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

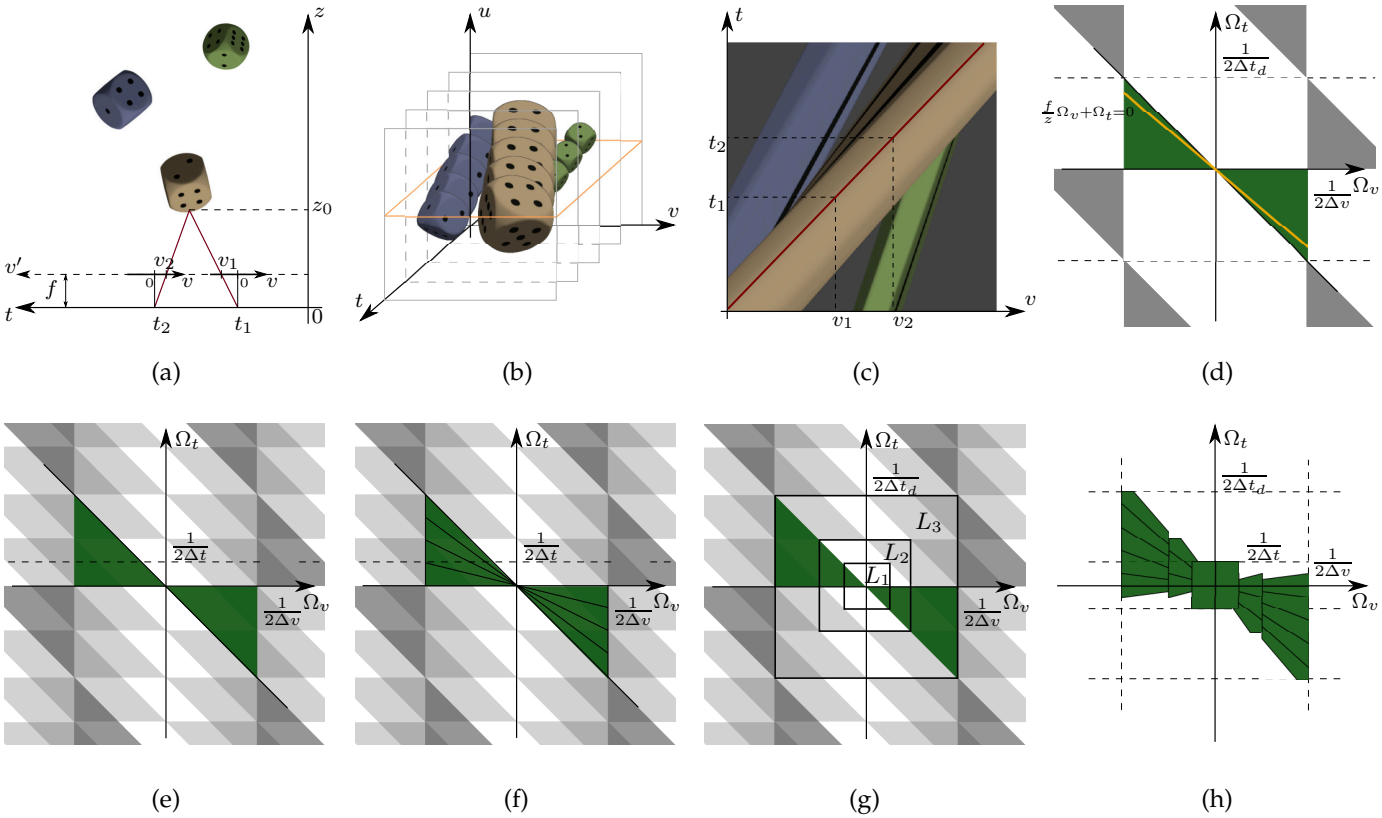


Fig. 1. Epipolar-plane image (EPI) formation and its frequency domain properties. (a) Capturing setup and EPI formation, a scene point is observed by two pinhole cameras positioned at t_1, t_2 at image coordinates v_1 and v_2 respectively; (b) Stack of captured images; an epipolar plane is highlighted 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

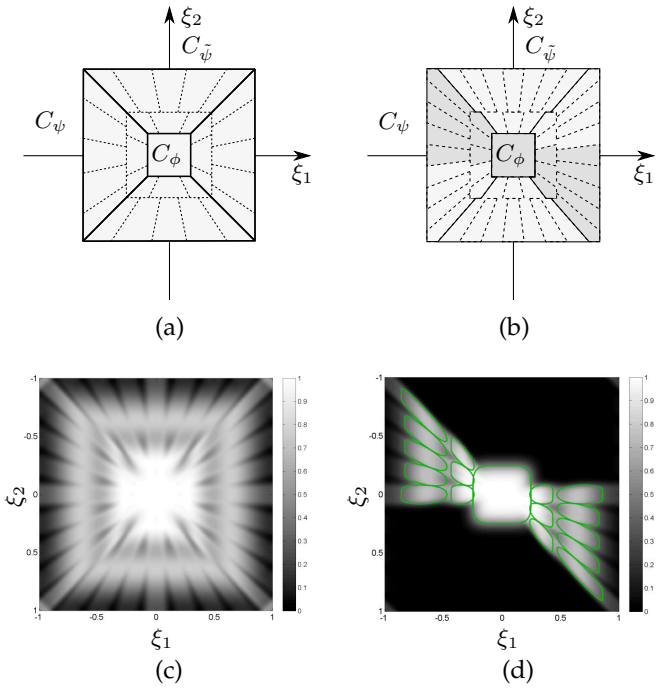


Fig. 2. (a) Frequency plane tilting by shearlet transform. C_ψ,$C_{\tilde{\psi}}$ are cone-like regions and C_ϕ is low-frequency region. (b) Desirable frequency domain tilting by proposed reconstruction algorithm. Gray color region includes transform elements used for reconstruction; other transform 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

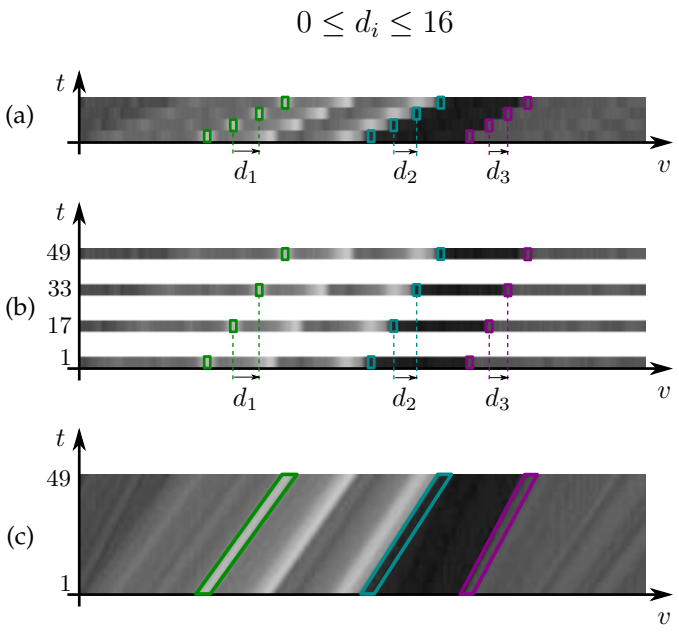


Fig. 3. The given 4 views with maximal disparity 16px between consecutive views are interpreted as every 16th view in the target densely sampled LF. (a) EPI for coarsely sampled LF over t-axis; (b) corresponding partially defined densely sampled EPI; (c) ground truth densely sampled EPI. Three different 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

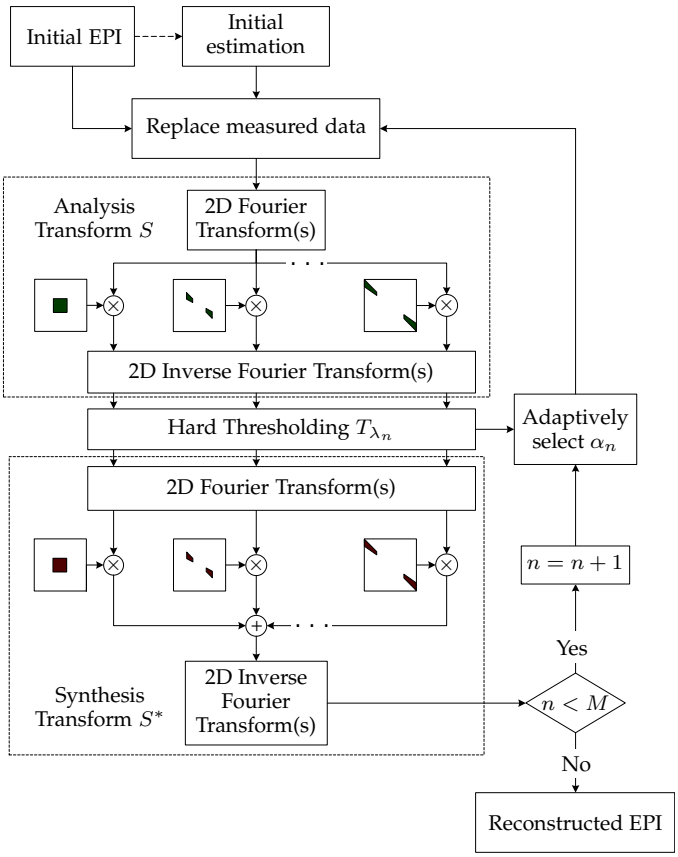


Fig. 4. Diagram of the EPI reconstruction algorithm.

$y, y^* \in \mathbb{R}^{N^2}, H \in \mathbb{R}^{N^2 \times N^2}$. 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

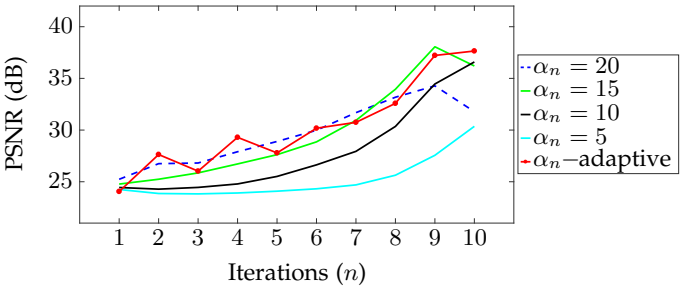


Fig. 5. Example of reconstruction performance dependence on choice of acceleration coefficients α_n. For constant value for all iterations $\alpha_n = \alpha$, increasing α brings accelerating convergence. After some value, reconstruction starts to diverge ($\alpha = 20$).

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

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

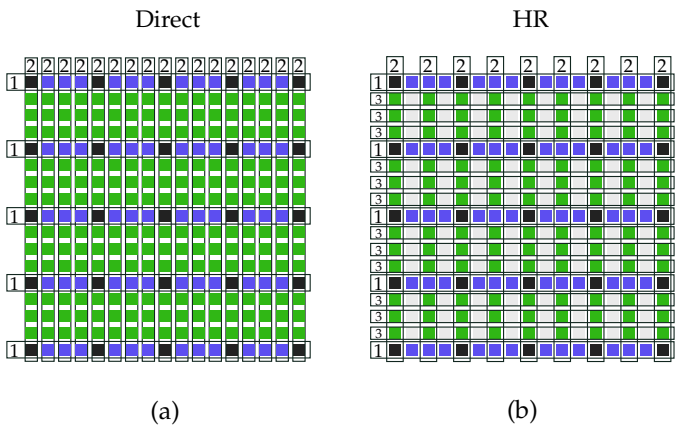


Fig. 6. Array of 17×17 views considered for reconstruction using 5×5 views highlighted with black color. (a) Direct reconstruction method. (b) Hierarchic order of reconstruction (HR).

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

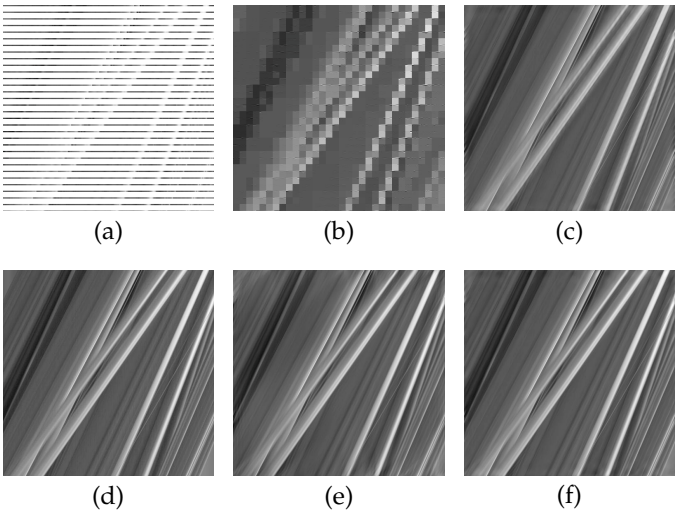


Fig. 7. (a) Input for reconstructing densely sampled EPI where only every 16-th row is available. (d) Densely sampled ground truth EPI. Reconstruction results using different transform are shown as follows (b) Haar 18.83dB, (c) Shearlab [36] 29.27dB, (e) FFST [42] 37.27dB, (f) Proposed modified shearlet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

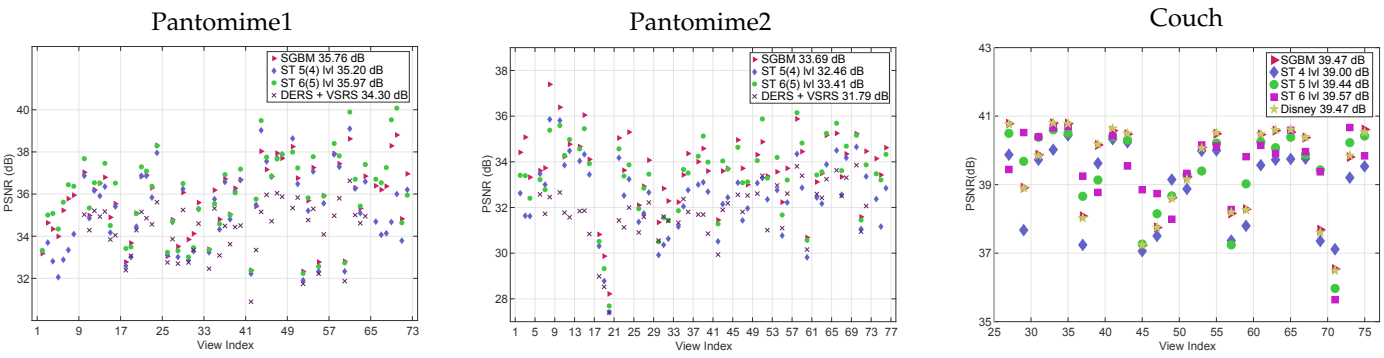


Fig. 8. Reconstruction results for different multiview datasets, error shown in PSNR for reconstructed views.

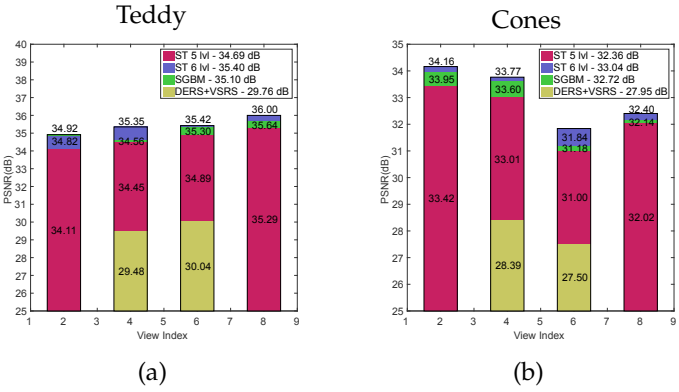


Fig. 9. Reconstruction quality for datasets (a) $Teddy$ and (b) $Cones$. Evaluation has been performed for proposed methods $ST(5), ST(6)$, DERS+VSRS, SGBM. Average PSNR of all reconstructed views presented in legend of the figures.

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

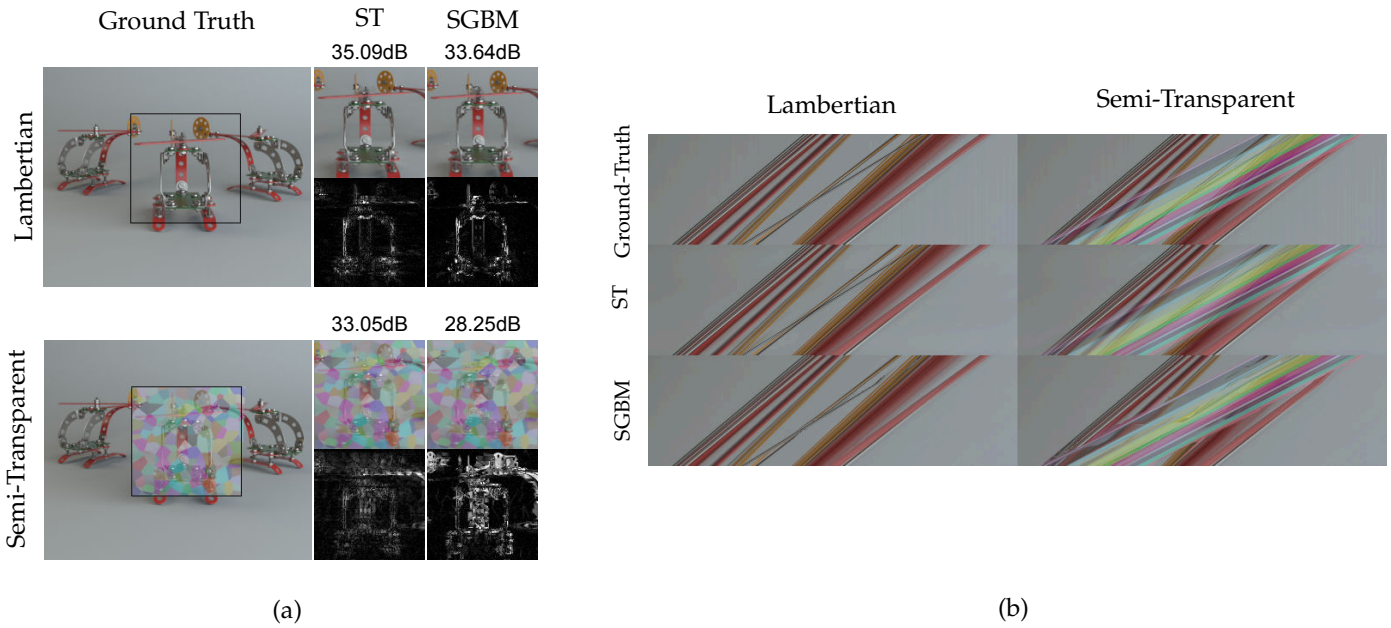


Fig. 10. (a) Considered scenes with and without semi-transparent plane in front. Reconstruction results are presented using the proposed method ($ST(5)$) and ($SGBM$). (b) Example of EPI of the scene and corresponding reconstructions.

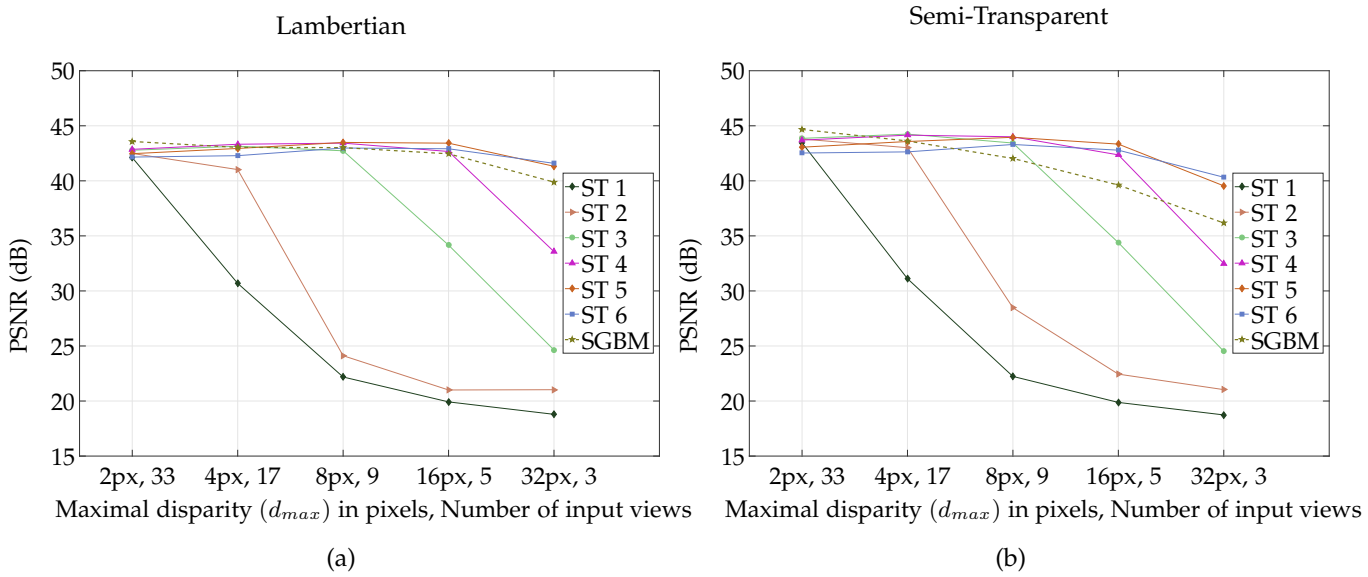


Fig. 11. Evaluation of the proposed method (ST) with different numbers of scaling and reference method using depth estimation [6] (SGBM). Average reconstruction quality of the methods for different decimation levels for the synthetic Lambertian scene (a) and semi-transparent object (b).

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

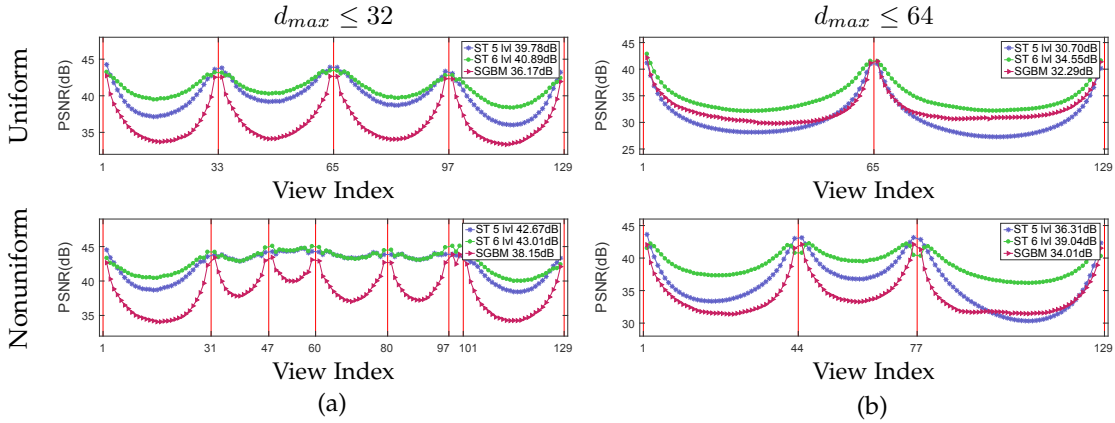


Fig. 12. Comparison of densely sampled LF reconstruction using the proposed method $(ST(5), ST(6))$ with $SGBM$ for scene with semi-transparent object in case of uniform and nonuniform sampling. Vertical red lines are representing sampling positions of the input dataset from densely sampled LFs.

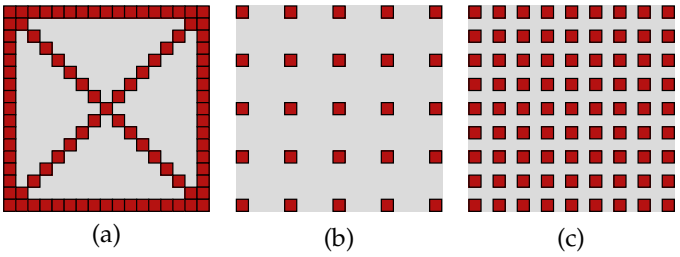


Fig. 13. Sampling pattern where every rectangle represents one view from the LF consisting of 17×17 views. (a) box and two diagonals pattern consisting of 93 views used for method [25]. (b), (c) uniformly decimated setup consisting of 5×5 and 9×9 views respectively.

TABLE 2
Full parallax LF reconstruction quality is presented by average PSNR in dB and speed is given in seconds per view (in parentheses)

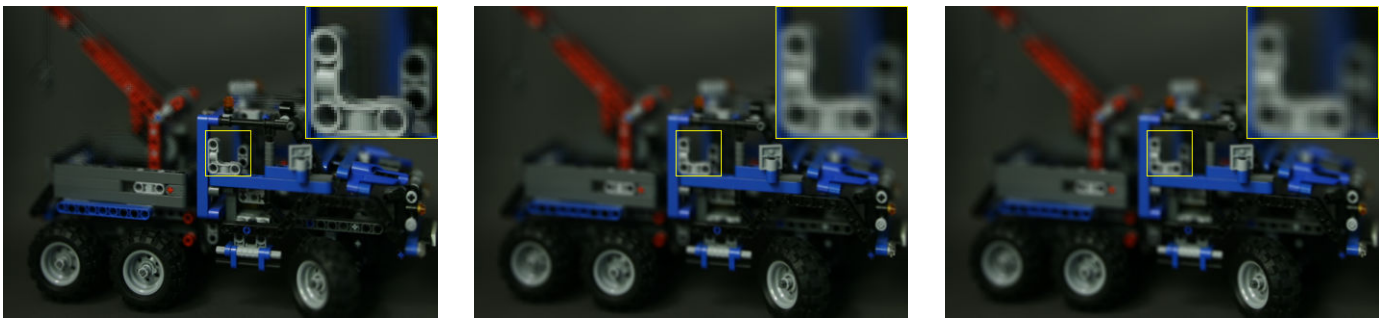
Datasets	Truck	Bunny	Helicopter
SFFT [25]	35.45 (87.2)	38.56 (87.2)	40.87 (87.2)
LBVS 9×9 **[23]**	37.65 (8)	38.16 (10)	38.39 (10)
LBVS 5×5 **[23]**	35.31 (8)	36.45 (10)	36.12 (10)
ST 9×9	40.93 (5)	41.29 (5.3)	46.43 (5.3)
ST 5×5	40.69 (9.2)	39.97 (10)	44.24 (10)
ST (HR) 5×5	40.46 (7.6)	39.57 (8.6)	44.03 (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

$$\psi_{j,k,m}(x) = 2^{j/2}\psi\left(S_k A_{2^j} x - M_{c_j} m\right), \tag{7}$$

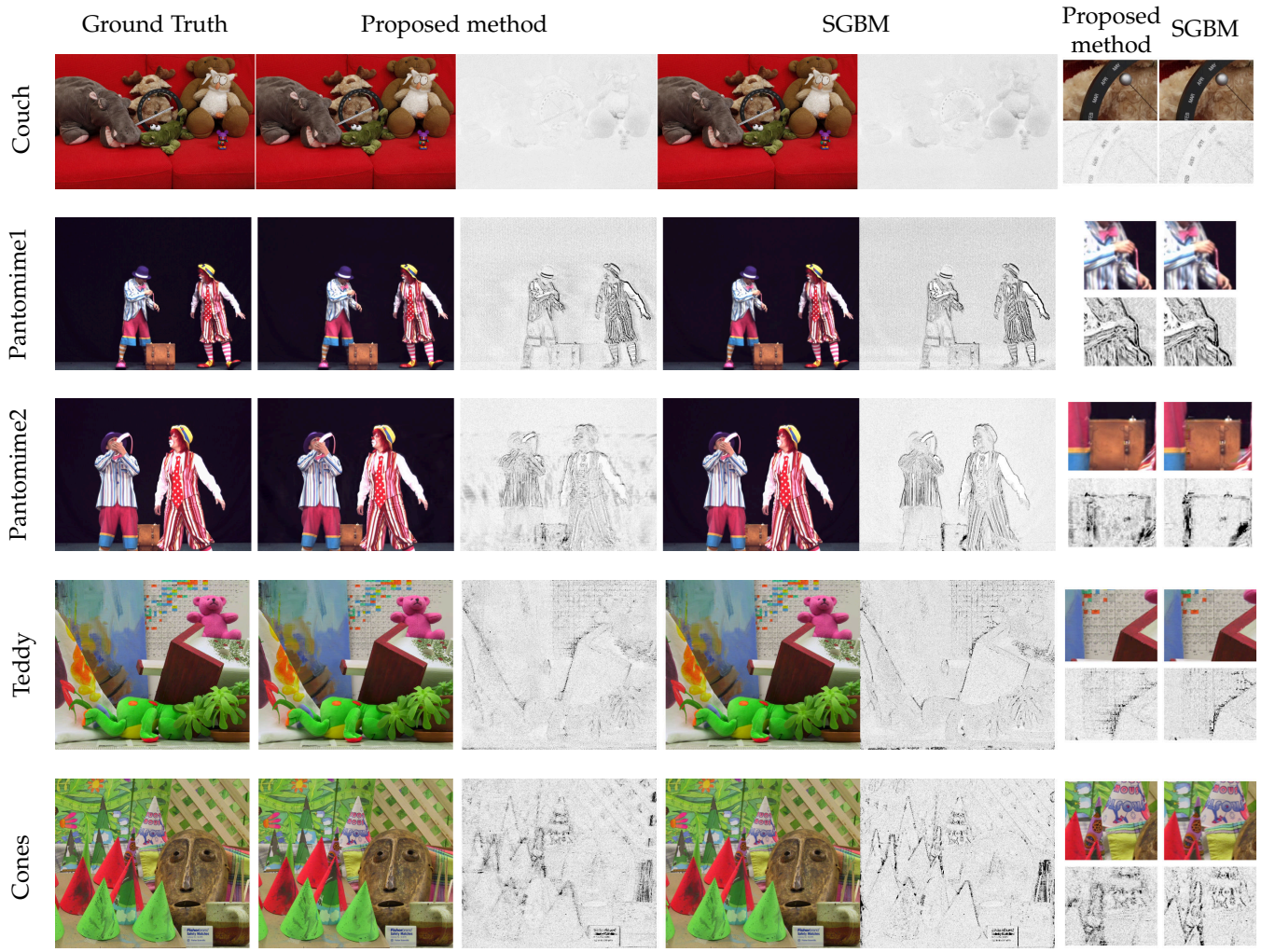


Fig. 15. Examples of the reconstructed views for several different multiview datasets, particularly view number 34 is presented for dataset $Couch$, 18 for $Pantomime1$, 51 for $Pantomime2$ and 4 for $Teddy$ and $Cones$. In the first column are presented ground truth of corresponding reconstructed view. 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 $\phi^d = (h_J \otimes h_J)(m)$.

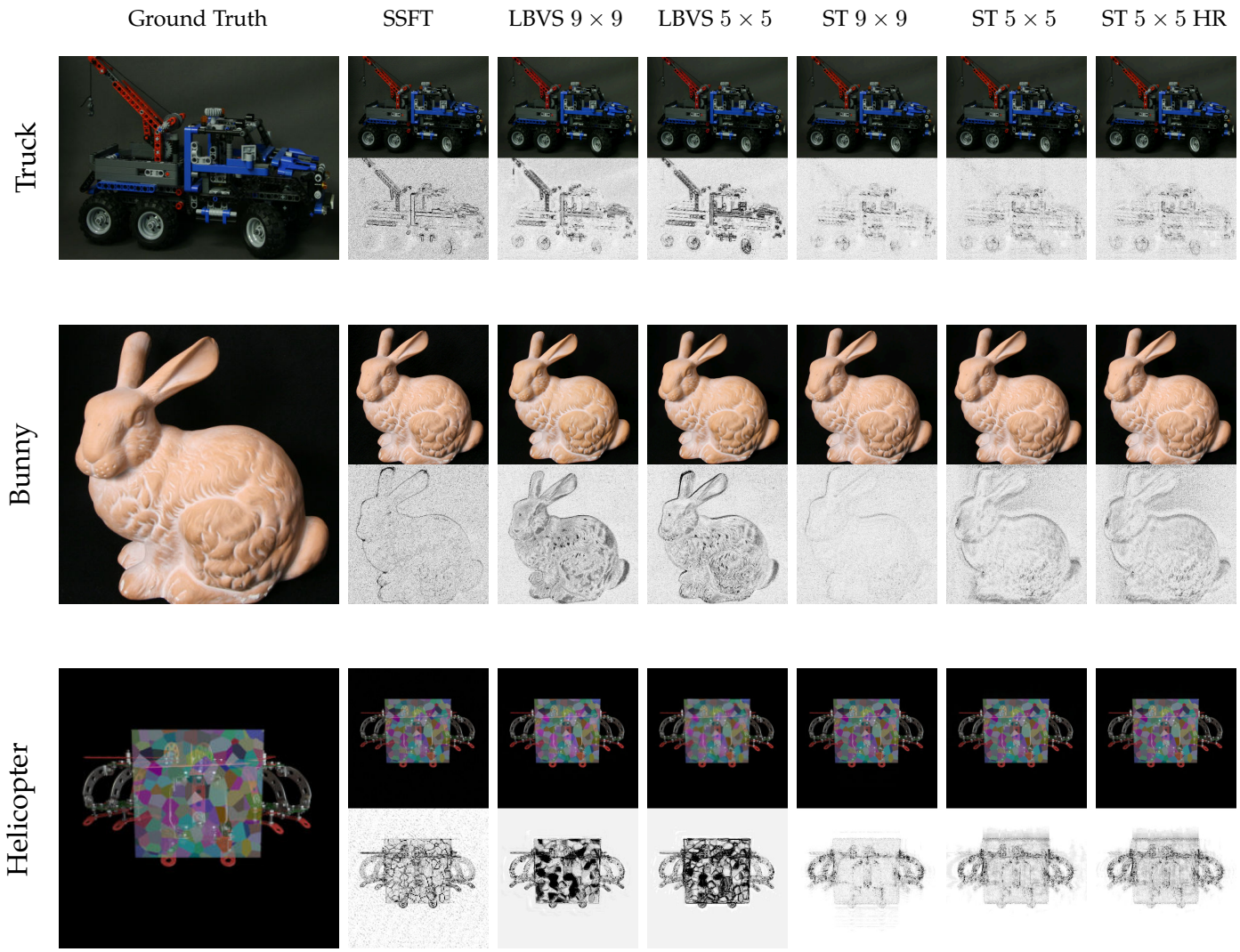


Fig. 16. Reconstruction results for full parallax datasets. Obtained results are presented with difference maps for the methods SSFT [25], LBVS [23] and the proposed method using direct and hierarchic processing order.

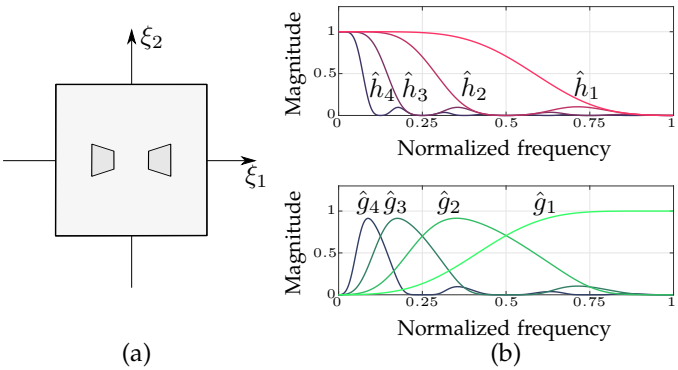


Fig. 17. (a) Shearlet support in Fourier domain; (b) Frequency responses of the scaling and wavelet filters $h_j, g_j, j = 1, , 4$.

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.

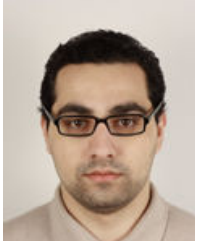
[45] V. Vaish and A. Adams, "The (New) Stanford Light Field Archive," *http://lightfield.stanford.edu*, 2008.

[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.