Spatial and Angular Variational Super-Resolution of 4D Light Fields

Sven Wanner and Bastian Goldluecke

Heidelberg Collaboratory for Image Processing

Abstract. We present a variational framework to generate super-resolved novel views from 4D light field data sampled at low resolution, for example by a plenoptic camera. In contrast to previous work, we formulate the problem of view synthesis as a continuous inverse problem, which allows us to correctly take into account foreshortening effects caused by scene geometry transformations. High-accuracy depth maps for the input views are locally estimated using epipolar plane image analysis, which yields floating point depth precision without the need for expensive matching cost minimization. The disparity maps are further improved by increasing angular resolution with synthesized intermediate views. Minimization of the super-resolution model energy is performed with state of the art convex optimization algorithms within seconds.

1 Introduction

In constrast to a finite collection of 2D images, the complete 4D light field of a scene is defined on a continuous domain of camera view points. The continuous disparity space admits non-traditional approaches to the multiview stereo problem that do not rely on feature matching [1,2]. However, in practice, it used to be difficult to achieve a dense enough sampling of the full light field. A few years ago, this was usually performed with expensive custom setups like camera arrays [3] or gantry constructions consisting of a moving camera [4], which can only capture static scenes. Nowadays, plenoptic cameras are commercially available, which makes a dense sampling even of light field videos feasible for real-world applications. Naturally, this creates high demand for robust and efficient light field analysis algorithms.

However, plenoptic cameras usually have to deal with a trade-off between spatial and angular resolution. Since the total sensor resolution is limited, one can either opt for a dense sampling in the spatial (image) domain with sparse sampling in the angular (view point) domain [5], or vice versa [6,7,8]. Recent commercial cameras tend to favor a small number of view points, for example the Raytrix [5] captures a collection of 9 × 9 views simultaneously. Increasing angular resolution is therefore a paramount goal if one wants to make efficient use of plenoptic cameras. It is equivalent to the synthesis of novel views from new viewpoints, which has also been a prominent research topic in the computer graphics community [9,4]. In this work, we will unify it with state-of-the-art spatial super-resolution research in computer vision [10,11].

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part V, LNCS 7576, pp. 608–621, 2012.

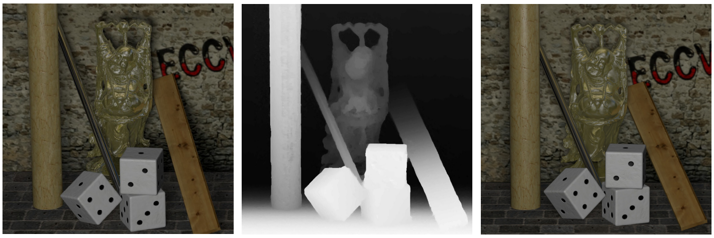

Input view (1/24) Estimated disparity Super-res. novel view

Fig. 1. Our variational framework allows the synthesis of super-resolved views of a light field from arbitrary view points. The novel view above was generated from 24 input views with a resolution of 768×768 pixels at 3×3 super-resolution, yielding an output resolution of 2304×2304.

Related Work. The *plenoptic function* is defined on a seven-dimensional space and describes the entire information about light emitted by a scene, storing an intensity value for every 3D point, direction, wavelength and time [12]. A reduction to a plane and directional information leads to the four-dimensional *Lumigraph* [13,4], which is usually expressed in the two-plane parametrization we also adopt in this work. Any collection of images of a scene can be interpreted as a sparse sampling of the plenoptic function. Consequently, image-based rendering approaches [14] treat the creation of novel views as a resampling problem, circumventing the need for any explicit geometry reconstruction [9,4,15]. However, this approach ignores occlusion effects, and therefore is only really suitable for synthesis of views reasonably close to the original ones.

Thus, it quickly became clear that one faces a trade-off, and interpolation of novel views in sufficient enough quality requires either an unreasonably dense sampling or knowledge about scene geometry [16]. A different line of approaches to light field rendering therefore tries to infer at least some geometric knowledge about the scene. They usually rely on image registration, for example via robust feature detecting and tracking [17], or view-dependent depth map estimation based on color consistency [18].

The creation of *super-resolved* images requires subpixel-accurate registration of the input images. Approaches which are based on pure 2D image registration [10] are unsuitable for the genereation of novel views, since a reference image for computing the motion is not available yet. Super-resolved depth maps and images are inferred in [7] with a discrete super-resolution model tailored to a particular plenoptic camera. A full geometric model with a super-resolved texture map is estimated in [11] for scenes captured with a surround camera setup. Our approach is mathematically closely related to the latter, since [11] is also based on continuous geometry which leads to correct point-wise weighting of the energy gradient contributions. However, we do not perform expensive computation of a global model and texture atlas, but instead compute the target view directly.

Contributions. This paper simultaneously addresses the problems of spatial and angular super-resolution. We extend the mathematical framework of *variational light field analysis* [2], and propose a variational inverse problem whose solution is the synthesized super-resolved novel view. Since we work in a continuous setting, we can correctly take into account foreshortening effects caused by the scene geometry. The method requires an initial geometry estimate with subpixel accuracy. Recent algorithms tailored to light field data [2] can achive this, and allow to synthesize high-quality novel views. Source code for the method as well as our data sets are available online[1].

2 Super-Resolution View Synthesis Model

In this section, we propose a variational model for the synthesis of super-resolved novel views, to our knowledge the first of this kind which works directly in view space. Since the model is continuous, we will be able to derive Euler-Lagrange equations which correctly take into account foreshortening effects of the views caused by variations in the scene geometry. This makes the model essentially parameter-free. The framework is in the spirit of [11], which computes super-resolved textures for a 3D model from multiple views, and shares the same favourable properties. However, it has substantial differences, since we do not require a complete 3D geometry reconstruction and costly computation of a texture atlas. Instead, we only make use of depth maps on the input images, and model the super-resolved novel view directly.

The following mathematical framework is fully general, and formulated for views with arbitrary projections. However, an implementation in full generality would be highly difficult to achieve. We therefore specialize to the scenario of a 4D light field in the subsequent section, and leave a generalization of the implementation for future work.

For the remainder of the section, assume we have images $v_i : \Omega_i \to \mathbb{R}$ of a scene available, which are obtained by projections $\pi_i : \mathbb{R}^3 \to \Omega_i$. Each pixel of each image stores the integrated intensities from a collection of rays from the scene. This subsampling process is modeled by a blur kernel b for functions on Ω_i, and essentially characterizes the point spread function for the corresponding sensor element. It can be measured for a specific imaging system [19]. In general, the kernel may depend on the view and even on the specific location in the images. We omit the dependency here for simplicity of notation.

The goal is to synthesize a view $u : \Gamma \to \mathbb{R}$ of the light field from a novel view point, represented by a camera projection $\pi : \mathbb{R}^3 \to \Gamma$, where Γ is the image plane of the novel view. The basic idea of super-resolution is to define a physical model for how the subsampled images v_i can be explained using high-resolution information in u, and then solve the resulting system of equations for u. This inverse problem is ill-posed, and is thus reformulated as an energy minimization problem with a suitable prior or regularizer on u.

Image Formation and Model Energy. In order to formulate the transfer of information from u to v_i correctly, we require geometry information [16]. Thus,

[1] http://hci.iwr.uni-heidelberg.de/HCI/Research/LightField/

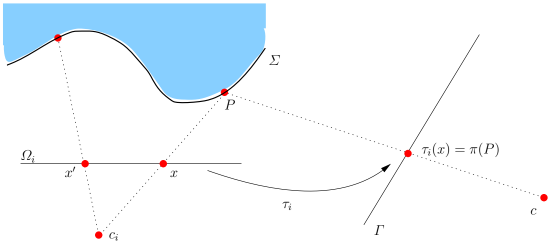


Fig. 2. Transfer map τ_i from an input image plane Ω_i to the image plane Γ of the novel view point. The scene surface Σ can be inferred from the depth map on Ω_i. Note that not all points $x \in \Omega_i$ are visible in Γ due to occlusion, which is described by the binary mask m_i on Ω_i. Above, $m_i(x) = 1$ while $m_i(x') = 0$.

we assume we know (previously estimated) depth maps d_i for the input views. A point $x \in \Omega_i$ is then in one-to-one correspondence to a point P which lies on the scene surface $\Sigma \in \mathbb{R}^3$. The color of the scene point can be recovered from u via $u \circ \pi(P)$, provided that x is not occluded by other scene points, see figure 2.

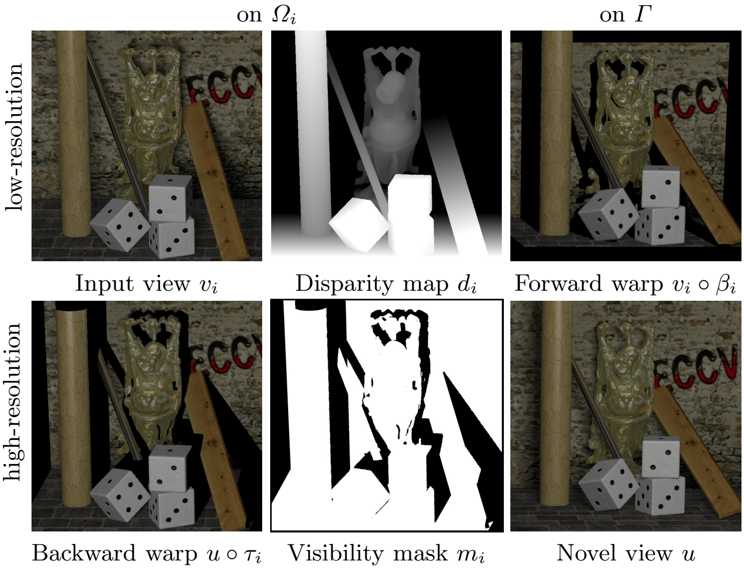
The process explained above induces a *backwards warp map* $\tau_i : \Omega_i \to \Gamma$ which tells us where to look on Γ for the color of a point, as well as a binary *occlusion mask* $m_i : \Omega_i \to \{0, 1\}$ which takes the value 1 if and only if a point in Ω_i is also visible in Γ. Both maps only depend on the scene surface geometry as seen from v_i, i.e. the depth map d_i. The different terms and mappings appearing above and in the following are visualized for an example light field in figure 3.

Having computed the warp map, one can formulate a model of how the values of v_i within the mask can be computed, given a high-resolution image u. Using the downsampling kernel, we obtain $v_i = b*(u\circ\tau_i)$ on the subset of Ω_i where $m_i = 1$, which consists of all points in v_i which are also visible in u. Since this equality will not be satisfied exactly due to noise or inaccuracies in the depth map, we instead propose to minimize the energy

$$E(u) = \sigma^2 \int_\Gamma |Du| + \sum_{i=1}^{n} \underbrace{\frac{1}{2} \int_{\Omega_i} m_i (b * (u \circ \tau_i) - v_i)^2 \, \mathrm{d}x}_{=:E^i_{\text{data}}(u)} . \tag{1}$$

which is the MAP estimate under the assumption of Gaussian noise with standard deviation σ on the input images. It resembles a classical super-resolution model [19], which is made slightly more complex by the inclusion of the warp maps and masks.

In the energy (1), the total variation acts as a regularizer or objective prior on u. Its main tasks are to eliminate outliers and enforce a reasonable inpainting of regions for which no information is available, i.e. regions which are not visible in any of the input views. It could be replaced by a more sophisticated prior for natural images, however, the total variation leads to a convex model which can be very efficiently minimized.



Input view v_i Disparity map d_i Forward warp $v_i \circ \beta_i$

Backward warp $u \circ \tau_i$ Visibility mask m_i Novel view u

Fig. 3. Illustration of the terms in the super-resolution energy. The figure shows the ground truth depth map for a single input view and the resulting mappings for forward- and backward warps as well as the visibility mask m_i. White pixels in the mask denote points in Ω_i which are visible in Γ as well.

Functional Derivative. The functional derivative for the inverse problem above is required in order to find solutions. It is well-known in principle, but made slightly more complicated by the different domains of the integrals. Note that τ_i is one-to-one when restricted to the visible region $V_i := \{m_i = 1\}$, thus we can compute an inverse *forward warp map* $\beta_i := (\tau_i|_{V_i})^{-1}$, which we can use to transform the data term integral back to the domain Γ, see figure 3. We obtain for the derivative of a single term of the sum in (1)

$$dE^i_{\text{data}}(u) = |\det D\beta_i| \ \big(m_i\bar{b} * (b * (u \circ \tau_i) - v_i)\big) \circ \beta_i \tag{2}$$

The determinant is introduced by the variable substitution of the integral during the transformation. A more detailed derivation for a structurally equivalent case can be found in [11].

The term $|\det D\beta_i|$ in equation (2) introduces a pointwise weight for the contribution of each image to the gradient descent. However, β_i depends on the depth map on Γ, which needs to be inferred and is not readily available. Furthermore, for efficiency it needs to be pre-computed, and storage would require another high-resolution floating point matrix per view. Memory is a bottleneck in our method, and we need to avoid this. For this reason, it is much more efficient to transform the weight to Ω_i and multiply it with m_i to create a single weighted mask. Note that

$$|\det D\beta_i| = \left|\det D\tau_i^{-1}\right| = |\det D\tau_i|^{-1} \circ \beta_i. \tag{3}$$

Thus, we obtain a simplified expression for the functional derivative,

$$dE^i_{\text{data}}(u) = \big(\tilde{m_i}\,\bar{b} * (b * (u \circ \tau_i) - v_i)\big) \circ \beta_i \tag{4}$$

with $\tilde{m_i} := m_i \left|\det(D\tau_i)\right|^{-1}$. An example weighted mask is visualized in figure 3.

3 Specialization to 4D Light Fields

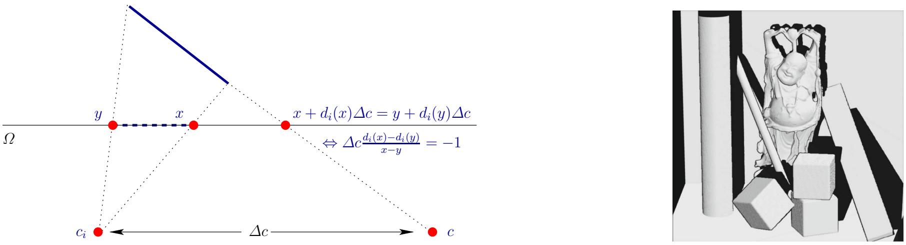
The model introduced in the previous section is hard to implement efficiently in fully general form. This paper, however, focuses on the setting of a 4D light field, where we can make a number of significiant simplifications. The main reason is that the warp maps between the views are given by parallel translations in the direction of the view point change. The amount of translation is proportional to the *disparity* of a pixel, which is in one-to-one correspondence to the depth.

4D Light Field Structure, Views and Depth Estimation. A *4D light field* or *Lumigraph* is the collection of all pinhole views whose focal points lie in a plane Π which is parallel to a common image plane Ω. We write it as a map

$$L : \Pi \times \Omega \to \mathbb{R}, \qquad (c, x) \mapsto L(c, x), \tag{5}$$

which is an assignment of an intensity value to each ray $R_{c,x}$ emanating from the focal point $c \in \Pi$ and passing through $x \in \Omega$. In reality, we will not know the intensity value for every possible ray, but have a more or less sparse sampling available. In particular, we have captured the light field only for a finite collection of vantage points $c_i \in \Pi, 1 \leq i \leq n$, which yield the input views $v_i = L(c_i, \cdot)$ to our super-resolution algorithm with projections π_i corresponding to a projection through the center c_i.

The 3D structure of a scene is strongly related to the internal structure of a light field, which can be exploited for depth reconstruction. This becomes obvious when considering *epipolar plane images (EPIs)*. An epipolar plane image is a 2D cut along a line in Ω through the full stack of views obtained by moving the camera along a direction in Π [1]. Points in space project onto a line in an EPI, yielding a characteristic structure which can be observed in figure 12. Notably, the slope of the line is inversely proportional to the distance of the corresponding point, and called its *disparity*. In [2], a method was presented how to robustly and efficiently obtain dense disparity maps by analyzing the structure tensor of the EPI. As we will see later, disparity maps obtained by this method have subpixel accuracy and are thus suitable to obtain super-resolution. However, a look at figure 12 already suggests that the estimate will be more accurate when the angular sampling of the light field is more dense. An idea is therefore to increase angular resolution and improve the depth estimate by synthesizing intermediate views. In the experimental section, we will demonstrate that this is actually a viable strategy, although it looks circular at first glance.


y
x
$x + d_i(x)\Delta c = y + d_i(y)\Delta c$
Ω
$\Leftrightarrow \Delta c \frac{d_i(x) - d_i(y)}{x-y} = -1$
c_i
Δc
c

(a) Light field geometry (b) Weighted mask $\tilde{m}_i$

Fig. 4. (a) The slope of the solid blue line depends on the disparity gradient in the view v_i. If $\Delta c \cdot \nabla d_i = -1$, then the line is projected onto a single point in the novel view u. (b) Visualization of the weighted mask $\tilde{m}_i$ obtained for the view reconstruction in figure 3.

View Synthesis in the Light Field Plane. The warp maps required for view synthesis become particularly simple when the target image plane Γ lies in the common image plane Ω of the light field, and π resembles the corresponding light field projection through a focal point $c \in \Pi$. In this case, τ_i is simply given by a translation proportional to the disparity,

$$\tau_i(x) = x + d_i(x)(c - c_i), \tag{6}$$

see figure 4. Thus, one can compute the weight in equation (4) to be

$$|\det D\tau_i|^{-1} = |1 + \nabla d_i \cdot (c - c_i)|^{-1} \tag{7}$$

There are a few observations to make about this weight. Disparity gradients which are not aligned with the view translation $\Delta c = c - c_i$ do not influence it, which makes sense since it does not change the angle under which the patch is viewed. Disparity gradients which are aligned with Δc and tend to infinity lead to a zero weight, which also makes sense since they lead to a large distortion of the patch in the input view and thus unreliable information.

A very interesting result is where the location of maximum weight lies. The weights become larger when $\Delta c \cdot \nabla d_i$ approaches -1. An interpretation can be found in figure 4. If $\Delta c \cdot \nabla d_i$ gets closer to -1, then more information from Ω_i is being condensed onto Γ, which means that it becomes more reliable and should be assigned more weight. The extreme case is a line segment with a disparity gradient such that $\Delta c \cdot \nabla d_i = -1$, which is projected onto a single point in Γ. In this situation, the weight becomes singular. This does not pose a problem: From a theoretical point of view, the set of singular points is a null set according to the theorem of Sard, and in practice, it means that we have occlusion and the mask m_i is zero anyway.

Note that formula (7) is highly non-intuitive, but the correct one to use when geometry is taken into account. We have not seen anything similar being used in previous work. Instead, weighting factors for view synthesis are often imposed according to measures based on distance to the interpolated rays or matching similarity scores, which are certainly working, but also somewhat heuristic strategies [4,18,15,10].

4 Discretization and Optimization

Writing a correct implementation of a super-resolution algorithm is a highly complex task which requires much attention to detail. We will therefore provide a clean implementation of our method on our web page after publication. In this section, we sketch the necessary steps and clarify some of the fine points.

Grids and Transformations. The input views are given on low-resolution grids with P pixels in the domains Ω_i, and are represented as vectors $v_i \in \mathbb{R}^P$ of grayscale values. Using bilinear filtering, we first upsample them to a high-resolution grid of $M = P \cdot K$ pixels, where K is the desired magnification factor. We then compute disparity maps $d_i \in \mathbb{R}^M$, using the local structure tensor analysis described in [2]. Note that this step is quite fast and performs at near real-time frame rates.

The depth maps induce the warp maps $\tau_i : \Omega_i \to \Gamma$, thus we can use them to look up values on Γ with bilinear interpolation and compute the backwards warp $u \circ \tau_i$. In the discrete setting, we write the backwards warp as a matrix multiplication $T_i u$ with a sparse matrix $T_i \in \mathbb{R}^{M \times M}$, which has at most four non-zero entries per row. During computation of T_i, we can set up the weighted masks $\tilde{m_i}$ and a lookup table for the backwards warp in form of a sparse matrix W_i such that $W_i v_i = v_i \circ \beta_i$. This pre-computation step is currently relatively expensive since we have not yet been able to fully parallelize it. It takes about half a second per input view at resolution 768 × 768.

Subsampling and Functional Derivative. The subsampling kernel b induces a linear map $B : \mathbb{R}^M \to \mathbb{R}^P$, which maps a high-resolution image to the subsampled low-resolution one. It is given by a sparse matrix, where each row has at most K non-zero entries. Note that B must take into account visibility, i.e. filter only over pixels x for which $m_i(x) = 1$. Putting all this together, we can now compute the functional derivative of the data term on the high-resolution grid Γ via the discretized form of equation (4)

$$dE^i_{\text{data}}(u) = W_i \left[\tilde{m_i} B^T (B T_i u - v_i) \right]. \tag{8}$$

Optimization of the Functional. The energy (1) is convex, since the integral transformation preserves convexity. The derivative of the data term is Lipschitz continuous, from (8), we see that a Lipschitz constant for the discretized form for each term is given by the operator norm of $W_i \tilde{m_i} B^T B T_i$. It is hard to compute in practice, so we determined a conservative upper bound experimentally. All ingredients are thus available to minimize the energy via the fast iterative shrinkage and thresholding algorithm (FISTA) found in [20], see figure 5. Convergence takes about 15 seconds on a recent GPU, at target resolution 2304 × 2304 and input resolution 768 × 768 with 24 views. Most of the time is spent computing

```
Initialize
  fields on $\Gamma$:
    functions $u_0 = 0$, $\bar{u}_0 = 0$
    a vector field $\boldsymbol{\zeta}_0 = 0$
  real numbers:
    step size $\tau = \frac{1}{8}$
    Lipschitz const. $L = 5n$
    overrelaxation factor $t_0 = 1$

Iterate
  $g_n = \bar{u}_n - \frac{1}{L} \sum_{i=1}^{n} dE^i_{\text{data}}(u),\ u_{n+1} = u_n$
  repeat $k$ times
      $\boldsymbol{\zeta}_{n+1} = \Pi_{\sigma^2 \mathbb{E}}(\boldsymbol{\zeta}_n + \tau \nabla u_{n+1})$
      $u_{n+1} = g_n + \frac{1}{L} \operatorname{div}(\boldsymbol{\zeta}_{n+1})$
  $t_{n+1} = (1 + \sqrt{1 + 4t_n^2})/2$
  $\bar{u}_{n+1} = u_n + \frac{t_n - 1}{t_{n+1}}(u_{n+1} - u_n)$
```

Fig. 5. Super-resolution algorithm for minimization of energy (1). The above method is a specialization of FISTA [20], where the inner loop computes a proximation for the total variation using the Bermùdez-Moreno algorithm [21]. The operator $\Pi_{\sigma^2 \mathbb{E}}$ denotes a point-wise projection onto the ball of radius σ^2.

the gradient terms, in particular the warping. Computation time scales roughly linearly with the number of input views and pixels.

5 Experiments

View Interpolation and Super-Resolution. In a first set of experiments, we show the quality of view interpolation and super-resolution, both with ground truth as well as estimated depth. In table 6, we synthesize the center view of a light field with our algorithm using the remaining views as input, and compare the result to the actual view. We compute results both with ground truth disparities to show the maximum theoretical performance of the algorithm, as well as for the usual real-world case that depth needs to be estimated. This estimation is performed using the local method in [2].

	Conehead			*Buddha*			*Mona*			
Views	1*x*1	3*x*3	*IP*	1*x*1	3*x*3	*IP*	1*x*1	3*x*3	*IP*	
5 × 5	31.59	29.30	26.50	32.24	28.91	26.54	30.13	28.29	26.42	GT
9 × 9	31.58	29.43	26.45	32.20	29.05	26.45	30.04	28.29	26.32	
17 × 17	31.19	30.38	26.02	31.75	30.19	27.17	30.17	28.87	26.49	
5 × 5	31.05	29.30	25.77	27.96	28.91	24.34	26.44	28.29	23.84	ED
9 × 9	31.38	29.43	26.23	30.68	29.05	27.70	28.87	28.29	25.13	
17 × 17	31.49	30.86	24.27	31.42	29.54	26.81	29.49	28.30	25.80	

Fig. 6. Reconstruction error for the data sets obtained with a ray-tracer. The table shows the PSNR of the center view without super-resolution, at super-resolution magnification 3 × 3, and for bilinear interpolation to 3 × 3 resolution (IP) as a comparison. The set of experiments is run with both ground truth (GT) and estimated disparities (ED). The estimation error for the disparity map can be found in figure 7. Input image resolution is 384 × 384.

Views	*Cone*	*Buddha*	*Mona*
5 × 5	4.53	0.88	2.13
9 × 9	0.066	0.08	0.19
17 × 17	0.014	0.014	0.04

Disparity MSE

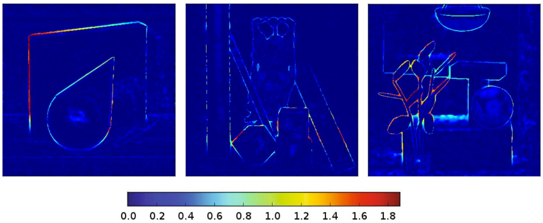
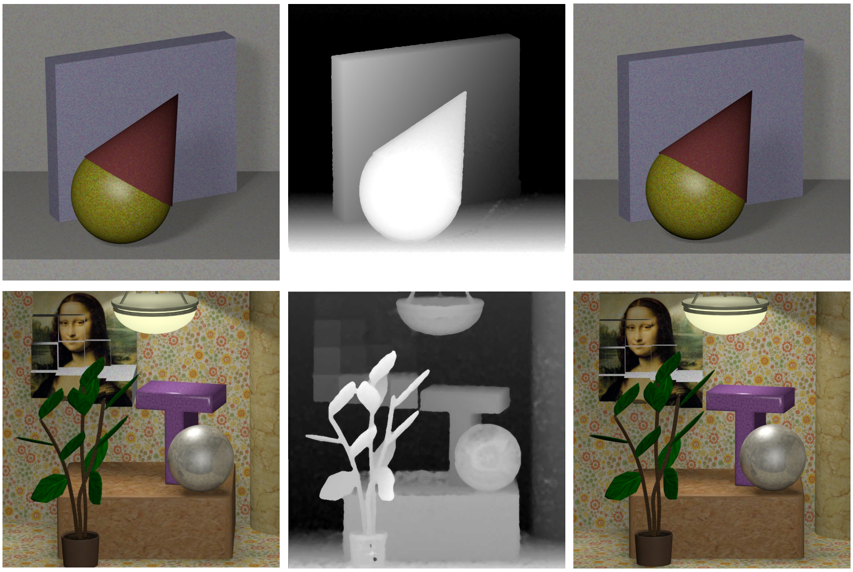


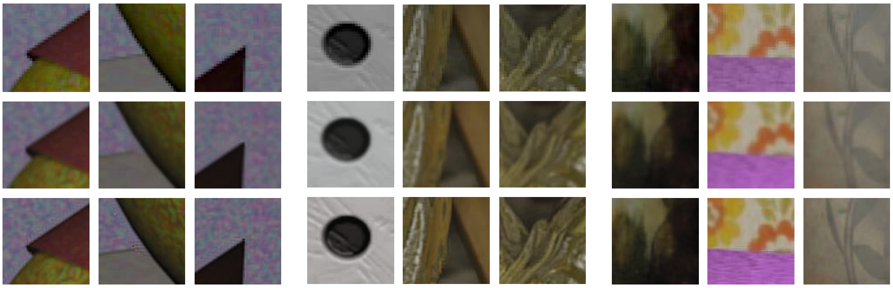
Fig. 7. Disparity reconstruction error. The table shows the mean squared error for the depth maps reconstructed with [2] at different angular resolutions. We see that depth maps clearly get better the higher the angular resolution is. The images show the distribution of the error, which is concentrated around depth discontinuities. Note that these regions do not influence the final result by much, since the weight of their contribution is small, see figure 4(b).

In order to test the quality of super-resolution, we compute the 3 × 3 super-resolved center view and compare with ground truth. For reference, we also compare the result of a bilinear interpolation (IP) of the center view synthesized in the first experiment. While the reconstruction with ground truth disparities is almost perfect, we can clearly see that in the case of estimated depth, the result strongly improves with higher angular resolution due to better depth estimates, figure 7. Super-resolution is definitely superior to simple bilinear upsampling. This also emphasizes the sub-pixel accuracy of the disparity maps, since without



Input view (1/24) Estimated disparity Super-resolved novel view

Fig. 8. Super-resolution results for the data sets *Conehead* and *Mona*. Computed from 24 input views with a resolution of 768 × 768 pixels at 3 × 3 super-resolution, yielding an output resolution of 2304 × 2304. See figure 9 for closeups.



(a) Conehead (b) Buddha (c) Mona

Fig. 9. Closeups of the results in figure 8 for the light fields generated with a ray tracer. From top to bottom: low-resolution center view (not used for reconstruction), high resolution center view obtained by bilinear interpolation of a low-resolution reconstruction from 24 other views, super-resolved reconstruction. Note the increased sharpness and details of the super-resolved result.

Views	*Cone*	*Buddha*	*Mona*
input 5 × 5	4.534	0.883	2.125
SR 9 × 9	1.084	0.559	1.058
input 9 × 9	0.066	0.080	0.192
SR 17 × 17	0.044	0.066	0.105

Disparity MSE

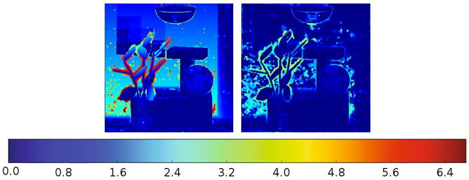


Fig. 10. Iterative disparity improvement. With initial depth maps for the original angular sampling of the input data set, one can compute intermediate views in order to increase the resolution of the epipolar plane images, see figure 12. This in turn leads to an improved disparity estimate when using the algorithm in [2]. The table shows mean squared error for the depth maps at original and super-resolved (SR) angular resolution, the images illustrate the distribution of the depth error before and after super-resolution and the final depth map.

Method	*Demo*	*Motor*
1 × 1	36.91	35.36
3 × 3	30.82	31.72
IP	23.89	22.84

Reconstruction PSNR

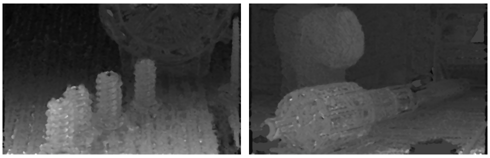
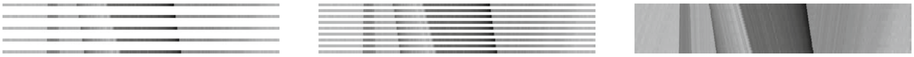
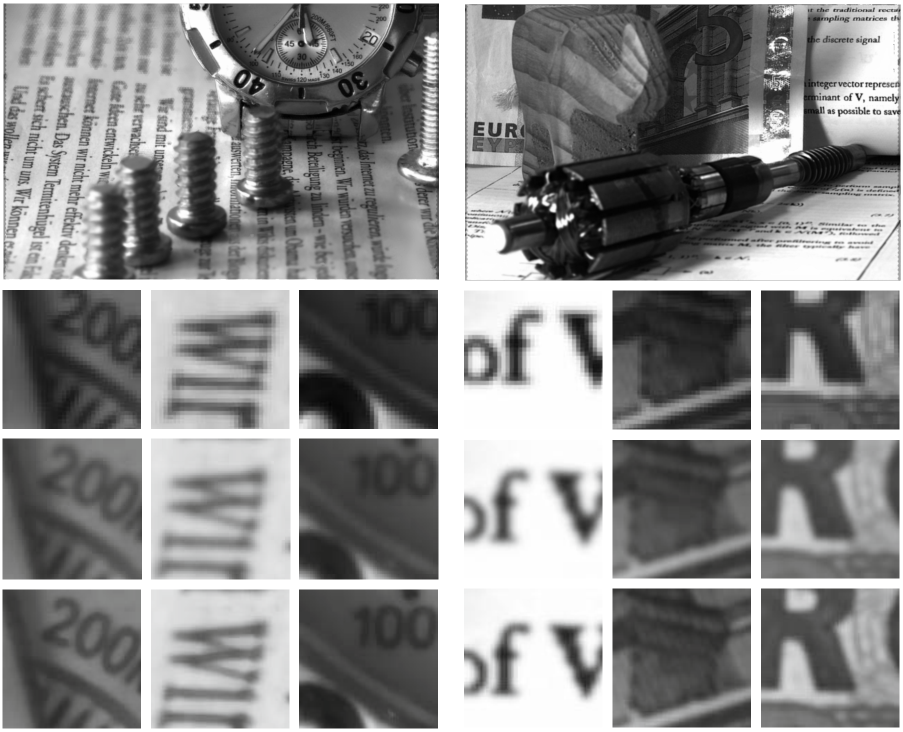


Fig. 11. Reconstruction error for light fields captured with the Raytrix plenoptic camera. The table on the left shows PSNR for the reconstructed input view at original resolution as well as 3 × 3 super-resolution and 3 × 3 interpolation (IP) for comparison. Since no ground truth for the scene is available, the input views were downsampled to lower resolution before performing super-resolution and compared against the original view. The images on the right show the estimated disparity maps for the two scenes.



5 × 5 input views super-resolved to 9 × 9 super-resolved to 17 × 17

Fig. 12. *Upsampling of epipolar plane images (EPIs).* The left image shows the five layers of an epipolar plane image of the input data set with 5 × 5 views. We generate intermediate views using our method to achieve angular super-resolution. One can observe the high quality and accurate occlusion boundaries of the resulting view interpolation. Indeed, they are accurate enough such that using the upsampled EPIs leads to a further improvement in depth estimation accuracy, see figure 10.



(a) Demo (b) Motor

Fig. 13. *Super-resolution view synthesis using light fields from a plenoptic camera.* Scenes were recorded with a Raytrix camera at a resolution of 962 × 628 and super-resolved by a factor of 3 × 3. The light field contains 9 × 9 views. Numerical quality of the estimate is computed in figure 11. The top image shows the full novel view, in the bottom rows can bee seen (from top to bottom): closeups of a low-resolution input view, the high-resolution view obtained by bilinear interpolation, and the super-resolved result. One can clearly make out additional detail, for example the diagonal stripes in the Euro note, which were not visible before.

accurate matching, super-resolution would not be possible. Figures 1 and 8 show images from the input light fields, estimated depth maps and super-resolved novel views. Note that due to file size limit, the PDF resolution is not high enough to observe the increased details of the super-resolved results in these figures, instead see figure 9 for closeups.

Figures 11 and 13 show the results of the same set of experiments (without ground truth data, which is not available) for two real-world scenes captured with the Raytrix plenoptic camera. We can see that the algorithm allows to accurately reconstruct both subpixel disparity as well as a high-quality super-resolved intermediate view.

Depth Refinement. In figure 7, the accuracy of the disparity maps is shown numerically in a comparison to the ground truth. They clearly become better with higher angular resolution. Consequently, we apply the idea of re-computing the disparity maps with synthesized intermediate views. We first synthesize novel views to increase angular resolution by a factor of 2 and 4. Figure 12 shows resulting epipolar plane images, which can be seen to be of high quality with accurate occlusion boundaries. Nevertheless, it is highly interesting that the quality of the disparity map increases significantly when recomputed with the super-resolved light field, figure 10. This is a striking result, since one would expect that the intermediate views reflect the error in the original disparity maps. However, they actually provide more accuracy than a single disparity map, since they represent a consensus of all input views.

6 Conclusions

We substantially extended the mathematical framework for *variational light field analysis* by introducing a variational model for super-resolution view synthesis. Compared to previous work on novel view generation, this allows us to analytically derive weighting factors for the contributions of the input views caused by foreshortening effects due to scene geometry. Experiments on synthetic ground truth as well as real-world images from a recent plenoptic camera give numerical evidence for the high quality of the method. In conjunction with a non-classical approach to disparity estimation which exploits the continuous structure of the disparity space [2], we compute super-resolved disparity maps at sub-pixel precision even for complex light fields with highly specular objects. Notably, the quality of view synthesis is good enough to further improve the disparity estimate, which improves with higher angular resolution.

References

1. Bolles, R., Baker, H., Marimont, D.: Epipolar-plane image analysis: An approach to determining structure from motion. International Journal of Computer Vision 1, 7–55 (1987) 1, 6
2. Authors: -. In: Additional Material, reference_2.pdf (2012, under review) 1, 3, 6, 8, 9, 10, 11, 12

3. Vaish, V., Wilburn, B., Joshi, N., Levoy, M.: Using plane + parallax for calibrating dense camera arrays. In: Proc. International Conference on Computer Vision and Pattern Recognition (2004) 1
4. Levoy, M., Hanrahan, P.: Light field rendering. In: Proc. SIGGRAPH, pp. 31–42 (1996) 1, 2, 8
5. Perwass, C., Wietzke, L.: The next generation of photography (2010) 1, www.raytrix.de
6. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Technical Report CSTR 2005-02, Stanford University (2005) 1
7. Bishop, T., Favaro, P.: Full-Resolution Depth Map Estimation from an Aliased Plenoptic Light Field. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part II. LNCS, vol. 6493, pp. 186–200. Springer, Heidelberg (2011) 1, 2
8. Georgiev, T., Lumsdaine, A.: Focused plenoptic camera and rendering. Journal of Electronic Imaging 19, 021106 (2010) 1
9. McMillan, L., Bishop, G.: Plenoptic modeling: An image-based rendering system. In: Proc. SIGGRAPH, pp. 39–46 (1995) 1, 2
10. Protter, M., Elad, M.: Super-resolution with probabilistic motion estimation. IEEE Trans. Image Processing 18, 1899–1904 (2009) 1, 2, 8
11. Goldluecke, B., Cremers, D.: Superresolution texture maps for multiview reconstruction. In: Proc. ICCV (2009) 1, 2, 3, 5
12. Adelson, E., Bergen, J.: The plenoptic function and the elements of early vision. Computational Models of Visual Processing 1 (1991) 2
13. Gortler, S., Grzeszczuk, R., Szeliski, R., Cohen, M.: The Lumigraph. In: Proc. SIGGRAPH, pp. 43–54 (1996) 2
14. Shum, H., Chan, S., Kang, S.: Image-based rendering. Springer, New York (2007) 2
15. Kubota, A., Aizawa, K., Chen, T.: Reconstructing dense light field from array of multifocus images for novel view synthesis. IEEE Trans. Image Processing 16, 269–279 (2007) 2, 8
16. Chai, J.X., Tong, X., Chany, S.C., Shum, H.Y.: Plenoptic sampling. In: Proc. SIGGRAPH, pp. 307–318 (2000) 2, 4
17. Siu, A., Lau, E.: Image registration for image-based rendering. IEEE Trans. Image Processing 14, 241–252 (2005) 2
18. Geys, I., Koninckx, T.P., Gool, L.V.: Fast interpolated cameras by combining a GPU based plane sweep with a max-flow regularisation algorithm. In: 3DPVT, pp. 534–541 (2004) 2, 8
19. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Trans. on Pattern Analysis and Machine Intelligence 24, 1167–1183 (2002) 3, 4
20. Beck, A., Teboulle, M.: Fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sciences 2, 183–202 (2009) 8, 9
21. Bermùdez, A., Moreno, C.: Duality methods for solving variational inequalities. Comp. and Maths. with Appls. 7, 43–58 (1981) 9