Full-Resolution Depth Map Estimation from an Aliased Plenoptic Light Field

Tom E. Bishop and Paolo Favaro

Department of Engineering and Physical Sciences
Heriot-Watt University, Edinburgh, UK
{t.e.bishop,p.favaro}@hw.ac.uk

Abstract. In this paper we show how to obtain full-resolution depth maps from a single image obtained from a plenoptic camera. Previous work showed that the estimation of a low-resolution depth map with a plenoptic camera differs substantially from that of a camera array and, in particular, requires appropriate depth-varying antialiasing filtering. In this paper we show a quite striking result: One can instead recover a depth map at the same full-resolution of the input data. We propose a novel algorithm which exploits a photoconsistency constraint specific to light fields captured with plenoptic cameras. Key to our approach is handling missing data in the photoconsistency constraint and the introduction of novel boundary conditions that impose texture consistency in the reconstructed full-resolution images. These ideas are combined with an efficient regularization scheme to give depth maps at a higher resolution than in any previous method. We provide results on both synthetic and real data.

1 Introduction

Recent work [1, 2] demonstrates that the depth of field of Plenoptic cameras can be extended beyond that of conventional cameras. One of the fundamental steps in achieving such result is to recover an accurate depth map of the scene. [3] introduces a method to estimate a low-resolution depth map by using a sophisticated antialiasing filtering scheme in the multiview stereo framework. In this paper, we show that the depth reconstruction can actually be obtained at the full-resolution of the sensor, i.e., we show that one can obtain a depth value at each pixel in the captured light field image. We will show that the full-resolution depth map contains details not achievable with simple interpolation of a low-resolution depth map.

A fundamental ingredient in our method is the formulation of a novel photoconsistency constraint, specifically designed for a Plenoptic camera (we will also use the term light field (LF) camera as in [4]) imaging a Lambertian scene (see Fig. 1). We show that a point on a Lambertian object is typically imaged under several microlenses on a regular lattice of correspondences. The geometry of such lattice can be easily obtained by using a Gaussian optics model of the Plenoptic camera. We will see that this leads to the reconstruction of full-resolution views

R. Kimmel, R. Klette, and A. Sugimoto (Eds.): ACCV 2010, Part II, LNCS 6493, pp. 186–200, 2011.
© Springer-Verlag Berlin Heidelberg 2011

consisting of mosaiced tiles from the sampled LF. Notice that in the data that we have used in the experiments, the aperture of the main lens is a disc. This causes the square images under each microlens to be split into a disc containing valid measurements and 4 corners with missing data (see one microlens image in the right image in Fig. 1). We show that images with missing data can also be reconstructed quite well via interpolation and inpainting (see e.g. [5]). Such interpolation however is not necessary if one uses a square main lens aperture and sets the F-numbers of the microlenses and the main lens appropriately. More importantly, we will show that the reconstructed samples form tiled mosaics, and one can enforce consistency at tile edges by introducing additional boundary conditions constraints. Such constraints are fundamental in obtaining the full-resolution depth estimates. Surprisingly, no similar gain in depth resolution is possible with a conventional camera array. In a camera array the views are not aliased (as the CCD sensors have contiguous pixels), and applying any existing method on such data (including the one proposed here) only results in upsampling of the low-resolution depth map (which is just the resolution of one view, and not the total number of captured pixels). This highlights another novel advantage of LF cameras.

1.1 Prior Work and Contributions

The space of light rays in a scene that intersect a plane from different directions may be interpreted as a 4D light field (LF) [6], a concept that has been useful in refocusing, novel view synthesis, dealing with occlusions and non-Lambertian objects among other applications. Various novel camera designs that sample the LF have been proposed, including multiplexed coded aperture [7], Heterodyne mask-based cameras [8], and external arrangements of lenses and prisms [9]. However these systems have disadvantages including the inability to capture dynamic scenes [7], loss of light [8], or poor optical performance [9]. Using an array of cameras [10] overcomes these problems, at the expense of a bulky setup. *Light Field*, or *Plenoptic*, cameras [4, 11, 12], provide similar advantages, in more compact portable form, and without problems of synchronisation and calibration. LF cameras have been used for digital refocusing [13] capturing extended depth-of-field images, although only at a (low) resolution equal to the number of microlenses in the device. This has been addressed by super-resolution methods [1, 2, 14], making use of priors that exploit redundancies in the LF, along with models of the sampling process to formulate high resolution refocused images.

The same sampling properties that enable substantial image resolution improvements also lead to problems in depth estimation from LF cameras. In [3], a method for estimating a depth map from a LF camera was presented. Differences in sampling patterns of light fields captured by camera arrays and LF cameras were analysed, revealing how spatial aliasing of the samples in the latter case cause problems with the use of traditional multi-view depth estimation methods.

Simply put, the LF views (taking one sample per microlens) are highly undersampled for many depths, and thus in areas of high frequency texture, an error term based on matching views will fail completely. The solution proposed in [3]

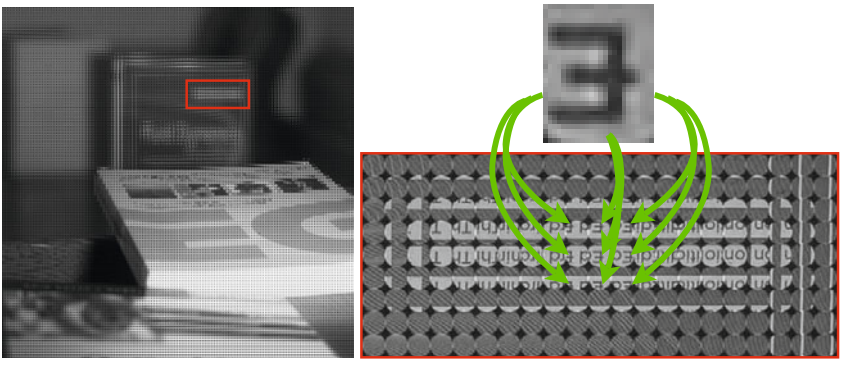

Fig. 1. Example of LF photoconsistency. Left: A LF image (courtesy of [15]) with an enlarged region (marked in red) shown to the right. Right: The letter E on the book cover is imaged (rotated by 180°) under nine microlenses. Due to Lambertianity each repetition has the same intensity. Notice that as depth changes, the photoconsistency constraint varies by increasing or decreasing the number of repetitions.

was to filter out the aliased signal component, and perform matching using the remaining valid low-pass part. The optimal filter is actually depth dependent, so an iterative scheme was used to update the filtering and the estimated depth map. Whilst this scheme yields an improvement in areas where strong aliasing corrupts the matching term, it actually throws away useful detail that could be used to improve matches if interpreted correctly.

Concurrent with this work, [16] proposed estimation of depth maps from LF data using cross-correlation, however this also only gives low resolution results.

In this paper we propose to perform matching directly on the sampled light-field data rather than restricting the interpretation to matching views in the traditional multi-view stereo sense. In this way, the concept of view aliasing is not directly relevant and we bypass the need to antialias the data. A key benefit of the new formulation is that we compute depth maps at a higher resolution than the sampled views. In fact the amount of improvement attainable depends on the depth in the scene, similar to superresolution restoration of the image texture as in [2]. Others have considered super-resolution of surface texture along with 3D reconstruction [17] and depth map super-resolution [18], however in the context of multi-view stereo or image sequences, where the sampling issues of a LF camera do not apply. Specifically, our contributions are:

1. A novel method to estimate a depth value at each pixel of a single (LF) image;
2. Analysis of the subimage correspondences, and reformulation in terms of virtual full-resolution views, i.e., mosaicing, leading to a novel photoconsistency term for LF cameras;
3. A novel penalty term that enforces gradient consistency across tile boundaries of mosaiced full-resolution views.

2 Light Field Representation and Methodology Overview

We consider a plenoptic camera, obtained by adding a microlens array in front of a conventional camera's sensor [1, 2, 4]. The geometry of this camera is almost identical to that of a camera array where each camera's aperture corresponds to a microlens. However, the LF camera has an additional main lens in front of the array, and this dramatically affects the sampling of the LF. We will see in sec. 3 that the precise study of how the spatial and angular coordinates sample the light field is critical to the formulation of matching terms for depth estimation.

Let us consider a version of the imaging system where the microlens apertures are small and behave as pinholes.[1] In the LF camera, each microlens forms its own *subimage* $S_{\boldsymbol{c}}(\boldsymbol{\theta})$ on the sensor, where $\boldsymbol{c} \in \mathbb{R}^2$ identifies the center of a microlens and $\boldsymbol{\theta} \in \mathbb{R}^2$ identifies the *local* (angular) coordinates under such a microlens. In a real system the microlens centers are located at discrete positions $\boldsymbol{c_k}$, but to begin the analysis we generalize the definition to *virtual* microlenses that are free to be centered in the continuous coordinates $\boldsymbol{c}$.

The main lens maps an object in space to a *conjugate object* inside the camera. Then, pixels in each subimage see light from the conjugate object from different vantage points, and have a one-to-one map with points on the main lens (see Fig. 2). The collection of pixels, one per microlens, that map to the same point on the main lens (*i.e.*, with the same local angle $\boldsymbol{\theta}$) form an image we call a *view*, denoted by $V_{\boldsymbol{\theta}}(\boldsymbol{c}) \equiv S_{\boldsymbol{c}}(\boldsymbol{\theta})$. The views and the subimages are two equivalent representations of the same quantity, the sampled light field.

We assume the scene consists of Lambertian objects, so that we can represent the continuous light field with a function $r(\boldsymbol{u})$ that is independent of viewing angle. In our model, $r(\boldsymbol{u})$ is called the radiance image or scene texture, and relates to the LF via a reprojection of the coordinates $\boldsymbol{u} \in \mathbb{R}^2$ at the microlens plane, through the main lens center into space. That is, $r(\boldsymbol{u})$ is the full-resolution all-focused image that would be captured by a standard pinhole camera with the sensor placed at the microlens plane. This provides a common reference frame, independent of depth. We also define the depth map $z(\boldsymbol{u})$ on this plane.

2.1 Virtual View Reconstructions

Due to undersampling, the plenoptic views are not suitable for interpolation. Instead we may interpolate within the angular coordinates of the subimages. In our proposed depth estimation method, we make use of the fact that in the Lambertian light field there are samples, indexed by $\boldsymbol{n}$, in different subimages that can be matched to the same point in space. In fact, we will explain how several estimates $\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})$ of the radiance can be obtained, one per set of samples with common $\boldsymbol{n}$. We term these full resolution virtual views, because they make use of more samples than the regular views $V_{\boldsymbol{\theta}}(\boldsymbol{c})$, and are formed from interpolation

[1] Depending on the LF camera settings, this holds true for a large range of depth values. Moreover, we argue in sec. 4 that corresponding defocused points are still matched correctly and so the analysis holds in general for larger apertures.

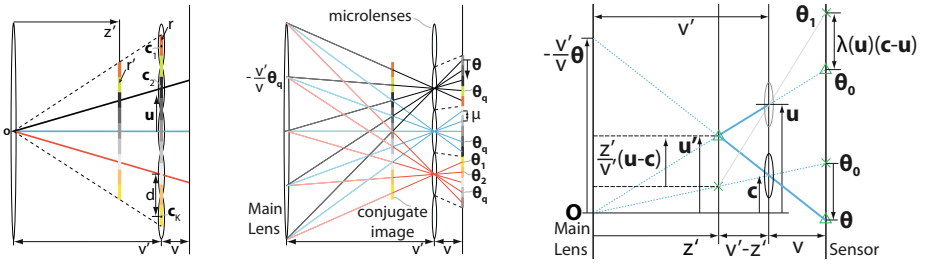

Fig. 2. Coordinates and sampling in a plenoptic camera. Left: In our model we begin with an image r at v', and scale it down to the conjugate image r' at z', where $\boldsymbol{O}$ is the main lens optical center. Center: This image at r' is then imaged behind each microlens; for instance, the blue rays show the samples obtained in the subimage under the central microlens. The central view is formed from the sensor pixels hit by the solid bold rays. Aliasing of the views is present: the spatial frequency in r is higher than the microlens pitch d, and clearly the adjacent view contains samples unrelated to the central view via interpolation. Right: Relation between points in r and corresponding points under two microlenses. A point $\boldsymbol{u}$ in r has a conjugate point (the triangle) at $\boldsymbol{u}'$, which is imaged by a microlens centered at an arbitrary center $\boldsymbol{c}$, giving a projected point at angle $\boldsymbol{\theta}$. If we then consider a second microlens positioned at $\boldsymbol{u}$, the image of the same point is located at the central view, $\boldsymbol{\theta} = \boldsymbol{0}$, on the line through $\boldsymbol{O}$.

in the non-aliased angular coordinates. Similarly to [1], these virtual views may be reconstructed in a process equivalent to building a mosaic from parts of the subimages (see Fig. 3). Portions of each subimage are scaled by a *magnification factor* λ that is a function of the depth map, and these tiles are placed on a new grid to give a mosaiced virtual view for each $\boldsymbol{n}$. As λ changes, the size of the portions used and the corresponding number of valid $\boldsymbol{n}$ are also affected. We will describe in detail how these quantities are related and how sampling and aliasing occurs in the LF camera in sec. 3, but firstly we describe the overall form of our depth estimation method, and how it uses the virtual views.

2.2 Depth Estimation via Energy Minimization

We use an energy minimization formulation to estimate the optimal depth map $\boldsymbol{z} = \{z(\boldsymbol{u})\}$ in the full-resolution coordinates $\boldsymbol{u}$. The optimal $\boldsymbol{z}$ is found as the minimum of the energy E defined as:

$$E(\boldsymbol{z}) = \sum_{\boldsymbol{u}} E_1(\boldsymbol{u}, z(\boldsymbol{u})) + \gamma_2 E_2(\boldsymbol{u}, z(\boldsymbol{u})) + \gamma_3 E_3(z(\boldsymbol{u})) \tag{1}$$

where $\gamma_2, \gamma_3 > 0$. This energy consists of three terms. Firstly, E_1 is a data term, penalizing depth hypotheses where the corresponding points in the light field do not have matching intensities. For the plenoptic camera, the number of virtual views $\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})$ depends on the depth hypothesis. We describe this term in sec. 4.

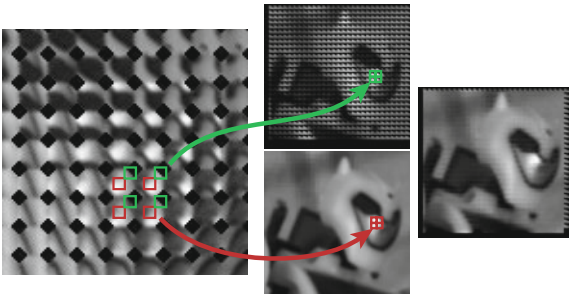

Fig. 3. Mosaiced full-resolution views generated from subimages. Each microlens subimage in a portion of the light field image shown on the left is partitioned into a grid of tiles. Rearranging the tiles located at the same offset from the microlens center into a single image produces mosaiced full-resolution views as shown in the middle. As we used a Plenoptic camera with a circular main lens aperture, some pixels are not usable. In the top view, the tiles are taken from the edge of each microlens and so there is missing data which is inpainted and shown on the image to the right. We use these full-resolution views to perform accurate data matching for depth reconstruction.

E_2 is a new matching term we introduce in sec. 4.1, enforcing the restored views to have proper boundary conditions. It effectively penalizes mismatches at the mosaiced tile boundaries, by imposing smoothness of the reconstructed texture (that also depends on the depth map $\boldsymbol{z}$).

The third term, E_3, is a regularization term that enforces piecewise smoothness of $\boldsymbol{z}$. Because the terms E_1 and E_2 only depend on $\boldsymbol{z}$ in a pointwise manner, they may be precomputed for a set of depth hypotheses. This enables an efficient numerical algorithm to minimize eq. (1), which we describe in sec. 4.3. Before detailing the implementation of these energy terms in sec. 4, we show in sec. 3 how such views are obtained from the sampled light field.

3 Obtaining Correspondences in Subimages and Views

We summarise here the image formation process in the plenoptic camera, and the relations between the radiance, subimages and views. We consider first a continuous model and then discretise, explaining why aliasing occurs. In our model, we work entirely inside the camera, beginning with the full-resolution radiance image $r(\boldsymbol{u})$, and describe the mapping to the views $V_{\boldsymbol{\theta}}(\boldsymbol{c})$ or microlens subimages $S_{\boldsymbol{c}}(\boldsymbol{\theta})$. This mapping is done in two steps: from $r(\boldsymbol{u})$ to the conjugate object (or image) at z', which we denote r', and then from r' to the sensor. This process is shown in Fig. 2. There is a set of coordinates $\{\boldsymbol{\theta}, \boldsymbol{c}\}$ that corresponds to the same $\boldsymbol{u}$, and we will now examine this relation.

We see in Fig. 2 that the projection of $r(\boldsymbol{u})$ to $r'(\boldsymbol{u}')$ through the main lens center $\boldsymbol{O}$ results in $\boldsymbol{u}' = \frac{z'(\boldsymbol{u})}{v'}\boldsymbol{u}$, where v' is the main lens to microlens array distance. This is then imaged onto the sensor at $\boldsymbol{\theta}(\boldsymbol{u}, \boldsymbol{c})$ through the microlens

at $\boldsymbol{c}$ with a scaling of $-\frac{v}{v'-z'(\boldsymbol{u})}$. Notice that $\boldsymbol{\theta}(\boldsymbol{u},\boldsymbol{c})$ is defined as the projection of $\boldsymbol{u}'$ through $\boldsymbol{c}$ onto the sensor relative to the projection of $\boldsymbol{O}$ through $\boldsymbol{c}$ onto the sensor; hence, we have

$$\begin{aligned}\boldsymbol{\theta}(\boldsymbol{u},\boldsymbol{c}) &= \frac{v}{v'-z'(\boldsymbol{u})}\frac{z'(\boldsymbol{u})}{v'}(\boldsymbol{c}-\boldsymbol{u}) \\ &= \lambda(\boldsymbol{u})(\boldsymbol{c}-\boldsymbol{u}),\end{aligned} \tag{2}$$

where we defined the *magnification factor* $\lambda(\boldsymbol{u}) = \frac{v}{v'-z'(\boldsymbol{u})}\frac{z'(\boldsymbol{u})}{v'}$, representing the signed scaling between the regular camera image $r(\boldsymbol{u})$ that would form at the microlens array, and the actual subimage that forms under a microlens. We can then use (2) to obtain the following relations:

$$\begin{aligned} r(\boldsymbol{u}) &= S_{\boldsymbol{c}}(\boldsymbol{\theta}(\boldsymbol{u},\boldsymbol{c})) = S_{\boldsymbol{c}}\Big(\lambda(\boldsymbol{u})(\boldsymbol{c}-\boldsymbol{u})\Big), \\ &= V_{\lambda(\boldsymbol{u})(\boldsymbol{c}-\boldsymbol{u})}(\boldsymbol{c}), \end{aligned} \qquad \forall \boldsymbol{c} \text{ s.t. } |\boldsymbol{\theta}(\boldsymbol{u},\boldsymbol{c})| < \frac{v}{v'}\frac{D}{2} \tag{3}$$

$$r(\boldsymbol{u}) = V_{\boldsymbol{\theta}}(\boldsymbol{c}(\boldsymbol{\theta},\boldsymbol{u})) = V_{\boldsymbol{\theta}}\left(\boldsymbol{u} + \frac{\boldsymbol{\theta}}{\lambda(\boldsymbol{u})}\right) \qquad \forall\, |\boldsymbol{\theta}| < \frac{v}{v'}\frac{D}{2}. \tag{4}$$

The constraints on $\boldsymbol{c}$ or $\boldsymbol{\theta}$ mean only microlenses whose projection $-\frac{v'}{v}\boldsymbol{\theta}$ is within the main lens aperture, of radius $\frac{D}{2}$, will actually image the point $\boldsymbol{u}$. Equations (3) and (4) show respectively how to find values corresponding to $r(\boldsymbol{u})$ in the light field for a particular choice of either $\boldsymbol{c}$ or $\boldsymbol{\theta}$, by selecting the other variable appropriately. This does not pose a problem if both $\boldsymbol{c}$ and $\boldsymbol{\theta}$ are continuous, however in the LF camera this is not the case and we must consider the implications of interpolating sampled subimages at $\boldsymbol{\theta} = \lambda(\boldsymbol{u})(\boldsymbol{c}-\boldsymbol{u})$ or sampled views at $\boldsymbol{c} = \boldsymbol{u} + \frac{\boldsymbol{\theta}}{\lambda(\boldsymbol{u})}$.

3.1 Brief Discussion on Sampling and Aliasing in Plenoptic Cameras

In a LF camera with microlens spacing d, only a discrete set of samples in each view is available, corresponding to the microlens centers at positions $\boldsymbol{c} = \boldsymbol{c_k} \doteq d\boldsymbol{k}$, where $\boldsymbol{k}$ indexes a microlens. Also, the pixels (of spacing μ) in each subimage sample the possible views at angles $\boldsymbol{\theta} = \boldsymbol{\theta_q} \doteq \mu\boldsymbol{q}$, where $\boldsymbol{q}$ is the pixel index; these coordinates are local to each microlens, with $\boldsymbol{\theta_0}$ the projection of the main lens center through each microlens onto the sensor. Therefore, we define the discrete observed view $\hat{V}_{\boldsymbol{q}}(\boldsymbol{k}) \doteq V_{\boldsymbol{\theta_q}}(\boldsymbol{c_k})$ at angle $\boldsymbol{\theta_q}$ as the image given by the samples for each $\boldsymbol{k}$. We can also denote the sampled subimages as $\hat{S}_{\boldsymbol{k}}(\boldsymbol{q}) \doteq V_{\boldsymbol{\theta_q}}(\boldsymbol{c_k})$.

For a camera array, the spatial samples are not aliased, and a virtual view $V_{\boldsymbol{\theta}}(\boldsymbol{c})$ may be reconstructed accurately, which allows for the sub-pixel matching used in regular multi-view stereo. However, in the LF camera, there is clearly a large jump between neighboring samples from the same view in the conjugate image (see Fig. 2). Even when the microlenses have apertures sized equal to their spacing, as described in [3], the resulting microlens blur size changes with

depth. Equivalently there is a depth range where the integration region size in the conjugate image is less than the sample spacing. This means that the sampled views $\hat{V}_{\boldsymbol{q}}(\boldsymbol{k})$ may be severely aliased, depending on depth. As a result, simple interpolation of $\hat{V}_{\boldsymbol{q}}(\boldsymbol{k})$ in the spatial coordinates $\boldsymbol{k}$ is not enough to reconstruct an arbitrary virtual view $V_{\boldsymbol{\theta}}(\boldsymbol{c})$ at $\boldsymbol{c} \neq \boldsymbol{c_k}$.

3.2 Matching Sampled Views and Mosaicing

In an LF camera we can reconstruct $V_{\boldsymbol{\theta}}(\boldsymbol{c})$ and avoid aliasing in the sampled views by interpolating instead along the angular coordinates of the subimages, and using the spatial samples only at known $\boldsymbol{k}$. Using eq. (3), we can find the estimates $\hat{r}(\boldsymbol{u})$ of the texture by interpolating as follows:

$$\begin{aligned} & r(\boldsymbol{u}) = V_{\lambda(\boldsymbol{u})(\boldsymbol{c_k}-\boldsymbol{u})}(\boldsymbol{c_k}) \\ \Rightarrow\ & \hat{r}(\boldsymbol{u}) = \hat{V}_{\frac{\lambda(\boldsymbol{u})}{\mu}(\boldsymbol{c_k}-\boldsymbol{u})}(\boldsymbol{k}), \quad \forall \boldsymbol{k} \ \text{ s.t. } \ |\lambda(\boldsymbol{u})(\boldsymbol{c_k}-\boldsymbol{u})| < \frac{v}{v'}\frac{D}{2} \end{aligned} \tag{5}$$

Let us now expand $\boldsymbol{c_k}$ as $\boldsymbol{u} + \boldsymbol{\Delta u} + \boldsymbol{n}d$, where $\boldsymbol{\Delta u} \doteq \boldsymbol{c}_{\boldsymbol{k}_0} - \boldsymbol{u}$ is the vector from $\boldsymbol{u}$ to the closest microlens center $\boldsymbol{c}_{\boldsymbol{k}_0} \doteq \boldsymbol{u} - \text{ mod } (\boldsymbol{u}, d)$. We can now enumerate the different possible reconstructed images $\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})$ for all valid choices of $\boldsymbol{n} \in \mathbb{Z}^2$ that lead to the condition on $\boldsymbol{c_k}$ being satisfied. Then, we find

$$\hat{r}_{\boldsymbol{n}}(\boldsymbol{u}) = \hat{V}_{\frac{\lambda(\boldsymbol{u})}{\mu}(\boldsymbol{\Delta u}+\boldsymbol{n}d)}\left(\boldsymbol{k}_0 + \boldsymbol{n}\right), \quad \forall \boldsymbol{n} \ \text{ s.t. } \ |\lambda(\boldsymbol{u})(\boldsymbol{\Delta u}+\boldsymbol{n}d)| < \frac{v}{v'}\frac{D}{2} \tag{6}$$

The reconstructed virtual views $\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})$ for different $\boldsymbol{n}$ should match at position $\boldsymbol{u}$, if we have made the right hypothesis for the depth $z(\boldsymbol{u})$ and hence $\lambda(\boldsymbol{u})$ (notice that given $\lambda(\boldsymbol{u})$ one can immediately obtain $z(\boldsymbol{u})$ and vice versa). This interpretation is similar to that of the method used in [1], except that rather than averaging together the full-resolution mosaics $\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})$, we keep them separated.

4 Energy Terms for Full-Resolution Depth Maps

We now describe the energy terms in sec. 2.2. Firstly, we define the matching term E_1 that penalizes differences between the different virtual views[2] $\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})$ as

$$E_1\left(\boldsymbol{u}, z(\boldsymbol{u})\right) = \frac{1}{W(\boldsymbol{u})} \sum_{\substack{\forall \boldsymbol{n} \neq \boldsymbol{n}' \\ \boldsymbol{n}, \boldsymbol{n}' \in \mathbb{Z}^2 \\ \boldsymbol{n} \text{ s.t. } |\lambda(\boldsymbol{u})(\boldsymbol{\Delta u}+\boldsymbol{n}d)| < \frac{v}{v'}\frac{D}{2} \\ \boldsymbol{n}' \text{ s.t. } \left|\lambda(\boldsymbol{u})(\boldsymbol{\Delta u}+\boldsymbol{n}'d)\right| < \frac{v}{v'}\frac{D}{2}}} \left|\hat{r}_{\boldsymbol{n}}(\boldsymbol{u}) - \hat{r}_{\boldsymbol{n}'}(\boldsymbol{u})\right|. \tag{7}$$

$W(\boldsymbol{u})$ is a normalization term, counting the number of valid pairs in the sum.

The number of possible $\boldsymbol{n}$, and hence number of virtual views, changes based on the magnification factor λ, according to the depth hypothesis. Therefore

[2] Interestingly, unlike matching low resolution views in multi-view stereo, the virtual views are already aligned to each other in regions with the correct depth assumption.

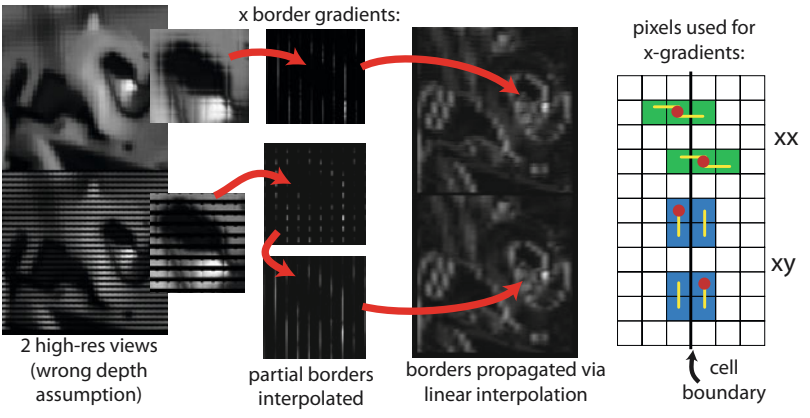

Fig. 4. Left: The energy term E_2 penalizes unnatural discontinuities in the reconstructed full-resolution views, at the mosaic cell boundaries. Interpolation is used to complete partial views, and interior pixels from the borders. **Right:** The x-border gradients at the red pixels use the supports shown, using the difference between the pairs of gradients given between the yellow pairs.

$W(\boldsymbol{u})$ varies with both position and depth hypothesis. Also, when using a Plenoptic camera with a circular main lens aperture, we need to deal with missing data. In this case, parts of some of the views will be missing where tiles were taken from the borders of a microlens (see Fig. 3). For this reason, we consider using inpainting (we use a PDE implementation in MATLAB[3]) to slightly reduce the amount of missing data, *i.e.*, to extend the field of view of each microlens based on extrapolating the data contained in neighboring lenses.

Furthermore, rather than comparing all views with each other as in eq. (7), we have found that using a penalty in E_1 that compares each mosaic $\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})$ (at valid locations) with the median obtained by inpainting each $\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})$ can be more robust, and also considerably faster for large W.

We have made the assumption that microlens blur is insignificant at working depths. If blur is present, it will be the same amount for each point being matched, and will not cause any difference for reasonably planar surfaces. In the case of a depth transition, highly textured surfaces will appear slightly larger than they really are, due to matching occurring at all points within the true point blur disc. Ideally, we could also deconvolve each mosaic independently before matching, however, this is beyond the scope of this paper.

4.1 Mosaiced Boundary Condition Consistency

Given the full-resolution mosaiced views, $\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})$, we may also establish a self-consistency criterion in E_2. This term says that each reconstructed $\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})$ should have consistent boundary conditions across blocks within each view. That is, the mosaicing process should not introduce unnatural jumps if the depth is estimated

[3] `http://www.mathworks.com/matlabcentral/fileexchange/4551`

correctly. Therefore we expect the x- and y-derivatives to be mostly smooth across boundaries (assuming the probability of an image edge coinciding with a mosaic tile along all its edges to be low), and set

$$E_2\left(z(\boldsymbol{u})\right) = \sum_{\substack{\boldsymbol{n}\in\mathbb{Z}^2 \\ \left|\boldsymbol{\theta}_{\boldsymbol{c_k}(\boldsymbol{n})}(\boldsymbol{u})\right| < \frac{v}{v'}\frac{D}{2} \\ \boldsymbol{u}\in\mathcal{R}_{\text{boundary-x}}}} |\nabla_{xx}\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})| + |\nabla_{xy}\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})| + \sum_{\substack{\boldsymbol{n}\in\mathbb{Z}^2 \\ \left|\boldsymbol{\theta}_{\boldsymbol{c_k}(\boldsymbol{n})}(\boldsymbol{u})\right| < \frac{v}{v'}\frac{D}{2} \\ \boldsymbol{u}\in\mathcal{R}_{\text{boundary-y}}}} |\nabla_{yx}\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})| + |\nabla_{yy}\hat{r}_{\boldsymbol{n}}(\boldsymbol{u})|, \tag{8}$$

where ∇_{ab} denotes the 2^{nd} order derivative with respect to the coordinates a and b, $\mathcal{R}_{\text{boundary-x}}$ includes the pixels on the left and right tile boundaries, and $\mathcal{R}_{\text{boundary-y}}$ the top and bottom boundaries (only non-inpainted regions). The restriction of this constraint to the boundary of the tiles is due to eq. (6). We observe that for a fixed $\boldsymbol{n}$, eq. (6) tells us that, as we vary $\boldsymbol{u}$, several consecutive samples are taken from the same microlens subimage at $\boldsymbol{k}_0$, but there is a discontinuity present in the reconstruction as $\boldsymbol{u}$ moves past a microlens boundary and $\boldsymbol{k}_0$ then jumps to a different microlens $\boldsymbol{k}$. Changing $\boldsymbol{n}$ will choose a different offset portion of each subimage to reconstruct (effectively a full-resolution virtual view, centered at an angle $\boldsymbol{q}_{\boldsymbol{n}} = \frac{d\lambda(\boldsymbol{u})}{\mu}\boldsymbol{n}$). Within each tile the data is continuous and just comes from a scaled version of the same subimage and so we should not expect to penalize differently for a different scaling/depth.

We can instead define $E_2(z(\boldsymbol{u}))$ for $\boldsymbol{u} \notin \mathcal{R}_{\text{boundary}}$ via bilinear interpolation of the values at the boundary. While this term does not help super-resolving the depth map, it helps enforcing its overall consistency.

4.2 Regularization of the Depth Map

Due to the ill-posedness of the depth estimation problem, we introduce regularization. Our focus here is not on the choice of prior, so for simplicity we use a standard total variation term (∇ denotes the 2D gradient with respect to $\boldsymbol{u}$):

$$E_3(z(\boldsymbol{u})) = \|\nabla z(\boldsymbol{u})\|_1. \tag{9}$$

4.3 Numerical Implementation

At this point we have all the necessary ingredients to work on the energy introduced in eq. (1). In the following we use the notation introduced earlier on where the depth map z is discretized as a vector $\boldsymbol{z} = \{z(\boldsymbol{u})\}$. Then, we pose the depth estimation problem as the following minimization

$$\hat{\boldsymbol{z}} = \arg\min_{\boldsymbol{z}} E(\boldsymbol{z}) = \arg\min_{\boldsymbol{z}} E_{\text{data}}(\boldsymbol{z}) + \gamma_3 E_3(\boldsymbol{z}) \tag{10}$$

where we combine E_1 and E_2 in the data term $E_{\text{data}} = E_1 + \gamma_2 E_2$; in our experiments the tradeoff between E_3 and E_{data} is fixed by $\gamma_2 = 1$ and $\gamma_3 = 10^{-3}$. We then minimize this energy by using an iterative solution. We first notice that the term E_{data} can be written as a sum of terms that depend on a single entry of $\boldsymbol{z}$ at a time. Hence, we find a first approximate solution by

ignoring the regularization term and performing a fast brute force search for each entry of $\boldsymbol{z}$ independently. This procedure yields the initial solution $\boldsymbol{z}_0$. Then, we approximate E_{data} via a Taylor expansion up to the second order, *i.e.*,

$$E_{\text{data}}(\boldsymbol{z}_{t+1}) \simeq E_{\text{data}}(\boldsymbol{z}_t) + \nabla E_{\text{data}}(\boldsymbol{z}_t)(\boldsymbol{z}_{t+1} - \boldsymbol{z}_t) + \frac{1}{2}(\boldsymbol{z}_{t+1} - \boldsymbol{z}_t)^T H E_{\text{data}}(\boldsymbol{z}_t)(\boldsymbol{z}_{t+1} - \boldsymbol{z}_t) \tag{11}$$

where ∇E_{data} is the gradient of E_{data}, HE_{data} is the Hessian of E_{data}, and $\boldsymbol{z}_t$ and $\boldsymbol{z}_{t+1}$ are the solutions at iteration t and $t+1$ respectively. To ensure that our local approximation is convex we let HE_{data} be diagonal and each element be positive by taking its absolute value (component-wise) $|HE_{\text{data}}(\boldsymbol{z}_t)|$. In the case of the regularization energy we use a Taylor expansion of its gradient up to the linear term. When we compute the Euler-Lagrange equations of the approximate energy E with respect to $\boldsymbol{z}_{t+1}$ this linearization results in

$$\nabla E_{\text{data}}(\boldsymbol{z}_t) + |HE_{\text{data}}(\boldsymbol{z}_t)|(\boldsymbol{z}_{t+1} - \boldsymbol{z}_t) - \gamma_3 \nabla \cdot \frac{\nabla(\boldsymbol{z}_{t+1} - \boldsymbol{z}_t)}{|\nabla \boldsymbol{z}_t|} = 0 \tag{12}$$

which is a linear system in the unknown $\boldsymbol{z}_{t+1}$, and can be solved efficiently via the Conjugate Gradient method.

5 Experiments

We begin by testing the performance of the method on artificially generated light fields, before considering data from a real LF camera.

Synthetic Data. We simulated data from a LF camera, with main lens of focal length $F = 80mm$ focused at $0.635m$, with parameters $d = 0.135mm$, $\mu = 0.09mm$, $v = 0.5mm$, $Q = 15$ and microlens focal length $f = 0.5mm$. To compare with the antialiasing method of [3], we estimate the view disparity $s(\boldsymbol{u}) \doteq \frac{\mu}{d\lambda(\boldsymbol{u})}$, in pixels per view, for a set of textured planes at known depths (the disparity between the views is zero for the main lens plane in focus, and increases with depth). We used the *Bark* texture from the Brodatz texture dataset, and in each case just used the energy without regularization. We plot the true value of s for each plane against the mean value from the depth maps found by both the method of [3] and the proposed method, along with error bars in Fig. 5.

Clearly in the case of no noise the proposed algorithm performs a lot more reliably with a smaller variance of the error, although with increasing noise the benefit over the method of [3] reduces. Both methods perform poorly around the main lens plane in focus: in the case of the method of [3], the disparity becomes smaller than the interpolation accuracy, whereas the proposed method begins to fail as the magnification factor tends towards 1, and there are not enough virtual views to form the median. This is not surprising as it relates to the undersampling of the space by the LF camera around these depths.

Real Data. We obtained datasets from a portable LF camera with the same system parameters as above. The scene range was 0.7–0.9m, corresponding to disparities of 0.16–0.52 pixels per view. One of the low resolution views is shown,

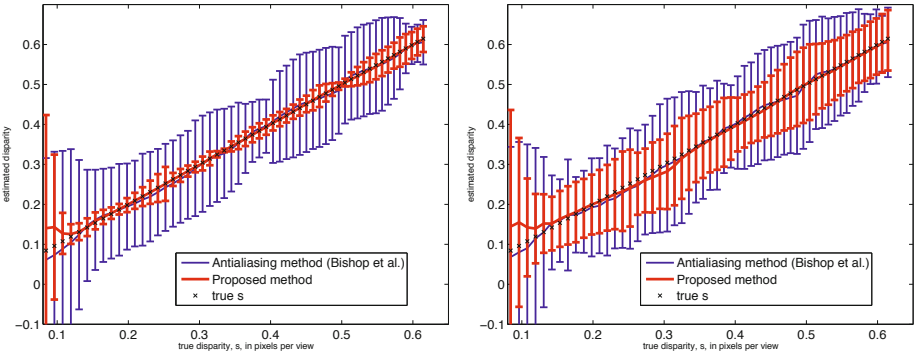

Fig. 5. Results on synthetic data. We plot the estimated disparity from a textured plane versus the true disparity s (which depends on the depth of the plane) for our method and the method from [3]. Error bars shown at 3 standard deviations. Left: data with no noise. Right: Data with noise standard deviation 2.5% of signal range.

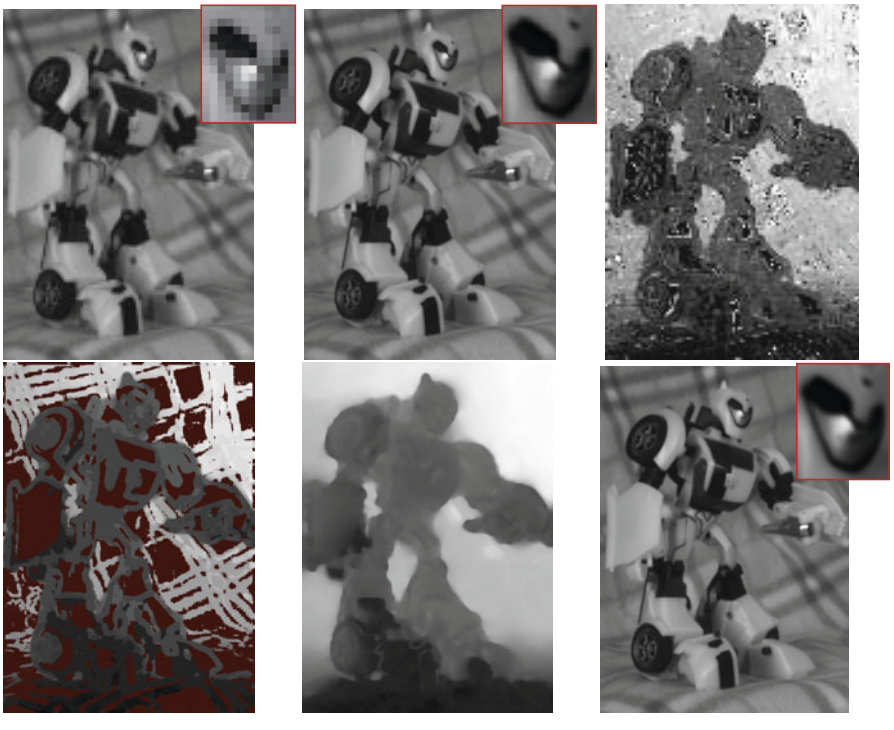

Fig. 6. Results on real data. Top row, left to right: One low resolution view from the light field; One full-resolution virtual view, with depth hypothesis in the middle of the scene range (note blocking artefacts); Low resolution depth map using multiview stereo method [3]; Bottom row: Unregularized full-resolution depth map (note that the textureless areas shown in red do not yield a depth estimate); After regularization; High resolution all-in-focus image obtained using the estimated depth map.

Fig. 7. Results on real data. From left to right: One low resolution view from the light field; Regularized depth estimate; High resolution image estimate from the light field, using estimated depth map. Source LF data in bottom row courtesty of [16].

along with a full-resolution virtual view used in the algorithm, and the estimated depthmap before and after regularization in Fig. 6 and Fig. 7. We also applied the method to data from [16] in the bottom row of Fig. 7. In each case, before applying regularization, we automatically select regions where the confidence is too low due to lack of texture (the energy has very small values for all depths at these pixels, but is estimated very reliably elsewhere). For this size image (2100×1500 pixel virtual views in the case of Fig. 6), our MATLAB implementation takes about half an hour to estimate E_1 and E_2, using up to $n = 37$ virtual views at the maximum depths, although solving eq. (12) only takes a couple of minutes. Notice the main burden is not in the regularisation scheme but simple interpolation and difference operations between a large number of views (e.g. 40*37=1480 views if using $n = 37$ and 40 depth levels), which could be done at least an order of magnitude faster using, for example, GPU acceleration.

6 Conclusions

We have demonstrated how depth can be estimated at each pixel of a single (LF) image. This is an unprecedented result that highlights a novel quality of Plenoptic cameras. We have analysed the sampling tradeoffs in LF cameras and shown

how to obtain valid photoconsistency constraints, which are unlike those for a regular multi-view matching. We also presented a novel boundary consistency constraint in reconstructed full-resolution views to penalize incorrect depths, and proposed a novel algorithm to infer full-resolution depth maps using these terms, with efficient regularization.

Acknowledgement. This work has been supported in part by EPSRC grant EP/F023073/1(P), grant EP/I501126/1, and SELEX/HWU/2010/SOW3.

References

1. Lumsdaine, A., Georgiev, T.: The focused plenoptic cameras. In: ICCP 2009 (IEEE International Conference on Computational Photography) (2009)
2. Bishop, T.E., Zanetti, S., Favaro, P.: Light field superresolution. In: ICCP 2009 (IEEE Int. Conference on Computational Photography) (2009)
3. Bishop, T.E., Favaro, P.: Plenoptic depth estimation from multiple aliased views. In: ICCV 2009 Workshops: 3DIM (The 2009 IEEE International Workshop on 3-D Digital Imaging and Modeling), Kyoto, Japan (2009)
4. 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)
5. Weickert, J., Welk, M.: Tensor field interpolation with pdes. In: Weickert, J., Hagen, H. (eds.) Visualization and Processing of Tensor Fields, pp. 315–325. Springer, Berlin (2006)
6. Levoy, M., Hanrahan, P.: Light field rendering. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, pp. 31–42. ACM Press, New York (1996)
7. Liang, C.K., Lin, T.H., Wong, B.Y., Liu, C., Chen, H.H.: Programmable aperture photography: Multiplexed light field acquisition. ACM Transactions on Graphics (Proc. SIGGRAPH 2008) 27 (2008)
8. Veeraraghavan, A., Raskar, R., Agrawal, A.K., Mohan, A., Tumblin, J.: Dappled photography: mask enhanced cameras for heterodyned light fields and coded aperture refocusing. ACM Trans. Graph (Proc. SIGGRAPH 2007) 26, 69 (2007)
9. Georgeiv, T., Zheng, K.C., Curless, B., Salesin, D., Nayar, S., Intwala, C.: Spatio-angular resolution tradeoff in integral photography. In: Akenine-Mller, T., Heidrich, W. (eds.) Eurographics Symposium on Rendering (2006), Adobe Systems, University of Washington, Columbia University (2006)
10. Vaish, V., Levoy, M., Szeliski, R., Zitnick, C., Kang, S.B.: Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In: Proc. CVPR 2006, vol. 2, pp. 2331–2338 (2006)
11. Adelson, E.H., Wang, J.Y.: Single lens stereo with a plenoptic cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence 14, 99–106 (1992)
12. Fife, K., El Gamal, A., Wong, H.S.: A 3d multi-aperture image sensor architecture. In: Custom Integrated Circuits Conference, CICC 2006, pp. 281–284. IEEE, Los Alamitos (2006)
13. Ng, R.: Fourier slice photography. ACM Trans. Graph. 24, 735–744 (2005)
14. Levin, A., Freeman, W.T., Durand, F.: Understanding camera trade-offs through a bayesian analysis of light field projections. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 88–101. Springer, Heidelberg (2008)

15. Georgiev, T., Lumsdaine, A.: Depth of field in plenoptic cameras. In: Alliez, P., Magnor, M. (eds.) Eurographics 2009 (2009)
16. Georgiev, T., Lumsdaine, A.: Reducing plenoptic camera artifacts. Computer Graphics Forum 29, 1955–1968 (2010)
17. Goldluecke, B., Cremers, D.: A superresolution framework for high-accuracy multiview reconstruction. In: IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, pp. 342–351 (2009)
18. Li, F., Yu, J., Chai, J.: A hybrid camera for motion deblurring and depth map super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8 (2008)