Article Identifier
DOI: 10.22032/dbt.58927
URN: urn:nbn:de:gbv:ilm1-2023isc-070:0

60th ILMENAU SCIENTIFIC COLLOQUIUM
Technische Universität Ilmenau, 4 – 8 September 2023
URN: urn:nbn:de:gbv:ilm1-2023isc:1

INVESTIGATIONS OF CLOSED SOURCE REGISTRATION METHOD OF DEPTH SENSOR TECHNOLOGIES FOR HUMAN-ROBOT COLLABORATION

Christina Junger[a] *and 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

ABSTRACT

Productive teaming is the new form of human-robot interaction. The multimodal 3D imaging has a key role in this to gain a more comprehensive understanding of production system as well as to enable trustful collaboration from the teams. For a complete scene capture, the registration of the image modalities is required. Currently, low-cost RGB-D sensors are often used. These come with a closed source registration function. In order to have an efficient and freely available method for any sensors, we have developed a new method, called Triangle-Mesh-Rasterization-Projection (TMRP). To verify the performance of our method, we compare it with the closed-source projection function of the Azure Kinect Sensor (Microsoft). The qualitative comparison showed that both methods produce almost identical results. Minimal differences at the edges indicate that our TMRP interpolation is more accurate. With our method, a freely available open-source registration method is now available that can be applied to almost any multimodal 3D/2D image dataset and is not like the Microsoft SDK optimized for Microsoft products.

Index Terms - human-robot interaction, collaboration, productive teaming, registration, projection, depth completion; Triangle-Mesh-Rasterization-Projection; Azure Kinect SDK; Microsoft

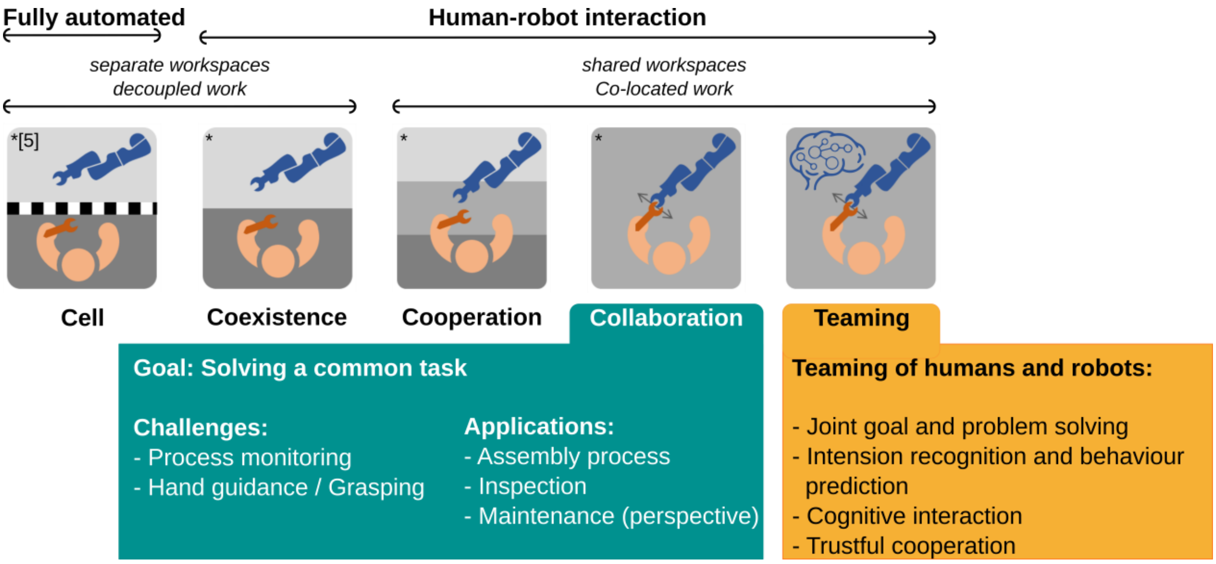
1. INTRODUCTION AND RELATED WORK

1.1 Human-robot interaction

Nowadays, in almost all industry sectors **various industrial robots** are represented. The choice of the robot depends on the field of application (s. Figure 9 in the appendix). Robots can be used **as work support** or **as a platform for sensors** for monitoring. In addition to robots with wheels or caterpillar drive (tile-laying robot [1] [2], pick-and-place robots), there are also robots with a movement capability that enables them to reach hard-to-reach places[1]. Among the latter is the Boston Dynamics Spot, which can be used as a platform for optical metrology to capture the environment in three dimensions, e.g. for forest monitoring or litter collection along the highway (s. Figure 9).

There are currently four **human-robot interaction** (HRI) forms [3] (see Figure 1): (1) coexistence, (2) cooperation, (3) collaboration [4] and (4) productive teaming. Collaborative robots, in short cobots, focus on safety and user-friendliness so as not to injure humans. Another goal is the awareness[2] that it is considered in the form of productive teamwork (Figure 1). Here, with a common understanding of robots and humans, a flexible, variant and non-automated

[1] Stairs or terrain insurmountable for wheels (forest or safety-critical production halls)
[2] observability, predictability and controllability



Goal: Solving a common task

Challenges:
- Process monitoring
- Hand guidance / Grasping

Applications:
- Assembly process
- Inspection
- Maintenance (perspective)

Teaming of humans and robots:
- Joint goal and problem solving
- Intension recognition and behaviour prediction
- Cognitive interaction
- Trustful cooperation

Figure 1: ***Four forms of human-robot interaction*** *(source according to [5]). The newest form is productive teaming, which focuses on human-centered collaboration between people and machines.*

work process is to be executed efficiently. To ensure human safety and human-centered collaboration, it is important to select the appropriate sensors. For example, in a robot-human collaboration (HRC), safety must be ensured in the event of simultaneous grasping of un-/known objects [4]. On the one hand, this is ensured via a lower robot speed and the (mobile/stationary) sensor technology. In most cases, human-robot collaboration uses multimodal 3D sensors to enable trustful cooperation. Multimodal 3D imaging has a key role, as the quality and processing speed of the image analysis depends on the output data of the sensor.

1.2 Registration of a multimodal 3D system

Multimodal 3D imaging is enormously important nowadays [6] and is present in various industries. E.g. for safe robot-human cooperation [4] or in scene analysis [7, 8, 9]; for interactive robot teaching [10]; for autonomous navigation [11]; for medical applications [12, 13]; and for quality control [14]. A newer area is the detection of optically uncooperative surfaces (in the visual wavelength range) [15, 16, 17, 18, 19]. An important step in multimodal image processing is the **registration**. The goal of registration is optimal data fusion of the different image modalities into a coherent coordinate system for a more comprehensive understanding of the assembly process. Various challenges can arise here [20]: non-commensurability, different resolutions, number of dimensions, noise, missing data, conflicting, contradicting or inconsistent data. We focus on one challenge, the **different resolutions**. The depth sensor is the most important sensor of all, since it determines the geometry. However, this sensor usually has a low resolution. Figure 2 shows the resolutions from an RGB-D sensor. Here the depth image has a resolution of 0.09 MPx and the RGB image has a resolution of 3.7 MPx. In order not to discard any information, all acquired data is transformed and projected to the maximum sensor resolution of the multimodal 3D system. In our example this would be the RGB camera. The low resolution point cloud is first transformed into the high-resolution target image coordinate system and then projected into the high-resolution 2D raster target image. Depending on the projection method, four challenges (A)-(D) can arise here. For more details see our publication [21].

(A) gap due to physical limitation of source sensor technology → not fully considered

(B) false gap in raster (neighboring 3D coordinates X/Y are further than one pixel apart) [22]

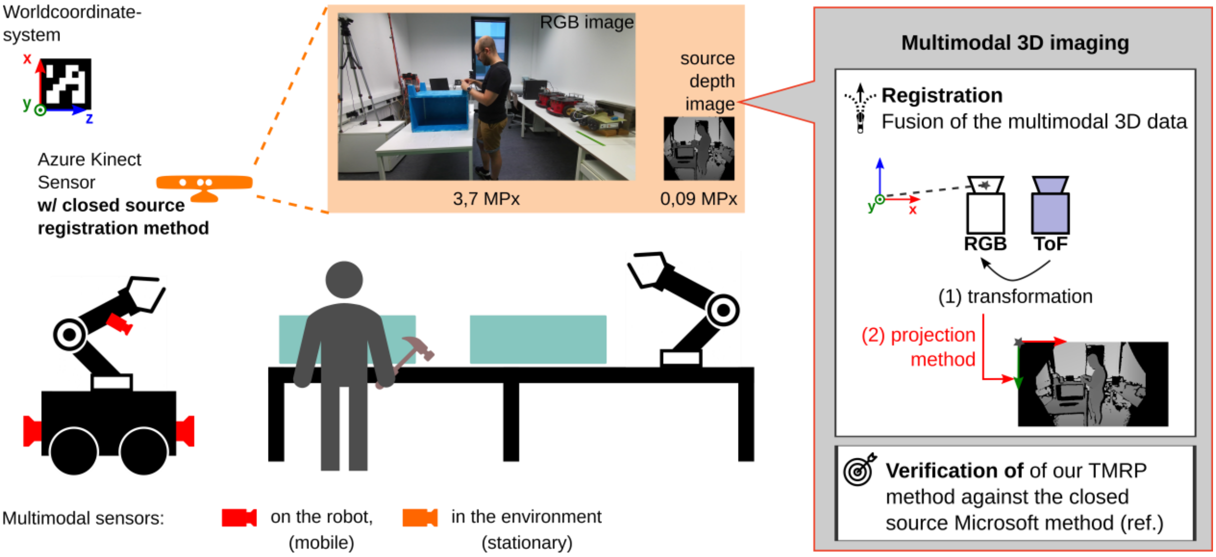


*Figure 2: **Robot-based assistance in manufacturing and production factories with multimodal sensors**. Key role: Registration based on Triangle-Mesh-Rasterization-Projection (TMRP) method and closed source Microsoft method as reference (ref.). Frame: 1624005922901327000 of ATTACH data set [23].*

(C) false neighbor in raster (neighboring 3D coordinates X/Y are less than one pixel apart) → fattening problem near depth discontinuities [22, 24, 25, 26]

(D) superposition of foreground and background in the target image (ambiguities) [27]

An overview of the different projection methods and their advantages and disadvantages are described in detail in our publication [21]. More complex projection methods have integrated up-sampling by interpolation. The accuracy of the up-sampling depends on the projection method used. Here a rough distinction is made between (i) *conventional methods*: [26, 27, 28, 22, 29, 30]; (ii) *polygon based method* [27, 6, 21]; and (iii) *deep learning based up-sampling*: [31, 22, 32, 25, 33, 34]. Depth sensor manufacturers usually provide **closed sourced Software Development Kits (SDK) or drivers** with Depth-to-RGB registration functions. Since registration of data is the starting point for various image analyses[3], we have developed our own projection method to be independent of sensor manufacturers. Our method is polygon based and is called Triangle-Mesh-Rasterization-Projection (TMRP) [21]. Our method is open source and freely available for a wide range of multimodal sensors.

1.3 The main contributions of our paper

To investigate the performance of our new Triangle-Mesh-Rasterization-Projection method, we compare it with the closed-source projection function from Microsoft (Azure Kinect Sensor SDK, v1.4.1) with respect to challenges (A)-(D). In addition, we show the transferability and long-term benefits of this method for robot-human interactions.

2. POLYGON BASED PROJECTION METHOD

Polygon based projection method is a research area with very little scientific literature [27]. Advantage of this method is the interpolation of all points, which is independent of the distance between the points of a triangle. With this method, the density is *XYZ* independent. The disadvantage is the higher computational effort compared to conventional methods. The best known representative is the Delaunay triangulation with nearest neighbor interpolation [6, 27]. Another representative is our Triangle-Mesh-Rasterization-Projection (TMRP) method [21]. Our interpolation is more accurate (s. details in [21]).

[3] e.g. object recognition; grip pose estimation [4]; action recognition [23]

3. METHODOLOGY

3.1 Overview

Our goal is to investigate the performance of our Triangle-Mesh-Rasterization-Projection (TMRP) method. For this, we compare it with the closed source projection function from Microsoft (Azure Kinect Sensor SDK, v1.4.1). Figure 3 shows our methodology using the Azure Kinect Sensor. Above is the image acquisition as well as the registration function based on the Azure-Kinect-Sensor Software Development Kit (SDK). Below is our open source registration. This consists of a coordinate transformation and our Triangle-Mesh-Rasterization-Projection (TMRP) method [21]. Intrinsic and extrinsic calibration parameters are required for registration.

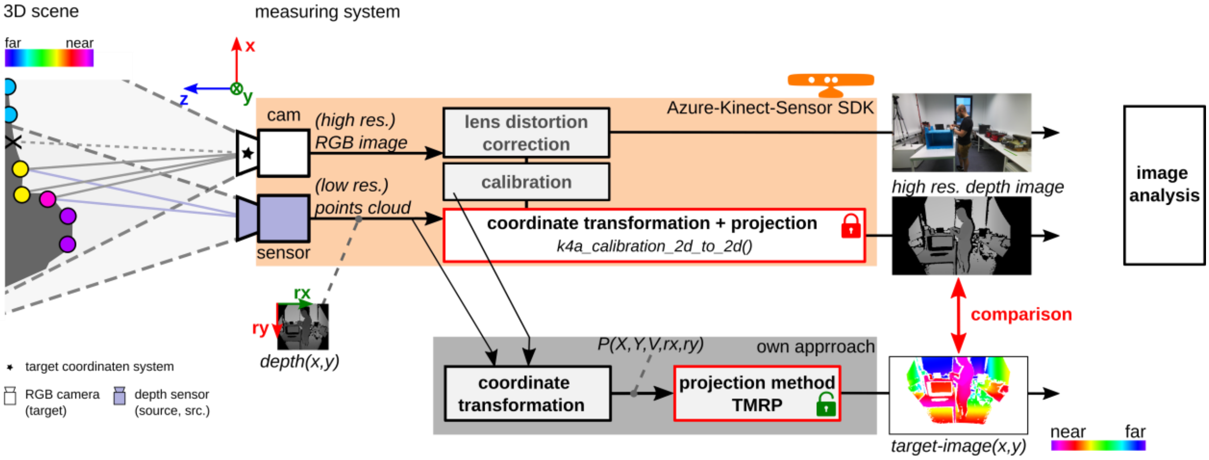


*Figure 3: **Overview of the pipeline based on the Azure Kinect Sensor.** Closed-source registration method based on Azure-Kinect-Sensor SDK (top) and an our open-source registration method based on our Triangle-Mesh-Rasterization-Projection (TMRP) method [21] (bottom).*

3.2 Our registration method

3.2.1 Coordinate transformation

Before projection, the low resolution depth image (2D) / low resolution point cloud (3D) must be transformed into the coordinate system of the RGB camera. In addition, the 2D neighborhood (rx, ry) of the low resolution depth image or point cloud must be determined in order to apply our TMRP method (section 2.2.2). Figure 4 shows the coordinate transformations of the depth sensor.

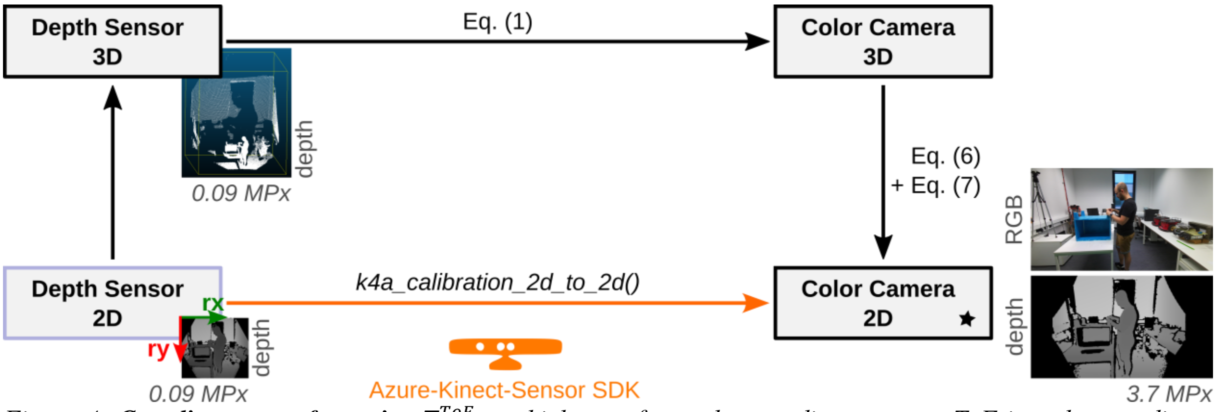


*Figure 4: **Coordinate transformation T_{Color}^{ToF}**, which transforms the coordinate system ToF into the coordinate system Color.*

Equation (1) describe the transformation of the low resolution point cloud (Depth Sensor 3D) into the Color Camera 3D coordinate system. The transformation is described by the rotation matrix $\mathbf{R}_{RGB\rightarrow depth}$ and the translation vector $\mathbf{t}_{RGB\rightarrow depth}$.

$$\begin{pmatrix} X_{\text{transf-depth-3D}} \\ Y_{\text{transf-depth-3D}} \\ Z_{\text{transf-depth-3D}} \\ 1 \end{pmatrix} = \begin{bmatrix} \mathbf{R}^{-1}_{RGB\rightarrow depth} & -\mathbf{R}^{-1}_{RGB\rightarrow depth}\mathbf{t}_{RGB\rightarrow depth} \\ 0_{1\times3} & 1 \end{bmatrix} \cdot \begin{pmatrix} X_{depth3D} \\ Y_{depth3D} \\ Z_{depth3D} \\ 1 \end{pmatrix} \tag{1}$$

Lens distortion correction is performed using radial distortion coefficients $k_1 - k_6$ (Eq. (3)) and tangential distortion coefficients p_1, p_2 (Eq. (4)) [35]. Equation
(6) describes the converted depth points into a 2D point cloud in the color camera 2D coordinate system.

$$X' = \frac{X_{\text{transf-depth-3D}}}{Z_{\text{transf-depth-3D}}}, \qquad Y' = \frac{Y_{\text{transf-depth-3D}}}{Z_{\text{transf-depth-3D}}}, \qquad r = \sqrt{X'^2 + Y'^2} \tag{2}$$

$$\delta_{rad} = \frac{1 + k_1 \cdot r^2 + k_2 \cdot r^4 + k_3 \cdot r^6}{1 + k_4 \cdot r^2 + k_5 \cdot r^4 + k_6 \cdot r^6} \tag{3}$$

$$\begin{aligned} \delta_{tan_x} &= 2 \cdot p_1 \cdot X' \cdot Y' + p_2 \cdot (r^2 + 2 \cdot X'^2) \\ \delta_{tan_y} &= p_1 \cdot (r^2 + 2 \cdot Y'^2) + 2 \cdot p_2 \cdot X' \cdot Y' \end{aligned} \tag{4}$$

$$X'' = X' \cdot \delta_{rad} + \delta_{tan_x}, \qquad Y'' = Y' \cdot \delta_{rad} + \delta_{tan_y} \tag{5}$$

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{bmatrix} f_{\text{x-RGB}} \cdot X'' + c_{\text{x-RGB}} \\ f_{\text{y-RGB}} \cdot Y'' + c_{\text{y-RGB}} \\ Z \end{bmatrix} \tag{6}$$

3.2.2 Triangle-Mesh-Rasterization-Projection

Our Triangle-Mesh-Rasterization-Projection (TMRP) method enables the generation of dense 2D raster images from multimodal 3D data $P(X, Y, V)$[4] with different source resolutions. Instead of considering only spatial coordinates, our method uses a 5-dimensional representation by combining source 2D raster information (rx, ry). The source 2D neighborhood information (rx, ry) can be acquired simultaneously with almost all 3D measurement methods. This additional information is used for neighborhood determination resp. triangulation (see Figure 5). Based on the determined 2D three- and four-neighborhoods, the triangular interpolation can be performed quickly. With this efficient polygon based up-sampling method, no false gaps and false neighbors are created in the target image. In addition, valid gaps in the original 3D survey are fully preserved. Valid gaps are imperfections that are present in the source image resp. source 3D point cloud due to the physical limitations of the source sensor technology. With a

[4] V...any modality

ToF sensor, valid gaps occur with optically uncooperative surfaces (transparency) or surfaces aligned parallel to the optical axis. Invalidation of values also occurs on the Azure Kinect Sensor when: (1) outside the active IR illumination mask; (2) saturated IR signal; (3) low IR signal; (4) filter outliers; and (5) multipath interference [36]. Due to the triangles, our method is also XYZ independent. Ambiguities are also taken into account by a filter (none/min/max), so that foreground and background are clearly separated. To fully understand the components of our TMRP, we wrote down the process and mathematics as pseudocode in the paper [21].

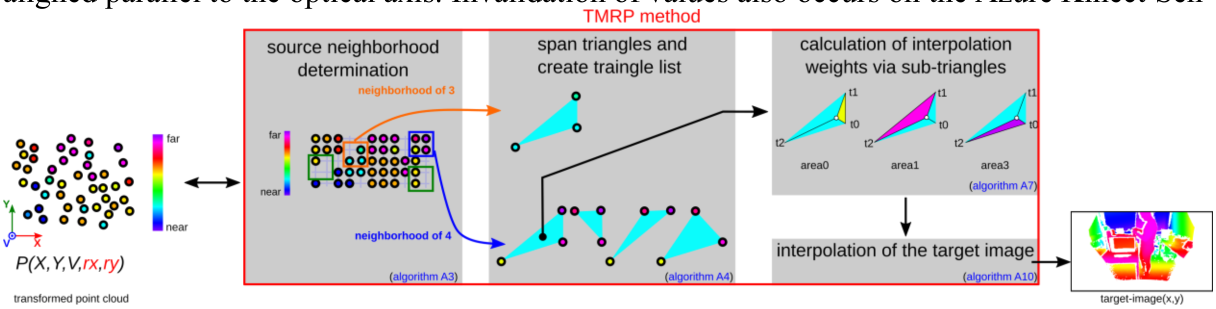


Figure 5: ***Brief overview of how our Triangle Mesh Rasterization Projection (TMRP) method works.*** *Input: transformed point cloud* (X, Y, Z, rx, ry) *with source raster information* (rx, ry)*. Algorithm reference refers to the publication [21].*

$$\text{target-image(x,y)} = TMRP\left(X, Y, V, rx, ry, \text{height}_{\text{target}}, \text{width}_{\text{target}}, \text{<filter>}\right) \qquad (7)$$

3.3 Azure Kinect SDK registration method

The transformation function *k4a_calibration_2d_to_2d()* [37] transform a 2D pixel coordinate with an associated depth value of the source camera into a 2D pixel coordinate of the target camera. Here, a triangle mesh is transformed from the geometry of the depth camera to the geometry of the color camera. The triangle mesh avoids incorrect gaps in the transformed depth image. "A Z-buffer ensures that occlusions are handled correctly. GPU acceleration is enabled for this function by default." [38]

4. SPECIFICATION OF UTILIZED DATA

Table 1 shows the specifications of the utilized data.

Table 1: Specifications of utilized data. Sensor: Azure Kinect (Microsoft).

Experiment	Source of data	Source image resolution (ToF sensor)	Target image resolution (RGB camera)
#1 in general	ATTACH data [23]	320 px × 288 px (0.09 MPx)	2560 px × 1440 px (3.7 MPx)
#2 test specimen	own data	640 px × 576 px (0.4 MPx)	4096 px × 3072 px (12.6 MPx)

5. EXPERIMENTS

5.1 Density and accuracy

Figure 6 shows the registered depth map based on the registration function of the closed source SDK (section 3.3) as well as based on our TMRP method (section 3.2). When registering the depth map ourselves using the calibration data, we do not get a congruent registered depth map. This is because the SDK internally uses three additional constants that are unknown to us: (i) scaling factor, (ii) x-axis offset and (iii) y-axis offset. Due to the unknown constants, we can only qualitatively evaluate the two registered depth maps. Figure 6 shows the qualitative comparison with regard to the challenges (A)-(D), s. section 1.2:

- **Ambiguities — challenge (D):** Both methods separate foreground and background.
- **False gaps**[5] **— challenge (B):** Figure 6 shows that both methods use efficient up-sampling, avoiding false gaps and producing 100 % dense areas. For comparison, a registered depth map based on the simple stand-of-the-technique projection method is also shown.

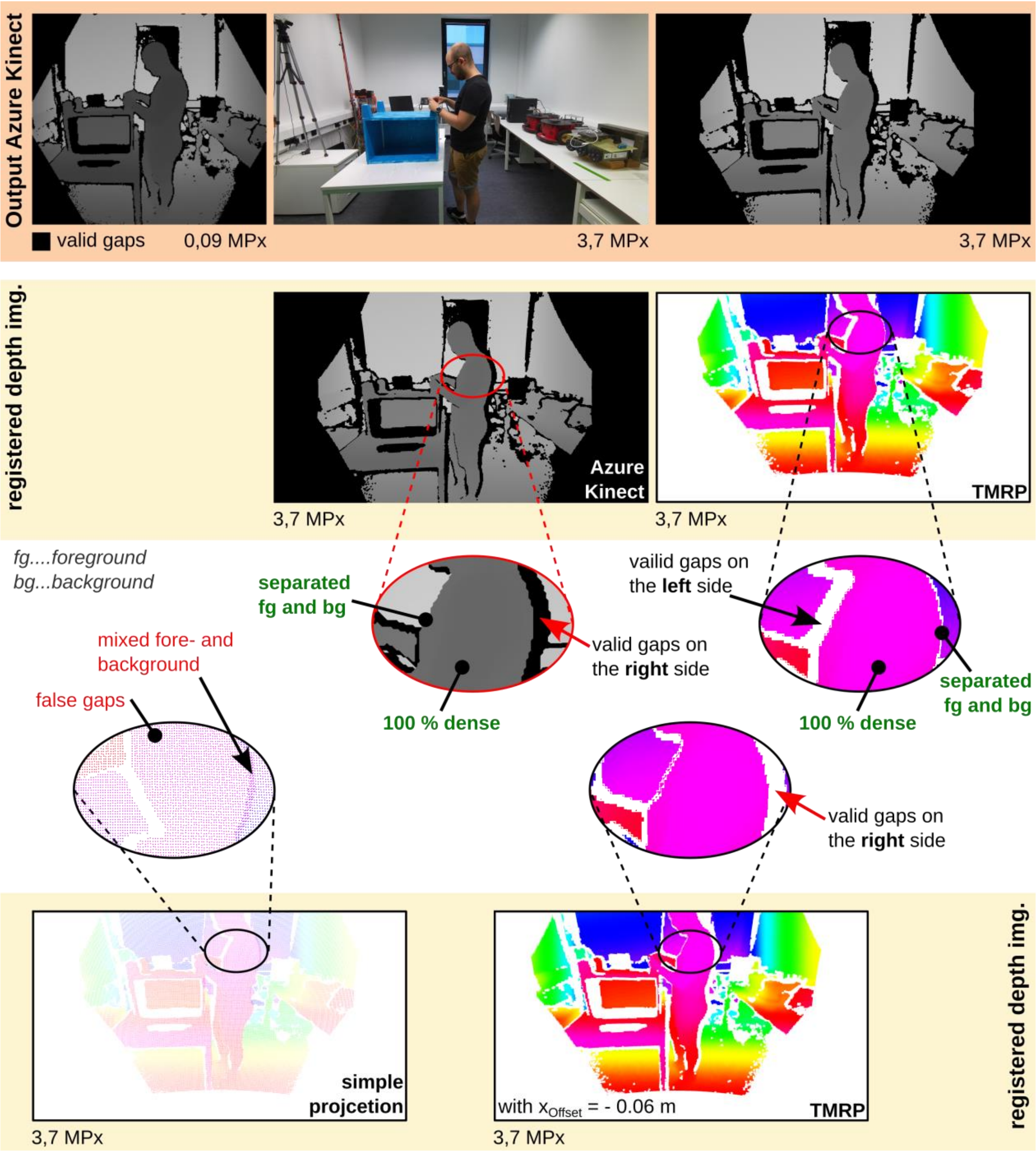


Figure 6: ***Qualitative comparison of the registered depth images (of experiment #1) based on closed source Azure Kinect SDK, TMRP and simple projection method.*** *(top) Output data of Azure Kinect: low resolution depth image, high resolution RGB image (undistorted) and registered depth image (undistorted). (bottom) Qualitative comparison on the basis of selected Region of Interest (ROI).*

[5] False gaps only occur in a registration where low-resolution source data is transformed and projected into a high-resolution target data.

- **Valid gaps[6] not fully considered & false neighbor— challenge (A) & (C):** Compared to the low-resolution depths, the registered depth map contains wider “shadows” resp. valid gaps at edges in the depth geometry such as the person's back (see Figure 6, ROIs). This is caused by the fact that objects are shifted by a different amount during the transformation depending on their distance from the camera. This can lead to a far away object being shifted more than a neighboring object nearby, creating a gap where no depth information is available.

  Figure 6 (expt. #1) shows that the valid gaps are at different positions. In the depth map based on the Microsoft method, the shadow is located at the back of the person (sagittal plane). In the depth map based on our TMRP method, it is on the front side of the person. This difference may be caused by the unknown constant (x-offset) (see Figure 6, (bottom-right)). With an x-axis offset, the "shadow" shifts to the right side, as in the Microsoft-based depth image.

  Figure 8 (expt. #2) shows differences in valid gaps / false neighbors. Our method considers valid gaps (see ROI, gaps size of 1 px). However, we cannot make a statement about the Microsoft method because we do not know whether a filter is applied afterwards to close small gaps.

Comparison of both interpolation methods:

Both projection methods are polygon based. To find out whether the same interpolation method is used, we look at the edges (Figure 7). Due to the three unknown constants, this consideration is also only possible qualitatively. The edges in the entire registered depth image (Microsoft) are exclusively smooth. The edges in the registered depth image (TMRP) are smooth and have

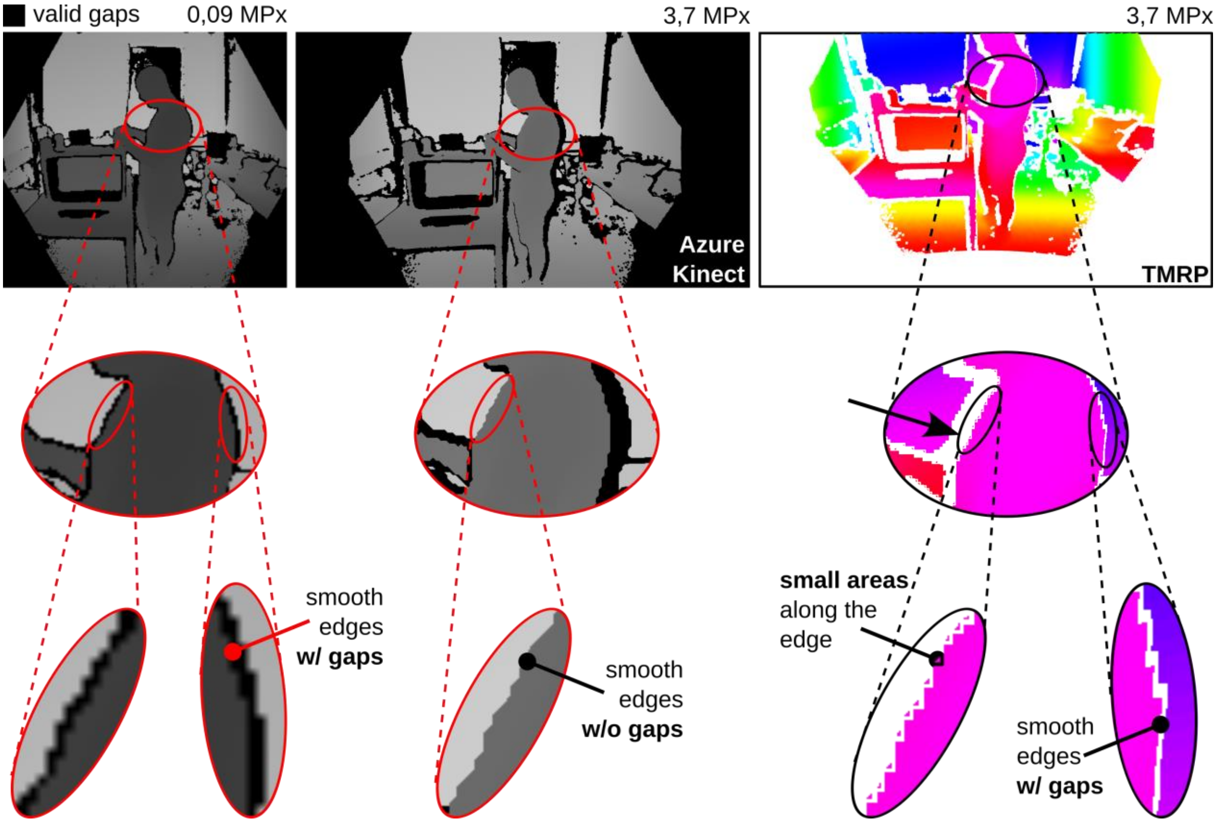


*Figure 7: **Qualitative examination of the edges to draw conclusions about the interpolation method** (images of experiment #1).*

[6] Invalidation of values occurs on the Azure Kinect Sensor when: (1) outside the active IR illumination mask; (2) saturated IR signal; (3) low IR signal; (4) filter outliers; and (5) multipath interference [36].

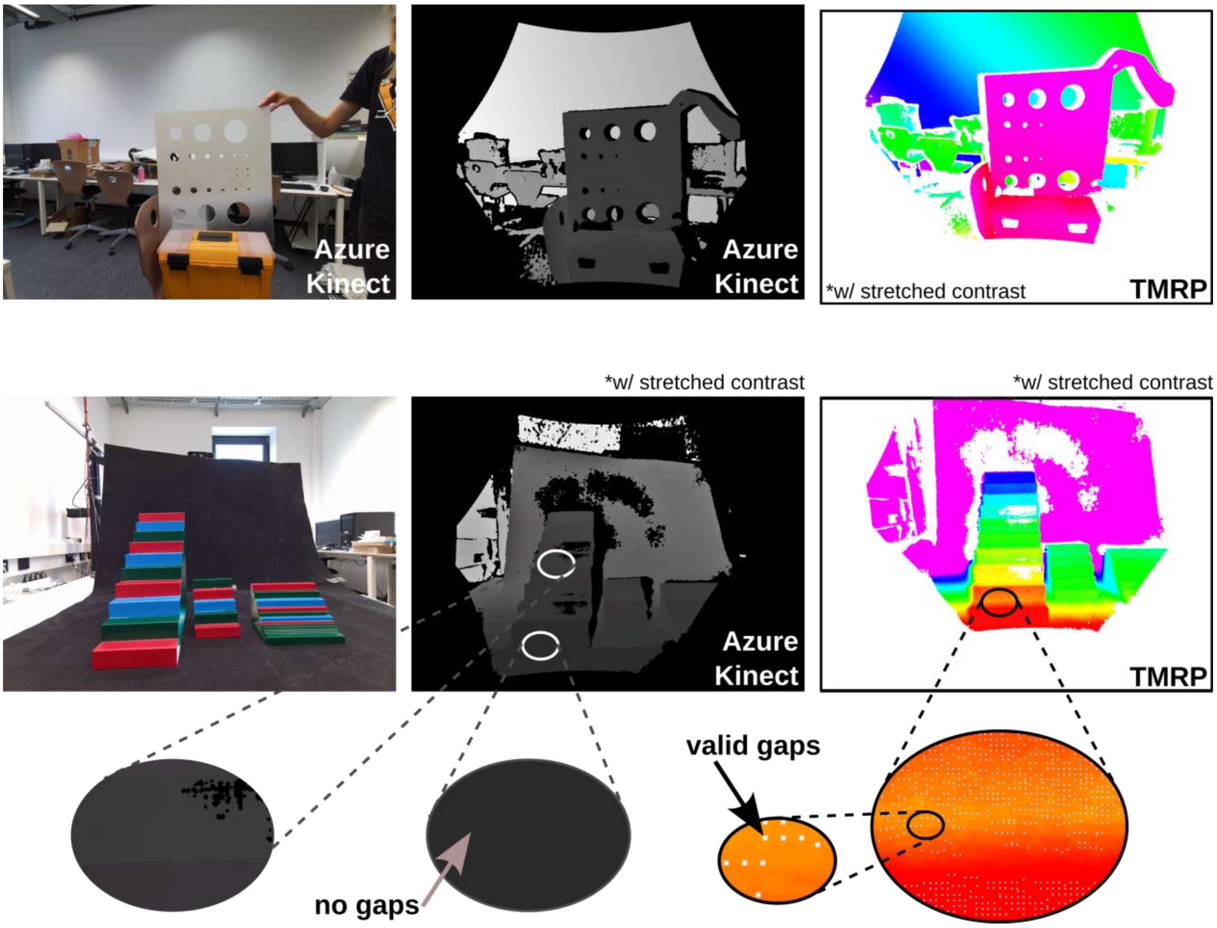


*Figure 8: **Qualitative comparison of the images from experiment #2.** Test specimen: (top) white matt plane with cutouts* $(r1/2/3\ = 3/4/5\ mm)$*; make with a laser-cutter. (bottom) RGB coloured staircase; made with an FDM printer (w/ painting). In the ROIs, the differences in terms of maintaining valid gaps/false neighbor become visible.*

either no or adjacent small areas (cf. Region-of-Interests in Figure 7). This indicates that either different interpolation methods are used or that Microsoft performs efficient edge-preserving smoothing afterwards. In the former case, our interpolation method would be more accurate (s. [21]).

5.2 Computation time and memory usage

Table 2 shows a quantitative comparison of the computation time and memory consumption of our TMRP method (sequential) at different target image resolutions. With approximately the

*Table 2: **Quantitative comparison of computation time and memory usage of our TMRP algorithm (sequential)**. Avg. computation time and max. resident set size (RSS) on processing unit (i9): Intel Core i9-7960X CPU @ 2.80 GHz and (i7): Intel Core i7-6700X CPU @ 4.00 GHz. **Input data:** transformed points: Figure 6 (#1) and Figure 8 (#2, top).*

expt.	**image resolution (MPx)**		**computation time**		**max. RSS**	**density (in %)**	
	source	target	uint (i9)	unit (i7)		visual	dense accurate
#1	0.09	3.7	0.5 s	0.4 s	244.1 MiB	39.6*	100
#2	0.4	12.6	1.9 s	1.7 s	804.9 MiB	40.4*	100

* *Why not 100 %? Valid gaps (A) are not taken into account here. The value is also dependent on the three unknown constants.*

same visual density of the target image (experiment #1 and #2), quadrupling the number of pixels of the source and target images increases the computation time of the algorithm by a factor of four. The TMRP algorithm thus has a linear time complexity in relation to the image resolution.

6. CONCLUSION, LIMITATION, TRANSFERABILITY AND FUTURE WORK

6.1 Conclusion

To ensure safe and trustful cooperation in robot-human interactions (e.g. collaboration or teaming), multimodal 3D imaging is required. The key role of multimodal 3D technology is to gain a more comprehensive understanding of the production as well as to enable a trustful cooperation of the teams. For image analysis, the data must be registered, i.e. transferred into a coherent coordinate system. The registration consists of two steps: the coordinate transformation (s. Figure 4) of the source data into the target coordinate system and the projection of the point cloud into the target image. One challenge here is the different resolutions of the sensors. Thermal imaging sensors, for example, have a low resolution because the special semiconductor chips required are very expensive. To solve this challenge, we have developed our own polygon-based projection method, named Triangle-Mesh-Rasterization-Projection (TMRP). Our method is also independent of sensor and modality. To verify the performance of our TMRP method, we compare our registered depth maps with those generated by Microsoft's closed source projection function (Azure Kinect SDK). The results were examined in terms of (i) accounting for valid gaps, (ii) avoiding false gaps, (iii) avoiding false neighbors, (iv) separating foreground and background, and (v) accounting for XYZ dependence of density.

Due to unknown constants (scaling, x-offset and y-offset) in the coordinate transformation, our resulting planar point cloud differs from the Azure Kinect Sensor (AKS) based one. Therefore, our methods can only be compared qualitatively. Table 3 shows a comparison of the results. Research has shown that in qualitative comparison with Microsoft's Azure Kinect Sensor SDK, our algorithm achieves almost identical results in terms of (ii) avoiding false gaps, (iv) separating foreground and background and (v) accounting for XYZ dependence of density (s. Figure 6). Due to the three unknown constants (s. Figure 7), we cannot say with 100 % certainty whether valid gaps are fully considered (i) and false neighbors are generated (iii). Looking at the edges, it is noticeable that either (I) both methods use a different interpolation or that (II) Microsoft performs additional edge preserving smoothing and filtering afterwards. In the first case (I), our method is even more accurate. Since our interpolation method (s. Figure 5) refers to the 2D neighborhoods and includes one (for neighborhoods of 3) or four (for neighborhoods of 4) interpolation weights depending on the neighborhood.

Summary: The qualitative comparison showed that both methods produce almost identical results. Minimal differences at the edges indicate that both projection methods either use slightly different interpolation methods or Microsoft performs efficient edge-preserving smoothing and filtering afterwards. With the former, our TMRP interpolation would be more accurate. With the TMRP method, there is now a **freely-available open source**[7] projection method that can be **applied to almost any multimodal 3D / 2D image dataset** and not like the Microsoft SDK optimized for Microsoft products.

[7] https://github.com/QBV-tu-ilmenau/Triangle-Mesh-Rasterization-Projection

Table 3: ***Qualitative comparison of Azure-Kinect Sensor (AKS) projection method and our TMRP method.*** *(left-to-right) State-of-the-art projection (SOTA proj.), closed source AKS SDK (v1.4.1) projection function and our TMRP method.*

Properties	**SOTA proj.**	**AKS SDK** (Microsoft)	**TMRP** (our)
Creates false neighbors in raster (A) & (C)	low	(low)	**never**
Creates false gaps in raster, (B)	often	**never**	**never**
Resolution of ambiguities, (D)	no	**yes**	**yes**
Density is independent of $\boldsymbol{XYZ}$	no	**yes**[†]	**yes**[†]
Required input	**w/o** rx, ry	w/ rx, ry	w/ rx, ry
Computing effort	**low**	middle-high	high[‡]
Frame Rates	-	30/15 fps[§] *Azure Kinect Sensor SoC*	approx. 2/0.5 fps *Intel Core i7/i9*
Code available	**open source, free**	closed source	**open source, free** [21]
For any sensor & any modality	**yes**	no	**yes** [21]

[†] because this method interpolates all points regardless of the distance between the points of a triangle
[‡] however highly parallelizable
[§] incl. image capture and other things

6.2 Limitation

Currently, our TMRP method [21] cannot be used in real-time applications because it requires a relatively high computing time due to its serial implementation. However, this limitation can be overcome by parallelization, as our TMRP algorithm is highly parallelizable.

6.3 Transferability

Our efficient TMRP method is applicable to any 3D/2D sensors and modalities [21]. We believe that this method will also enrich the field of human-robot interaction in the long term. Figure 9 shows possible robot applications.

- Robot perception / robot manipulation ability in general
- To ensure trustful **human-robot collaboration and teaming**, additional thermal cameras are used [4, 10]. Thermal cameras usually have low resolution because the special semiconductor chips required are very expensive.
- **In the forest sector:** monitoring; litter collection or wood mass determination; forest renewal

6.4 Future work

- In the future, we want to parallelize our current serial algorithm. Parallelizing the algorithm will significantly reduce the processing speed so that our TMRP method can be integrated into various real-time applications in the future.

- In cooperation with the Fraunhofer Institute for Applied Optics and Precision Engineering we are currently developing a new measurement principle *TranSpec3D* to generate a stereo data set for transparent and specular surfaces in the VIS range without object painting. Our TMRP method is used in this measurement principle.

ACKNOWLEDGMENTS

This research was funded by the Carl-Zeiss-Stiftung as part of the project Engineering for Smart Manufacturing (E4SM) – Engineering of machine-learning-based assistance systems for data-intensive industrial scenarios. Many thanks to my project partner Mr. Benedict Stephan for the joint data acquisition of the test specimens as well as for providing the ATTACH data set [23].

REFERENCES

[1] R. Illmann, M. Rosenberger and G. Notni, "Multi-channel supported surveying of industrial floor tiles," in *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2020.

[2] R. Illmann, "Robotik im Handwerk - Fliesenlegeautomat," *5. Thüringer Maschinenbautag,* 2021.

[3] C. Junger, Y. Zhang and G. Notni, *Forschungsblog zur Mensch-Roboter-Kollaboration,* 2022. https://www.tu-ilmenau.de/en/university/departments/department-of-mechanical-engineering/profile/institutes-and-groups/group-for-quality-assurance-and-industrial-image-processing/research/mensch-roboter-kollaboration

[4] Y. Zhang, S. Müller, B. Stephan, H.-M. Gross and G. Notni, "Point Cloud Hand–Object Segmentation Using Multimodal Imaging with Thermal and Color Data for Safe Robotic Object Handover," *Sensors,* vol. 21, 2021.

[5] Institut für angewandte Arbeitswissenschaft e. V., "Mensch-Roboter-Kollaboration (MRK)", 2023. https://www.arbeitswissenschaft.net/fileadmin/Bilder/Angebote_und_Produkte/ifaa_Einsatz_kollaborierender_Roboter_Interaktionsformen.png

[6] A. Asvadi, L. Garrote, C. Premebida, P. Peixoto and U. J. Nunes, "Multimodal vehicle detection: fusing 3D-LIDAR and color camera data," *Pattern Recognition Letters,* vol. 115, pp. 20-29, 2018.

[7] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld and H.-M. Gross, "Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[8] Z. Zheng, D. Xie, C. Chen and Z. Zhu, "Multi-resolution Cascaded Network with Depth-similar Residual Module for Real-time Semantic Segmentation on RGB-D Images," in *2020 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, 2020.

[9] Y. Zhang, xianda guo, M. Poggi, Z. Zhu, G. Huang and S. Mattoccia, *CompletionFormer: Depth Completion with Convolutions and Vision Transformers,* 2023.

[10] Y. Zhang, R. Fütterer and G. Notni, "Interactive robot teaching based on finger trajectory using multimodal RGB-D-T-data," *Frontiers in Robotics and AI,* vol. 10, 2023.

[11] P. Kolar, P. Benavidez and M. Jamshidi, "Survey of Datafusion Techniques for Laser and Vision Based Sensor Integration for Autonomous Navigation," *Sensors,* vol. 20, 2020.

[12] L. Svoboda, J. Sperrhake, M. Nisser, C. Zhang, G. Notni and H. Proquitté, "Contactless heart rate measurement in newborn infants using a multimodal 3D camera system," *Frontiers in Pediatrics,* vol. 10, 2022.

[13] C. Zhang, I. Gebhart, P. Kühmstedt, M. Rosenberger and G. Notni, "Enhanced Contactless Vital Sign Estimation from Real-Time Multimodal 3D Image Data," *Journal of Imaging,* vol. 6, 2020.

[14] E. Gerlitz, M. Greifenstein, J.-P. Kaiser, D. Mayer, G. Lanza and J. Fleischer, "Systematic Identification of Hazardous States and Approach for Condition Monitoring in the Context of Li-ion Battery Disassembly," *Procedia CIRP,* vol. 107, pp. 308-313, 2022.

[15] J. Jiang, G. Cao, J. Deng, T.-T. Do and S. Luo, *Robotic Perception of Transparent Objects: A Review,* 2023.

[16] S. S. Sajjan, M. J. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng and S. Song, "Clear Grasp: 3D Shape Estimation of Transparent Objects for Manipulation," *2020 IEEE International Conference on Robotics and Automation (ICRA),* pp. 3634-3642, 2019.

[17] Y. Tang, J. Chen, Z. Yang, Z. Lin, Q. Li and W. Liu, "DepthGrasp: Depth Completion of Transparent Objects Using Self-Attentive Adversarial Network with Spectral Residual for Grasping," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[18] F. Erich, B. Leme, N. Ando, R. Hanai and Y. Domae, *Learning Depth Completion of Transparent Objects using Augmented Unpaired Data,* 2023.

[19] M. Landmann, S. Heist, P. Dietrich, H. Speck, P. Kühmstedt, A. Tünnermann and G. Notni, "3D shape measurement of objects with uncooperative surface by projection of aperiodic thermal patterns in simulation and experiment," *Optical Engineering,* vol. 59, p. 094107, 2020.

[20] D. Lahat, T. Adalý and C. Jutten, "Challenges in multimodal data fusion," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014.

[21] C. Junger, B. R. Buch and G. Notni, "Triangle-Mesh-Rasterization-Projection (TMRP): An algorithm to project a point cloud onto a consistent, dense and accurate 2D raster image," *Sensors,* vol. 23, 2023.

[22] J. You and Y.-K. Kim, "Up-Sampling Method for Low-Resolution LiDAR Point Cloud to Enhance 3D Object Detection in an Autonomous Driving Environment," *Sensors,* vol. 23, 2023.

[23] D. Aganian, B. Stephan, M. Eisenbach, C. Stretz and H.-M. Gross, *ATTACH Dataset: Annotated Two-Handed Assembly Actions for Human Action Understanding,* 2023.

[24] Y. Li, T. Xue, L. Sun and J. Liu, "Joint Example-Based Depth Map Super-Resolution," in *2012 IEEE International Conference on Multimedia and Expo*, 2012.

[25] Q. Yang, R. Yang, J. Davis and D. Nister, "Spatial-Depth Super Resolution for Range Images," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[26] J. Kopf, M. F. Cohen, D. Lischinski and M. Uyttendaele, "Joint Bilateral Upsampling," *ACM Trans. Graph.,* vol. 26, p. 96–es, July 2007.

[27] C. Premebida, L. Garrote, A. Asvadi, A. P. Ribeiro and U. Nunes, "High-resolution LIDAR-based depth mapping using bilateral filter," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2016.

[28] S. Fadnavis, "Image Interpolation Techniques in Digital Image Processing: An Overview," *Int. Journal of Engineering Research and Applications,* vol. 4, pp. 70-73, 2014.

[29] L. Chen, Y. He, J. Chen, Q. Li and Q. Zou, "Transforming a 3-D LiDAR Point Cloud Into a 2-D Dense Depth Map Through a Parameter Self-Adaptive Framework," *IEEE Transactions on Intelligent Transportation Systems,* vol. 18, pp. 165-176, 2017.

[30] M. R. Akhtar, H. Qin and G. Chen, "Velodyne LiDAR and monocular camera data fusion for depth map and 3D reconstruction," in *Eleventh International Conference on Digital Image Processing (ICDIP 2019)*, 2019.

[31] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox and A. Geiger, "Sparsity Invariant CNNs," in *2017 International Conference on 3D Vision (3DV)*, 2017.

[32] J. Chang and Y. Chen, "Pyramid Stereo Matching Network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, 2018.

[33] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao and A. Geiger, "Unifying Flow, Stereo and Depth Estimation," *arXiv preprint arXiv:2211.05783,* 2022.

[34] Y. Zhang and T. Funkhouser, "Deep Depth Completion of a Single RGB-D Image," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[35] openCV, *Camera Calibration and 3D Reconstruction,* 2014.

[36] Microsoft Corporation, *Tiefenkamera in Azure Kinect DK,* 2023. *https://learn.microsoft.com/de-de/azure/kinect-dk/depth-camera*

[37] Microsoft Corporation, *Documentation for k4a_calibration_2d_to_2d(),* 2022. *https://microsoft.github.io/Azure-Kinect-Sensor-SDK/master/group___functions_ga3b6bf6dedbfe67468e2f895dcce68ed4.html#ga3b6bf6dedbfe67468e2f895dcce68ed4*

[38] Microsoft Corporation, *Use Azure Kinect Sensor SDK image transformations,* 2022. *https://learn.microsoft.com/en-us/azure/Kinect-dk/use-image-transformation*

[39] R. Illmann, R. Fütterer, M. Rosenberger and G. Notni, "Investigation into the implementation of a multimodal 3D measurement system for a forestry harvesting process," Ilmenau Scientific Colloquium, 2023.

[40] D. Müller, "Pfadplanung und Simulation eines robotergesteuerten Kalibrierprozesses für ein Multi-View-Stereosystem," Master thesis, TU Ilmenau, 2022.

CONTACTS

Christina Junger	email: christina.junger@tu-ilmenau.de ORCID: https://orcid.org/0000-0002-5310-495X
Univ.-Prof. Dr. rer. nat. Gunther Notni	email: gunther.notni@tu-ilmenau.de

APPENDIX A. ROBOT APPLICATIONS

Figure 9 shows robotic applications of our department.

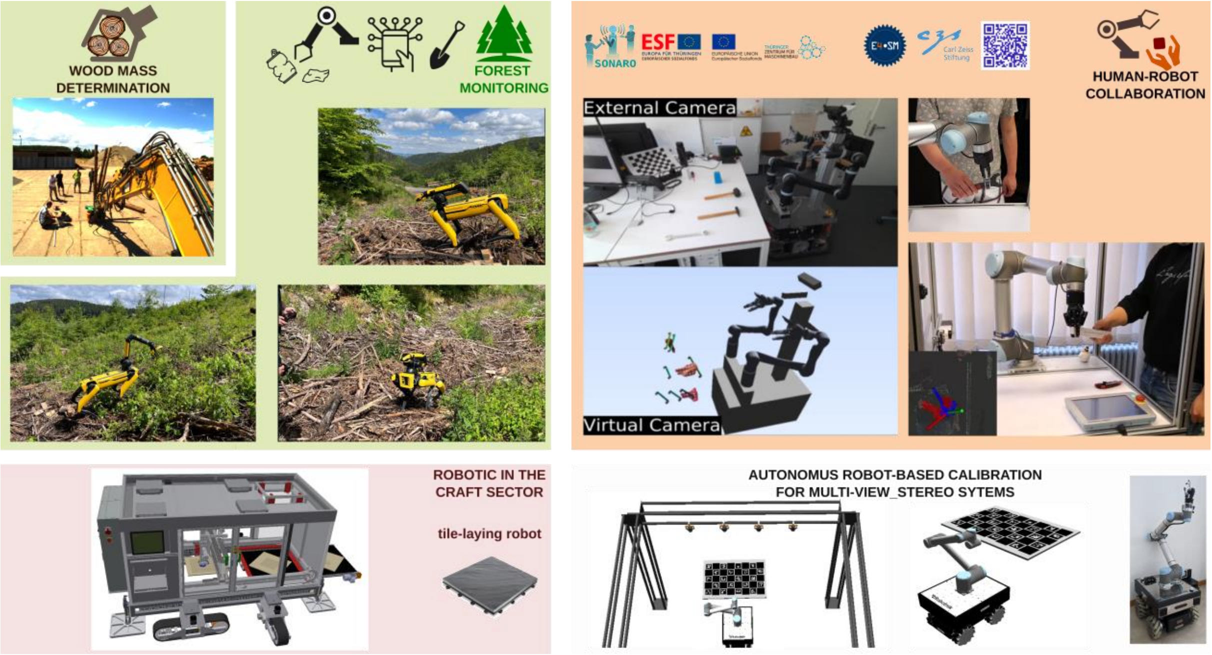

WOOD MASS DETERMINATION
FOREST MONITORING
SONARO
ESF
HUMAN-ROBOT COLLABORATION
External Camera
Virtual Camera
ROBOTIC IN THE CRAFT SECTOR
tile-laying robot
AUTONOMUS ROBOT-BASED CALIBRATION FOR MULTI-VIEW_STEREO SYTEMS

Figure 9: ***Robotic applications at our department.*** *(top-left) Current [39] and future work in the forest sector; Boston Dynamics spot for forest monitoring or litter collection along the highway (image source: Cooperation with FH Erfurt (Mr. P. Voigt) and Holz21 Regio https://www.holz-21-regio.de/). (top-right) human-robot collaboration in assembly process (image source: project partner in E4SM https://www.e4sm-projekt.de/: Mr. B. Stephan, Department of Neuroinformatics and Cognitive Robotics, TU Ilmenau) and safe Robotic in HRC [10, 4] (bottom-left) Robotic in the craft sector: Tile-laying robot [1, 2]. (bottom-right) Autonomous robot-based calibration of multi-view stereo systems [40].*