Acquisition system for dense lightfield of large scenes

Matthias Ziegler, Ron op het Veld, Joachim Keinert, Frederik Zilly

Fraunhofer Institute for Integrated Circuits IIS,
Am Wolfsmantel 33, 91058 Erlangen, Germany

ABSTRACT

Capturing high resolution and high density lightfield is classically done using precise gantry systems and a DSLR camera. The overall baseline of available systems is small: Scene realism and change in perspective is consequentially limited. This work presents a system for acquisition of dense lightfield of large scenes using precise linear axes and a high quality camera. In contrast to former systems, our presented system can capture lightfield from natural scenes with dense sampling and significant change in perspective. Width and height of the scene can be several meters. Furthermore, for calibration of captured images, we propose a novel self-calibration method. The obtained data may serve as ground-truth reference images for evaluation of lightfield reconstruction methods, novel view synthesis algorithms and many more.

Index Terms **—** Lightfield acquisition, camera calibration, self-calibration, rectification

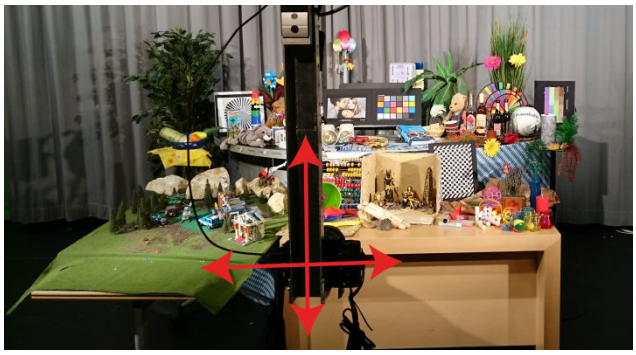

Figure 1: Camera and scene setup. The gantry system is partially visible in the foreground with the camera pointing to the scene.

1. INTRODUCTION

The advent of lightfield into the industry made many researchers develop more efficient techniques for lightfield processing. Big corporations like Google, Lytro and Samsung have presented first market ready products for acquisition and playback.An important aspect in this context concerns generation and acquisition of high quality testdata. In case of VR applications this data should have significant change in perspective and is typically captured using multiple cameras. Applications like plenoptic capture require footage that has high overlap and a very dense spatial sampling. Although different cameras are available, their respective optical configuration is fixed and simulation of different configurations is not possible. Testing and improving many types of algorithms also benefits from dense sampling. This allows testing depth estimation algorithms, comparing image- and depth-image based rendering (IBR, DIBR), comparing virtual views to a ground-truth reference and many more. Additionally, the influence of camera baseline can be evaluated.

Such testdata can be generated with 3D rendering software like Blender. However, such artificial images do not reproduce the properties of natural images and available datasets [1] are of limited resolution (9x9 views, 512x512 pixel). Another option is to use precise mechanical setups like sliders or gantries. In this case the scene in front of the camera is typically small as dimensions of such systems are limited [1], [2].

Based on our knowledge neither system nor dataset is available that combines high resolution imagery, dense 2D sampling and significant change in perspective for natural scenes. In this work we present a system for acquisition of dense lightfield at large scale. The system consists of two industrial cantilever axis and a high resolution mirrorless camera. The camera can be positioned within a range of 4m horizontally and 0.5m vertically. Repositioning error is only 80µm. A new, self-calibration algorithm is used to align captured images. In addition a set of 9 new, dense datasets is published.

The paper is structured as follows: In section 2 we will briefly review current datasets followed by on our gantry system and the calibration in section 3. Section 4 will evaluate our proposed calibration and describe the dataset.

2. PREVIOUS WORK

Various multi-camera datasets for different types of image processing have been presented in literature. The intended and tested applications include stereo-matching, DIBR, IBR and epipolar imaging. Many of them are available online.

Lumsdaine and Georgiev [3] provide a series of high resolution images captured with a plenoptic camera. Each microlens image shows only a small detail of the overall scene. When it comes to novel view synthesis, these datasets are limited as change in perspective is limited. Significant change in perspective requires camera systems with larger baseline.

For pairs of cameras stereo matching is a key for novel views with high quality. The well-known Middlebury database [4] provides stereo image pairs with ground truth disparity maps. This works well for evaluation of stereo matching algorithms. Testing novel view algorithms is possible but limited as only two views are available and parts of the scene are occluded.

A dense, one dimensional sampling of several outdoor scenes is presented by Kim et al. in [5]. In their setup the authors use a 1.5m slider. They show how this data can be used for scene reconstruction. Novel view synthesis with relevant change in perspective is feasible but only along one dimension.

A two dimensional system build from Lego bricks was presented online [6] and can carry a DSLR camera. While not explicitly mentioned, the images allow deriving an approximate baseline of 10-15cm. Images have been calibrated as presented in [7]. Strictly speaking, sampling positions are not on a strict grid. In consequence, estimating and merging disparity maps across many views is more complex.

A similar system is proposed in [2]. The reported repositioning error is in the micrometer range. As before exact physical dimensions are not given but should be within 10-15 cm.Besides real imagery, computer generated datasets are available. They can be generated using 3D rendering software like Blender. Wetzstein [8] provides sets of 5x5 views and 7x7 views on his website. His datasets features transparent objects, specularities and variations of Depth-of-Field (DoF). Honauer et al. [1] recently presented a similar dataset in combination with ground-truth depth information. As mentioned before, those synthetic scenes clearly differ from natural scenes.

Besides high quality images, precise and efficient image processing requires calibrated cameras. To obtain calibration information for systems such as the proposed one, classical approaches use calibration objects such as checkerboards in the scene. However, having such an object in the scene is not always desired. Another method proposed by Vaish et al. [7] requires a reference plane in the scene. Alternatively this plane can also frame the scene. For large scenes, this is challenging since the objects forming the reference plane need to be aligned precisely and preferably should be positioned in the background of the scene. In their work, Wanner et al. [2] calibrate the intrinsic camera parameters beforehand using a calibration chart. The authors point out that the positioning accuracy of the electrical drives actually surpasses the pattern based calibration.

In this context our proposed method goes beyond this, as it determines the camera's intrinsic parameters purely image based, without relying on a calibration chart.

3. LIGHTFIELD CAPTURE SYSTEM

3.1 Mechanical setup

The robot system we present is constructed from two industrial cantilever axes. The 4m horizontal axis is mounted along a standard supporting structure in a height of about 2.4m. The vertical axis is mounted on the slider of the horizontal axis and can translate 50cm. Detailed information on the axes can be found in [10]. Both axes are equipped with electric drives and controllers for positioning. The design allows for a repeat accuracy of 80µm for the horizontal axis and 50µm for the vertical axis. The acceptable payload of the system is about 1kg and allows mounting a high quality camera and matching lens.

Since the camera mount is stable, only a slight but constant uncertainty in the cameras' orientation remains during the acquisition process. This deviation needs to be determined in an offline calibration process.

3.2 Scene configuration

We wish to capture lightfield that feature dense sampling as well as large parallax. These goals are contradictive and a compromise has to be made. In this section we derive a model that brings together camera and scene properties. Due to constraints like scene dimension we manipulate our configuration to take care of all constraints.

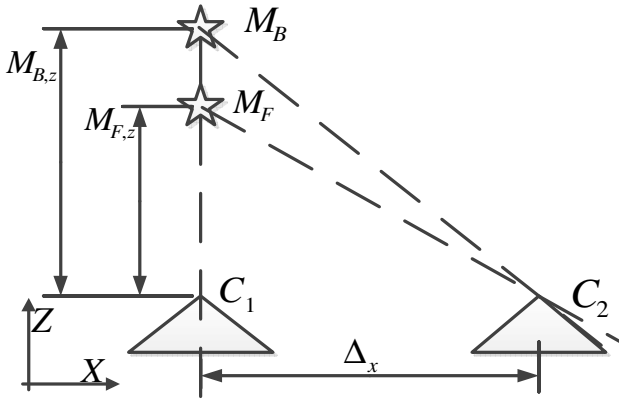

Figure 2: A stereo camera pair with parallel orientation. Both world points M_B and M_F are exactly in front of C_1. In the image of C_1, M_B is occluded by M_F. In the image of C_2 both points are visible.

Figure 2 shows a schematic model of the scene seen from above. Points M_F and M_B denote world points closest and farthest from the acquisition plane Z_0=0, respectively. Camera 1 is in the origin of the coordinate frame.

Projective geometry eases computation of the horizontal pixel coordinate u of a 3D-point $M=(M_x, M_y, M_z, 1)^T$. M is being projected on the image plane as given by eq. (1). s_p denotes the size of a pixel. The maximum amount of occlusion is computed from the distance between M_F and M_B in the projected image of C_2 as given in eq. (2). with $M_x=\Delta_x$.

$$u = \frac{1}{s_p} \cdot \frac{f \cdot M_x}{M_z} = C_0 \cdot \frac{1}{M_z} \text{ with } C_0 = f \cdot \frac{M_x}{s_p} \quad (1)$$

$$d = u_1 - u_2 = C_0 \cdot \frac{M_{F,z} - M_{B,z}}{M_{F,z} \cdot M_{B,z}} \quad (2)$$

The quantity *d* is typically known as scene parallax and may be given in pixel or as relative quantity with respect to the image width.

Since many computer vision algorithms require images with high DoF this is an important aspect and should be taken into account in terms of scene and camera configuration. The scene should be configured such that all objects in the scene are in focus. The hyperfocal distance D_{Hyp} of a lens with focal length *f*, an aperture value N and a given diameter c of the circle of confusion can be given as:

$$D_{Hyp} = f + \frac{f^2}{Nc} \approx \frac{f^2}{Nc} \quad (3)$$

With the focal distance s, the near-limit distance D_N and the far-limit distance D_F can be given as:

$$D_N = \frac{D_{Hyp} \cdot s}{D_{Hyp} + s} \quad (4) \qquad D_F \approx \frac{D_{Hyp} \cdot s}{D_{Hyp} - s} \text{ for } s < H \quad (5)$$

For a given scene and lens we need to determine the required aperture value N. This can be approximated as:

$$N = \frac{f^2}{c} \cdot \frac{D_F - D_N}{2 \cdot D_F \cdot D_N} \quad (6)$$

Repositioning accuracy ε of our system is given as 80µm. We need to ensure that ε is low compared to the sampling distance: $\varepsilon \ll \Delta$. This requirement may also be expressed in terms of projected points according to eq. (1) as:

$$\tilde{\varepsilon} = \frac{1}{s_p} \cdot \frac{f \cdot \varepsilon}{M_z} \quad (7)$$

In our opinion it is sufficient if $\tilde{\varepsilon} < 0.5px$ since this is lower than the cameras' optical resolution.

Based on these considerations we can plan and setup the scene: The camera available and compatible with our software is a Sony Alpha 7RII camera (7952x5304px) and a 50mm lens.

Due to the Bayer-pattern the true image resolution is only half the resolution of the sensor. Then, we get s_p = 9.06 µm. With Formula 7 solved for M_z we can compute the minimum scene distance $M_{F,z} \approx$ 1.77m. Figure 1 shows one of the scenes we setup in our studio. It is built from various natural objects that feature fine details as well as highlights and

rich color. A reasonable scene width we could setup is about 2.4m at the back in combination with a total depth of about 3m.

Then, $M_{B,z} \approx M_{F,z} + 3m = 3{,}77m$ and N=15.3 using c=0.029mm being a common value for the circle of confusion for the given sensor size. Finally, we needed to limit ourselves to a sampling density of 4mm horizontally and 6mm vertically with 101x21 sampling positions resulting in 2121 images per dataset. This was related to the number of images we could capture within one workday. Then, maximum achieved parallax from left to right is 20.8%.

3.3 Post processing

Images have initially been captured in RAW format and subsequently been processed in several steps in order to obtain high quality images:

1. Debayering
2. Correct radial distortion and chromatic aberration
3. Color correction
4. Shading
5. De-Flickering
6. Rectification
7. Downsampling

Steps 1 and 2 have been performed using Adobe Photoshop with the appropriate lens profile. In step 3, white- and blackpoint of the scene have been selected. In Step 4, correction of lens-specific shading and other distortions like small particles on the lens have been applied. Step 5 matched global luminance using one reference image. This step was necessary as global luminance of individual images slightly varied from view to view, possibly caused by slight flickering of the employed lights. Rectification in step 6 ensures horizontal and vertical alignment of the images, which will be further explained in section 3.4. Step 7 finally resampled the images from full resolution to half.

In all processing steps, the cameras dynamic range of about 14 f/stop has been preserved. Using PNG compression each dataset comes to about 100GB. Transferring the whole dataset over internet would be at significant time and costs. Hence, we decided to compress the data using HEVC with very high quality settings (CRF 8, no chroma subsampling, 12-Bit). When traversing the dataset in a meander like fashion (left to right and back to left in the line below), this type of data is excellently suited for a video codec like HEVC. Like this, each dataset is reduced to about 1GB relating to a compression ratio of about 1:100.

3.4 Calibration

As discussed in section 3.2, scene and camera configuration are designed such that positional inaccuracy of the capturing system can be neglected with no need of optimization in a calibration algorithm. Moreover, due to the rigid mounting of the camera in the gantry system, orientation of the camera remains constant during the acquisition process.

In case of a rectangular sampling, as carried out with the gantry system, we like to ensure that adjacent images are aligned horizontally and vertically. Mathematically, these requirements can be formulated as two optimization problems that need to be solved simultaneously:

$$\underset{H}{\operatorname{argmin}} \sum_{l} \left(v\left(H \cdot m_{hor,l}\right) - v\left(H \cdot m_{hor,l}{}'\right) \right)^2 \tag{8}$$

$$\underset{H}{\operatorname{argmin}} \sum_{l} \left(u\left(H \cdot m_{ver,l}\right) - u\left(H \cdot m_{ver,l}{}'\right) \right)^2 \tag{9}$$

In Formula 8 and 9 H is a 3x3 homography matrix, m_{hor} and $m_{hor}{}'$ denote horizontally matching features. m_{ver} and $m_{ver}{}'$ denote vertical matches. l denotes the number of horizontal and vertical matches, respectively. *u()* and *v()* extract the horizontal / vertical component of a feature pair.

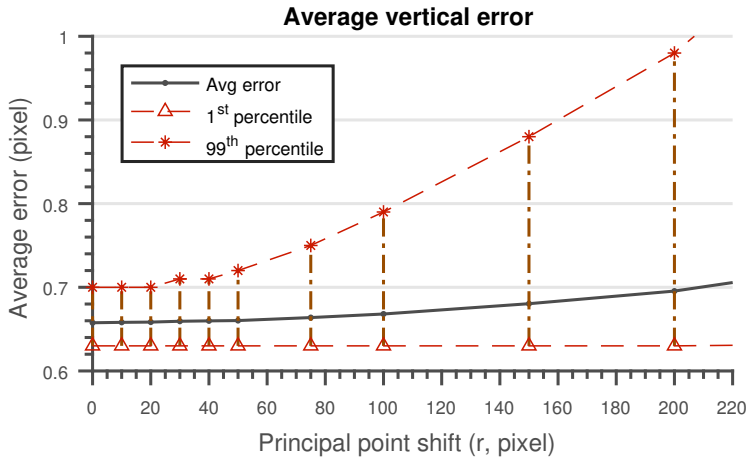

Figure 3: Average error after rectification on synthetic data along a simulated shift in principal point (PP). For PP-shift < 50px, over 99% of all probes yield average error values of 0.72 px or smaller.

H is constructed from a rotation matrix R and a camera matrix K as $H = R \cdot K$. The homography rotates images such that the optical axis is perpendicular to the sampling plane and that the x and y axes of the images are in parallel with the coordinate system of the gantry system. The camera matrix K is defined in equation 10:

$$K = \begin{bmatrix} f & s & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix} \tag{10}$$

As focal length of the camera is constant during acquisition no optimization is required. That is why we set f=1. We may further set s=0 as the camera has square pixels. Therefore only the camera's principle point (PP) and its orientation remain unknown. However, if the camera's deviation in its orientation as well as the deviation in its PP is small, calibration can ignore the PP deviation and only optimizes for three angular parameters.

Pairs of matching featurepoints can be obtained from two images in the same line or column using algorithms like SIFT/SURF [10] or similar algorithms. A choice that is open concerns the baseline between image pairs used for calibration. For simplicity we subsample the dense 21x101 images to obtain a 2x3 array.

4. EXPERIMENTAL EVALUATION

4.1 Calibration accuracy

We evaluate our proposed calibration method by simulation as well as on our real data. For the simulation, parameters are set such that the simulated system matches the real system as written in section 3. By projection, we obtain precise feature pairs from simulated 3D points. Uncertainty due to positional inaccuracy is modeled using additive Gaussian noise (AGN) with std. $\sigma_C = 80\mu m$. Additional noise in feature pairs is also modeled using AGN with std. $\sigma_p = 0.5$px. For the camera's orientation we assume that the maximum deviation is within 5° with Gaussian distribution (σ_R=2.5°). For synthetic evaluation, we simulate deviations of PP. We set $p_y = \sin(\varphi) \cdot r$ and $p_x = \cos(\varphi) \cdot r$. φ is uniformly selected from $0 \ldots 2\pi$. The simulation probes values of r. Figure 3 shows the remaining, averaged error per pixel. This error is within 0.66-0.71px for the evaluated range. In addition, we plot curves for the 1th and 99th percentile.
The lower curve marks the 1st percentile; the upper one marks the 99th percentile. I.e. for r=50px, 99% of all probes yielded an average error of about 0.72px or less.

For most applications, this should be acceptable. This also shows that optimization of the PP is not required if its deviation from the camera center is small. For high quality cameras, such as the employed one, this should be the case.

In case of real data the remaining error for our datasets is about 0.75px. Compared to the synthetic results this is slightly higher but still below 1px and in our opinion, acceptable for many applications.

Figure 4: Sample images from 9 datasets captured and processed with the presented acquisition system. The scenes contain objects with fine details as well as specular and transparent objects.

4.2 Dataset description

So far, we have captured 9 dense datasets we would like to share with the scientific community. 8 of the datasets comprise 21x101 views, one comprises 21x99 views. Image processing including our proposed self-calibration has been carried out as described in section 3. Figure 3 shows one sample image for each of the datasets. The scenes contain various different objects like plants with thin structures, translucent objects, objects with strong depth-discontinuities like charts and also glasses or metallic objects causing reflections and highlights.

Figure 5: A false color disparity map as overlay on one of the captured and provided scenes.

The video streams can be downloaded at: **www.iis.fraunhofer.de/en/ff/bsy/tech/lichtfeld.html**

As an example application, Figure 5 shows the false color version of a disparity map as an overlay on one of the captured scenes. Especially the plant in the foreground is challenging in wide-baseline stereo matching methods.

5. CONCLUSION

In this work we presented our robot system for capturing dense lightfield of large scenes. The system consists of two industrial cantilever axes with electrical drives and a high quality camera. The specified repositioning error of the system is only 80µm.

Selecting views with small baseline allows testing and evaluating algorithms intended for plenoptic cameras. Selecting larger baseline between views allows testing stereo matching and multi-camera disparity estimation. Novel view synthesis methods can possibly be evaluated by comparison to ground truth. Calibration is performed using a novel self-calibration algorithm. A set of 9 datasets has been captured and published

This work was supported by the Fraunhofer and the Max Planck cooperation program within the framework of the German pact for research and innovation (PFI).

The work in this paper was funded from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging

6. REFERENCES

[1] K. Honauer, O. Johannsen, D. Kondermann, B. Goldluecke, "A Dataset and Evaluation Methodology for Depth Estimation on 4D Light Fields" in Asian Conference on Computer Vision (pp. 19-34), Springer, 2016

[2] S. Wanner, S. Meister, B. Goldluecke, "Datasets and benchmarks for Densely Sampled 4D Light Fields", in VMV, pp. 225-226, 2013, http://klimt.iwr.uni-heidelberg.de/HCI/Research/LightField/lf_benchmark.php

[3] A. Lumsdaine, T. Georgiev, "Full resolution lightfield rendering", Indiana University and Adobe Systems, Technical Report, 2008, www.tgeorgiev.net

[4] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth", in *Pattern Recognition* (pp. 31-42). Springer, 2014

[5] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, "Scene reconstruction from high spatio-angular resolution light fields" in *ACM Trans. Graph.*, *32*(4), 73-1, 2013

[6] http://lightfield.stanford.edu/index.html

[7] V. Vaish, B. Wilburn, N. Joshi and M. Levoy, "Using plane + parallax for calibrating dense camera arrays" in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on (Vol. 1, pp. I-2). IEEE.

[8] http://web.media.mit.edu/~gordonw/SyntheticLightFields/index.php

[9] https://www.festo.com/cat/de_de/data/doc_engb/PDF/EN /EGC-TB_EN.PDF

[10] Lowe, David G. "Distinctive image features from scale-invariant keypoints." *International journal of computer vision* 60.2 (2004): 91-110.