Fakultät für Physik und Astronomie

Ruprecht-Karls-Universität Heidelberg

Masterarbeit

Im Studiengang Physik

vorgelegt von

(Vor- und Zuname)

geboren in (Geburtsort)

(Jahr der Abgabe)

(Titel)

(der)

(Masterarbeit)

Die Masterarbeit wurde von (Vorname Name)

ausgeführt am

(Institut)

unter der Betreuung von

(Frau/Herrn Prof./Priv.-Doz. Vorname Name)

Department of Physics and Astronomy

University of Heidelberg

Master thesis

in Physics

submitted by

(name and surname)

born in (place of birth)

(year of submission)

(Title)

(of)

(Master thesis)

This Master thesis has been carried out by (Name Surname)

at the

(institute)

under the supervision of

(Frau/Herrn Prof./Priv.-Doz. Name Surname)

(Titel der Masterarbeit - deutsch):

(Abstract in Deutsch, max. 200 Worte. Beispiel: **?**)
Lorem ipsum dolor sit amet, consectetur adipisici elit, sed eiusmod tempor incidunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquid ex ea commodi consequat. Quis aute iure reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint obcaecat cupiditat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.
Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

(Title of Master thesis - english):

(abstract in english, at most 200 words. Example: **?**)
Lorem ipsum dolor sit amet, consectetur adipisici elit, sed eiusmod tempor incidunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquid ex ea commodi consequat. Quis aute iure reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint obcaecat cupiditat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.
Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Contents

1 Theory

1.1 Depth from focus

One advantage of using lightfields for depth measure is its ability to get a two-dimensional image of the scene at any depth. Integrating the views of the light field camera array has the same effect as the integration of a focussed lense camera, as the lense is simply integrating slightly different viewpoints of the same scene point when focussed on the correct depth.
Obtaining the refocussed integrated image is a synthetic process that only requires shifting the view coordinates artificially. Given a full four-dimensional light field $L(u, v, x, y)$ we can refocus the light field as described in Ng et al.:

$$L'(u, v, x, y) = L(u(1 - d'), v(1 - d'), x, y), \tag{1.1}$$

where d' describes the relative pixel shift. The disparity is directly related to the absolute depth of the focus (relate to PICTURE) if the relevant camera parameters are known. Given the baseline b in meters and the focal length f in pixels, the depth Z is given as

$$Z = \frac{f \cdot b}{d}. \tag{1.2}$$

We obtain

$$\bar{L}(x, y) = \frac{1}{N_{u,v}} \int \int L'(u, v, x, y) du dv = \frac{1}{N_{u,v}} \sum_u \sum_v L'(u, v, x, y) \tag{1.3}$$

Once we can focus at any range, one can adopt *depth-from-focus*-techniques as described in Watanabe and Nayar [1998] for depth measure. If the scene point at a given image coordinate (x, y) in the center view is in focus, the contrast in the integrated image $\bar{L}(x, y)$ is high, thus a contrast measure at each pixel combined with stepwise refocussing yields a depth map.
For measuring the contrast, one has different options: The most straight forward approach is calculating the first derivative of the grey-value image. At high contrast structure the local intensity changes are expected to be high. Alternatively one could measure the second derivative laplacian that eventually results in higher robustness. The implementation and tests of those techniques for the benchmark dataset can be found in section **??**.
Using a pinhole camera array allows us to go further and find a response value that shows higher consistency. Taking the absolute difference between the center view of

the camera array and the refocussed image yields to promising results as shown in Tao et al. [2017]. Under the assumption of lambertian surfaces the RGB- value of any scene point should be the same under all angles. Thus when refocussed on the correct depth, summing over all angles should result in a value that ideally is the same as in the center view alone. This is referred as *photo consistency*; for more information read Tao et al. [2017]. The response value at a given depth is obtained from

$$D'(x,y) = \frac{1}{|W_D|} \sum_{x',y' \in W_D} \left| \bar{L}(x',y') - P(x',y') \right|, \tag{1.4}$$

where $P(x,y)$ is the center view. For more robustness, it is averaged over a small window. We refer to this measuring technique as *photo consistency* in the following. Note that calculating the absolute results in a 1-channel-image while the input images are RGB-images.
Tao et al. propose another measure that they refer to as *angular correspondence*. It follows the same principle, but instead of integrating the refocussed lightfield followed by comparing it to the center view, they directly take the difference of each viewpoint to the center view and sum up those differences:

$$D'(x,y) = \frac{1}{N_{u,v}} \sum_u \sum_v \left| L'(u,v,x,y) - P(x,y) \right|. \tag{1.5}$$

We tested those methods against the common contrast measures mentioned above, the results are found in section results.

1.2 Semi-Global Matching

1.2.1 Semi - Global Matching for Stereo Vision

In contrast to Light field depth estimation techniques Stereo systems often suffer from mismatching pixels between the left and right images. Many attemps have been made to smoothen bad pixels, resulting in blurred edges or long calculation times. One promising attempt to imporove matching results was published in 2005 by Heiko Hirschmüller (Hirschmuller [2005]) that was described as „a very good trade off between runtime and accuracy “ (Hirschmüller [2011]): we speak of Semi-Global Matching.
In general, matching of two stereo images means shifting the disparity over the predefined disparity range and comparing both images (pixel- or blockwise) until we have a cost value at each image point for each discrete disparity. We assign to each pixel $\vec{p}$ the disparity value $D_{\vec{p}}$ which is related to the lowest cost $C(\vec{p}, D_{\vec{p}})$. This matching does not have to be unique, resulting in errorneous pixel disparities. To

overcome this one wants to minimize a global cost function of the form

$$E(D) = \sum_{\vec{p}} \left(C(\vec{p}, D_{\vec{p}}) + \sum_{q \in N_p} \begin{cases} P1 & \text{if } |D_{\vec{p}} - D_{\vec{q}}| = 1 \\ P2 & \text{if } |D_{\vec{p}} - D_{\vec{q}}| \geq 1 \\ 0 & \text{else} \end{cases} \right). \tag{1.6}$$

The fist term sums all matching costs over the whole image, while the second term forces continuity by comparing the disparity of all neighbour pixels N_q to the disparity D_p; if a small discontinuity is detected ($D_{\vec{p}} - D_{\vec{q}} = 1$), a small penalty is added to the global cost function. Since a small discontinuity can be found essentially at any tilted plane, only a small error is added. A bigger disparity difference indices a clear discontinuity in the disparity map. Note that the penalty $P2$ can be divided by the gradient of the original image to allow a disparity discontinuity when we find edges in the image; at these points we expect the disparity to be discontinuous.
However, minimizing the global cost function involves computational cumbersome algorithms as it is a NP-complete Problem (Hirschmüller [2011]). Semi-Global Matching however chooses another approach by minimizing the global cost function along one-dimensional lines – this can indeed be calculated in polynomial time. The new smoothed cost function $S(\vec{p}, D_{\vec{p}})$ at pixel $\vec{p}$ is then given as the sum of all 1D minimum cost paths that are ending in $\vec{p}$. The minimal cost L'_r along the path r is defined recursively as

$$L'_r(\vec{p}, D) = C(\vec{p}, D) + \min \begin{cases} L'_r(p_{\vec{\text{before}}}, D) \\ L'_r(p_{\vec{\text{before}}}, D+1) + P1 \\ L'_r(p_{\vec{\text{before}}}, D-1) + P1 \\ \min_i L'_r(p_{\vec{\text{before}}}, i) + P2 \end{cases} \tag{1.7}$$

By always adding the minimum path cost of the previous pixel on the scanline we are looking at, we solve equation 1.6 in one dimension. It is to mention that the rolling sum can reach quite high numbers that are unpleasant to handle on the computer; a normalization is implemented by substracting $\min_D L'_r(p_{\vec{\text{before}}}, D)$ from all pixel cost values $L'_r(\vec{p}, D)$. The position of the minimum cost function at pixel $\vec{p}$ is unaffected by that normalization.
Summing along at least 8 path directions (crosshair + diagonals) results in disparity maps with reduced error pixel while maintaining clean edges. Neither a blur filter, a median filter or a bilateral filter would preserve those features.

1.2.2 Semi - Global Matching for Light fields

Even though Hirschmüller describes Semi - Global Matching (SGM) as a complete algorithm to obtain a disparity map from a stereo image input, we further refer to SGM as the true novelty of his work: the implementation of an approximation to the global solution of the cost function (equation 1.6). Indepent from the method one uses to calculate a disparity map, one needs a cost function defined in disparity space

for each pixel to make use of SGM. Similar to the Stereo Matching depth estimation, the structure tensor depth estimation pipeline for Lightfield data sets produces a disparity map and a coherence value at each disparity shift. This implies, that the SGM algorithm can be adapted to improve the results of the structure tensor pipeline. However, there are some significant differences between those two methods:

1. The structure tensor algorithm is tuned to a much smaller disparity range. While in Hirschmuller [2005] Hirschmüller scans a disparity range of 32 pixels , The benchmark data sets for light fields mostly include close-up views of objects, with a disparity range between 2 and 10 pixels. In figure 1.1 one can see the different values that are allocated in memory for each pixel of the image.

2. The subpixel accuracy using the structure tensor is a lot higher than the stereo matching subpixel accuracy. A simple adaption of the algorithm to the structure tensor pipeline would require to give up the best feature that is provided by the ST, its subpixel accuracy.

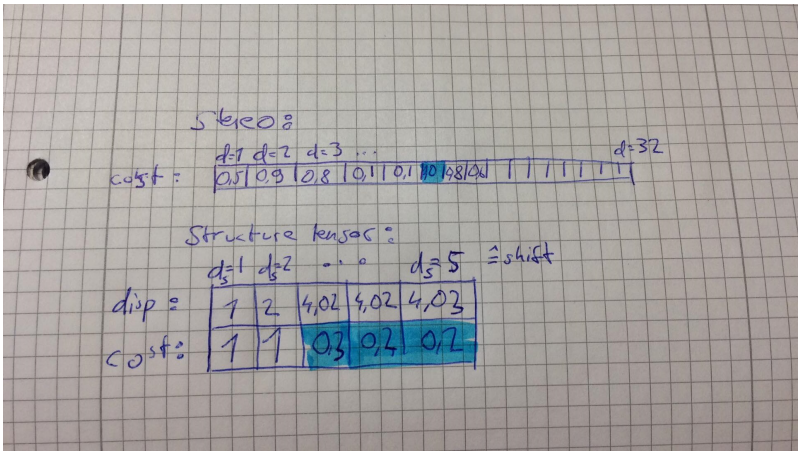

Figure 1.1: One point $\vec{p}$ contains different values (values here: example values): For Stereo matching, the resolution is given by the discrete disparity steps. Each disparity value has a cost value assigned to it. Using the ST, we have a different subpixel accuracy for every disparity shift, while the subpixel accuracy can differ from the shift by up to 1.2

To handle those problems, we do not throw away the subpixel accuracy: instead we use the float-value disparities to decide whether we penalize a disparity discontinuity or not. As one can see in figure 1.1, we have to process an additional information, since the exact disparity value is not implicitely given by the index of the allocated cost value (in contrast to the original algorithm). Switching to a continuous space as depicted in figure 1.2 requires a new definition of the error propagation defined in equation 1.6.

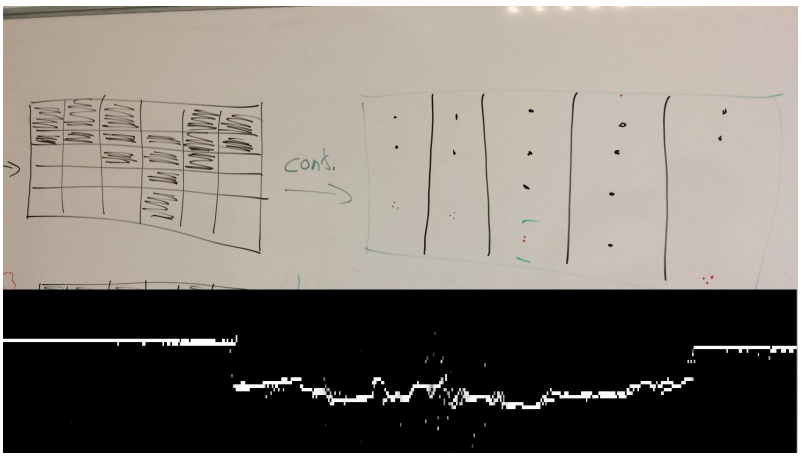

Figure 1.2: In this figure one can see the skizze of one arbitrary scanline of the structure tensor method. Under the Assumption that the ST algorithm recognizes the structure of the EPI perfectly, we either have two or three disparity shifts that have a high coherence (colored white) (a) and result in approximately the same final disparity. This can be seen if we plot the exact disparity values in a continuous space(b). In (c) real data scanline cost is plotted with a resolution of ca. 100 pixels.

In the following we refer to s as the disparity shift in the ST algorithm. Note that we replaced D by d to clarify that the disparity is no longer discrete:

$$E(d) = \sum_{\vec{p}} \left(C(\vec{p}, d_{\vec{p}}) + \sum_{q \in N_p} \begin{cases} P1 \cdot |d_{\vec{p}} - d_{\vec{q}}| & \text{if } |d_{\vec{p}} - d_{\vec{q}}| \leq 1 \\ P2 & \text{if } |d_{\vec{p}} - d_{\vec{q}}| > 1 \end{cases} \right). \tag{1.8}$$

The recursive 1-d form to solve the global constraint on a scanline then changes to:

$$L'_r(\vec{p}, s) = C(\vec{p}, s) + \min_i \begin{cases} L'_r(\vec{p}_{\text{before}}, s_i) + P1 \cdot |d_{\vec{p},s} - d_{\vec{p}_{\text{before}},s_i}| & \text{if } |d_{\vec{p},s} - d_{\vec{p}_{\text{before}},s_i}| \leq 1 \\ L'_r(\vec{p}_{\text{before}}, s_i) + P2 & \text{if } |d_{\vec{p},s} - d_{\vec{p}_{\text{before}},s_i}| > 1 \end{cases} \tag{1.9}$$

The biggest difference lies in the fact that the small factor that is smoothing the image linearely increases with the distance. This change is necessary under the assumption that the disparity space is continuous. In other words we cluster disparity differences between two neighbouring points as either part of one surface ($|d_{\vec{p},s} - d_{\vec{p}_{\text{before}},s_i}| \leq 1$) that gets smoothed by the linearly increasing penalty, or assume a real disparity discontinuity that is penalized regardless of the size of the jump - the second error remains constant as it is in Stereo matching. Note that in

our implementation, $P2$ is modified by

$$P2' = \frac{P2}{\sqrt{(Im_b^2 + Im_r^2 + Im_g^2)}}, \tag{1.10}$$

with Im being the center view of the lightfield and $Im_{b,g,r}$ being the 3 color channels. If the color intensity changes, the penalty for a disparity discontinuity is lowered.

1.2.3 Occlusion awareness in SGM for light Fields

Having a look at the results of the the benchmark test of Honauer et al. [2016] one realizes that most Light field depth estimation algorithms suffer from large error near depth discontinuities. Since the center view pixels close to the edge of a depth discontinuity are at least partly occluded, this behaviour is to be expected. The ST almost always produces a systematic error near discontinuities, leading to a „magnification "of the object closer to the camera in the depth map, see figure 1.3. The reason for this error has its origin in the smoothing of the EPI as part of the algorithm. In figure (INPUT EPI) one can identify that after smoothing , the structure tensor near a occlusion edge calculates the structure of the object in the foreground, since it is

- not occluded at all, providing his structure over the whole Angular spectrum
- under better illumination in most cases
- provided with higher structure resolution.

If we take a close look at figure 1.2, one can see that at least at some discontinuities the ST pipeline manages to calculate the depth of the background structure near boundaries with good coherence, but the foreground structure is overlapping and quantitatively measured with higher coherence. Once we know that at least at some edges an improvement can be made by adapting the evaluation function in a sense that the highest coherence does not necessarily measure the right depth, we realize that SGM is doing the job already. The simple heuristic approach is to change the global minimization function 1.8 such that a positive disparity jump is less punished than a negative one. In fact, the function changes to

$$E(d) = \sum_{\vec{p}} \left(C(\vec{p}, d_{\vec{p}}) + \sum_{q \in N_p} \begin{cases} P1 \cdot |d_{\vec{p}} - d_{\vec{q}}| & \text{if } |d_{\vec{p}} - d_{\vec{q}}| \leq 1 \\ P2 & \text{if } d_{\vec{p}} - d_{\vec{q}} > 1 \\ P3 & \text{if } d_{\vec{p}} - d_{\vec{q}} < -1 \end{cases} \right). \tag{1.11}$$

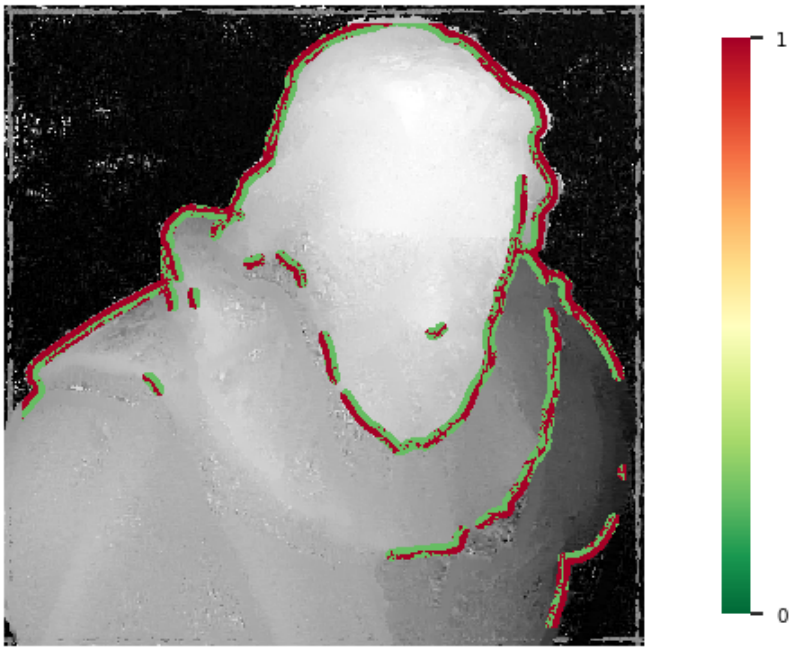

Figure 1.3: Evaluation of the Error at the depth discontinuity for scene "cotton". The red border indicates that the depth map is errorneous at the outside of the edge.

2 results

2.1 Depth from focus

The depth measure using epipolar plane analysis requires iterative calculation of the structure tensor for each EPI at each disparity. A way to overcome this is to generate a preestimate of the depth before actually calculating the correct depth. This could also help to prevent possible errors due to periodic scene characteristics which can lead to mismatch errors when calculating the structure tensor. Therefore the depth pre-estimate should fulfil the following criteria:

1. It should be *consistent*, meaning that the number of pixels with low confidence should be the lowest possible.
2. It should result in a *fast* measure, ideally faster then it would take to do the full iterative structure tensor algorithm.
3. It does not have to be subpixel accurate, since it only serves as a pre-estimate.

The methods that are tested are described in section 1.1. We test four different ways to obtain a depth map using depth from focus:

Photo consistency This measure takes advantage of the fact that the difference between the refocussed two-dimensional image and the center view is close to zero when refocussed to the correct depth. Response value:

$$D'(x,y) = \frac{1}{|W_D|} \sum_{x',y' \in W_D} \left| \bar{L}(x',y') - P(x',y') \right|, \tag{2.1}$$

Angular correspondence In contrast to the *Photo consistency* - measure, it first calculates the absolute difference between each camera array view and the center view followed by the summation of those deviations. The response value is given as in equation (1.5)

$$D'(x,y) = \frac{1}{N_{u,v}} \sum_u \sum_v |L'(u,v,x,y) - P(x,y)| \tag{2.2}$$

First derivative The first derivative is calculated for contrast measure by applying the sobel filter onto the refocussed image I:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \cdot I \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \cdot I \tag{2.3}$$

The directional gradients are simply added up to the response value

$$D'(x,y) = |G_x(x,y)| + |G_y(x,y)| \tag{2.4}$$

Laplace Here we calculate the second derivative laplacian by appling the sobel operator twice:

$$D'(x,y) = \text{Laplace}(I)(x,y) = \frac{\partial^2 I}{\partial x^2}(x,y) + \frac{\partial^2 I}{\partial y^2}(x,y) \tag{2.5}$$

In the following we are going to compare the method qualitatively and quantitatively. In figure 2.1 one can see the pixel response value refocussed at different disparities for all the methods at example points in the testscene „complextestscene“. We chose points close to edges as well as points on clear surface with less structure on it. One can see that the pixel response of the Angular correspondence and the Photoconsistency method for those points show a more consistent behaviour, meaning that only one clear maximum can be seen. The derivative method as well as the Laplace method both seem to have trouble especially on pixel coordinates close to edges. This can be seen most clearly at the edge of the melon, where The first derivative shows 2 Maxima, while the Angular correspondence and Photo consistency measure find a clean maximum indicating at which depth the point can be found. However, on surfaces with less structure as for example on the potato, the Photo consistency measure struggles to find a clear maximum – still it has one at the right position. Having a look at the actual disparity maps produced by the different techniques we can already capture that the angular correspondence and photo consistency method produce the more consistent, smooth disparity maps. A quantitative evaluation confirms this impression: We measure the mean relative error (MRE), which is defined as the mean squared error over all pixels $\vec{p}$ divided by the maximum disparity range (the maximum error possible).

$$MRE = \sum_{\vec{p}} (d_{\text{ground truth}} - d_{\vec{p}})^2 / (\text{max. disp - min. disp}) \tag{2.6}$$

2.1.1 Using depth-from-refocus as a preestimate for the ST

The long-term aim of this work is to make the structure tensor pipeline more robust. We achieve this by using the depth-from-refocus method as a pre

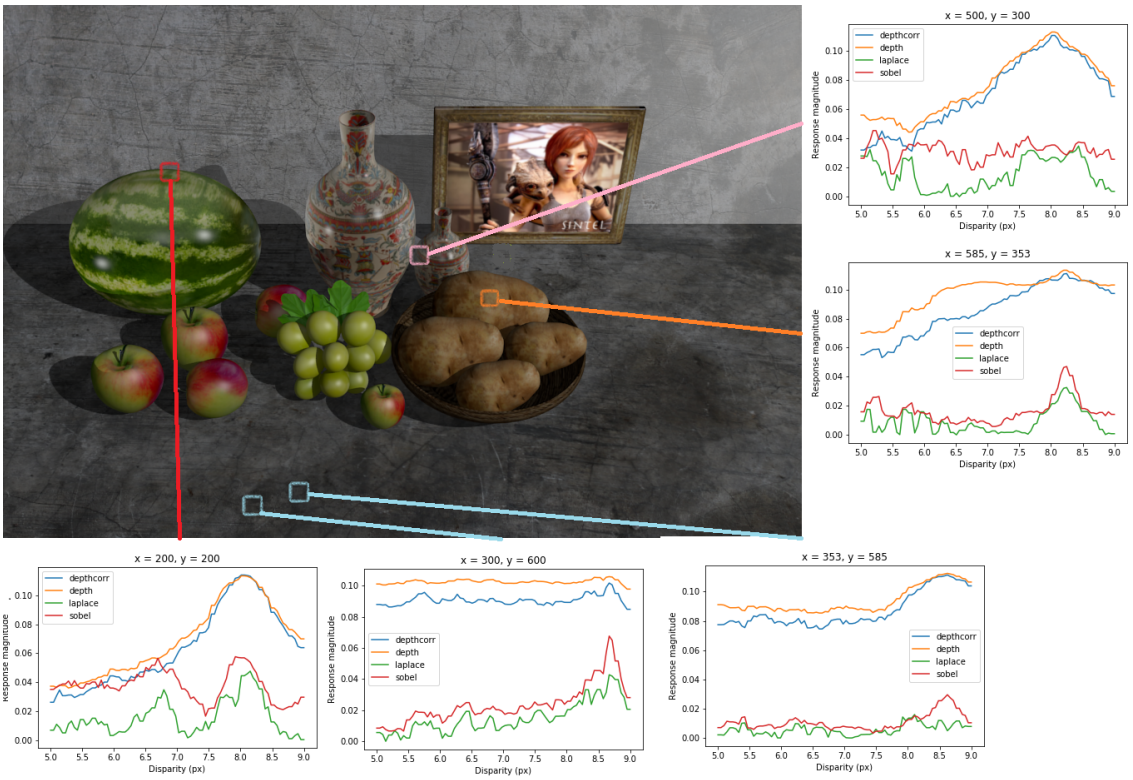

Figure 2.1: At different pixel positions we take a look on how the Pixel response value behaves for the four different depth-from-focus techniques: Photo consistency (orange), Angular correspondence (blue), First derivative (red) and The Laplace method (green) A high value means high confidence (low cost).

Part I

Appendix

A Lists

A.1 List of Figures

A.2 List of Tables

B Bibliography

Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 807–814. IEEE, 2005.

Heiko Hirschmüller. Semi-global matching-motivation, developments and applications. *Photogrammetric Week 11*, pages 173–184, 2011.

Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*. Springer, 2016.

Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera.

Michael W Tao, Pratul P Srinivasan, Sunil Hadap, Szymon Rusinkiewicz, Jitendra Malik, and Ravi Ramamoorthi. Shape estimation from shading, defocus, and correspondence using light-field angular coherence. *IEEE transactions on pattern analysis and machine intelligence*, 39(3):546–560, 2017.

Masahiro Watanabe and Shree K Nayar. Rational filters for passive depth from defocus. *International Journal of Computer Vision*, 27(3):203–225, 1998.

Erklärung:

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den (Datum) .....