NARROW FIELD-OF-VIEW VISUAL ODOMETRY BASED ON A FOCUSED PLENOPTIC CAMERA

N. Zeller[a, b, *], F. Quint[a], U. Stilla[b]

[a] Faculty of Electrical Engineering and Information Technology, Karlsruhe University of Applied Sciences, 76133 Karlsruhe, Germany - niclas.zeller@hs-karlsruhe.de; franz.quint@hs-karlsruhe.de
[b] Department of Photogrammetry and Remote Sensing, Technische Universität München, 80290 Munich, Germany - stilla@tum.de

KEY WORDS: Depth estimation, featureless, focused plenoptic camera, light-field, SLAM, visual odometry

ABSTRACT:

In this article we present a new method for visual odometry based on a focused plenoptic camera. This method fuses the depth data gained by a monocular Simultaneous Localization and Mapping (SLAM) algorithm and the one received from a focused plenoptic camera. Our algorithm uses the depth data and the totally focused images supplied by the plenoptic camera to run a real-time semi-dense direct SLAM algorithm. Based on this combined approach, the scale ambiguity of a monocular SLAM system can be overcome. Furthermore, the additional light-field information highly improves the tracking capabilities of the algorithm. Thus, visual odometry even for narrow field of view (FOV) cameras is possible. We show that not only tracking profits from the additional light-field information. By accumulating the depth information over multiple tracked images, also the depth accuracy of the focused plenoptic camera can be highly improved. This novel approach improves the depth error by one order of magnitude compared to the one received from a single light-field image.

1. INTRODUCTION

Even though the concept of a plenoptic camera has been developed more than hundred years ago (Ives, 1903, Lippmann, 1908), only for the last years research based on plenoptic cameras became more and more popular. One reason therefore was the appearance of the Lytro (Ng, 2006) and Raytrix (Perwaß and Wietzke, 2012) cameras, which are the first commercially available plenoptic cameras. Besides, today's graphic processor units (GPUs) are capable to process the recordings of a plenoptic camera with acceptable frame rates.

Both, the Lytro and the Raytrix camera capture the light-field of a scene as a 4D function based on a micro lens array (MLA) in front of the sensor. Nevertheless, both therefore follow a slightly different concept. While the Lytro camera is an "unfocused" plenoptic camera, which has a high angular resolution but a low spatial resolution (Ng, 2006), the Raytrix camera is a focused plenoptic camera as described for the first time by (Lumsdaine and Georgiev, 2008). The focused plenoptic camera, which is also called plenoptic camera 2.0, captures the light-field with higher spatial resolution. Therefore it is paid by angular resolution. The high spatial resolution is beneficial in estimating depth from the recorded light-field as described in (Perwaß and Wietzke, 2012).

Even though the Raytrix camera supplies depth information, the accuracy is rather low for a distance of a few meters compared to other depth sensors, like Time-of-Flight (ToF) cameras or stereo camera systems, at least at a comparable field of view (FOV). The depth accuracy and range of a focused plenoptic camera strongly decays when reducing the focal length. Thus, a trade-of between wide FOV and long depth range has to be found (Zeller et al., 2014a).

The advantages of a plenoptic camera lie more in its small dimensions which are similar to those of a conventional camera. In future there will also be miniaturized light-field sensors available, which will be assembled in smartphones (Venkataraman et al., 2013). In addition, a Raytrix camera offers a much larger depth of field (DOF) compared to a standard camera at the same aperture. Thus, a Raytrix camera has a much closer short range limit than e.g. a stereo camera system.

In many navigation applications such small sensors are profitable, for example on unmanned aerial vehicles (UAVs), where space and weight is limited. But also for indoor navigation or blind people assistance, where bulky sensors can be annoying, such small and light sensors are beneficial.

For this kin of applications today mostly monocular visual odometry (or Simultaneous Localization and Mapping (SLAM)) systems are used, which gain depth information from motion (Structure from Motion (SfM)). However, such monocular systems come with some drawbacks. One drawback of a monocular visual odometry system is its scale ambiguity. Thus, especially in navigation applications additional sensors are needed to gather metric dimensions. Another disadvantage of monocular systems is that no depth can be estimated when rotating around the camera's optical center as well as for structures which are homogeneous along their epipolar lines.

Thus, a plenoptic camera seems to be a perfect compromise between a monocular and a stereo camera or depth sensor based system for visual odometry. Since for a plenoptic camera based system rough depth information is available for each single frame, tracking becomes much more robust compared to a monocular system. Therefore, if the camera moves not to far from one frame to the next, a smaller section of the scene is sufficient for reliable tracking, as will be shown in the sequel. Hence, depending on the application a narrow FOV camera can be used and so the depth accuracy further be improved.

1.1 Related Work

SLAM systems can be divided into feature-based and direct methods. Some of them are based on depth or stereo image sensors,

*Corresponding author.

while other are pure monocular.

In feature-based SLAM or SfM systems features are extracted from the recorded 2D images based on some feature detector. Those features are matched between the corresponding images. In a second step the appropriate camera positions and the 3D feature coordinates are estimated based on the feature correspondences. This reduces the complexity of the problem since most image information is disregarded (Klein and Murray, 2007, Li and Mourikis, 2013). Nevertheless, it also effects the robustness of the tracking negatively, especially if the images do not match the feature criteria. Thus, different feature types are used (Klein and Murray, 2008, Eade and Drummond, 2009, Concha and Civera, 2014). From a feature-based method itself only a sparse point cloud is received. To obtain a dense point cloud it has to be estimated after feature-based pose estimation using multi view stereo methods (Newcombe and Davison, 2010).

Direct methods avoid the feature extraction by performing tracking and mapping directly on the recorded images (Forster et al., 2014). Thus, tracking becomes much more robust since all image data is used. Because the complete images are used, dense depth maps can be directly estimated (Newcombe et al., 2011). Such direct, dense methods are very complex. The complexity can be reduced by performing semi-dense direct tracking and mapping algorithms (Engel et al., 2013, Engel et al., 2014). Semi-dense means, that only image regions of high contrast are considered for tracking and mapping and all homogeneous regions are neglected. These semi-dense methods are capable to run in real-time on today's standard central processing units (CPUs) or even on smartphones (Schöps et al., 2014).

The use of multiple cameras or depth sensors strongly simplifies the SLAM problem. Here depth information is already received without motion. Besides, the scale of the scene is received directly from the recorded images without using any additional sensors (Akbarzadeh et al., 2006, Izadi et al., 2011, Dansereau et al., 2011, Kerl et al., 2013).

1.2 Our Contribution

In this paper we present the advantages of visual odometry based on a focused plenoptic camera. On one side the tracking robustness of a monocular SLAM algorithm can be improved by the introduction of light-field information. More than this, for the case that sequent images have sufficient overlap, convergence of the tracking algorithm and thus, gaining depth information can be assured even for a narrow FOV. In monocular visual odometry tracking converges only at a wide FOV. On the other side we demonstrate that the depth information of a plenoptic camera is considerably improved by tracking multiple frames and combining their information.

In Section 2 we briefly present the concept of a focused plenoptic camera. Section 3 presents the monocular SLAM algorithm on which our light-field based method relies. Our method for visual odometry based on a plenoptic camera is described in Section 4. The method is evaluated in Section 5 and the corresponding results are presented in Section 6. Section 7 draws conclusion.

2. THE FOCUSED PLENOPTIC CAMERA

As already mentioned in the introduction, a plenoptic camera records the light-field of a scene as 4D function. Since in free space the intensity along a light-ray does not change, here the definition of the light-field as 4D function is sufficient as shown

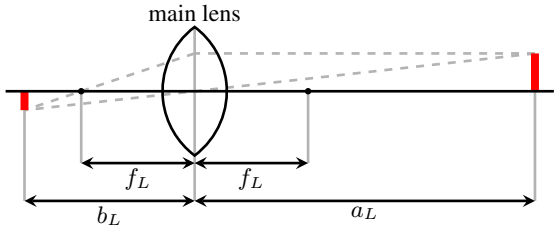

Figure 1. Optical path of a thin lens. An object in the distance a_L in front of the main lens results in a focused image in the distance b_L behind the main lens.

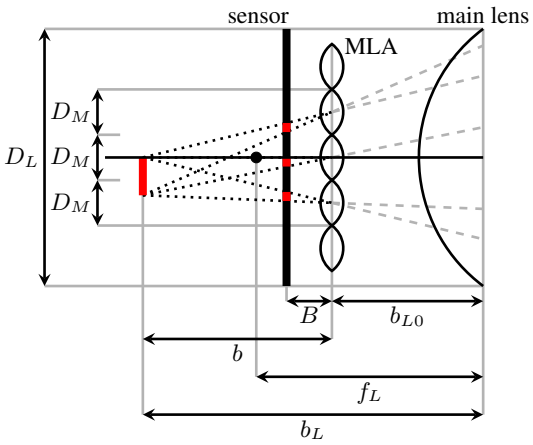

Figure 2. Optical path inside a Raytrix camera. The virtual image which would occur behind the sensor in the distance b_L to the main lens is projected by the MLA to multiple micro images on the sensor.

in (Gortler et al., 1996). Thus, a ray in the light-field is defined by two position coordinates and two angle coordinates.

The concept of the plenoptic camera of Raytrix, which is used in our research, easily can be derived from a thin lens projection as shown in Figure 1. The relation between the object distance a_L and the image distance b_L is defined by the thin lens equation given in eq. (1).

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

Different to a conventional camera, in a Raytrix camera the sensor is not placed on the image plane, in the distance b_L behind the main lens. Instead, the sensor is placed closer to the main lens. Besides, a MLA is placed in front of the sensor and focuses the virtual image, which would occur behind the sensor, on the sensor, as shown in Figure 2. A distinct feature of Raytrix cameras is that the MLA consists of three different types of micro lenses with different focal lengths. Each type focuses on a different image distances. This increases the DOF of the camera.

Within its DOF each micro lens can be considered as a pinhole and thus each pixel in a micro image represents the corresponding central ray of the micro lens.

Since a point of the virtual image occurs focused in more than one micro image and in each from a slightly different perspective, the distance b between the MLA and the corresponding virtual image point can be calculated by triangulation, as given in eq. (2).

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

In eq. (2) d is the length of the base line between the two micro

lenses considered for triangulation and p_x is the parallax measured in the corresponding micro images. The distance B between MLA and sensor is not specified. Thus, the distance to the virtual image b is calculated as a relative measure, the so called virtual depth v, as defined in eq. (3).

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

The relationship between the virtual depth v and the metric object distance relies on the thin lens equation and some camera specific parameters. This relationship has to be estimated in a calibration procedure as presented in (Zeller et al., 2014a) for instance.

The depth for a virtual image point can only be estimated if this point can be found in more than one micro image. Thus, the virtual depth v can only be estimated for regions of high contrast. For all other regions the depth map has to be filled by interpolation.

For each virtual image point for which the virtual depth is known, the accompanied pixel on the sensor can be determined. Thus, based on a dense virtual depth map, a totally focused, central perspective image can be calculated from the raw light-field image since it is known which light-rays contribute to which virtual image point. To calculate the totally focused image it is sufficient to have a reliable virtual depth only in regions of sufficient contrast since for homogeneous regions it does not matter which pixels belong to the same virtual image point.

Virtual image points which have a long image distance b_L (short object distance a_L) occur in more micro images than points with a short image distance. Thus, the totally focused image has a higher effective resolution in regions with a small virtual depth (short image distance b_L) than in regions with a high virtual depth (long image distance b_L).

For a more detailed description on depth estimation and image synthesis we refer to (Perwaß and Wietzke, 2012) and (Zeller et al., 2014a).

3. MONOCULAR DIRECT SLAM

Monocular direct SLAM methods do not perform any feature extraction on the recorded images, but work directly on the intensity images recorded by a conventional camera. Based on the recorded images such methods track the current camera position and also build a 3D map of the environment. In the open source project LSD-SLAM (Engel et al., 2014) a monocular direct SLAM algorithm is implemented. Different from other direct SLAM methods which used the complete image information for tracking and mapping (Newcombe et al., 2011), LSD-SLAM works only on image regions with sufficient contrast. This allows to ignore homogeneous image regions which carry less information suitable for pose and depth estimation. Working only on image regions with sufficient contrast strongly reduces the amount of data and consequently the complexity of the problem. Thus, the complete algorithm is capable to run in real-time on a standard CPU.

In this article we are mainly interested in the tracking of new frames and the depth estimation algorithm which we improve by introducing light-field information. Thus, in this article we will not discuss key-frame selection, global map optimization or other aspects of SLAM algorithms.

The following paragraphs describe very briefly the probabilistic depth model, the tracking and the depth estimation of LSD-SLAM. For a more detailed explanation we refer to (Engel et al., 2013, Engel et al., 2014).

3.1 Inverse Depth Map

In LSD-SLAM the depth map of a frame is not just defined as a 2D map of depth values, but as a 2D map of random variables. For the case that the camera rotation between two frames is small, the estimated depth is approximately inverse proportional to the estimated disparity. Since the disparity can be considered to be disturbed by additive Gaussian noise, the *inverse* depth map is also defined as a map of Gaussian distributed random variables. Each pixel $\boldsymbol{p}_i$ in the inverse depth map D is defined by the expected inverse depth value $d_i = D(\boldsymbol{p}_i)$ and the corresponding inverse depth variance $\sigma_{di}^2 = V(\boldsymbol{p}_i)$. The inverse depth map D is continuously refined when new images are added. Thus, with each new observation the inverse depth variance σ_{di}^2 decays and the inverse depth becomes more reliable.

3.2 Tracking

For each new frame its pose with respect to a reference frame has to be estimated. The transform between a reference coordinate system $\boldsymbol{x}_R$ (camera coordinates of the reference frame) and the camera coordinates of the new frame $\boldsymbol{x}_C$ is defined by a rigid transform $\boldsymbol{G} \in \mathrm{SE}(3)$, as given in eq. (4).

$$\boldsymbol{x}_C = \begin{pmatrix} x_C \\ y_C \\ z_C \\ 1 \end{pmatrix} = \boldsymbol{G} \cdot \boldsymbol{x}_R = \boldsymbol{G} \cdot \begin{pmatrix} x_R \\ y_R \\ z_R \\ 1 \end{pmatrix} \tag{4}$$

The rigid transform $\boldsymbol{G}$ is defined as the combination of a rotation and a translation in 3D space, as given in eq. (5).

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{pmatrix} \quad \text{with} \quad \boldsymbol{R} \in \mathrm{SO}(3) \text{ and } \boldsymbol{t} \in \mathbb{R}^3. \tag{5}$$

The Matrix $\boldsymbol{G}$ has six degrees of freedom which have to be estimated. Those are the three rotation angles ϕ, ω, and κ as well as the coefficients of the translation vector t_x, t_y, and t_z. In LSD-SLAM those six degrees of freedom are estimated based on the intensity images by minimizing the photometric error. The minimization is done based on a weighted Gauss-Newton optimization. In the algorithm, the Gauss-Newton optimization is performed iteratively on different pyramid levels, starting from very low image resolution up to full image resolution.

Because of the scale-ambiguity of the monocular SLAM scale-drifts can occur during tracking. To handle such drifts the camera pose between reference frames ($\boldsymbol{x}_R^{(1)}$ and $\boldsymbol{x}_R^{(2)}$) is represented as a scale-aware 3D similarity transform $\boldsymbol{S} \in \mathrm{Sim}(3)$ instead of a rigid transform, as defined in eq. (6) and (7).

$$\boldsymbol{x}_R^{(2)} = \boldsymbol{S} \cdot \boldsymbol{x}_R^{(1)} \tag{6}$$

$$\boldsymbol{S} = \begin{pmatrix} s\boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{pmatrix}$$
$$\text{with} \quad \boldsymbol{R} \in \mathrm{SO}(3), \boldsymbol{t} \in \mathbb{R}^3 \text{ and } s \in \mathbb{R}^+. \tag{7}$$

Since this definition of the pose between two camera position results in an additional degree of freedom, the average inverse depth map is set to one for each reference frame.

3.3 Depth Estimation

The inverse depth map D is updated based on each new frame which was successfully tracked. Therefore, for each pixel with

sufficient contrast, stereo-matching along the epipolar line is performed. If there does not already exist a depth hypothesis for the current pixel, searching is performed over the full disparity range. Otherwise, if there exist a hypothesis for the pixel, the search interval is limited to $d_i \pm 2\sigma_{di}$. Here d_i is the mean of the inverse depth pixel and σ_{di} the corresponding standard deviation. Limiting the search interval for stereo-matching to $\pm 2\sigma_{di}$ prevents the algorithm from finding multiple matched when the baseline increases.

In LSD-SLAM the error of an observed disparity is modeled by two different error sources. One error source is the geometric error, which results from noise on the estimated camera pose and the intrinsic camera parameters, and affects the position and orientation of the epipolar line. The second error source is the photometric error, which results from noise in the intensity image. It is considered that both errors are Gaussian distributed and additively interfere the observed disparity and thus the observed inverse depth. Thus, based on both error sources the variance of the inverse depth observation is defined. For a detailed definition of the two error sources we refer to (Engel et al., 2013).

The obtained observation of the inverse depth is incorporated into the already existing inverse depth hypothesis as known from the update step of a Kalman filter, as given in eq. (8) and (9).

$$d_i^{(n+1)} = \frac{\left(\sigma_{di}^{(n)}\right)^2 \cdot d_o + \sigma_o^2 \cdot d_i^{(n)}}{\left(\sigma_{di}^{(n)}\right)^2 + \sigma_o^2} \tag{8}$$

$$\left(\sigma_{di}^{(n+1)}\right)^2 = \frac{\left(\sigma_{di}^{(n)}\right)^2 \cdot \sigma_o^2}{\left(\sigma_{di}^{(n)}\right)^2 + \sigma_o^2} \tag{9}$$

In eq. (8) and (9) $d_i^{(n)}$ is the expected value of the inverse depth after n incorporated observations and $\sigma_{di}^{(n)}$ the corresponding standard deviation. d_o and σ_o are the inverse depth and standard deviation of the new observation.

4. PLENOPTIC CAMERA BASED VISUAL ODOMETRY

A plenoptic camera seems to be perfectly suited for visual odometry since it supplies much more information about the recorded scene than just a standard camera. Even though a plenoptic camera gathers more information of a scene than a standard camera, both cameras are similarly in size. From the 4D light-field recorded by a plenoptic camera depth information can be obtained for regions with sufficient contrast.

In consideration of the capabilities of a plenoptic camera it is worthwhile to run a direct SLAM with additional light-field information. Here, it has to be investigated if tracking of the SLAM algorithm as well as the depth map accuracy of the focused plenoptic camera can both be improved.

In the case of LSD-SLAM the algorithm starts from a completely random depth map. Here two problems arise. On one hand for an accurate tracking of new camera poses a depth map is already needed. Since the depth map is initially random, it can last for many frames until tracking converges. On the other hand accurate depth from two images can only be estimated when the corresponding camera position is known with sufficient precision. This is quite a vicious circle.

Cameras with a wide FOV gather much more of a scene in their images compared to cameras with a narrow FOV. Due to this, for wide angle cameras the tracking of the monocular SLAM algorithm converges even without an accurate depth map after a reasonable amount of translation. However, this is not the case for narrow FOV cameras, for which the SLAM algorithm does not reach convergence.

Our approach is to overcome this issue by performing visual odometry and SfM based on a focused plenoptic camera. The method we present is basically divided into two steps. In a first step metric depth is calculated only based on light-field information (Section 4.1). Afterwards, this depth map is improved by visual odometry, which benefits from the larger baseline compared to a single light-field image (Section 4.2).

4.1 Depth and Image Synthesis from Light-Field

For each recorded frame the virtual depth map is calculated solely based on the light-field information. Using the virtual depth map for each frame a central perspective totally focused image is calculated.

To calculate the totally focused image all pixels on the sensor corresponding to a virtual image point have to be combined. Therefore, based on the virtual depth v of a virtual image point a radius R can be defined, as given in eq. (10). Here D_M is the diameter of a micro lens and B the distance between MLA and sensor (see Figure 2).

$$R = \frac{|v| \cdot D_M}{2 \cdot B} \tag{10}$$

The radius R defines a circular area around the orthogonal projection of a virtual image point on the sensor plane. This area comprises all micro images in which the virtual image point actually occurs. Besides, from the virtual depth value v it is known in which of the three types of micro lenses the virtual image point occurs focused. The resulting intensity value of the virtual image point finally is calculated as a weighted mean of all corresponding pixels in the selected micro images (Perwaß and Wietzke, 2012). The weights in the mean calculation balance the vignetting of the micro lenses.

Since the main lens of the plenoptic camera does not perform an perfect central projection, but adds distortion, during an image calibration the intrinsic camera parameters as well as distortion parameters have to be estimated. For the experiments presented in Section 5 a camera model consisting of a camera constant f, the principal point (c_x, c_y), three radial symmetric and two radial asymmetric distortion parameters was defined. Based on this model the totally focused image as well as the virtual depth map are rectified.

After the image rectification, the virtual depth map is further processed to result in a metric depth map by using the camera parameters obtained in a prior depth map calibration. As transform from virtual depth v to object distance a_L, the behavioral model as described by (Zeller et al., 2014b) and defined in eq. (11) is used.

$$a_L = \frac{v \cdot c_1 + c_2}{1 - v \cdot c_0} \tag{11}$$

Thus, additionally to a central perspective intensity image, for each frame a rough metrical depth map is available.

4.2 Improving Depth with visual odometry

Due to the small baseline between the micro images of the plenoptic camera the metric depth estimated only based on light-field information is not accurate enough for certain applications (Zeller

Figure 3. For the shown scene an image sequence of 400 frames was recorded by the Raytrix R5 camera and evaluated by visual odometry.

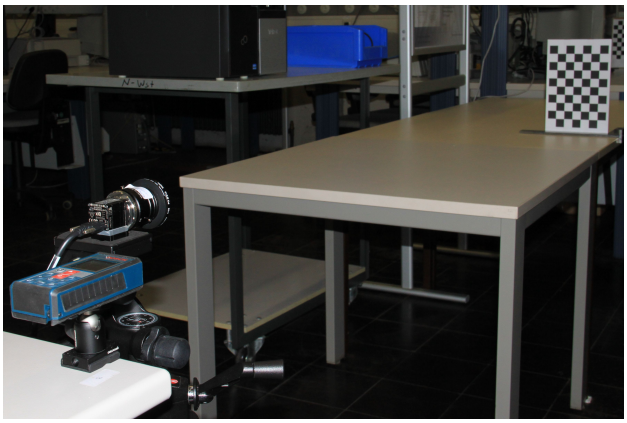

Figure 4. Setup to measure depth accuracy. A chessboard target in a defined distance is recorded by the Raytrix camera. A laser range finder measures the distance to the target for reference.

et al., 2014c). Hence, we improve the depth accuracy by applying visual odometry to a sequence of light-field images.

4.2.1 Initializing Inverse Depth For initialization of the light-field based visual odometry the first frame of the sequence is set as reference frame. Different from LSD-SLAM, where the inverse depth map D of this frame is initialized randomly, D is initialized by the inverse metric depth map calculated from the light-field information of this frame. Here only the pixels with sufficient contrast are initialized. Besides, the inverse depth map variance V is initialized by the mean square error of the inverse metric depth map of the plenoptic camera. The mean square error is defined as a function of the inverse depth and is obtained arithmetically from the data recorded during the depth calibration.

4.2.2 Updating Inverse Depth For all pixels in the totally focused image, which have sufficient contrast and for which a correspondence in the next frame has been established, a new inverse metric depth observation is received. The new observation is incorporated into the existing inverse depth hypothesis using an algorithm similar to (Engel et al., 2014), as described in Section 3.3. At one point, according to the increasing baseline, the depth accuracy received from stereo-matching outperforms the one received from light-field information.

For pixels which appear for the first time in the image sequence (e.g. due to occlusion or changed viewport) no correspondence can be established and thus, no depth hypothesis exists. These pixels are initialized with the light-field based inverse depth and are treated like other depth pixels in the following frames. Thus, different from LSD-SLAM, never an exhaustive search along the complete epipolar line has to be performed and hence, ambiguous pixel correspondences can be prevented.

4.3 Benefits

Due to the light-field information an inverse depth map is available already with the first recorded frame and therefore from the very beginning of a sequence robust tracking is possible. Thus, the light-field based algorithm converges faster than monocular SLAM. Furthermore, as the experiments will reveal, in some applications, SLAM is possible even for images with narrow FOV, for which monocular SLAM fails. Nevertheless, here it has to be guaranteed that between consecutive frames the viewport does not change to much. Since the plenoptic camera has a very large DOF and supplies rough depth information for each pixel, lost or inconsistent pixels immediately can be newly initialized. Thus, the light-field based SLAM can handle strong occlusions in the scene between consecutive frames quite well.

Besides, by taking advantage of the larger baseline between consecutive images a more accurate depth value can be calculated than from the light-field information.

5. EXPERIMENTS

To evaluate the method for visual odometry based on a focused plenoptic camera several experiments were performed. In these experiments we focused on measuring the accuracy of recorded objects in 3D space. Besides, we want to demonstrate the capabilities of the plenoptic camera based visual odometry in a sample scene compared to the monocular approach. Since our interest lies in the depth map of the scene, we did not measure a ground truth for the performed trajectories and thus, the tracking itself was not evaluated explicitly.

All our experiments were performed by using a Raytrix R5 camera with a focal length of the main lens of 35 mm. This results in a FOV of approximately 18° horizontally as well as vertically.

5.1 3D Reconstruction

As a proof for the improved tracking capabilities of our method compared to the monocular case, we recorded an image sequence composed of 400 frames by the Raytrix camera. This corresponds to a video length of approximately 8 s in which the camera was moved free hand by roughly 3 m. As will be seen, this short sequence is enough to estimate a 3D point cloud of the scene with good accuracy.

The sequence is evaluated once by applying the standard monocular LSD-SLAM algorithms only to the sequence of totally focused images and once using our approach with additional depth information from the light-field. An image of the recorded scene is shown in Figure 3.

5.2 Depth Accuracy

Two experiments were performed to evaluate the depth accuracy of our method. We evaluate the depth over time as well as over distance. For both experiments the same setup was used as presented in the following paragraph.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-II-3-W4-285-2015

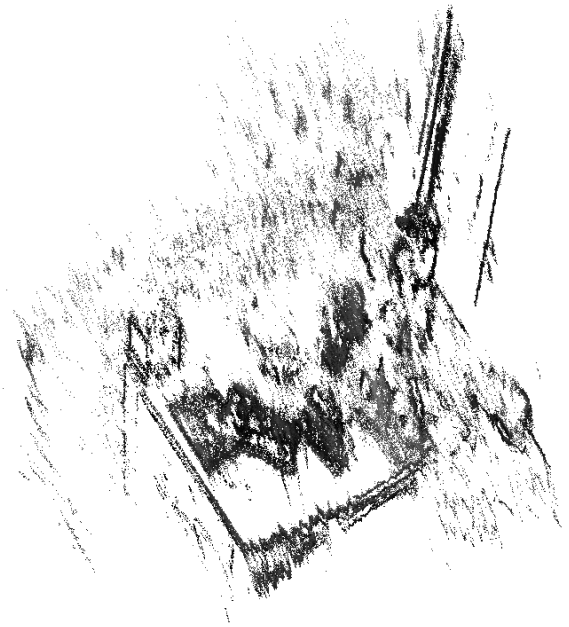

Figure 5. 3D point cloud of a sample scene recorded by a Raytrix R5 camera after applying the LSD-SLAM algorithm to the recorded sequence totally focused images without using the light-field based depth information.

5.2.1 Measurement Setup The setup we used to measure the depth accuracy is shown in Figure 4. Here the Raytrix camera is assembled on a tripod. Parallel to the image plane of the camera a chessboard target is placed in a certain distance. Besides, a laser range finder is placed close to the camera to measure a reference distance.

In the experiments not only the accuracy of the depth map received from the Raytrix camera itself has to be evaluated. Also the depth calculated by visual odometry will be measured. Thus, an image sequence is recorded, while the camera is translated in vertical direction. For each object distance a vertical movement of 20 cm was performed while recording the image sequence.

5.2.2 Accuracy as Function of Sequence Length In the first experiment the depth accuracy over a sequence of images, while the camera is moving in vertical direction, is evaluated for object distances from approx. 2.6 m to 5.3 m with a spacing of 30 cm. Exemplary we present the results for an object distance of 3.183 m. The calculated metric depth map at each frame is read out and analyzed. Since the camera is moved more or less uniformly over time, this evaluation is equivalent to measuring the accuracy as function of baseline distance.

5.2.3 Accuracy as Function of Object Distance The second experiment is performed to evaluate the depth accuracy improvement of the light-field based visual odometry compared to the depth calculated from pure light-field information of a single image of the plenoptic camera. For the pure light-field based depth, the standard deviation of the Raytrix camera was measured for object distances from 1 m to 5.3 m. For our visual odometry based approach, the standard deviation of the depth was evaluated for the 10 object distances in the range from 2.6 m to 5.3 m, after a vertical translation of 20 cm.

6. RESULTS

This section presents and discusses the results of the experiments described in Section 5.

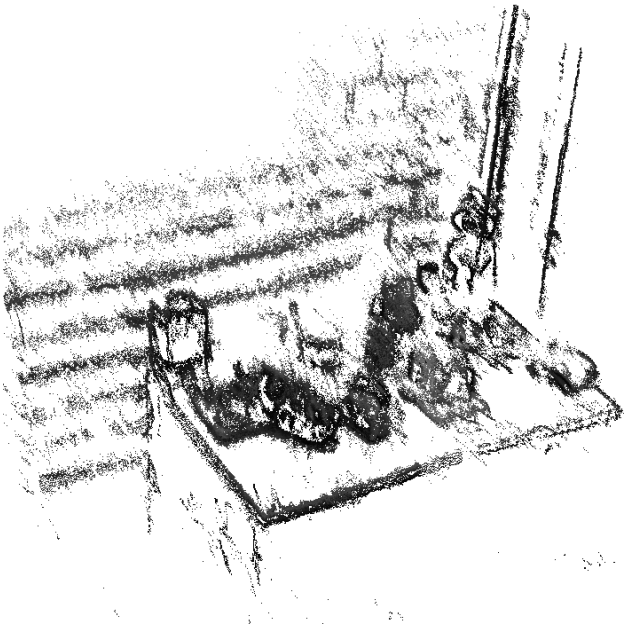

Figure 6. 3D point cloud of a sample scene recorded by a Raytrix R5 camera after applying our light-field based visual odometry method to the recorded light-field sequence.

6.1 3D Reconstruction

Figure 5 shows the 3D point cloud of the sample scene (Figure 3) which was calculated when applying the monocular LSD-SLAM algorithm (i.e. without using light-field information) to the sequence of totally focused images. Figure 6 shows the corresponding results when applying our light-field based approach. Of course, those two point clouds give only a qualitative measure for the two approaches. Nevertheless, in the point cloud received from the light-field based approach one can see, that the rectangular shape of the table is kept as well as the straight edges of the wall in the back. Besides, the objects on the table are modeled quite well.

For the pure monocular case, where tracking starts from a totally random inverse depth map, the algorithm is not capable to track the camera pose appropriately. This is quite obvious since according to the narrow FOV, the camera captures only a small section of the scene. This section is not sufficient to find the correct camera pose without any depth information. Since for the light-field based visual odometry each new depth hypothesis is initialized by its inverse metric depth calculated from the light-field information, tracking is much more robust and stable.

6.2 Depth Accuracy

6.2.1 Accuracy as Function of Sequence Length Figure 7 shows the course of the depth's standard deviation over all frames in the recorded sequence and thus as a discrete function over time. Since the sequence was recorded for a more or less homogeneous movement in vertical direction, the standard deviation can also be considered as a function of the baseline distance to the first frame. As already mentioned, the baseline between the first and the last frame is 20 cm in length. Besides, in Figure 7 the indices of the start and end frame of the movement are marked. Those frames were detected visually from the complete sequence of images

One can see from Figure 7 that the curve has approximately $1/x$ behavior. This conforms to the theoretical depth accuracy of a pair of stereo images recorded for the simplified case. Here one

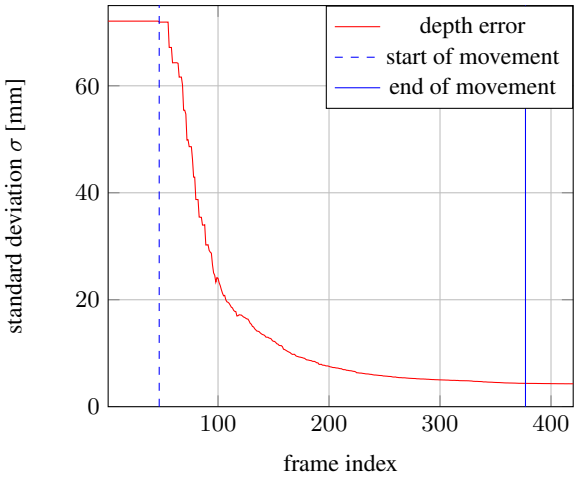

Figure 7. Standard deviation of the measured depth for a chessboard target in 3.183 m distance to the camera. From the first to the last frame the camera was translated by 20 cm in vertical direction.

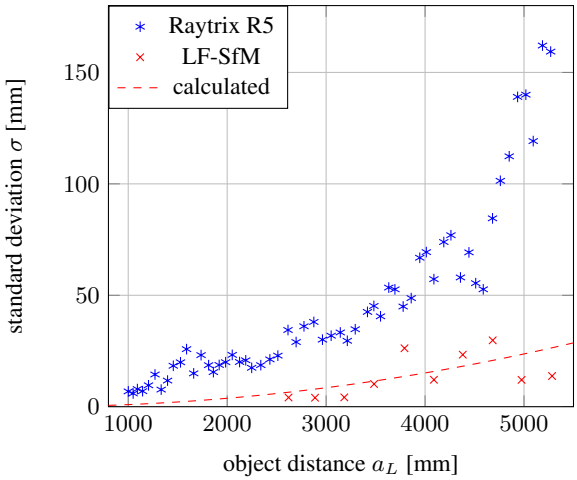

Figure 8. Standard deviation measured over object distance. Blue asterisk: Standard deviation measured by a Raytrix R5 camera. Red cross: Calculated standard deviation of light-field based visual odometry after a translation of 20 cm. Red dashed line: Standard deviation for a stereo camera pair with baseline distance of 20 cm, intrinsic parameters similar to the Raytrix camera, at a disparity standard deviation of 0.3 pixel.

can derive the depth accuracy based on the theory of propagation of uncertainty, as given in eq. (12).

$$\sigma_Z = \frac{f \cdot B}{p_x^2} \cdot \sigma_{p_x} = \frac{Z^2}{f \cdot B} \cdot \sigma_{p_x} \tag{12}$$

In eq. (12) f represents the camera constant, B the baseline distance, Z the object distance and p_x the measured parallax. The standard deviation of the parallax σ_{p_x} can be considered as constant.

After the first moving frame there is still a range of about eight frames where the standard deviation does not decay. In this range the baseline to the first frame is too short to improve the depth of the Raytrix camera and thus, no improvement in the depth accuracy is achieved here. Thereafter, the larger baseline built by subsequent frames leads to a quite steep descent of the depth's standard deviation. After approximately 200 frames, corresponding to as low as approx. 10 cm of movement, the standard deviation asymptotically reaches its minimum value. Thus, our visual odometry algorithm considerably improves the depth calculated solely from light-field information. This will be analyzed in more detail in the next paragraph. However, one has to mention that the monocular SLAM (i.e. without light-field information) does not converge in this measurement setup. The reason therefore is, that the FOV for a main lens with a focal length of 35 mm is too narrow to reach convergence in tracking without any depth information.

6.2.2 Accuracy as Function of Object Distance Figure 8 shows the results for the chessboard plane recorded for different object distances a_L. Here the blue asterisks show the standard deviation of the depth received from a rigid Raytrix R5 camera without any SfM. Those measurements conform quite well to previously made measurements in (Zeller et al., 2014a). The red crosses show the depth standard deviation after the light-field based visual odometry with a translation of 20 cm. Here the measured standard deviation at an object distance of $a_L = 3.79$ m represents an outlier where the tracking did not work perfectly. This can be assumed when looking at the depth map behavior over the sequence of frames, as shown in Figure 9. Here, the standard deviation is not a constantly descending function, but has maximums and minimums in between which probably come from imperfect tracking.

Nevertheless, the graph in Figure 8 shows, that the depth accuracy of a focused plenoptic camera can be extremely improved by SfM. The red dashed line represents the theoretical depth standard deviation for a stereo camera pair with baseline distance of 20 cm, intrinsic parameters similar to the Raytrix camera, and a disparity standard deviation of 0.3 pixel. This curve can be calculated from eq. (12). Thus, the measured values conform to the theoretical limits.

A clear description for the good measurements at object distances of around 5 m could not be found. One explanation might be, that the effective spatial resolution in the image for far away objects is higher than for close objects and thus the disparity can be measured more accurate.

7. CONCLUSION

In this paper we improved the accuracy of depth information by adapting a monocular visual odometry algorithm to work on image sequences of a focused plenoptic camera. We achieve considerable improvements both with respect to 3D data from light-field only and 3D data from visual odometry only. The main improvement compared to monocular visual odometry is that we were able extend operability of SLAM algorithms towards smaller FOVs and respectively larger focal lengths. Furthermore, tracking is more stable and the depth estimation converges faster. Another main improvement is that our plenoptic camera based visual odometry also measures scale and thus metric tracking and mapping is possible.

Compared to a static plenoptic camera the depth accuracy could be increased by an order of magnitude. Especially for large object distances the improvement is such that depth information can now reliably be used for object segmentation. By this we were able to extend the range of operation of a plenoptic camera towards larger object distances.

Compared to a standard camera, the hardware effort stays pretty much the same, since the plenoptic camera differs only by the MLA in front of the sensor. The computational effort however increases compared to visual odometry since light-field processing has to be done. However, since GPUs are already capable to

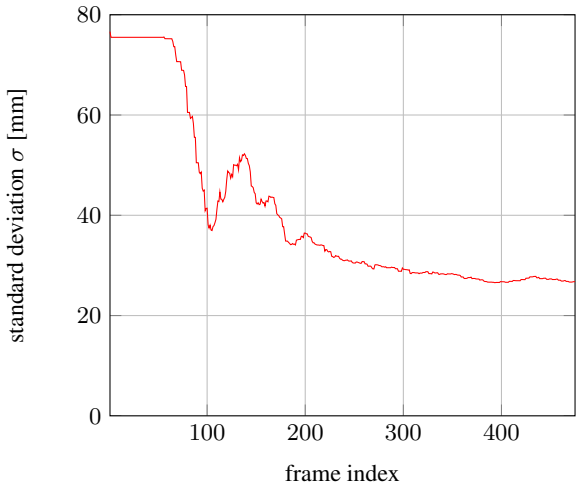

Figure 9. Standard deviation of measured depth for a chessboard target in 3.79 m distance to the camera. Tracking errors resulted in an not constantly descending standard deviation.

calculate depth from light-field images with high frame rates, it is already possible to have plenoptic camera based SLAM systems which run in real time.

REFERENCES

Akbarzadeh, A., Frahm, J.-M., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Merrell, P., Phelps, M., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewenius, H., Yang, R., Welch, G., Towles, H., Nister, D. and Pollefeys, M., 2006. Towards urban 3d reconstruction from video. In: *Proc. Third International Symposium on 3D Data Processing, Visualization, and Transmission*, pp. 1–8.

Concha, A. and Civera, J., 2014. Using superpixels in monocular slam. In: *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pp. 365–372.

Dansereau, D., Mahon, I., Pizarro, O. and Williams, S., 2011. Plenoptic flow: Closed-form visual odometry for light field cameras. In: *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4455–4462.

Eade, E. and Drummond, T., 2009. Edge landmarks in monocular slam. *Image and Vision Computing* 27(5), pp. 588–596.

Engel, J., Schöps, T. and Cremers, D., 2014. Lsd-slam: Large-scale direct monocular slam. In: *Proc. European Conference on Computer Vision (ECCV)*, pp. 834–849.

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

Forster, C., Pizzoli, M. and Scaramuzza, D., 2014. Svo: Fast semi-direct monocular visual odometry. In: *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15–22.

Gortler, S. J., Grzeszczuk, R., Szeliski, R. and Cohen, M. F., 1996. The lumigraph. In: *Proc. 23rd annual conference on computer graphics and interactive techniques, SIGGRAPH*, ACM, New York, NY, USA, pp. 43–54.

Ives, F. E., 1903. Parallax stereogram and process of making same.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A. and Fitzgibbon, A., 2011. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In: *Proc. 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, ACM, New York, NY, USA, pp. 559–568.

Kerl, C., Sturm, J. and Cremers, D., 2013. Dense visual slam for rgb-d cameras. In: *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2100–2106.

Klein, G. and Murray, D., 2007. Parallel tracking and mapping for small ar workspaces. In: *Proc. IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, Vol. 6, pp. 225–234.

Klein, G. and Murray, D., 2008. Improving the agility of keyframe-based slam. In: *Proc. European Conference on Computer Vision (ECCV)*.

Li, M. and Mourikis, A. I., 2013. High-precision, consistent ekf-based visual-inertial odometry. *The International Journal of Robotics Research* 32(6), pp. 690–711.

Lippmann, G., 1908. Epreuves reversibles. photographies integrales. *Comptes Rendus De l'Academie Des Sciences De Paris* 146, pp. 446–451.

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

Newcombe, R. A. and Davison, A. J., 2010. Live dense reconstruction with a single moving camera. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1498–1505.

Newcombe, R. A., Lovegrove, S. J. and Davison, A. J., 2011. Dtam: Dense tracking and mapping in real-time. In: *Proc. IEEE International Conference on Computer Vision (ICCV)*.

Ng, R., 2006. Digital light field photography. PhD thesis, Stanford University, Stanford, USA.

Perwaß, C. and Wietzke, L., 2012. Single lens 3d-camera with extended depth-of-field. In: *Proc. SPIE 8291, Human Vision and Electronic Imaging XVII*, Burlingame, California, USA.

Schöps, T., Engel, J. and Cremers, D., 2014. Semi-dense visual odometry for ar on a smartphone. In: *Proc. IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 145–150.

Venkataraman, K., Lelescu, D., Duparre, J., McMahon, A., Molina, G., Chatterjee, P., Mullis, R. and Nayar, S., 2013. Picam: An ultra-thin high performance monolithic camera array. *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia 2013* 32(6), pp. 1–13.

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

Zeller, N., Quint, F. and Stilla, U., 2014b. Kalibrierung und genauigkeitsuntersuchung einer fokussierten plenoptischen kamera. In: *34. Wissenschaftlich-Technische Jahrestagung der DGPF (DGPF Tagungsband 23 / 2014)*, Vol. 23, HafenCity Universität, Hamburg, Germany.

Zeller, N., Quint, F., Zangl, C. and Stilla, U., 2014c. Edge segmentation in images of a focused plenotic camera. In: *Proc. 11th International Symposium on Electronics and Telecommunications*, Universitatea Politehnica, Timisoara, Romania, pp. 269–272.