Measurement: Sensors 24 (2022) 100425

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

FPGA-based multi-view stereo system with flexible measurement setup

Christina Junger [a,*], Richard Fütterer [a], Maik Rosenberger [a], Gunther Notni [a,b]

[a] Technische Universität Ilmenau, Department of Mechanical Engineering, Group for Quality Assurance and Industrial Image Processing, Ilmenau, Germany
[b] Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany

ARTICLE INFO

Keywords:
multi-view stereo system
FPGA
System-on-Chip devices
camera calibration
Semi-Global matching
point cloud

ABSTRACT

In recent years, stereoscopic image processing algorithms have gained importance for a variety of applications. To capture larger measurement volumes, multiple stereo systems are combined into a multi-view stereo (MVS) system. To reduce the amount of data and the data rate, calculation steps close to the sensors are outsourced to Field Programmable Gate Arrays (FPGAs) as upstream computing units. The calculation steps include lens distortion correction, rectification and stereo matching. In this paper a FPGA-based MVS system with flexible camera arrangement and partly overlapping field of view is presented. The system consists of four FPGA-based passive stereoscopic systems (Xilinx Zynq-7000 7020 SoC, EV76C570 CMOS sensor) and a downstream processing unit (Zynq Ultrascale ZU9EG SoC). This synchronizes the sensor near processing modules and receives the disparity maps with corresponding left camera image via HDMI. The subsequent computing unit calculates a coherent 3D point cloud. Our developed FPGA-based 3D measurement system captures a large measurement volume at 24 fps by combining a multiple view with eight cameras (using Semi-Global Matching for an image size of 640 px × 460 px, up to 256 px disparity range and with aggregated costs over 4 directions). The capabilities and limitation of the system are shown by an application example with optical non-cooperative surface.

1. Introduction

In recent years, stereoscopic image processing algorithms have gained importance for a variety of applications in the field of automation and quality assurance. These include automotive applications [1], integrated assembly solutions [2,3] (turnover gain of 11 % to 7.1 billion euros in Germany, 2021 [4]), holistic monitoring [5] and medical technology [6]. Each application field has its own requirements for the measuring system.

Some applications (e.g. assembly process) require a wide measurement range. To cover this completely, a multi-view stereo (MVS) system can be used. The object(s) or scene is captured by combining several partial surfaces from different perspectives. As the number of stereo systems increases, therefore also rise the challenges of managing data volumes, camera connection requirements and synchronization. To overcome these challenges, sensor close stereo systems based on embedded systems, e.g. Field Programmable Gate Arrays (FPGA), can be used. For the implementation of sensor-close real-time computation of dense disparity maps, embedded systems - based on application-specific integrated circuits (ASICs) [7], FPGAs [1] and graphics processing units (GPUs) [8] - are particularly efficient in terms of low latency and low power consumption compared to self-contained systems. The performance is scalable. ASICs are more efficient than FPGAs, but can no longer be reprogrammed. The first FPGA implementation of dense stereo matching was in 1997 [9]. By using Semi-Global Matching (SGM) [10], the accuracy of FPGA implementations improved significantly [1,11,12]. SGM achieves more accurate results compared to local operators and achieves a higher speed increase than global optimisation methods. However, the disadvantage is the large memory requirement. There are alternative approaches that improve the frame rate but at the expense of accuracy. Nowadays, closely coupled CPU/FPGA system-on-chip (SoC) devices are used for implementation [13].

The advantage of an FPGA-based heterogeneous MVS system is that pre-processing steps and especially the computationally intensive correspondence point search, also called stereo matching, are outsourced to one or more FPGAs reducing the data streams [14]. However, the development effort of a hardware-near processing is very high. In the following is a list of reasons for an FPGA-based system.

(i) By outsourcing calculation steps - especially the computationally intensive stereo matching - to an SoC, the following calculation unit is relieved with regard to camera connection req. and

* Corresponding author.
E-mail address: Christina.Junger@tu-ilmenau.de (C. Junger).

https://doi.org/10.1016/j.measen.2022.100425
Received 9 March 2022; Received in revised form 17 August 2022; Accepted 19 August 2022
Available online 5 September 2022

computing power. This is particularly relevant when several modules are connected.

(ii) The number of data streams is reduced per sensor near processing module. This results in optimum synchronicity and lower requirements on the hardware interfaces of the following computing unit. Due to the additional lower communication between sensor and computing unit, the latency is further reduced.

(iii) The ARM/FPGA sensor module is compact and therefore particularly flexible in its use e.g. in industrial environments. Furthermore, the developed FPGA sensor module is a plug and play system, which makes handling in an industrial environment easier. Furthermore, these systems usually have a lower power consumption.

This paper presents an FPGA-based MVS system by combining a multiple view with eight cameras (Fig. 1). Furthermore, capabilities and limitations of the system are shown by an application example with optically non-cooperative surface.

2. Depth sensing technologies

There are a couple of technologies for optical three dimensional detection of objects or scenes with high lateral resolution at medium distances (m) to long distances (km):

(a) triangulation-based approaches (active and passive stereo methods, light field camera),
(b) runtime-based methods (lidar or time-of-flight, radar, interferometric methods) or
(c) image-based methods with prior knowledge (use of motion, perspective, occlusion).

Special, optimised inspection systems for optical 3D detection must be developed for each application. Depending on the requirements, the systems and measuring arrangement must be selected. When selecting systems, the combination of the number of recorded object points per second and lateral as well as longitudinal resolution with a given surface condition is decisive. The measurement arrangement depends on the scene or object to be captured. Multi-view arrangements are useful for (i) large measurement volumes or for (ii) a capturing of the entirety of an object or scene. Here, all captured partial volumes are registered in a global coordinate system. For case (i), arrangement in a row with multiple views could be used. For case (ii), an arrangement with multiple views around the object would be possible.

This publication presents our FPGA-based passive multi-view stereo system. Passive stereo systems have the advantage that they do not require active lighting (pattern projection). Multiview stereo systems based on passive stereo systems have the great advantage that the measurement setup can be changed flexibly without much effort. A combination of two or more active stereo systems generates an overlay of different patterns (each system has its own pattern). This overlay leads to problems with the assignment of correspondence points in the image pairs. Active stereo systems also have higher energy consumption due to the lighting. On the other hand, a disadvantage of a passive stereo system is that the accuracy of the depth information strongly depends on the stereo matching algorithm used. Traditional stereo macthing algorithms usually have limitations with, for example, texture-poor or optical non-cooperative surfaces [15,16]. Sec. 4.2 describes the calculation of the lateral $\Delta x_{\text{lateral}}$ and longitudinal $\Delta z_{\text{longitudinal}}$ resolution of a stereo camera. Sec. 4.3 describes the depth error ΔZ.

3. Overview of our multi-view stereo system

Fig. 1 shows an overview of our modular FPGA-based multi-view stereo system. With this system, the geometry and texture of a 3D scene can be captured. The captured 3D scenes are displayed as cubes. The MVS system consists of four sensor-related and FPGA-based passive stereo modules ($S1$, $S2$, $S3$, $S4$). Each module (see Sec. 6) contains two cameras cam_A and cam_B. Each stereo module synchronously captures a 3D scene and calculate a disparity map (see Sec. 5.2). The disparity map contains the depth of the scene points, i.e. its distance from the stereo module. This processing unit (Zynq Ultrascale, Sec. 6) synchronizes the

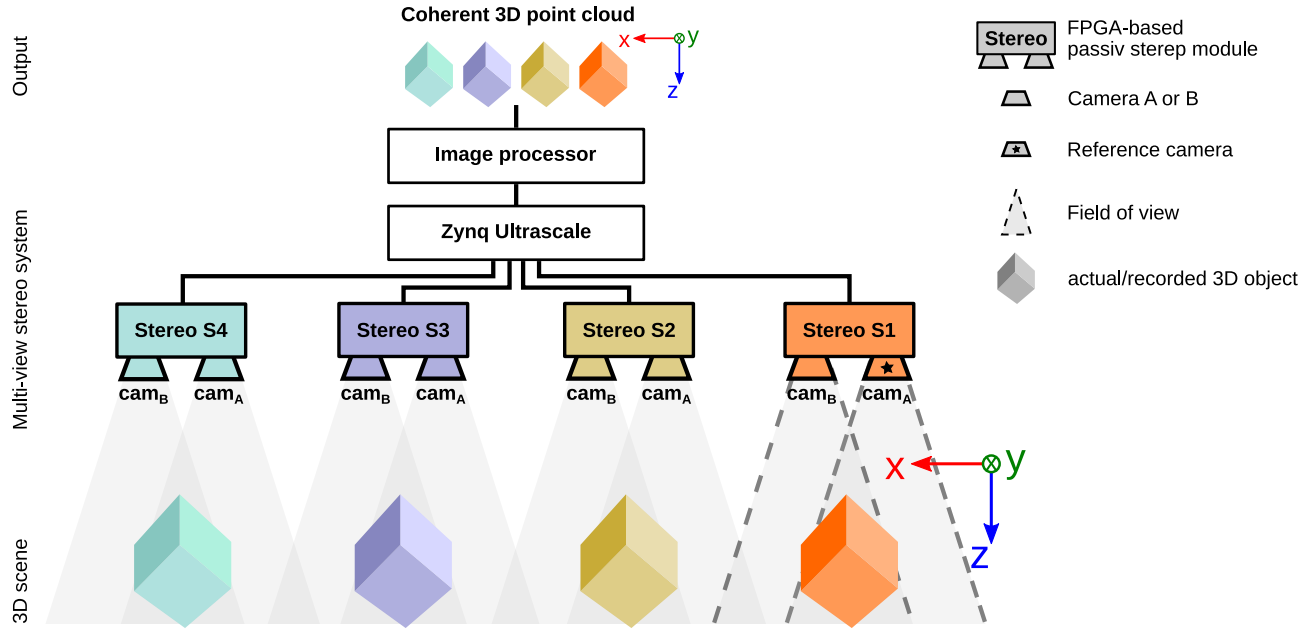

Fig. 1. Overview of the presented multi-view stereo (MVS) system. The MVS system captures a 3D scene (geometric as well as texture) represented as four 3D objects. The system consists of four FPGA-based passive stereo modules ($S1$, $S2$, $S3$, $S4$). Two cameras each cam_A and cam_B are integrated in this module. These modules are synchronized by the downstream processing unit (Zynq Ultrascale). In addition, the result image (disparity map with cam_A image) are merged and forwarded to an image processor. The image processor computes 3D point clouds and registers them to a coherent 3D point cloud. The reference of this 3D point cloud is the cam_A of the stereo module $S1$.

image acquisition of the four stereo modules (*S*1, *S*2, *S*3, *S*4) and merges all four output data (disparity map and image of cam_A). A coherent 3D point cloud is generated from the four partial volumes in a subsequent image processor (see Sec. 5.4). The reference of this point cloud is the cam_A of the sensor module *S*1.

The MVS system has a modular structure. Thus, the number of stereo modules as well as the measuring arrangement can be adapted according to the application. For a simplified representation, a stereo arrangement in a row has been chosen in Fig. 1.

4. Basics of a stereo camera system

4.1. Camera arrangement

Fig. 2 (left) shows a stereo camera system with parallel camera arrangement. The only advantage of a parallel camera arrangement over a convergent array is that fewer resources (memory) are required for pre-processing (see Sec. 5.1, the vertical transformation maps contains smaller values). In FPGA based systems, this is a very important point, since technical limitations (BRAM) are often a limiting factor. Eq. (1) describes the calculation of the angle resolution $\Delta\alpha$, which depends on the pixel width of the sensor w_{pixel} and the focal length f. Eq. (2) shows the horizontal field of view FOV_h with sensor width w_{sensor}.

$$\Delta\alpha = \arctan\left(\frac{w_{\text{pixel}}}{f}\right) \tag{1}$$

$$\text{FOV}_h = 2\cdot\arctan\left(\frac{w_{\text{sensor}}}{2\cdot f}\right) \tag{2}$$

4.2. Derivation of longitudinal and lateral resolution

Eq. (4) describes the max. longitidinal resolution $\Delta z_{\text{longitudinal}}$, which is determined by three triangles $\Delta H_B P H_A$, $\Delta H_B N H_A$ and $\Delta H_B P' H_A$. Fig. 2 shows their geometric relationship. Eq. (3) describes the triangles $\Delta H_B P H_A$, $\Delta H_B N H_A$ and $\Delta H_B P' H_A$ mathematically. Angle resolution $\Delta\alpha$ of Eq. (1) and α of Eq. (3) are inserted into Eq. (4).

$$\begin{aligned}
\Delta H_B N H_A: &\quad \tan(\alpha) = \frac{2\cdot Z_i}{b} \rightarrow \alpha = \arctan\left(\frac{2\cdot Z_i}{b}\right)\\
\Delta H_B P H_A: &\quad \tan\left(\alpha - \frac{\Delta\alpha}{2}\right) = \frac{2\cdot Z_P}{b} \rightarrow Z_P = \tan\left(\alpha - \frac{\Delta\alpha}{2}\right)\cdot\frac{b}{2}\\
\Delta H_B P' H_A: &\quad \tan\left(\alpha + \frac{\Delta\alpha}{2}\right) = \frac{2\cdot Z_{P'}}{b} \rightarrow Z_{P'} = \tan\left(\alpha + \frac{\Delta\alpha}{2}\right)\cdot\frac{b}{2}
\end{aligned} \tag{3}$$

$$\Delta z_{\text{longitudinal}} = \left[\tan\left(\alpha + \frac{\Delta\alpha}{2}\right) - \tan\left(\alpha - \frac{\Delta\alpha}{2}\right)\right]\cdot\frac{b}{2} \tag{4}$$

Eq. (5) describes the physical maximal lateral resolution $\Delta x_{\text{lateral}}$.

$$\Delta\alpha = \arctan\left(\frac{\Delta x_{\text{lateral}}}{Z_i}\right) \quad \rightarrow \quad \Delta x_{\text{lateral}} = Z_i\cdot\tan(\Delta\alpha) \tag{5}$$

Fig. 3 shows the physical max. lateral resolution $\Delta x_{\text{lateral}}$ (bottom) and longitudinal resolution $\Delta z_{\text{longitudinal}}$ depending on the baseline b of the two cameras (top). The following system parameters at parallel camera arrangement were chosen: Focal length $f = 9$ mm and sensor pixel width $w_{pixel} = 4.5$ µm (EV76C570, 1600 × 1200 pixels). If the baseline b increases, the $\Delta x_{\text{lateral}}$ decreases (see Fig. 3, top). However, the baseline cannot be chosen arbitrarily large, since the overlapping measurement volume of the two cameras decreases as the baseline increases. As a result, the stereo field of view would decrease.

4.3. Derivation of the depth error

Using stereo matching algorithms, disparity d in horizontal direction can be calculated from undistorted and rectified image pairs with overlapping areas. The disparity d and its measurement uncertainty Δd is determined by the used stereo matching algorithm and partly by the image resolution. Eq. (7) describes the disparity $d_{u,v}$ on reference image position u, v.

$$u_l = f\cdot\frac{X_{x,y}}{Z_{x,y}} \qquad u_r = f\cdot\frac{X_{x,y} - b}{Z_{x,y}} \tag{6}$$

$$d_{u,v} = u_l - u_r = f\cdot\frac{X_{x,y} - X_{x,y} + b}{Z_{x,y}} = \frac{f\cdot b}{Z_{x,y}} \tag{7}$$

Uncertainties in the determination of disparity values lead to depth

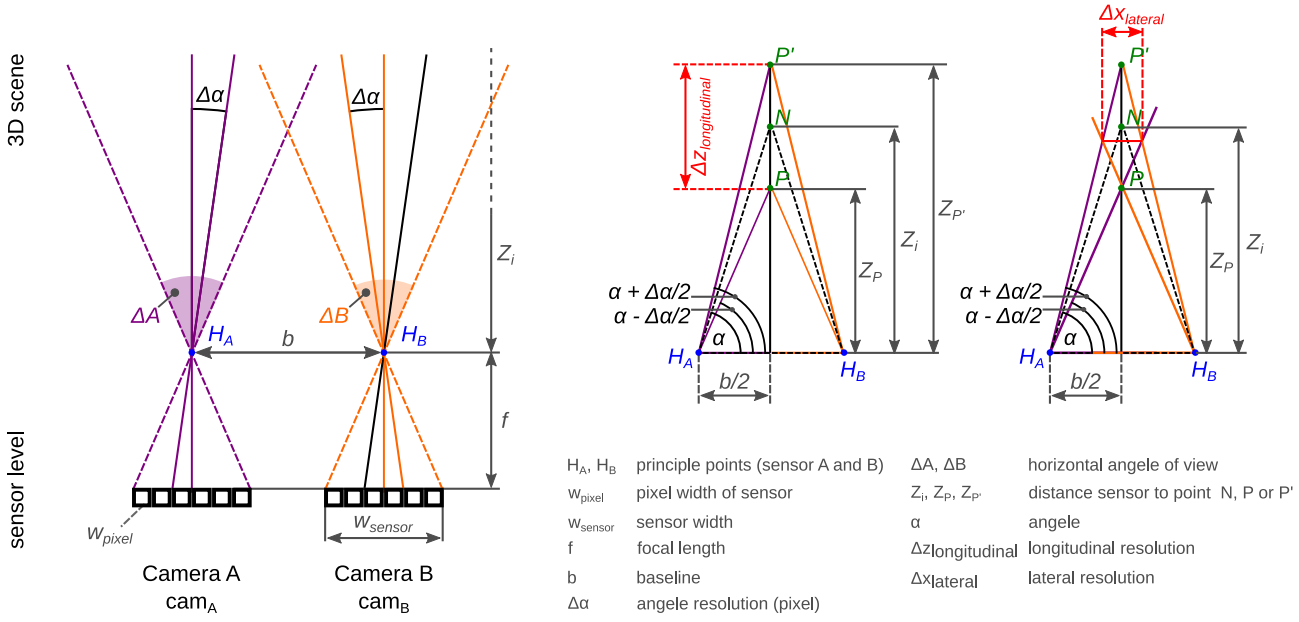

Fig. 2. Passive stereo system *S*1 with camera A cam_A and B cam_B in top view (*left*) with sensor width w_{sensor}, pixel width of the sensor w_{pixel}, baseline b, focal length f, angle resolution $\Delta\alpha$, principle points H_A and H_B, object distance to the camera Z_i and horizontal angle of view ΔA and ΔB. Max. longitudinal resolution $\Delta z_{\text{longitudinal}} = Z_{P'} - Z_P$ and lateral resolution $\Delta x_{\text{lateral}}$ (*right*) with object distance to the camera Z_i.

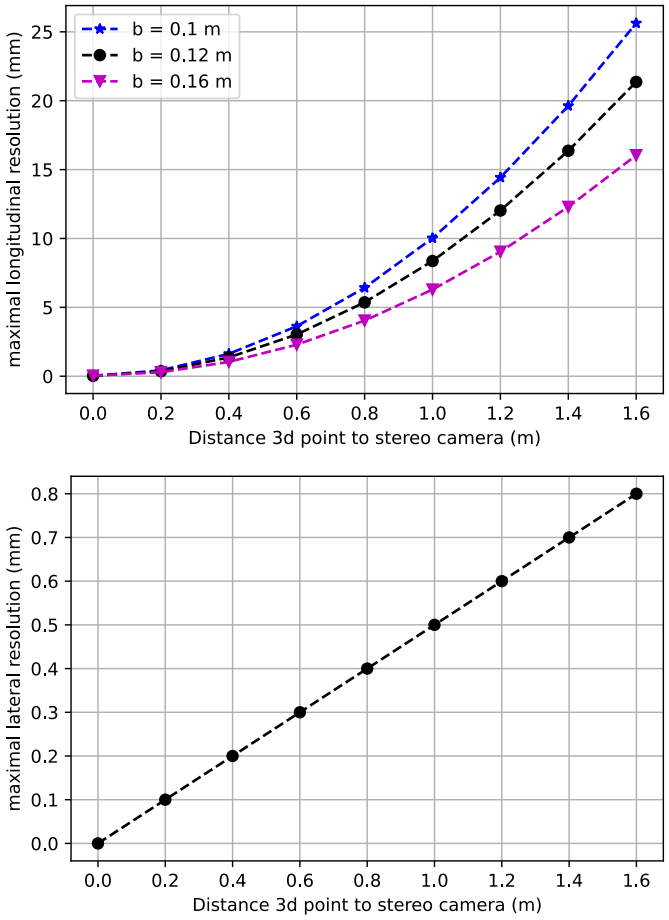

Fig. 3. Maximal longitudinal $\Delta Z_{\text{longitudinal}}$ (top; Eg. (4)) and lateral $\Delta x_{\text{lateral}}$ (bottom; Eg. (5)) resolution at parallel camera arrangement, focal length $f = 9$ mm and sensor pixel width of $w_{\text{pixel}} = 4.5$ µm.

errors ΔZ [17]. Eqs. (8) and (9) describe the estimated depth $\widehat{Z}$ and the actual depth Z with estimated $\widehat{d}$ and the actual disparity d values. Eq. (9) shows the relationship between the uncertainty Δd (see Eq. (10)) in determining a disparity value and the depth error ΔZ.

$$Z = \frac{b \cdot f}{d} \qquad \widehat{Z} = \frac{b \cdot f}{\widehat{d}} \tag{8}$$

$$\Delta Z = |\widehat{Z} - Z| = |\frac{b \cdot f \cdot (d - \widehat{d})}{\widehat{d} \cdot d}| = \frac{\widehat{Z}}{\frac{\widehat{d}}{\Delta d} - 1} \tag{9}$$

For large disparities, an approximation (see Eq. (10)) is assumed. Eq. (11) describes the approximation of the depth error ΔZ (in mm). For the calculation of the depth error one can assume that Δd is at least 0.5 px. However, it is better to choose Δd more generously. Fig. 4 shows the depth error ΔZ (in mm) with varying distance of the measuring point Z (in m) to the stereo system with varying disparity errors Δd (in px) and fixed base distance $b = 0.12$ m, pixel width of the sensor $w_{\text{pixel}} = 4.5$ µm and focal length $f = 9$ mm. As the distance to the stereo system Z increases, the depth error ΔZ increases.

$$\widehat{d} \cdot d \approx d^2, \qquad \frac{b \cdot f}{d} = Z, \qquad d - \widehat{d} = \Delta d \tag{10}$$

$$\Delta Z \approx \frac{b \cdot f}{d^2} \cdot \Delta d = \frac{Z_{x,y}^2}{b \cdot f} \cdot \Delta d \tag{11}$$

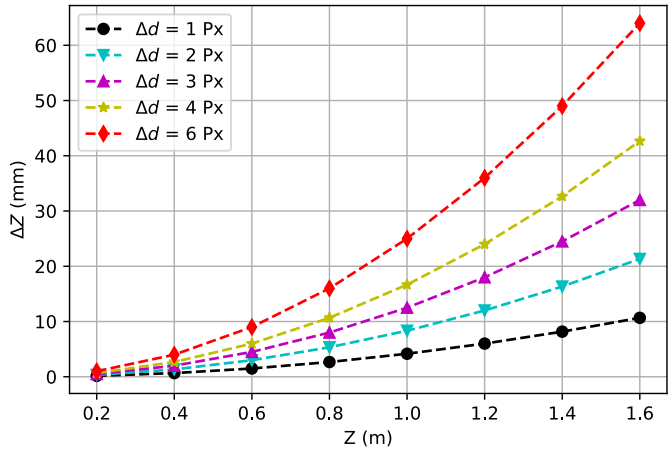

Fig. 4. Approximation of depth error ΔZ (in mm) with varying distance of the measuring point Z (in m) to the stereo system with varying disparity errors Δd (in px). See Eq. (11) at parallel camera arrangement, focal length $f = 9$ mm and sensor pixel width of $w_{\text{pixel}} = 4.5$ µm.

5. Methodology - traceability and calculation of a coherent 3D point cloud

Multi-view stereo (MVS) systems consist of four FPGA-based passive stereo modules ($S1$, $S2$, $S3$, $S4$) (see Fig. 1). The partial 3D volumes acquired by the stereoscopic systems are merged into a coherent 3D point cloud (see Sec. 5.4). Thus, for example, a larger measuring range can be recorded. To transfer the measuring points into the global reference coordinate system, the system must be calibrated with a traceable standard. For our system we use two planar calibration targets with asymmetric circles and with ChArUco markers (see Fig. 5). In Sec. 5.1, the necessary initial steps are explained. These must be carried out again each time the measuring arrangement or the environmental parameters are changed.

5.1. Calibration process and traceability

Fig. 5 shows the two-step calibration process of each stereo module S_n for $n \in \mathbb{N}[1, 4]$ (left) and the registration process of the MVS system (right) for a measurement setup in a row. The two-step calibration process can be done separately or in one step, whereby the two-step method provides more accurate results. After the calibration [18] of each stereo module S_n, intrinsic and extrinsic parameters are available for each system. With this information, a depth map can be calculated for each steroscopic system in the subsequent measurement process (Sec. 5.2 and 5.3). Fig. 6 shows the initial calibration process for a stereo module S_n using Zhang's [20] method. In the first step, images are acquired which show the traceable standard in different positions in the respective measuring field. Using the acquired image pairs, the intrinsic and extrinsic parameters are calculated. We used the following OpenCV functions [19] to calculate the intrinsic and extrinsic parameters:

- `findCirclesGrid()`: Determination of the image points of the circle grid (see Fig. 6, mid)
- `calibrateCamera()`: Using for estimate the intrinsic parameters for each camera (see Fig. 5, left-top; Fig. 6, mid).
- `stereoCalibrate()`: Using for stereo camera calibration to estimate the extrinsisc parameters of each stereo camera system (see Fig. 5, left-bottom).

To speed up the calculation of the correspondence point search during the measurement process (see Sec. 5.2), the image pairs are rectified. For this purpose, undistorted and rectified transformation maps are calculated. Two maps are calculated per camera. The maps

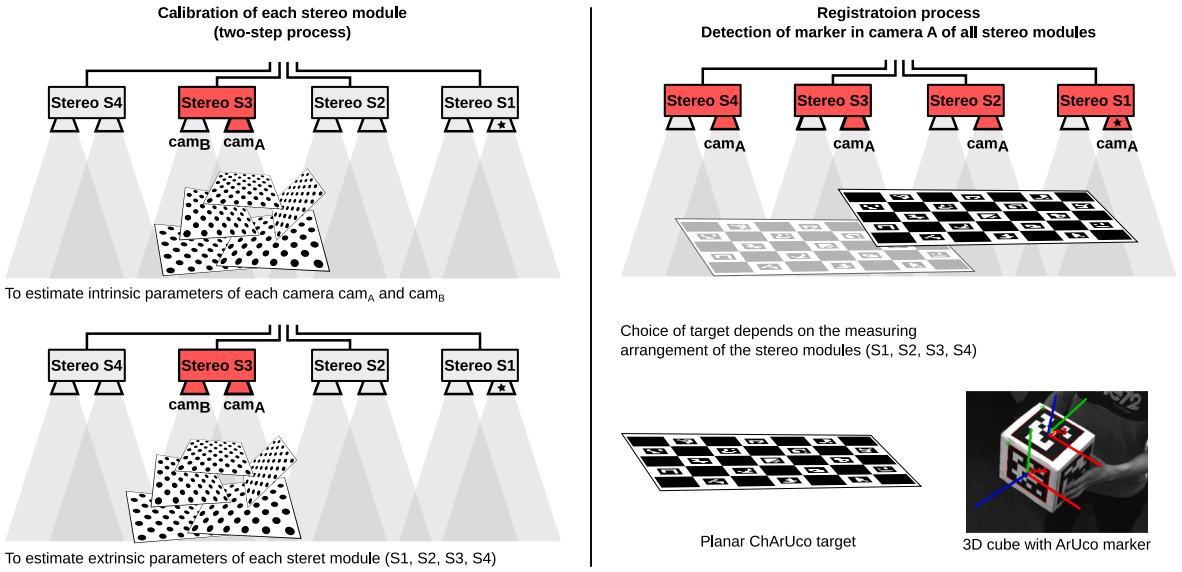

Fig. 5. (*left*) Calibration process and (*right*) registration process. (*left*) Two-step calibration process of each stereo module (*S*1, *S*2, *S*3, *S*4). (*left-top*) Separate camera calibration of each cam_A and cam_B - to estimate intrinsic parameters of each individual camera; (*left-bottom*) Stereo calibration of each stereo module - to estimate extrinsic parameters; (*right*) Initial steps for multi-view stereo registration process using a target with unique markers - to estimate extrinsic parameters (rotation $\mathbf{R_{12/13/14}}$ and translation $\mathbf{T_{12/13/14}}$ matrices) of each cam_A-cam_A stereo pair. (*right-top*) At the same time, ChArUco pattern is visible in the camera A cam_A of stereo module [*S*1, *S*2, *S*3] or in [*S*3, *S*4]. (*right-bottom*) Depending on the arrangement of the stereo modules (*S*1, *S*2, *S*3, *S*4), a planar target (partly overlapping view of the stereo modules) or a 3D cube with ArUco marker (global overlapping view) must be used.

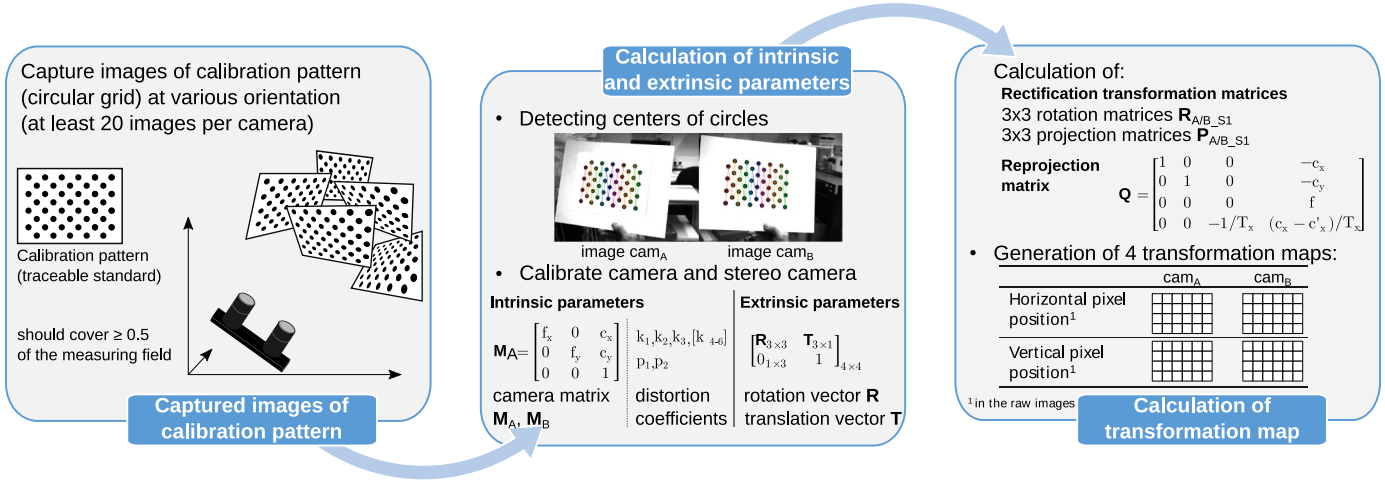

Fig. 6. Pre-processing steps for a stereoscopic system *S*1. (left) Capturing images of calibration pattern (circular grid, ChArUco or ArUco cube). (mid) Calculation of intrinsic and extrinsic parameters [18,19]. (right) Calculation of rotation matrices $\mathbf{R_{A\_S1}}$ and $\mathbf{R_{B\_S2}}$ (undistorted and rectified coordinate system), the reprojection matrix **Q** as well as undistorted and rectified transformation maps.

contain for each pixel (*u*, *v*) in the (corrected and rectified) target image the corresponding (horizontal or vertical) coordinates in the source image (raw image). We used the following OpenCV functions [19] to create these maps:

- `stereoRectify()`: Calculation of the rotation $\mathbf{R}_{A\_S_n}$, $\mathbf{R}_{B\_S_n}$ and projection matrices $\mathbf{P}_{A\_S_n}$, $\mathbf{P}_{B\_S_n}$ for stereo pairs and the 4 × 4 reprojection matrix **Q** (necessary for 3D reconstruction).
- `initUndistortRectifyMap()`: Generation of transformation maps using the parameters calculated by previously described function. Two maps (horizontal and vertical pixel position) for each camera containing the floating point pixel positions in the raw images (see Fig. 6, right)

If an FPGA is used as the computational unit, compress the transformation matrix for each camera to increase data throughout. The compressed transformation matrices are stored on the respective secure digital (SD) memory card of the FPGA peripheral board (see Sec. 6).

The registration process (Fig. 5, right) is to estimate the rotation $\mathbf{R_{12/13/14}}$ and translation matrices $\mathbf{T_{12/13/14}}$ between the cameras cam_A of each stereo module S_n so that the partial volumes of each S_n can be merged into one (see Sec. 5.4). Due to the measurement arrangement in a row, we need a larger target with unique markers that can be seen in several cameras cam_A at the same time. We use a planar ChArUco pattern of the size DIN A4. This can be seen, for example, simultaneously in the camera cam_A images of stereo module [*S*1, *S*2, *S*3] or in [*S*3, *S*4] (see Fig. 5, right-top). For a measurement arrangement with different viewing directions and a common global measurement volume, a three-

dimensional target (e.g. ArUco cube) can be used instead of a planar target. The advantage of a three-dimensional target is that all left cameras can capture the target at the same time.

5.2. Pre-processing and stereo matching

Fig. 7 shows the image processing chain of a stereo camera system, which is calculated on the FPGA. The main steps are the lens distortion correction with rectification and the stereo matching. With the undistortion and rectification transformation maps, a reverse transformation can be performed, whereby the raw image pairs are corrected from the lens distortion and are also rectified in one step. The rectification is necessary because it reduces the computational effort of the subsequent stereo matching (1D correspondence point search). Since both images have line-corresponding content in the overlapping region (i.e. horizontal epipolar lines on both images with the same y-coordinate), the complexity of the correspondence point search is reduced [14,19]. We use the well-known Semi-Global Matching (SGM) algorithm by Hirschmüller [10] for calculated a dense disparity map. The method offers a very good compromise between runtime and accuracy. Especially at object boundaries and fine structures [21].

After stereo matching, the left image (texture) and the disparity map (geometry) are merged and output via HDMI. This is necessary so that texture and geometry information are transferred to the subsequent processing unit.

5.3. Calculation of 3D point cloud

Equations (13) and (12) describe the calculation of a 3D point on image position x, y based on the disparity map $d_{x,y}$ and the reprojection matrix $\mathbf{Q}$, which contains the focal length f in px and the principal point (c_x, c_y) in px [17]. To get the disparity vales from such fixed-point representation, each element must be divided by 16.

$$Z_{X,Y} = \frac{b \cdot f}{d_{x,y}} \tag{12}$$

$$X = (x - c_x) \cdot \frac{Z_{X,Y}}{f}, \qquad Y = (y - c_y) \cdot \frac{Z_{X,Y}}{f} \tag{13}$$

5.4. Coherent 3D point cloud

This section describes the registration of the four 3D point clouds using unique target markers. To obtain a coherent 3D point cloud, the extrinsic parameters of the partial 3D point clouds must be determined. A traceable normal is used to determine the positions and orientation of each left camera. Depending on the measuring arrangement, a different traceable normal is required [22–25]. Regardless of which traceable normal is used, it must have uniquely definable points. Three to four unique points on the normal are registered simultaneously by all stereoscopic systems. The corresponding points determine rotation and transformation matrices, allowing the merging of 3D surfaces [26].

Due to our horizontally aligned stereoscopic systems with low overlapping areas (Fig. 5, right), we use a planar ChArUco (chessboard with ArUco) pattern in DIN A0 size. The ArUco markers on the target have uniquely definable points (rotation invariant marker). Depending on the measurement setup, the calibration target is not acquired by all stereo modules S_n at the same time. In this case, the calibration target must be acquired in several positions. At least two stereoscopic systems must be able to detect the normal simultaneously. E.g. in Fig. 5 (right) two positions are required – target is visible in the camera cam_A of stereo module [$S1$, $S2$, $S3$] or in [$S3$, $S4$].

Eq. (14)/(15)/(16) shows the 3D point transformation from the stereo camera coordinate system $S2/S3/S4$ into the global reference coordinate system $S1$ (see Fig. 8). Sec. 5.1 describes the rotation $\mathbf{R}$ and translation $\mathbf{T_{12/13/14}}$ matrices.

$$\begin{pmatrix} X_{21} \\ Y_{21} \\ Z_{21} \end{pmatrix} = \mathbf{R_{A\_S1}} \cdot \left(\mathbf{R_{12}^{-1}} \cdot \mathbf{R_{A\_S2}^{-1}} \cdot \begin{pmatrix} X_{S2} \\ Y_{S2} \\ Z_{S2} \end{pmatrix} - \mathbf{T_{12}} \right) \tag{14}$$

$$\begin{pmatrix} X_{31} \\ Y_{31} \\ Z_{31} \end{pmatrix} = \mathbf{R_{A\_S1}} \cdot \left(\mathbf{R_{13}^{-1}} \cdot \mathbf{R_{A\_S3}^{-1}} \cdot \begin{pmatrix} X_{S3} \\ Y_{S3} \\ Z_{S3} \end{pmatrix} - \mathbf{T_{13}} \right) \tag{15}$$

$$\begin{pmatrix} X_{41} \\ Y_{41} \\ Z_{41} \end{pmatrix} = \mathbf{R_{A\_S1}} \cdot \left(\mathbf{R_{14}^{-1}} \cdot \mathbf{R_{A\_S4}^{-1}} \cdot \begin{pmatrix} X_{S4} \\ Y_{S4} \\ Z_{S4} \end{pmatrix} - \mathbf{T_{14}} \right) \tag{16}$$

6. Modular multi-view stereo system

Our multi-view stereo (MVS) system consists of four FPGA-based passive stereoscopic modules (Xilinx Zynq-7000 7020 SoC) and a downstream computing unit (Zynq UltraScale ZU9EG SoC). Preprocessing and stereo matching (see Sec. 5.2) are outsourced to the stereoscopic systems [12]. This enables data reduction close to the

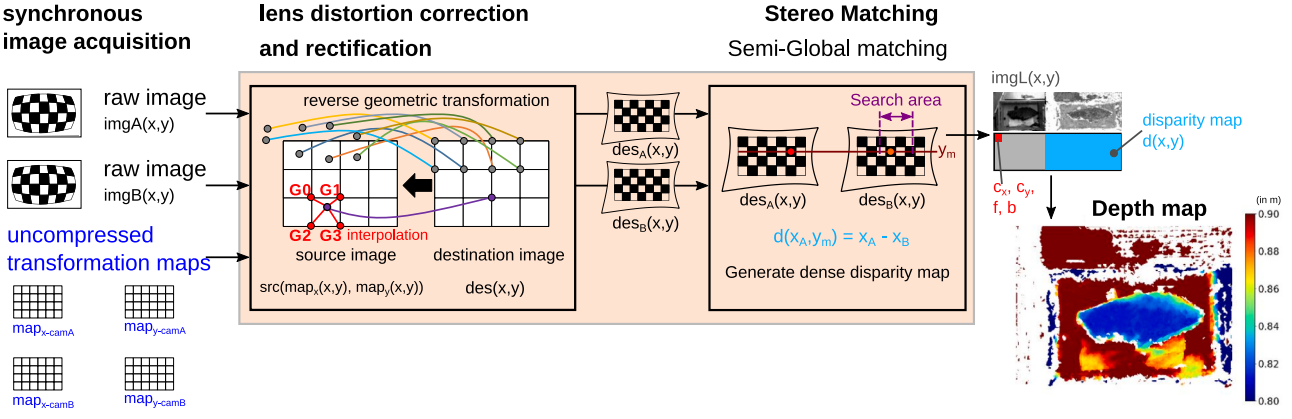

Fig. 7. Image processing chain of a stereo module S_n. First, the images are undistorted and rectified in one step. Here, a generic geometric transformation with bilinear interpolation of the neighbouring pixels $G0$, $G1$, $G2$ and $G3$ is applied to the images. The source image is transformed using the undistortion and rectification transformation maps map_x and map_y. Correspondence points are now searched for in the generated destination images des_A and des_B. The purple marked area indicates the search area on the red line y_m. The disparity map is obtained by subtracting the x-position of the point in the left image $des_A(x_1, y)$ and that of the right image $des_B(x_2, y)$. As a stereo matching algorithm, we use the Semi-Global Matching algorithm by Hirschmüller [10]. The output is a dense disparity map $d(x, y)$ (in px). This can be reprojected into the 3D space with the reprojection matrix $\mathbf{Q}$. The depth map (in m) is shown at the bottom right. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

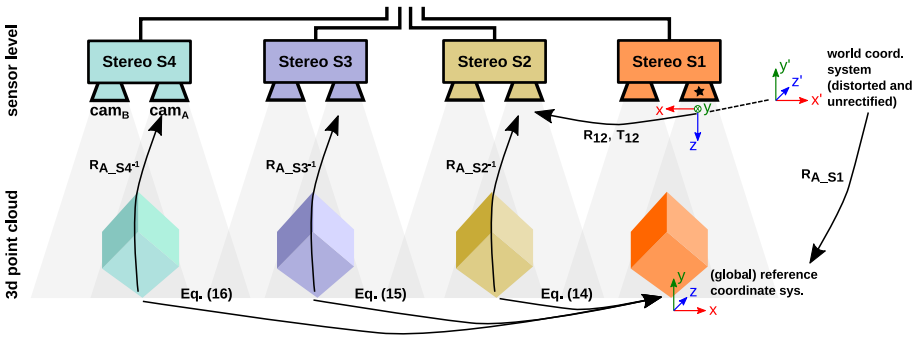

Fig. 8. Registration of four 3D point clouds into the reference coordinate system. Eqs. (14)–(16) describe mathematically the transformation of a 3D point from the respective stereo coordinate system to the reference coordinate system (left camera of stereo system $S1$).

sensor. Fig. 9 shows a schematic diagram of one stereo module S_n. The output of each stereoscopic system is a composite image consisting of the left stereo image and the corresponding calculated disparity map (8 bit and 4 bit subpixel). Relevant calibration values (baseline b, focal length f, and principle point c_x, c_y) of the reprojection matrix **Q** (see Fig. 6) are stored in the first pixels of the left image. Thus, the stored image contains all values to calculate a depth map. The output image is delivered synchronously via HDMI to the downstream processing unit (Zynq UltraScale). The concept of this unit is described by Hänsel et al. in Ref. [27]. Another downstream computing (image processor) unit pulls in the merged output image and generates depth maps and a coherent 3D point cloud (see Sec. 5.4).

The modularity of the MVS system allows a versatile use of the measuring system. The cameras cam_A and cam_B of each stereo module S_n are arranged in parallel with a fixed baseline of $b = 0.120$ m. Compared to a convergent camera arrangement, a parallel arrangement requires a fewer memory (smaller ring buffer) because the relative position values in the vertical direction in the transformation maps are smaller (Sec. 5.1 and 5.2).

7. Capabilities and limitation of the system by an application example with optical non-cooperative surface

Some objects have challenging surfaces that are difficult or impossible to detect with a passive stereo system. These include, for example, objects with a texture-poor surface (metal), optically non-cooperative surfaces such as asphalt (highly absorbent surface) or glass.

The capabilities and limitations of the presented system are demonstrated by an application example. For this we want to capture the surface of a piece of asphalt, which is optically non-cooperative. The asphalt surface is very rough (micro- and macrotexture), whereby the ambient light is reflected and absorbed differently. In addition, backscattering increases massively when the asphalt is wet. Fig. 10 shows the measurement object, a piece of asphalt with naturally formed fissures as well as artificially added holes. The measurement object was aligned horizontally to the measurement system.

Fig. 11 shows the left image (only ambient light) with superimposed mask, based on depth values $Z(x, y)$. The depth values in the yellow-marked area correspond to the plane distance Z_{plane}. The created colour mask consists maximum of four colours: Lying in plane distance (yellow), above (cyan) and below the plane distance (red), as well as invalid pixel (white). Measurement points that fall below $(< Z_{\text{plane}})$ or exceed $(> Z_{\text{plane}})$ this distance value are marked based on the size of the connected area, conclusions can be drawn about the surface. For example, fissures can be recognized on smaller red contiguous pixels, surrounded by yellow pixels. Interesting areas can be examined in more detail afterwards, for example by analysing the normalized depth values along a line.

Due to the high absorption of the asphalt, the system was extended with constant lighting. Depending on the sensitivity of the camera

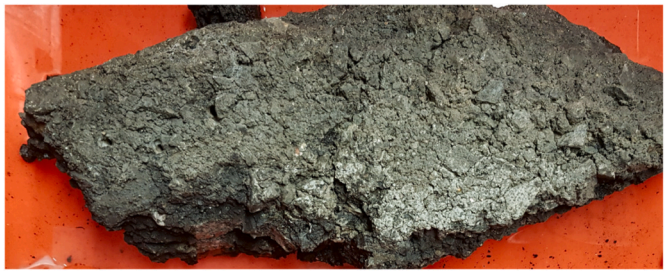

Fig. 10. Application example: Piece of asphalt with holes and naturally formed fissures.

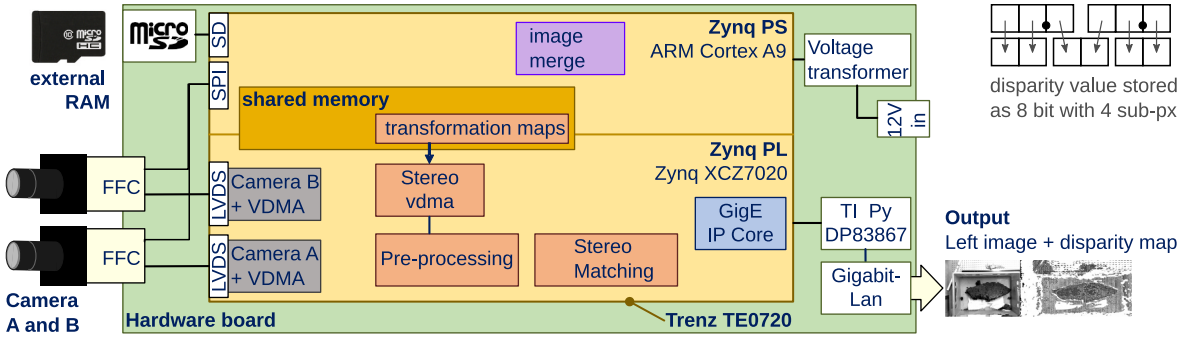

Fig. 9. Hardware board of FPGA-based stereoscopic system Trenz TE0720 with Zynq processing system (PS), programmable logic (PL) and shared memory. Synchronous image acquisition, pre-processing (lens distortion correction and rectification by transformation maps, see Fig. 7), stereo matching (using Semi-Global Matching), merging of left image (texture) and disparity map (geometry), merged images as output. The first pixels of the left image contain calibrating parameters (principle point, focal length and baseline). Disparity map contains subpixel values.

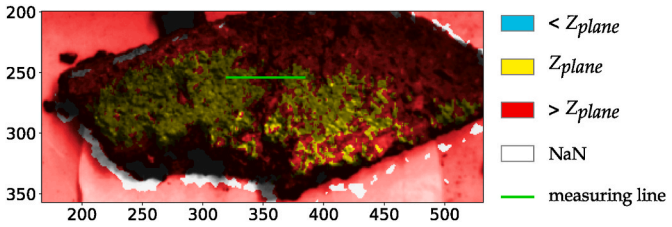

Fig. 11. Left image with superimposed colour mask. (Large and small) coherent areas, e.g. smaller coherent areas correspond to holes or fissure. Fig. 13 shows depth values along the marked green line. The measuring line runs orthogonal to a slit. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

sensor, a preliminary investigation was carried out to determine in which wavelength range the asphalt is still optically cooperative. This preliminary investigation has shown that high-power LEDs in the red to infrared wavelength range (625 nm, 660 nm, 730 nm and 850 nm) are suitable for detecting bigger fissures in asphalt. The preliminary investigations were carried out in the laboratory with one stereo module S_n. Fig. 12 shows the measurement setup. The main emission directions of the LEDs are orthogonal to the measuring surface. The measured object was recorded in dry, damp as in wet condition. Fig. 13 shows a depth values along the line $y = 254$ px (see Fig. 11, green line), that lies across a crack in the surface. The piece of asphalt is not a standard unit for height. The cracks are therefore determined with a measuring probe (ZEISS Jena ABBE P01 length measuring device 100/0.001 mm). In contrast to the diffuse ambient light, the detection of the fissure depth is better with the directional LEDs orthogonal to the surface. With the presented system it is not possible to measure correct depth values of damp or wet (water film) asphalt. The water film on the asphalt surface acts as an additional boundary layer.

The undamaged asphalt surface reflects enough light to measure depth values. However, deep cracks in the asphalt in particular absorb the light, which means that no correct depth values can be measured. However, the area of the cracks on the surface can be detected because the measuring points at these areas deviate from the plane distance ($< Z_{\text{plane}}$). Wet asphalt, is visually non-cooperative despite illumination (when viewed orthogonally).

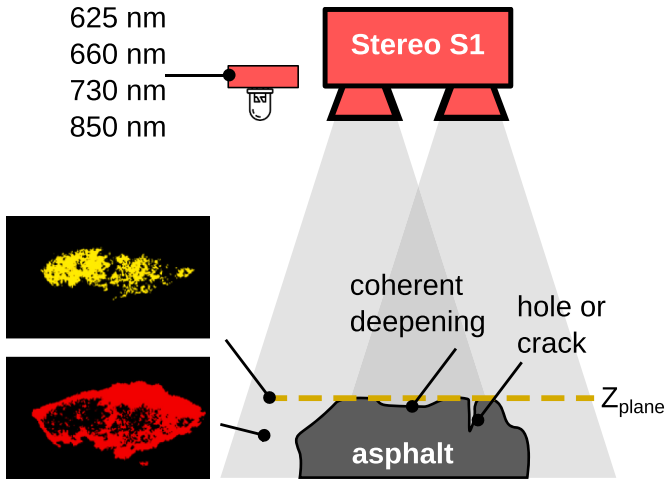

Fig. 12. Sketch of the measuring arrangement. Z_{plane} is the distance from the stereo module $S1$ to the asphalt surface plane. Measurement points $Z(y, x)$ that fall below ($< Z_{\text{plane}}$) or exceed this distance ($> Z_{\text{plane}}$) are marked.

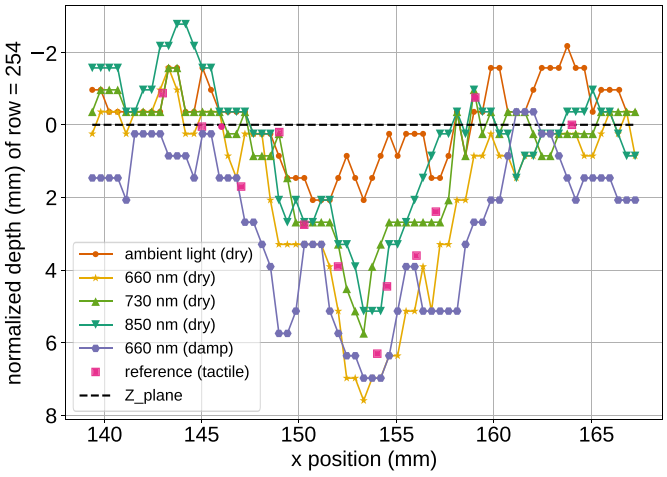

Fig. 13. Normalized depth points (mm) along the image line $y = 254$ (see Fig. 11) with distance Z_{plane}. Measurement curves at dry as well as damp asphalt surfaces with different illumination. Position of the crack at 154 mm. Actual crack depth: 6.320 mm (measuring probe).

8. Conclusions and future work

8.1. Real-time sensor-closed multi-view stereo system

We have shown an real-time FPGA-based multi-view stereo (MVS) system (Fig. 1) with flexible measurement arrangement and partial overlapping field of view (FOV). With this system, objects or scenes can be captured in 3D with texture by combining a multiple view with eight cameras. Our MVS system produces coherent point clouds with a maximum number of 7 37 280 points per frame using Semi-Global Matching algorithm (by Hirschmüller [10]) at a disparity range of 256 px, an image resolution of 640 px × 460 px and with aggregated costs over 4 directions. The merged point cloud is recorded and stored at 24 fps at 12 bit depth resolution.

This system consists of four parallel stereo modules S_n (using Xilinx-Zynq-7000 7020 SoC system), which supply disparity maps (8 bit and 4 bit subpixel) synchronously to a downstream processing unit (Zynq Ultrascale) via HDMI. The disparity maps are calculated to depth maps and a coherent 3D point cloud on a image processor unit below. The advantages of the system are the modularity and the low power consumption (about 6 W) of the stereo modules. To reduce the data rate and data volume, each stereo module transmits only the left image and the disparity map (with relevant calibration parameters). This is achieved by outsourcing calculation steps to the FPGA close to the camera sensor. Due to the use of HDMI interface, greater distances between the stereo systems are possible. Due to the modular 3D measurement system, the stereo system arrangements can easily be adapted to a corresponding application. In contrast to active MVS systems (pattern projection pro stereo system, see Sec. 2), the measurement arrangement of our system can be flexibly selected depending on the application. This is because a combination of two or more active stereo systems generates an overlay of different patterns (each system has its own pattern), which leads to problems in assigning correspondence points in the image pairs. Depending on the application, the measuring arrangement of the stereo modules ($S1$, $S2$, $S3$, $S4$) can be flexibly adapted. A measuring arrangement in a row is suitable, for example, for recording a larger measuring volume (e.g. assembly process in a production hall). A measuring arrangement with different viewing directions is useful if an object has to be captured holistically. The utilisation of the FPGA fabric depends on image resolution, camera alignment, disparity range, matching cost function and cost aggregation. To minimise core calculation time, calculation clock, base clock, and grad of parallelization were optimised within their respective limits. With a parallelization

factor of 32, the FPGA logic is almost 80 % utilised. With a more powerful closely FPGA/ARM SoC, the computing time can be increased [12]. The performance of FPGA-based stereo systems ([1,11,13]) is only comparable to a limited extent due to different hardware and matching parameters, which require different amounts of logical resources.

8.2. Limitations - optically non-cooperative surface

We have shown on an optically non-cooperative surface (asphalt) that it is useful to extend the system with an additional constant illumination unit (high-power LEDs with different wavelengths) to make the surface more visually cooperative (Fig. 13). However, the presented system reaches its limits for some surfaces, such as damp asphalt. Based on the size of the contiguous area at a certain depth, conclusions can be drawn about cracks or shallow depressions. However, due to the reflections of the light, the depth of deep fissures cannot be correctly detected in either dry or damp conditions. Tactile measurement methods also reach their limits in the case of deep cracks. Passive stereo systems are not suitable for detecting the surface of wet asphalt. For qualitative evaluation, the surface of the measured object was determined with a tactile measuring device.

In many applications, e.g. manufacturing processes, there are visually non-cooperative objects (e.g. non or repetitive textures). Traditional (not AI-based) stereo matching algorithms have their limitations in 3D detection of texture-poor or optically non-cooperative surfaces [16]. Some surfaces can be made more visually cooperative with the addition of constant illumination (one or more wavelengths). Others (e.g. glass or wet asphalt) are simply not three-dimensionally detectable with a passive stereo system. To overcome the limitations of conventional correspondence point analysis, among others, many different efficient stereo matching algorithms based on deep learning have been developed recently [15,16]. However, data-driven methods require a large amount of training data (with disparity ground truth). The creation of this is very expensive and time-consuming. Moreover, this system would be strongly limited to the trained use case, conditioned by the training data set resp. pre-trained model.

CRediT authorship contribution statement

Christina Junger: Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, All authors have read and agreed to the published version of the manuscript. **Richard Fütterer:** Methodology, Software, All authors have read and agreed to the published version of the manuscript. **Maik Rosenberger:** Supervision, Project administration, All authors have read and agreed to the published version of the manuscript. **Gunther Notni:** Conceptualization, Writing – review & editing, Supervision, Project administration, All authors have read and agreed to the published version of the manuscript.

Acknowledgement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] S. Gehrig, F. Eberli, T. Meyer, A Real-Time Low-Power Stereo Vision Engine Using Semi-global Matching, vol. 5815, 2009, pp. 134–143, https://doi.org/10.1007/978-3-642-04667-4_14.

[2] M. Eisenbach, D. Aganian, M. Köhler, B. Stephan, C. Schröter, H. Groß, Visual scene understanding for enabling situation-aware cobots, ilmedia (2022), https://doi.org/10.22032/dbt.51471.

[3] Y. Zhang, S. Müller, B. Stephan, H.-M. Gross, G. Notni, Point cloud hand–object segmentation using multimodal imaging with thermal and color data for safe robotic object handover, Sensors, 21 (16), 10.3390/ s21165676, URL, https://www.mdpi.com/1424-8220/21/16/5676.

[4] VDMA, Robotik und Automation mit vollen Auftragsbüchern, aber gestörten Lieferketten, 06 22, https://www.vdma.org/viewer/-/v2article/render/52710516. (Accessed 21 July 2022).

[5] G. Straube, C. Zhang, A. Yaroshchuk, S. Lübbecke, G. Notni, Modelling and calibration of multi-camera-systems for 3D industrial supervision applications, in: Photonics and Education in Measurement Science 2019, vol. 11144, SPIE, 2019, pp. 68–76, https://doi.org/10.1117/12.2532320. International Society for Optics and Photonics.

[6] C. Zhang, I. Gebhart, P. Kühmstedt, M. Rosenberger, G. Notni, Enhanced contactless vital sign estimation from real-time multimodal 3d image data, Journal of Imaging 6 (2020) 123, https://doi.org/10.3390/jimaging6110123.

[7] K.J. Lee, K. Bong, C. Kim, J. Jang, K.-R. Lee, J. Lee, G. Kim, H.-J. Yoo, A 502-gops and 0.984-mw dual-mode intelligent adas soc with real-time semiglobal matching and intention prediction for smart automotive black box system, IEEE J. Solid State Circ. 52 (1) (2017) 139–150, https://doi.org/10.1109/JSSC.2016.2617317.

[8] D. Hernandez-Juarez, A. Chacón, A. Espinosa, D. Vázquez, J. Moure, A. López, Embedded real-time stereo estimation via semi-global matching on the GPU, in: International Conference on Computational Science 2016, ICCS, 2016, pp. 143–153, https://doi.org/10.1016/j.procs.2016.05.305, 6-8 June 2016, San Diego, California, USA, 2016.

[9] J. Woodfill, B. Von Herzen, Real-time stereo vision on the parts reconfigurable computer, in: Proceedings. The 5th Annual IEEE Symposium on Field-Programmable Custom Computing Machines Cat. No.97TB100186), 1997, pp. 201–210, https://doi.org/10.1109/FPGA.1997.624620.

[10] H. Hirschmüller, Accurate and efficient stereo processing by semi-global matching and mutual information, in: CVPR 2005, vol. 2, IEEE, 2005, pp. 807–814.

[11] C. Banz, S. Hesselbarth, H. Flatt, H. Blume, P. Pirsch, Real-time stereo vision system using semi-global matching disparity estimation: architecture and fpga-implementation, in: 2010 International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation, 2010, pp. 93–101, https://doi.org/10.1109/ICSAMOS.2010.5642077.

[12] R. Fütterer, M. Schellhorn, G. Notni, Implementation of a multiview passive-stereo-imaging system with SoC technology, in: Photonics and Education in Measurement Science 2019, vol. 11144, SPIE, 2019, pp. 164–168, https://doi.org/10.1117/12.2530721. International Society for Optics and Photonics.

[13] O. Rahnama, D. Frost, O. Miksik, P.H. Torr, Real-time dense stereo matching with ELAS on FPGA-accelerated embedded devices, IEEE Rob. Autom. Lett. 3 (3) (2018) 2008–2015, https://doi.org/10.1109/lra.2018.2800786.

[14] C. Junger, A. Heß, M. Rosenberger, G. Notni, FPGA-based lens undistortion and image rectification for stereo vision applications (Erratum), International Society for Optics and Photonics, in: Photonics and Education in Measurement Science 2019, vol. 11144SPIE, 2021, https://doi.org/10.1117/12.2595882, 352 – 352.

[15] X. Cheng, Y. Zhong, M. T. Harandi, Y. Dai, X. Chang, T. Drummond, H. Li, Z. Ge, Hierarchical Neural Architecture Search for Deep Stereo Matching, ArXiv abs/2010.13501.

[16] C. Junger, G. Notni, Optimisation of a stereo image analysis by densify the disparity map based on a deep learning stereo matching framework, in: Dimensional Optical Metrology and Inspection for Practical Applications XI, vol. 12098, SPIE, 2022, pp. 91–106, https://doi.org/10.1117/12.2620685. International Society for Optics and Photonics.

[17] H. Fradi, J.-L. Dugelay, Improved depth map estimation in stereo vision, Proc. SPIE-Int. Soc. Opt. Eng. 7863. doi:10.1117/12.872544.

[18] W. Jakob, Calibration best practices, URL, https://calib.io/blogs/knowledge-base/calibration-best-practices, 2018, 11.

[19] D. van Heesch, openCV - Camera Calibration and 3D Reconstruction (v4.5.2).

[20] Z. Zhang, A flexible new technique for camera calibration, IEEE Trans. Pattern Anal. Mach. Intell. 22 (11) (2000) 1330–1334, https://doi.org/10.1109/34.888718.

[21] H. Hirschmüller, Semi-global matching - motivation, developments and applications, in: Photogrammetric Week, 2011, pp. 173–184. Wichmann.

[22] F. Abedi, Y. Yang, Q. Liu, Group geometric calibration and rectification for circular multi-camera imaging system, Opt Express 26 (23) (2018) 30596–30613, https://doi.org/10.1364/OE.26.030596.

[23] M. Feng, X. Jia, J. Wang, S. Feng, T. Zheng, Global calibration of multi-cameras based on refractive projection and ray tracing, Sensors 17 (2017) 2494, https://doi.org/10.3390/s17112494.

[24] Z. Liu, Z. Meng, N. Gao, Calibration of the relative orientation between multiple depth cameras based on a three-dimensional target, Sensors 19 (2019) 3008, https://doi.org/10.3390/s19133008.

[25] J. Sun, H. He, D. Zeng, Global calibration of multiple cameras based on sphere targets, Sensors 16 (1). doi:10.3390/s16010077.

[26] V. Matiukas, D. Miniotas, Point cloud merging for complete 3d surface reconstruction, Electron. Electr. Eng. 113. doi:10.5755/j01.eee.113.7.616.

[27] M. Hänsel, T. Scholz, M. Rosenberger, G. Notni, Modular embedded lightfield system for road condition assessment, J. Phys. Conf. 1065 (2018), 032018, https://doi.org/10.1088/1742-6596/1065/3/032018.