59th ILMENAU SCIENTIFIC COLLOQUIUM
Technische Universität Ilmenau, 11 – 15 September 2017
URN: urn:nbn:de:gbv:ilm1-2017iwk-054:6

FPGA implementation of a multi-view stereo approach for depth estimation and image reconstruction for plenoptic cameras

M. Hänsel, M. Rosenberger, G. Notni

Technische Universität Ilmenau, Group for Quality Assurance and Industrial Image Processing

ABSTRACT

In this paper a concept for an algorithm for depth estimation and image reconstruction for a plenoptic camera is presented. The algorithm follows a multi-view stereo approach and is intended for an FPGA-based Xilinx Zynq Ultrascale+ SoC platform to allow for real-time processing in an embedded environment. The micro-lens array separates a complete image in many micro-images. The micro-images are considered as individual cameras and the processing is calculated in a multi-view stereo approach. To accomplish an adequate frame rate and a reasonable resolution efficient processing steps and fixed-point integer calculation are chosen. The conceptual algorithm will be implemented and tried out in an experimental setting in 2019.

Index Terms – plenoptic camera, depth estimation, FPGA, embedded, multi-view stereo

1. INTRODUCTION

A recent trend in image processing is the reconstruction of 3d geometry of a scene with means of 2d CMOS active pixel sensors. Applications of 3d image processing reach over a great range of different fields, whether they are academic (e.g. 3d measurement of archaeological finds), industrial (e.g. quality assurance) or artistic (e.g. 3d movies).

The techniques applied for 3d image reconstruction are various and can loosely be divided into passive and active methods. Active methods require a specific and well-defined illumination setting, e.g. an array of equidistant spots. Passive methods include the classical stereo approach, sometimes utilizing more than two cameras (multi-view stereo), and movement-based methods. A more or less recent development in the field of passive 3d image acquisition are light-field cameras with microlens arrays. The optical system does not consist of the objective lenses in front of the camera only but also a microlens array which is placed directly on top of the active pixels of a CMOS sensor chip (as illustrated in Figure 1). Because of the microlenses the acquired image is divided into smaller micro-images, one for each microlens. If the system is set up correctly the micro-images show overlapping contents of the scene which allows for triangulation and by that the estimation of the depth (distance between camera and object).

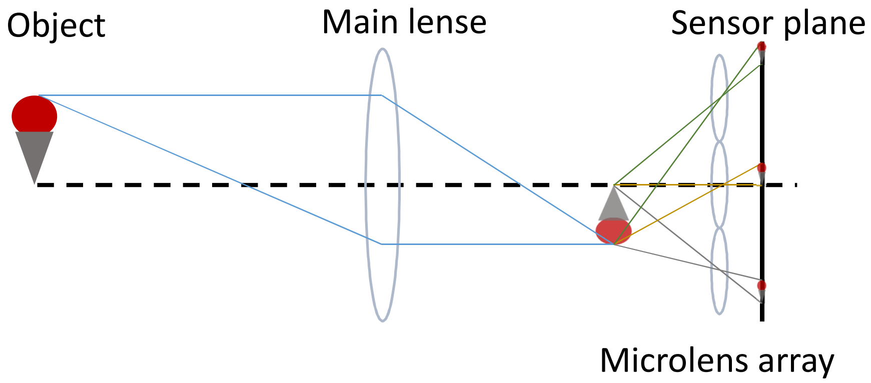

Figure 1: Simplified 2d-model for a plenoptic camera. The illustration is not true to scale. Modified image from [1].

At the moment the reconstruction requires a high-end graphics card to run in real-time. In the project "Light-field camera-based Embedded Sensor Platform" TU Ilmenau will implement the image reconstruction for a Raytrix R8 camera in a Xilinx Zynq Ultrascale+ board. The project partners in parallel develop software applications for ambient assisted living and road condition measurement. Both fields are economically lucrative and would profit from depth information to extend the capabilities of the up to now 2d-based systems. In this paper the focus will lie on the implementation of 3d reconstruction of the scene geometry rather than the applications.

The depth calculation is analogous to multi-view stereo approaches and shares some of their conditions and requirements although the magnitude of their influence is not always identical. In section "Requirements for passive depth estimation" the boundaries of light-field cameras and multi-view-stereo approaches are explained and illustrated.

Although the basic idea for the depth estimation is the same for stereoscopy and for light-field cameras, there are differences in the scope. In the classical stereo approach two images are compared to find matching points for triangulation. In the light-field approach *micro*-images are compared. Because of this the search for the matching points can be simplified. The epipolar lines are more numerous but thinner and different photo-consistency measures can be applied. In section "Detection of matching points" the algorithms for the embedded implementation as well as the reasoning behind their selection are explained. The triangulation, i.e. the estimation of depth on basis of the matching points, requires calibration. The differences of the distance of a pixel to the center of the corresponding micro-image for the various matching points is inversely proportional to the depth. The triangulation and the calibration measurements conducted to transform the disparity in a metric distance are introduced in section "Calibration and triangulation".

The section "Implementation on embedded system" will contain information about the way the depth estimation Is intended to be implemented on embedded platform. The computation of the search for matching points as well as the triangulation will be performed in the FPGA-based programmable logic (PL) of the Zynq Ultrascale+ system. To satisfy real-time criteria the implementation will have a pipeline structure. The resulting depth map is a point cloud and does not provide a depth value for every pixel in the calculated image. To close the spaces between the points a filling algorithm will be applied. The resulting depth map will be saved in shared DDR memory which is accessible from the processing system (PS) of the Zynq. This way the project partners are provided with the 3d data required for their applications.

Utilizing the depth map an all-in-focus image with extended depth of field can be calculated. Due to the redundancy in the micro-images the (lateral) resolution of this image is reduced compared to a plain 2d image taken with the same sensor chip but without the micro-lenses. On the other hand, this is counter-balanced since the increased depth of field results in less blurring caused by defocus.

For the end of the project late 2019 a prototype for one or both of the embedded applications is announced.

2. REQUIREMENTS FOR PASSIVE DEPTH ESTIMATION

Various conditions have to be met to provide a basis for accurate passive depth estimation with the multi-view stereo triangulation approach.

Most importantly the images, or rather micro-images in this case, need overlapping content. Without overlapping content triangulation is impossible. Light reflections are required to be diffuse rather than specular. Furthermore, the images need some kind of contrast or features. This could be for example edges or inhomogeneous textures. Obviously the images need to be sharply focused by the lens system. Otherwise the contrast in the scene is lost in the image. For homogenous surfaces (random) light-spot illumination can provide contrast for triangulation. The downside to that is that the texture is not preserved. The spots are visible in the image and cannot be removed without a priori knowledge. If the texture is not of interest active depth estimation approaches (e.g. light section) could be more suitable.

If the textures are periodic, matching points might be ambiguous. This problem is less pronounced for plenoptic cameras than for classical (multi-view) stereo settings since the micro-images are usually much smaller than the full images of even low-resolution cameras. Fine periodic textures can lead to erroneous triangulation nonetheless.

3. DETECTION OF MATCHING POINTS

The search area for matching points can be limited to the epipolar lines. In conventional (multi-view) stereo vision the epipolar lines have to calculated beforehand or the images have to be rectified. Plenoptic cameras do not require micro-image rectification, since the micro-images are all sampled from various very small subsections of the same sensor plane. Only neighboring subsections are considered to show overlapping image content and main lens objective can reasonably be assumed to not significantly distort neighboring micro-images.

The target of the depth estimation and image reconstruction algorithm is a Raytrix R8. Raytrix plenoptic cameras utilize a micro-lens array with hexagonal tiling. Every micro-image has up to six neighboring micro-images (see Figure 2). Since the system is set up in a way that the overlap of neighboring micro-images is at most 50% for every point in a micro-image, up to 3 matching points in adjacent micro-images can be found. As mentioned before, the search area is limited to the epipolar lines. The applicable epipolar lines for a hexagonal grid are illustrated in Figure 2.

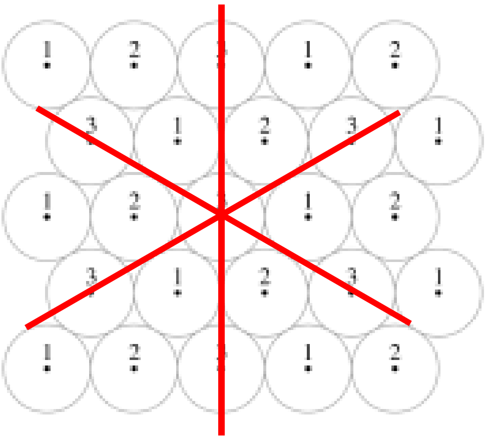

Figure 2: Hexagonal micro-lens for Raytrix lightfield cameras.The micro-lenses are from 3 different lens types, which are designated by the numbers 1 to 3. The red lines exemplarily illustrate the epipolar lines for a pixel near the center of a micro-lens of type 3. Modified image from [2].

To detect matching points at least one photo-consistency measure has to be applied. Various photo-consistency measures have been published. An overview over the most common ones can be found in [2]. In the context of embedded real-time processing with FPGA-based programmable logic SAD (sum of absolute differences) and SSD (sum of squared differences) are the most suitable because of their low computational complexity. More complex measures, e.g. NCC (normalized cross-correlation), offer more robustness against differences in lighting and exposure conditions. This robustness is required for multi-camera settings but not for a plenoptic camera. Lighting and exposure are virtually identical for all adjacent micro-images. The optimal size of the support grid for SAD and SSD will be determined empirically. Larger support grids might increase the overall accuracy of the matching point detection. On the other hand, they are computationally more expensive and result in more flawed boundary pixels. Since every micro-image has its own borders, this can accumulate to large loss of information in the image.

When matching points are detected, their different positional offsets are put in relation. The differences in pixel coordinates of the matching points, the disparities, are saved in a matrix with the same dimensions as the images, the disparity map. For standard stereo approaches usually one image is the master image. The disparity of two matching points is saved in the disparity map at the indices of the point in the master image.

For multi-view stereo the disparity is in some way derived from partial disparities between the master image and each of the other images. Possible methods are e.g. the mean of the partial disparities or a majority vote (requires numerous sub-images/cameras).

In the approach for plenoptic cameras, the disparity will be calculated in relation to the distance of the micro-image centers. E.g. a pair of matching points that are each in the centers of their respective micro-images get a (partial) disparity value of ‘0’ assigned. The total disparity is the mean value of the partial disparities of matching point pairs. Each micro image is considered master image once. The averaged disparity values are saved at the pixel coordinates of the macro-image.

4. CALIBRATION AND TRIANGULATION

After the detection of matching points, you get, depending on the applied photo-consistency measure and the scene, a more or less densely populated disparity map. Pixels without a valid disparity can be interpolated with a filling algorithm.

Disparity is inversely proportional to distance. Points close to the camera have a high disparity, points far away from camera have low disparity value (as illustrated in Figure 3). By calibration each disparity value can be mapped to a corresponding distance or depth range. The relation between disparity and depth is assumed to be inversely linear proportional. For this a scene with a known geometry and defined camera alignment is required. Obviously, the scene also has to satisfy the requirements presented in the previous section of this paper.

A particular mapping is only valid for a specific setting of main objective focus and aperture. If the setting is changed, an update to the calibration is necessary. If the disparity was not interpolated with a filling algorithm, this could instead be done on the depth map.

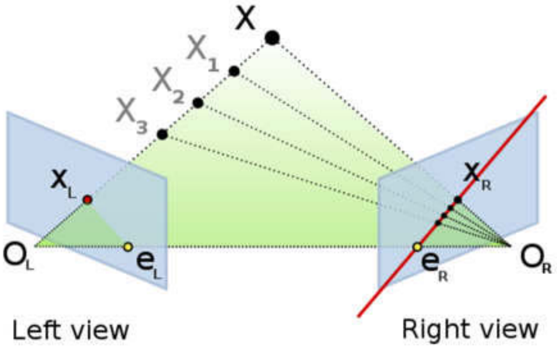

Figure 3: Illustration of the relation between disparity and distance. In this example the left view provides the master image and the red line depicts the epipolar line. Image taken from [3].

5. IMPLEMENTATION ON EMBEDDED SYSTEM

The matching point detection and calculation of a depth map is intended to be implemented in the programmable logic of a Zynq Ultrascale+ device, which is a SoC combining FPGA-based programmable logic (PL), an ARM-based processing system (PS) and a Mali GPU. PS and PL are connected via various interfaces, e.g. direct GPIO connections and shared DDR4 RAM. The combination of these different architectures supports the combination of diverse programming paradigms, implementation approaches and applications.

Filling and image reconstruction are probably to be calculated by the application processor unit, which will be operated by an embedded GNU/Linux. Depth maps and raw images will be saved in shared DDR4 RAM to allow access from both processor system applications and programmable logic. Completion of a depth map for a frame will be signaled to the processing system by an interrupt. That way the depth maps can be saved on a persistent storage in time.

The input data will be streamed to the depth estimation core in the PL. The stream may originate from a camera or a data stream coming from the shared DDR4 RAM. The goal is a pipelined implementation of the depth estimation with same clock rate as the input pixel clock. The computation time for each pixel is intended to be a constant multiple of the pixel clock period.

The buffering of the image data utilizes block RAM to keep the impact on logic slice utilization low. Division by a variable divisor will be avoided, because it is expensive in computation time and FPGA resources. The calculation will be carried out with integers and, if necessary, fixed-point numbers. Floating-point arithmetic is not advantageous in this case.

6. CONCLUSIONS AND OUTLOOK

In this paper a concept for a multi-view stereo approach for depth estimation for plenoptic cameras was presented. Basics and requirements for stereoscopy were introduced. The approach will be implemented on a Zynq Ultrascale+ device – a highly versatile SoC combining FPGA-based programmable logic and an ARM-processing system. The majority of the calculations for the depth estimation are intended to be executed in the programmable logic.

The implementation presented in this paper will be a component for an embedded system for lightfield cameras, which can be applied for different use cases. Partners of TU Ilmenau are currently studying the feasibility for road condition identification and advanced information systems for ambient assisted living.

A prototype system for either or both applications will be presented in late 2019.

7. ACKNOWLEDGMENTS

The research presented in this paper is funded by the German Federal Ministry of Education and Research (BMBF) under the project 01IS16005B titled “Lichtfeldkamera-basierte eingebettete Sensorplattform” (light-field camera-based embedded sensor platform).

SPONSORED BY THE

Federal Ministry
of Education
and Research

REFERENCES

[1] G. Straube, *Funktionsweise plenoptischer Kameras.* [Presentation]. Steinbeis Qualitätssicherung und Bildverarbeitung GmbH, 2017.

[2] Y. Furukawa und C. Hernández, „Multi-view Photo-consistency“ in *Multi-View Stereo: A Tutorial*, Boston; Delft, Now, 2015, pp. 17-37.

[3] A. Nordmann, "Wikimedia Commons: Epipolar geometry.svg" 2007. [Online]. Available: https://commons.wikimedia.org/wiki/File:Epipolar_geometry.svg. [Accessed 7-12-2017].

[4] C. Perwaß und L. Wietzke, „Single Lens 3D-Camera with Extended Depth-of-Field“ 2012. [Online]. Available: https://www.raytrix.de/?ddownload=1709.