A Convex Solution to Disparity Estimation from Light Fields via the Primal-Dual Method

Mahdad Hosseini Kamal[1], Paolo Favaro[2], and Pierre Vandergheynst[1]

[1] Ecole Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland,
{mahdad.hosseinikamal,pierre.vandergheynst}@epfl.ch
[2] University of Bern, Bern 3012, Switzerland
paolo.favaro@iam.unibe.ch

Abstract. We present a novel approach to the reconstruction of depth from light field data. Our method uses dictionary representations and group sparsity constraints to derive a convex formulation. Although our solution results in an increase of the problem dimensionality, we keep numerical complexity at bay by restricting the space of solutions and by exploiting an efficient Primal-Dual formulation. Comparisons with state of the art techniques, on both synthetic and real data, show promising performances.

Keywords: Light fields, multi-view stereo, primal-dual formulation

1 Introduction

The estimation of a disparity map from multiple images is one of the very well studied problems in computer vision. Some of the most dramatic improvements in this field occurred with the introduction of novel numerical frameworks and their corresponding theory. A non-exhaustive list of such breakthroughs are the early work on space carving [20], the level set formulation and the variational framework [10], the Markov random field framework with polynomial-complexity solvers [6], the L_1-Total Variation optimization framework [35] and, more recently, convex formulations that aim for global optimality [25]. In this paper, we look at a novel approach based on recent primal-dual optimization techniques. Our approach is also convex as in the most recent developments, but we work with discrete labels (the possible disparity values).

Our formulation is based on a linear model of the data where a patch in an image is written as a linear combination of patches in other views. The key idea is that ideal Lambertian objects generate views that look alike (modulo foreshortening) and therefore corresponding patches live approximately on a 1D manifold. When objects are not Lambertian, they generate effects, such as specularities, that change with the pose of the camera. One can notice, however, that these effects are typically rare (*i.e.*, they happen only on some of the views) and spatially local. Hence, a natural way to model image patches of non Lambertian objects is by using an additive model where one of the two factors is sparse and the other is low-rank. If a finite set of possible depth candidates for a patch is

available, one can then verify which hypothesis best fits the low-rank + sparse model. Our strategy is therefore a competition between the different disparity hypotheses. We essentially allow the data to be explained by a simultaneous linear combination of **all** low-rank + sparse models. However, we force coefficients to focus on only a few of the models (where each model corresponds to a single disparity hypothesis) via group-sparsity penalty terms. We expect that coefficients be mostly non zero at the true disparity as this is the case that gives the fit with the sparsest set of outliers. Notice that the individual coefficients of each linear combination are not important, and indeed, typically, infinite solutions might be possible especially at the correct disparity. However, as long as coefficients have most non zero values at only one group, we can still correctly identify the disparity.

While this approach seems straightforward, in practice it faces considerable dimensionality challenges because data is replicated several times due to the patch-based model and the number of disparity hypotheses. This makes operations such as matrix inversion, often encountered in optimization schemes, impossible to carry out. To address these challenges we propose a primal-dual approach that results in simple element-wise thresholding operations and 2 (global) matrix multiplications at each step.

Contributions: We propose a framework to address the disparity estimation problem of light fields. In particular, we make the following contributions:

- We present a novel model for light field disparity estimation to represent a light field image patch as a linear combination of other light field patches. This representation satisfies a group sparse model and depends only on a group of light field patches of the same disparity.
- Occlusions are handled uniformly in our framework as a sparse component and this brings more robustness than in traditional matching methods.
- We introduce a robust and globally optimal solution for light field patch matching based on a preconditioned primal-dual algorithm [24], which allows to match a light field patch in all the views to estimate the disparity map.

2 Related work

Light field disparity estimation: Light fields can be captured using a camera array [30] or lenslet arrays [23] or as a sequence of images. One of the first approaches to compute light field depth exploits linear structures in light fields through a line fitting algorithm [5]. Other methods use more traditional stereo reconstruction techniques to match the corresponding pixels in light field images, such as block-matching techniques [4] or clustering methods to identify similar pixel matches [3, 11]. Ziegler et al. [36] proposed a Fourier-based technique to compute depth values. To achieve higher global coherence, light field depth estimation methods employ a global cost function to impose smoothness on the estimated depth values [8, 19, 32]. A limitation common to all these methods is that they optimize a global cost function that is not convex. Therefore, the estimated depth map depends on the initial input. Moreover, fine details are lost

because a coarse-to-fine multi-resolution technique is often used to avoid ending in weak local minima. Our approach overcomes these limitations by introducing a convex formulation.

Multiview stereo methods: Multiview techniques require detecting and handling outliers [2,16]. The difficulty of outlier modeling is due to the unstructured nature of errors produced by outliers. However, these errors can only influence a small part of the image and are therefore sparse in a canonical basis [2, 33]. An alternative to explicit occlusion modeling is to match only reliable pixels and fill the unmatched correspondences via regularization [18, 27]. However, as explained in [28], these methods are prone to artifacts. Multiview stereo methods employ a large number of images [13, 17] to compute the full geometry of a scene and often yield a smooth geometry. Our light field disparity estimation yields a representation that falls in the middle: it is more complete than in stereo techniques, but less than in multiview stereo.

Sparse representation: The similarity of image structures in a dataset is used in data clustering [9, 22] to determine the low-dimensional subspace of high dimensional data. Many schemes exploit data similarity to represent image correspondences in a dataset [21, 33]. In contrast to these clustering techniques, our proposed disparity estimation scheme looks for the best representation of each patch within a set of clusters. The clusters are generated from a number of disparity hypotheses, such that the members of a cluster are either chosen or discarded together. To achieve this we introduce a coupling term between the coefficients via group sparsity.

In this paper, we estimate disparity from light fields by representing patches of a desired light field view with an overcomplete dictionary. The elements of the dictionary are patches of other views reprojected back onto a reference view for a given set of disparity candidates. If sufficiently many patch samples are available, patches of the reference view can be written as a linear combination of patches from the correct disparity hypothesis. This representation is naturally group sparse, since only a single disparity candidate of the dictionary can be assigned to a given patch. This representation can be recovered efficiently via group sparsity minimization [34].

3 Multiple views and light fields

We consider capturing several images of the same static scene by translating a camera on the $x - y$ plane, where z is aligned to the camera optical axis, or, equivalently, by employing a camera array, or a plenoptic camera, where all the camera sensors lie on the same plane. More in general, we can describe the captured data as a 4D light field $L : \Omega \times \Theta \mapsto [0, +\infty)$ where $\Omega \equiv \mathbb{R}^{N \times M}$ denotes the spatial domain (the pixel coordinates within each image) and Θ the angular domain (the camera center coordinates). We consider cameras arranged in a regular lattice and denote with $\Delta = [\Delta_x \ \Delta_y]^T \in \mathbb{R}^2$ the displacement between a camera and its north-west neighbor. Then, we define $\Theta = \{[\Delta_x i \ \Delta_y j]^T | i = 1 \dots n, j = 1 \dots m\}$ as the 3D camera center of the (i, j)-th camera is located at

$[\Delta_x i \ \Delta_y j \ 0]^T$. For simplicity, we use the notation $L_{i,j}(x,y)$ to denote $L(x,y,i,j)$. A visible plane in the scene, parallel to the images planes of the cameras, will generate images in the light field L that are related to each other by a shift or *disparity* $\rho : \Omega \mapsto [0, +\infty)$, for simplicity we denote $\rho(x,y)$ by ρ. In formulas, this can be written as

$$L_{i,j}(x,y) = L_{p,q}(x - \rho\Delta_x(p-i), y - \rho\Delta_y(q-j)) \tag{1}$$

for all (x,y) that fall within the spatial domain of both light field views and for all (i,j) and (p,q) camera pairs.
A common approach to estimating the disparity ρ is then to pose a variational problem of the form

$$\min_{\rho} \sum_{\substack{i,j,p>i \\ q>j,x,y}} \Phi(L_{i,j}(x,y) - L_{p,q}(x - \rho(p-i)\Delta_x, y - \rho(q-j)\Delta_y)) + \Gamma(\rho), \tag{2}$$

where Φ is some robust penalty term for departures from zero and Γ is a regularization term for the unknown disparity ρ such as total variation. This problem is non convex and therefore finding the global optimum is a very challenging task. While good solutions have been obtained for the above problem, recent efforts have produced convex variational formulations [12, 25] with high-quality disparity reconstructions. Both of these methods work with continuous representations. However, one of the key differences between these two methods is that, while [25] achieves convexity by increasing the problem dimensionality, [12] achieves convexity by fixing the structure tensor with some initial approximate disparity estimate. Our method follows the strategy of the first approach and also results in a high-dimensional representation. However, we do not rely on any initial estimate (although it might considerably speed up the convergence). Moreover, as we describe in the next sections, our convex formulation is entirely in the discrete domain and exploits the quantization of the disparity values.

4 A patch-based image formation model

Our first step is to rewrite the problem (2) as a patch matching problem. Let us define the *patch operator* $\mathcal{P}_{x,y}$ as the mapping that extracts the $W \times W$ patch whose top-left corner lies at (x,y) of an image I, *i.e.*,

$$\mathcal{P}_{x,y}(I) = \{I(x + x_0, y + y_0)\}_{x_0,y_0=0,\ldots,W-1}. \tag{3}$$

We define the output of the patch operator to be a patch rearranged as a column vector whose W^2 elements have been rearranged in lexicographical order. Consider extracting one patch from each view of a light field, except for the (i_0, j_0)-th one (for example, this could be the central view), given a disparity ρ and collecting all the patches in a matrix $Q^{\rho}_{x,y} \in \mathbb{R}^{W^2 \times (nm-1)}$. This operation can be described via

$$Q^{\rho}_{x,y} = \{\mathcal{P}_{x-\rho\Delta_x(p-i_0), y-\rho\Delta_y(q-j_0)}(L_{p,q}) : \forall (p,q) \neq (i_0, j_0)\}. \tag{4}$$

If ρ is the true disparity of a fronto-parallel object in space, then all the columns in $Q^{\rho}_{x,y}$ will be identical to each other (in the ideal Lambertian case) and identical to the column vector $\mathcal{P}_{x,y}(L_{i_0,j_0})$. We also denote the latter vector with the symbol $Y_{x,y}$. More in general however, noise, non Lambertianity, shadows, occlusions, inter reflections and so on need to be taken into account. Since we believe that most of the time the Lambertian approximation will hold, we consider all the other image distortions as infrequent and use a sparse representation to model them, *i.e.*,

$$Y_{x,y} = Q^{\rho}_{x,y} C^{\rho}_{x,y} + E_{x,y} \tag{5}$$

where $C^{\rho}_{x,y}$ is a $nm-1$ column vector and $E_{x,y}$ is a W^2 column vector with few nonzero entries. The coefficients in $C^{\rho}_{x,y}$ determine the linear combination of vectors in $Q^{\rho}_{x,y}$ that generate $Y_{x,y}$. When the disparity ρ corresponds to the true solution, any $C^{\rho}_{x,y}$ such that $\mathbf{1}^T C^{\rho}_{x,y} = 1$ will satisfy the above equation. Vice versa, when the disparity is incorrect and the scene has sufficiently rich texture, there should not exist any vector $C^{\rho}_{x,y}$ that satisfies (5). Thus, we propose to force the disparity ρ to take values only from the set $\{\rho_1, \rho_2, \ldots, \rho_D\}$ and extend (5) to

$$Y_{x,y} = [Q^{\rho_1}_{x,y}\ Q^{\rho_2}_{x,y}\ \ldots\ Q^{\rho_D}_{x,y}][C^{\rho_1}_{x,y} C^{\rho_2}_{x,y} \cdots C^{\rho_D}_{x,y}]^T + E_{x,y} \doteq Q_{x,y} C_{x,y} + E_{x,y} \tag{6}$$

where the $W^2 \times (nm-1)D$ matrix $Q_{x,y}$ and the $(nm-1)D$ vector $C_{x,y}$ are implicitly defined by the equation to the right.

5 Depth estimation

Based on the model (5), a first formulation for estimating disparity through patch matching is

$$\min_{C,E} \frac{1}{2} \sum_{x,y} \|Y_{x,y} - Q_{x,y} C_{x,y} - E_{x,y}\|_2^2 + \mu \|E_{x,y}\|_1 \tag{7}$$

where $\mu > 0$ is a constant determining the degree of sparsity of $E_{x,y}$, $\|E_{x,y}\|_1$ denotes the ℓ^1 norm of $E_{x,y}$, and C and E are the column vectors obtained by stacking vertically all the vectors $C_{x,y}$ and $E_{x,y}$ respectively. Since the total number of patches within the image domain is $\tilde{M}\tilde{N}$, where $\tilde{M} = M - W + 1$ and $\tilde{N} = N - W + 1$, the E vector has $\tilde{M}\tilde{N}W^2$ elements and the C vector has $\tilde{M}\tilde{N}(nm-1)D$ elements.

As explained in the previous section, we aim at concentrating the coefficients of $C_{x,y}$ on the patches belonging to just one disparity hypothesis. If this is the case, then, given $C_{x,y}$, one can estimate the disparity at a pixel (x, y) by using

$$\hat{\rho} = \operatorname*{argmax}_{\rho \in \{\rho_1, \ldots, \rho_D\}} \|C^{\rho}_{x,y}\|. \tag{8}$$

The same problem can be written in the following compact form

$$\min_{C,E} \frac{1}{2}\|Y - QC - E\|_2^2 + \mu\|E\|_1 \tag{9}$$

where the column vector Y has been obtained by stacking all the $Y_{x,y}$, and Q is a block diagonal matrix whose blocks are the matrices $Q_{x,y}$. To encourage the concentration of nonzero entries in a single disparity block of $C_{x,y}$ we propose to minimize the mixed ℓ_1/ℓ_2 norm of $C_{x,y}$, which is defined as $\|C\|_{1,2} \doteq \sum_{x,y}\sum_{k=1,\dots,D}\|C_{x,y}^{\rho_k}\|_2$. Finally, since the disparity is a smooth map, we add a vector-valued isotropic total variation (TV) regularization term

$$\|\nabla C\|_{1,2} \doteq \sum_{x,y}\sqrt{\|C_{x,y} - C_{x+1,y}\|_2^2 + \|C_{x,y} - C_{x,y+1}\|_2^2} \tag{10}$$

where ∇ denotes the finite gradient in the spatial domain (and can be written in matrix form). By minimizing this term we encourage C coefficients to be similar across the spatial domain. The complete minimization problem can be written as follows

$$\min_{C,E} \frac{1}{2}\|Y - QC - E\|_2^2 + \mu\|E\|_1 + \lambda\|\nabla C\|_{1,2} + \gamma\|C\|_{1,2} \tag{11}$$

where $\lambda, \gamma > 0$ are two constants. This is a convex problem and therefore it has the desirable property of converging to the same global optimum given any initialization. The minimization of problem (11) presents several challenges due to its high dimensionality, which we address in the next section.

6 Primal-dual formulation

One immediate issue of a primal solver for problem (11) is that it requires inverting very large matrices that are not easily diagonalized. To avoid such computational difficulties, we consider the primal-dual method, which is a first order algorithm, it does not require matrix inversions and enjoys fast convergence rates [25].

Firstly, we rewrite problem (11) in a more compact way by combining all the unknowns C and E into a single variable X, and by defining 3 new functions F_1, F_2, and F_3 as follows

$$F_1(AX - Y) \doteq \frac{1}{2}\|Y - QC - E\|_2^2 \tag{12}$$

$$F_2(\Pi_E X) \doteq \|E\|_1 \tag{13}$$

$$F_3(BX) \doteq \|\nabla C\|_{1,2} + \frac{\gamma}{\lambda}\|C\|_{1,2} \tag{14}$$

where $A \doteq [Q\ I_d]$, with I_d the identity matrix, $\Pi_E X \doteq E$ and $B \doteq [\nabla^T\ \frac{\gamma}{\lambda}I_d]^T\Pi_C$, with $\Pi_C X \doteq C$. Notice that all the above functions are convex in the variable X. Then, our primal formulation becomes

$$\min_X F_1(AX - Y) + \mu F_2(\Pi_E X) + \lambda F_3(BX). \tag{15}$$

To solve the primal problem we can compute the gradients of the cost function and set it to zero. An immediate observation is that the gradient will yield in the best case linear systems with non-diagonal matrices. For example, the first term $F_1(AX - Y)$ yields

$$\frac{\partial}{\partial X} F_1(AX - Y) = A^T AX - A^T Y \tag{16}$$

which requires dealing with the matrix $A^T A$. To avoid that, we use the primal-dual method. This method is based on the Legendre-Fenchel (LF) transform. Given a function F, the LF transform yields a conjugate function F^* such that

$$F^*(Z) \doteq \sup_X \langle X, Z \rangle - F(X). \tag{17}$$

The conjugate function F^* is by construction convex and when F is also convex, then the LF transform F^{**} of the conjugate F^* is again F. When the conjugate functions F_1^*, F_2^*, and F_3^* can be computed easily and possibly in closed-form, then it is convenient to consider the primal-dual problem

$$\begin{aligned} \min_X \max_{Z_1, Z_2, Z_3} & < AX - Y, Z_1 > - F_1^*(Z_1) + \mu < \Pi_E X, Z_2 > - \mu F_2^*(Z_2) \\ & + \lambda < BX, Z_3 > - \lambda F_3^*(Z_3). \end{aligned} \tag{18}$$

which we write in more compact form as

$$\min_X \max_Z < KX, Z > - \hat{F}(Z) \tag{19}$$

where $K \doteq [A^T \ \mu \Pi_E^T \ \lambda B^T]^T$, $Z \doteq [Z_1^T \ Z_2^T \ Z_3^T]^T$, and $\hat{F}(Z) \doteq F_1^*(Z_1) + \mu F_2^*(Z_2) + \lambda F_3^*(Z_3)$. To solve the above saddle point problem, we need to define the *proximity operator*, which is our fundamental computational tool to deal with the conjugate functions.

6.1 Proximity operator

A proximity operator $\text{prox}_{\sigma F}$, with $\sigma > 0$, takes as input a convex and lower semicontinuous function F and maps it to the following function

$$\text{prox}_{\sigma F}(Z) = \underset{X}{\text{argmin}} \ \frac{1}{2} \|Z - X\|_2^2 + \sigma F(X), \quad \forall Z, \tag{20}$$

see for more information the review paper [7]. The main result that we will exploit here is Moreau's identity. Given the conjugate F^* of F we have that

$$\text{prox}_{\sigma F^*}(Z) = Z - \sigma \text{prox}_{F/\sigma}(Z/\sigma) \tag{21}$$

and hence we can compute the proximity operator of the conjugate function F^* directly by using the proximity operator of the function F.

6.2 Primal-dual algorithm

The primal-dual algorithm to solve problem (19) is

$$\boxed{\begin{aligned} Z_1^{n+1} &= \text{prox}_{\sigma F_1^*}(Z_1^n + \sigma(A\bar{X}^n - Y)) \\ Z_2^{n+1} &= \text{prox}_{\sigma\mu F_2^*}(Z_2^n + \sigma\mu \Pi_E \bar{X}^n) \\ Z_3^{n+1} &= \text{prox}_{\sigma\lambda F_3^*}(Z^n + \sigma\lambda B\bar{X}^n) \\ X^{n+1} &= X^n - \tau K^T Z^{n+1} \\ \bar{X}^{n+1} &= X^{n+1} + \theta(X^{n+1} - X^n) \end{aligned}} \tag{22}$$

where n is the iteration index, $\theta \in (0, 1]$ and $\tau\sigma\|K\|^2 < 1$. While the bottom two iterations are straightforward, the first one on the dual variable Z requires computing the proximity operator of the conjugate functions F_1^*, F_2^*, and F_3^*. The first two functions are relatively easy to obtain as the conjugate functions can be computed in closed-form

$$F_1^*(Z_1) = \frac{1}{2}\|Z_1\|_2^2, \qquad \{F_2^*(Z_2)\}_s = \begin{cases} 0 & \text{if } |\{Z_2\}_s)| \leq \mu \\ +\infty & \text{otherwise} \end{cases} \tag{23}$$

where $s = 1, \ldots, \tilde{M}\tilde{N}W^2$. Hence, we can readily obtain the first two steps of the primal-dual algorithm

$$Z_1^{n+1} = \frac{1}{\sigma + 1}(Z_1^n + \sigma(A\bar{X}^n - Y)), \qquad \{Z_2^{n+1}\}_s = \mathcal{H}_{\sigma\mu}\left(\left\{\frac{Z_2^n}{\sigma\mu} + \Pi_E \bar{X}^n\right\}_s\right) \tag{24}$$

where $s = 1, \ldots, \tilde{M}\tilde{N}W^2$ and $\mathcal{H}_{\sigma\mu}$ denotes the element-wise thresholding operator

$$\mathcal{H}_{\sigma\mu}(z) \doteq \min\{\sigma\mu, |z|\}\,\text{sign}(z). \tag{25}$$

The last term F_3^* is more involved. We compute the update equation by exploiting Moreau's identity

$$\text{prox}_{\sigma\lambda F_3^*}(Z_3^n + \sigma\lambda B\bar{X}^n) = Z_3^n + \sigma\lambda B\bar{X}^n - \sigma\lambda \text{prox}_{F_3/(\sigma\lambda)}(Z_3^n/(\sigma\lambda) + B\bar{X}^n) \tag{26}$$

so that we only need to compute $\text{prox}_{F_3/(\sigma\lambda)}$. Notice that $F_3(Z_3)$ is the ℓ_1/ℓ_2 norm $\|Z_3\|_{1,2}$. Thus, we need to evaluate

$$\text{prox}_{F_3/(\sigma\lambda)}(Z_3^n/(\sigma\lambda) + B\bar{X}^n) = \underset{Z}{\text{argmin}}\ \frac{1}{2}\left\|\frac{1}{\sigma\lambda}Z_3^n + B\bar{X}^n - Z\right\|_2^2 + \frac{1}{\sigma\lambda}\|Z\|_{1,2}. \tag{27}$$

The solution is computed in closed-form and results in a block soft-thresholding

$$\text{prox}_{F_3/(\sigma\lambda)}(Z_3^n/(\sigma\lambda) + B\bar{X}^n) = \mathcal{S}_{1/(\sigma\lambda)}\left(\frac{1}{\sigma\lambda}Z_3^n + B\bar{X}^n\right) \tag{28}$$

with

$$\left\{\mathcal{S}_{1/(\sigma\lambda)}(Z_3)\right\}_b = \{Z_3\}_b \max\left\{0, 1 - \frac{1}{\sigma\lambda\|\{Z_3\}_b\|_2}\right\} \tag{29}$$

and where blocks are indexed by $b = 1, \ldots, (3\tilde{M}\tilde{N} - \tilde{M} - \tilde{N})D$, since Z_3 is a $(3\tilde{M}\tilde{N} - \tilde{M} - \tilde{N})D(nm-1)$ dimensional vector.[3] Finally, by plugging the last expression in the proximity operator of F_3^*, the last update equation becomes

$$\begin{aligned}\{\mathrm{prox}_{\sigma\lambda F_3^*}(Z_3^n + \sigma\lambda B\bar{X}^n)\}_b =& \{Z_3^n + \sigma\lambda B\bar{X}^n\}_b \\ &\cdot \left(1 - \max\left\{0, 1 - \frac{1}{\|\{Z_3^n + \sigma\lambda B\bar{X}^n\}_b\|_2}\right\}\right)\end{aligned} \tag{30}$$

where $b = 1, \ldots, (3\tilde{M}\tilde{N} - \tilde{M} - \tilde{N})D$.
In all update equations there are no matrix inversions and calculations are therefore highly parallelizable. The final algorithm is summarized in Table 1.

Table 1. Primal-dual algorithm for disparity estimation from light field data. Notice that $Z_1 \in \mathbb{R}^{\tilde{M}\tilde{N}W^2 \times 1}$, $Z_2 \in \mathbb{R}^{\tilde{M}\tilde{N}W^2 \times 1}$, and $Z_3 \in \mathbb{R}^{(3\tilde{M}\tilde{N} - \tilde{M} - \tilde{N})D(nm-1)\times 1}$.

$$\begin{aligned}
Z_1^{n+1} &= (Z_1^n + \sigma(A\bar{X}^n - Y))/(\sigma + 1) \\
\{Z_2^{n+1}\}_s &= \mathcal{H}_{\sigma\mu}(\{Z_2^n/(\mu\sigma) + \Pi_E\bar{X}^n\}_s) \\
\{Z_3^{n+1}\}_b &= \{Z_3^n + \sigma\lambda B\bar{X}^n\}_b \left(1 - \max\left\{0, 1 - \frac{1}{\|\{Z_3^n + \sigma\lambda B\bar{X}^n\}_b\|_2}\right\}\right) \\
X^{n+1} &= X^n - \tau K^T Z^{n+1} \\
\bar{X}^{n+1} &= X^{n+1} + \theta(X^{n+1} - X^n) \\
s &= 1, \ldots, \tilde{M}\tilde{N}W^2 \\
b &= 1, \ldots, (3\tilde{M}\tilde{N} - \tilde{M} - \tilde{N})D
\end{aligned}$$

6.3 Implementation details

Because of the discretization, the dimensionality of the problem is quite high. One approach to managing such dimensionality is to use block coordinate descent [29], where one works iteratively on different subsets of the variables. In this paper we consider a simple and efficient approximation: we consider restricting the possible disparities $\rho_1, \ldots, \rho_D$ to a small but carefully selected subset and

[3] The total variation term introduces 2 blocks for any pixel in Ω except for the left hand side column and the bottom row of pixels (total blocks is $2(\tilde{M}-1)(\tilde{N}-1)$). These two rows of pixels, except for the bottom right corner, introduce only one block (total blocks $(\tilde{M}-1)+(\tilde{N}-1)$). Finally, the block sparsity term introduces $\tilde{M}\tilde{N}D$ blocks.

always work with that subset. To gain additional freedom, at each pixel (x, y) we make a different choice of such subset. Our strategy is to evaluate the function

$$g_{x,y}(\rho) = \sum_{i,j} \sum_{p>i, q>j} \Phi(L_{i,j}(x,y) - L_{p,q}(x - \rho\Delta_x(p-i), y - \rho\Delta_y(q-j))) \quad (31)$$

for as many ρ values as possible. Then, we sort $g_{x,y}$ in ascending order and take the disparities corresponding to the first 5 values of $g_{x,y}$. We then also add 5 more disparity candidates by selecting the disparities of neighboring pixels (in a 4-neighborhood structure) corresponding to the smallest cost. The purpose of this second group of disparity candidates is to allow (spatially) smooth disparity estimates.

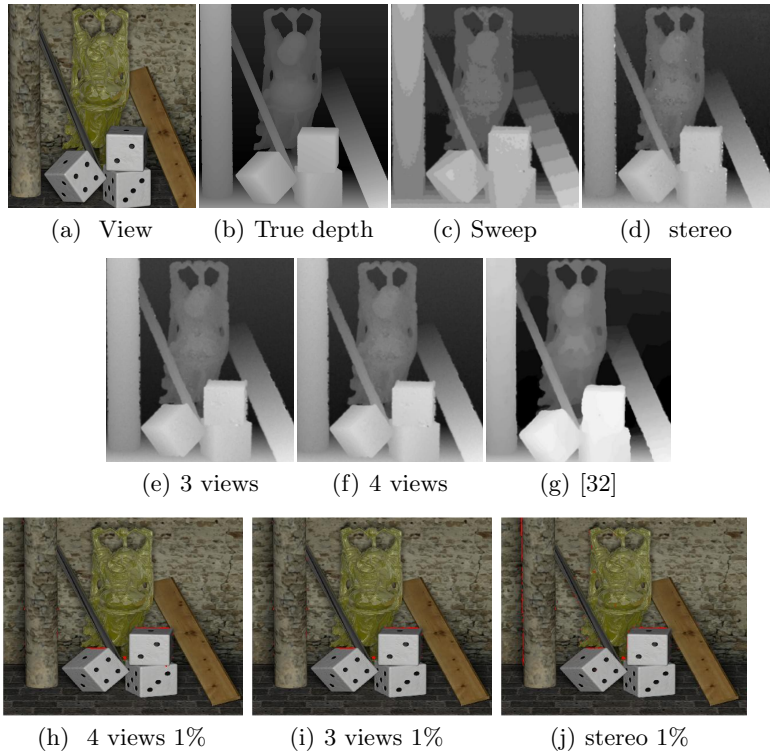
7 Experimental results

We study the performance and robustness of our light field disparity estimation framework on different datasets, Buddha [31], Watch [1], Amethyst and Truck from the Stanford light field archive.[4] We compare our results with two light field depth estimation schemes [19,32], and convex formulations [26]. Our parameters are: $\mu = 0.6$ and $\gamma = 1$ for all datasets, and $\lambda = 0.1$ for Amethyst and Truck. We work with 5×5 pixels patches ($W = 5$). Our algorithm is also demonstrated in the limit case where there are only two views (stereo). The group sparsity constraint can still work quite successfully. Another important factor is the input image size. We find that the method works better with high resolution images. However, it can also perform reasonably well on low-resolution data. In contrast, [19, 32] are challenged with few views and/or low-resolution images. The runtime of our algorithm is higher than [19]. If parallelism is fully exploited the ideal running time is about 1-3 minutes depending on the resolution and number of views. In our experiments we search through 200 disparity candidates to determine the 10 candidates. Figure 1 compares our scheme with simple plane sweep disparity search (independently at each pixel). We observe that our scheme imposes the global smoothness on the estimated disparity while the plane sweep fails to provide a smooth disparity map. As expected, the number of views used in the disparity estimation problem improves the depth estimate considerably. In our approach an increase in the number of views results in more samples per disparity candidate in the Q matrix, and therefore a better chance of fitting data more reliably. This is clearly noticeable in Fig. 1 and Fig. 2. We compare qualitatively our disparity estimation algorithm with the techniques introduced in [14, 32] in Table 2. It is clear that our scheme provides a better reconstruction quality. In Fig. 4 we illustrate how the patch size W has an immediate effect on the recovered depth map. As is well known, the larger the patch, the less noisy the depth estimate is. However, increases in patch size also affect the performance of the algorithm in the recovery of small details. More comparisons are included in [15].

[4] See `http://lightfield.stanford.edu`.

Table 2. Qualitative results for Buddha shown in Fig. 1. The table shows the percentage of pixels with relative depth error of more than 0.2%, 0.5% and 1%.

4 views			3 views			stereo			[14]			[32]	
1%	0.5%	0.2%	1%	0.5%	0.2%	1%	0.5%	0.2%	1%	0.5%	0.2%	1%	0.2%
0.13	0.33	1.9	0.139	0.33	1.99	0.42	0.85	3.26	1.15	2.44	15.05	2.9	60.4



(a) View (b) True depth (c) Sweep (d) stereo

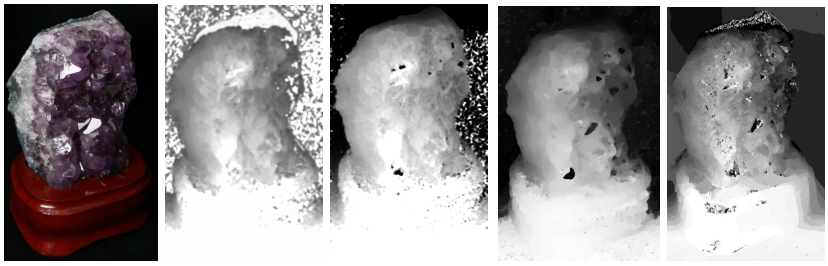
(e) 3 views (f) 4 views (g) [32]

(h) 4 views 1% (i) 3 views 1% (j) stereo 1%

Fig. 1. Buddha dataset: Comparison of the depth maps obtained from our method with the ground truth. From left to right, top row shows: the center view, the ground truth, the depth map obtained by plane sweep depth search (independently at each pixel). Middle row: the estimated depth map using different number of views, and the depth map obtained from [32]. Bottom: the estimated disparity in areas with error more than 1% are highlighted in red. We observe that an increase in the number of views improves the reconstruction quality and our scheme provides sharpe edges while the depth map estimated using [32] blurs the edges and has staircasing artifacts.

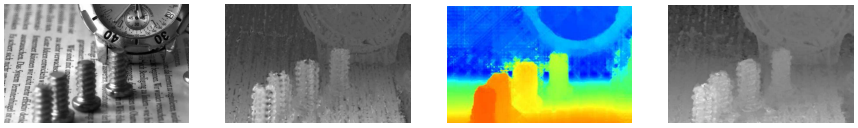
8 Conclusions

We have presented a novel convex formulation to estimate depth from light field data. The method is based on a careful discretization of disparity values and exploits a linear patch-based formulation to represent patches in one view with



(a) Input (b) Initial (c) 4 views (d) [32] (e) [25]

Fig. 2. Amethyst dataset. (a) One of the input images. (b) Initial depth estimate (plane sweep depth search) (c) Estimated disparitity using our scheme. (d-e) Estimated depth map using [32] and [25]. Notice how we obtain a reasonable estimate of the top part of the stone, while competing methods either fail or obtain a noisier estimate.



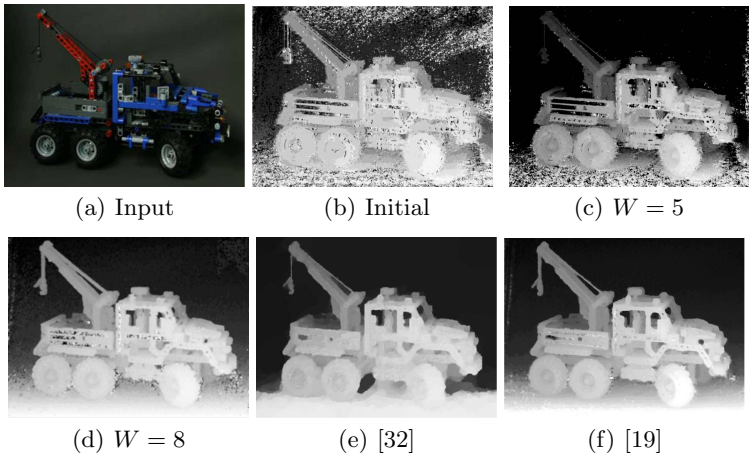
(a) Center view (b) 4 views (c) Manufacturer (d) [32]

Fig. 3. Depth estimation with the Raytrix plenoptic camera (handheld light field camera). We compare our algorithm with the reference depth provided by the manufacturer and [32]. Our scheme on a handheld light field camera yields a more detailed depth map.

patches in other views. The proposed model can easily be extended to handle simple departures from the ideal Lambertian model. For example, the current model can already handle contrast changes due to illumination (these changes would be reflected in the magnitude of the coefficients of C). The problem of depth estimation is cast as a minimization problem subject to group sparsity constraints and spatial smoothing. To gain computational efficiency we use the primal-dual method. This results in an algorithm where each dual variable update can be computed easily, independently and efficiently. Our experiments show that this method competes well with the state of the art.

References

1. Raytrix. `http://www.raytrix.de/`
2. Ayvaci, A., Raptis, M., Soatto, S.: Sparse occlusion detection with optical flow. IJCV (2012)
3. Basha, T., Avidan, S., Hornung, A., Matusik, W.: Structure and motion from scene registration. In: CVPR. IEEE (2012)
4. Bishop, T., Favaro, P.: The light field camera: extended depth of field, aliasing and superresolution. PAMI (2012)



(a) Input (b) Initial (c) $W = 5$

(d) $W = 8$ (e) [32] (f) [19]

Fig. 4. Truck dataset. We assess the influence of patch size in our scheme. Increasing the patch size results in a less noisy, but also smoother, depth map. In comparison to [19, 32], our algorithm provides sharper edges with a noisier background. This is due to two main reasons: 1) The initial 10 disparity candidates selected among 200 candidates do not contain the true disparity value, which can be improved by working on 200 candidates using block coordinate descent [29]. 2) The selection of the highest coefficients in C may lead to noisy disparity which can be addressed by imposing smoothness in the final estimation of the disparity from the coefficients of C.

5. Bolles, R.C., Baker, H.H., Marimont, D.H.: Epipolar-plane image analysis: An approach to determining structure from motion. IJCV (1987)
6. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI (2001)
7. Combettes, P.L., Pesquet, J.C.: Proximal Splitting Methods in Signal Processing. In: Fixed-Point Alg. for Inv. Prob. in Science and Eng. (2011)
8. Donatsch, D., Bigdeli, S.A., Robert, P., Zwicker, M.: Hand-held 3d light field photography and applications. The Visual Computer (2014)
9. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: CVPR. IEEE (2009)
10. Faugeras, O., Keriven, R.: Variational principles, surface evolution, PDE's, level set methods and the stereo problem. IEEE (2002)
11. Fitzgibbon, A.W., Wexler, Y., Zisserman, A., et al.: Image-based rendering using image-based priors. In: ICCV. vol. 3, pp. 1176–1183 (2003)
12. Goldluecke, B., Cremers, D.: An approach to vectorial total variation based on geometric measure theory. In: CVPR (2010)
13. Goldluecke, B., Magnor, M.A.: Joint 3d-reconstruction and background separation in multiple views using graph cuts. In: CVPR. IEEE (2003)
14. Heber, S., Ranftl, R., Pock, T.: Variational shape from light field. In: EMMCVPR. Springer (2013)
15. Hosseini Kamal, M., Favaro, P., Vandergheynst, P.: A Convex Solution to Disparity Estimation from Light Fields via the Primal-Dual Method. oai:infoscience.epfl.ch:202076 (2014)
16. Humayun, A., Mac Aodha, O., Brostow, G.J.: Learning to find occlusion regions. In: CVPR. IEEE (2011)

17. Kang, S.B., Szeliski, R.: Extracting view-dependent depth maps from a collection of images. IJCV 58(2), 139–163 (2004)
18. Kang, S.B., Szeliski, R., Chai, J.: Handling occlusions in dense multi-view stereo. In: CVPR. IEEE (2001)
19. Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., Gross, M.: Scene reconstruction from high spatio-angular resolution light fields. SIGGRAPH (2013)
20. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. IJCV (2000)
21. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: Sift flow: Dense correspondence across different scenes. In: ECCV. Springer (2008)
22. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: ICML (2010)
23. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. CSTR (2005)
24. Pock, T., Chambolle, A.: Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In: ICCV. pp. 1762–1769. IEEE (2011)
25. Pock, T., Cremers, D., Bischof, H., Chambolle, A.: Global solutions of variational models with convex regularization. SIAM J. on Imag. Sciences (2010)
26. Pock, T., Schoenemann, T., Graber, G., Bischof, H., Cremers, D.: A convex formulation of continuous multi-label problems. In: ECCV. Springer (2008)
27. Sun, X., Mei, X., Zhou, M., Wang, H., et al.: Stereo matching with reliable disparity propagation. In: 3DIMPVT. IEEE (2011)
28. Szeliski, R., Scharstein, D.: Symmetric sub-pixel stereo matching. In: ECCV. Springer (2002)
29. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. (2001)
30. Vaish, V., Wilburn, B., Joshi, N., Levoy, M.: Using plane+ parallax for calibrating dense camera arrays. In: CVPR. IEEE (2004)
31. Wanner, S., Meister, S., Goldluecke, B.: Datasets and benchmarks for densely sampled 4d light fields. In: Vision, Modelling and Visualization (VMV) (2013)
32. Wanner, S., Goldluecke, B.: Globally consistent depth labeling of 4d light fields. In: CVPR. IEEE (2012)
33. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. PAMI 31(2), 210–227 (2009)
34. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. of the Royal StatistSociety: Series B (Stat. Meth.) (2006)
35. Zach, C., Pock, T., Bischof, H.: A globally optimal algorithm for robust TV $-\ell_1$ range image integration. In: ICCV. pp. 1–8. IEEE (2007)
36. Ziegler, R., Bucheli, S., Ahrenberg, L., Magnor, M., Gross, M.: A bidirectional light field-hologram transform. In: Computer Graphics Forum. vol. 26, pp. 435–446. Wiley Online Library (2007)