Vision, Modeling, and Visualization (2015)
D. Bommes, T. Ritschel, and T. Schultz (Eds.)

Variational Separation of Light Field Layers

Ole Johannsen, Antonin Sulc and Bastian Goldluecke

University of Konstanz

Abstract

Images of scenes which contain reflective or transparent surfaces are composed of different layers which are observed at different depths. Analyzing such a scene requires separating the image into its individual layers, which remains a challenging and important problem. While the problem is very much ill-posed when only a single image is considered, recent work has shown that depth estimation for two layers becomes quite tractable when one instead captures a 4D light field of the scene. In this paper, we propose a novel variational approach to layer separation which is based on these ideas. We formulate a linear generative model to reconstruct the light field from disparity and luminance information for the individual layers on the center view. Comparing the model with the observerd data yields a convex variational problem for layer reconstruction, which can be solved to global optimality with a primal-dual scheme. Layer disparity is estimated in a first step, for which we improve upon a model based on second order structure tensors on the epipolar plane images. In contrast to previous work, the resulting approach is robust enough to be able to deal with light fields from the Lytro Illum camera, for which we obtain a compelling separation of the reflectance layer in real-world scenes.

Categories and Subject Descriptors (according to ACM CCS): I.4.4 [Image Processing and Computer Vision]: Restoration—I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Shape

1. Introduction

Partially reflecting and transparent surfaces are omnipresent in the real world. Images of such surfaces will typically show a complex mixture of multiple layers. For example, when looking through a window, one will usually observe objects behind, as well as the reflection of objects in front of the window, resulting in two superimposed layers with different luminance. In cases of textured or very dirty glass, one might even get contributions of a third layer. Separating those layers again is a very difficult problem, but also an important step when dealing with real-world data, as many algorithms based on feature detection and correspondence search require Lambertian surfaces.

Given only a single image, separating the different layers is a highly ill-posed problem and in some cases even complicated for a human observer to solve. Therefore, most existing methods use multiple images of the same scene captured under different imaging modalities. These include focus stacks to estimate the different superimposed layers [SKB00], using a polarizer to vary the intensity of the reflection [SSK99, KTS14] or statistical approaches which maximize the probability that the estimated layers generate the input data [FA99, BBZZ03]. One notable approach is even capable of separating the two layers from a single image [LZW04] by finding a decomposition that minimises the total number of edges and corners. However, this idea reqiures that only limited amount of texture is present in the image. Gai et al. [GSZ12] learn a statistical descriptor of real world images and are capable of estimating the number of superimposed layers as well as reconstucting those layers from two images only. The prior assumption is that the different layers perform rigid motions, and the method otherwise relies on learning image statistics to be successful.

Another main class of approaches to layer separation utilizes multiview stereo images and estimates separate motion fields between the input images for the individual layers. These employ a generative model, where the layers that are to be estimated are warped and superimposed according to the inter-frame motion estimates to form the candidate observed images. In an energy minimization framework, both layer motion as well as layer images are then optimized to match the input images [SAA00, TKS06, SKG*12].

DOI: 10.2312/vmv.20151268

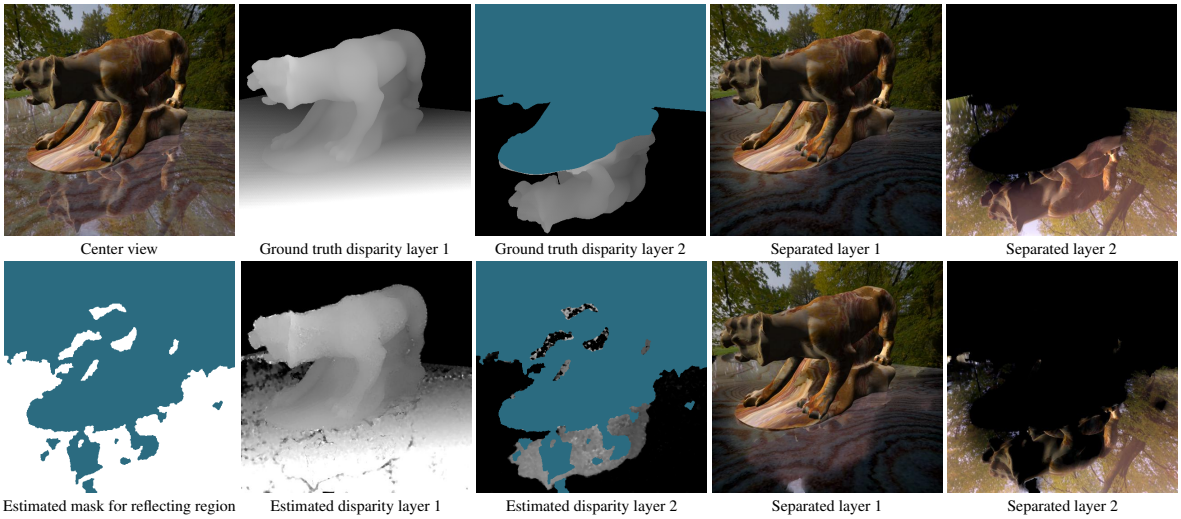

Figure 1: *Comparison of layer separation with ground truth (top row) versus estimated (bottom row) disparity. Areas which are masked out from the respective estimates as no reflection was determined are shown in blue. Layers can only be recovered up to a constant offset (see main text), which leads to intensity variations. In regions where reliable disparity estimation for both layers is possible, however, the result is quite accurate and close to the actual ground truth.*

In our paper, we utilise a related approach, which is however adapted to match the specific structure of a 4D light field in the two-plane parametrization. In particular, we will demonstrate that a single shot from a plenoptic camera is sufficient to separate the superimposed layers. While estimating layers and their individual motions looks like a chicken-and-egg problem at first glance, it turns out that in the 4D light field setting, the disparity of each individual layer in the scene can be reliably estimated using a second order structure tensor on the epipolar plane images. This approach was previously proposed in [WG13], and allows to perform layer disparity estimation as a pre-processing step to layer separation.

Contributions. While the focus of our work lies in the actual separation of the layers once individual disparity has been estimated, we also propose improvements to the multi-layer disparity estimation algorithm [WG13]. Specifically, the previous work dealt with the estimates from different slices through the 4D light field volume (epipolar plane images) in a heuristic manner, while we give a theoretical justification that they can be merged into a single tensor. Experiments demonstrate that this substantially increases robustness, in particular for real-world data.

Our main contribution is a novel variational model for layer separation given the disparity information of the individual layers. In our framework, we identifiy the pixels in each view that correspond to a certain position in the respective layers and formulate a generative model which composes the complete 4D light field from individual layers on the center view. It turns out that this leads to a deconvolution-like problem to obtain the layers. A variational energy minimization framework then balances the difference of the model to the observation with state-of-the art regularization terms. Optimization is performed with a well-known first order primal-dual scheme using optimal preconditioning [CP10,PC11]. We demonstrate the precision of our approach on multiple synthetic and Gantry data sets with ground truth available. In addition, we demonstrate in experiments with 4D light fields from a Lytro Illum plenoptic camera [Ng06] the feasibility of the approach for real-world data sets.

2. The 4D Light Field and Epipolar Plane Images

We first briefly review notation commonly used in light field analysis, and describe the problem of layer motion estimation in the context of epipolar plane images. In light field imaging, we usually resort to the two-plane parametrization [LH96] to parameterize the rays captured by a light field camera. A useful way to visualize this $4D$ representation is as a collection of pinhole cameras with focal points in a common plane Π and common image plane Ω, see figure 2. The focal plane Π is parameterized by spatial coordinates (s,t), the image plane Ω by angular coordinates (x,y). The $4D$ light field L is then a map describing the luminance of each ray (x,y,s,t) passing through both planes,

$$\begin{aligned} L : \Omega \times \Pi &\to \mathbb{R}, \\ (x,y,s,t) &\mapsto L(x,y,s,t). \end{aligned} \tag{1}$$

For imaging of light fields in the two-plane parametrization, several methods are in common use. An obvious capturing method are camera arrays, where cameras are positioned equidistantly in a grid with parallel optical axes. Such arrays

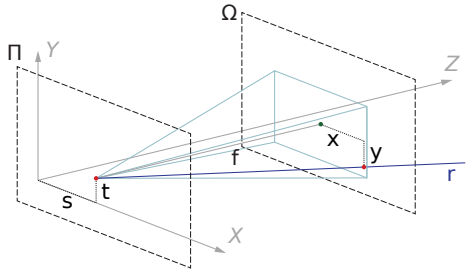

Figure 2: Light field parametrization. *An incident ray* $\boldsymbol{r}$ *is parametrized by its intersections with the* focal plane Π *and the* image plane Ω *(red dots). The planes are parallel with distance equal to the focal length* f*. The intersection coordinates* (s,t) *are given in relation to the origin of the world coordinate system. The coordinates* (x,y) *are given relative to the intersection of the optical axis of a virtual camera placed at* $(s,t,0)$ *in Z direction with the second plane (green dot). Each of these virtual cameras gives a subaperture view of the light field.*

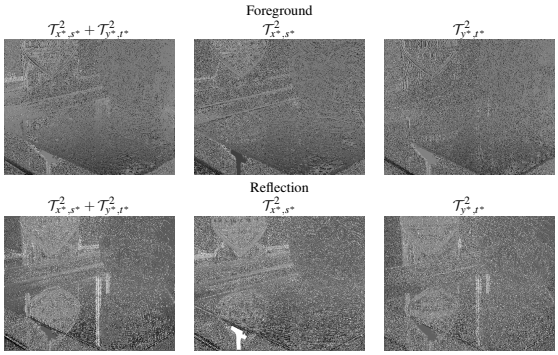

Figure 3: Raw disparity estimates from second order structure tensors. *Images show disparities using different second order structure tensors on light field data captured with the Lytro Illum plenoptic camera (central view is depicted in figure 4). The top row contains estimates for foreground, bottom row background. Results from the proposed scheme which uses a combined structure tensor are in the leftmost column, and visibly more robust than the estimates from individual EPIs (second and rightmost column).*

are now commercially available in miniature form in mobile phones and tablets for example from the company Pelican Imaging, which reduces the traditionally considerable efforts regarding hardware requirements. For static scenes, gantries can be employed, where images are captured sequentially with a camera moving in a 2D plane. Finally, commercially available plenoptic cameras have been making rapid progress recently. Well known are the hand-held comsumer camera Lytro Illum, which we employ to capture real-world light fields in this work, and the offerings by Raytrix targeted at industrial applications.

In this work, we consider the motion of the projections of 3D points into the light field for layer separation. These can best be captured by considering epipolar plane images (EPIs) [BBM87], which are $2D$ slices through the $4D$ light field. To describe such an EPI, we fix both a 1D view point coordinate (either t^* or s^*) as well as the corresponding 1D image coordinate (y^* or x^*). This leads to EPIs $f_{y^*,t^*}(x,s) = L(x,y^*,s,t^*)$ in coordinates (x,s) or EPIs $f_{x^*,s^*}(y,t) = L(x^*,y,s^*,t)$ in coordinates (y,t), respectively, which exhibit a characteristic structure consisting of overlapping lines, see figure 4.

The reason for these patterns is that the projection of a 3D world point into an epipolar plane image is a line [BBM87]. Indeed, if the camera coordinate changes linearly, this leads to a linear change of projected coordinates according to the pinhole camera projection equations. Specifically, if Z is the distance to the image plane and f the focal length, i.e. distance between image and focal plane, a $3D$ point will be projected onto a line with slope $\frac{f}{Z}$ in both horizontal as well as vertical EPIs. The slope is called the disparity of the 3D points' projection [GW13]. Thus, reconstruction of depth information is equivalent to detecting orientation of patterns in the EPI. This insight is exploited in a number of recent publications in order to infer depth [CKS*05, WG14, KZP*13].However, they rely on the assumption that along the lines, the luminance is constant, which implies a Lambertian reflection model. Thus, they completely fail for surfaces which are for example strongly reflective or transparent.

The problem we thus have to address in our scenario is to deal with ambiguous orientations. In the case of reflections or transparencies, there are superimposed patterns with different orientations which correspond to points at different depths which are visible simultaneously. These need to be separated in order to infer the respective layer disparities. This problem was investigated in [WG13] based on the second order structure tensor, which was proposed in [AMS*06] for the analysis of superimposed oriented patterns. It was shown that the framework ideally fits the proposed scenario. In the following section, we will give a brief overview of the ideas, and propose improvements to make the method more robust for the difficult real-world data from light field cameras.

3. Disparity estimation with superimposed layers

We first briefly state the main results from [WG13] to recover the two disparities in an EPI which consists of two different layers (i.e. reflecting surface plus reflected scene), see figure 4.

Two superimposed layers on a single EPI. Assume a region Ω where the EPI f is the superposition $f = f_u + f_v$ of two layers f_u and f_v with disparities λ_u and λ_v, respectively. The model is valid only for planar reflection surfaces because reflection EPIs must consist of lines. We encode the

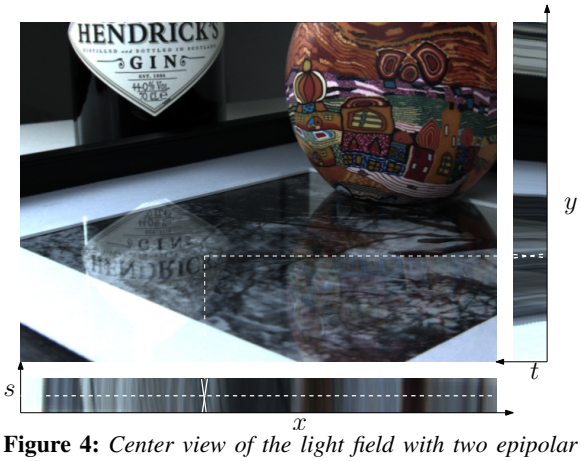

Figure 4: *Center view of the light field with two epipolar plane images extracted along the dotted lines shown in the margins. The two orientations are visualized with intersecting white lines on the EPIs.*

disparities in a mixed-orientation parameters (MOP) vector $a = (\lambda_u\lambda_v, \lambda_v + \lambda_u, 1)^T$, which can be decomposed again into the disparities after it has been estimated [WG13]. The first key observation [AMS*06] is that a satisfies

$$a^T (d_f d_f^T) a = 0 \text{ on } \Omega, \tag{2}$$

with the spatially varying vector $d = (f_{xx}, f_{xy}, f_{yy})^T$ of second order derivatives. In practice, the equation will not be satisfied exactly everywhere. To recover a, [AMS*06] thus minimize the quadratic form

$$\begin{aligned} Q(a) &= \int_\Omega a^T (d_f d_f^T) a \, \mathrm{d}x = a^T \left(\int_\Omega d_f d_f^T \, \mathrm{d}x \right) a \\ &=: a^T \mathcal{T}^2 a. \end{aligned} \tag{3}$$

The 3×3 matrix $\mathcal{T}^2$ is called the second order structure tensor. In practice, the integral is a weighted summation over a square window around the pixel under consideration, often weighted with a Gaussian to decrease the influence of derivatives further away. According to (3), the MOP vector a and thus the two disparities can be recovered as the Eigenvector to the smallest Eigenvalue of $\mathcal{T}^2$.

Merging contributions from different EPIs. For each pixel of the center view, one obtains two estimates for disparities - one from vertical EPI slices, one for the horizontal ones. Both need to merged into a single disparity map for each layer. In [WG13], a heuristic strategy was proposed which was based on comparison of the outputs of the different models, selecting disparities which agree in both EPIs. This strategy also yields a binary map detecting the regions in the image where two orientations can reliably be detected.

Unfortunately, it turns out that for real world data from the Lytro, the previous approach completely breaks down, since the data from the different channels is just too unreliable and noisy, see figure 3. We thus propose a new approach which constructs a single tensor from the contributions of the individual EPIs. This automatically merges all available information, and yields an overall much more robust result.

Let (s^*, t^*) be the focal point of the center view, and (x^*, y^*) a fixed image coordinate. From the EPI f_{x^*,s^*}, we obtain the second order structure tensor $\mathcal{T}^2_{x^*,s^*}$, from the EPI f_{y^*,t^*}, the second order structure tensor $\mathcal{T}^2_{y^*,t^*}$, respectively. The key observation is that since disparities only depend on the Z-coordinates of 3D points, the MOP vector a for both EPIs will be the same, and in the ideal case zeroes both quadratic forms $a^T \mathcal{T}^2_{x^*,s^*} a$ as well as $a^T \mathcal{T}^2_{y^*,t^*} a$. We thus propose to minimimize

$$Q'(a) = a^T \mathcal{T}^2_{x^*,s^*} a + a^T \mathcal{T}^2_{y^*,t^*} a = a(\mathcal{T}^2_{x^*,s^*} + \mathcal{T}^2_{y^*,t^*})a, \tag{4}$$

i.e. compute a as the Eigenvector to the smallest Eigenvalue of $\mathcal{T}^2_{x^*,s^*} + \mathcal{T}^2_{y^*,t^*}$. Figure 3 demonstrates that this gives more robust results compared to the contributions from [WG13].

4. Generative model for EPIs from center view data

The different superimposed layers in a scene containing e.g. reflections have different disparities. The central idea is to build a model to generate a complete epipolar plane from data in the center view only, namely the (yet unknown) layer luminances and the layer disparity values inferred using the methods in the previous secion. The multiple observations of the superimposed layers under different motions give the necessary information for layer reconstruction.

Propagation of center view information. To mathematically define a method to reconstruct EPIs from the center view data only, we first consider one individual epipolar plane image and one individual layer, for which we assume a Lambertian reflectance model. The idea is that the color at (most) points on this EPI can be derived from disparity and color information of the center view. On the EPI, this data can be found on a single line with fixed s or t coordinate, respectively, passing through the midpoint of the EPI. As can be seen in figure 5, the disparity of pixels at the center line defines the epipolar lines (dashed lines), each of which consists of projections of the same 3D point. In particular, the color of all pixels along such a line should be equal to the color at the center view in the occlusion-free case. Thus, in the most simple scenario, the color at a point on the EPI (e.g. red dot) can simply be approximated by interpolating the constant color values of the closest epipolar lines.

However, care must be taken in regions where occlusions occur (green dots). There are two different cases to be distinguished. In the first case, there are multiple epipolar lines with different slope close to the point, as in the case of the top green dot. Here, one needs to identify which of the epipolar lines is closer to the observer and thus occluding the other one. This will be the one with larger disparity (red lines). In the second case, there is no information about the point we are considering available in the center view, as it is occluded

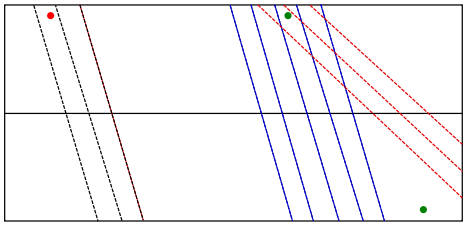

Figure 5: Construction of an EPI from the data on the center view (solid line). *The disparity at each point on the central view yields an epipolar line (dotted) on the EPI which passes through the respective point. Neglecting occlusion (red dot), the color value at any position in the EPI can be inferred by linear interpolation from the neigbouring epipolar lines. For a detailed description of how to deal with occlusions (green dots) see the text.*

by other 3D points (bottom green dot). Here, the EPI can not be reconstructed and the area needs to be masked out from further consideration.

Mathematical model. To formalize the above ideas, let us consider an EPI E of size $N \times K$. On the EPI, we define a binary mask M which will be zero for all pixels for which no information is available on the center view (second occlusion case). For all other pixels, the mask is set to one, and color can be reconstructed by finding the closest non-occluded epipolar lines to the left and to the right, and then linearly interpolating between the color of these two. Thus, a grayscale EPI E can be reconstructed by matrix multiplication $\bar{E} = Gu$. Here, $\bar{E}$ is a vector of length $N \cdot K$ obtained by stacking the columns of E on top of each other, and G a sparse matrix of size $N \cdot K \times N$. The vector $u \in \mathbb{R}^n$ contains the luminance values on the center view for this particular EPI. Thus, each row of the sparse matrix G has reconstruction information for a single pixel of the EPI. Only the two entries corresponding to the closest left and right epipolar line are non-zero, and they contain the linear interpolation weights. In the case of a color EPI, the matrix G is the same and each channel is reconstructed individually.

Implementation details. Algorithmically, the matrix G can be constructed by iterating over the N pixels on the central view and their epipolar lines in order of increasing disparity. For each epipolar line under consideration, the rows in G corresponding to pixels immediately to the left and to the right of the line are updated with the respective interpolation weights. The process can be sped up by maintaining extra buffers for the indices and interpolation weights for the closest left and right epipolar lines for each pixel. Iterating in the order of increasing disparity assures that the occlusion order of epipolar lines is respected. All rows in G for which all entries are still zero correspond to pixels which are not visible in the center view. These are masked out, i.e. their entry in M is zero. For the remaining pixels, their entry in M is one.

5. Variational layer decomposition

The previous section modelled formation of a single EPI for a single layer. Assume we have observed a (Lambertian) epipolar plane image f, and have reconstructed disparity values d of the center view, and the center line has intensity values u. The central idea for layer decomposition is the observation that by our modeling assumption, the error

$$\varepsilon(u,d,f) = \|M_d \odot [G_d u - f]\|_p^p, \tag{5}$$

for any choive of p-norm should be small. Above, the symbol $\odot$ denotes point-wise multiplication. We write M_d and G_d instead of just M and G to emphasize that both matrices depend on the disparities (and only on these). Note that while u and d are only 1D functions (they live on a line in the center view), equation (5) gives a distance of 2D EPIs.

We will now extend the model from a single epipolar plane image for a single layer to multiple layers on the complete light field. For this, first consider a single EPI f which is formed from two superimposed patterns f_u and f_v. The natural assumption for the image formation process is $f = f_u + f_v$, see e.g. [WG13]. Given the disparity at the center view for both layers, one can calculate the two matricies G_{d_u} and G_{d_v} and the respective masks M_{d_u} and M_{d_v}, where d_u, d_v denote the respective disparities. In the ideal noise-free case for perfect disparities, $f_u = G_{d_u} u$ and $f_v = G_{d_v} v$. However, this model will never be exactly satisfied in practice, so we propose to minimize the data cost

$$D_{EPI}(u,v) = \|C(u,v)]\|_p^p, \tag{6}$$

$$C(u,v) = M_{d_u} \odot M_{d_v} \odot [G_{d_u} u + G_{d_v} v - f] \tag{7}$$

for each individual EPI.

This cost only accounts for a single EPI, corresponding to an individual 1D slice though the center view whose layers are to be reconstructed. Let us now assume we have $y = 1, \ldots H$ rows and $x = 1, \ldots, W$ columns in the center view. Each one corresponds to one epipolar plane image, thus we obtain data terms D_y and D_x for each of the rows and columns, respectively. In order to estimate the decomposition into two layers for the complete center view, we extend the data term to the total cost

$$D(u,v) = \sum_{x=1}^{W} D_x(u_x, v_x) + \sum_{y=1}^{H} D_y(u_y, v_y). \tag{8}$$

where u_x, v_x denote column x and u_y, v_y row y of the respective unknown matrices.

While for ground truth depth maps close to no regularisation is required, in the case of real world data with noise in the light field as well as imperfect disparity estimation we employ a state-of-the-art regulariser. We use the second order Total Generalised Variation (TGV), which favors piecewise linear solutions instead of piecewise constant ones like standard total variation [BKP10].

Putting all together, we need to minimize the energy

$$E(u,v) = D(u,v) + \lambda(J(u) + J(v)), \tag{9}$$

where J denotes the regularisation term on u and v, respectively, and $\lambda \geq 0$ is the constant user-defined regularization weight.

In order to minimise this energy, we employ the well-known primal-dual algorithm by Chambolle and Pock [CP10]. To be able to apply the algorithm, we rewrite the energy (9) in its primal-dual form. The primal-dual for the TGV2-regularizer is well-known [BKP10]. For the primal-dual of the data term (8), we require dual variables q_x and q_y for each of the horizontal and vertical EPIs. Each q_x, q_y is a vectorial function on the EPI with as many channels as there are color channels, whose values are restricted to the unit ball. The resulting primal-dual form for the minimization of (8) is

$$\min_{u,v} \max_{\substack{\|q_x\|_2 \leq 1 \\ \|q_y\|_2 \leq 1}} \left\{ \sum_{x=1}^{W} \langle C_x(u,v), q_x \rangle + \sum_{y=1}^{H} \langle C_y(u,v), q_y \rangle \right\}. \tag{10}$$

In the same notation as for D, the residuals C_x, C_y for each EPI are defined via equation (7).

To improve the speed of convergence, we apply preconditioning [PC11]. The step sizes are restricted by the row and column sum norms of the matrices G_d, as well as the counterparts from the regularizer. For details, we refer to [PC11].

6. Results and experiments

For our experiments, we use synthetic data as well as real-world data captured with a gantry [WMG13] and a Lytro Illum light field camera, respectively. The Lytro light fields where processed with the light field suite [DPW13] to obtain subaperture images and camera calibration information. We obtain 15×15 subaperture views with resolution 434×625 pixels each. Outer views in corners are ignored due to vignetting effects.

Accuracy of disparity estimation. To validate the quality of the depth estimates, we use a synthetic light field rendered with 17×17 sub-aperture views at resolution 515×512 pixels, for which ground truth disparity is known. We compared our disparity estimates using the proposed combined $\mathcal{T}^2_{x^*,s^*} + \mathcal{T}^2_{y^*,t^*}$ structure tensor with disparity estimates from separate tensors $\mathcal{T}^2_{x^*,s^*}$ and $\mathcal{T}^2_{y^*,t^*}$ with the ground truth data, see table 1. To separate foreground from reflection, we use the measure $c = 1 - \left(\frac{\lambda-\mu}{\lambda+\mu}\right)^2$, where λ and μ are the smallest eigenvalues of second and first order structure tensors, respectively. While only a heuristic measure, it yields a good estimate for confidence in the double orientation model in practice, see figure 1. While $\mathcal{T}^2_{x^*,s^*}$ and $\mathcal{T}^2_{y^*,t^*}$ gave slightly worse disparities of foregrounds, the proposed method performs significantly better on the reflection layer in all cases.

Reflection coefficient	$\mathcal{T}^2_{x^*,s^*} + \mathcal{T}^2_{y^*,t^*}$		$\mathcal{T}^2_{x^*,s^*}$		$\mathcal{T}^2_{y^*,t^*}$	
	front	back	front	back	front	back
$\alpha = 0.1$	**0.119**	**0.182**	0.124	0.278	**0.119**	0.282
$\alpha = 0.3$	**0.116**	**0.0927**	0.122	0.189	0.123	0.183
$\alpha = 0.5$	**0.127**	**0.065**	0.133	0.148	0.145	0.155
$\alpha = 0.7$	**0.156**	**0.061**	0.159	0.142	0.186	0.146
$\alpha = 0.9$	0.235	**0.095**	**0.231**	0.195	0.266	0.219

Table 1: *MSE of point-wise disparity estimates compared to ground truth data for different reflection coefficients* α *(*$f = (1-\alpha) f_u + \alpha f_v$*). We compared results of the previous method with separate structure tensors* $\mathcal{T}^2_{x^*,s^*}$ *and* $\mathcal{T}^2_{y^*,t^*}$ *with our proposed combined structure tensor* $\mathcal{T}^2_{x^*,s^*} + \mathcal{T}^2_{y^*,t^*}$ *with same parameter setting. The new method overall achieves much more accurate results, see text.*

For the $\alpha = 0.9$ we got slightly worse results for foreground with our method in comparison to $\mathcal{T}^2_{x^*,s^*}$.

Separation of Reflection Layers. First, we demonstrate the quality of the algorithm on synthetic data. We use two images and generate two different constant depth maps to generate a single light field with superimposed layers which perfectly fits the image formation model. This can be thought of as two overlaying posters where one is semitransparent. The results are close to perfect as can be seen in figure 6. The MSE as well as the energy converges, and - as evident from the lower two images on the right half of figure 6 - most errors accure either at edges and are due to regularisation, or seem to be caused by a constant offset. This is an inherent problem which arises from an ambiguity of the dataterm - adding a constant offset will not change the energy as long as none of the superimposed layers have values closer to pure black or white than the offset value. Thus, layer separation is in general only possible up to an additive constant on both layers, which explains intensity variation visible in some of the experiments. For the synthetic light field for which ground truth disparity was available, we compare the results from layer separation with ground truth and estimated depth maps in figure 1.

In addition, we performed experiments on real world data generated with a gantry. Results can be observerd in figure 7. Due to high quality of the images as well as high precision of the camera positions the decomposition works remarkably well. As a final experiment, we captured a reflecting surface with a Lytro Illum camera, see figure 8. Although the light field is quite inaccurate due to currently poor calibration of the camera, the presented algorithms are capable of estimating the depth for both layers as well as seperating the two layers. For both real world experiments, the available data was unfortunately of insufficient accuracy to estimate a reliable segmentation in reflecting and Lambertian surfaces. This is left for future work, at the moment, those masks are manually drawn.

Regarding computational efficiency, the generation of the matrices G is computationally expensive and takes around

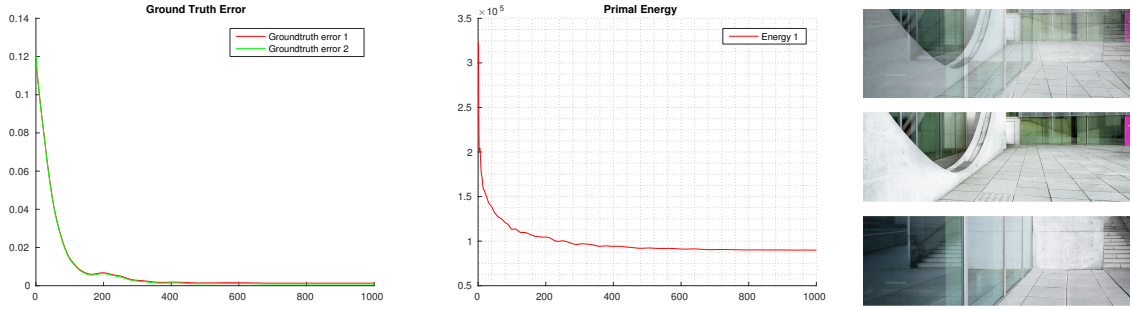

Figure 6: Decomposition of a synthetic light field, one transparent poster in front of another poster. *From left to right: convergence of MSE for estimated layers over iterations, primal energy, the center view of the input light field as well as the two resulting layers. The MSE converges and reaches a constant level after around 400 iterations, while the primal energy still decreases until it reaches a near constant level at around 800 iterations. The resulting images show, that the model is capable of separating layers with high precision, independent whether texture is present or not.*

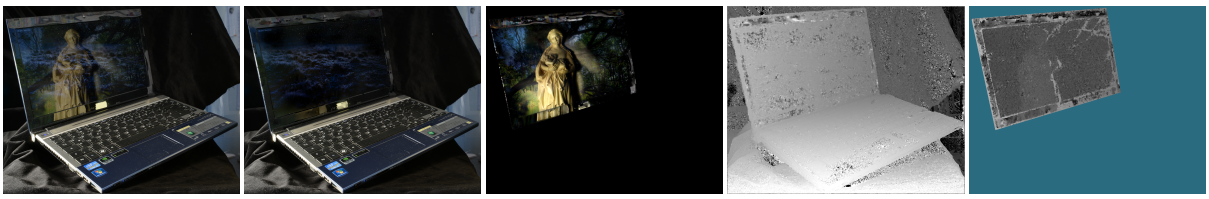

Figure 7: Results from real-world light field captured by a Gantry. *While for a human observer it is hard to separate the to superimposed layers on the laptop's screen, the proposed algorithm is capable of estimating the disparity for both layers as well as separating them accurately. For better visualization, the reflection layer intensity is scaled by a factor of two. The blue part in the second disparity map is masked out as no reflection is present there.*

0.1 seconds per matrix resulting in a runtime of arround 2 minutes for a whole lightfield. As each matrix has a size of 9375×625 and 6510×434, respectively, while only few entries are nonzero, we used MATLAB's sparse matrix operator to store these matrices. Otherwise storing all matrices completely would need up 20GB of memory. However, there is no GPU implementation of this sparse matrix operator, hence, in each iteration the matrices u and v have to be copied from the GPU to the CPU, where the matrix multiplication is performed and then moved back to the GPU, which again is time intensive and not optimal. Thus, runtimes can be significanlty improved by moving to a full GPU implementation. Performing one iteration of the primal-dual scheme using a NVIDIA GTX TITAN Black and an Intel i7-4770 takes just below 2 seconds, resulting in a total runtime in the scope of several minutes.

7. Conclusion

We propose a novel variational approach to separate a light field into multiple layers. For this, we first locally estimate disparity from the orientations of superimposed patterns on the epipolar plane images based on the framework in [WG13] and [AMS*06]. While they treat horizontal and vertical epipolar plane images individually, we make the approach more robust by constructing a joint second order structure tensor to recover the two orientations. The improved performance is demonstrated numerically on synthetic data, and visually on real-world light fields captured with a Lytro Illum plenoptic camera, which turn out to be very challenging for reconstruction.

The main contribution of the paper is the novel approach to segment the light field into layers from this input data. We first formulate a generative model to generate the complete light field from layer data on the center view. Based on this, we set up a variational inverse problem to optimize the fit of this model to the actually observed light field data. The problem is solved with a primal-dual scheme to recover the seperated layers. For synthetic data, this approach leads to reconstruction results wich are very close to ground truth. In addition, we show the feasibility of the approach on different types of captured datasets. In particular, the approach is robust enough to yield visually compelling results for the challenging data sets captured with a plenoptic camera.

Acknowledgements

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

Center view | Separated layer 1 | Separated layer 2 | Disparity layer 1 | Disparity layer 2

Figure 8: Reflection separation for real world light field captured with a Lytro Illum. *The disparity estimation was performed with the proposed algorithm, to identify the part of the image which contains a reflection a ground truth mask was used. The reflection of the bottle is seperated accurately, while the reflection of the ball object is only separated completely in the lower parts of the image. This is due to the fact that the disparity is very similar for both layers if object and reflecting surface are close together. Note that the calibration of the Lytro Illum is currently still work in progress, we believe the results can be much better once that is improved.*

References

[AMS*06] AACH T., MOTA C., STUKE I., MUEHLICH M., BARTH E.: Analysis of superimposed oriented patterns. *IEEE Transactions on Image Processing 15*, 12 (2006), 3690–3700. 3, 4, 7

[BBM87] BOLLES R., BAKER H., MARIMONT D.: Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision 1*, 1 (1987), 7–55. 3

[BBZZ03] BRONSTEIN A. M., BRONSTEIN M. M., ZIBULEVSKY M., ZEEVI Y. Y.: Blind separation of reflections using sparse ICA. In *Proc. Int. Conf. ICA* (2003), pp. 227–232. 1

[BKP10] BREDIES K., KUNISCH K., POCK T.: Total generalized variation. *SIAM Journal on Imaging Sciences 3*, 3 (2010), 492–526. 5, 6

[CKS*05] CRIMINISI A., KANG S., SWAMINATHAN R., SZELISKI R., ANANDAN P.: Extracting layers and analyzing their specular properties using epipolar-plane-image analysis. *Computer vision and image understanding 97*, 1 (2005), 51–85. 3

[CP10] CHAMBOLLE A., POCK T.: A first-order primal-dual algorithm for convex problems with applications to imaging. *preprint* (2010). 2, 6

[DPW13] DANSEREAU D. G., PIZARRO O., WILLIAMS S.: Decoding, Calibration and Rectification for Lenselet-Based Plenoptic Cameras. In *Proc. International Conference on Computer Vision and Pattern Recognition* (2013), pp. 1027–1034. 6

[FA99] FARID H., ADELSON E. H.: Separating reflections and lighting using independent components analysis. In *Proc. International Conference on Computer Vision and Pattern Recognition* (1999), vol. 1. 1

[GSZ12] GAI K., SHI Z., ZHANG C.: Blind separation of superimposed moving images using image statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 1 (2012), 19–32. 1

[GW13] GOLDLUECKE B., WANNER S.: The Variational Structure of Disparity and Regularization of 4D Light Fields. In *Proc. International Conference on Computer Vision and Pattern Recognition* (2013). 3

[KTS14] KONG N., TAI Y.-W., SHIN J.: A physically-based approach to reflection separation: from physical modeling to constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*, 2 (2014), 209–221. 1

[KZP*13] KIM C., ZIMMER H., PRITCH Y., SORKINE-HORNUNG A., GROSS M.: Scene Reconstruction from High Spatio-Angular Resolution Light Fields. *ACM Transactions on Graphics (Proc. SIGGRAPH) 32*, 4 (2013). 3

[LH96] LEVOY M., HANRAHAN P.: Light Field Rendering. In *Proc. SIGGRAPH* (1996), pp. 31–42. 2

[LZW04] LEVIN A., ZOMET A., WEISS Y.: Separating reflections from a single image using local features. In *Proc. International Conference on Computer Vision and Pattern Recognition* (2004), vol. 1, pp. I–306. 1

[Ng06] NG R.: *Digital Light Field Photography*. PhD thesis, Stanford University, 2006. 2

[PC11] POCK T., CHAMBOLLE A.: Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *International Conference on Computer Vision (ICCV 2011)* (2011). 2, 6

[SAA00] SZELISKI R., AVIDAN S., ANANDAN P.: Layer extraction from multiple images containing reflections and transparency. In *Proc. International Conference on Computer Vision and Pattern Recognition* (2000), vol. 1, pp. 246–253. 1

[SKB00] SCHECHNER Y., KIRYATI N., BASRI R.: Separation of transparent layers using focus. *International Journal of Computer Vision 39*, 1 (2000), 25–39. 1

[SKG*12] SINHA S. N., KOPF J., GOESELE M., SCHARSTEIN D., SZELISKI R.: Image-based rendering for scenes with reflections. *ACM Transactions on Graphics 31*, 4 (2012), 100. 1

[SSK99] SCHECHNER Y., SHAMIR J., KIRYATI N.: Polarization-based decorrelation of transparent layers: The inclination angle of an invisible surface. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on* (1999), vol. 2, IEEE, pp. 814–819. 1

[TKS06] TSIN Y., KANG S. B., SZELISKI R.: Stereo matching with linear superposition of layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 2 (2006), 290–301. 1

[WG13] WANNER S., GOLDLUECKE B.: Reconstructing reflective and transparent surfaces from epipolar plane images. In *German Conference on Pattern Recognition (Proc. GCPR)* (2013). 2, 3, 4, 5, 7

[WG14] WANNER S., GOLDLUECKE B.: Variational Light Field Analysis for Disparity Estimation and Super-Resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*, 3 (2014), 606–619. 3

[WMG13] WANNER S., MEISTER S., GOLDLUECKE B.: Datasets and benchmarks for densely sampled 4D light fields. In *Vision, Modelling and Visualization (VMV)* (2013). 6