Inline Double Layer Depth Estimation with Transparent Materials

Christian Kopf\textsuperscript{1(✉)}, Thomas Pock\textsuperscript{1}, Bernhard Blaschitz\textsuperscript{2}, and Svorad Štolc\textsuperscript{3}

\textsuperscript{1} Graz University of Technology, Graz, Austria
\{christian.kopf, pock\}@icg.tugraz.at
\textsuperscript{2} AIT Austrian Institute of Technology GmbH, Vienna, Austria
\textsuperscript{3} Photoneo s.r.o., Bratislava, Slovakia

Abstract. 3D depth computation from stereo data has been one of the most researched topics in computer vision. While state-of-art approaches have flourished over time, reconstruction of transparent materials is still considered an open problem. Based on 3D light field data we propose a method to obtain smooth and consistent double-layer estimates of scenes with transparent materials. Our novel approach robustly combines estimates from models with different layer hypotheses in a cost volume with subsequent minimization of a joint second order TGV energy on two depth layers. Additionally we showcase the results of our approach on objects from common inspection use-cases in an industrial setting and compare our work to related methods.

1 Introduction

Reconstructing 3D depth information through multi-view stereo methods is a well researched topic of computer vision and has lead to countless variations and approaches over the years. In order to obtain depth from correspondences in different views, the majority of approaches rely on the assumption that scene points reflect light in all directions uniformly, \textit{i.e.} the Lambertian assumption \cite{11}. Unfortunately this assumption is violated for scene points from transparent materials which poses a special case that is rarely considered in state-of-the-art approaches. Moreover little to no surface structure and refractive effects make the task even more difficult. Despite the challenging nature of the problem, computing depth from scenes with transparent materials is a desirable ability for industrial applications. The amount of publications regarding depth estimation for transparent materials is very sparse. Earlier publications explore different approaches to tackle the problem as for example through visible light tomography \cite{20} or polarized light \cite{15}. The trend for recent approaches is towards specialized hardware such as the utilization of Time-of-Flight (ToF) measurements \cite{7, 19} from RGB-D cameras and structured light setups \cite{10, 16, 25}. With the increasing popularity of learning based methods recent approaches show full 3D reconstructions of transparent objects. As an example recent work of Li \textit{et al.} \cite{14} extends the range of capable approaches by computing 3D transparent shapes.
with a small set of images from a mobile phone camera and a known environment map. Another learning-based publication of recent years is the single-image approach by Chen et al. [5] which is able to produce visually convincing transparent shapes from learned refractive flow of transparent objects. Despite the impressive results of these approaches, they are not applicable to our intentions. Our goal is to present a method in a more restricted environment. Firstly, we limit our approach to passive (multi-view) data only. We also assume a front-parallel multi-view alignment along exactly one image axis which is common for stereo data. To allow for applications in an industrial setting, e.g. objects on a conveyor belt, our approach is also required to process scenes which are subjected to linear motion while guaranteeing reasonably fast runtime. The intention is to allow continuous acquisition with parallel processing of the data. The scene structure modeled by our approach consists of a transparent object in front of an opaque background. Thus the goal is to estimate two depth estimates globally across an entire scene, i.e. a front layer for non-opaque surfaces and an opaque back-layer for scene points that may be underneath, see Fig. 1 for an example. In the context of passive (multi-view) data, a popular method to describe densely sampled scenes is given by the plenoptic function [2] which enables the description of 3D scene geometry in a light field (LF) [6]. A simplified variant of a light field can be interpreted intuitively through a discrete volume in 3D, comprised of a stack of images from different views. The acquisition of such light fields is commonly subject to epipolar constraints, thus enabling depth estimation through light field analysis. For scene points of transparent materials, multiple depth cues superimpose which results in local multi-orientation structures in light field data. In [22,23] Wanner and Goldlücke show how local structure tensor (ST) analysis [1] in the epipolar domain of light fields can be utilized to solve correspondence problems for Lambertian and non-Lambertian surfaces. The major drawback of this approach is that the depth estimates are bound to their corresponding model hypothesis, which implies that depth estimates are only justified in regions of a scene where the respective hypothesis is valid. Johannsen et al. [8,9] present a different approach by solving coding problems with a dictionary of pattern orientations in a local neighborhoods to compute multi-layer depth estimates.
based on 4D light fields. Our experiments have shown that conventional stereo matching and line fitting in the epipolar domain gives bad results when it comes to transparent materials. In comparison we have found that the double orientation structure tensor approach is well suitable for transparent materials since this model is very sensitive to depth cues from non-Lambertian surfaces. In this work we will adopt the basic idea of depth estimation with structure tensors on narrow baseline light field data from an inline setup with the aforementioned requirements. We will start by reviewing our setup and explain the light field structure in Sect. 2. Furthermore we will give a brief introduction to structure tensor analysis in the context of light fields and in Sect. 4 follow up with our proposed method to combine estimates in a cost volume. After that we will explain how we refine our double-layer results through a novel variational method.

2 Setup and Light Fields

Our assumption on the general scene structure is that a transparent object of arbitrary shape is located in front of an opaque second layer as in Fig. 2. This example shows a clear tape spanned over a coin. The objective is to reconstruct the depth of the tape surface as well as the depth of the coin underneath. Acquisition rays that traverse through the transparent object carry composite depth information of at least two surfaces. This leads to some areas for which a single-layer depth hypothesis is valid and some areas for which a double-layer hypothesis is valid. In addition the problem becomes even more complicated due to opaque structures on the transparent object which might occlude the subjacent background. As the final result our goal is to obtain two depth images of size $M \times N$ where the depth estimates among both images coincide in single-layer regions and differ in double-layer regions. The input data is assumed to be 3D light field data which is a mapping of the form

$$\mathcal{L} : (x, y, s) \mapsto \mathbb{R}^3,$$

(1)
Fig. 3. Illustration of the acquisition system. A line-scan camera is set up in line with a linear transport stage. Through the alignment of the transport direction and the horizontal image axis each scan line acquires a view of the entire scene incrementally while the object on the transport stage passes the camera. The baseline of the resulting light field is determined by the distance between scan lines.

where $x \in \{1, \ldots, M\}$, $y \in \{1, \ldots, N\}$ denote the image coordinates and $s \in \{1, \ldots, P\}$ the view-index. More general 4D light field data can be adapted to this assumption by dropping the dimension corresponding to the movement along the second image axis. To illustrate the content of LF data in this paper we will depict the central view with the view index

$$s_{\text{ref}} = \frac{P + 1}{2}$$

as a substitute for the whole light field. All data that is used in this paper is listed in Table 1. To evaluate our proposed method we will use the publicly available “Maria” dataset from [24] and adopt it to our 3D light field setup by discarding all but the 9 central views in the horizontal direction of movement. In addition we acquired multiple light fields of transparent objects with an inline acquisition system based on the principles of [21]. An Illustration and a short Explanation of the setup can be found in Fig. 3.

3 Depth from Light Field Analysis

The depth information of a LF scene can be computed through local orientation analysis in epipolar plane images (EPI)

$$\mathcal{E}(s, y) : (s, y) \mapsto \mathbb{R}^3.$$  

(3)

EPIs are images comprised of projections from acquisitions rays sharing common epipolar planes. By choosing a fixed value $x = \tilde{x}$ thus one obtains

$$\mathcal{E}_{\tilde{x}}(s, y) := \mathcal{L}(\tilde{x}, y, s).$$  

(4)

An example of such an EPI is depicted in Fig. 2. From this figure it can be seen how different depth layers from the scene impose multi-orientation patterns with certain angles in the epipolar domain of a LF. Note that the angle of these
patterns is linked to the depth of the corresponding scene points. To obtain depth estimates we thus apply an orientation analysis through the aforementioned structure tensor approach [22]. Since the orientation analysis is performed locally at each point in the light field \((M \times N \times P)\) we obtain a result stack with the same dimensions.

**Single Orientation Structure Tensor (SOST):** The purpose of the single-orientation model [1] is to compute depth estimate for opaque or occluded scene points, i.e. single-layer regions. By finding the local orientation \(w(\xi)\) in a local neighborhood \(\Omega\) through minimization of the least squares energy term

\[
\min_{\xi} E_1(\xi) = \int_{\Omega} (w(\xi)^T \nabla \mathcal{E}_x)^2 \, dp ,
\]

we obtain an angular estimate \(\xi\), where \(p = [s, y]^T\) is the image pixel of the EPI \(\mathcal{E}_x\) and

\[
\nabla \mathcal{E}_x = \left[ \frac{\partial \mathcal{E}_x}{\partial s}, \frac{\partial \mathcal{E}_x}{\partial y} \right]^T
\]

denotes the gradient vector in the EPI domain. Reformulating Eq. (7) leads to the definition of the single orientation structure tensor \(\mathcal{S}\),

\[
\min_{\xi} E_1(\xi) = w(\xi)^T \left( \int_{\Omega} \nabla \mathcal{E}_x (\nabla \mathcal{E}_x)^T \, dp \right) w(\xi) = w(\xi)^T \mathcal{S} w(\xi) .
\]

From the eigenvalue analysis on \(\mathcal{S}\), \(w(\xi)\) is obtained as the eigenvector corresponding to the smaller eigenvalue. The resulting estimate stack for the SOST model will be denoted by \(\mathcal{H}_1(x, y, s)\).

**Second Order Double Orientation Structure Tensor (SODOST):** By extending the single-orientation case to the double-orientation case [1] with an additive composition of patterns in a local neighborhood \(\Omega\), the single orientation model can be extended to the double orientation model. In a similar fashion to the SOST model, the optimal solution for a given multi-orientation pattern patch \(\Omega\) can be computed by minimizing

\[
\min_{m(\theta, \gamma)} E_2(m(\theta, \gamma)) = m(\theta, \gamma)^T \left( \int_{\Omega} (D^2 \mathcal{E}_x)(D^2 \mathcal{E}_x)^T \, dp \right) m(\theta, \gamma)
\]

\[
= m(\theta, \gamma)^T \mathcal{T} m(\theta, \gamma) ,
\]

where \(m(\theta, \gamma)\) denotes the MOP vector [1] and

\[
D^2 \mathcal{E}_x = \left[ \frac{\partial^2 \mathcal{E}_x}{\partial s^2}, \frac{\partial^2 \mathcal{E}_x}{\partial y \partial s}, \frac{\partial^2 \mathcal{E}_x}{\partial y^2} \right]^T .
\]

Because the structure tensor model \(\mathcal{T}\) is comprised of second order derivatives we will refer to this model as the second order double orientation structure tensor (SODOST) model. Through an eigenvalue analysis with a subsequent root solving on this structure tensor we obtain the two orientations \(u(\theta)\) and \(v(\gamma)\) or the corresponding disparities where it is assumed that \(u(\theta)\) denotes the estimate
closer to the camera. By performing this local analysis based on this model we thus obtain two further estimate volumes $H_2(x, y, s)$ (front) and $H_3(x, y, s)$ (back). The quality of the structure tensor estimates greatly depends on the choice of the inner and outer scale for Gaussian smoothing. This is explained in more detail in [22]. For the “Maria” dataset we chose to use the same parameters as the referenced paper (i.e. $\sigma_i = \sigma_o = 0.8$) and for all other datasets we chose to use a slightly larger outer scale $\sigma_o = 1.2$.

In our experiments we have observed that the structures in EPIs are a lot more prominent on opaque surfaces compared to transparent scene points. Prior to computing the structure tensor estimates we therefore normalize the contrast of the entire light field to enhance the structure imposed by transparent surfaces. Let $I \in \mathbb{R}^{M \times N \times P}$ denote a single channel of the light field, we apply the normalization filtering

$$I_n = \frac{I - k_\sigma * I}{c(k_\sigma * (I - k_\sigma * I)^2)^{\frac{1}{2}} + \epsilon} + \frac{1}{2}, \quad (10)$$

where $*$ is the volumetric convolution and $k_\sigma$ denotes a Gaussian kernel. The parameter $c$ is used to scale the normalized output to $[0, 1]$. In our implementation we used $c = 6$. We chose $k_\sigma$ to be a 2D kernel in spatial dimensions with a standard deviation $\sigma = 2$.

4 Double-Layer Cost Volume

The major drawback of the plain structure tensor results are the affiliation to either a single-layer or double-layer hypothesis which is valid in certain regions of the scene. Our approach combines the results from all models and implicitly rules out incorrect or weakly supported estimates. To achieve this in a robust manner we create a cost volume $V \in \mathbb{R}^{M \times N \times D}$ where $D$ is the number of elements in a finite set {$d_0, d_1, ..., d_{D-1}$} of disparity hypotheses. As depicted in Fig. 2 we can observe that the angle of the sloped pattern lines in the epipolar domain of the light field correspond to different depths and subsequently disparities in a light field. Recall that we compute depth estimates at each point in a local neighborhood of the light field, such that we attain the estimate volumes $H_1$, $H_2$ and $H_3$. Because of the locality of the structure tensor operation, we can observe that the estimates along the corresponding sloped line in the light field ideally are constant. If the variance along this sloped line is high we want to penalize this with a high cost. To analyze the estimates along any sloped line we thus formulate a ray which intersects the reference view $s_{\text{ref}}$ at the coordinates $x, y$ with an angle $\alpha$. Since disparities can be translated to angles we can therefore use the set of disparity hypotheses from above to determine the intersected voxel in each view for each hypothesis ray. The relative shift for a certain view $s_i$ with respect to $s_{\text{ref}}$ can therefore be computed with

$$\Delta y_{i \rightarrow \text{ref}} = \lfloor \tan(\alpha)(s_i - s_{\text{ref}}) \rfloor. \quad (11)$$
In theory each ray thus intersects up to $P$ voxels. We define that all these voxels of a single ray form a finite set $R_{x,y,d_i}$. In a geometric perspective we can alter the thickness of this ray by also selecting neighboring voxels in each view such that we can set its radius of influence resulting in more or fewer voxels in the set. An illustration of this principle is given in Fig. 4. With this in mind the cost for each point of the cost volume can therefore be attained through

$$V(x, y, d_i) = \frac{1}{|R_{x,y,d_i}|} \sum_{r \in R_{x,y,d_i}} \min_j h(H_j(r), d_i),$$  \hspace{1cm} (12)

where $|R_{x,y,d_i}|$ denotes the number of elements in the set $R_{x,y,d_i}$ and $r$ denotes a point in light field coordinates. For the inner cost metric we chose to use a stable absolute distance from the disparity hypothesis $d_i$

$$h(z, d_i) = \max \{|z - d_i|, \tau\},$$  \hspace{1cm} (13)

where $\tau$ is a hyperparameter. We remark that the structure tensor estimates $H_j(r)$ from each model may be incorrect and noisy in certain regions. With the given cost measure above a combined robust cost over all three structure tensor models in a local neighbourhood along a hypothesis ray is provided. The cost volume in Fig. 5 shows the unimodal and bimodal distribution along $d$ for any location $x, y$ on the spatial grid. Naturally the distribution can have more than two modes but in our testing we have found that either one or two prominent modes are present in the cost volume. The task now is to find the index $d$ and the corresponding cost for one or two prominent modes on mostly smooth distributions. To isolate the minimal points we use a simple non-maximum-suppression algorithm denoted in Algorithm 1 and define that a local minimum
is only valid if it is below a certain threshold and no local maximum is within a certain range along \( d \). The algorithm is based upon a local gray value dilation operation with a subsequent comparison to the unchanged cost volume.

For each \( x, y \) determining the index of the smallest minimum forms an image which is denoted \( g_1 \). For areas with a valid second best minimum we likewise determine the indices which form \( g_2 \). For single-layer coordinates we use the corresponding entry from \( g_1 \). Since the estimates in both images have no clear affiliation to the front-layer or the back-layer we sort both images such that

\[
(g_1)_i \geq (g_2)_i \quad \forall i = 1, ..., MN ,
\]

where \((g_j)_i\) denotes the i-th element of a flattened version of \( g_j \).

5 Joint TGV\(^2\)–\(\ell_1\) Refinement

In this section, we propose a joint refinement approach to refine noisy depth estimates for the foreground surface \( u \) and a background surface \( v \). As a regularizer we make use of the Total Generalized Variation (TGV) regularizer \([3]\) of second order which has been shown to be very suitable for the refinement of depth images \([12, 17]\). The TGV regularizer of second order is defined as

\[
\text{TGV}^2(x) = \min_{\tilde{x}} \lambda_0 \left\| \tilde{D} \tilde{x} \right\|_{2,1} + \lambda_1 \left\| D x - \tilde{x} \right\|_{2,1} ,
\]

where \( x \in \mathbb{R}^{M \times N} \) is the image and \( \tilde{x}, \in \mathbb{R}^{M \times N \times 2} \) is an auxiliary vector field. \( \tilde{D} : \mathbb{R}^{M \times N \times 2} \to \mathbb{R}^{M \times N \times 4} \) and \( D : \mathbb{R}^{M \times N} \to \mathbb{R}^{M \times N \times 2} \) are the finite differences approximations of the gradient operator. For robust joint depth refinement we propose to minimize the following variational energy

\[
\min_{u,v} \text{TGV}^2(u) + \text{TGV}^2(v) + \lambda_u \left\| c_1 \odot (u - g_1) \right\|_1 + \lambda_v \left\| c_2 \odot (v - g_2) \right\|_1 \quad \text{s.t. } u \geq v ,
\]

where the optimal solution for the foreground \( u \) is constrained to be at least as close to the camera as the background \( v \). Here \( c_1, c_2 \in \mathbb{R}^{M \times N} \) denote the
Algorithm 1. Implementation of the non-maximum suppression. The $\odot$ operator denotes the Hadamard product and $\delta_B$ is the 1-0 indicator function on condition $B$.

1: **procedure** Non-Maximum Suppression($V(x,y,d)$)  
2: $ V_d(x,y,d) \leftarrow \text{dilate}(V(x,y,d)) $ \quad \triangleright \text{Dilate along } d \text{ with width=3} 
3: $ V_{\text{NMS}}(x,y,d) \leftarrow V(x,y,d) \odot \delta_{V(x,y,d)=V_d(x,y,d)} $. 
4: **return** $V_{\text{NMS}}(x,y,d)$ 
5: **end procedure**

reciprocal values of the residual cost from the solutions $g_1$ and $g_2$ to steer data fidelity locally and the $\odot$ operator denotes the Hadamard product. By transforming Eq. (16) into the corresponding saddle-point notation in the general form of

$$
\min_x \max_y (Kx)^T y + h(x) - f^*(y),
$$

the problem can adequately be solved through the PDHG approach by Chambolle and Pock [4]. To apply the algorithm to our problem, the proximal operators $\text{prox}_{\tau h}$ and $\text{prox}_{\sigma f}$ have to be determined. Note that both can be computed element-wise. Since the convex conjugate of $f : \| \cdot \|_{2,1}$ is given by the indicator function on the $2, \infty$-norm ball, the proximal map in dual space is given by

$$
\text{prox}_{\sigma f^*}(y)_i = \frac{y_i}{\max\{1,\|y_i\|_2\}}.
$$

Unfortunately the operator in primal-space is not so straight forward since we need to pay attention to the constraint $u \geq v$. We can formulate the term $h(x)$ from the saddle point notation in the following way

$$
h(u,v) = \lambda_u \|c_1 \odot (u - g_1)\|_1 + \lambda_v \|c_2 \odot (v - g_2)\|_1 + \delta_{u \geq v}(u,v),
$$

where the last term is the indicator function of the constraint

$$
\delta_{u \geq v}(u,v) = \begin{cases} 
0 & u \geq v \\
\infty & \text{else}
\end{cases}.
$$

As a first step we compute the solutions $\hat{u}$ and $\hat{v}$ by the well known shrinkage operator

$$
\text{prox}_{\tau h}(x)_i = (g_j)_i + \max(0,|x_i - (g_j)_i| - \lambda_x \tau(c_j)_i) \text{sign}(x_i - (g_j)_i).
$$

In case $\hat{u} < \hat{v}$ the shrinkage operator is invalid. The closest valid solution is given by $u = v$. The joint proximal map in primal space thus becomes

$$
\arg \min_u \lambda_u (\|c_1 \odot (u - g_1)\|_1 + \|c_1 \odot (u - g_2)\|_1) + \frac{1}{2\tau}(\|u - \bar{u}\|_2^2 + \|u - \bar{v}\|_2^2),
$$

(22)
where \( \bar{u} \) and \( \bar{v} \) are the solutions from the previous step. Since this minimization problem is of the form

\[
\underset{u}{\text{arg min}} \sum_{i=1}^{K} w_i |u - g_i| + F(u),
\]

we can solve this through the median formula proposed by Li and Osher [13].

6 Results

To the best of our knowledge there is no dedicated benchmark for transparent depth estimation. Nevertheless we compare our results quantitatively on the publicly available “Maria” dataset [24] and show results on more difficult self acquired datasets.

Quantitative Results: In Fig. 6 the results of our quantitative evaluation are shown. These results show that our procedure is at least on par with comparable methods on the mentioned dataset. However, we use only a small fraction of the data since we deliberately operate in an inline context. As a baseline we consider the results from only the SODOST model with subsequent variational denoising [4,18] of the central view which is closely related to the proposed method in [23]. While the transparent plane is already estimated well, these results lack consistency in single layer regions which is a major requirement of

|       | MSE  | BadX X = 0.2 | BadX X = 0.5 | BadX X = 1.0 |
|-------|------|--------------|--------------|--------------|
| SODOST| Front| 1.71         | 5.73         | 2.10         | 0.14         |
|       | Back | 7.69         | 12.87        | 6.09         | 2.88         |
|       | Combined | 3.83 | 8.26        | 3.51         | 1.11         |
| SODOST + TV-\(\ell_2\)| Front | 0.45 | 2.26       | 0.15         | 0.03         |
|       | Back | 4.75         | 9.96         | 5.07         | 1.51         |
|       | Combined | 1.97 | 4.98       | 1.89         | 0.55         |
| [9]   | Front | 4.89 | -          | -            | -            |
|       | Back | 3.20         | -            | -            | -            |
|       | Combined | -   | -          | -            | -            |
| [8]   | Front | 0.79 | -          | -            | -            |
|       | Back | 1.45         | -            | -            | -            |
|       | Combined | -   | -          | -            | -            |
| Ours  | Front | 1.26         | 5.21         | 0.24         | 0.01         |
|       | Back | 0.73         | 3.29         | 0.06         | 0.00         |
|       | Combined | 1.07 | 4.53       | 0.18         | 0.01         |

Fig. 6. Quantitative results on the “Maria” dataset [24]. The metrics shown on the left are the mean squared disparity error MSE times 100 and BadX. The later describes the percentage of pixels with a disparity error greater than a threshold \(X\). On the right the corresponding images including the ground truth are depicted. The stated scores are masked (2\textsuperscript{nd} row on the right) where only white mask pixels are considered. To provide better comparability we also state “Combined” values which are the average masked metrics across both layers (front and back).
Fig. 7. Results on acquired data. The key information regarding each dataset is stated in Table 1. Since transparent materials in the depicted data may be hard to identify the second column shows the manual semantic segmentation for each scene in green (single-layer) and blue (double-layer). We refer the reader to the online version of this paper for the color coding. The third column shows the result from the single orientation structure tensor model and the 4th and 5th columns show results from the double orientation model. Our final results are shown in the 6th and 7th column.

our application as stated in Sect. 2. [23] solves this problem through consistency checks among results from horizontal and vertical EPIs, which is not applicable in our setup (3D vs. 4D LF). Unfortunately we were not able to provide a proper
consistency metric because single orientation regions of the dataset are masked. We therefore leave the interpretation of this aspect to the reader and point to further qualitative results from our acquired datasets. Another crucial point is that we were also able to achieve a low runtime. The runtime for this dataset in our python testing environment is 11.3 seconds. For our larger light fields of size $1150 \times 1172 \times 33$ and 6 channels the average runtime is $\approx 35$ s. These times naturally depend on the chosen hyperparameters. Most significantly the size of the cost volume has a major impact on this. For these experiments we chose $D = 80$. These runtime numbers are based on a AMD Ryzen 3900X platform with the support of a NVidia RTX Titan GPU.

**Qualitative Results:** Based on acquisition with the setup in Fig.3 we also present qualitative results in Fig.7 on various use-cases e.g. objects in bags or differently shaped transparent objects which we also consider as a valuable contribution. In case of the “Maria” dataset the transparent object is ideally conditioned for concept studies with a high separation between the front-layer and the back-layer and little to no reflections on a planar structure. However we think that the problem becomes more interesting with the use-cases given by our acquisitions. With the depicted results we show that our approach also works well for more complex shapes of transparent objects in real world data. As an example consider the datasets D7 and D8. It is remarkable that although both examples have a large amount of reflections, the shape of the objects from inside the clear bag can be reconstructed. Also depth of the clear bag itself is estimated well. Similarly the front SODOST estimate of the “Coin & Tape” example (D1) shows heavy reflective effects on the coin and noise at the background. By incorporating single orientation estimates from the single orientation structure tensor, our model is able to circumvent this issue. With our approach it is possible to attain a smooth separation between both layers while maintaining consistency for the rest of the scene. All hyperparameters of our method that have not been stated explicitly have been chosen empirically based on the individual datasets.

**7 Conclusion**

In this paper we presented a method to obtain smooth double-layer results for a constrained inline setup based on 3D light field data from transparent objects. We covered the preliminaries on various structure tensor models and explained how we combine these estimates robustly in a cost volume. We furthermore presented a refinement procedure based on a joint $\text{TV}^2-\ell_1$ regularizer which enables us to handle two layers simultaneously while keeping a global depth ordering. Additionally we presented quantitative numbers and a qualitative evaluation on acquired data to demonstrate the applicability of our approach. To build a basis for future publications, the ability to benchmark depth reconstruction could be improved. Possible future work could therefore include the acquisition of a large high quality dataset with transparent objects which also could enable the development of learned methods. We believe that future approaches can improve the runtime and accuracy further by incorporating learned descriptors for non-Lambertian surfaces.
References

1. Aach, T., Mota, C., Stuke, I., Muhlich, M., Barth, E.: Analysis of superimposed oriented patterns. IEEE Trans. Image Process. 15(12), 3690–3700 (2006)
2. Adelson, E.H., Bergen, J.R.: The plenoptic function and the elements of early vision. In: Landy, M.S., Movshon, J.A. (eds.) Computational Models of Visual Processing, pp. 3–20. The MIT Press, Cambridge (1991)
3. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM J. Imaging Sci. 3(3), 492–526 (2010). https://doi.org/10.1137/090769521
4. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. 40(1), 120–145 (2011). https://doi.org/10.1007/s10851-010-0251-1
5. Chen, G., Han, K., Wong, K.Y.K.: TOM-Net: learning transparent object matting from a single image. In: CVPR (2018)
6. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, pp. 43–54. Association for Computing Machinery, New York (1996). https://doi.org/10.1145/237170.237200
7. Ji, Y., Xia, Q., Zhang, Z.: Fusing depth and Silhouette for scanning transparent object with RGB-D sensor. Int. J. Opt. 2017, 1–11 (2017). https://doi.org/10.1155/2017/9796127
8. Johannsen, O., Sulc, A., Goldlücke, B.: Occlusion-aware depth estimation using sparse light field coding. In: German Conference on Pattern Recognition (Proceedings of the GCPR) (2016)
9. Johannsen, O., Sulc, A., Goldlücke, B.: What sparse light field coding reveals about scene structure. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
10. Kim, J., Reshetouski, I., Ghosh, A.: Acquiring axially-symmetric transparent objects using single-view transmission imaging. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1484–1492 (2017)
11. Koppal, S.J.: Lambertian reflectance. In: Ikeuchi, K. (ed.) Computer Vision: A Reference Guide, pp. 441–443. Springer, Boston (2014). https://doi.org/10.1007/978-0-387-31439-6_534
12. Kuschk, G., Cremers, D.: Fast and accurate large-scale stereo reconstruction using variational methods. In: Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops, pp. 700–707 (2013)
13. Li, Y., Osher, S.: A new median formula with applications to PDE based denoising. Commun. Math. Sci. 7(3), 741–753 (2009). https://projecteuclid.org:443/euclid.cms/1256562821
14. Li, Z., Yeh, Y.Y., Chandraker, M.: Through the looking glass: neural 3D reconstruction of transparent shapes. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
15. Miyazaki, D., Ikeuchi, K.: Inverse polarization raytracing: estimating surface shapes of transparent objects. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 2, pp. 910–917 (2005)
16. Qian, Y., Gong, M., Yang, Y.: 3d reconstruction of transparent objects with position-normal consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), pp. 4369–4377 (2016)
17. Ranftl, R., Gehrig, S., Pock, T., Bischof, H.: Pushing the limits of stereo using variational stereo estimation. In: 2012 IEEE Intelligent Vehicles Symposium, pp. 401–407 (2012)

18. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D 60(1), 259–268 (1992). https://doi.org/10.1016/0167-2789(92)90242-F. http://www.sciencedirect.com/science/article/pii/016727899290242F

19. Tanaka, K., Mukaigawa, Y., Kubo, H., Matsushita, Y., Yagi, Y.: Recovering transparent shape from time-of-flight distortion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), pp. 4387–4395 (2016)

20. Trifonov, B., Bradley, D., Heidrich, W.: Tomographic reconstruction of transparent objects. In: ACM SIGGRAPH 2006 Sketches, SIGGRAPH 2006, p. 55-es. Association for Computing Machinery, New York (2006). https://doi.org/10.1145/1179849.1179918

21. Štrolc, S., Huber-Mörk, R., Holländer, B., Soukup, D.: Depth and all-in-focus images obtained by multi-line-scan light-field approach. In: Electronic Imaging (2014)

22. Wanner, S., Goldlücke, B.: Globally consistent depth labeling of 4D lightfields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)

23. Wanner, S., Goldlücke, B.: Reconstructing reflective and transparent surfaces from epipolar plane images. In: German Conference on Pattern Recognition (Proceedings of the GCPR, Oral Presentation) (2013)

24. Wanner, S., Meister, S., Goldlücke, B.: Datasets and benchmarks for densely sampled 4D light fields. In: Vision, Modelling and Visualization (VMV) (2013)

25. Wu, B., Zhou, Y., Qian, Y., Cong, M., Huang, H.: Full 3D reconstruction of transparent objects. ACM Trans. Graph. 37, 1–11 (2018). https://doi.org/10.1145/3197517.3201286