the performance of the colorization algorithm is to show the ISNR in comparison to the reference algorithm for the same time, using interpolation between iteration points. Fig. 12 gathers the performance for each individual dataset, along with the mean and median values. For short processing times, colorization is to be preferred as most of the sequences show positive ISNR values. As the processing time gets longer, the values cluster around the zero ISNR line (as shown also by the mean and median curves), while there are a few sequences still enjoying better performance with the colorization method and other sequences showing worsening results. Apparently, among the former group these are datasets with relatively simpler color and depth distributions.

To simplify the next experiments, we limit the comparison of algorithms exploring the inter-EPI similarities and decorrelation to comparing the results for the Y channel only, assuming that the RGB color channels can be efficiently reconstructed by the Y channel.

For the wavelet transform based acceleration approach, presented in Section III-D, we perform $L = 3$ levels of the $CDF\ 9/7$ transform. Here, the issue is to find the best proportion between the processing time allocated for obtaining the initial estimate and the processing time allocated for refining the EPI reconstruction based on this initial estimate. In order to illustrate the trend in convergence, we perform experiments where the initial estimate is obtained by reconstructing the coarse wavelet coefficients with $10, 25, 50, 75$ iterations. The obtained initial estimate is then refined for the same number of iterations, as the direct (reference) algorithm. Fig. 13 depicts the trends. Naturally, the time needed for obtaining the initial estimates, shift the initial curve points to the right, e.g. the curve corresponding to 75 iterations allocated for getting the initial estimate is the rightmost in the figure. Then, the curves corresponding to wavelet-based initialization get better and saturate faster, with the case of 50 iterations for the initial estimate showing the best performance.

The algorithm based on grouping similar EPIs and processing them together as presented in Section III-C does not show consistent results for all datasets. This is to be attributed to the

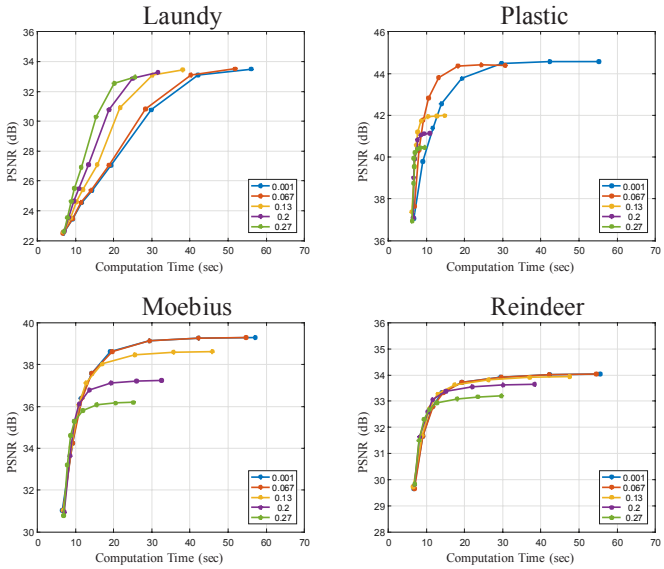


Fig. 14. Comparison of the reconstruction trends depending on different thresholding value in method utilizing groups formed based on similarity of the EPIs. In the legend of the figure presented corresponding thresholding values.

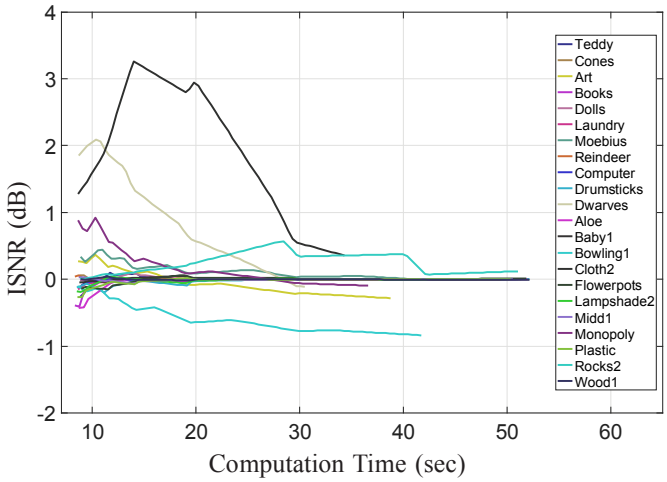


Fig. 15. Average reconstruction performance for the method utilizing grouping based on similarity between EPIs for different datasets for fixed thresholding value.

strong dependence on the threshold value, which determines which EPIs are sufficiently similar. In general, increasing the threshold value leads to increasing the size of formed groups and therefore decreases the computation time. However, the effect on quality is very much content-dependent. Results for several datasets with different threshold values are presented in Fig. 14. For the dataset *Laundry*, one can get significant acceleration, while e.g. for the dataset *Moebius*, the reconstruction quality is always inferior compared with the reference algorithm. Selecting one of the well-performing thresholds, i.e. the value of 0.067, one can get the performance for each individual dataset, as shown in Fig. 15.

Some examples of synthesized views are presented in Fig. 16, 17 with the corresponding quality in terms of PSNR. The DORE-based algorithm provides consistently better convergence compared to the original method [13]. The colorization-based algorithm achieves good reconstruction results when the Y channel manages to get the important structure of the scene (edge information) existing in R, G and B channels. For the particular scene *Laundry*, the algorithm based on wavelet transform provides better convergence compared to the reference RGB algorithm.

V. Conclusions

In this article, we have addressed the problem of accelerating the DSLF reconstruction algorithm, which originally uses sparse camera views and works on each EPI independently by employing regularized iterative reconstruction in shearlet transform domain. In order to speed up the algorithm, we proposed modifications in three categories. First, we aimed at improving the algorithm itself by using double relaxation in the iterative procedure. Second, we explored the similarities between color channels within the same EPI in the flavor of colorization based approaches. Third, we aimed at avoiding redundant processing and reducing the overall reconstruction time through exploiting similarities between EPIs. Furthermore, our implementation employed GPUs allowing for an efficient parallelized computation of the iterative procedure and the underlying shearlet transform.

We have generated experimental results on a wide set of test sequences and analyzed the performance of the considered approaches. The new reconstruction method based on double overrelaxation shows better convergence speed in comparison with the original algorithm. We favor the use of colorization as the approach catches well the color dependences in natural images. The benefit of using similarities between EPIs is very much content dependent. The wavelet based approach shows a marginal improvement in terms of convergence rate, which is still worth employing. As of the algorithm based on grouping of similar EPIs and group processing, it provides acceleration only for scenes where significant amount of EPIs are similar.

The modifications employ structured similarities within EPI and between EPIs were integrated within the DSLF reconstruction algorithm. However, they are perfectly applicable also in other LF image processing algorithms, where DSLF reconstruction is not the main goal. Such potential applications include LF depth estimation, compression, segmentation, and matting.

References

[1] N. S. Holliman, N. A. Dodgson, G. E. Favalora, and L. Pockett, "Three-dimensional displays: a review and applications analysis," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 362–371, 2011.

[2] B. Javidi and F. Okano, *Three-dimensional television, video, and display technologies*. Springer Science & Business Media, 2002.

[3] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," 2017.

[4] Z. Lin and H.-Y. Shum, "A geometric analysis of light field rendering," *Int'l J. of Computer Vision*, vol. 58, no. 2, pp. 121–138, 2004.

[5] C. Fehn, "Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv," vol. 5291, 2004, pp. 93–104.

[6] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, "Unstructured lumigraph rendering," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '01. New York, NY, USA: ACM, 2001, pp. 425–432. [Online]. Available: http://doi.acm.org/10.1145/383259.383309

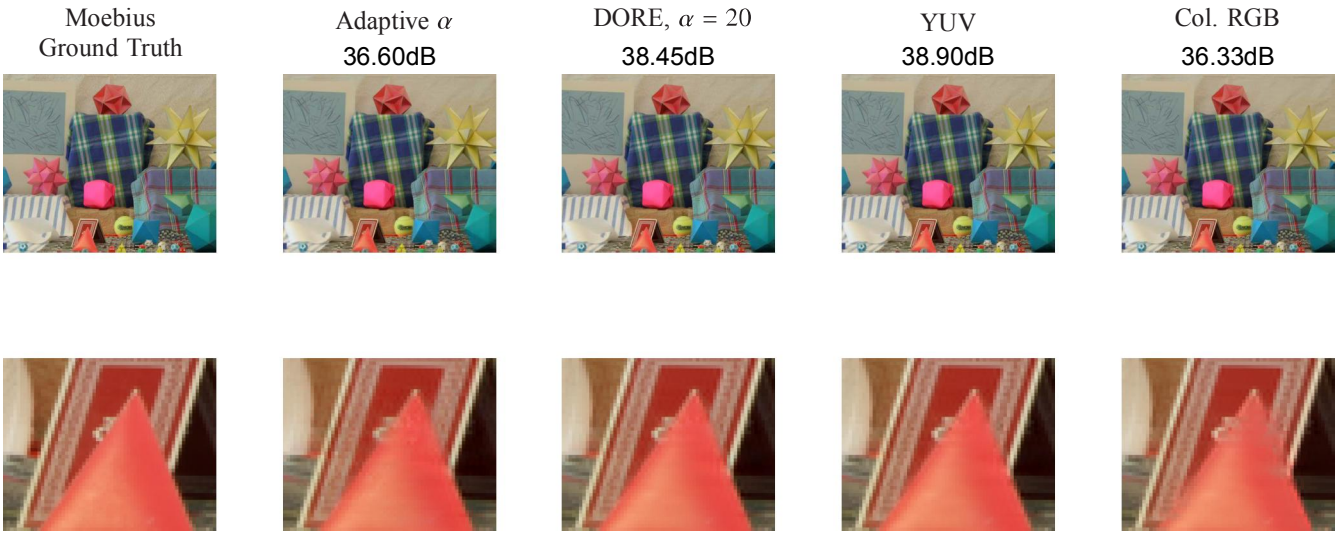


Fig. 16. Example of synthesized views for datasets $Moebius$ using different algorithms for approximately same computation time. Presented algorithms are used for accelerated color space reconstruction.

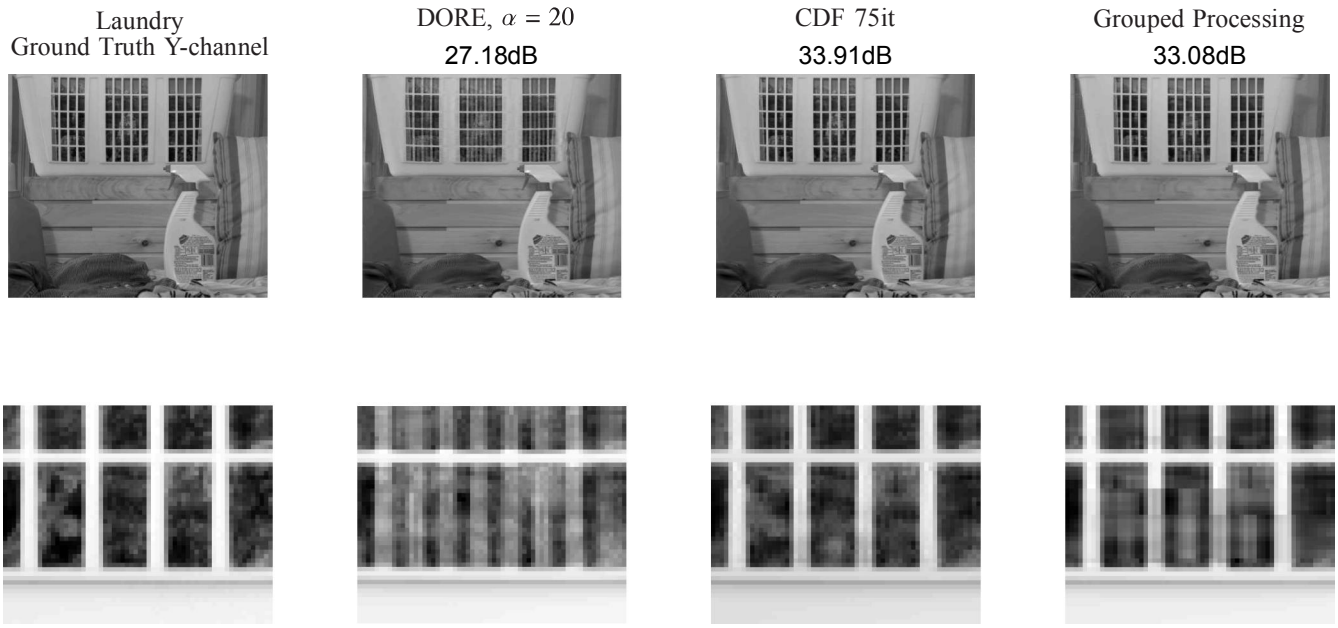


Fig. 17. Example of synthesized views for only Y-channel of dataset $Laundry$ using different decorrelation algorithms for approximately same computation time.

[7] 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 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, ser. CVPR '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 519–528. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.19

[8] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 1068–1080, June 2008.

[9] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, "Global solutions of variational models with convex regularization," *SIAM Journal on Imaging Sciences*, vol. 3, no. 4, pp. 1122–1145, 2010.

[10] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 73:1–73:12, Jul. 2013.

[11] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous fourier domain," *ACM Trans. on Graphics (TOG)*, vol. 34, no. 1, p. 12, 2014.

[12] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Image based rendering technique via sparse representation in shearlet domain," in *Image Processing (ICIP), 2015 IEEE International Conference on*, Sept 2015, pp. 1379–1383.

[13] ——, "Light field reconstruction using shearlet transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.

[14] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Depth estimation reference software (ders) 5.0," *ISO/IEC JTC1/SC29/WG11 M*, vol. 16923, 2009.

[15] M. Tanimoto, T. Fujii, and K. Suzuki, "Reference software of depth estimation and view synthesis for ftv/3dv," *ISO/IEC JTC1/SC29/WG11 M*, vol. 15836, 2008.

[16] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, Feb 2008.

[17] C.-K. Liang, Y.-C. Shih, and H. Chen, "Light field analysis for modeling image formation," *IEEE Trans. Image Processing*, vol. 20, no. 2, pp. 446–460, Feb 2011.
[18] K. Qiu and A. Dogandzic, "Double overrelaxation thresholding methods for sparse signal reconstruction," in *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, March 2010, pp. 1–6.
[19] Y. Chen and P. Hao, "Integer reversible transformation to make jpeg lossless," in *Signal Processing, 2004. Proceedings. ICSP'04. 2004 7th International Conference on*, vol. 1. IEEE, 2004, pp. 835–838.
[20] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," in *ACM Transactions on Graphics (ToG)*, vol. 23, no. 3. ACM, 2004, pp. 689–694.
[21] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, "Natural image colorization," in *Proceedings of the 18th Eurographics Conference on Rendering Techniques*, ser. EGSR'07. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2007, pp. 309–320. [Online]. Available: http://dx.doi.org/10.2312/EGWR/EGSR07/309-320
[22] J. Wang and M. F. Cohen, "Image and video matting: A survey," *Found. Trends. Comput. Graph. Vis.*, vol. 3, no. 2, pp. 97–175, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1561/0600000019
[23] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 228–242, Feb 2008.
[24] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Tree-structured algorithm for efficient shearlet-domain light field reconstruction," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2015, pp. 478–482.
[25] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable parallel programming with cuda," *Queue*, vol. 6, no. 2, pp. 40–53, 2008.
[26] V. Podlozhnyuk, "Fft-based 2d convolution," *NVIDIA white paper*, p. 32, 2007.
[27] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. I–195–I–202, June 2003.
[28] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, June 2007.
[29] V. Vaish and A. Adams, "The (new) stanford light field archive," *http://lightfield.stanford.edu*, 2008.

Suren Vagharshakyan Suren Vagharshakyan received the MSc in mathematics from Yerevan State University (2008). He is a PhD student at the Department of Signal Processing at Tampere University of Technology since 2013. His research interests are in the area of light field capture and reconstruction.

Robert Bregovic Robert BregoviÄĞ (Member, IEEE) received the Dipl. Ing. and MSc degrees in electrical engineering from University of Zagreb, Zagreb, Croatia, in 1994 and 1998, respectively, and the Dr. Sc. (Tech.) degree (with honors) in information technology from Tampere University of Technology, Tampere, Finland, in 2003. From 1994 to 1998, he was with the Department of Electronic Systems and Information Processing of the Faculty of Electrical Engineering and Computing, University of Zagreb. Since 1998, he is with the Laboratory of Signal Processing, Tampere University of Technology. His research interests include the design and implementation of digital filters and filterbanks, multirate signal processing, and topics related to acquisition, processing, modeling, and visualization of 3D content.

Atanas Gotchev Atanas Gotchev (Member, IEEE) received the M.Sc. degrees in radio and television engineering (1990) and applied mathematics (1992) and the Ph.D. degree in telecommunications (1996) from the Technical University of Sofia, and the D.Sc.(Tech.) degree in information technologies from the Tampere University of Technology (2003). He is a Professor at the Laboratory of Signal Processing and Director of the Centre for Immersive Visual Technologies at Tampere University of Technology. His research interests have been in sampling and interpolation theory, and spline and spectral methods with applications to multidimensional signal analysis. His recent work concentrates on algorithms for multisensor 3-D scene capture, transform-domain light-field reconstruction, and Fourier analysis of 3-D displays.

PUBLICATION

IV

Light Field Reconstruction Using Shearlet Transform

S. Vagharshakyan, R. Bregovic and A. Gotchev

IEEE Transactions on Pattern Analysis and Machine Intelligence 40.1 (2018), 133–147

Light Field Reconstruction Using Shearlet Transform

Suren Vagharshakyan, Robert Bregovic, *Member, IEEE,* and Atanas Gotchev, *Member, IEEE*

Abstract—In this article we develop an image based rendering technique based on light field reconstruction from a limited set of perspective views acquired by cameras. Our approach utilizes sparse representation of epipolar-plane images (EPI) in shearlet transform domain. The shearlet transform has been specifically modified to handle the straight lines characteristic for EPI. The devised iterative regularization algorithm based on adaptive thresholding provides high-quality reconstruction results for relatively big disparities between neighboring views. The generated densely sampled light field of a given 3D scene is thus suitable for all applications which require light field reconstruction. The proposed algorithm compares favorably against state of the art depth image based rendering techniques and shows superior performance specifically in reconstructing scenes containing semi-transparent objects.

Index Terms—Image-based rendering, light field reconstruction, shearlets, frames, view synthesis.

1 Introduction

Synthesis of intermediate views from a given set of captured views of a 3D visual scene is usually referred to as image-based rendering (IBR) [1]. The scene is typically captured by a limited number of cameras which form a rather coarse set of multiview images. However, denser set of images (i.e. intermediate views) is required in immersive visual applications such as free viewpoint television (FVT) and virtual reality (VR) aimed at creating the perception of continuous parallax.

Modern view synthesis methods are based on two, fundamentally different, approaches. The first approach is based on the estimation of the scene depth and synthesis of novel views based on the estimated depth and the given images, where the depth information works as correspondence map for view reprojection. A number of depth estimation methods have been developed specifically for stereo images [2], and for multiview images as well [3], [4], [5], [6], [7], [8], [9]. In all cases, the quality of depth estimation is very much content (scene) dependent. This is a substantial problem since small deviations in the estimated depth map might introduce visually annoying artifacts in the rendered (synthesized) views. The second approach is based on the concept of plenoptic function and its light field (LF) approximation [10], [11]. The scene capture and intermediate view synthesis problem can be formulated as sampling and consecutive reconstruction (interpolation) of the underlying plenoptic function. LF based methods do not use the depth information as an auxiliary mapping. Instead, they consider each pixel of the given views as a sample of a multidimensional LF function, thus the unknown views are function values that can be determined after its reconstruction from samples. In [12], different interpolation kernels utilizing available geometrical information are discussed. As shown there, established interpolation algorithms such as linear interpolation require a substantial number of samples (images) in order to obtain synthesized views with good quality.

The required bounds for sampling the LF of a scene have been defined in [13]. In order to generate novel views without ghosting effects by using linear interpolation, one needs to sample the LF such that the disparity between neighboring views is less than one pixel [13]. Hereafter, we will refer to such sampling as dense sampling and to the correspondingly sampled LF as densely sampled LF. In order to capture a densely sampled LF, the required distance between neighboring camera positions can be estimated based on the minimal scene depth (z_{min}) and the camera resolution. Furthermore, camera resolution should provide enough samples to properly capture highest spatial texture frequency in the scene [14].

Densely sampled LF is an attractive representation of scene visual content, particularly for applications, such as refocused image generation [15], dense depth estimation [16], object segmentation [17], novel view generation for FVT [18], and holographic stereography [19]. However, in many practical cases one is not able to sample a real-world scene with sufficient number of cameras to directly obtain a densely sampled LF. Therefore, the required number of views has to be generated from the given sparse set of images by using IBR.

An approach for LF reconstruction from undersampled LFs has been presented in [20]. It combines a band-limited filtering with wide-aperture reconstruction which is essentially a directional edge-preserving filtering. The problem of upsampling camera arrays has been cast as a directional super-resolution in 4D space with no use of depth information [21]. The generation of the desired perspective views is performed through patch matching and the effect of sampling patterns has been studied. In [22], convolutional neural networks have been utilized to predict depth from LF data. The method learns an end-to-end mapping between the LF and a representation of the corresponding 4D depth field in terms of 2D hyperplane orientations. The obtained prediction is then further refined in a post processing step by applying a higher-order regularization. In [23], view synthesis technique has been presented based on learning-based approach using two convolutional neural networks for disparity and color estimation. Four corner views from

the light fields are used to synthesise an intermediate view. This method has been aimed at increasing angular resolution of the light field captured by Lytro Illum camera.

The work [14] has discussed the effective use of the depth limits (z_{min}, z_{max}) in order to reconstruct desired views from a limited number of given views using appropriate interpolation filters. Use has been made of the so-called epipolar-plane image (EPI) and its Fourier domain properties [24]. Further benefits in terms of improved rendering quality has been achieved by using depth layering [4], [14]. More recently, another approach to LF reconstruction has been proposed [25]. It considers the LF sampled by a small number of 1D viewpoint trajectories and employs sparsity in continuous Fourier domain in order to reconstruct the remaining full-parallax views.

The problem of reconstructing a piecewise-smooth function using its given incomplete measurements has been addressed in the context of natural images through sparse approximation provided by some appropriately constructed transforms [26], [27], [28]. The general aim has been to design frames or other over-complete image representations and to study their performance by the asymptotic decay speed of the approximation error obtained using only N largest coefficients of the decomposition. Within this context, wavelets have been found less efficient for representing images and other systems have been designed with better approximation properties. The sought transforms have targeted good directional sensitivity in order to tackle singularities in images, which are usually distributed over smooth curves being borders between smooth image regions. Examples include adaptive triangle based approximation [29], tight curvelet frames [27], contourlets [30], and shearlets [31]. Among the designed transforms, shearlets have been shown to be optimally sparse and getting very close to the ideal adaptive image decomposition [31], [32].

In this article, we advance the concepts of LF sparsification and depth layering with the aim to develop an effective reconstruction of the LF represented by EPIs. The reconstruction seeks to utilize an appropriate transform providing sparse representation of the EPI. We assume that a good sparse transform should incorporate scene representation with depth layers, which are expected to be sparse. Based on the observation that the anisotropic property of the EPI is caused by a shear transform, we favor the shearlet transform as the sought sparsifying transform and develop an inpainting technique working on EPI, in a fashion similar to how shearlets have been applied for seismic data reconstruction [33].

Preliminary results of novel view synthesis by using shearlet transform have been presented in [34]. In this paper, we extend the ideas presented in [34] by including the underlying analysis, describing in detail the construction of the used shearlet transform and the corresponding view synthesis algorithm for the cases of horizontal and full parallax and evaluating the efficiency of the proposed algorithm on various datasets. Furthermore, we present experiments for the cases of non-equidistant camera positions and reconstruction of scenes containing semi-transparent objects.

The outline of this paper is as follows. The LF and EPI concepts are presented in Section 2. The same section discusses the shearlet transform, its properties and construction for the given case. The reconstruction algorithm is presented in Section 3. The algorithm evaluation for different datasets and a comparison with the state of the art is presented in Section 4. Finally, the work is concluded in Section 5.

2 Light Field Formalization and Representation

2.1 Light Field Representation

The propagation of light in space in terms of rays is fully described by the $7D$ continuous plenoptic function $R(\theta_1, \theta_2, \omega, \vartheta, V_x, V_y, V_z)$, where (V_x, V_y, V_z) is a location in the 3D space, (θ_1, θ_2) are propagation angles, ω is wavelength, and ϑ is time [10]. In more practical considerations, the plenoptic function is simplified to its 4D version, termed as 4D LF or simply LF. It quantifies the intensity of static and monochromatic light rays propagating in half space. In this representation, the LF ray positions are indexed either by their Cartesian coordinates on two parallel planes, the so-called two-plane parameterization $L(u, v, s, t)$, or by their one plane and direction coordinates $L(u, v, \theta_1, \theta_2)$ [35].

Consider a pinhole camera, with image plane (u, v) and focal length f, moving along the (s, t) plane. This is an important practical consideration, which associates the parameterizing planes with LF acquisition and multiview imagery and relates LF sampling with discrete camera positions and a discrete camera sensor. The case is illustrated in Fig. 1 (a) where the z axis represents the scene depth and the plane axes s and u are considered perpendicular to the figure and omitted for simplicity. Constraining the vertical camera motion by fixing $s = s_0$ and moving the camera along the t-axis, leads to so-called horizontal parallax only (HPO) multiview acquisition. Images captured by successive camera positions $t_1, t_2, \ldots$ can be stacked together which is equivalent to placing the t-axis perpendicular to the (u, v) plane. The corresponding LF $L(u, v, s_0, t)$ is illustrated in Fig. 1 (b).

2.2 EPI Representation and Sampling Requirements

The LF data organization as in Fig. 1 (b) leads to the concept of EPIs pioneered by Bolles et al. in [24]. Assume an ideal horizontal camera motion (or, equivalently, perfectly rectified perspective images). Gathering image rows for fixed $u = u_0$ along all image positions forms an LF slice $E(v, t) = L(u_0, v, s_0, t)$. Such LF slice is referred to as EPI and is given in Fig. 1 (c). In the EPI, relative motion between the camera and object points manifests as lines with depth depending slopes. Thus, EPIs can be regarded as an implicit representation of the scene geometry. In comparison with regular photo images, an EPI has a very well defined structure. Any visible scene point appears in one of the EPIs as a line whose slope depends on the distance of the point from the capture position and the measured intensity over the line reflects the intensity of emanated light from that scene point. The Lambertian reflectance model (any point in the scene emanates light in different direction with same intensity) leads to an EPI with even more definitive structure – each line in the EPI has a constant intensity proportional to the intensity of the point. For a scene point at depth z_0 measured from the capture plane (s_0, t), the

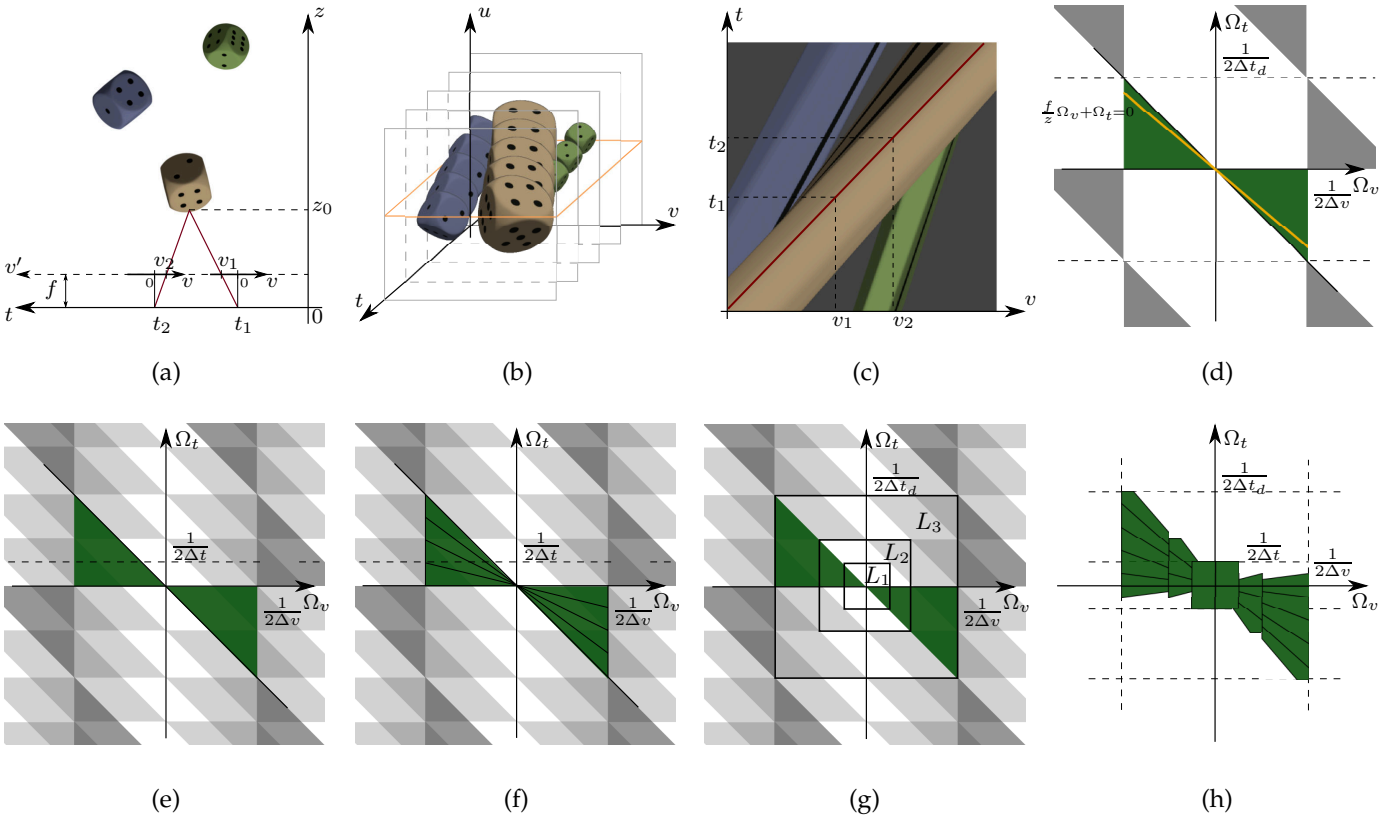


Fig. 1. Epipolar-plane image (EPI) formation and its frequency domain properties. (a) Capturing setup and EPI formation, a scene point is observed by two pinhole cameras positioned at t_1, t_2 at image coordinates v_1 and v_2 respectively; (b) Stack of captured images; an epipolar plane is highlighted for fixed vertical image coordinate u; (c) Example of EPI; red line represents a scene point in different cameras; (d) Frequency support of a densely sampled EPI; green area represents the baseband bounded by min and max depth; yellow line corresponds to a depth layer, the slope determines the depth value; (e) Frequency domain structure of an EPI being insufficiently sampled over t-axis, the overlapping regions represent aliasing; (f) Desirable frequency domain separation based on depth layering; (g) Frequency domain separation based on dyadic scaling; (h) Composite directional and scaling based frequency domain separation for EPI sparse representation.

disparity in the image plane (u_0, v) between two cameras positioned at t_1 and t_2 is [14]

$$\Delta v = v_2 - v_1 = \frac{f}{z_0}(t_2 - t_1) = \frac{f}{z_0}\Delta t,$$

where f is the camera focal length. This is illustrated by the red lines in Fig. 1 (a), which show a point projected on cameras at t_1 and t_2. The same point appears as the red line in Fig. 1 (c).

By assuming a horizontal sampling interval Δv satisfying the Nyquist sampling criterion for scene's highest texture frequency, one can relate the required camera motion step (sampling interval) with the scene depth. For given z_{min} the sampling interval Δt should be such that

$$\Delta t \leq \frac{z_{min}}{f}\Delta v \tag{1}$$

in order to ensure maximum 1 pixel (px) disparity between nearby views [13], [14]. Fig. 1 (d) shows the frequency domain support of a densely sampled EPI, which is of bow-tie shape. The baseband (in green) is limited by the minimum and maximum depth and its replicas are caused by the sampling intervals Δv and Δt. In Fourier domain, the frequency support of a depth layer (i.e. all scene points at a certain depth z_0, which in EPI appear as lines with same slope) is confined to a line. An example is given by the yellow line in Fig. 1 (d). By selecting equality for Δt in (1), which is denoted in Fig. 1 as Δt_d, we effectively place the z_{min} line at 45 degrees in the frequency domain plane. This maximizes the baseband support, which helps in designing linear reconstruction filters.

2.3 Motivation

Our problem in hand is to reconstruct densely sampled EPIs (and thus the whole LF) from their decimated and aliased versions produced by a coarser camera grid determined by a higher interval Δt. The problem is illustrated in Fig. 1 (e). The figure shows a case, where a densely sampled EPI has been decimated by a factor of 4, which means that every 4th row has been retained while the others have been zeroed. As seen in the figure, aliased replicas (gray) and the baseband (green) overlap, hence a band-limited reconstruction is infeasible with a classical filtering method. Therefore, the work [14] has specified requirements for the LF sampling density for given z_{min} and z_{max} in order to allow a band-limited reconstruction. Reconstruction of more complex scenes (e.g. piecewise-planar or tilted-plane) would require additional information about scene depth and depth layering [4], [14]. For real scenes it is natural to assume that objects are distributed at a finite, rather small number of depths. In our approach, we aim at implicitly determining those sparse depth layers by analyzing the given aliased

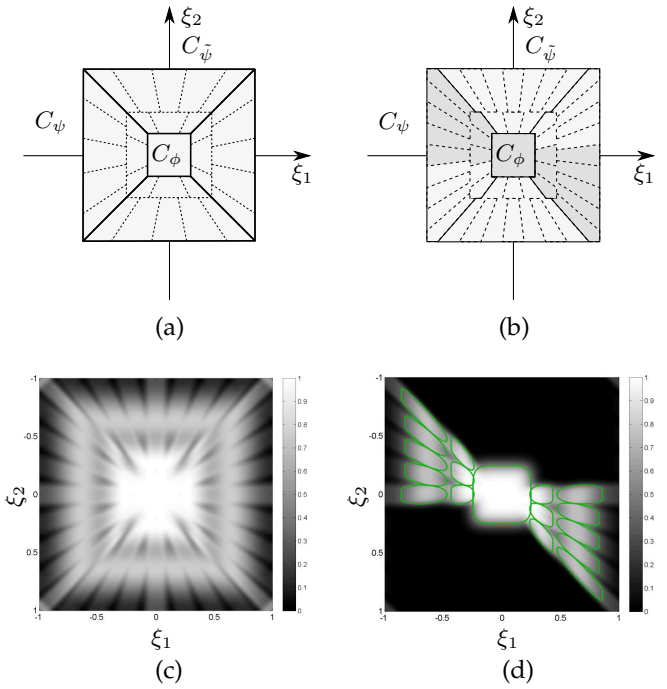


Fig. 2. (a) Frequency plane tilting by shearlet transform. C_ψ,$C_{\tilde{\psi}}$ are cone-like regions and C_ϕ is low-frequency region. (b) Desirable frequency domain tilting by proposed reconstruction algorithm. Gray color region includes transform elements used for reconstruction; other transform elements are not associated with valid shear values (disparities) in EPI. (c) $\hat{\Psi}^d$ corresponding to constructed shearlet transform for $J = 2$. (d) Frequency domain support of shearlet transform elements used in reconstruction algorithm corresponding to gray color region in (b). Green contour regions in (d) represent significant parts of transform elements support in frequency domain.

EPIs in frequency domain using depth guided filters. This is equivalent to applying a proper frequency plane tiling. The case in Fig. 1 (e) is further analyzed in Fig. 1 (f), which highlights a frequency plane tiling by 4 depth layers, with 1px disparity range in each layer. If those depth layers are given, they are sufficient to guide the interpolation of EPIs without aliasing artifacts. Furthermore, by an additional dyadic separation of the frequency plane, i.e. a multiresolution analysis, one can process each region differently and utilize a more efficient analysis tool. Fig. 1 (g) illustrates a wavelet based separation of the frequency plane for the same aliased EPI. It is easy to notice that the L_1 region does not contain any aliasing. Therefore by applying a low-pass filter corresponding to the L_1 region on the aliased EPI will reconstruct the desirable densely sampled EPIs frequencies in that region. In other words, the procedure of low-pass filtering followed by decimation can be interpreted as increasing the pixel size, which directly decreases the disparity between the given rows. In this manner, fewer depth layering directions will have to be distinguished from each other in order to efficiently reconstruct the full EPI. Based on the above discussion, the desirable frequency plane tiling with elemental filters for the case of densely sampled EPI reconstruction from its 4th row subsampled version is given in Fig. 1 (h). The construction of such set of filters is closely related to the construction of shearlet frames as presented in the next section.

2.4 Shearlet Transform

The shearlet system is our main tool for EPI sparsification. We establish the following general notations. We deal with two-dimensional functions $f(x) \in L^2(\mathbb{R}^2), x = (x_1, x_2)$. The corresponding Fourier transform is denoted by $\hat{f}(\xi), \xi = (\xi_1, \xi_2)$. The discretized version of $f(x)$ is denoted by $f^d(m), m \in \mathbb{Z}^2, m = (m_1, m_2)$. In frequency domain, discrete sequences generate trigonometric polynomials, which, for brevity, are also denoted by the $\hat{}$ sign. The conjugate of a function f is denoted by $\bar{f}$. While processing EPIs, the spatial axes (x_1, x_2) correspond to (v, t) parameters of the plenoptic function, and the frequency domain variables (ξ_1, ξ_2) correspond to the frequency axes (Ω_v, Ω_t).

We are specifically interested in the so-called cone-adapted shearlet system, which can generate the directed multi-scale frequency bands as conceptualized in Fig. 1 (h) [28], [36]. Consider two cone-like regions $C_\psi, C_{\tilde{\psi}}$ complemented by a low-pass region C_ϕ as highlighted in Fig. 2 (a). For their effective tiling, one needs shearlet system elements (atoms) generated by a scaling function $\phi \in L^2(\mathbb{R}^2)$ and two shearlets $\psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$.

The shearlet system is generated by the translation of the scaling function and translation, shearing and scaling of the shearlet transform

$$SH(c; \phi, \psi, \tilde{\psi}) = \begin{cases} \phi_m = \phi(\cdot - c_1 m), m \in \mathbb{Z}^2, \\ \psi_{j,k,m} = 2^{(j+\lfloor j/2 \rfloor)/2j} \psi(S_k A_{2^j} \cdot - M_c m), \\ \tilde{\psi}_{j,k,m} = 2^{\frac{j+\lfloor j/2 \rfloor}{2} j} \tilde{\psi}(S_k^T \tilde{A}_{2^j} \cdot - \tilde{M}_c m), \end{cases}$$

where $S_k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}$ is a shear matrix, $M_c = \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \end{pmatrix}$, $\tilde{M}_c = \begin{pmatrix} c_2 & 0 \\ 0 & c_1 \end{pmatrix}$, $c = (c_1, c_2)$ are sampling densities of the translation grid and A_{2^j} and $\tilde{A}_{2^j}$ are scaling matrices, which for the case of EPI take the form

$$A_{2^j} = \begin{pmatrix} 2^j & 0 \\ 0 & 2^{-1} \end{pmatrix}, \tilde{A}_{2^j} = \begin{pmatrix} 2^{-1} & 0 \\ 0 & 2^j \end{pmatrix}.$$

This particular form of the scaling matrices supports the desirable number of shears in each scale and provides scaling only by one axis, therefore it is well suited for representing the EPI singularities distributed over straight lines. It can be considered as a special case of a more general shearlet transform called universal shearlet [28], [36].

The transform maps $f \in L^2(\mathbb{R}^2)$ to the sequence of coefficients

$$f \to \langle f, \tau \rangle, \tau \in SH(c; \phi, \psi, \tilde{\psi}).$$

The properties of the shearlet transform highly depend on the design of the generator functions $\phi, \psi, \tilde{\psi}$. A specific design of compactly supported scaling function and shearlets is discussed in Appendix A.

In order to handle discrete data by the continuous shearlet transform, we assume that the given samples $f_J^d(n), n \in \mathbb{Z}^2$ correspond to samples of the continuous function, for some sufficiently large $J \in \mathbb{N}$

$$f(x) = \sum_{n \in \mathbb{Z}^2} f_J^d(n) 2^J \phi(2^J x - n).$$

The particular choice of J depends on the given input data and will be discussed in Section 3.1.

For the efficient implementation of the transform, one needs its representation in the form of digital filters $\psi^d_{j,k,m}$ corresponding to $\psi_{j,k,m}$. The discretization is not trivial and technical details are provided in Appendix B.

As the frame elements are not orthogonal, one needs also the dual frame elements. They can be constructed based on the shift invariance properties of the shearlet frame. First, we set

$$\hat{\Psi}^d = |\hat{\phi}^d|^2 + \sum_{j=0,\dots,J-1} \sum_{|k|\leq 2^{j+1}} (|\hat{\psi}^d_{j,k}|^2 + |\hat{\tilde{\psi}}^d_{j,k}|^2).$$

Then, the dual shearlet filters are defined in Fourier domain, as follows:

$$\hat{\varphi}^d = \frac{\hat{\phi}^d}{\hat{\Psi}^d}, \hat{\gamma}^d_{j,k} = \frac{\hat{\psi}^d_{j,k}}{\hat{\Psi}^d}, \hat{\tilde{\gamma}}^d_{j,k} = \frac{\hat{\tilde{\psi}}^d_{j,k}}{\hat{\Psi}^d}.$$

The constructed frame guarantees stable reconstruction, if $A \leq \hat{\Psi}^d \leq B$ is satisfied for some finite bounds $0 < A, B < \infty$ [37]. An illustration of the obtained $\hat{\Psi}^d$ for $J = 2$ is presented in Fig. 2 (c). In this case, the upper and lower bounds are numerically found to be $0.03 < \hat{\Psi}^d < 1.03$.

Since we are going to use shearlet transform for processing EPIs, we are interested only in shear operation with a positive sign, i.e. $0 \leq k \leq 2^j + 1$. The corresponding frame elements cover the frequency plane region highlighted by gray in Fig. 2 (d). The resulting direct transform S for discrete values f^d_J and $j = 0, ..., J-1, k = 0, ..., 2^j + 1, m \in \mathbb{Z}^2$ is

$$S(f^d_J) = \left\{ s_{j,k}(m) = (f^d_J * \bar{\psi}^d_{j,k})(m), s_0(m) = (f^d_J * \bar{\phi}^d)(m) \right\}.$$

The corresponding inverse transform is then

$$S^*\left(\{s_{j,k}, s_0\}\right) = \sum_{\substack{j=0,\dots,J-1 \\ k=0,\dots,2^{j+1}}} (s_{j,k} * \gamma^d_{j,k})(m) + (s_0 * \phi^d)(m).$$

The frequency-domain support of the elements selected from the frame in Fig. 2 (c) is shown in Fig. 2 (d).

3 Reconstruction Algorithm

In this section we present the developed LF reconstruction algorithm, which utilizes EPI sparse representation in shearlet domain. We first present the main features for the case of horizontal parallax only and then discuss the specifics of the full parallax implementation.

3.1 Horizontal Parallax

Usually, a setup of uniformly distributed, parallel positioned and rectified cameras is used for capturing a 3D scene. The horizontal parallax between views limits the motion associated with the depth of the objects in horizontal axis only. This allows us to perform intermediate view generation over EPI independently. In order to formulate the reconstruction algorithm in discrete domain we assume that the starting coarse set of views are downsampled version of the unknown densely sampled LF we try to reconstruct. The uniformly distributed cameras imply the possibility of

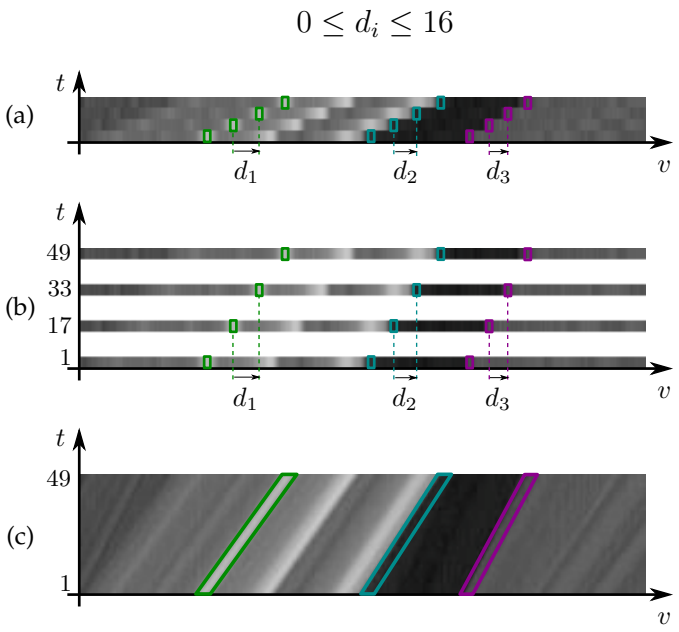


Fig. 3. The given 4 views with maximal disparity 16px between consecutive views are interpreted as every 16th view in the target densely sampled LF. (a) EPI for coarsely sampled LF over t-axis; (b) corresponding partially defined densely sampled EPI; (c) ground truth densely sampled EPI. Three different points from given input images forming traces are highlighted in the coarsely (a) and densely (c) sampled EPIs. Only in (c) they are revealed as a straight lines.

estimating a common upper bound d_{max} for disparities between nearby views. Thus, the given coarse set of views are regarded as taken at each $d_{max} = \lceil d_{max} \rceil$-th view of a densely sampled LF. Thus, in every densely sampled EPI, all unknown rows should be reconstructed assuming given every d_{max}-th row. An example is presented in Fig.3 (a), where EPI representation of four views with 16px disparity is given. Therefore, the targeted densely sampled EPI is to be constructed in such a way that the available data will appear in rows with 16px distance (Fig.3 (b)). Fig.3 (c) shows the same rows with respect to the fully reconstructed EPI, where successive rows appear at disparity less than or equal to 1px. EPI lines are not distinguishable in Fig.3 (a). The lines start to form when the views are properly arranged, as in Fig.3 (b), and they get fully reconstructed in the densely sampled EPI. A set of non-equidistant cameras implying non-uniform down-sampling of densely sampled LF can be handled likewise, as far as the given views are arranged properly with respect to the global d_{max}.

Without loss of generality we assume that the densely sampled EPI is a square image denoted by $y^* \in \mathbb{R}^{N^2}$, where $N = (K-1)d_{max} + 1$ and K is the number of available views. The samples $y \in \mathbb{R}^{N^2}$ of y^* are obtained by

$$y(i,j) = H(i,j)y^*(i,j), \tag{2}$$

where $H \in \mathbb{R}^{N^2}$ is a measuring matrix, such that $H(kd_{max}, \cdot) = 1, k = 1, \dots, K$ and 0 elsewhere. The measurements y form an incomplete EPI where only rows from the available images are presented, while everywhere else EPI values are 0. Eq. (2) can be rewritten in the form $y = Hy^*$ by lexicographically reordering the variables

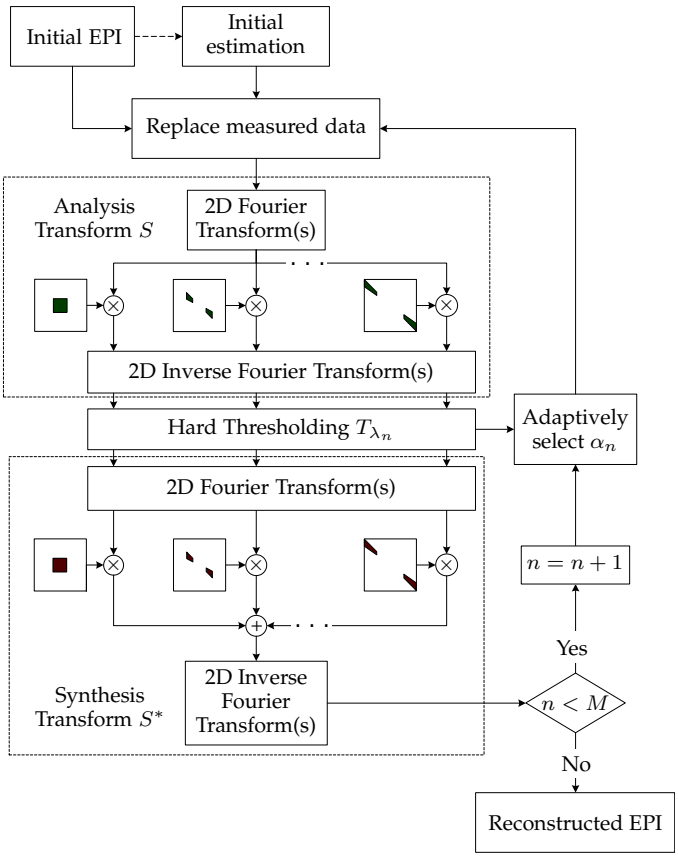


Fig. 4. Diagram of the EPI reconstruction algorithm.

$y, y^* \in \mathbb{R}^{N^2}, H \in \mathbb{R}^{N^2 \times N^2}$. The shearlet analysis and synthesis transforms are defined as $S : \mathbb{R}^{N^2} \to \mathbb{R}^{N^2 \times \eta}, S^* : \mathbb{R}^{N^2 \times \eta} \to \mathbb{R}^{N^2}$, where η is the number of all translation invariant transform elements.

The reconstruction of y^* given the sampling matrix H and the measurements y can be cast as an inpainting problem, with constraint to have solution which is sparse in the shearlet transform domain, i.e.

$$x^* = \underset{x \in \mathbb{R}^{N^2}}{\arg\min} \|S(x)\|_1, \text{subject to } y = Hx. \tag{3}$$

We make use of the iterative procedure within the morphological component analysis approach, which has been originally proposed for decomposing images into piecewise-smooth and texture parts [38], [39]. In particular, we aim at reconstructing the EPI y^* by performing regularization in the shearlet transform domain. Solution is sought in the form of the following iterative thresholding algorithm

$$x_{n+1} = S^* \left(T_{\lambda_n} (S(x_n + \alpha_n (y - Hx_n))) \right), \tag{4}$$

where $(T_\lambda x)(k) = \begin{cases} x(k), |x(k)| \geq \lambda \\ 0, |x(n)| < \lambda \end{cases}$ is a hard thresholding operator applied on transform domain coefficients and α_n is an acceleration parameter. The thresholding level λ_n decreases with the iteration number linearly in the range $[\lambda_{max}, \lambda_{min}]$. After sufficient number of iterations, $x_n \to x^*$ reaches a satisfying solution of the problem (3). The diagram of the reconstruction method is given in Fig. 4.

The rate of convergence is controlled by the parameter α_n. For $\alpha_n = 1$ the convergence is slow and can be accelerated by selecting $\alpha_n > 1$. However, selecting alpha too high can cause instability. The case is illustrated in Fig. 5 where the convergence speed benefits from fixing a higher value $\alpha_n = \alpha$ up to some value where the algorithm starts to diverge. Best values for fixed α are different for different EPIs. This motivates us to apply an iteration-adaptive selection of the parameter α_n, which can be applied to all EPIs. We devise the adaptation procedure in the way as proposed in [40]. Let us define Γ_n as the support of $S(x_n)$. The adaptive selection of the acceleration parameter is

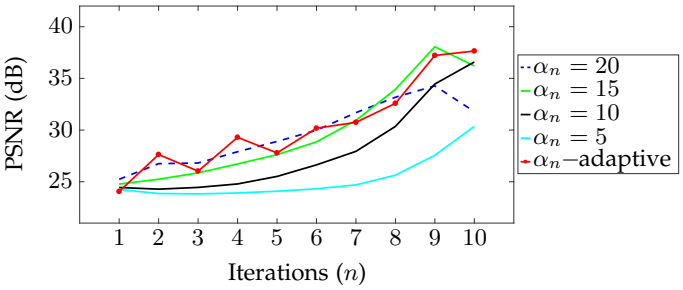


Fig. 5. Example of reconstruction performance dependence on choice of acceleration coefficients α_n. For constant value for all iterations $\alpha_n = \alpha$, increasing α brings accelerating convergence. After some value, reconstruction starts to diverge ($\alpha = 20$).

$$\alpha_n = \frac{\|\beta_n\|_2^2}{\|HS^*(\beta_n)\|_2^2},$$

where $\beta_n = S_{\Gamma_n}(y - Hx_n)$ and S_{Γ_n} is the shearlet transform decomposition only for coefficients from Γ_n. The convergence rate for the adaptive selection of the acceleration parameter is illustrated in Fig. 5. As can be seen in the figure, the adaptation provides high convergence speed and stable reconstruction.

The initial estimate f_0 can be chosen either 0 everywhere or as the result of a low-pass filtering of the input y using the central separable filter ϕ^d only.

As discussed previously we are not obliged to use all general shearlet transform atoms. We favor the use of atoms which are associated with valid directions in EPI, i.e. only those having support in frequency domain enclosed in the region highlighted in Fig. 1 (d). An example of such subset is presented in Fig. 1 (h). The scales of the shearlet transform are constructed in dyadic manner, therefore we can select the number of scales as follows

$$J = \lceil \log_2 d_{max} \rceil. \tag{5}$$

For every scale we select $2^{j+1} + 1$ shears ($j = 0, \ldots, J - 1$) to cover the region presented in Fig. 1 (g) associated with $s_k = \frac{k}{2^{j+1}}, k = 0, \ldots, 2^{j+1}$ shears (i.e. disparities). The role of J is two-side. Selecting higher J will guarantee better refinement however for the price of more computations. Related with this, d_{max} has to be specified rather correctly in order to avoid unnecessary computations. Choosing lower value for J than the one suggested by (5) will drastically decrease the reconstruction quality because of the lack of shearing atoms.

The parameter d_{max} itself has to be fixed at the stage of sampling (multiview acquisition) or can be estimated from

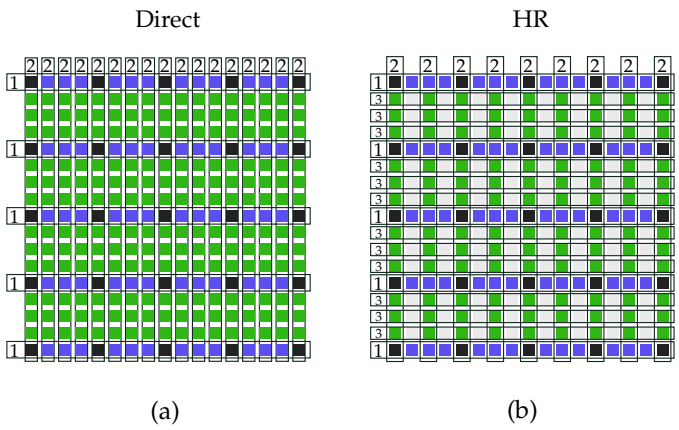


Fig. 6. Array of 17×17 views considered for reconstruction using 5×5 views highlighted with black color. (a) Direct reconstruction method. (b) Hierarchic order of reconstruction (HR).

an already captured imagery by some fast sparse feature-based or coarse-to-fine disparity estimation methods. In our implementation, we have used the method developed in [41], which was modified for the case of multi-view images.

3.2 Full Parallax

The method for reconstruction of HPO LFs can be generalized for the case of full parallax in a straightforward manner by directly reconstructing the vertical parallax views after all horizontal parallax views have been reconstructed. We illustrate this direct approach by Fig. 6 (a). The figure represents an array of 17×17 full parallax views to be reconstructed out of 5×5 views marked in black. The views marked in blue are the views reconstructed first (in the horizontal parallax reconstruction step) and the views marked in green represent the views reconstructed in the second (vertical parallax) reconstruction step.

The direct full parallax reconstruction is computationally demanding. Therefore, as a second approach we propose performing the reconstruction in a specific order that, from iteration to iteration, gradually reduces the maximum disparity between input views. This, in turn, reduces the number of scales in the shearlet transform and thereby speeds up the algorithm. We refer to this algorithm as hierarchical reconstruction (HR). We illustrate it by means of the same example, where we aim at reconstructing 17×17 views out of 5×5 given views. Let us assume that the maximum disparity is 12. We perform the reconstruction in 3 steps, as illustrated in Fig. 6 (b).

1) Views in rows $1, 5, 9, 13, 17$ are reconstructed first using (4) and shearlet transform with four number of scales ($ST(4)$), since the assumed maximal disparity is 12, hence, $\lceil log_2(12) \rceil = 4$. This step reconstructs views marked in blue in Fig. 6(b).
2) Views in columns $1, 3, 5, \ldots, 17$ are reconstructed, again using $ST(4)$ since the disparity is the same as in Step 1. This step reconstructs views marked in green in Fig. 6(b).
3) Missing views in rows $2, 3, 4, 6, 7, 8, \ldots, 18$ are reconstructed. Since there are more vertical views available than in the initial set, the disparity in this reconstruction step has been reduced to 6. Therefore, one can use $ST(3)$.

For other cases where more intermediate views have to be reconstructed, one can further alternate between reconstructing horizontal and vertical views. At each step, the disparity reduces by two, thus gradually decreasing the required number of scales of shearlet transform.

4 Evaluation

In this section we provide details about the implementation of the proposed algorithm and evaluate its performance using wide range of datasets. As evident from Section 2.4 the direct and inverse shearlet transforms involve a good number of digital filtering operations applied at each iteration of the reconstruction algorithm. We opt for implementing them by circular convolution in Fourier domain as presented in the diagram in Fig. 4. In this implementation, one should consider reasonable padding with zeros for the input signal such that the border artifacts are tackled. Increasing the padding region increase the size of the convolved signals with an effect on computation time. We have used GPU implementation of the proposed reconstruction algorithm and the experiments presented in this paper were executed on a GeForce GTX Titan X. The computation time mainly depends on the time for computing 2D FFT for large-size arrays. The reconstruction of an LF might vary from few minutes to a couple of hours depending on the number of scales, the desirable number of iterations and the given resolution of images in the dataset.

We quantify the reconstruction performance for different test sets using leave N out tests. The experimental setup considers downsampled versions of a number of given multiview test sets, where every $(N+1)$-th view is kept and the others are dropped. The downsampled versions are used as input to the algorithm, which is supposed to reconstruct all dropped views.The reconstruction quality is assessed by calculating the PSNR between the original and the reconstructed views. Along with figures and tables in the article, we present supplementary videos at the journal web site, illustrating the performance of the proposed method.

4.1 Evaluation of Sparsifying Transforms

First, we demonstrate the performance of the reconstruction algorithm with respect to different sparsifying transforms [36], [42]. We compare Haar wavelets, the compactly supported shearlets as constructed in [36] and the fast finite shearlet transform [42]. The ground truth densely sampled EPI (Fig. 7(d)) has been obtained using properly generated views of a synthetic scene. Every 16-th row has been used as input for the reconstruction algorithm as in Fig. 7 (a), and interpreted in similar fashion as presented in Fig. 3. The obtained reconstruction results using the algorithm in Section 3.1 are presented in Fig. 7. The reconstruction using Haar wavelet transform is not properly revealing straight lines and the performance is poor. Directional sensitive transforms are showing better reconstruction performance, while the proposed shearlet transform outperforms the others. The proposed transform combines two properties, compact support in horizontal direction in spatial domain

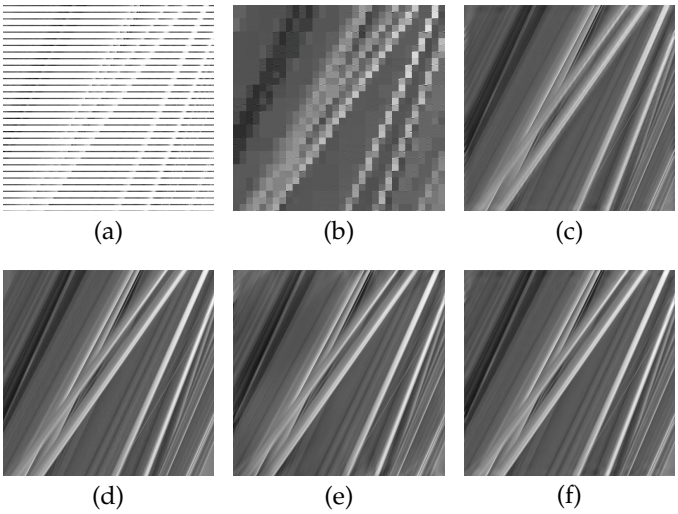


Fig. 7. (a) Input for reconstructing densely sampled EPI where only every 16-th row is available. (d) Densely sampled ground truth EPI. Reconstruction results using different transform are shown as follows (b) Haar 18.83dB, (c) Shearlab [36] 29.27dB, (e) FFST [42] 37.27dB, (f) Proposed modified shearlet 40.75dB.

and tight distribution of transform elements near the low-pass region of the Fourier plane which affect the reconstruction performance. The shearlet transform as developed in Section 2.4 can handle the reconstruction of EPIs from highly decimated versions. The proposed construction provides an optimal size of atoms compared to the other methods and at same time preserves the desirable Fourier plane tiling.

4.2 Multiview Datasets

We compare our approach against established depth based approaches. These include the reference methods and software used by the MPEG community for the development of new multiview video compression methods, namely DERS (depth estimation reference software) [46] and VSRS (view synthesis reference software) [47], and a state of the art method for disparity estimation employing semi-global stereo matching (SGBM) [6]. DERS is applied for every three consecutive views in order to estimate the disparity map collocated with the middle view. Using a stack of given images with corresponding estimated disparity maps, the desired intermediate views are generated using VSRS. In the case of SGBM, we obtain disparity maps for every pair of consecutive views in the given stack and warp the views by linear interpolation to obtain the intermediate views.

We have used a number of publicly available multiview datasets, as presented in Table 1. The table summarizes also some specifications of the sequences such as spatial resolution, number of views of the provided dataset, processing color space. For some of the datasets ($Couch$, $Teddy$, $Cones$) we also applied shearing on input views by d_{min} in order to compensate the minimum disparity d_{min} such that the maximum disparity in the sheared datasets can be considered as $d_{range} = d_{max} - d_{min}$. In all test cases, our algorithm is applied independently on every EPI to reconstruct the missing intermediate views. The adaptive acceleration parameter, as described in Section 3.1, has been applied. Typically, 100 iterations is used with λ thresholding

TABLE 1
Multiview Data Sets Details

Dataset	Resolution	Number of views	Leave N out	d_{range}
Couch [3]	2768×4020	51	1	12(RGB)
Pantomime1 [43]	640×480	73	7	24(Y),16(UV)
Pantomime2 [43]	640×480	77	3	28(Y),16(UV)
Teddy [44]	450×375	9	1	20(RGB)
Cones [44]	450×375	9	1	20(RGB)
Truck [45]	384×512	17×17	1, 3	6,12(RGB)
Bunny [45]	512×512	17×17	1, 3	6,12(RGB)

value linearly decreasing in the range of $[5, 0.02]$ per EPI in each dataset to obtain the presented results.

Fig. 8 presents the comparative results for two $Pantomime$ and $Coach$ sequences. As seen in the figures, for the $Pantomime$ sequences we used shearlet transform with 5 and 6 number of scales, denoted as $ST(5)$ and $ST(6)$, with $ST(6)$, in average, outperforming other competing algorithms. In the case of the $Couch$ sequence, the performance of all algorithms is similar. For this particular test sequence, we also compare the results with the method presented in [3], referred to as Disney in Fig. 8(c). It should be pointed out that in [3] the disparity maps are estimated using the full set of images, not only the downsampled one. Thus, the depth maps are expected to be of higher quality than the one that can be achieved if only the downsampled views are given. Surprisingly enough, the results of the method by Disney and SGBM are identical, while the latter is more general in the sense that it requires only stereo pairs from the decimated views as an input. This motivates us to further use SGBM as depth-based reference method. The comparison reveals that our algorithm reconstructs views with competitive quality without the need of any disparity / depth estimation. It is interesting to observe that in some sequences, there are views that were problematic for all algorithms, e.g. view 20 in $Pantomime2$. For this particular case, the cause is that the input data contains hardly pronounced EPI structures, which are insufficient for generating the particular view.

For the datasets $Teddy$ and $Cones$ containing originally 9 views we consider every second view as input, or 5 views in overall. The obtained disparity range is estimated to be 20px for both datasets. As seen in Fig. 9, the proposed method with shearlet transform using 5 number of scales ($ST(5)$) is in par with SGBM. However, when using 6 number of scales ($ST(6)$), which corresponds to 64 depth layers, the proposed method consistently performs better than the methods using SGBM or DERS. These results show that using higher number of scales is beneficial in the case of complex scenes. For the $Teddy$ dataset, we also compare the proposed method with the one presented in [4] which is an IBR method utilizing depth layering. For the purpose of comparison, we average the performance of the proposed method over all four reconstructed views. The average reported in [4] shows 33.25dB, while our method gives 35.29dB in the case where d_{min} has not been compensated and the reconstruction has been applied assuming $d_{min} = 0$.

Reconstruction results for the multiview datasets are

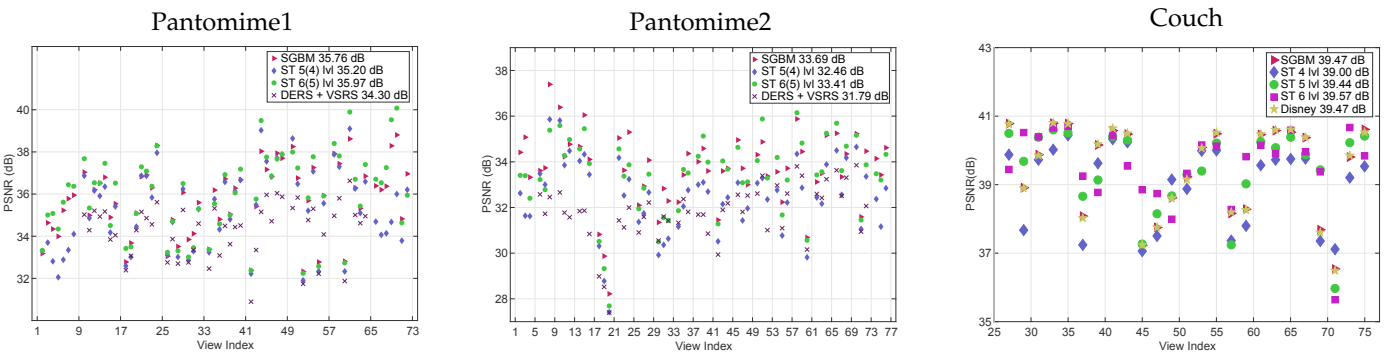


Fig. 8. Reconstruction results for different multiview datasets, error shown in PSNR for reconstructed views.

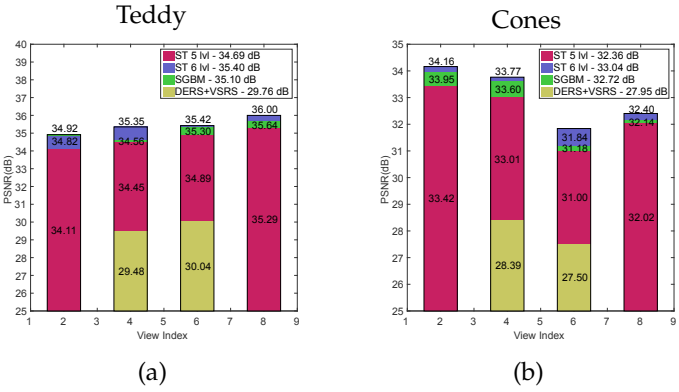


Fig. 9. Reconstruction quality for datasets (a) $Teddy$ and (b) $Cones$. Evaluation has been performed for proposed methods $ST(5), ST(6)$, DERS+VSRS, SGBM. Average PSNR of all reconstructed views presented in legend of the figures.

illustrated in Fig. 15.

4.3 Semi-Transparent Objects

Next, we demonstrate the superiority of our algorithm for the case of scenes with semi-transparent objects. These constitute a particular case of non-Lambertian scenes containing semitransparent materials that are positioned at different depths. For such scenes, textures of different depth layers are fused in the captured views. Reconstruction methods based on depth estimation (such as [6]) fail on such scenes since a point in the scene (or in a particular view) on a semitransparent object cannot be associated with a unique depth value and therefore a reliable depth map cannot be estimated. On the contrary, the proposed reconstruction method is based on regularization in a linear space of functions, thus one can expect a good reconstruction quality for a scene consisting of depths layers not only occluding each other, but also being fused in the captured views, as in the case of semitransparent materials.

For the evaluation of the proposed method for scenes containing semitransparent objects we created two synthetic scenes made in Blender [48]. The corresponding two densely sampled LFs, both with $d_{max} = 32$px, have been generated: the first scene is purely Lambertian and contains no semi-transparent objects, while the second scene is the same as the first one with the addition of a semi-transparent plane in front. One view from each scene, as rendered in blender, is shown in Fig. 10(a). This figure also shows the performance of the proposed method versus SGBM. For the first scene, the differences between the reconstructed views are negligible, while for the second scene, the proposed method generates better results. The same trend can also be noticed in the EPI images. An example is given in Fig. 10(b). As seen, the proposed method preserves better the semitransparent property of overlapping EPI lines.

4.4 Required Number of Scales

In (5) we gave the relation between the number of scales J and the maximal disparity d_{max}. In this section we analyze the behavior of the reconstruction algorithm for varying decimation factors and varying number of transform scales. The evaluation has been done for the same synthetic scenes as in Section 4.3. Fig. 11 summarizes the obtained results. It is important to mention that the performance of the reconstruction shows a direct correlation between the decimation factor and the number of scales of the shearlet transform. The relation confirms the importance of selecting $J \geq \lceil \log_2 d_{max} \rceil$ number of scales. Choosing higher number of scales improves the reconstruction results, in some cases only marginally. The same trend can be observed for the scene with semi-transparent object (Fig. 11 (b)). However, in this case the proposed method returns significantly better performance for high decimation factor.

4.5 Nonuniform Sampling

All so-far experimental settings assumed equidistant camera and uniformly downsampled number of views. However, as commented earlier, the proposed method is not limited to such sampling strategy. Indeed, it can process nonuniformly sampled LFs by properly interpreting the corresponding EPI slices as being on sampling positions of a densely sampled LF with the maximum disparity between all adjacent views being less than or equal to d_{max}. While uniform sampling has to be favoured because it provides the least number of capturing positions for a given fixed d_{max}, the non-uniform sampling case might arise in some capture settings and therefore is worth discussing it.

Again the scene with the semi-transparent front object as in Fig. 10 has been used. Two different experimental setups

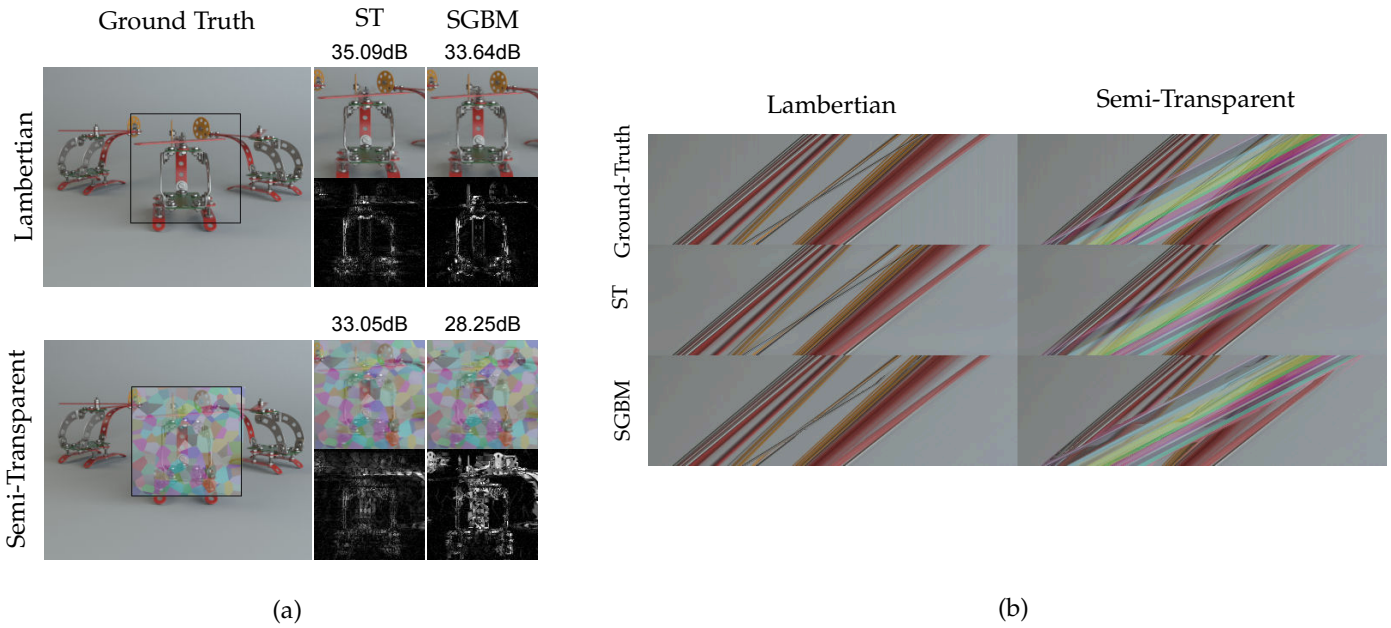


Fig. 10. (a) Considered scenes with and without semi-transparent plane in front. Reconstruction results are presented using the proposed method ($ST(5)$) and ($SGBM$). (b) Example of EPI of the scene and corresponding reconstructions.

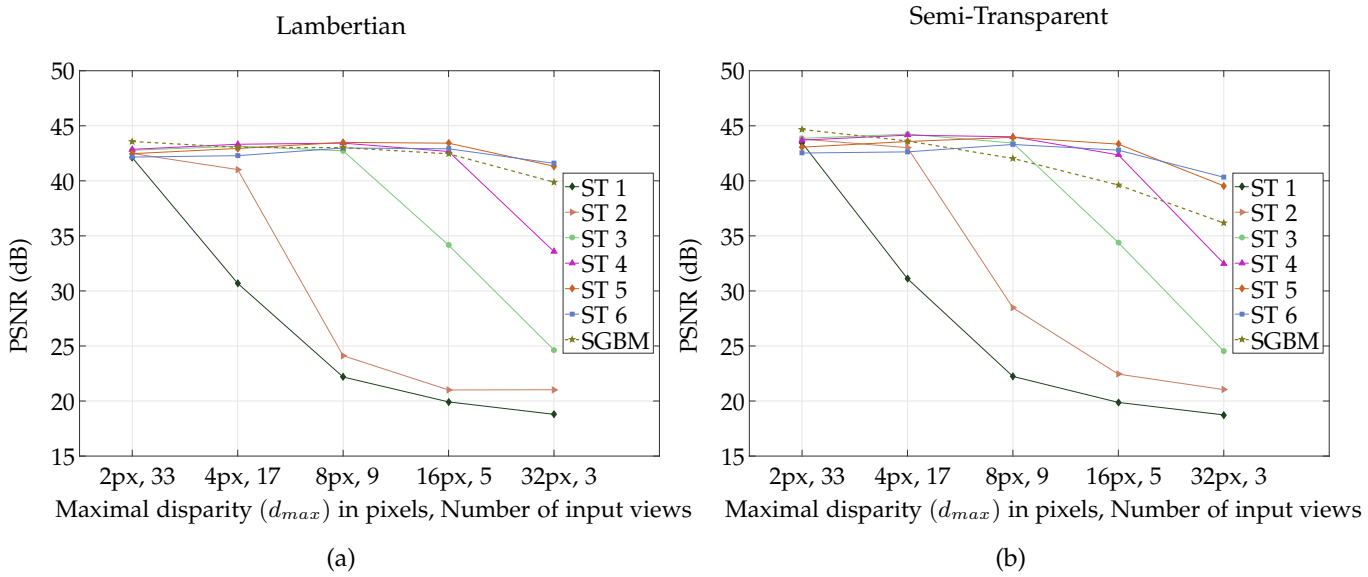


Fig. 11. Evaluation of the proposed method (ST) with different numbers of scaling and reference method using depth estimation [6] (SGBM). Average reconstruction quality of the methods for different decimation levels for the synthetic Lambertian scene (a) and semi-transparent object (b).

are studied and summarized in Fig. 12. In the first setup, the scene has been sampled at 5 equidistant positions (see Fig. 12(a)), which leads to $d_{max} = 32$. In the second experiment, namely a nonuniform sampling, we used 8 views positioned at camera positions $(1, 31, 47, 60, 80, 97, 101, 129)$. The distances between adjacent views are different, with $d_{max} \leq 32$. In order to reconstruct such input datasets one has to replace the uniform positions of input rows in the masking matrix H in (2) by the provided nonuniform sampling positions. Following the same approach as in Section 3 we reconstruct all intermediate views. As shown in Fig. 12(a) the reconstruction quality decreases depending on the distance between the reconstructed view and available input views.

In the second setup, the scene has been sampled at 3 equidistant points (see Fig. 12(b)) which leads to $d_{max} = 64$. Reconstruction using $ST(5)$ performs poor due to insufficient number of scales in the shearlet transform. We need to use $ST(6)$ instead. In overall, the experiments show that the method can handle non-uniform setups well.

4.6 Full Parallax

The last tests deal with full parallax imagery. The proposed method is compared with two state of the art methods. The first one is the learning-based view synthesis method (LBVS) proposed in [23]. The second one is the LF reconstruction method presented in [25], which utilizes sparsity

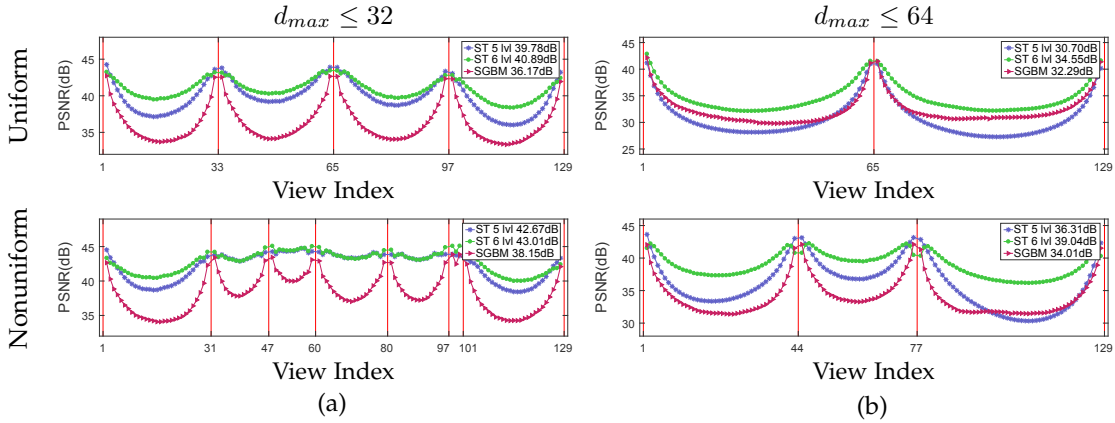


Fig. 12. Comparison of densely sampled LF reconstruction using the proposed method $(ST(5), ST(6))$ with $SGBM$ for scene with semi-transparent object in case of uniform and nonuniform sampling. Vertical red lines are representing sampling positions of the input dataset from densely sampled LFs.

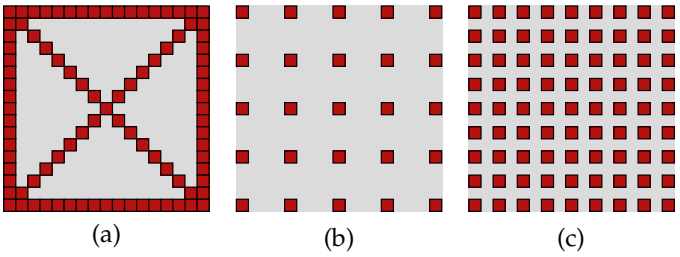


Fig. 13. Sampling pattern where every rectangle represents one view from the LF consisting of 17×17 views. (a) box and two diagonals pattern consisting of 93 views used for method [25]. (b), (c) uniformly decimated setup consisting of 5×5 and 9×9 views respectively.

TABLE 2
Full parallax LF reconstruction quality is presented by average PSNR in dB and speed is given in seconds per view (in parentheses)

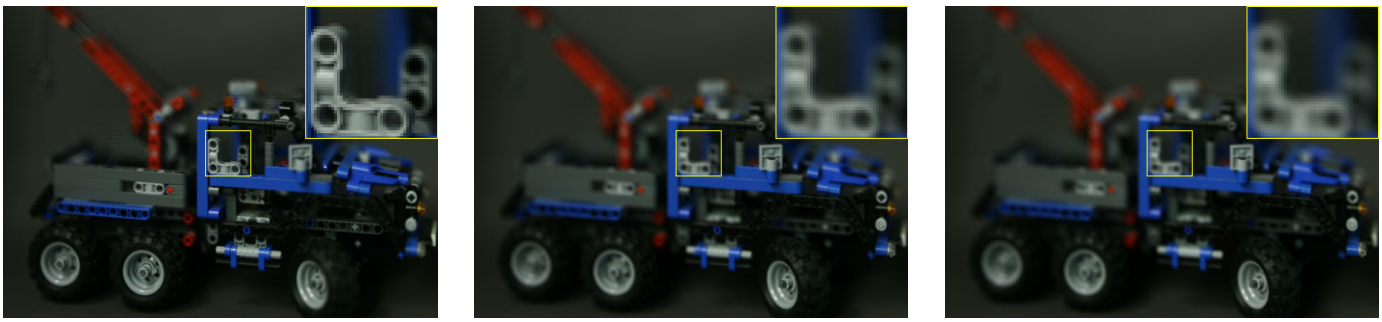
Datasets	Truck	Bunny	Helicopter
SFFT [25]	35.45 (87.2)	38.56 (87.2)	40.87 (87.2)
LBVS 9×9 **[23]**	37.65 (8)	38.16 (10)	38.39 (10)
LBVS 5×5 **[23]**	35.31 (8)	36.45 (10)	36.12 (10)
ST 9×9	40.93 (5)	41.29 (5.3)	46.43 (5.3)
ST 5×5	40.69 (9.2)	39.97 (10)	44.24 (10)
ST (HR) 5×5	40.46 (7.6)	39.57 (8.6)	44.03 (8.6)

of full parallax LF in continuous Fourier domain. The method is claimed useful for the reconstruction of both Lambertian and non-Lambertian scenes. It requires a set of views obtained from a set of 1D viewpoint trajectories [25]. We compared reconstruction results for the dataset $Bunny$ and $Truck$ [45] consisting of 17×17 views, which are representing Lambertian scenes, thus suitable for the proposed method and the method in [25]. In addition we used dataset of a synthetic scene for evaluating both methods in the case of non-Lambertian reflection generated by a semi-transparent plane. Two experiments with different number of input views have been considered for every dataset, one with 25 views and one with 81 views out of 289 for processing with the proposed method. In the case of 25 input views the direct processing and the HR processing as presented in Section 3.2 have been employed. The method in [25] uses 93 views as input. The view patterns used as inputs for different methods are illustrated in Fig. 13. The average PSNR for reconstructed views is presented in Table 2. In the table, computation times per one view for all experiments are presented. For the proposed method these are based on the GPU setting described in the beginning of Section 4. As seen in the table, the proposed HR approach decreases the computation time by about 15% compared to the direct computation for the price of a rather small loss of average reconstruction quality. For the method SFFT [25] and LBVS [23], the computations have been employed on CPU using parallelization with 36 cores. Reconstruction using SFFT takes considerably longer time, e.g. the dataset $Bunny$ was processed overall for about 7 hours to obtain all intermediate 17×17 views. The method presented in [23] considers processing every 4 adjacent views from the input datasets to synthesis intermediate views. An available implementation of the method with already trained neural networks was used in order to obtain results for the datasets with 9×9 and 5×5 views. Examples of reconstructed views with difference maps with respect to ground truth are shown in Fig. 16. While the method from [25] shows capability of reconstructing intermediate views of the scene with semi-transparent objects, our proposed approach seems to perform better also for this case.

One of the applications of full parallax LF is to construct digitally refocused images in post-processing. Fig. 14 shows digitally refocused images corresponding to the central view for differently sampled LFs. As expected, the lack of available views results in strong artifacts in the synthesized refocused image Fig. 14 (a) where only 5×5 subset of views is used, while for the up-sampled (reconstructed) LF consisting of 49×49 views, small disparity between the reconstructed views causes smooth blurring in the refocused image areas. Fig. 14 (c) presents the result of similar refocusing for the original dataset Fig. 14 (b).

5 Conclusions

We have presented a method for reconstructing densely sampled LF from a small number of rectified multiview images taken with a wide baseline. The reconstructed LF bears the property that the disparity between adjacent views



(a) (b) (c)

Fig. 14. Example of refocused images generated from differently sampled dataset $Truck$ [45] using linear interpolation for shearing operation. (a) Refocused image generated for central view using 5×5 views from original dataset, every 4-th view has been chosen. (b) Refocused image generated using all 17×17 views. (c) Refocused image generated from reconstructed LF (49×49 views) based on decimated LF (5×5 views).

is 1px at most while the input views can be with quite high disparity. The method utilizes a sparse representation of the underlying EPIs in shearlet domain and employs an iterative regularized reconstruction. We have constructed a shearlet frame specifically for the case of EPIs and proposed an adaptive tuning for the parameter controlling the convergence in the iterative procedure. Experiments with various datasets compare our method favorably against the MPEG's DERS+VSRS, the state of the art SGBM and the state of the art in IBR for full parallax reconstruction. The method is particularly successful when dealing with non-Lambertian scenes consisting of semi-transparent objects. The method reconstructs all LF views and therefore can be used in applications which require densely sampled views such as refocusing, wide field of view LF displays and digital holographic printing.

As the regularization constraints are limited within the viewing frustum, the frame elements are also spatially concentrated there. Therefore, the LF reconstruction offers only some limited extrapolation due to the elements found near the frustum border. The extrapolation problem can be further addressed by analyzing the parameters of the frame elements near the borders in terms of their scale and directional indexes and generating similar elements by proper translation. This is a topic of future research.

Although the implementation of the algorithm reported in this paper is limited to scenes with Lambertian properties or non-Lambertian scenes generated by semi-transparent objects, it is possible to extend the algorithm such that it will be able to reconstruct reflective non-Lambertian scenes as well. This will, primarily, requires modification of the bases used in reconstruction since different parts of the frequency domain have to be covered, in comparison to the Lambertian case. Also, the regularization procedure has to be tuned to better handle the case of conflicting directions, which might arise from reflective non-Lambertian scenes. This extension is a topic of future research.

Appendix A
Construction of Compactly Supported Shearlet System

The construction of compactly supported shearlet frame elements starts with defining a 1-D multi-resolution analysis with scaling and wavelet functions $\phi_1, \psi_1 \in L^2(\mathbb{R})$

$$\phi_1(x_1) = \sum_{n_1 \in \mathbb{Z}} h(n_1)\sqrt{2}\phi_1(2x_1 - n_1)$$
$$\psi_1(x_1) = \sum_{n_1 \in \mathbb{Z}} g(n_1)\sqrt{2}\phi_1(2x_1 - n_1),$$

where $h(n_1)$ and $g(n_1)$ are appropriately-designed half-band filters. The 2-D generator scaling function ϕ is constructed in a separable manner as

$$\phi(x_1, x_2) = \phi_1(x_1)\phi_1(x_2). \tag{6}$$

However, constructing the shearlet generator $\psi(x_1, x_2)$ in a separable manner is not efficient as it would generate an over-redundant frame with poor directional selectivity [49]. A better approach is to utilize a non-separable directional filter [32]. Then, the non separable shearlet generator is defined in Fourier domain as

$$\hat{\psi}(\xi_1, \xi_2) = P(\xi_1/2, \xi_2)\hat{\psi}_1(\xi_1)\hat{\phi}_1(\xi_2),$$

where the trigonometric polynomial P represents a 2D directional fan filter [30] which is used to approximate the 2D non-separable filter with essential support in frequency domain bounded within the region shown in Fig. 17 (a).

Appendix B
Discrete Implementation

Assume the continuous function $f(x), x \in \mathbb{R}^2$ to be reconstructed, is represented by its samples $f_J^d(n), n \in \mathbb{Z}^2$ at the finest (sufficiently large) scale $J \in \mathbb{N}$, i.e.

$$f(x) = \sum_{n \in \mathbb{Z}^2} f_J^d(n) 2^J \phi(2^J x - n),$$

where $\phi(x)$ is defined as in (6).

The shearlet system consists of the functions

$$\psi_{j,k,m}, |k| \leq 2^{j+1}, j = 0, \ldots, J-1,$$

where

$$\psi_{j,k,m}(x) = 2^{j/2}\psi\left(S_k A_{2^j} x - M_{c_j} m\right), \tag{7}$$

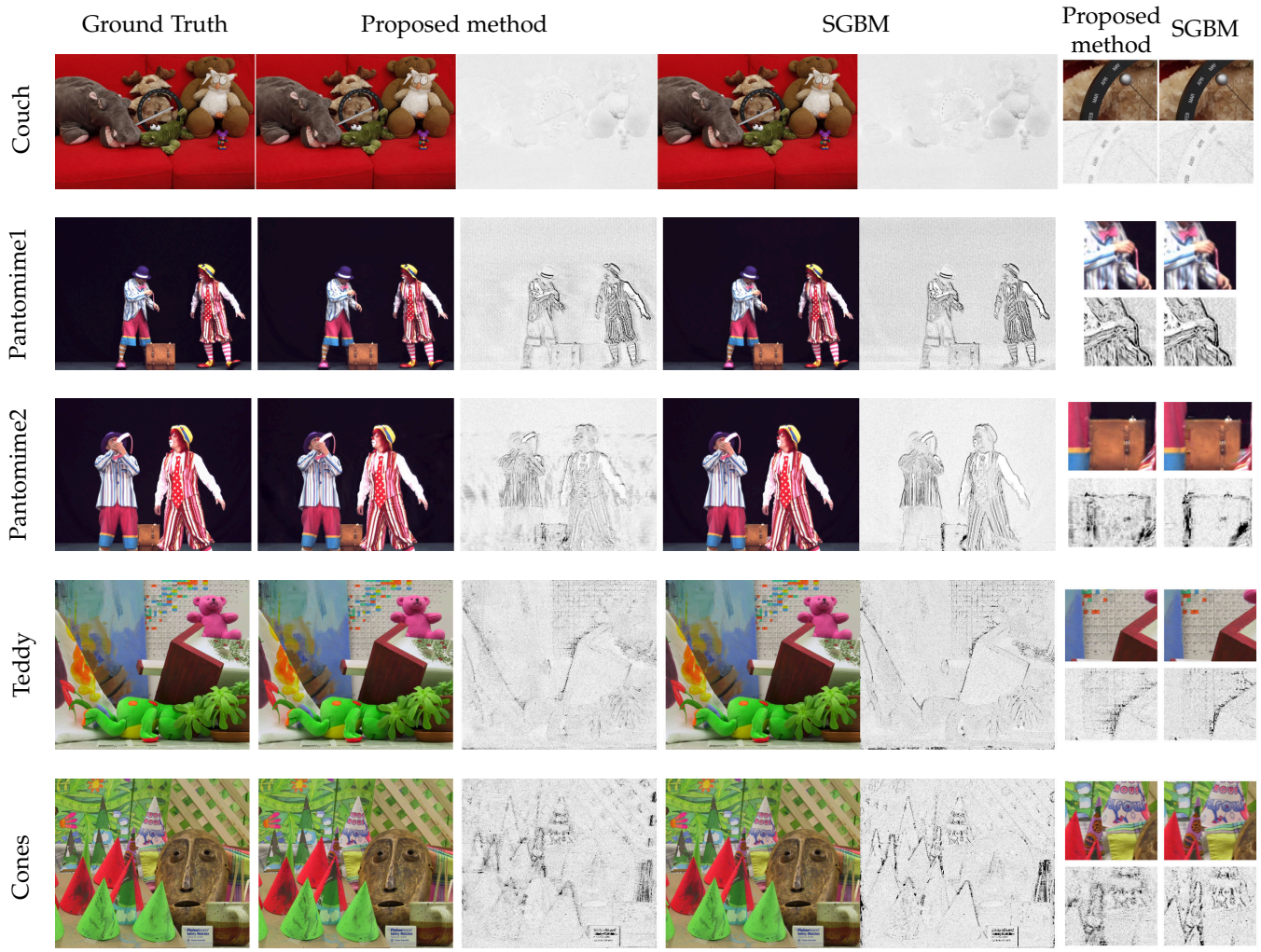


Fig. 15. Examples of the reconstructed views for several different multiview datasets, particularly view number 34 is presented for dataset $Couch$, 18 for $Pantomime1$, 51 for $Pantomime2$ and 4 for $Teddy$ and $Cones$. In the first column are presented ground truth of corresponding reconstructed view. In the following columns are presented proposed and SGBM based reconstruction results together with scaled difference maps. Zoomed in regions from different reconstructed images are presented in the last column.

and $c_j = (c_1^j, c_2^j)$ are sampling constants for translation. Easy to notice

$$\psi_{j,k,m}(x) = \psi_{j,0,m}\left(S_{\frac{k}{2^{j+1}}}x\right). \tag{8}$$

Following the same methodology as in [36], it can be shown that the digital filter corresponding to $\psi_{j,0,m}$ has the form

$$\psi^d_{j,0}(m) = \left(p_j * \left(g_{J-j} \otimes h_{J+1}\right)\right)(m), \tag{9}$$

where $\otimes$ denote tensor product such that $\left(g_{J-j} \otimes h_{J+1}\right)(m) = g_{J-j}(m_1)h_{J+1}(m_2)$, $\{p_j(n)\}_{n\in\mathbb{Z}}$ are the Fourier coefficients of the trigonometric polynomial $P(2^{J-j-1}\xi_1, 2^{J+1}\xi_2)$, $\{h_j(n)\}_{n\in\mathbb{Z}}$ and $\{g_j(n)\}_{n\in\mathbb{Z}}$ are the Fourier coefficients of the respective trigonometric polynomials

$$\hat{h}_j(\xi) = \prod_{k=0,\ldots,j-1} \hat{h}(2^k\xi),$$
$$\hat{g}_j(\xi) = \hat{g}(2^{j-1}\xi)\hat{h}_{j-1}(\xi)$$

and $\hat{h}_0 \equiv 1$. Fig. 17 (b) illustrates the frequency responses of the digital filters h_j, g_j for $j = 1, \ldots, 4$.

The shear transform $S_{k2^{-j}}, j \in \mathbb{N}, k \in \mathbb{Z}$ does not preserve the regular grid $\mathbb{Z}^2$, therefore its digitalization is not trivial. The solution of the problem presented in [49], is to refine the $\mathbb{Z}^2$ grid along the x_1-axis by a factor 2^j. In that case, the grid $2^{-j}\mathbb{Z} \times \mathbb{Z}$ is invariant under the $S_{k2^{-j}}$ transform. Thus, for an arbitrary $r \in l^2(\mathbb{Z}^2)$, the shear transform $S_{k2^{-j}}$ can be implemented as a digital filter

$$S^d_{k2^{-j}}(r) = \left((2^j r_{\uparrow 2^j} *_1 \tau_j)(S_k \cdot) *_1 \bar{\tau}_j\right)_{\downarrow 2^j}, \tag{10}$$

where τ_j represents a digital low-pass filter with normalized cutoff frequency at 2^{-j}, $*_1$ is 1D convolution along x_1 axis and $\uparrow 2^j, \downarrow 2^j$ are upsampling and downsampling operators corresponding to 2^j factor.

Using (7), (9), (10) the discrete filter $\psi^d_{j,k}$ corresponding to $\psi_{j,k,m}$ takes the form

$$\psi^d_{j,k} = (S^d_{k2^{-(j+1)}}(p_j * g_{J-j} \otimes h_{J+1}))(m).$$

The digital filter ϕ^d corresponding to the scaling function ϕ, is constructed in a separable manner $\phi^d = (h_J \otimes h_J)(m)$.

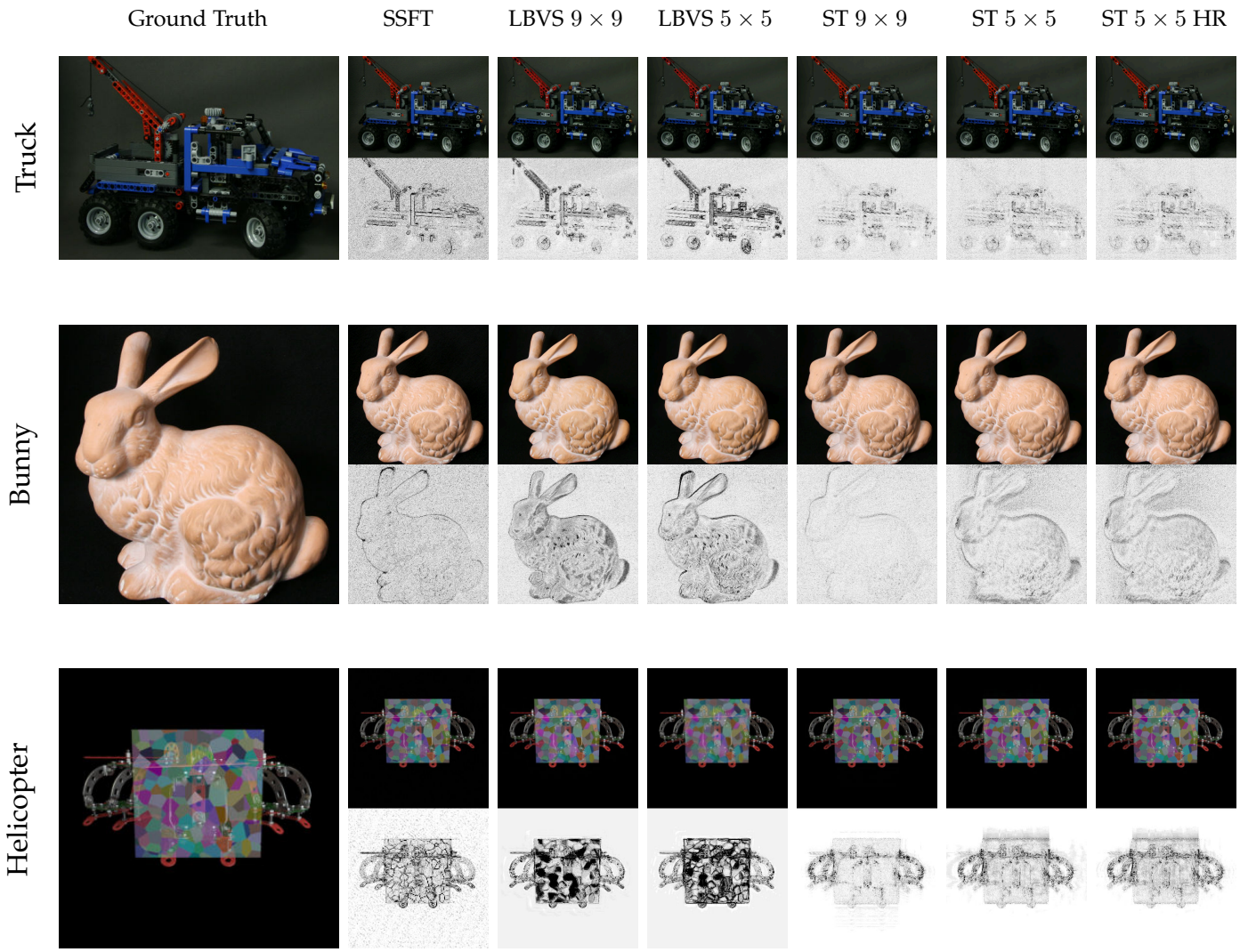


Fig. 16. Reconstruction results for full parallax datasets. Obtained results are presented with difference maps for the methods SSFT [25], LBVS [23] and the proposed method using direct and hierarchic processing order.

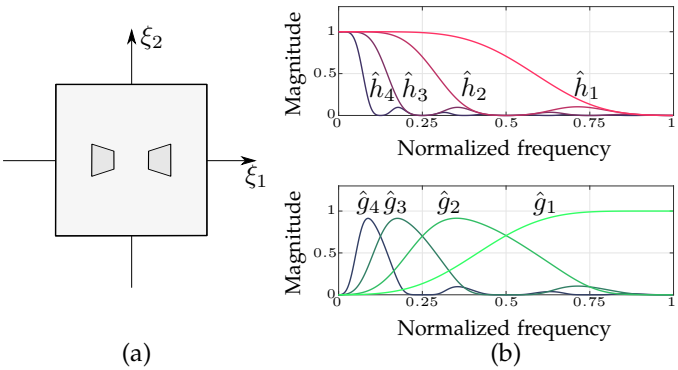


Fig. 17. (a) Shearlet support in Fourier domain; (b) Frequency responses of the scaling and wavelet filters $h_j, g_j, j = 1, , 4$.

ACKNOWLEDGMENT

The research leading to these results has received funding from the PROLIGHT-IAPP Marie Curie Action of the People programme of the European Unions Seventh Framework Programme, REA grant agreement 32449 and from the Academy of Finland, grant No. 137012: High-Resolution Digital Holography: A Modern Signal Processing Approach.

REFERENCES

[1] H. Shum, S. Chan, and S. Kang, *Image-Based Rendering*. Springer-Verlag, 2007.

[2] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Int'l J. Computer Vision*, vol. 47, no. 1-3, pp. 7–42, Apr. 2002.

[3] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene Reconstruction from High Spatio-Angular Resolution Light Fields," *ACM Trans. on Graphics*, vol. 32, no. 4, pp. 1–12, Jul. 2013.

[4] J. Pearson, M. Brookes, and P. Dragotti, "Plenoptic Layer-Based Modeling for Image Based Rendering," *IEEE Trans. Image Processing*, vol. 22, no. 9, pp. 3405–3419, Sept. 2013.

[5] S. Wanner and B. Goldluecke, "Variational Light Field Analysis for Disparity Estimation and Super-Resolution," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[6] H. Hirschmuller, "Stereo Processing by Semiglobal Matching and Mutual Information," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[7] S. N. Sinha, D. Scharstein and R. Szeliski, "Efficient High-Resolution Stereo Matching Using Local Plane Sweeps," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1582–1589, June 2014.

[8] G. Zhang, J. Jia, T. Wong, and H. Bao, "Consistent Depth Maps Recovery from a Video Sequence," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 974–988, June 2009.

[9] S. Wanner and B. Goldluecke, "Globally Consistent Depth Labeling of 4D Light Fields," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 41–48, June 2012.

[10] E. Adelson and J. Bergen, "The Plenoptic Function and The Elements of Early Vision", *Computational Models of Visual Processing*, vol. 1, MIT Press, 1991.

[11] M. Levoy and P. Hanrahan, "Light field rendering," *Proc. ACM SIGGRAPH*, pp. 31–42, 1996.

[12] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The Lumigraph," *Proc. ACM SIGGRAPH*, pp. 43–54, 1996.

[13] Z. Lin and H.-Y. Shum, "A Geometric Analysis of Light Field Rendering," *Int'l J. of Computer Vision*, vol. 58, no. 2, pp. 121–138, 2004.

[14] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic Sampling," *Proc. ACM SIGGRAPH*, pp. 307–318, 2000.

[15] R. Ng, "Fourier Slice Photography," *Proc. ACM SIGGRAPH*, vol. 24, no. 3, pp. 735–744, July 2005.

[16] I. Tosic and K. Berkner, "Light Field Scale-Depth Space Transform for Dense Depth Estimation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops* (CVPRW), pp. 441–448, June 2014.

[17] K. Yücer, A. Sorkine-Hornung, O. Wang, and O. Sorkine-Hornung, "Efficient 3D Object Segmentation from Densely Sampled Light Fields with Applications to 3D Reconstruction," *ACM Trans. on Graphics*, , vol. 35, no. 3, 2016.

[18] M. Tanimoto, "Overview of FTV (free-viewpoint television)," *Proc. IEEE Conf. Multimedia and Expo* (ICME 2009), pp. 1552–1553, June 2009.

[19] J. Jurik, T. Burnett, M. Klug, and P. Debevec, "Geometry-Corrected Light Field Rendering for Creating a Holographic Stereogram," *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops* (CVPRW), pp. 9–13, 2012.

[20] J. Stewart, J. Yu, S. J. Gortler, and L. McMillan, "A New Reconstruction Filter for Undersampled Light Fields," *Proc. 14th Eurographics Workshop on Rendering* (EGRW '03), pp. 150–156, 2003.

[21] D. C. Schedl, C. Birklbauer, and O. Bimber, "Directional Super-Resolution by Means of Coded Sampling and Guided Upsampling," *Proc. IEEE Conf. Computational Photography* (ICCP), pp. 1–10, 2015.

[22] S. Heber and T. Pock, "Convolutional Networks for Shape From Light Field," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2016.

[23] N. K. Kalantari, T.-C. Wang and R. Ramamoorthi, "Learning-Based View Synthesis for Light Field Cameras," *ACM Trans. on Graphics*, vol. 35, no. 6, 2016.

[24] R. Bolles, H. Baker, and D. Marimont, "Epipolar-Plane Image Analysis: An Approach to Determining Structure from Motion," *Int'l J. of Computer Vision*, vol. 1, no. 1, pp. 7–55, 1987.

[25] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light Field Reconstruction Using Sparsity in the Continuous Fourier Domain," *ACM Trans. on Graphics*, vol. 34, no. 1, 2014.

[26] E. J. Candes, D. L. Donoho *et al.*, *Curvelets: A Surprisingly Effective Nonadaptive Representation for Objects with Edges*, Stanford University, 1999.

[27] E. J. Candès and D. L. Donoho, "New Tight Frames of Curvelets and Optimal Representations of Objects with Piecewise C^2 Singularities," *Comm. Pure Appl. Math.*, vol. 57, no. 2, pp. 219–266, 2004.

[28] G. Kutyniok *et al.*, *Shearlets: Multiscale Analysis for Multivariate Data*, Birkhuser Basel, 2012.

[29] D. L. Donoho, "Sparse Components of Images and Optimal Atomic Decompositions," *Constructive Approximation*, vol. 17, no. 3, pp. 353–382, 2001.

[30] M. Do and M. Vetterli, "The Contourlet Transform: An Efficient Directional Multiresolution Image Representation," *IEEE Trans. Image Processing*, vol. 14, no. 12, pp. 2091–2106, Dec. 2005.

[31] G. Easley, D. Labate, and W.-Q. Lim, "Optimally Sparse Image Representations Using Shearlets," *Proc. Fortieth Asilomar Conf. Signals, Systems and Computers* (ACSSC '06), pp. 974–978, Oct. 2006.

[32] G. Kutyniok and W.-Q. Lim, "Compactly Supported Shearlets are Optimally Sparse," *J. of Approximation Theory*, vol. 163, no. 11, pp. 1564 – 1589, 2011.

[33] S. Hauser and J. Ma, "Seismic Data Reconstruction via Shearlet-Regularized Directional Inpainting," http://www.mathematik.uni-kl.de/uploads/tx_sibibtex/seismic.pdf, May 2012.

[34] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Image Based Rendering Technique via Sparse Representation in Shearlet Domain," *IEEE Int'l Conf. Image Processing*, pp. 1379–1383, Sept. 2015.

[35] C.-K. Liang, Y.-C. Shih, and H. Chen, "Light Field Analysis for Modeling Image Formation," *IEEE Trans. Image Processing*, vol. 20, no. 2, pp. 446–460, Feb. 2011.

[36] G. Kutyniok, W.-Q. Lim, and R. Reisenhofer, "ShearLab 3D: Faithful Digital Shearlet Transforms Based on Compactly Supported Shearlets," *ACM Trans. on Mathematical Software*, vol. 42, no. 1, 2015.

[37] S. Mallat, *A Wavelet Tour of Signal Processing : The Sparse Way*, 3rd ed., Academic Press, 2008.

[38] J.-L. Starck, Y. Moudden, J. Bobin, M. Elad, and D. L. Donoho, "Morphological Component Analysis," *Proc. SPIE 5914 Wavelets XI*, 59140Q, May 2005.

[39] J. Fadili, J.-L. Starck, M. Elad, and D. Donoho, "Mcalab: Reproducible Research in Signal and Image Decomposition and Inpainting," *IEEE Computing in Science & Engineering*, vol. 12, no. 1, pp. 44–63, 2010.

[40] T. Blumensath and M. Davies, "Normalized iterative hard thresholding: Guaranteed stability and performance," *IEEE J. Sel. Topics Signal Processing*, vol. 4, no. 2, pp. 298–309, April 2010.

[41] S. Smirnov, A. Gotchev, and M. Hannuksela, "A Disparity Range Estimation Technique for Stereo-Video Streaming Applications," *IEEE Int'l Conf. Multimedia and Expo Workshops* (ICMEW), pp. 1–4, July 2013.

[42] S. Häuser and G. Steidl, "Fast Finite Shearlet Transform," *arXiv:1202.1773*, 2014.

[43] S. Toyohiro, "Nagoya University Multi-View Sequences," *http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data.*

[44] D. Scharstein and R. Szeliski, "High-Accuracy Stereo Depth Maps Using Structured Light," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 195–202, June 2003.

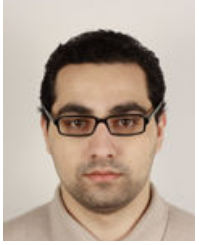
[45] V. Vaish and A. Adams, "The (New) Stanford Light Field Archive," *http://lightfield.stanford.edu*, 2008.

[46] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Depth Estimation Reference Software (DERS 5.0)," *ISO/IEC JTC1/SC29/WG11 M*, vol. 16923, 2009.

[47] M. Tanimoto, T. Fujii, and K. Suzuki, "View Synthesis Algorithm in View Synthesis Reference Software 2.0 (VSRS 2.0)," *ISO/IEC JTC1/SC29/WG11 M*, vol. 16090, 2009.

[48] Blender Online Community, "Blender - a 3d Modelling and Rendering Package", http://www.blender.org.

[49] W.-Q. Lim, "Nonseparable shearlet transform," *IEEE Trans. Image Processing*, vol. 22, no. 5, pp. 2056–2065, May 2013.



Suren Vagharshakyan Suren Vagharshakyan received the MSc in mathematics from Yerevan State University (2008). He is a PhD student at the Department of Signal Processing at Tampere University of Technology since 2013. His research interests are in the area of light field capture and reconstruction.

Robert Bregovic Robert Bregović received the MSc in electrical engineering from University of Zagreb (1998) and the Dr.Sc.(Tech) in information technology from Tampere University of Technology (2003). He has been working at Tampere University of Technology since 1998. His research interests include the design and implementation of digital filters and filterbanks, multirate signal processing, and topics related to acquisition, processing/modeling and visualization of 3D content.

Atanas Gotchev Atanas Gotchev received the M.Sc. degrees in radio and television engineering (1990) and applied mathematics (1992) and the Ph.D. degree in telecommunications (1996) from the Technical University of Sofia, and the D.Sc.(Tech.) degree in information technologies from the Tampere University of Technology (2003). He is a Professor at Tampere University of Technology. His recent work concentrates on algorithms for multisensor 3-D scene capture, transform-domain light-field reconstruction, and Fourier analysis of 3-D displays.