Establishing a Probabilistic Depth Map from Focused Plenoptic Cameras

Niclas Zeller Franz Quint
Karlsruhe University of Applied Sciences
Karlsruhe, Germany
{niclas.zeller,franz.quint}@hs-karlsruhe.de

Uwe Stilla
Technische Universität München
Munich, Germany
stilla@tum.de

Abstract

In this paper we propose a novel method for depth estimation based on a single recording of a focused plenoptic camera. The presented algorithm is based on multiple stereo-observations within the multi-view micro images of the focused plenoptic camera. Here, pixel correspondences are found based on local intensity error minimization. Since our algorithm works directly on the micro images, no sub-aperture or epipolar plane images have to be synthesized. Due to the fact that we perform stereo matching based on local criteria we only estimate depth for pixels with sufficient gradient. Thus, we reduce the complexity of the problem, while neglecting uncertain stereo correspondences. Our algorithm incorporates multiple stereo-observations of the same point in a probabilistic depth map. We will show, that this (inverse) depth map can be modeled as a map of Gaussian distributed random variables. Thus, each depth pixel consists of an estimated depth and a corresponding variance, which gives a measure for the uncertainty of the estimation. This uncertainty information can be used in subsequent filtering methods.

1. Introduction

Plenoptic cameras capture the light-field of a scene as a 4D function [1, 6]. Thus, a plenoptic camera gathers much more information about the recorded scene than a traditional monocular camera.

Due to the fact, that an image point is not only represented by a single, but by multiple sensor pixels (or light rays), for instance, the image distance and thus also the object distance of a point can be estimated from one single shot of a plenoptic camera. Hence, a plenoptic camera offers a passive depth sensor alternatively to a binocular stereo camera for example.

One big advantage of a plenoptic camera compared to a binocular stereo camera system are the small dimensions in which it can be realized. A plenoptic camera basically has the same dimensions as a traditional monocular camera.

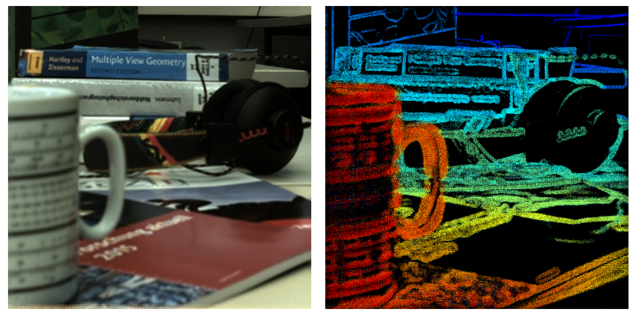


Figure 1. Virtual depth map calculated from a single shot of a focused plenoptic camera. Left: RGB-image calculated by the Raytrix software. Right: Color-coded virtual depth map calculated based on our probabilistic approach.

The only trade-off is, that for the 4D light-field information is payed by less image resolution [5].

In this paper we present a novel approach for estimating depth from a single recording of a focused plenoptic camera [11, 14].

1.1. Related Work

For the last years various algorithms for depth estimation based on the recordings of plenoptic cameras or other light-field representations were published. First methods were published even more than 20 years ago [1].

Since light-field based depth estimation represents a multi-dimensional optimization problem, always a trade-off between low complexity and high accuracy or consistency has to be chosen. In [19, 18] for instance a globally consistent depth labeling is presented which is performed on the epipolar plane images (EPIs) of the 4D light-field and results in a dense depth map. In [8] the phase-shift theorem of the Fourier transform is used to calculate a dense disparity map with sub-pixel accuracy, while in [7] principal component analysis is used to find the optimum depth map. Other methods make use of geometric structures, like 3D line segments [20], to improve the estimate and to reduce the complexity. In [17] the use of a so called scale-depth space is presented, which provides a coarse depth map for uniform regions and a fine one for textured regions. Other methods reduce complexity by the use of local instead of global constrains and thus result in a sparse depth map. This sparse map supplies depth only for textured regions [2]. The methods presented in [16] and [9] additionally make used of the focus cues which are supplied by a plenoptic camera.

In our approach we are mostly interested in low complexity and real-time applicability. Therefor we were inspired by a monocular camera based multi-view stereo approach presented in [3].

1.2. Outline of Our Work

The contribution of this work is a new, focused plenoptic camera based depth estimation algorithm which establishes a semi-dense depth map. Therefor no sub-aperture images or EPIs are calculated beforehand. Our approach incorporates multiple stereo-observations of a point in a probabilistic depth estimate, similar to [3] where this idea has been used to gain depth in a monocular visual odometry approach. Beside the estimated depth, a measure of uncertainty is supplied for each pixel. Based on this additional information selective filters and outlier removals can be applied. Using the measure of uncertainty, after establishing the depth map it can be easily chosen between a very dense but less reliable or a more reliable but sparser depth map.

2. The Focused Plenoptic Camera

This section describes the concept of a focused plenoptic camera as it was presented for the first time in [11]. This concept differs slightly from the one of a traditional plenoptic camera [13, 12].

A focused plenoptic camera can be realized in two different configurations which are often referred to as Keplerian configuration and Galilean configuration.

In the Keplerian configuration [10, 11] a micro lens array (MLA) and the sensor are placed behind the focused image which is created by the main lens. Here, the focal length of the micro lenses is chosen such that multiple focused sub-images (micro images) of the main lens image occur on the image sensor.

In the Galilean configuration [10, 11] MLA and sensor are placed in front of the focused image which would be created by the main lens behind the sensor (Fig. 2). Subsequently we will call this image behind the sensor the virtual image. Similarly to the Keplerian configuration, the focal length of the micro lenses is chosen such that multiple sub-images of the virtual main lens image occur focused on the image sensor.

A Raytrix camera [14] is a focused plenoptic camera based on the Galilean configuration. While a plenoptic camera has already a larger depth of field (DOF) than a monocular camera at the same main lens aperture [4, 14], in a Raytrix camera the DOF is further increased by using

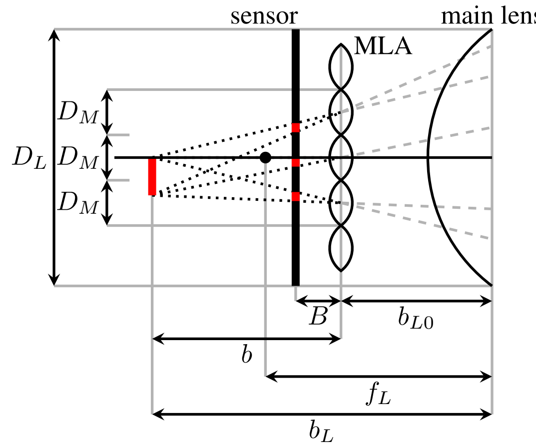


Figure 2. Optical path inside a focused plenoptic camera based on the Galilean configuration. The MLA and the image sensor lay in front of the virtual image created by the main lens. A virtual image point in distance b behind the MLA results in multiple focused micro images on the sensor.

an interlaced MLA in a hexagonal arrangement (see Figure 3). This MLA consists of three different micro lens types, where each type has a different focal length and thus focuses a different virtual image distance on the sensor. The DOFs of the three micro lens types are chosen such that they are just adjacent to each other. Thus, the effective DOF of the camera is increased by a factor of three compared to an MLA with only one type of micro lenses.

In the following we will only discuss a focused plenoptic camera which relies on the Galilean configuration. Nevertheless, for the Keplerian configuration similar relations can be derived.

If we consider the main lens to be an ideal thin lens, the relationship between the object distance a_L of an object point and the image distance b_L of the corresponding image point is defined by the thin lens equation, given in eq. (1). For a thick lens only slight changes have to be made in this equation.

$$\frac{1}{f_L} = \frac{1}{a_L} + \frac{1}{b_L} \tag{1}$$

Here f_L is the main lens focal length. Thus, if the image distance b_L of an image point is known, the object distance a_L of the corresponding object point can be calculated based on the lens equation.

As one can see from Figure 2, a virtual image point in distance b_L behind the main lens is projected to multiple micro images on the sensor. If we now consider the micro images as ideal central perspective images, the distance b between MLA and a virtual image point can be calculated by triangulation, as derived in [21]. Thereby it follows that

Figure 3. Section of the micro lens images (raw image) of a Raytrix camera. It shows on the left a part of the handle of the white cup and on the right a part of the headphone frame (see Fig. 1).

the distance b is calculated as given in eq. (2).

$$b = \frac{d \cdot B}{p_x} \tag{2}$$

Here, p_x is the disparity between the corresponding points in the micro images, d the baseline distance between the micro lenses used for triangulation and B the distance between MLA and image sensor. If we consider two adjacent micro images, the stereo baseline d is just the diameter of a micro lens D_M, as defined in Figure 2. For further apart micro images the baseline distance is a multiple of that diameter $d = k \cdot D_M$ ($k \geq 1$). Since d defines the euclidean distance between any two micro image centers, k is not mandatory an integer. This is the case for any regular tessellation.

The disparity p_x and the baseline distance d are both defined in pixels and can be measured from the recorded raw image, while the distance B between MLA and sensor is a metric dimension which can not be measured precisely. Thus, the distance b is estimated relative to the distance B. This relative distance, which is free of any unit is called virtual depth [14] and will be denoted by v in the following.

$$v = \frac{b}{B} = \frac{d}{p_x} \tag{3}$$

From Figure 2 one can see that virtual image points which have a large virtual depth occur in more micro images than points with a small virtual depth. Thus, one can make use of the larger baseline distance d between micro lenses which are further apart and thus improve the virtual depth estimate.

3. Virtual Depth Map Estimation

The estimation of the virtual depth v can be considered as a multi-view stereo problem since each virtual image point occurs in multiple micro images. Besides, the problem simplifies since all micro lenses have the same orientation by construction and thus, the micro images are already rectified. Due to the small dimensions of the micro images with respect to the pixel pitch (about 23 pixel in diameter), distortions of the micro images are negligible. Nevertheless, if there occurred significant distortion in the micro images a prior MLA calibration and rectification would have to be performed.

One approach to solve such a multi-view stereo problem would be to find correspondences between multiple micro images and then solve for the virtual depth. However, therefor pixel correspondences with sub-pixel accuracy have to be found across multiple micro images. Due to the very small micro images, feature extraction and matching seems to be quite difficult. Thus, we follow a different approach which is based on multiple depth observation received from different micro image pairs. Instead of feature matching we determine pixel correspondences by intensity error minimization along the epipolar line. For each depth observation an uncertainty measure is defined and thus, a probabilistic virtual depth map is established similar to the depth map in [3] where it is used to gain depth in a monocular visual odometry approach.

3.1. Probabilistic Virtual Depth

We define the inverse virtual depth $z = v^{-1}$, which is obtained from eq. (3). The inverse virtual depth z is proportional to the estimated disparity p_x, as given in eq. (4).

$$z = \frac{1}{v} = \frac{p_x}{d} \tag{4}$$

Since we determine pixel correspondences by matching pixel intensities, we consider the sensor noise to be the main error source which effects the disparity estimation and thus the inverse virtual depth z. Thereby, we neglect for instance misalignment of the MLA with respect to the image sensor or offsets on the micro lens centers. Furthermore, as one can see from eq. (4), the estimate of z relies only on the baseline distance d and the disparity p_x which result both as differences of absolute 2D positions in pixel coordinates. Thus, at least within a local region, the estimate of z is invariant of alignment errors on the MLA.

The sensor noise is usually modeled as additive white Gaussian noise (AWGN). Since pixel correspondences are estimated based on intensity values, the disparity p_x and thus the estimated inverse virtual depth z can also be considered as Gaussian distributed. This projection will be derived mathematically in Section 3.3.2.

In the following we will denote the inverse virtual depth hypothesis of a pixel by the random variable $Z \sim \mathcal{N}(z, \sigma_z^2)$ defined by the distribution function $f_Z(x)$ as given in eq. (5).

$$f_Z(x) = \frac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(x-z)^2}{2\sigma_z^2}} \tag{5}$$

Since the random variable Z is Gaussian distributed, it is completely defined by its mean z and its variance σ_z^2.

3.2. Graph of Baselines

For stereo matching we define a graph of baselines. This graph defines which micro images are matched to

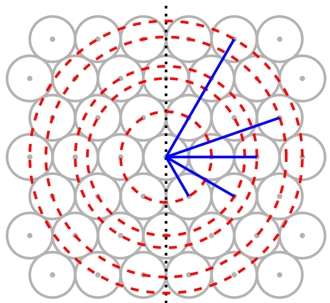


Figure 4. Five shortest baseline distances in a hexagonal micro lens grid. For a micro lens stereo matching is only performed with neighbors for which the baseline angle ϕ is in the range $-90° \leq \phi < 90°$.

each other. Each baseline in the graph is defined by its length d as well as its 2D orientation on the MLA plane $\boldsymbol{e}_p = (e_{px}, e_{py})^T$. Since we consider the micro images to be rectified, the orientation vector of the baseline is equivalent to that of the epipolar line. Thus, $\boldsymbol{e}_p$ defines the epipolar line for each pixel of the micro lens pair. In the following we will always consider $\boldsymbol{e}_p$ to be normed to unity ($\|\boldsymbol{e}_p\| = 1$ pixel).

In the graph the baselines are sorted in ascending order with respect to their length. This is also the order in which stereo matching will be performed. Matching is performed in that order since for short baselines it is more likely to find a unique match, while a long baseline improves the accuracy of the estimation but is also more likely to result in ambiguous matches. Thus, the matching result for short baselines can be used as prior knowledge for micro image pairs which are connected by a longer baseline.

Stereo matching is performed for each pixel on the raw sensor image separately. Here, we want to assure that corresponding pixels in different micro images are only matched once to each other. Thus, it is sufficient to perform matching only with respect to micro images right to the reference micro image. All micro images to the left will establish correspondence when they are considered as reference. Thus only baselines or epipolar lines with an angle $-90° \leq \phi < 90°$ are considered.

Figure 4 shows the five shortest baseline distances in a hexagonal MLA grid. Here the red dashed circles represent the respective baseline distance around the micro lens of interest. The solid blue lines show one example baseline for each distance, while only baselines right of the dotted line are used for stereo matching. The epipolar line $\boldsymbol{e}_p$ is defined such that it points away from the centered micro lens.

3.3. Virtual Depth Observation

The inverse virtual depth estimation is performed for each pixel $\boldsymbol{x}_R = (x_R, y_R)^T$ in the raw (sensor) image $I(\boldsymbol{x}_R)$.

As already mentioned, the depth observation is performed starting from the shortest baseline up to the largest possible baseline. Based on each new observation, the inverse depth hypothesis of a raw image pixel $Z(\boldsymbol{x}_R)$ is updated and thus becomes more reliable.

To reduce computational effort, for each baseline it is checked first, if the pixel under consideration $\boldsymbol{x}_R$ has sufficient contrast along the epipolar line, as defined in eq. (6).

$$|\boldsymbol{g}_I(\boldsymbol{x}_R)^T \boldsymbol{e}_p| \geq T_H \tag{6}$$

Here $\boldsymbol{g}_I(\boldsymbol{x}_R)$ represents the intensity gradient vector at the coordinate $\boldsymbol{x}_R$ (eq. (7)) and T_H some predefined threshold.

$$\boldsymbol{g}_I(\boldsymbol{x}_R) = \boldsymbol{g}_I(x_R, y_R) = \begin{pmatrix} \frac{\partial I(x_R, y_R)}{\partial x_R} & \frac{\partial I(x_R, y_R)}{\partial y_R} \end{pmatrix}^T \tag{7}$$

3.3.1 Stereo Matching

To find the pixel in a certain micro image which corresponds to our pixel of interest $\boldsymbol{x}_R$ we search for the minimum intensity error along the epipolar line in the corresponding micro image.

If there was no inverse virtual depth observation obtained yet for the pixel of interest $\boldsymbol{x}_R$, an exhaustive search along the epipolar line has to be performed. For that case the search range is limited on one end by the micro lens border and on the other end by the coordinates of $\boldsymbol{x}_R$ with respect to the micro lens center. A pixel on the micro lens border results in the maximum observable disparity p_x and thus in the minimum observable virtual depth v, while a pixel at the same coordinates as the pixel of interest in the corresponding micro image equals a disparity $p_x = 0$ and thus a virtual depth $v = \infty$.

If there exists already an inverse virtual depth hypothesis $Z(\boldsymbol{x}_R)$, the search range can be limited to $z(\boldsymbol{x}_R) \pm n\sigma_z(\boldsymbol{x}_R)$, where n is usually chosen to be $n = 2$.

$$Z(\boldsymbol{x}_R) \sim \mathcal{N}(z(\boldsymbol{x}_R), \sigma_z^2(\boldsymbol{x}_R)) \tag{8}$$

In the following we define the search range along the epipolar line as given in eq. (9)

$$\boldsymbol{x}_R^s(p_x) = \boldsymbol{x}_{R0}^s + p_x \cdot \boldsymbol{e}_p \tag{9}$$

Here $\boldsymbol{x}_{R0}^s$ is defined as the coordinate of a point on the epipolar line at the disparity $p_x = 0$, as given in eq. (10).

$$\boldsymbol{x}_{R0}^s = \boldsymbol{x}_R + d \cdot \boldsymbol{e}_p \tag{10}$$

Within the search range we calculate the sum of the squared intensity error e_{ISS} over a 1-dimensional pixel patch ($1 \times N$) along the epipolar line, as defined in eq. (11).

$$e_{ISS}(p_x) = \sum_{k=-\frac{N-1}{2}}^{\frac{N-1}{2}} [I(\boldsymbol{x}_R + k\boldsymbol{e}_p) - I(\boldsymbol{x}_R^s(p_x) + k\boldsymbol{e}_p)]^2 \tag{11}$$

The best match is the disparity p_x which minimizes $e_{ISS}(p_x)$. For the experiments presented in Section 4 we set $N = 5$. In the following we refer to the estimated disparity by $\hat{p}_x$, which defines the corresponding pixel coordinate $\boldsymbol{x}_R^s(\hat{p}_x)$.

3.3.2 Observation Uncertainty

As described before, the sensor noise n_I is the main error source which effects the estimated disparity $\hat{p}_x$ and thus the inverse virtual depth observation.

While the variance of the sensor noise σ_N^2 can be considered to be the same for each pixel $\boldsymbol{x}_R$, it effects the disparity estimation differently. This effect can be derived mathematically. Therefore we formulate the stereo matching by the minimization problem given in eq. (12), where the estimated disparity $\hat{p}_x$ is the one which minimizes the squared intensity error $e_I(p_x)^2$. For simplification of the mathematical derivation, $e_I(p_x)^2$ is defined without the averaging over several pixels.

$$\begin{aligned}\hat{p}_x &= \min_{p_x}\left(e_I(p_x)^2\right) \\ &= \min_{p_x}\left(\left(I(\boldsymbol{x}_R) - I(\boldsymbol{x}_R^s(p_x))\right)^2\right)\end{aligned} \tag{12}$$

Calculating the first derivative with respect to p_x, as given in eq. (13) and setting it to zero results in the condition given in eq. (14) as long as $g_I(p_x) \neq 0$ holds.

$$\begin{aligned}\frac{\partial e_I(p_x)^2}{\partial p_x} &= \frac{\partial\left(I(\boldsymbol{x}_R) - I(\boldsymbol{x}_R^s(p_x))\right)^2}{\partial p_x} \\ &= 2\left(I(\boldsymbol{x}_R) - I(\boldsymbol{x}_R^s(p_x))\right)\cdot\left(-g_I(p_x)\right)\end{aligned} \tag{13}$$

$$I(\boldsymbol{x}_R) - I(\boldsymbol{x}_R^s(p_x)) \stackrel{!}{=} 0 \tag{14}$$

Here, the intensity gradient along the epipolar line $g_I(p_x)$ is defined as follows:

$$g_I(p_x) = g_I\left(\boldsymbol{x}_R^s(p_x)\right) = \frac{\partial I(\boldsymbol{x}_{R0}^s + p_x\boldsymbol{e}_p)}{\partial p_x} \tag{15}$$

Based on the chain rule for derivatives it can be derived that $g_I(p_x)$ is given as follows, where $\boldsymbol{g}_I(\boldsymbol{x}_R)$ is defined as given in eq. (7).

$$g_I(p_x) = \boldsymbol{g}_I(\boldsymbol{x}_R^s(p_x))^T\boldsymbol{e}_p \tag{16}$$

After approximating eq. (14) by its first order Taylor-series it can be solved for p_x as given in eq. (17).

$$\hat{p}_x = \frac{I(\boldsymbol{x}_R) - I(\boldsymbol{x}_R^s(p_{x0}))}{g_I\left(\boldsymbol{x}_R^s(p_{x0})\right)} + p_{x0} \tag{17}$$

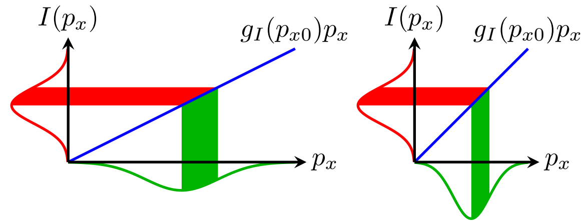


Figure 5. Camera sensor noise n_I can be considered as additive white Gaussian noise (AWGN) which disturbed the intensity values $I(\boldsymbol{x}_R)$ and thus effects the disparity observation as AWGN. As shown on the left, for a low image gradient along the epipolar line, the influence of the sensor noise n_I is stronger than for a high image gradient.

If we now consider $I(\boldsymbol{x}_R)$ in eq. (17) as an Gaussian distributed random variable, the variance $\sigma_{p_x}^2$ of the disparity p_x can be derived as given in eq. (18).

$$\begin{aligned}\sigma_{p_x}^2 &= \frac{\mathrm{Var}\{I(\boldsymbol{x}_R)\} + \mathrm{Var}\{I(\boldsymbol{x}_R^s(p_{x0}))\}}{g_I(\boldsymbol{x}_R^s(p_{x0}))^2} \\ &= \frac{2\sigma_N^2}{g_I(\boldsymbol{x}_R^s(p_{x0}))^2}\end{aligned} \tag{18}$$

Similarly, Figure 5 illustrates how the gradient g_I effects the estimation of p_x. Here the blue line represents the tangent at the disparity p_{x0} at which the intensity values are projected onto the disparities.

The variance $\sigma_{p_x}^2$ considers only the stochastic noise which is produced by the sensor and assumes that aside from that noise the corresponding image regions around $\boldsymbol{x}_R$ and $\boldsymbol{x}_R^s(\hat{p}_x)$ are identical. In reality this is not the case and especially not for a Raytrix camera, where neighboring micro lenses have different focal lengths and thus do not focus on the same virtual depth (see Fig. 3). Thus, beside the variance $\sigma_{p_x}^2$ we define a second error source which we call the focus uncertainty. In this focus uncertainty we take into account the obvious thought that a small intensity error e_{ISS} gives a more reliable disparity estimate than a large intensity error. Thus, we define the focus uncertainty as follows:

$$\sigma_f^2 = \alpha \cdot \frac{e_{ISS}(p_x)}{g_I(\boldsymbol{x}_R(p_x))^2} \tag{19}$$

Here α is a constant scaling factor which defines the weight of σ_f^2 with respect to $\sigma_{p_x}^2$. We chose α such that for micro images with a different focus plane σ_f^2 equals on average $\sigma_{p_x}^2$.

The observation uncertainty σ_z^2 results as the sum of $\sigma_{p_x}^2$ and σ_f^2 since we consider both error sources as uncorrelated. From eq. (4) one can see that z is the disparity p_x scaled by d^{-1}. Thus, for σ_z^2 the scaling factor d^{-2} has to be introduced, as given in eq. (20).

$$\sigma_z^2 = d^{-2}\cdot\left(\sigma_{p_x}^2 + \sigma_f^2\right) \tag{20}$$

3.4. Updating Virtual Depth Hypothesis

As described in Section 3.2 the observations for the inverse virtual depth z are performed starting from the shortest baseline up to the largest possible baseline, for which a virtual image point is still seen in both micro images. In that way for each pixel an exhaustive stereo matching over all possible micro images is performed leading to multi-view stereo. In our algorithm we incorporate new inverse virtual depth observations similar to the update step in a Kalman filter. Thus, the new inverse virtual depth distribution $\mathcal{N}(z, \sigma_z^2)$ results form the previous distribution $\mathcal{N}(z_p, \sigma_p^2)$ and the new inverse depth observation $\mathcal{N}(z_o, \sigma_o^2)$ as given in eq. (21).

$$\mathcal{N}(z, \sigma_z^2) = \mathcal{N}\left(\frac{\sigma_p^2 \cdot z_o + \sigma_o^2 \cdot z_p}{\sigma_p^2 + \sigma_o^2}, \frac{\sigma_p^2 \cdot \sigma_o^2}{\sigma_p^2 + \sigma_o^2}\right) \tag{21}$$

The baseline distance d is more or less proportional to the virtual depth $v = z^{-1}$. From eq. (20) one can see that the inverse virtual depth variance σ_z^2 is inverse proportional to d^2. Besides, the number of observations increases with the virtual depth v since one point occurs in more micro images. For the case that all depth observations are statistically independent and have the same variance $\sigma_i^2 = \sigma_o^2$ ($i \in 1, 2, \ldots, N$), the variance of the incorporated depth estimate σ_z^2 is just N times smaller then the observation variance σ_0^2, as defined in eq. (22).

$$\frac{1}{\sigma_z^2} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} = \frac{N}{\sigma_o^2} \tag{22}$$

Thus, one can assume that the inverse virtual depth variance σ_z^2 improves approximately proportional to v^3.

3.5. Calculating a Virtual Depth Map

Based on the observed inverse virtual depth z, a pixel in the raw image, defined by the coordinates $\boldsymbol{x}_R$, can be projected in a 3D space which we will call the virtual image space, denoted by the coordinates $\boldsymbol{x}_V = (x_V, y_V, v = z^{-1})^T$. Based on the main lens projection (eq. (1)) the virtual image space can be projected into a metric object space. Nevertheless, therefore a prior metric calibration as presented in [21] for instance is needed. The transform of raw image coordinated $\boldsymbol{x}_R$ to virtual image coordinates $\boldsymbol{x}_V$ is defined as given in eq. (23).

$$\begin{pmatrix} z \cdot x_V \\ z \cdot y_V \\ z \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & c_x & -c_x \\ 0 & 1 & c_y & -c_y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_R \\ y_R \\ z \\ 1 \end{pmatrix} \tag{23}$$

Here $\boldsymbol{c} = (c_x, c_y)^T$ is the center of the micro lens under which the pixel $\boldsymbol{x}_R$ lies. From eq. (23) the coordinates x_V and y_V result as follows.

$$x_V = (x_R - c_x)z^{-1} + c_x \tag{24}$$
$$y_V = (y_R - c_y)z^{-1} + c_y \tag{25}$$

For the following experiments we will define the virtual image $\boldsymbol{x}_V$ as a 2D depth map $v(x_V, y_V) = v(\boldsymbol{x}_V)$ or $z(x_V, y_V) = z(\boldsymbol{x}_V)$ respectively.

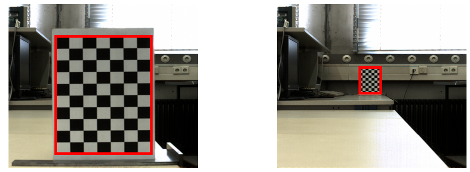


Figure 6. Test setup. Depth estimation algorithms are evaluated based on a planar chessboard target for different object distances. Only depth values in the red marked region of interest are evaluated.

4. Experiments & Results

In this section we want to present the evaluation of our proposed method. Therefore we performed experiments, where we compare our multi-view stereo (MVS) algorithm to the classical virtual depth estimation implemented in the Raytrix software [14]. Additionally, we compared both methods based on realistic data sets which are supplied by Raytrix [15].

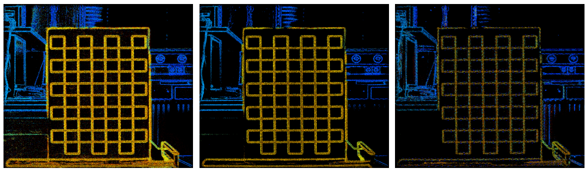
4.1. Experiments

All experiments were performed based on a Raytrix R5 camera with a main lens focal length of $f_L = 35$ mm. To evaluate the depth estimation methods a planar target is recorded for different object distances, as shown in Figure 6. Here, both images show the totally focused RGB-image of the recorded target for two different target distances. Since the target is placed frontal to the plenoptic camera for a perfect estimation one would expect a constant virtual depth across the complete plane.

For each of the recorded frames a virtual depth map was calculated, once using our probabilistic method and once using the classical algorithm [14]. Since we only want to evaluate the depth estimation algorithm itself without any post processing, all post processing steps like filtering or hole filling are disabled in the Raytrix software.

As one will see form the results in Section 4.2, our method results in a much denser depth map than [14]. Thus, to receive an as dense as possible depth map, both, the resolution as well as the sensitivity for the classical approach were set to high. Thereby, the algorithm uses a step size of 0.3 pixel and a correlation patch diameter between three and four pixel.

Since our method offers additionally to the virtual depth $v = z^{-1}$ an inverse virtual depth variance σ_z^2, two different



(a) (b) (c)

Figure 7. Color coded virtual depth maps calculated from the raw image of a Raytrix R5 camera. (a) Depth map calculated based on our MVS algorithm. All valid depth pixels are considered. (b) Depth map calculated based on our MVS algorithm. Only depth pixels with a variance $\sigma_z^2 < T(z)$ are considered. (c) Depth map calculated based on the classical algorithm [14]. High sensitivity as well as resolution was set.

depth maps were calculated based on our MVS algorithm. While the first depth map considers all valid depth pixel disregarding their variance, the second depth map considers only these depth pixel which have a variance σ_z^2 underneath a certain threshold $T(z)$, as defined in eq. (26).

$$\sigma_z^2(\boldsymbol{x}_V) < T(z) = \beta \cdot z(\boldsymbol{x}_V)^3 \tag{26}$$

The threshold $T(z)$ is chosen as a third order function of z due to the thoughts made in Section 3.4. In eq. (26) β is just a scaling factor, which defines the point density of the resulting depth map. In our experiments a scaling factor $\beta = 0.1$ was chosen. This resulted in a more or less equal point density for our approach compared to [14].

It is important to emphasize, that here no low-pass filtering is performed and just uncertain estimates are removed.

4.2. Results

Figure 7 exemplary shows the depth maps calculated for an object distance $a_L \approx 1.2\,\text{m}$. These depth maps correspond to the recoded scene which is shown on the left side in Figure 6. In Figure 7, (a) and (b) show the results of our MVS algorithm. Here, (a) includes all valid depth pixels, while (b) includes only those which have a variance $\sigma_z^2(\boldsymbol{x}_V) < T(z)$, as defined in eq. (26). The depth map (c) in Figure 7 shows the results of the classical algorithm [14].

From Figure 7 one can see already, that the outliers in our method are drastically reduced by introducing the threshold $T(z)$, while most of the details are kept. Besides, one can see that the depth map of [14] is much sparser than the raw depth map resulting from our approach. In addition it seems that the outliers of the method [14], especially on the chessboard plane are not statistically independent, but occur in clusters.

Beside the qualitative evaluation based on the depth maps some statistics were calculated for different object distances a_L. In this Section we present the results for $a_{L1} \approx 1.2\,\text{m}$, $a_{L2} \approx 3.1\,\text{m}$, and $a_{L3} \approx 5.1\,\text{m}$. For all three object distances Table 1 shows the depth pixel density of the corresponding algorithm. The depth pixel density

Method	depth pixel density a_{L1}	a_{L2}	a_{L3}
MVS (all depths)	0.3075	0.5041	0.5638
MVS ($\sigma_z^2(\boldsymbol{x}_V) < T(z)$)	0.1788	0.3900	0.4760
Classical alg. [14]	0.1305	0.3109	0.4216

Table 1. Depth pixel density across the chessboard target for different object distances a_L.

Method	standard deviation a_{L1}	a_{L2}	a_{L3}
MVS (all depths)	0.0366	0.0505	0.0386
MVS ($\sigma_z^2(\boldsymbol{x}_V) < T(z)$)	0.0104	0.0167	0.0170
Classical alg. [14]	0.0889	0.0632	0.0651

Table 2. Empirical standard deviation of the inverse virtual depth z for different object distances a_L.

is defined as the ratio between the number of valid depth pixels and the total number of pixels within the region of interest. Here one can see, that our method has a higher depth pixel density than the classical approach for all object distances.

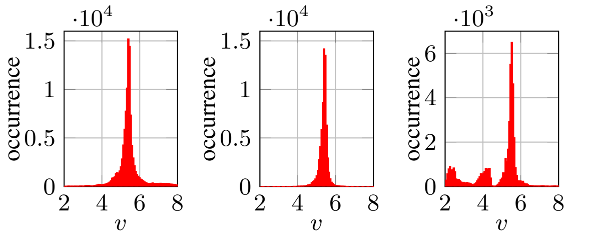
For all three object distances we calculated the empirical standard deviation of the inverse virtual depth values $z = v^{-1}$ across the chessboard target. The results are shown in Table 2. As one can see, the standard deviation of our MVS approach is better than that of the classical algorithm for all three object distances, even without removing outliers. After removing outliers, we achieve a standard deviation which is at least three times better than that of [14], while still having a higher depth pixel density (see Tab. 1). Also quite interesting to see is, that only sightly reducing the depth pixel density, by introducing the threshold $T(z)$, highly reduces the empirical standard deviation of the inverse virtual depth.

Figure 8 and 9 shows the virtual depth histograms across the chessboard target for the object distances $a_{L1} \approx 1.2\,\text{m}$ and $a_{L3} \approx 5.1\,\text{m}$. Especially from Figure 8 one can see that the outliers of the classical algorithm have some systematic characteristic instead of been uniformly distributed. Besides, the histograms again show quite well how the outliers in our approach are removed by introducing the threshold $T(z)$.

4.3. Results on Realistic Data Sets

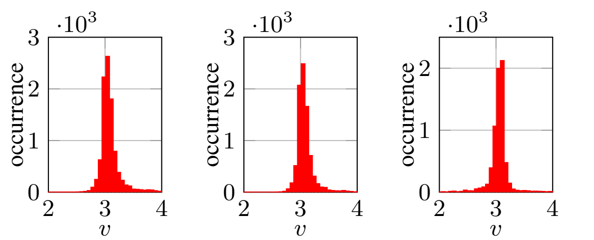
For a qualitative evaluation the depth maps for two sample scenes (Fig. 10) were calculated. Here the settings for the classical algorithm [14] were set similar to the experiments in Section 4.1.

We ran both algorithms on a NVIDIA GeForce GTX TITAN and measured the run-times given in Table 3. Scene 1 has a raw image resolution of $4016\,\text{pixel} \times 2688\,\text{pixel}$ and



(a) (b) (c)

Figure 8. Virtual depth histograms for object distance $a_{L1} \approx 1.2\,\mathrm{m}$. (a) Histogram of our MVS algorithm including all valid depth pixels. (b) Histogram of our MVS algorithm including all depth pixels with $\sigma_z^2 < T(z)$. (c) Histogram of the classical algorithm [14].



(a) (b) (c)

Figure 9. Virtual depth histograms for object distance $a_{L1} \approx 5.1\,\mathrm{m}$. (a) Histogram of our MVS algorithm including all valid depth pixels. (b) Histogram of our MVS algorithm including all depth pixels with $\sigma_z^2 < T(z)$. (c) Histogram of the classical algorithm [14].

scene 2 of $4008\,\mathrm{pixel} \times 2664\,\mathrm{pixel}$. We want to mention that the classical algorithm can be sped up by changing the settings. Nevertheless, this likely results in less quality of the depth map. The MVS run-times are highly dependent on the scene since for high virtual depths more observations are received than for low ones.

Without any ground truth it is difficult to evaluated the absolute accuracy. However, one can see that at regions of depth discontinuities (e.g. at the edges of the screws in scene 2) the MVS performs very well.

5. Summary and Our Contribution

In this paper we proposed a virtual depth estimation algorithm for a focused plenoptic camera. We introduce a graph of baselines which defines the multiple micro lens pairs in the MLA. Based on this graph multiple stereo-observations are obtained, starting from a short up to a long baseline. These observations are incorporated in a probabilistic depth map.

We expressed mathematically how the camera noise effects the disparity estimation. Thus, the estimated inverse virtual depths can be defined as Gaussian distributed random variables. The multiple inverse virtual depth observations of the same pixel are considered to be statistically independent and are incorporated one after another into the probabilistic depth map.

Based on the probabilistic depth map it is possible to remove outliers without any low-pass filtering by setting a threshold for the inverse virtual depth variance. Thus, discontinuities in the depth map are preserved.

The performed experiments showed that our algorithm outperforms the classical algorithm in accuracy and run-time.

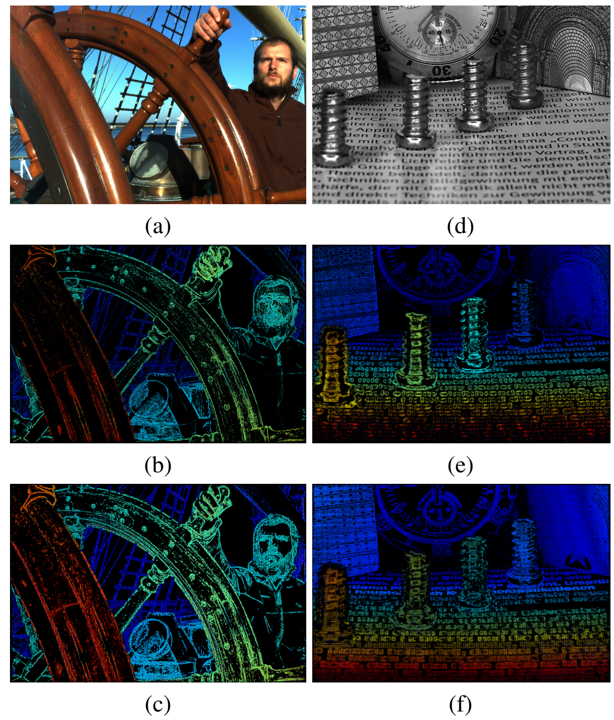


Figure 10. Depth estimation applied to sample scenes (data sets available at [15]). (a) and (d) show the totally focused images. (b) and (e) show the results of our MVS algorithm. (c) and (f) show the results of the classical algorithm [14].

Method	Scene 1 (pilot)	Scene 2 (watch)
MVS	58 ms	93 ms
Classical alg. [14]	513 ms	404 ms

Table 3. Run-times measured for depth estimation

Acknowledgement

This research is funded by the Federal Ministry of Education and Research of Germany in its program ”IKT 2020 Research for Innovation”.

References

[1] E. H. Adelson and J. Y. A. Wang. Single lens stereo with a plenoptic camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):99–106, February 1992. 1

[2] T. E. Bishop and P. Favaro. Full-resolution depth map estimation from an aliased plenoptic light field. In *Computer Vision ACCV 2010*, volume 6493 of *Lecture Notes in Computer Science*, pages 186–200. Springer Berlin Heidelberg, 2011. 2

[3] J. Engel, J. Sturm, and D. Cremers. Semi-dense visual odometry for a monocular camera. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1456, Dec 2013. 2, 3

[4] T. Georgiev and A. Lumsdaine. Depth of field in plenoptic cameras. In *Eurographics*, 2009. 2

[5] T. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. Nayar, and C. Intwala. Spatio-angular resolution tradeoffs in integral photography. In *Proc. 17th Eurographics conference on Rendering Techniques*, EGSR'06, pages 263–272, Aire-la-Ville, Switzerland, 2006. Eurographics Association. 1

[6] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. In *Proc. 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH*, pages 43–54, New York, NY, USA, 1996. ACM. 1

[7] S. Heber and T. Pock. Shape from light field meets robust pca. In *Proc. European Conference on Computer Vision (ECCV)*, 2014. 1

[8] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon. Accurate depth map estimation from a lenslet light field camera. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1547–1555, 6 2015. 1

[9] M.-J. Kim, T.-H. Oh, and I. S. Kweon. Cost-aware depth map estimation for lytro camera. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 36–40, 10 2014. 2

[10] A. Lumsdaine and T. Georgiev. Full resolution lightfield rendering. Technical report, Adobe Systems, Inc., 2008. 2

[11] A. Lumsdaine and T. Georgiev. The focused plenoptic camera. In *Proc. IEEE International Conference on Computational Photography (ICCP)*, pages 1–8, San Francisco, CA, April 2009. 1, 2

[12] R. Ng. *Digital light field photography*. PhD thesis, Stanford University, Stanford, USA, July 2006. 2

[13] R. Ng, M. Levoy, M. Brédif, G. Guval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. Technical report, Stanford University, Computer Sciences, CSTR, 05 2005. 2

[14] C. Perwaß and L. Wietzke. Single lens 3d-camera with extended depth-of-field. In *Proc. SPIE 8291, Human Vision and Electronic Imaging XVII*, Burlingame, California, USA, January 2012. 1, 2, 3, 6, 7, 8

[15] Raytrix GmbH. Test scenes, last accessed: July 30, 2015. http://www.raytrix.de/index.php/Research.html. 6, 8

[16] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 673–680, Dec 2013. 2

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

[18] S. Wanner and B. Goldlücke. Variation light field analysis fo disparity estimation and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):606–619, 3 2014. 1

[19] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4d lightfields. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[20] Z. Yu, X. Guo, H. Lin, A. Lumsdaine, and J. Yu. Line assisted light field triangulation and stereo matching. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2792–2799, 12 2013. 1

[21] N. Zeller, F. Quint, and U. Stilla. Calibration and accuracy analysis of a focused plenoptic camera. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3:205–212, 09 2014. 2, 6