What Sparse Light Field Coding Reveals about Scene Structure

Ole Johannsen, Antonin Sulc and Bastian Goldluecke
University of Konstanz

Abstract

In this paper, we propose a novel method for depth estimation in light fields which employs a specifically designed sparse decomposition to leverage the depth-orientation relationship on its epipolar plane images. The proposed method learns the structure of the central view and uses this information to construct a light field dictionary for which groups of atoms correspond to unique disparities. This dictionary is then used to code a sparse representation of the light field. Analyzing the coefficients of this representation with respect to the disparities of their corresponding atoms yields an accurate and robust estimate of depth. In addition, if the light field has multiple depth layers, such as for reflective or transparent surfaces, statistical analysis of the coefficients can be employed to infer the respective depth of the superimposed layers.

1. Introduction

In the scope of this work, light fields are dense collections of views of a scene with view point shifting parallel to the image plane. Image plane coordinates are pairs (x, y), view point coordinates are pairs (s, t), so a light field is four-dimensional. Let us look at the structure of such a light field and its epipolar plane images (EPIs), which are slices in the (x, s) and (y, t) planes, see figure 1. One can immediately observe that the data exhibits a large amount of redundancy, as patches in any of the views reappear slightly shifted in multiple neighbouring views. For a light field of a Lambertian scene, the amount of shift from one view to the next depends linearly on the disparity of the patch, which yields the well-known orientation-depth relationship on epipolar plane images [2].

This inherently sparse structure of the light field has been exploited in several lines of research. A natural application is light field compression [15] and compressive sensing [17], where the redundancy is used to generate an efficient light field coding scheme or reduce the amount of data one has to record to capture one. Indeed, the key idea is that if one knows the depth (and thus disparity) for all of the points in any of the 2D views, one can perfectly reconstruct the light field except for occlusions. The latter work [17] also employs patch dictionaries to learn the 4D structure and improve reconstruction.

Figure 1. *Epipolar plane images (EPIs).* The picture shows the center view of a light field parametrized by image coordinates x and y. On the bottom and right, the epipolar plane images for the white lines in the center view are shown, where s and t describe varying view point coordinates. As the camera moves, 3D scene points trace straight lines on the EPIs, whose slope is inversely proportional to the distance of the point [2]. Thus, orientation on the EPI is related to local depth.

Just like traditional 2D image patch dictionaries [7], 4D light field dictionaries can also be used for regularizing inverse problems. In particular, these have been employed for light field denoising and deconvolution, inpainting, and super-resolution [14]. In [18], they solve similar problems by modeling light field patches as Gaussian random variables conditioned on disparity to construct a GMM prior. Similarly to our work, they generate patches synthetically based on disparity, but not to create dictionaries. In contrast, the papers which also explicitly employ light field dictionaries [17, 14] learn structure directly on 4D light field patches. While this allows sparse coding and construction of priors, it can not be expected that the orientation-depth relationship on the EPIs is preserved in the atoms.

However, the correspondence between depth and orientation is exactly what has been leveraged in a lot of work on lightfield-based depth estimation, e.g. [13, 24, 26, 30]. We

This work was supported by the ERC Starting Grant "Light Field Imaging and Analysis" (LIA 336978, FP7-2014).

will discuss this line of research in depth in the next section, as it is most closely related to our work.

Contributions. In this work, for the first time, we unify the idea of orientation-based depth reconstruction with sparse light field coding based on generating a depth-based dictionary. In contrast to previous work, we first learn a lower-dimensional dictionary on the center view only. Then, the base atoms of the center view dictionary are "lifted" into the 4D light field domain based on a generative model such that the resulting light field atoms have a unique, known disparity. It turns out that with a simple averaging strategy, the sparse coding coefficients for the lifted dictionary already allow to compute a reliable per-pixel estimate of disparity for Lambertian surfaces. However, in contrast to orientation analysis using the structure tensor of EPIs [30], disparities can be much larger than one pixel.

Moreover, statistical analysis of the coefficients reveals whether the light field contains multiple depth layers caused by transparent or reflective surfaces. The disparity of these layers can be reconstructed, substantially surpassing existing state-of-the-art [29, 12] for multi-orientation estimation in accuracy and robustness, in particular on real-world light fields from plenoptic cameras.

2. Related Work

There has been a substantial amount of recent work and great progress in depth estimation for Lambertian surfaces in light fields. The pioneering work which introduced epipolar volumes is [2], where they analyze slopes of lines by line fitting to estimate disparity. Based on these ideas, [5] perform shearing of the epipolar volume to subsequently extract the lines with the smallest color variations. In [13], they refine the idea of line extraction and obtain very accurate results on extremely large scale light fields. The first order structure tensor is used in [30] to compute orientation on the EPIs and exploit the orientation-depth relationship. They also propose a variational framework to optimize results with respect to occlusion constraints. In [24], they calculate depth by shearing EPIs and measuring defocus responses. The work [11] employs the phase-shift theorem to match sub-aperture images in order to deal with the narrow baseline of light field cameras. The idea of a scale-depth space is pursued in [26] to find the best disparity. A data term based on active wavefront sampling is considered in [10] within a variational stereo framework. None of the above methods employ sparse coding of the light field for the purpose of depth reconstruction. However, in [9] they use the idea of redundancy of sub-aperture views and used sparsity of the RPCA as new matching term. Likewise, [18] employ sparsity ideas to model light field patches as Gaussian random variables conditioned on its disparity value. They construct a patch prior and can estimate disparity by finding the nearest PCA subspace.

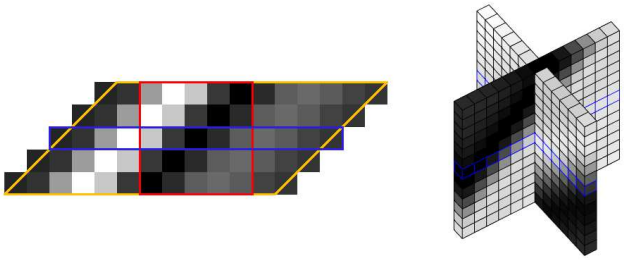

Figure 2. *Dictionary lifting. Left:* 2D epipolar plane patches are generated from line shaped atoms on the center view. The trained patches are extrapolated by slightly shifting the base patch according to the chosen disparity when moving from layer to layer (yellow). Afterwards, the final patch (red) is extracted from the region where all data is valid. *Right:* illustration of the same procedure for crosshair-shaped base patches. Please note that the 3D visualization is insufficient to show all aspects of what happens, as the underlying data is 4D. Thus, the values at the intersection of the two EPI slices do not match outside of the blue base patch area.

Much fewer work has been pursued on the topic of multi-layer light fields, which appear in the context of reflective or transparent surfaces, but also to some extent in the case of specular reflections. Again based on the idea of orientation estimation, [29] use the second order structure tensor to estimate disparity values for superimposed patterns. The work is again extended in [12] with an improvement to handle different contributions from horizontal and vertical EPIs for a slightly better accuracy. It can also separate the two layers from each other. While both methods work well on rendered and gantry datasets, it turns out in section 6 that compared to ours, they are very sensitive to noise and calibration inaccuracies. Similarly, [23, 27] optimize for two overlaid matching models for an epipolar volume using graph cuts or semi-global matching, respectively. In [25], they describe depth estimation for glossy surfaces with light field cameras. Light sources are estimated in order to separate the diffuse and specular part of signal. Again, none of these works explore the relationships between disparity estimation and sparse coding, which will be our focus in the remainder of the work.

3. Sparse light field coding

We first briefly review ideas and notation for sparse coding and dictionary learning, and afterwards specialize to our scenario of light field coding.

The central idea is to represent a signal as a linear combination of elements from a dictionary. Consider a set of n patches $x_1, \dots, x_n$ written as m-dimensional vectors $x_i \in \mathbb{R}^m$, and a patch dictionary $D \in \mathbb{R}^{m \times k}$ where k is the number of elements, usually with $k \gg n$. The problem of l_1-*sparse coding* or Lasso [20] is to find

$$\underset{\alpha_i \in \mathbb{R}^k}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| x_i - D\alpha_i \right\|_2^2 + \lambda \left\| \alpha_i \right\|_1 \right\}, \ \forall 1 \leq i \leq n, \quad (1)$$

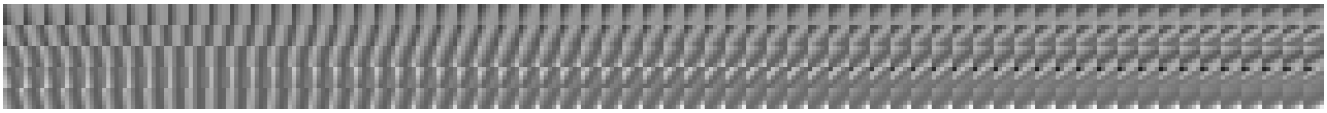

Figure 3. *Visualization of a subset of the atoms of a light field patch dictionary with lines as base atoms.* Each light field atom represents a 2D 5×5 epipolar plane image patch generated from a specific center view atom and using an individual disparity value. Every row of patches corresponds to one center view base atom, while each column of patches corresponds to a distinct disparity. A total of 320 light field atoms is visible, corresponding to 64 disparities and 5 different center view atoms.

where λ is a regularization parameter. The columns of D represent the dictionary elements to approximate the input signal and are called atoms. To ensure comparability between the different atoms and input patches both are normalised, i.e. rescaled to a zero-mean and constant variance.

A real sparse solution would be obtained for the regulariser $\|\alpha\|_0$ - i.e. the number of non-zero elements - which is not convex. Although there is no analytical link between the l_1 and l_0 norm, the l_1 norm is widely used instead and gives sparse solutions similar to the l_0 norm. The problem of dictionary learning is closely related to (1), but instead of optimizing over α given a dictionary D, the task is to find a dictionary D which is optimal in terms of representing the signal with a sparse α [16].

Creating light field dictionaries. When considering the problem of sparse coding for light fields, one needs to decide for the shape of the patches x_i and thus the dictionaries' atoms. Remember that the key idea of the paper is to encode light field dictionaries in such a way that their sparse coding coefficients α yield information for disparity estimation. For this, we leverage the fact that if disparity is constant, the center view of a light field completely determines all other views. With this in mind, we create light field dictionary atoms according to a generative model. They are generated from atoms which are learned by training a standard image dictionary on the center view. Each center view dictionary atom generates a large number of light field dictionary atoms, one for each discrete disparity value under consideration. In particular, each light field atom which is generated corresponds to a unique disparity value. In the remainder of this section, we will formalize this process and discuss several possibilities to choose the patch shape of the base dictionary.

1D base patches, 2D EPI patches. The most straightforward way to create light field atoms is to lift a center view dictionary made up of lines. We train the base dictionary on all horizontal and vertical patches of a fixed constant length p of the center view. Every 1D-atom of the base dictionary can then be lifted to atoms for an epipolar plane image as follows. Consider a shift in view point parallel to the line, and assume the points on the line have constant disparity. Since for fixed disparity, there is a linear relationship between view point and image plane coordinates of the projection, all pixels in the base patch will shift by the same amount. To simplify notation, we assume disparity units are chosen such that when shifting to the next view, the pixel shift is exactly equal to the disparity value. We repeat this shifting for every view point coordinate, and thus generate a 2D patch where one coordinate is along the line in the image plane, the other coordinate is along the line of view points parallel to it, see figure 2.

In effect, this describes a generative model which creates an EPI patch from a 1D image patch, similar to [12], but without the need to model occlusion. To obtain the final light field dictionary atom, we cut out the valid area which lies exactly above the base line. To generate the complete light field dictionary, we repeat this for every base atom and every disparity label $d = 1, \ldots, L$. Note that disparities itself need not be integer, each integer label corresponds to an actual disparity value $\lambda_d \in \mathbb{R}$. Figure 3 shows an exemplary 2D patch dictionary created for a light field. We use the same dictionary to solve the Lasso 1 for both horizontal and vertical EPI patches, and all color channels.

In the following, we refer to this light field dictionary type as the `2D` dictionary. It is light-weight, simple and efficient to compute. As a drawback, there is no inherent correlation between the horizontal and vertical EPI patches corresponding to a pixel, while of course disparity should be the same for both. Thus, we also consider two alternatives.

Crosshair base patch, 2×2D EPI patches. The first alternative is designed to enforce disparity consistency between horizontal and vertical EPIs. Here, the base patch has the shape of a crosshair of width and height p. For simplicity of implementation, the center pixel is duplicated and the horizontal and vertical lines lifted separately as above to create horizontal and vertical EPI patches. Thus, a single light field atom consists of a pair of orthogonal 2D patches in EPI space with consistent disparity, see figure 2.

The resulting Lasso problems (1) are higher-dimensional and thus computationally more expensive, however, only one has to be solved for horizontal and vertical EPI together. Thus, the different contributions do not have to be made consistent in the later optimization pass, which is conceptually more satisfying. However, as we will later see in the results, this approach is of higher accuracy for smooth areas but has more problems at occlusions, especially if the occlusion boundary is close to being vertical or horizontal. We will refer to this type of light field dictionary as `cross`.

Square base patch, 4D EPI patches. The most ambitious implementation employs complete 4D light field patches as atoms. Each base center view atom is a $p \times p$ square, which is lifted from the center view into every other

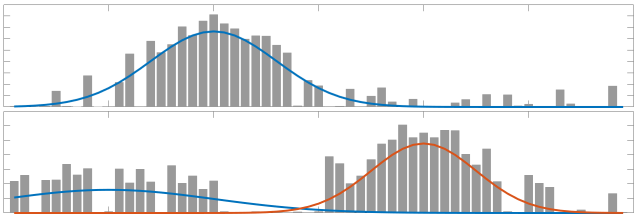

Figure 4. *Different coding coefficient distributions.* The graphs illustrate typical distributions for the sparse coding coefficients $a_d(x)$ as grey bars over the disparity range. *Top:* this distribution is likely to have only a single mode, we can see that a single normal distribution is a good fit. *Bottom:* distribution with two modes and a two-component GMM fitted to it using the EM-algorithm [21].

of the light field views to create the 4D light field atom. As we will also see later on, this type of light field dictionary, referred to as `4D`, yields a high accuracy but is computationally very expensive. It also has a tendency to smear edges because of an inherently built-in spatial smoothing. In our experience, it is often advantageous to keep the point-wise results potentially more noisy but also more precise, and leave regularization to a global optimization step which is actually designed for it.

4. Sparse coding for disparity estimation

As explained in section 3, estimating depth from epipolar plane images is equivalent to estimating the slope of the linear structures, as in the Lambertian case each line corresponds to the projection of a single 3D point. The non-Lambertian case will be discussed in the next section. Since each dictionary atom by construction corresponds to a well-defined slope, depth estimation from sparse dictionary learning comes down to analysing the result of the Lasso (1) for each pixel in the center view. To simplify notation, we first fix one pixel and drop indices related to the pixel position.

For each different disparity label d, we collect the responses of all coefficients α_i in (1) which correspond to atoms of disparity d in a vector A_d. As our dictionary is grayscale, if we consider color light fields, we need to solve several instances of (1) per pixel, one for each channel. Additional instances are required if horizontal and vertical EPIs are analyzed separately and not e.g. with the crosshair dictionary. In the cases of multiple instances, we just add all responses from all problem instances and atoms of disparity d to the vector A_d as separate elements.

From the vector A_d of responses for disparity d, we now compute a single number $a_d \in \mathbb{R}$ by taking the sum of the absolute values in A_d. Doing this for every pixel yields the final response function

$$a : \Omega \ni x \mapsto (a_1(x), \dots, a_d(x), \dots, a_L(x)) \in \mathbb{R}^L \quad (2)$$

which returns a real number $a_d(x)$ for every pixel x and disparity label d. According to our dictionary model, the number will be larger the more likely it is that the pixel has a disparity of d. We now analyze the distribution of the coefficients over d for every pixel in order to obtain the final disparity estimate.

Lambertian Case. In the case of a Lambertian surface, for every pixel the positive values of $a_d(x)$ will be clustered around the correct disparity value, see figure 4, graph on top. Thus, we fit a Gaussian to the data at each pixel by computing mean and standard deviation as

$$\mu(x) = \sum_{d=1}^{L} d\, a_d(x), \;\; \sigma^2(x) = \sum_{d=1}^{L} a_d(x)(d - \mu(x))^2, \quad (3)$$

respectively. Although the disparity values are quantized when creating the dictionary, this strategy achieves a point-wise sub-label accurate estimate, as oriented patterns of slope between e.g. d and $d+1$ will be a weighted mixture of the atoms corresponding to the discrete labels.

In cases that an estimate in a pixel is completely inaccurate, we will likely have a very high standard deviation. Also, in a textureless region, the response from all atoms will be close or equal to zero. Thus, we inpaint unreliable estimates and perform an overall smoothing by solving the L^1-inpainting problem with weighted second order total generalized variation (TGV) [3]

$$\underset{u:\Omega\to[1,L]}{\operatorname{argmin}} \left\{ \lambda\, \mathrm{TGV}_g(u) + \frac{m}{2} \|u - \mu\|_2^2 \right\} \quad (4)$$

to obtain the final disparity map u. The point-wise regularizer weight g is adapted to the edges in the center view I, and defined in a standard fashion as

$$g(x) = \exp(-K \|\nabla I(x)\|^2), \quad (5)$$

where we set $K = 5$. The inpainting mask m is zero whenever the variance $\sigma^2(x)$ is larger than one-fourth its maximum value, and one otherwise.

By design, the method sketched above assumed Lambertian surfaces, since it assumes the presence of a single orientation in every pixel. In the following section, we will generalize this to a second disparity layer to handle more difficult materials.

5. Twin peaks: two disparity layers

In [29], it is discussed how flat reflecting or transparent surfaces give rise to layered epipolar plane images, which consist of two differently oriented superimposed patterns, see figure 6 for an illustration. In this section, we show how to analyze this situation with the help of the sparse coding coefficients obtained using the generated dictionary. In contrast to previous work, we show how to compute an accurate estimate for the regions where the two-layer model

applies. Furthermore, we establish an optimization framework to obtain disparity estimates for both layers which far surpass previous work [29, 12] in quality.

Computing the mask for the two-layer model. In a first step, we need to detect the region $T \subset \Omega$ in the center view where the two-layer model actually applies. Let a_x^l be the sparse coding coefficients of pixel x for disparity labels $l = 1, \dots, L$ computed with the method described in section 4.

If pixel x belongs to a reflective or transparent surface, the surrounding regions on the epipolar plane images should exhibit two superimposed orientations. Thus, the question of whether x belongs to T is a question to the distribution of the coefficients a_x^l. If $x \in T$, then we should see two distinct peaks corresponding to the two different disparities of the two layers, otherwise we should see only a single peak. Figure 4 shows a few examples taken from our light fields.

In order to assess whether there are two layers, we perform different statistical tests. First, we perform a pixel-wise fit of a two-component Gaussian Mixture Model (GMM) using a GPU implementation of the EM-algorithm [21] which runs in parallel on all the pixels. We use a fixed number of fifty iterations, where the means are initialized with $\mu_- = 0$ and $\mu_+ = L + 1$, respectively, and both with standard deviation $L/4$. Let $\mu_1 \leq \mu_2$ be the estimated means of the mixture components.

We now construct a data term $\rho : \Omega \to \mathbb{R}$ for a global binary segmentation. First, the data term should be negative where we have a preference for the one-layer model. A good test whether a distribution has only a single mode is to check whether $\gamma^2 - \kappa \leq \frac{5}{6}$, where γ is the skewness and κ the kurtosis of the distribution [22]. Formulas to compute these are formed similar to μ and σ in (3). Second, ρ should be positive in case of a preference for the two-layer model. There is a preference for this if the initial estimates μ_1 and μ_2 are both valid and substantially different from each other.

We thus define the data term ρ to strike a balance between these two indications,

$$\rho := -1_{\{\gamma^2 - \kappa \leq \frac{5}{6}\}} + \tau\, 1_{\{1 \leq \mu_1, \mu_2 \leq L\}} (\mu_2 - \mu_1). \quad (6)$$

Above, the notation 1_S denotes the characteristic function of the set S, and $\tau > 0$ is a constant, which we set at $\tau = 1$ throughout the experiments. A special case occurs if $\sum_l a_x^l = 0$, which happens when the region around x is completely devoid of texture. In this case, a valid estimate is not possible, and we flag x to belong to a region $I \subset \Omega$ which will later be inpainted during optimization.

First, we compute the final mask T denoting the region flagged for the two-layer model by solving the binary segmentation problem with weighted length regularity

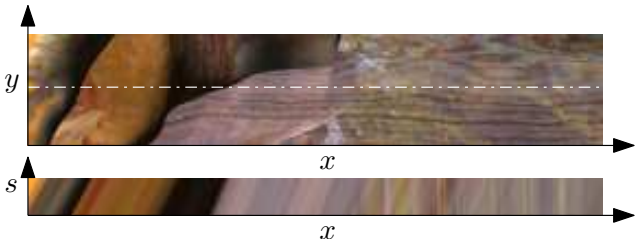

Figure 6. *Light field with two disparity layers.* The top image shows a close-up of the center view of the *tiger* dataset, see figure 5, courtesy of [29]. The EPI corresponding to the white line is depicted below. Parts of the surface are non-Lambertian and show a reflection, the corresponding regions of the EPI in turn exhibit two superimposed oriented patterns.

$$\operatorname*{argmin}_{t:\Omega \to \{0,1\}} \left\{ \int_\Omega g \left\| \nabla t(x) \right\|_2 + \rho(x) t(x) \, \mathrm{d}x \right\}. \quad (7)$$

for its characteristic function $t = 1_T$. We achieve global optimality with relaxation to functions taking values in $[0, 1]$, optimization via the primal-dual algorithm [4], and subsequent thresholding, The point-wise regularizer weight g is defined as in (5).

An example two-layer mask as a result of this optimization can be observed in figure 5. Note that such a detection was not reliably possible with any of the previous methods [29, 12].

Optimizing disparities for the two-layer model. In the region $\Omega \setminus T$, one can compute disparities with the method described in the preceding section. Within T, however, two disparity maps v, u need to be extracted from the coefficient distributions. To facilitate this, we introduce the notion of a separating disparity label $s(p)$ assigned to each pixel. We demand $v \leq s \leq u$. The key idea is that the coefficients below s should explain u, the coefficients above s should explain v. We allow fractional separation in the sense that in case of $s(p)$ lying in between the integers l and $l+1$, then the coefficient a_x^l is split, and a fraction $s(p) - l$ is assigned to the upper layer u, the rest to the lower layer v.

Let us formalize this idea a bit: for every pixel, we extend a_x^l to a function $A_x(l)$ over the continuous interval $l \in (0, L]$ by setting $A_x(l) := A_x^{\lceil l \rceil}$, where $\lceil l \rceil$ denotes rounding up, see figure 7. Then, we define the lower expectation value $\bar{v}_s(p)$ and upper expectation value $\bar{u}_s(p)$ as

$$\bar{v}_s(p) = \int_0^{s(p)} l A_x(l) \, \mathrm{d}l \text{ and } \bar{u}_s(p) = \int_{s(p)}^{L} l A_x(l) \, \mathrm{d}l, \quad (8)$$

respectively. Of course, this is just notation, in practice, the integrals can be computed by simple summation, correctly taking care of a possibly split coefficient.

In an ideal world, all observed data would be explained perfectly with the dictionary. In practice, however, we often have regions where the estimate is noisy or invalid due to

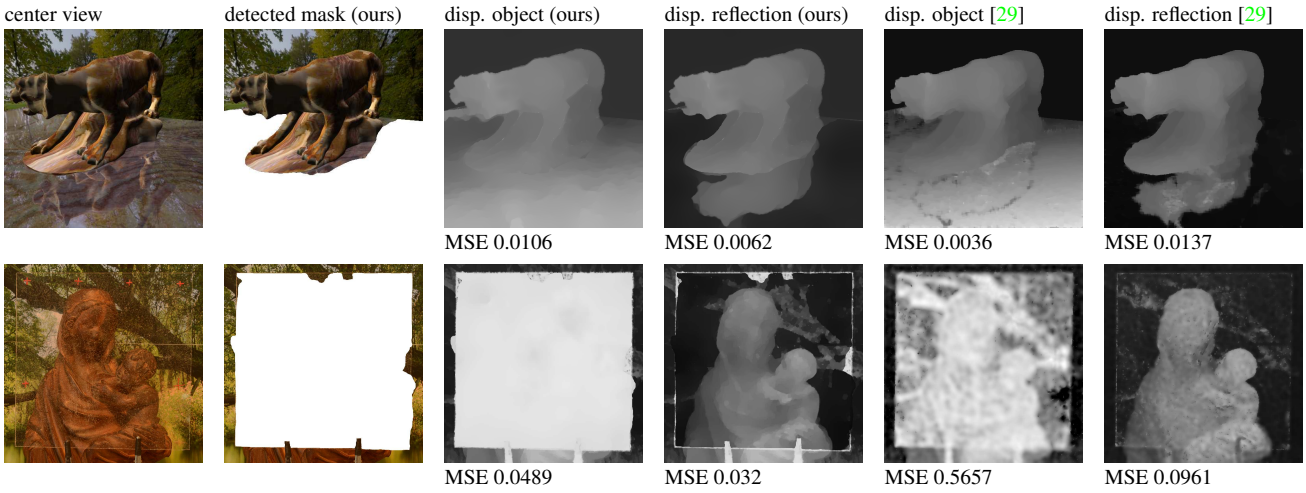

Figure 5. *Accuracy evaluation of two-layer disparity reconstructions.* We compare our method to [29] on their ray-traced *Tiger* dataset (top) with a reflective plane, and a data set captured with a gantry (bottom), for which ground truth was acquired with a laser scanner [29]. Note that in contrast to [29], our algorithm detects the reflection mask automatically (2nd column from left), while the results in [29] were obtained for the ground truth mask. For both methods, disparity maps are shown after smoothing with an L^2-data term using (4). The disparity maps are visually much better, confirmed by a better mean squared disparity error (MSE). We also experimentally verified that the error remains similar if we remove a subset of the views and thus increase maximum disparity to up to four pixels.

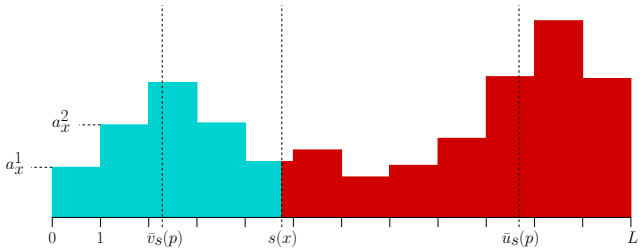

Figure 7. *Separation, upper and lower expectations.* The coordinate $s(x)$ separates the label range into an upper and lower part and thus the two disparity layers. The upper and lower estimates $\bar{u}_s(p)$ and $\bar{v}_s(p)$ are computed as expectation values over the red and blue distributions, respectively. Note that $s(x)$ is allowed to be fractional - if it lies between two disparity labels, it cuts a bar from the chart in two.

lack of texture. Thus, our algorithm consists of finding a local optimum of the functional

$$E(u,v,s) = \lambda R(u,v) + \int_T m_u(u-\bar{u}_s)^2 + m_v(v-\bar{v}_s)^2 \,\mathrm{d}x \quad (9)$$

for the three unknowns. Above, R is a regularizer for the disparity maps u and v, we use total generalized variation [3] with a point-wise weight g to account for image edges, defined as before. The functions m_u and m_v are masks depending on the coefficients and the estimated separator s. For some pixels, all coefficients above or below s might be zero, i.e. no information is available about u or v, respectively. In this case, we set the respective mask value and thus data term weight to zero, effectively performing inpainting. For all other pixels, the mask is set to one.

The optimization can only be performed up to a local minimum. We initialize u, v with the corresponding estimates μ_2 and μ_1 from the GMM, s as the mean of u, v, masks as defined above, and iterate the following steps:

1. Keep v, s fixed and minimize E for u, which is an instance of TGV-smoothing and inpainting (4).

2. Keep u, s fixed and minimize E for v, the same TGV-smoothing and inpainting (4).

3. Keep u, v fixed, ignore the masks, and compute a new estimate for s by optimizing E point-wise. Then update the masks as described above.

In our experiments, this scheme converges in about ten iterations to a steady state. Example results for the different steps can be observed in figure 5, several more final results are referenced in the following section.

6. Results and Evaluation

We conduct experiments with a Matlab implementation of our method, with a solver for the Lasso (1) from the SPAMS toolbox [16]. For a thorough evaluation, we have several data sets available of varying origin. The first type of data is ray-traced. Here, we use several data sets from the HCI light field benchmark [31], which has mostly Lambertian scenes with some mild specular reflections. On these, we only evaluate our method up to section 4, without the sophisticated scheme for two-layered light fields. As example light fields with multiple layers, we use one rendered light field with ground truth from [29], as well as one gantry light field with ground truth also from the benchmark [31]. To our knowledge, these are the only multi-layer light fields with ground truth available. They are of very high quality and present only a moderate challenge, so to test the limits

lightfield	EPI_C	EPI_G	ST_S	ST_G	Ours
buddha	**0.55**	0.62	0.78	0.90	0.57
buddha2	0.87	0.89	1.05	**0.68**	1.08
horses	2.21	2.67	1.85	**1.00**	3.26
medieval	1.10	1.24	0.91	**0.76**	0.84
monasRoom	0.82	0.93	1.05	0.79	**0.65**
papillon	2.52	2.48	2.92	3.65	**1.85**
stillLife	**2.61**	3.37	4.23	4.04	2.95
couple	**0.16**	0.19	0.24	0.30	0.42
cube	0.82	0.87	**0.51**	0.56	**0.51**
maria	**0.10**	0.11	0.11	0.11	0.11
pyramide	**0.38**	0.39	0.42	0.42	0.48
statue	0.29	0.35	**0.21**	**0.21**	0.62
average	**1.04**	1.18	1.19	1.12	1.11

Figure 8. *Comparison of different methods for disparity estimation*. Datasets are from the HCI light field benchmark [31]. The numbers show mean squared disparity error for the method described in section 4, which assumes Lambertian surfaces. Best results in each row are bold-faced. We achieve the best result on three of the data sets, and are second place on average. See text in section 6 for a description of the competing methods.

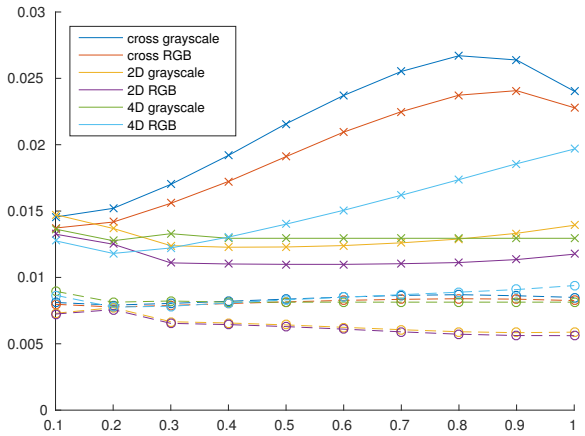

Figure 9. *MSE for different types of patches and regularization depending on the sparsity parameter* λ. Solid lines show the raw point-wise estimates, while dashed lines show results after TGV-L^2 regularisation and inpainting using 4. See text in section 3 for a description of the different light field patches used, and section 6 for a detailed discussion of the results.

of the methods, we also compare qualitatively on a data set captured with a Lytro Illum plenoptic camera [19].

We first verify on the synthetic benchmark [31] how different parameters and choices of dictionary for our method influence the quality of results. Graphs of the results can be seen in figure 9, we now proceed with a detailed discussion.

Influence of the dictionary size. First, we verify the influence of the dictionary size on the quality of the results. Two different factors have to be distinguished when in comes to the size of the dictionary. On the one hand, the number of disparity levels used, and on the other hand the number of trained atoms for the center view. Somewhat remarkably, the influence of the size of the dictionary on the quality of the estimation is diminutive. For both the number of disparity levels as well as the number of base atoms, it turns out that the results do not improve as long as the values are chosen above a certain threshold. In our experience, a good tradeoff between accuracy and run-time is to discretize disparity space such that the difference between two subsequent labels equals about one-third of a pixel. Finer disparity levels do not improve the results, and using fewer labels implies that the method becomes faster. For the number of center view atoms, four times the number of pixels in each patch seems to be a good rule of thumb. Further increasing the disparity resolution and the overall dictionary size gives close to no profit in case of Lambertian scenes, and we thus do not include a figure with detailed numbers.

Influence of patch shape. In theory, the patch shapes `cross` and `4D` have certain advantages. The shape `cross` enforces a coherent estimate over both epipolar plane image directions, while `4D` in addition enforces some spatial coherence. However, this does not reflect in performance, the best results are actually attained by `2D`, which returns coefficients for 2D patches for both directions separately. A reason could be that the more complex atoms are also less flexible to adapt to occlusions. In general, except for `4D` patches at higher values of λ, coding the individual color channels separately yields better quality than using aggregated grayscale information. Note that while the point-wise results for the different patch shapes can differ greatly, the results are much closer after inpainting and smoothing with (4).

Influence of the sparsity parameter λ. For the raw point-wise estimates, the sparsity norm weight λ has a sweet spot at around $\lambda = 0.3$. Again, after inpainting and smoothing with (4), this sweet spot is not visible anymore and the results of our method are quite robust with respect to different choices of λ. Thus, as the runtime decreases with larger values of λ, we suggest a value of around $\lambda = 0.8$ for general purpose depth estimation.

Accuracy under Lambertian assumption. Table 6 shows the mean squared disparity error for evaluation on the complete HCI light field database [29]. Results for competing methods were taken from the benchmark evaluation in [29]. As the light fields compared here are mostly Lambertian, we only employ our basic method described in section 4. The method EPI_C refers to [8] which enforces consistent depth labeling at occlusions. The globally optimal labeling scheme [30], which constructs a cost volume from structure tensor orientation estimates for horizontal and vertical epipolar plane images, is denoted EPI_G. The multi view stereo methods ST_S and ST_G compute a data term based on point-wise color consistency of all views. ST_S takes the point-wise optimum and performs simple smoothing, while ST_G performs global optimization of a continuous multi-label problem, respectively. See [29] for details on the methods.

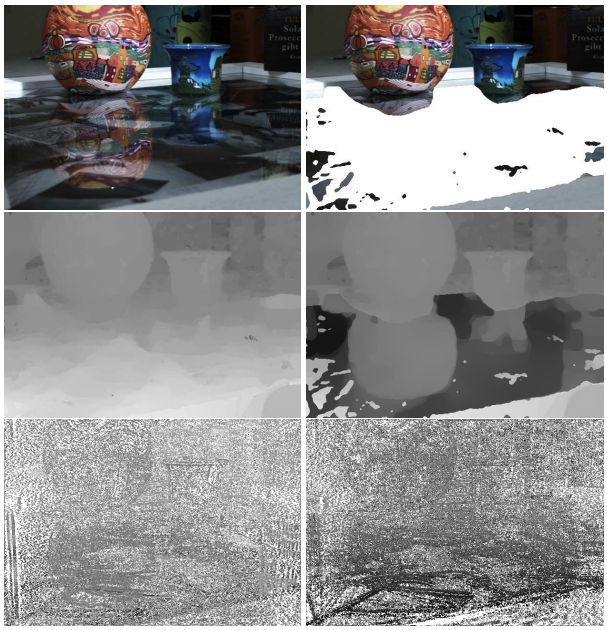

Figure 10. *Disparity estimation for two superimposed layers using Lytro data.* Our method can reliably estimate a mask for the superimposed region, as well as generate reasonably accurate disparity maps for both surface and reflection. The challenge imposed by this dataset can be appreciated by comparing to the raw estimates from the second order structure tensor [12] below, which do not turn out to be useful despite our best efforts at applying regularization (not shown). Contrast-enhanced for better visibility.

We can see that our method performs on par with the above methods, achieving second lowest MSE. Note that all of these are not occlusion-aware, which is a major drawback - indeed, more recent works [9, 28, 11] outperform these results. However, in contrast to the other methods, we are also capable of estimating two depth layers at the same time. This is what sets us truly apart from previous techniques, and where we make a big leap in quality.

Accuracy on reflective and transparent surfaces. For light fields with two layers, we compare to the method [29], which is based on decomposing the two orientations using superimposed pattern analysis [1]. We compute accuracy for both light fields where we have ground truth available, see figure 5, and note that we achieve both qualitatively as well as quantitatively far superior results. In addition, our method computes a robust segmentation into regions with and without multiple layers. Note that both light fields are of very high quality (rendered and from a gantry, respectively), so to test our limits we move to data from a plenoptic camera.

Plenoptic camera data. In order to process the data captured with the Lytro Illum, we employ the Lightfield Toolbox provided by [6] to construct an epipolar volume. This creates challenging data, quite noisy and with non-linear distortions from calibration inaccuracies, as can be observed in figure 1. We then compare our full algorithm with the second-order structure tensor approach [12], see figure 10. While the previous method [12], which is already an improvement over [29], works reasonably well on high-quality data as in figure 5, it completely breaks down on the Lytro data set and does not yield any useful estimate. In contrast, our method is still capable of estimating reflection masks as well as the disparities of the separate layers robustly and with visually convincing accuracy. Using the algorithm from [12], we can thus proceed with performing a separation of the two light field layers, see figure 11.

Figure 11. *Layer seperation of a real world light field.* Given the disparity estimates from our method for both layers and the mask for the two-layer region, see figure 10, we can seperate the two layers of the light field using the method described in [12]. Disparity maps obtained with [12, 29] are not accurate enough to perform this decomposition. Contrast-enhanced for better visibility.

7. Conclusions

In this paper, we present a novel approach for depth estimation from light fields. The key idea is to build a dictionary for sparse light field coding such that the disparity for every atom is known. For this, we first learn the structure of the center view using dictionary learning, and afterwards, lift the trained patches to the higher dimensional epipolar space using shifting proportional to disparity. The method supports different shapes of base patches, which capture different aspects of spatial coherence within the views and among epipolar plane images. Using the generated light field atoms, we then employ the Lasso (1) in order to compute sparse coding coefficients. Accumulating these with respect to the different disparities of the atoms allows to infer the depth of individual pixels of the center view, provided the light field is Lambertian.

Using statistical analysis, we are also able to detect regions where the Lambertian assumption is violated, and the light field is composed of different superimposed disparity layers. Experiments demonstrate that our method far surpasses previous work for multi-layered disparity estimation in robustness and accuracy. For purely Lambertian scenes, however, our method performs only on par with earlier methods which are not occlusion-aware. This is to be expected, as only disparity across the complete patch is considered, which decreases accuracy at object boundaries. We will remedy this in an upcoming work.

References

[1] T. Aach, C. Mota, I. Stuke, M. Muehlich, and E. Barth. Analysis of superimposed oriented patterns. *IEEE Transactions on Image Processing*, 15(12):3690–3700, 2006. 8

[2] R. Bolles, H. Baker, and D. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55, 1987. 1, 2

[3] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010. 4, 6

[4] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011. 5

[5] A. Criminisi, S. Kang, R. Swaminathan, R. Szeliski, and P. Anandan. Extracting layers and analyzing their specular properties using epipolar-plane-image analysis. *Computer vision and image understanding*, 97(1):51–85, 2005. 2

[6] D. Dansereau, O. Pizarro, and S. Williams. Decoding, calibration and rectification for lenselet-based plenoptic cameras. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 1027–1034, 2013. 8

[7] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006. 1

[8] B. Goldluecke and S. Wanner. The variational structure of disparity and regularization of 4D light fields. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2013. 7

[9] S. Heber and T. Pock. Shape from light field meets robust PCA. In *Proc. European Conference on Computer Vision*, 2014. 2, 8

[10] S. Heber, R. Ranftl, and T. Pock. Variational shape from light field. In *Int. Conf. on Energy Minimization Methods for Computer Vision and Pattern Recognition*, pages 66–79, 2013. 2

[11] H. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. Tai, and I. Kweon. Accurate depth map estimation from a lenslet light field camera. In *Proc. International Conference on Computer Vision and Pattern Recognition*, 2015. 2, 8

[12] O. Johannsen, A. Sulc, and B. Goldluecke. Variational separation of light field layers. In *Vision, Modelling and Visualization (VMV)*, 2015. 2, 3, 5, 8

[13] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 32(4), 2013. 1, 2

[14] Z. Li. *Image patch modeling in a light field*. PhD thesis, EECS Department, University of California, Berkeley, May 2014. 1

[15] M. Magnor and B. Girod. Data compression for light field rendering. *IEEE Trans. on Circuits and Systems for Video Technology*, 10(3):338–343, 2000. 1

[16] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010. 3, 6

[17] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics*, 32(4):46, 2013. 1

[18] K. Mitra and A. Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 22–28, 2012. 1, 2

[19] R. Ng. *Digital Light Field Photography*. PhD thesis, Stanford University, 2006. Note: thesis led to commercial light field camera, see also `www.lytro.com`. 7

[20] B. Olshausen and D. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997. 2

[21] S. Prince. *Computer Vision: Models Learning and Inference*. Cambridge University Press, 2012. 4, 5

[22] V. Rohatgi and G. Szekely. Sharp inequalities between skewness and kurtosis. *Statistics & Probability Letters*, 8:297–299, 1989. 5

[23] S. Sinha, J. Kopf, M. Goesele, D. Scharstein, and R. Szeliski. Image-Based Rendering for Scenes with Reflections. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 31(4):100:1–100:10, 2012. 2

[24] M. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *Proc. International Conference on Computer Vision*, 2013. 1, 2

[25] M. Tao, T.-C. Wang, J. Malik, and R. Ramamoorthi. Depth estimation for glossy surfaces with light-field cameras. In *Computer Vision-ECCV 2014 Workshops*, pages 533–547, 2014. 2

[26] I. Tosic and K. Berkner. Light field scale-depth space transform for dense depth estimation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 441–448, 2014. 1, 2

[27] Y. Tsin, S. Kang, and R. Szeliski. Stereo Matching with Linear Superposition of Layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):290–301, 2006. 2

[28] T. Wang, A. Efros, and R. Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3487–3495, 2015. 8

[29] S. Wanner and B. Goldluecke. Reconstructing reflective and transparent surfaces from epipolar plane images. In *German Conference on Pattern Recognition (Proc. GCPR)*, 2013. 2, 4, 5, 6, 7, 8

[30] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):606–619, 2014. 1, 2, 7

[31] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4D light fields. In *Vision, Modelling and Visualization (VMV)*, 2013. 6, 7