Vision, Modeling, and Visualization (2013)
Michael Bronstein, Jean Favre, and Kai Hormann (Eds.)

Epipolar Plane Image Refocusing for Improved Depth Estimation and Occlusion Handling

Max Diebold and Bastian Goldluecke

Heidelberg Collaboratory for Image Processing

Abstract

In contrast to traditional imaging, the higher dimensionality of a light field offers directional information about the captured intensity. This information can be leveraged to estimate the disparity of 3D points in the captured scene. A recent approach to estimate disparities analyzes the structure tensor and evaluates the orientation on epipolar plane images (EPIs). While the resulting disparity maps are generally satisfying, the allowed disparity range is small and occlusion boundaries can become smeared and noisy. In this paper, we first introduce an approach to extend the total allowed disparity range. This allows for example the investigation of camera setups with a larger baseline, like in the Middlebury 3D light fields. Second, we introduce a method to handle the difficulties arising at boundaries between fore- and background objects to achieve sharper edge transitions.

1. Introduction

Light field imaging takes advantage of the possibility to split the captured light into single rays. The light field of a static scene is described by the plenoptic function [AB91], which assigns an intensity value to rays defined by location and direction. In the case that the scene is contained entirely within a closed 2D surface, the plenoptic function has redundant information, because the intensity along rays outside the surface remains constant. Thus, the 4D function is sufficient to describe the light field outside this surface, which was a central idea of the Lumigraph model [LH96, GGSC96].

To capture a lumigraph, several methods have been established in the last couple of years. Earlier approaches employed multi camera arrays [VWJL04], where several independent cameras located on a common camera plane, capture the scene from slightly different positions. A similar approach and less cost intensive is the acquisition of light fields with gantries. A single camera, mounted on a gantry, moves on a 2D plane to different positions to capture light fields. However, this method is restricted to static scenes. Recently, plenoptic cameras have become commercially available [PW10, NLB*05, GL10], which employ multi-lens arrays in front of a single sensor to obtain angular information about the captured scene.

The transformation of the plenoptic function to a Lumigraph allows a wide range of applications. A major line of research investigates light field rendering [SCK07, MB95, KAC07], which deals with view interpolation, i.e. the generation of novel views from perspectives different from the recorded ones. Other applications are investigated in computational photography, such as virtual refocusing and reconstruction of occluded surfaces [VLS*06, VGT*05]. These approaches also allow disparity estimation based on the directional origin of the light beams by capturing different parts of the light field with different cameras.

Much research has also been devoted to disparity computation in the so called epipolar plane images (EPI), which emphasize a fundamental advantage of the 4D light field structure. Namely, a single 3D scene point is projected onto a single line in the EPI, which can be more robustly detected than a point correspondence [BBM87, CKS*05].

Our proposed method builds upon the EPI-based approach of Wanner and Goldluecke [WG12]. Here, the disparity is computed in the EPIs by evaluating the orientation as seen in Figure 3 using the structure tensor, which is known to yield robust and accurate results for orientation. However, due to the local nature of the involved derivative filters, this method is able to recover disparities only in a total range of two pixels, which implies small baselines between the cameras. Thus the depth analysis method severely restricts the possible camera setup parameters. To estimate disparities with an arbitrary camera parameter set, the range of two pixels is in most cases not sufficient, which we alleviate in this work using a refocusing scheme.

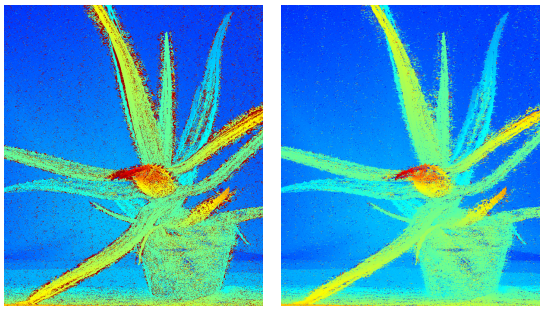

Figure 1: *The above disparity maps correspond to the center view of the Middlebury Aloe 3D (1x7) dataset. On the left is, the global disparity map superimposed by refocusing to several virtual depths. On the right is the result of improved disparity estimation with applied occlusion handling.*

A related approach to solve this problem is proposed in [ZGFN08, GFM*07]. They introduce a plane sweeping method for multi-view stereo algorithms, which generates depth maps for three surface-aligned sweeping directions. The resulting depth map is a per-pixel selection of this depth estimations, solved by a regularization using a global energy. The proposed approach is similar to the implemented algorithm in this paper, where also several local depths maps are superimposed using an reliability measure.

Another problem appearing in light field disparity estimation occurs at occlusion boundaries, which severely degrade the estimation of disparity values in the vicinity. This problem increases with larger baselines and smoothing kernels and becomes worse with increasing scene depth. If the underlying light fields have restricted depth ranges and are four-dimensional, this problem can be somewhat compensated for by merging the results for horizontal and vertical EPIs [WG12]. Datasets like the Middlebury stereo benchmark [SS02, SP07, SS03, HS07], however, include rendered light fields which are only 3D, i.e. have one direction of camera movement, with large baselines and a correspondingly large disparity range. These light fields have much less information than a densely sampled 4D light field, and thus the disparity estimation at boundaries is much more difficult.

The contribution of this paper is twofold. First, we lift the restriction on the disparity range by refocusing the EPIs to several virtual depth layers. and merging the results from all independent layers into a global disparity map, see Figure 1. Second, the proposed refocusing scheme allows to elegantly exploit the possibility to look slightly behind objects located in the foreground by shifting the view point. In occlusion areas, this allows to fill in regions of depth uncertainty from neighboring views which do not suffer from occlusion, and thus boundary transitions become sharper, see Figure 1.

2. Local disparity estimation

We describe a light field using the standard two-plane parametrization, see Figure 2. Rays are defined using two parallel planes Π and Ω. The first plane Ω denotes image coordinates $(x,y) \in \Omega$. The second plane Π contains the focal points $(s,t) \in \Pi$ of all cameras. An entire (gray scale) 4D light field can thus be described by a function

$$L : \Omega \times \Pi \to \mathbb{R} \qquad (s,t,x,y) \mapsto L(s,t,x,y), \tag{1}$$

where $L(s,t,x,y)$ defines the intensity of the corresponding ray defined by the intersection (x,y) with the image plane and (s,t) with the focal plane, respectively. Disparities are estimated locally on 2D slices Σ_{t^*,y^*} through the 4D light field structure, which arise from setting e.g. t to a fixed value t^* and y to a fixed value y^*. The restriction of L to such a slice is called an epipolar plane image

$$S_{t^*,y^*} : \Sigma_{t^*,y^*} \to \mathbb{R} \tag{2}$$

$$(x,s) \mapsto S_{t^*,y^*}(x,s) := L(s,t^*,x,y^*), \tag{3}$$

see Figure 3. Other slices with different fixed coordinates are defined analogously.

The local disparity estimation computes the orientation of lines of constant intensity in the EPI, which (for Lambertian scenes) yields information about the disparity for the respective scene point. Disparity for the complete light field can thus be computed by considering a set of EPIs which covers the complete ray space. However, the result can be made more robust by using redundant information, like the set of all horizontal and all vertical EPIs, which covers the ray space exactly twice [WG12].

In this paper, we focus on 3D light fields, where only one view point coordinate changes and thus epipolar plane images exist only for one focal coordinate. The following equations are specialized to 3D light fields but can be extended to 4D light fields in an obvious manner: Both EPI directions in a 4D light field can be computed independently and superimposed with a coherence merge afterwards.

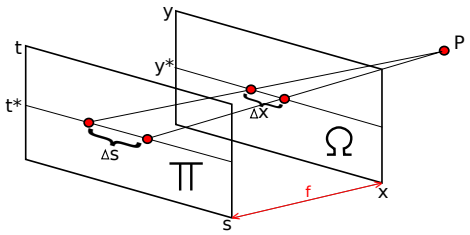

Figure 2: *Two-plane parametrization of a 4D light field by coordinates (x,y) in the image plane Ω and coordinates (s,t) the camera plane Π, which describes the projection of every 3D Point P into every camera.*

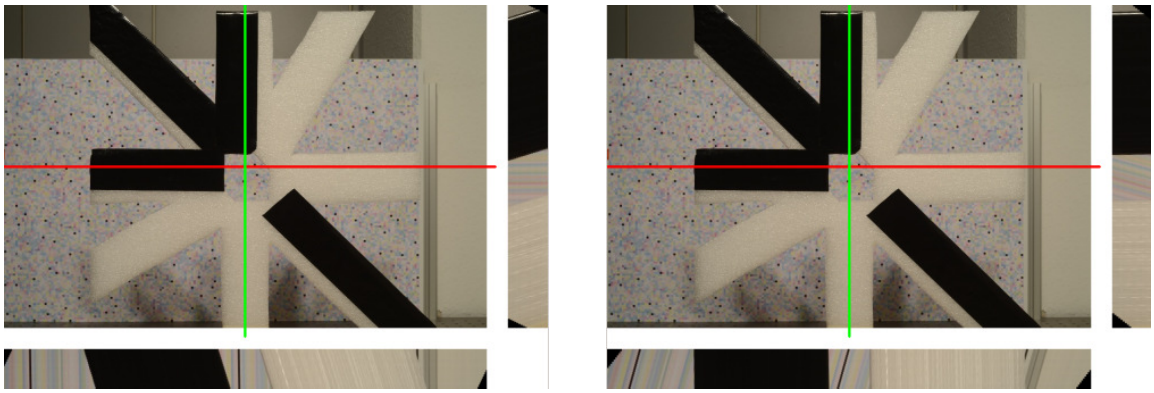

Figure 3: *The shown EPIs are related to the vertical(red) and horizontal(green) lines in the corresponding images. The right image, is focused on a virtual depth Z_1 and the left image is focused on a virtual depth Z_2. Objects placed at the focused depth have vertical orientations in the horizonal EPI and horizontal orientation vertical EPI. The left image is focused on the foreground and the right on the background.*

3. Global disparity estimation

The local disparity estimation based on a structure tensor delivers reliable solutions only within a two pixel range, since the underlying computation of the derivatives is based on a Scharr filter. The 3×3 kernel size of the Scharr filter restricts the range in which the estimation is possible. All lines corresponding to a scene point with an absolute disparity of more than one pixel cannot be detected properly since consecutive pixels belonging to the line do not fit in the filter window anymore. While the estimation of disparities larger than one pixel is still possible because of the pre-smoothing of the image with a Gaussian kernel, the error increases drastically, see Figure 4.

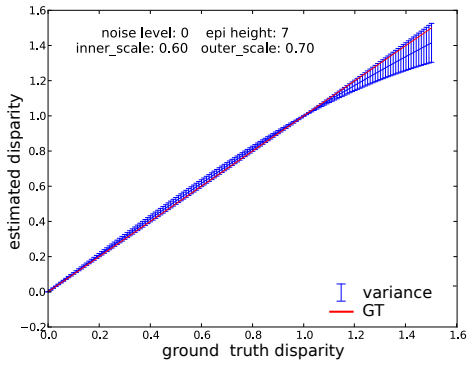

Figure 4: *The disparity estimation error of the disparity increases drastically for disparities more than 1 pixel. Plotted on the abscissa is the ground truth disparity and on the ordinate the estimated disparity.*

3.1. Refocusing

To solve this problem and get a higher reliability in the estimated disparity, we propose to refocus the EPIs, see Figure 3. The refocusing process references every EPI to a virtual depth layer

$$Z_i \in \{Z_1, ..., Z_N | Z_1 < ... < Z_N\}, \tag{4}$$

which is selected from a set of N predefined depth levels covering the scene.

To refocus onto the virtual depth Z_i, we need to translate the depth information to a disparity shift in every single EPI. For this, we select a reference view $(s_{ref}, t_{ref}) \in \Pi$, which is in this paper always assumed to be the center view. Some EPIs for different refocused depths are shown in Figure 5.

For the sake of simplicity, we also assume a symmetric setup in the sense that every camera has the same focal length f and the same baseline Δs with respect to the neighbouring views. The resulting disparity shift $\Delta x(s)$ related to the virtual depth Z_i is defined as

$$\Delta x(s) := (s_{ref} - s)\frac{\Delta s}{Z_i} f. \tag{5}$$

We now define the shifted EPI $\hat{S}^i_{t^*,y^*}$ transformed with respect to the virtual depth Z_i as

$$\hat{S}^i_{t^*,y^*} : \Sigma_{t^*,y^*} \to \mathbb{R} \tag{6}$$

$$(x,s) \mapsto \hat{S}^i_{t^*,y^*}(x,s) := L(s,t^*,x+\Delta x(s),y^*) \tag{7}$$

The refocusing on a different virtual depth Z_i influences the orientation of the structures in the EPI, see Figure 3. As one can see, some regions which had disparities outside of the two pixel range are now in vertical orientation, i.e. in focus. In contrast, previously vertically oriented regions now have orientations out of the two pixel range. The orientation

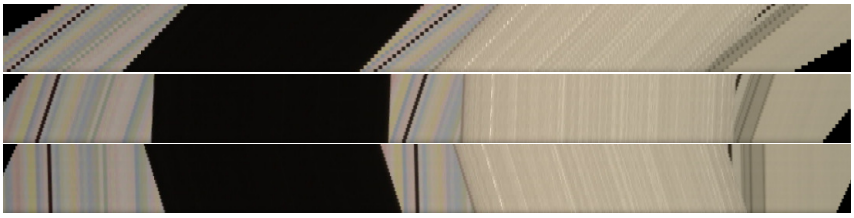

Figure 5: *Above is shown the refocusing process to three different virtual depths Z_1, Z_2, Z_3 with $Z_1 < Z_2 < Z_3$. The top image, focused on Z_1 has only orientations towards the left, no object is in focus. The middle image, focused on Z_2, has vertical lines for objects located at the selected depth. The bottom image is focused on Z_3. Object which where focused on Z_2 are now oriented towards the right and the background object, located at Z_3, is now in focus.*

of the new in focus regions can now be estimated more reliably than in not focused cases. It remains to be shown is how to superimpose the several orientation estimates from different focus planes using a reliability measure for every refocused EPI.

3.2. Superimposing to a global disparity map

Disparities are computed by analyzing the structure tensor $J_{\hat{S}_i}$ of a refocused EPI $\hat{S}^i_{t^*,y^*}$. It is defined as

$$J_i = \tau * \begin{pmatrix} \left(\frac{\partial \hat{S}_i}{\partial x}\right)^2 & \frac{\partial \hat{S}_i}{\partial x} \cdot \frac{\partial \hat{S}_i}{\partial s} \\ \frac{\partial \hat{S}_i}{\partial s} \cdot \frac{\partial \hat{S}_i}{\partial x} & \left(\frac{\partial \hat{S}_i}{\partial s}\right)^2 \end{pmatrix} =: \begin{pmatrix} J_{xx} & J_{xs} \\ J_{xs} & J_{ss} \end{pmatrix} \tag{8}$$

with the abbreviation

$$\hat{S}_i := \sigma * \hat{S}^i_{t^*,y^*}. \tag{9}$$

Above, σ is a Gaussian smoothing kernel called the inner kernel, which is applied to the EPI, while τ is another Gaussian smoothing kernel caller the outer kernel, which is applied component-wise to the derivative tensor.

The reconstructed local disparity map D_i can be computed from the structure tensors J_i using the formulas given in [WG13],

$$D_i = \sin\left(\arctan\left(\frac{J_{xs} - J_{xx}}{2J_{xs}}\right)\right). \tag{10}$$

As a reliability measure to assess the quality of the estimated disparities, we use the coherence of the structure tensor J_i as described in [BG87]

$$r_i := \sqrt{\frac{(J_{xx} - J_{xs})^2 + 4(J_{xs})^2}{(J_{xx} + J_{ss})^2}}. \tag{11}$$

This reliability measure defines how distinct the underlying structure is, and is a good estimate for the accuracy of the resulting disparity value.

The disparity map D_i constructed for this fixed EPI at coordinates t^*, y^* contributes a single line to the global disparity maps for each view s, t^*. A sensible way to construct the global disparity maps D_{s,t^*} is to superimpose all local disparity maps D_i based on the corresponding coherence maps, and select the estimate with highest reliability. In formulas,

$$D_{s,t^*}(x, y^*) = D_{I(x,s)}(x, s) \tag{12}$$

$$r_{s,t^*}(x, y^*) = r_{I(x,s)}(x, s), \tag{13}$$

where the index $I(x,s)$ of highest reliability at each EPI coordinate is given by

$$I(x,s) = \operatorname*{argmax}_i \{r_i(x,s)\}. \tag{14}$$

An example for a resulting global coherence map r_{s^*,t^*} for the center view can be observed in Figure 7. The resulting global disparity map D_{s^*,t^*} corresponding to the center view

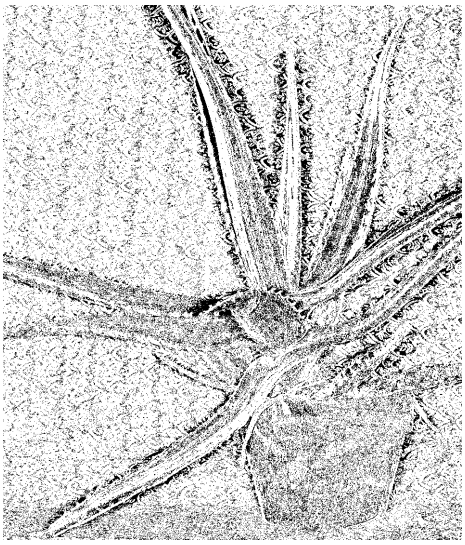

Figure 7: *The disparity estimation to the middle image seen in figure 6 reveals that at boundaries the estimation looks smoothed and noisy. This effects can also be seen in the corresponding coherence map. Estimations in noisy regions especially around boundaries have low coherence values shown as black spots.*

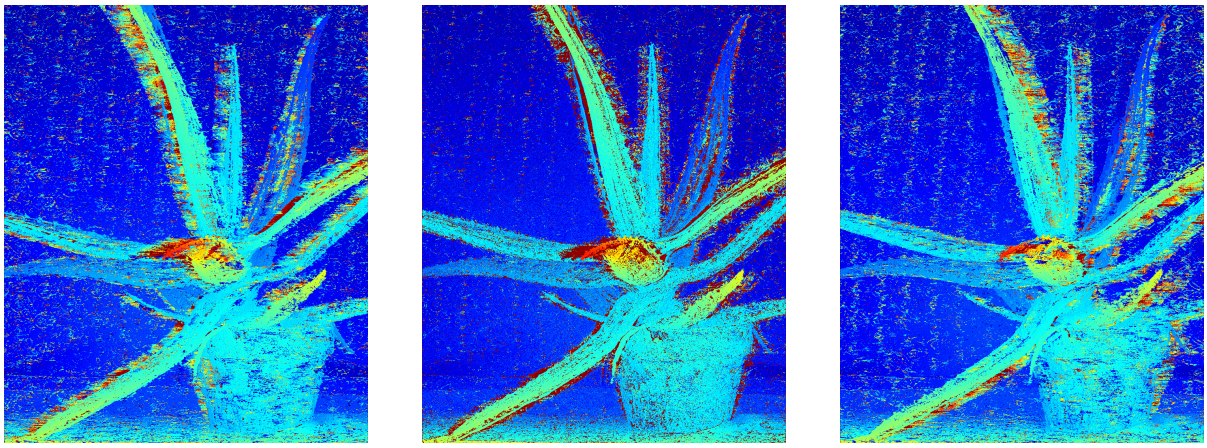

Figure 6: *Results for the Aloe Vera plant of the Middlebury dataset. This 3D light field contains seven images. Shown are from the left to the right the disparity maps of the second,fourth and sixth image. All disparity maps are equally aligned, but differs in the behavior at boundaries.*

is depicted in the center image of Figure 6. The results reveal an occlusion problem which mostly appears in images with a large baseline where objects are located close to the camera. These setups mostly result in disparities larger than 2 pixels. Transitions between foreground and background then appear smeared and noisy around the boundaries of objects in a range of the underlying object disparity between two neighboring views.

4. Boundary behavior in global disparity estimation

A larger disparity range in the EPI has negative impact on the accuracy of the boundaries because occlusion effects become stronger. Furthermore, larger smoothing kernels resulting in more robust reconstructions, but decrease the sharpness of the boundary transitions in the same time, see the center image of Figure 6. Here, the boundary between the object and background is not very clearly defined. However, the inaccuracy of the transitions can also be observed in the coherence map of the related view, see Figure 7. This correlation between the disparity estimation and the coherence map can be leveraged to handle the occlusion problem.

Refocusing on different virtual depth levels has the property that except for the reference view, the direct correspondence between cameras and disparity maps is lost. This effect occurs because objects close to the camera reach at first disparity estimations with high coherence values. With an restriction of the maximal allowed angle to less than 45 degree this values mostly belong to orientations close to zero. Thus, once the disparity values reaches a high coherence value it stays for the remaining global disparity estimation.
This effect is demonstrated in Figure 8, where two EPIs focused on different virtual depth are superimposed into a global EPI structure. The resulting global disparity maps for each view in a 3D light field show the same disparity maps just with different boundary behaviors.

In a 4D light field, this also happens with respect to the reference view, which usually is the center view of the 4D light field. From every view position around the reference view, we can see slightly behind objects located in the foreground. This makes it feasible to locate disparity transitions exactly on the correct object boundaries, as described in the following.

4.1. Occlusion handling by superimposing global disparity maps

The global disparity maps $D_{s,t}$ shown in Figure 6 depict the center view and two neighbor views. The location of the object in the disparity maps remains always the same as in the center view, however, the inaccuracies at the boundary transition differ substantially. The right image has sharp transitions on the left side of the boundary, while the image on the left has sharp transitions on the right side of the boundary.

The corresponding coherence maps fortunately again allow to distinguish good from bad disparity estimates, as coherence values are high exactly for those views where there is a sharp transition.

We therefore describe the final superposition $D(x,y)$ of the different global disparity maps as

$$D(x,y) = D_{s_{opt},t^*}(x,y) \tag{15}$$

with D_{s,t^*} and r_{s,t^*} defined in equation (12), and

$$s_{opt} = \underset{s}{\operatorname{argmax}}\{r_{s_{opt},t^*}(x,y)\}, \tag{16}$$

which again selects the estimate with the highest coherence from all candidate estimates.

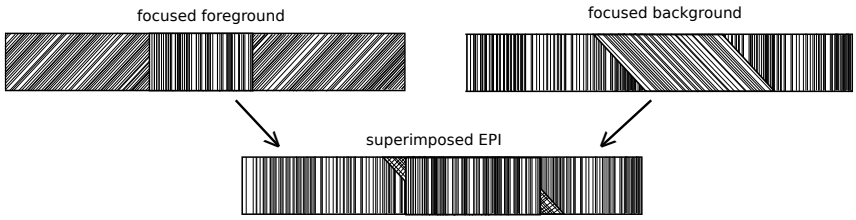

Figure 8: *Shown are two EPIs focused on different depth level. The right EPI shows structures focused on an object in the background. The left EPI shows structures focused on an object in the foreground. The superimposed EPI constructed by a a coherence based merge of both views, reveals for views beside the center view sharp boundary transition at respectively one side.*

For the sake of simplicity of equations, the superposition is only shown for 3D light fields. For 4D light fields, the computation can be adapted in a straight-forward manner by computing both EPI directions independently and then choosing the estimate with optimum coherence from both results.

5. Results

The proposed algorithm is tested on several different 3D Middlebury datasets, each containing seven images captured in a row with constant baseline. The Middlebury dataset also provides one 4D light field called "Tsukuba". We also evaluate our algorithm to this dataset to be able to compare the result with the provided ground truth data for the center view. The computed results to the 3D light fields of the Middlebury database can be seen in Figure 10. The image row on the left side show the center view images of the datasets we want to evaluate. All the shown results are obtained by refocusing to several virtual depths.

The occlusion problem is visible in the center column. Boundaries appear smeared and noisy.
In contrast, the right column shows the result we obtain with occlusion handling. The boundaries between fore- and background are sharper, less noisy and a reduction of the noise ratio is also visible. The light field evaluation shows similar results but has its general advantages in 4D light fields. The evaluation to the Tsukuba dataset is seen in Figure 9. The computed mean square error compared to the ground truth data is 2.45 pixel.
The solution of the middlebury dataset "Tsukuba" is been compared to the solution of a multi-view stereo algorithm. The used algorithm is a straight-forward stereo algorithm described in [WG13], where a local stereo matching cost function is computed for the single views. The cost function can then be used, integrated into a global energy functional, which is solved to global optimality using the method in [PCBC10].

6. Conclusion

We proposed a method to extend the feasible disparity range of an EPI-based disparity reconstruction method [WG12] by refocusing the light field to several different depth levels. In contrast to the previous approach, this makes it possible to evaluate light field setups with large baselines and objects placed closely in front of the camera. Another advantage we gain by refocusing is the possibility of handling occlusions by superimposing several views. Merging the neighboring views at occluded regions into the center view visibly sharpens the edge transitions and reduces the noise in the whole image. The results for 3D light fields show that disparity estimation works well even for only one dimensional view point changes and for large baselines.

References

[AB91] ADELSON E., BERGEN J.: The plenoptic function and the elements of early vision. *Computational models of visual processing 1* (1991). 1

[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. 1

[BG87] BIGÜN J., GRANLUND G. H.: Optimal orientation detection of linear symmetry. In *Proc. International Conference on Computer Vision* (1987), pp. 433–438. 4

[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. 1

[GFM*07] GALLUP D., FRAHM J.-M., MORDOHAI P., YANG Q., POLLEFEYS M.: al-time plane-sweeping stereo with multiple sweeping direction. In *CVPR* (2007). 2

[GGSC96] GORTLER S., GRZESZCZUK R., SZELISKI R., COHEN M.: The Lumigraph. In *Proc. SIGGRAPH* (1996), pp. 43–54. 1

[GL10] GEORGIEV T., LUMSDAINE A.: Focused plenoptic camera and rendering. *Journal of Electronic Imaging 19* (2010), 021106. 1

[HS07] HIRSCHMÜLLER H., SCHARSTEIN D.: Evaluation of Cost Functions for Stereo Matching. In *CVPR* (2007), IEEE Computer Society. 2

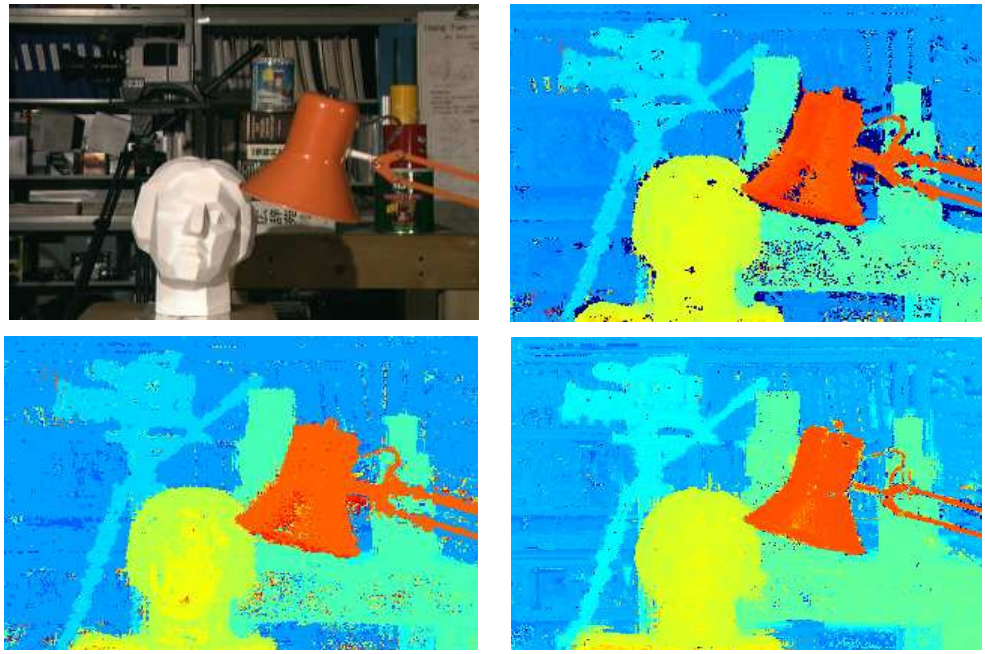

Figure 9: *Disparity estimation to the Middlebury 4D light field dataset "Tsubuka". The top left image is the center view image to which the estimated disparity maps correspond. The global disparity estimation is shown at the top right. The bottom left shows the global disparity with applied occlusion handling. As a reference, the bottom right shows the disparity estimate of a multi-view data term [WSG13].*

[KAC07] KUBOTA A., AIZAWA K., CHEN T.: Reconstructing Dense Light Field From Array of Multifocus Images for Novel View Synthesis. *IEEE Transactions on Image Processing 16*, 1 (2007), 269–279. 1

[LH96] LEVOY M., HANRAHAN P.: Light field rendering. In *Proc. SIGGRAPH* (1996), pp. 31–42. 1

[MB95] MCMILLAN L., BISHOP G.: Plenoptic modeling: An image-based rendering system. In *Proc. SIGGRAPH* (1995), pp. 39–46. 1

[NLB*05] NG R., LEVOY M., BRÉDIF M., DUVAL G., HOROWITZ M., HANRAHAN P.: *Light field photography with a hand-held plenoptic camera*. Tech. Rep. CSTR 2005-02, Stanford University, 2005. 1

[PCBC10] POCK T., CREMERS D., BISCHOF H., CHAMBOLLE A.: Global Solutions of Variational Models with Convex Regularization. *SIAM Journal on Imaging Sciences* (2010). 6

[PW10] PERWASS C., WIETZKE L.: The Next Generation of Photography, 2010. `www.raytrix.de`. 1

[SCK07] SHUM H., CHAN S., KANG S.: *Image-based rendering*. Springer-Verlag, New York, 2007. 1

[SP07] SCHARSTEIN D., PAL C.: Learning Conditional Random Fields for Stereo. In *CVPR* (2007), IEEE Computer Society. 2

[SS02] SCHARSTEIN D., SZELISKI R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision 47*, 1-3 (2002), 7–42. 2

[SS03] SCHARSTEIN D., SZELISKI R.: High-Accuracy Stereo Depth Maps Using Structured Light. In *CVPR* (2003), IEEE Computer Society, pp. 195–202. 2

[VGT*05] VAISH V., GARG G., TALVALA E.-V., ANTUNEZ E., WILBURN B., HOROWITZ M., LEVOY M.: Synthetic Aperture Focusing using a Shear-Warp Factorization of the Viewing Transform. In *CVPR* (2005), pp. 129–. 1

[VLS*06] VAISH V., LEVOY M., SZELISKI R., ZITNICK C. L., KANG S. B.: Reconstructing Occluded Surfaces Using Synthetic Apertures: Stereo, Focus and Robust Measures. In *CVPR* (2006), pp. 2331–2338. 1

[VWJL04] VAISH V., WILBURN B., JOSHI N., LEVOY M.: Using plane + parallax for calibrating dense camera arrays. In *Proc. International Conference on Computer Vision and Pattern Recognition* (2004). 1

[WG12] WANNER S., GOLDLUECKE B.: Globally consistent depth labeling of 4D light fields. In *Proc. International Conference on Computer Vision and Pattern Recognition* (2012), pp. 41–48. 1, 2, 6

[WG13] WANNER S., GOLDLUECKE B.: Variational Light Field Analysis for Disparity Estimation and Super-Resolution. *IEEE Transaction on Pattern Analysis and Maschine Intelligence* (2013). 4, 6

[WSG13] WANNER S., STRAEHLE C., GOLDLUECKE B.: Globally Consistent Multi-Label Assignment on the Ray Space of 4D Light Fields. In *Proc. International Conference on Computer Vision and Pattern Recognition* (2013). 7

[ZGFN08] ZACH C., GALLUP D., FRAHM J.-M., NIETHAMMER M.: Fast Global Labeling for Real-Time Stereo Using Multiple Plane Sweeps. In *Vision, Modeling and Visualization Workshop VMV 2008* (2008). 2

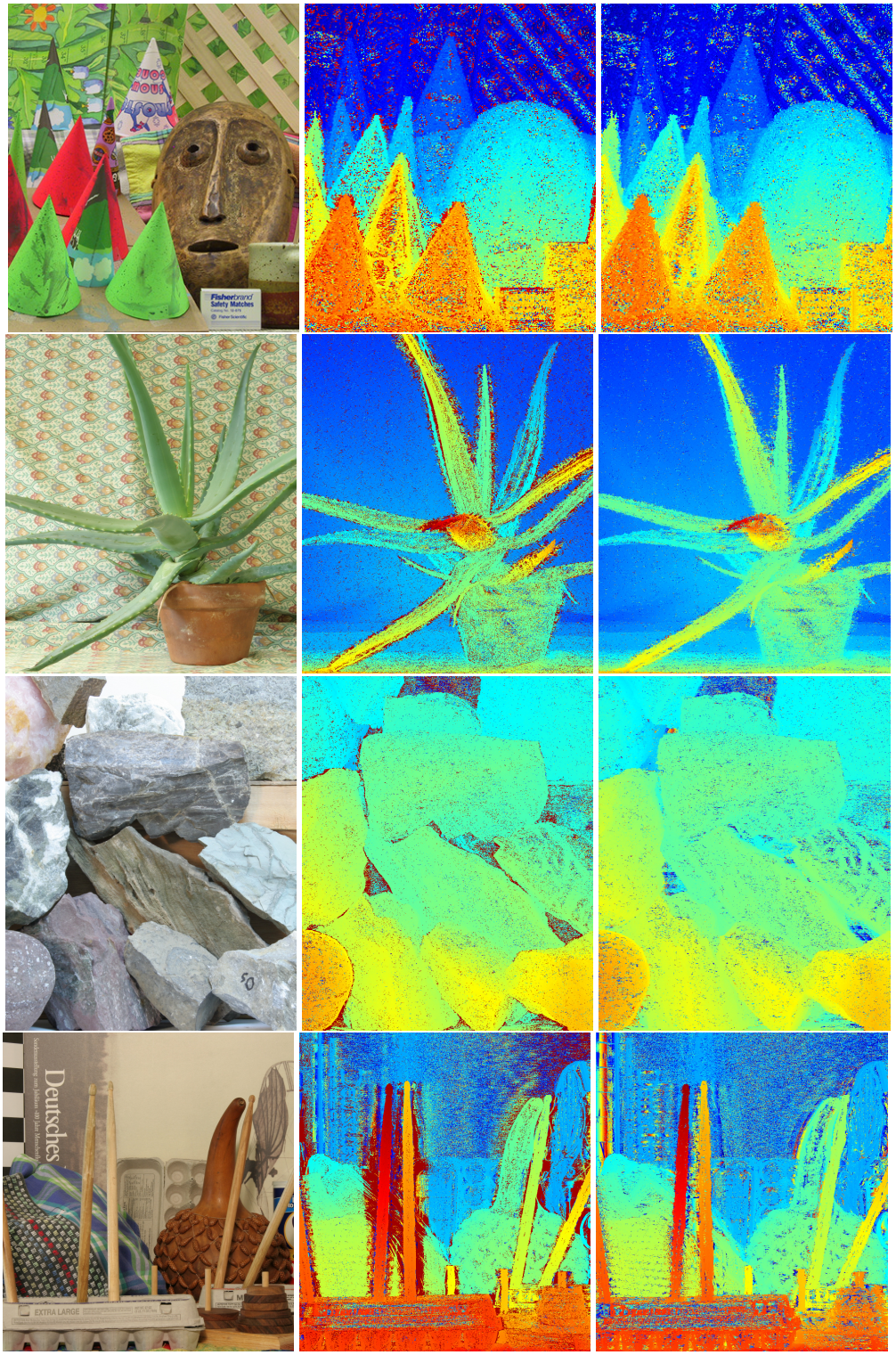

Figure 10: *Results for the Middlebury 3D light field datasets. The first column shows the original center view of the light field. The second column shows the global disparity maps computed by refocusing to several virtual depth layers. In the third column, we see the global disparity maps with occlusion handling.*