MAST: MASK-ACCELERATED SHEARLET TRANSFORM FOR DENSELY-SAMPLED LIGHT FIELD RECONSTRUCTION

Yuan Gao, Robert Bregovic, Atanas Gotchev and Reinhard Koch

Kiel University, Germany
{yga, rk}@informatik.uni-kiel.de

Tampere University, Finland
{robert.bregovic, atanas.gotchev}@tuni.fi

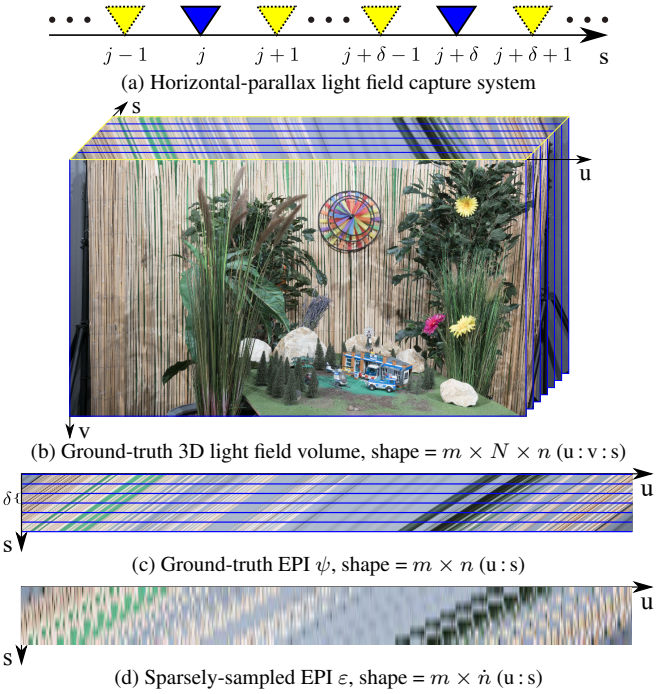
ABSTRACT

Shearlet Transform (ST) is one of the most effective algorithms for the Densely-Sampled Light Field (DSLF) reconstruction from a Sparsely-Sampled Light Field (SSLF) with a large disparity range. However, ST requires a precise estimation of the disparity range of the SSLF in order to design a shearlet system with decent scales and to pre-shear the sparsely-sampled Epipolar-Plane Images (EPIs) of the SSLF. To overcome this limitation, a novel coarse-to-fine DSLF reconstruction method, referred to as Mask-Accelerated Shearlet Transform (MAST), is proposed in this paper. Specifically, a state-of-the-art learning-based optical flow method, FlowNet2, is employed to estimate the disparities of a SSLF. The estimated disparities are then utilized to roughly estimate the densely-sampled EPIs for the sparsely-sampled EPIs of the SSLF. Finally, an elaborately-designed soft mask for a coarsely-inpainted EPI is exploited to perform an iterative refinement on this EPI. Experimental results on nine challenging horizontal-parallax real-world SSLF datasets with large disparity ranges (up to 35 pixels) demonstrate the effectiveness and efficiency of the proposed method over the other state-of-the-art approaches.

Index Terms— View Synthesis, Parallax View Generation, Densely-Sampled Light Field Reconstruction, Shearlet Transform, Mask-Accelerated Shearlet Transform

1. INTRODUCTION

Densely-Sampled Light Field (DSLF) is a discrete representation of the 4D approximation of the plenoptic function parameterized by two parallel planes (camera plane and image plane) [1], where multi-perspective camera views are arranged in such a way that the disparities between adjacent views are less than one pixel [2]. As can be seen in Fig. 1 (a), a horizontal-parallax light field capture system can be considered as a camera moving along the horizontal axis. All the parallax views captured by this camera constitute a ground-truth 3D light field volume as illustrated in Fig. 1 (b). This volume can then be turned into ground-truth Epipolar-Plane Images (EPIs), of which an example is shown in Fig. 1 (c). A Sparsely-Sampled Light Field (SSLF) for this horizontal-parallax light field dataset consists of views with blue borders. The virtual cameras represented by dash-line triangles with yellow color correspond to the target "unknown" views to be



(a) Horizontal-parallax light field capture system

(b) Ground-truth 3D light field volume, shape = $m \times N \times n$ (u : v : s)

(c) Ground-truth EPI ψ, shape = $m \times n$ (u : s)

(d) Sparsely-sampled EPI ε, shape = $m \times \dot{n}$ (u : s)

Fig. 1. Introduction to the DSLF reconstruction problem.

reconstructed, the number of which is decided by the interpolation rate δ. A sparsely-sampled EPI ε from the ground-truth EPI ψ is presented in Fig. 1 (d). The DSLF reconstruction for the SSLF can be treated as reconstructing a densely-sampled EPI f from the sparsely-sampled EPI ε. If the ground-truth EPI ψ is not densely sampled, it will be necessary to downsample the reconstructed densely-sampled EPI f to construct a target EPI with the same size as ψ (see Sect. 4.1).

Shearlet Transform (ST) [3, 4] is extremely effective in reconstructing a densely-sampled EPI from a sparsely-sampled EPI with a large disparity range. This algorithm typically needs to obtain the disparity range of the sparsely-sampled EPI to construct a specifically-tailored universal shearlet system [3, 5] with decent scales. Besides, the sparsely-sampled EPI also needs the disparity range information for shearing and padding in order to be correctly processed by this elaborately-designed shearlet system. Moreover, for DSLF reconstruction from SSLFs with large disparity ranges, this algorithm is prone to be time-consuming due to the high num-

ber of iterations of its iterative thresholding algorithm. Therefore, in this paper, a novel ST-based coarse-to-fine DSLF reconstruction method, referred to as Mask-Accelerated Shearlet Transform (MAST), is proposed to address these two problems. The presented MAST method takes full advantage of a state-of-the-art learning-based optical flow estimation approach, *i.e.* FlowNet2 [6], to estimate the disparities of the whole SSLF for resolving the first problem. In addition, the estimated disparities are also used to roughly restore a densely-sampled EPI from a sparsely-sampled EPI via inverse warping. The iterative estimation refinement algorithm in ST convergences faster by means of an elaborately-designed soft mask for the coarsely-inpainted densely-sampled EPI, thus tackling the second problem. Experimental results demonstrate the superior performance of MAST over the other state-of-the-art DSLF reconstruction methods on nine challenging horizontal-parallax real-world light field datasets with disparity ranges up to 35 pixels.

2. RELATED WORK

High-quality and high-fidelity Virtual Reality (VR) [7] and Free Viewpoint Video (FVV) [8] contents fundamentally rely on DSLFs for the reason that DSLFs can be turned into continuous light fields via linear interpolation [9]. However, due to the difficulty of directly capturing a DSLF, a DSLF is typically reconstructed from a SSLF. The challenging DSLF reconstruction problem has been tried to be solved by light field angular super-resolution-based approaches, most of which treat it as novel view synthesis problem and do not consider the disparity range of the input SSLF. Kalantari *et al*. propose a learning-based approach composed of a disparity estimator and a color predictor to synthesize novel views from four corner sub-aperture views of a micro-lens array-based light field camera [10]. Wu *et al*. utilize a residual-learning method to restore the angular detail of EPIs within a blur-deblur framework [11]. However, the maximum disparity of the SSLF that can be handled by this approach is only 5 pixels. Yeung *et al*. also design a learning-based view synthesis network consisting of view synthesis and refinement components to reconstruct DSLFs [12]. Nevertheless, for different interpolation rates, their network needs to be retrained. Gao and Koch utilize a state-of-the-art video frame interpolation method, *i.e.* adaptive Separable Convolution (SepConv) [13], and a fine-tuning strategy enhancing the convolution kernels of SepConv to reconstruct DSLFs in a recursive way [14].

3. METHODOLOGY

3.1. DSLF reconstruction using ST

The shearlet transform approach for DSLF reconstruction is originally proposed in [3] and extended in [4] with computational acceleration. Given a coarsely-sampled EPI $\varepsilon \in \mathbb{R}^{m\times\dot{n}}$ from a SSLF as shown in Fig. 1 (d), ST reconstructs a desired densely-sample EPI $f \in \mathbb{R}^{m\times\ddot{n}}$ by an iterative inpainting algorithm using the sparse representation of f in shearlet domain. The sampling interval of the desired EPI f for rearranging the rows of the input decimated EPI ε is denoted by τ as illustrated in Fig. 2 (a). Since the desired EPI f to be reconstructed is densely sampled, it is apparent that $\tau \geqslant d_{range}$ and d_{range} stands for the disparity range of the input decimated EPI ε, *i.e.* $d_{range} = d_{max} - d_{min}$. It should be noted that a pre-shearing process relying on d_{min} is typically necessary for the input decimated EPI ε in order to make sure that the new $d'_{min} = 0$ and $d'_{max} = d_{range}$. Besides, the vertical sizes of the input decimated EPI ε and reconstructed densely-sampled EPI f meet the condition that $\ddot{n} = (\dot{n} - 1)\tau + 1$.

The reconstruction of the desired densely-sampled EPI f is typically performed via an iterative inpainting process with t iterations, corresponding to the intermediate reconstructed EPI result f_i, $i \in [1,\, t] \cap \mathbb{Z}$. Besides, the shearlet analysis transform for reconstructing f is defined as $S : \mathbb{R}^{m\times\ddot{n}} \rightarrow \mathbb{R}^{\eta\times m\times\ddot{n}}$ and the shearlet synthesis transform is denoted by $S^* : \mathbb{R}^{\eta\times m\times\ddot{n}} \rightarrow \mathbb{R}^{m\times\ddot{n}}$. Additionally, f_0 stands for the coarse estimation of f, which is a zero-padded EPI for the input decimated EPI ε, *i.e.* $f_0(:\tau:,\,:) = \varepsilon$ as shown in Fig. 2 (a). The reconstruction of f_i during iteration i is performed using the double relaxation method [4], which has been demonstrated to be faster and more robust than the original hard-thresholding algorithm in [3]:

$$\begin{aligned} \hat{f}_i &= S^*\Big(T_{\lambda_i}\big(S\big(f_i + \alpha(f_0 - M\circ f_i)\big)\big)\Big)\,, \\ \tilde{f}_i &= \hat{f}_i + \beta_1(\hat{f}_i - f_{i-1})\,, \\ f_{i+1} &= \tilde{f}_i + \beta_2(\tilde{f}_i - f_{i-2})\,, \end{aligned} \tag{1}$$

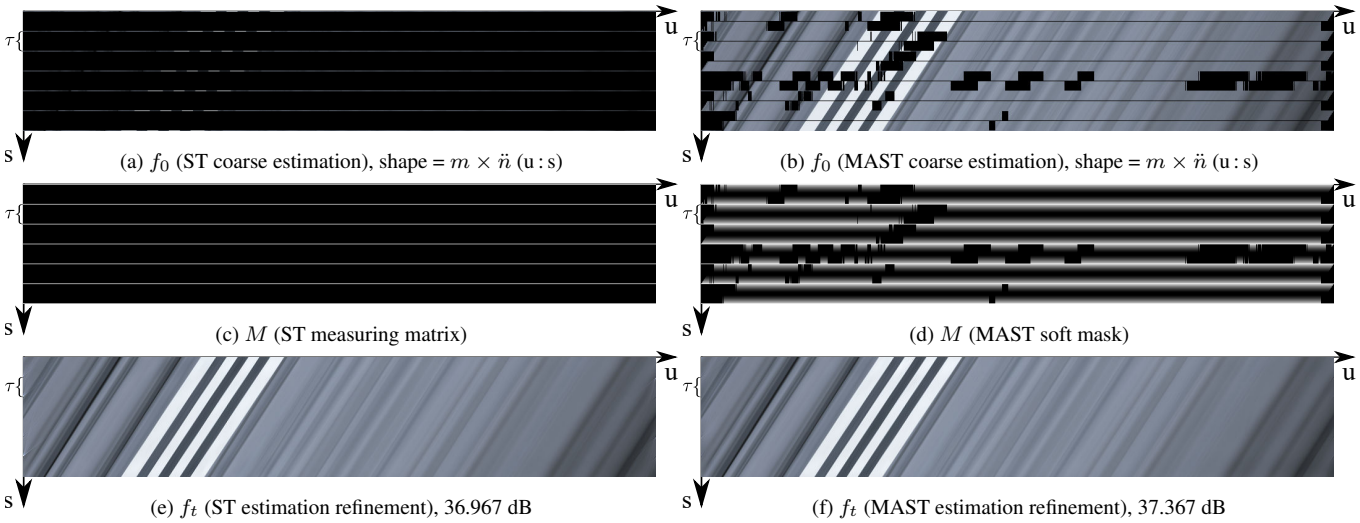
where

$$\begin{aligned} \beta_1 &= \frac{\mathrm{sum}\big((f_0 - \hat{f}_i)\circ M\circ(\hat{f}_i - f_{i-1})\big)}{\mathrm{sum}\big((\hat{f}_i - f_{i-1})\circ M\circ(\hat{f}_i - f_{i-1})\big)}\,, \\ \beta_2 &= \frac{\mathrm{sum}\big((f_0 - \tilde{f}_i)\circ M\circ(\tilde{f}_i - f_{i-2})\big)}{\mathrm{sum}\big((\tilde{f}_i - f_{i-2})\circ M\circ(\tilde{f}_i - f_{i-2})\big)}\,. \end{aligned} \tag{2}$$

Here, $\mathrm{sum}(\cdot)$ returns the sum of all the elements in the input matrix, α is a parameter for adjusting the convergence speed, '$\circ$' denotes the element-wise (Hadamard) product and M is a logical measuring matrix as shown in Fig. 2 (c), where ideally $f_t \circ M = f_0$. In addition, $T_{\lambda_i}(\cdot)$ is a hard-thresholding operator [15] for the threshold value λ_i, which linearly decreases from λ_{max} to λ_{min} with iteration i increasing from 1 to t. As can be seen from (1) and (2), the computation time of the ST approach above is linearly dependent on the maximum iteration number t. A reliable f_0, *i.e.* coarse estimation of f, makes it feasible to accelerate ST with a smaller t.

3.2. Mask-Accelerated Shearlet Transform (MAST)

In order to make a more reliable estimation of f_0 *w.r.t.* the desired densely-sampled EPI f, one of the state-of-the-art learning-based optical flow methods, *i.e.* FlowNet2 [6], is utilized to estimate bidirectional flow between adjacent views in a horizontal-parallax SSLF, $\mathcal{D}^{\mathrm{sslf}} = \{\mathcal{I}_i | 1 \leqslant i \leqslant \dot{n}\}$, of which the corresponding unknown DSLF is denoted by $\mathcal{D}^{\mathrm{dslf}} = \{\tilde{\mathcal{I}}_r | 1 \leqslant r \leqslant \ddot{n}\}$. Since a horizontal-parallax SSLF does not have vertical motions of image objects between any two neighboring views, only the horizontal component of the



(a) f_0 (ST coarse estimation), shape = $m \times \ddot{n}$ (u : s)
(b) f_0 (MAST coarse estimation), shape = $m \times \ddot{n}$ (u : s)
(c) M (ST measuring matrix)
(d) M (MAST soft mask)
(e) f_t (ST estimation refinement), 36.967 dB
(f) f_t (MAST estimation refinement), 37.367 dB

Fig. 2. The coarse estimation, measuring matrix (soft mask) and estimation refinement for a densely-sampled EPI f.

optical flow displacement vector is kept after the bidirectional flow estimation. The bidirectional flow between $\mathcal{I}_i$ and $\mathcal{I}_{i+1}$ in the $\mathcal{D}^{\text{sslf}}$ is represented by $F_{i\to i+1}$ and $F_{i+1\to i}$. A forward-backward consistency constraint [16] between $F_{i\to i+1}$ and $F_{i+1\to i}$ is applied here to roughly remove the inaccuracies caused by occlusions and large motions of image objects. Let $\dot{r} = (r-1)\%\tau$[1] and $i = 1 + (r - \dot{r} - 1)/\tau$, the estimated bidirectional flow is then used to perform a coarse estimation of the parallax images in $\mathcal{D}^{\text{dslf}}$ as follows[2]:

$$\tilde{\mathcal{I}}_r = \begin{cases} \mathcal{I}_i & \text{for} \quad \dot{r} = 0\,, \\ g(\mathcal{I}_i,\, -\frac{\dot{r}}{\tau} F_{i\to i+1}) & \text{for} \quad 0 < \dot{r} < \frac{\tau}{2}\,, \\ g(\mathcal{I}_{i+1},\, -\frac{(\tau-\dot{r})}{\tau} F_{i+1\to i}) & \text{for} \quad \frac{\tau}{2} < \dot{r} < \tau\,, \\ \mathbf{0} & \text{for} \quad \dot{r} = \frac{\tau}{2}\,. \end{cases} \tag{3}$$

Here, $g(\cdot,\,\cdot)$ is an inverse warping function using bicubic interpolation [17]. The roughly-estimated $\mathcal{D}^{\text{dslf}}$ is then turned into densely-sampled EPIs, such that the coarse estimation f_0 of f is partially restored as displayed in Fig. 2 (b). Note that the large missing areas are caused by the filtering of the unreliable optical flows using the bidirectional consistency check. However, the roughly inpainted areas in f_0 are not accurate enough for directly using ST. Specifically, due to the accumulation error of the optical flow in the interpolation algorithm in (3), horizontal lines of f_0 near the locations, *i.e.* $\dot{r} = \frac{\tau}{2}$, have larger inpainting errors than those near the ground-truth regions, *i.e.* $\dot{r} = 0$. Therefore, a novel ST-based method, Mask-Accelerated Shearlet Transform (MAST), is proposed to solve this problem by replacing the measuring matrix in (1) and (2) with an elaborately-designed soft mask, *i.e.*

$$M(r,\,c) = \begin{cases} 1.0 & \text{for} \quad \dot{r} = 0\,, \\ \omega(1 - \frac{2\dot{r}}{\tau})^2 & \text{for} \quad \dot{r} > 0\,,\ f_0(r,\,c) > 0\,, \\ 0 & \text{for} \quad \dot{r} > 0\,,\ f_0(r,\,c) = 0\,, \end{cases} \tag{4}$$

where $\omega \in (0,\,1)$, $r \in [1,\,\ddot{n}] \cap \mathbb{Z}$ and $c \in [1,\,m] \cap \mathbb{Z}$. An example soft mask corresponding to f_0 is illustrated in Fig. 2 (d). It can be seen that this mask suppresses the contributions of f_0 in the regions which are not inpainted or meet the condition that $\dot{r}$ is close to $\frac{\tau}{2}$; however, it enhances the contributions from the ground-truth nearby areas, thus effectively improving the initialization of the densely-sampled EPIs for the iterative double relaxation-based ST in Sect. 3.1.

[1]Here, '%' stands for the modulo operation.

[2]Assume that $\tau\%2 = 0$ for this $\mathcal{D}^{\text{dslf}}$.

4. EXPERIMENTS

4.1. Experimental settings

Datasets. The high density camera array dataset [18] is a real-world 4D light field dataset that can be utilized to evaluate light field angular super-resolution methods with large disparity ranges. Nine different scenes in this dataset are captured by a high-resolution and high-definition DSLR camera in a precise gantry system, such that nine corresponding light field sub-datasets are built. Eight of these sub-datasets have an angular resolution of 101×21. The remaining one has an angular resolution of 99×21. The spatial resolution of all the sub-datasets is 3976×2652 pixels. The raw images in each sub-dataset have black areas near the image borders, which is due to the calibration, and large disparities between neighboring views, which make it difficult to use these raw images as ground-truth light field data directly. To overcome this limitation, a cutting and scaling strategy is proposed as shown in Fig. 3 (j). In particular, a bottom-right $16:9$ image is cut from a raw image with preserving 95% of the width of this raw image. The cut image is then resized to 1024×576 pixels using bicubic interpolation. Finally, only the top 97 images after the process of the cutting and scaling strategy for each sub-dataset are kept and used as the ground-truth horizontal-parallax light field dataset $\mathcal{D}_\mu$, $\mu \in [1,\,9] \cap \mathbb{Z}$. In other words, $\mathcal{D}_\mu = \{\mathcal{I}_j^\mu | 1 \leqslant j \leqslant n\}$, $\mathcal{I}_j^\mu \in \mathbb{R}^{m\times N}$, where $n = 97$, $m = 1024$ and $N = 576$. The middle image, *i.e.* $\mathcal{I}_{49}^\mu$, of each ground-truth 3D light field dataset $\mathcal{D}_\mu$ is exhibited in Fig. 3 (a)-(i). The SSLF $\mathcal{D}_\mu^{\text{sslf}}$ from $\mathcal{D}_\mu$ is generated by using an interpolation rate δ ($= 16$) as shown in Fig. 1 (a) and (c), such that $\mathcal{D}_\mu^{\text{sslf}} = \{\mathcal{I}_i^\mu | 1 \leqslant i \leqslant \dot{n}\}$, $\dot{n} = (n-1)/\delta + 1$.

(a) $\mathcal{D}_1$: Books and charts (b) $\mathcal{D}_2$: Lego city (c) $\mathcal{D}_3$: Lightfield production (d) $\mathcal{D}_4$: Plants (e) $\mathcal{D}_5$: Table in the garden

(f) $\mathcal{D}_6$: Table top I (g) $\mathcal{D}_7$: Table top II (h) $\mathcal{D}_8$: Table top III (i) $\mathcal{D}_9$: Workshop (j) Cutting and scaling

Fig. 3. The middle views of nine evaluation datasets and (j) illustrates the image cutting and scaling strategy in Sect. 4.1.

Table I. The minimum and average per-view PSNR results (in dB, explained in Sect. 4.1) for the performance evaluation of different DSLF reconstruction methods on nine light field evaluation datasets.

Minimum per-view PSNR value (dB) of DSLF reconstruction on $\mathcal{D}_\mu$									
Method	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$
SepConv ($\mathcal{L}_1$) [13]	23.324	20.341	23.912	25.059	27.080	28.344	20.419	21.208	26.369
PIASC ($\mathcal{L}_1$) [14]	23.311	20.343	23.915	25.065	27.092	28.396	20.416	21.208	26.377
ST [4]	28.881	22.725	26.252	27.718	29.418	**32.485**	**23.186**	23.518	28.710
MAST	**30.167**	**22.965**	**26.866**	**27.920**	**29.541**	32.448	23.119	**23.847**	**29.001**
Average per-view PSNR value (dB) of DSLF reconstruction on $\mathcal{D}_\mu$									
Method	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$	$\mathcal{D}_5$	$\mathcal{D}_6$	$\mathcal{D}_7$	$\mathcal{D}_8$	$\mathcal{D}_9$
SepConv ($\mathcal{L}_1$) [13]	26.220	22.569	26.251	27.645	28.719	29.868	22.929	23.500	28.546
PIASC ($\mathcal{L}_1$) [14]	26.231	22.587	26.281	27.697	28.777	29.921	22.941	23.529	28.595
ST [4]	30.122	24.107	28.294	**29.487**	30.358	33.361	**24.431**	**25.417**	30.605
MAST	**31.286**	**24.214**	**28.740**	29.356	**30.371**	**33.768**	24.226	25.390	**30.624**

Table II. The average computation time of reconstructing a densely-sampled EPI (RGB channels) using ST and MAST.

Average computation time (s)		
Method	$\tau = 32$	$\tau = 48$
ST [4]	7.966	14.867
MAST	**2.813**	**5.073**

The DSLF to be reconstructed is $\mathcal{D}_\mu^{\text{dslf}} = \{\tilde{\mathcal{I}}_r^\mu | 1 \leqslant r \leqslant \ddot{n}\}$, $\ddot{n} = (\dot{n}-1)\tau + 1$ as described in Sect. 3.

Disparity estimation. The horizontal disparities between neighboring views in each $\mathcal{D}_\mu^{\text{sslf}}$ are calculated via the optical flow algorithm in Sect. 3.2. The estimated minimum disparity d_{min}, maximum disparity d_{max} and disparity range d_{range} are illustrated in Fig. 5. The sampling interval τ should be as small as possible in order to save computation time for both ST and MAST, while it has two constraints that $\tau \% \delta = 0$ and $\tau \geqslant d_{range}$ (see Sect. 3.1). Therefore, it can be seen from the figure that the best sampling interval τ for datasets $\mathcal{D}_\mu$, $\mu \in \{1, 2, 7\}$ is 32 and for the other six datasets, $\tau = 48$.

Evaluation criteria. The per-view PSNR for a ground-truth dataset $\mathcal{D}_\mu$ and the reconstructed $\mathcal{D}_\mu^{\text{dslf}}$ from $\mathcal{D}_\mu^{\text{sslf}}$ for it is described as below:

$$\text{MSE}_j^\mu = \frac{1}{3 \cdot m \cdot N} \sum_{x=1}^{m} \sum_{y=1}^{N} \left\| \tilde{\mathcal{I}}^\mu_{\frac{\tau(j-1)}{\delta}+1}(x, y) - \mathcal{I}_j^\mu(x, y) \right\|_2^2,$$

$$\text{PSNR}_j^\mu = 10 \log_{10} \left(\frac{255^2}{\text{MSE}_j^\mu} \right). \tag{5}$$

The minimum and average per-view PSNRs constitute the evaluation criteria for the evaluation of different DSLF reconstruction methods on a dataset $\mathcal{D}_\mu$.

Implementation details. For a dataset $\mathcal{D}_\mu$, the construction of a specifically-designed universal shearlet system [3] with ξ scales relies on the sampling interval τ of it, *i.e.* $\xi = \lceil \log_2 \tau \rceil$. The parameter ω in (4) is set to 0.1. The maximum threshold value λ_{max} and minimum threshold value λ_{min} are set to 8 and 0.04, respectively. Note that these two values are for the case of using a normalized coarsely-sampled EPI ε, *i.e.*

$$\varepsilon = \frac{\varepsilon - \min(\varepsilon)}{\max(\varepsilon) - \min(\varepsilon)}, \tag{6}$$

where $\max(\cdot)$ and $\min(\cdot)$ return the maximum value and the minimum value of an input matrix, respectively. The reconstructed f using this normalized ε is then rescaled back to original range of values via

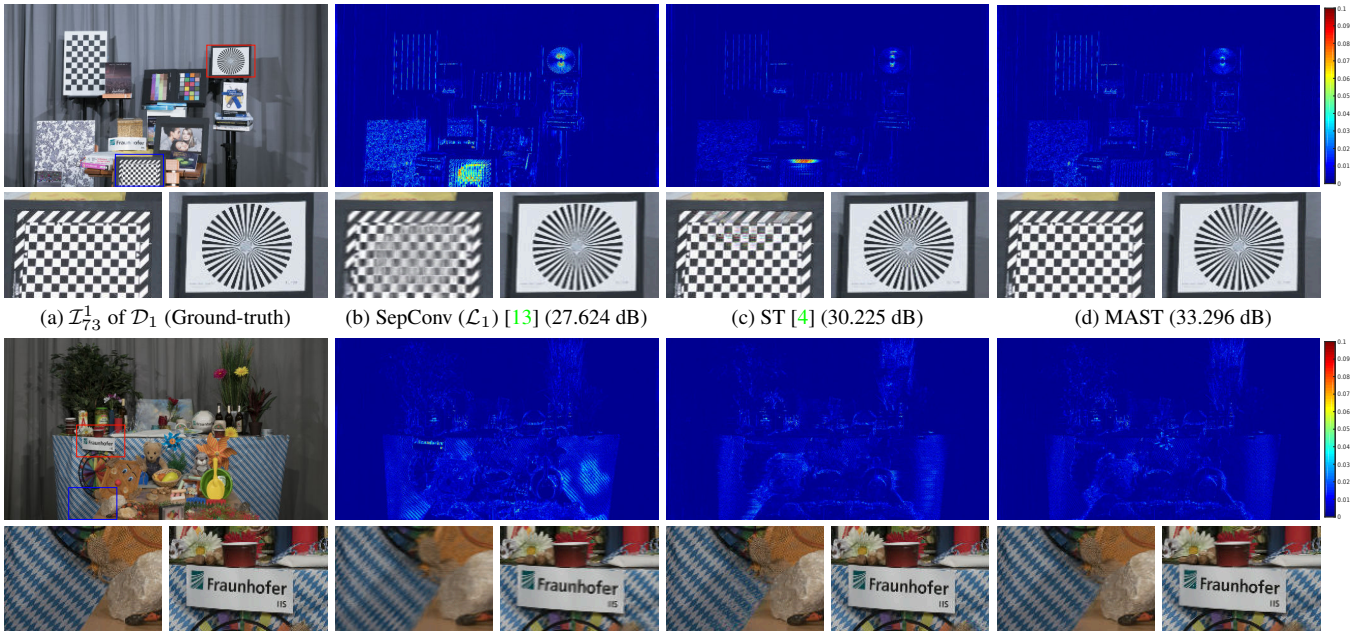
$$f = \big(\max(\varepsilon) - \min(\varepsilon)\big) f + \min(\varepsilon). \tag{7}$$

Besides, for the maximum iteration number of ST, $t = 100$ and for that of MAST, $t = 30$. Regarding the parameter controlling the convergence speed in (1), $\alpha = 30$. Both ST and MAST are implemented by using CUDA and executed on an Nvidia GeForce GTX Titan X 12 GB GPU.

4.2. Results and analysis

The proposed method and baseline approaches are evaluated quantitatively and qualitatively as follows:

Quantitative evaluation. The minimum and average per-view PSNR values of using different DSLF reconstruction methods on different horizontal-parallax light field datasets



(a) $\mathcal{I}^1_{73}$ of $\mathcal{D}_1$ (Ground-truth) (b) SepConv ($\mathcal{L}_1$) [13] (27.624 dB) (c) ST [4] (30.225 dB) (d) MAST (33.296 dB)

(e) $\mathcal{I}^6_{25}$ of $\mathcal{D}_6$ (Ground-truth) (f) SepConv ($\mathcal{L}_1$) [13] (28.863 dB) (g) ST [4] (33.326 dB) (h) MAST (34.024 dB)

Fig. 4. The visualization of the DSLF-reconstruction quality of using different methods.

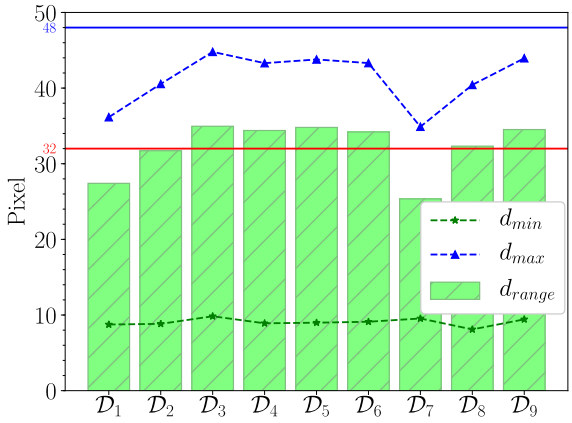


Fig. 5. The disparity estimations of $\mathcal{D}_\mu$.

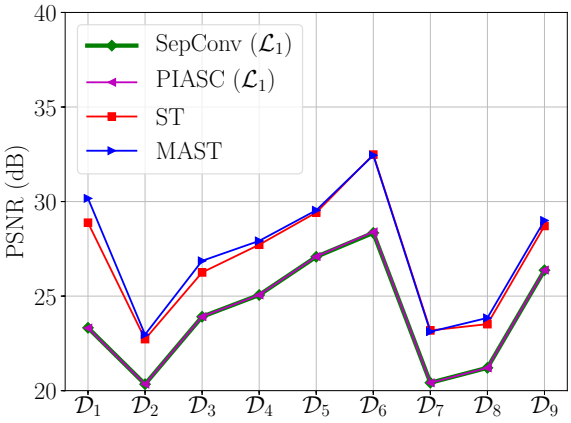


Fig. 6. The minimum per-view PSNR results on $\mathcal{D}_\mu$.

are presented in Table I. It can be seen from the minimum per-view PSNR data that the proposed MAST method achieves the best performance on most of the datasets except for $\mathcal{D}_6$ and $\mathcal{D}_7$. However, on these two datasets, the minimum per-view PSNR values of ST are only 0.037 dB and 0.067 dB higher than those of MAST. With regard to the average per-view PSNR data at the bottom of Table I, MAST still outperforms the other baseline methods on most datasets except for $\mathcal{D}_4$, $\mathcal{D}_7$ and $\mathcal{D}_8$, which demonstrates the effectiveness of the proposed DSLF reconstruction approach again. The computation efficiency of both ST and MAST is evaluated in terms of computation time as shown in Table II. The proposed MAST is significantly faster than ST, *i.e.* MAST requires only $\approx 35\%$ computation time of ST, which is mainly because MAST has 30 refinement iterations while ST needs 100 iterations for the DSLF reconstruction on the challenging horizontal-parallax real-world light field datasets. This also demonstrates that MAST is much more efficient than ST for DSLF reconstruction. Note that the computation time of optical flow estimation and inverse warping parts of MAST can be ignored compared with the computation time of the iterative EPI-refinement process since both of them are performed in real-time.

Qualitative evaluation. The minimum per-view PSNR data for all the DSLF reconstruction methods on different light field evaluation datasets are plotted in Fig. 6. It is apparent that both SepConv and PIASC have almost the same minimum per-view PSNR results on all the datasets, which are much lower than ST and MAST. This suggests that the two DSLF reconstruction methods using the state-of-the-art video frame interpolation technology are not appropriate for DSLF reconstruction from SSLFs with large disparity ranges. Besides, the proposed MAST approach outperforms ST on most of the challenging light field datasets, which indicates that MAST is more effective than ST for DSLF reconstruction. The reconstructed images using different DSLF reconstruc-

tion methods are visualized and compared in Fig. 4. Since SepConv and PIASC have similar DSLF reconstruction performance, only SepConv is compared here. For the top row, the image parts of the checkerboard and Siemens star on $\mathcal{I}_{73}^{1}$ of $\mathcal{D}_1$ are chosen as the interesting areas to be compared. It can be seen from Fig. 4 (b) that the reconstructed checkerboard using SepConv has serious blur artifacts, which is mainly because the the size of repetitive check patterns of the checkerboard is much smaller than the disparities of it, such that SepConv is incapable of knowing the true motion of this checkerboard. As can be seen from Fig. 4 (c), the recovered checkerboard using ST is slightly better than that of using SepConv, while the reconstructed Siemens star has obvious artifacts. In Fig. 4 (d), the proposed MAST method achieves the best reconstruction performance with visually correct and sharp results, which proves the effectiveness of the proposed MAST method composed of optical-flow-based coarse estimation and mask-assisted iterative estimation refinement for EPIs. Regarding the bottom row Fig. 4, part of the tablecloth with foreground and the Fraunhofer IIS logo are selected as the interesting areas from $\mathcal{I}_{25}^{6}$ of $\mathcal{D}_6$. Both of the reconstructed results in Fig. 4 (f) using SepConv are blur, which, on the one hand, is caused by the small size of the repetitive pattern of the tablecloth; on the other hand, the size of the convolution kernels of SepConv is only 51×51, restricting the performance of it in handling DSLF reconstruction from SSLFs with large disparity ranges. The DSLF reconstruction results of ST in Fig. 4 (g) do not have this kind of "blur" problem. However, the reconstructed tablecloth area has evident artifacts, which are well handled by the proposed MAST method as illustrated in Fig. 4 (h). It implies that the optical-flow-based coarse estimation and mask-assisted iterative estimation refinement in MAST are beneficial to improving the final DSLF reconstruction performance.

5. CONCLUSION

This paper presents a novel coarse-to-fine method, MAST, for DSLF reconstruction from SSLFs with large disparity ranges. The proposed MAST method fully leverages a state-of-the-art optical flow estimation method, *i.e.* FlowNet2, to roughly estimate a densely-sampled EPI from a sparsely-sampled EPI. Based on the coarsely-inpainted densely-sampled EPI and the inevitable error accumulation of any optical flow algorithm, a soft mask is elaborately designed for the iterative hard-thresholding-based estimation refinement approach in ST. Experimental results show that MAST achieves better performance than the other state-of-the-art DSLF reconstruction methods on nine challenging real-world horizontal-parallax light field datasets with large disparity ranges (up to 35 pixels). Moreover, MAST is a time-efficient algorithm that is nearly three times faster than ST.

Acknowledgments. The work in this paper was funded from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676401, European Training Network on Full Parallax Imaging, and the German Research Foundation (DFG) No. K02044/8-1. We thank Nvidia for their GPU donation.

6. REFERENCES

[1] M. Levoy and P. Hanrahan, "Light field rendering," in *SIGGRAPH*, 1996, pp. 31–42. 1

[2] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Image based rendering technique via sparse representation in shearlet domain," in *ICIP*, 2015, pp. 1379–1383. 1

[3] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform," *IEEE TPAMI*, vol. 40, no. 1, pp. 133–147, 2018. 1, 2, 4

[4] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Accelerated shearlet-domain light field reconstruction," *IEEE J-STSP*, vol. 11, no. 7, pp. 1082–1091, 2017. 1, 2, 4, 5

[5] M. Genzel and G. Kutyniok, "Asymptotic analysis of inpainting via universal shearlet systems," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 2301–2339, 2014. 1

[6] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, 2017, pp. 1647–1655. 2

[7] J. Yu, "A light-field journey to virtual reality," *IEEE MultiMedia*, vol. 24, no. 2, pp. 104–112, 2017. 2

[8] A. Smolic, "3D video and free viewpoint video - from capture to display," *Pattern Recognition*, vol. 44, no. 9, pp. 1958–1968, 2011. 2

[9] Z. Lin and H.-Y. Shum, "A geometric analysis of light field rendering," *IJCV*, vol. 58, no. 2, pp. 121–138, 2004. 2

[10] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM TOG*, vol. 35, no. 6, pp. 193:1–193:10, 2016. 2

[11] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on EPI," in *CVPR*, 2017, pp. 1638–1646. 2

[12] H.W.F. Yeung, J. Hou, J. Chen, Y.Y. Chung, and X. Chen, "Fast light field reconstruction with deep coarse-to-fine modelling of spatial-angular clues," in *ECCV*, 2018, pp. 138–154. 2

[13] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *ICCV*, 2017, pp. 261–270. 2, 4, 5

[14] Y. Gao and R. Koch, "Parallax view generation for static scenes using parallax-interpolation adaptive separable convolution," in *ICME Workshops*, 2018, pp. 1–4. 2, 4

[15] H. Lakshman, W.-Q. Lim, H. Schwarz, D. Marpe, G. Kutyniok, and T. Wiegand, "Image interpolation using shearlet based sparsity priors," in *ICIP*, 2013, pp. 655–659. 2

[16] J. Hur and S. Roth, "MirrorFlow: Exploiting symmetries in joint optical flow and occlusion estimation," in *ICCV*, 2017, pp. 312–321. 3

[17] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High quality estimation of multiple intermediate frames for video interpolation," in *CVPR*, 2018, pp. 9000–9008. 3

[18] M. Ziegler, R. op het Veld, J. Keinert, and F. Zilly, "Acquisition system for dense lightfield of large scenes," in *3DTV-CON*, 2017, pp. 1–4. 3