The 12th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
7-9 September, 2023, Dortmund, Germany

Stereo Vision and LiDAR Based Point Cloud Acquisition for Creating Digital Twins in Indoor Applications

Alexander Kuzminykh, Jerome Rohde, Phillip Oliver Gottschewski-Meyer, Volker Ahlers
University of Applied Sciences and Arts Hannover, Germany
{alexander.kuzminykh, jerome.rohde, phillip-oliver.gottschewski, volker.ahlers}@hs-hannover.de

***Abstract* — We present an approach towards a data acquisition system for digital twins that uses a 5G network for data transmission and localization. The current hardware setup, which utilizes stereo vision and LiDAR for 3D mapping, is explained together with two recorded point cloud data sets. Furthermore, a resulting digital twin comprised of voxelized point cloud data is shown. Ideas for future applications and challenges regarding the system are discussed and an outlook on further development is given.**

Keywords* — *stereo vision, lidar, 3d mapping, point clouds, digital twins, voxelization

I. Introduction

The fifth generation of cellular technology (5G) enables the implementation of use cases that could not be optimally covered utilizing previous standards. Some of the enhancements coming with 5G over 4G are increased bandwidth, higher coverage, lower latency and a greater capacity for connected devices. Another aspect of 5G is the possibility of accurate device localization, which is especially interesting for indoor scenarios [1]. These aspects propose to leverage 5G for digital twins.

A digital twin that is intended for purposes like area management in logistics or navigation of delivery drones requires a large amount of spatio-temporal sensor data to be transmitted [2]. It is necessary to have a near real-time space representation with high resolution, accuracy and coverage for time-critical and security-related use cases. High coverage implies sensors to be effective in different environmental conditions and to be able to move in space for varying perspectives. These requirements present a challenging task when the overall system must exhibit the flexibility to adapt to highly dynamic scenarios, all while maintaining an affordable price point. High cost as well as technical feasibility prevent organizations from benefiting from the advantages of using digital twins [3], [4]. To participate in this part of digital transformation, especially small and medium-sized enterprises (SMEs) need to find cost-effective and easy-to-use solutions [5].

This work was financially supported by the German Federal Ministry for Digital and Transport (BMDV), project *5GAPS* (grant no. 45FGU121_L).

The research project *Access to Public Spaces via 5G* (5GAPS) aims at the integration of 5G with use cases of digital twins via proof-of-concept applications. For that purpose, continuously localized sensors shall be used for real-time indoor space monitoring by delivering the required spatio-temporal data over a 5G network. The recorded data is sent to a database system which operates using a voxel grid structure, where each voxel represents a distinct unit of data with various attributes such as color or an assigned object type [6]. In addition to managing the current state of space, the database also stores previous states. By leveraging the historical data, we gain the ability to make informed predictions about future trends, patterns, and behavior that contribute to improved decision-making processes. Example indoor applications that are targeted by our project are the monitoring of the setup of industry fairs (cf. Sec. III-C) and the support of rescue forces in public buildings. Potential outdoor applications include the managing of parking lots, area management and the detection of environmental changes.

II. Related Work & Contribution

In recent years several application scenarios for digital twins of public or private spaces have been described and tested. In their extensive review with a focus on smart city applications Botín-Sanabria et al. identify five principal challenges, among them the high implementation costs due to the increased amount of sensors required for complex environments as well as the limited availability of 5G networks [3].

In most digital twin approaches rather expensive light detection and ranging (LiDAR) sensors are used, e.g., [7], [8]. Taurino and Villa, however, point out the necessity of cost-effective solutions to capture data for the creation of digital twins especially in SMEs [9].

Minos-Stensrud et al. thus present an approach using stereo cameras mounted to unmanned aerial vehicles (UAVs or drones) to create digital twins of SME factories [5]. They conclude that using stereo cameras instead of LiDARs the costs can be reduced by a considerable amount. Shi et al. use stereo cameras to create a digital twin for surveillance tasks [10]

DOI of the original publication: https://doi.org/10.1109/IDAACS58523.2023.10348709

Our approach follows a similar direction in using low-cost stereo cameras to create and update digital twins of public spaces. The stereo camera measurements are mapped to a ground truth captured once by LiDAR measurements. In addition, we mount the stereo cameras to regularly moving vehicles like forklift trucks in fair halls or warehouses and buses, thus allowing the digital twin to be updated with minimal extra effort.

III. Data Acquisition

Data plays an important role in testing the basic functionality of the mentioned database system. For that purpose, we have collected two data sets A and B, which serve various objectives. These objectives include assessing the efficiency of data transmission over 5G networks, evaluating the effectiveness of mapping and localization algorithms, and validating the capabilities of object recognition. Although the transmission and processing is tested with a previously collected data set instead of live sensor data, we expect to derive important information for further development of the database system and algorithms. At this stage of the project, one aspect yet to be determined is the size of voxels, as it significantly influences the performance requirements on all system components. However, we anticipate that voxels will have an edge length ranging from 0.1 to 0.3 meters. It is important to note that the edge length may vary due to the unique characteristics of potential use cases within and beyond the project scope.

There are two contexts in which we collect data: the static and the dynamic state of space. In a static state, we measure static geometry, like the ground and the walls of an empty room. Such data can be used as reference geometry in localization algorithms and, thus, must be recorded with high accuracy. This is especially important if no 5G localization is available, as was the case when acquiring data sets A and B. Hence, point cloud-based localization is an integral part of our work. Dynamic geometry, e.g., a moving object in a room, is measured in a dynamic state, which must be done highly frequently over time, so that object movement can be retraced. For that purpose, object attributes like identity are also determined in a dynamic state. Both contexts require appropriate scene coverage to fulfill their purpose, which implies that sensors are numerous and mobile. Combining the initial static state of a scene with the subsequent data collection of the dynamically changing state yields the basis for a digital twin.

A. Sensors

Data sets A and B were created using active stereo depth cameras and a spinning, scanning LiDAR sensor, which will be called cameras and lidar, respectively, from now on. The sensors were connected to laptop computers for data storage. Our decision for Intel RealSense D455 stereo cameras was driven by their cost-effectiveness. Moreover, these cameras are preferred for the inclusion of an infrared projector, which greatly enhances stereo vision performance in low-light environments and on surfaces lacking distinct features. The cameras record at a high frame rate and also provide color information of the measured geometry, making them applicable for usage in dynamic scenarios. However, due to the depth estimation method and sensor noise, the accuracy and range are limited. Based on our observations, we have found that the measured depth values tend to deviate, on average, by one voxel edge length from the actual geometry beginning at a distance of around 5 meters. Therefore, the voxel representation of geometry at such distances loses its connectivity. We intend to solve this problem by combining multiple sensors.

In order to effectively monitor the dynamic changes in the environment, the cameras were set to a frame rate of 30 frames per second (FPS), reducing motion blur but increasing the amount of redundant data. This frame rate selection aligns with the representation of objects as voxels in the database. Considering the expected relative movement speeds, the chosen FPS value prevents object translations by more than one voxel edge length between consecutive frames in most indoor use cases. However, if no relative movement is present, the recorded frames are redundant. This could be addressed by adjusting FPS based on observed or predicted movement. The cameras were set to record with 1280×720 pixel resolution for color and 640×360 for depth. The depth resolution was halfed for various reasons, especially to reduce data set size, but we intend to record at same resolution for depth as for color in a live sensor streaming setup. As the data is mapped to voxels, increasing resolution would not necessarily yield benefits beyond a certain point; however, stereo vision benefits from a higher resolution in general. Similar to a dynamic adjustment of FPS, dynamic adjustment of resolution might be beneficial. Each camera was fully utilizing the build-in infrared dot pattern projector and set to the highest accuracy preset in the Intel RealSense Viewer for estimating depth.

The lidar sensor used by us is the Velodyne Puck LITE. It has a range of around 100 meters, making it suitable for capturing data in large, open spaces such as empty halls or long pathways. Compared to the stereo cameras, the lidar has a lower frame rate but measures points with higher accuracy and, hence, is used to record reference data. More specifically, the reason not to use multiple lidar to monitor dynamic state are the high cost per device and the lack of ability to measure color with most if not all entry level devices. Another reason is the reduced coverage compared to cameras. While a camera captures light from an extended solid angle in a direction, the lidar measures geometry with a small number of point lines and, therefore, relies on relative sensor movement for good coverage, making it less flexible.

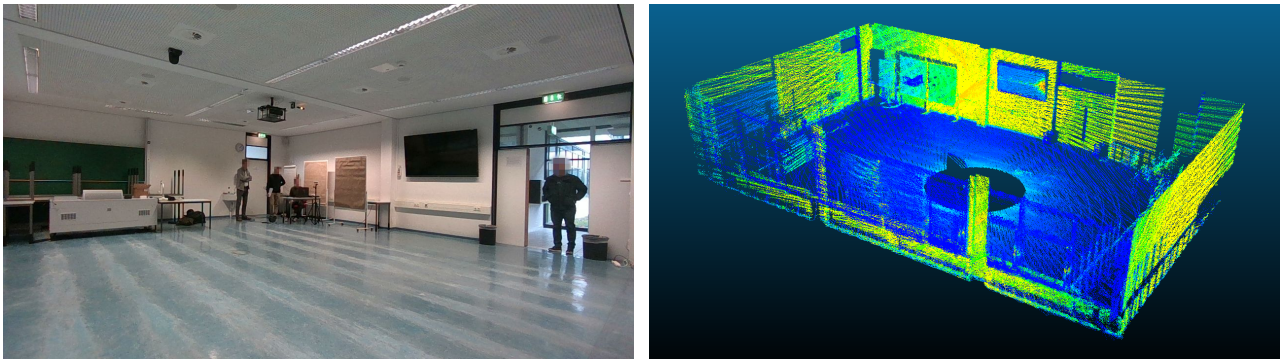

Figure 1. The captured room in data set A seen by a camera (left) and mapped with the lidar from one position (right). The lidar point cloud is colored by the reflectivity of measured geometry.

B. Data Set A

This data set covers an university room where a student project poster exhibition takes place. It includes point clouds of the empty room, i.e., the static state, collected with the lidar and shown by Figure 1. Further, it includes the subsequent build-up of poster stands, followed by the actual event, i.e., the dynamic state, recorded by the cameras. Before any mapping was done, the room dimensions were manually measured using a laser measuring tool. The mapping process via lidar was conducted at five manually localized places as no 5G localization was available. The localization was done relative to the room as the coordinate space, where the walls and ground served as the axis planes.

The data collection involved several steps. Initially, we utilize LidarView [11] to collect data at each of the five locations, saving it for further processing. Once data collection is complete at each location, LidarView's tools are utilized to generate a trajectory using the simultaneous localization and mapping (SLAM) filter. This trajectory serves as the reference to align each recorded frame, resulting in the creation of a dense point cloud, cf. Figure 1. In order to merge all five subset point clouds, each is relocated to approximately align with the manually measured location within an unified coordinate space. Then, a single point cloud is selected as a reference, and the remaining point clouds are automatically aligned to it using registration tools of CloudCompare [12]. This registration process results in the creation of a merged and highly dense reference point cloud. This point cloud is oriented using MeshLab [13] so that the walls and ground align with the axis planes. Finally, the result is subsampled to remove unnecessarily high density at certain areas. This process is tedious but, compared to a single measurement including the movement of the lidar, avoids distortions. The reader may find a single measurement enough for other applications.

Each camera was mounted on a stand and placed at one location in the room where it stayed for the rest of the recording session. Also, the cameras were oriented towards the room's middle area so that moving objects can be seen from different perspectives. Like with the lidar, the cameras were localized manually so that the collected point clouds can later be placed in the reference point cloud. The recording covers a time span of around three and a half hours resulting in over two terrabytes of data, making it especially interesting for stress testing the voxel database. We determined empirically that the accuracy of the recorded data is acceptable to a distance of 5 meters. However, data beyond that limit was captured anyways and is shown in Figure 2. The illustrated distortion, which grows with distance, limits the capability of point cloud based localization using stereo vision. Supportive information like estimated location via 5G or inertial measurement unit (IMU) data is required. As only three fixed cameras were used and each camera is only able to observe a relatively small volume of the room, the data set focuses on the middle.

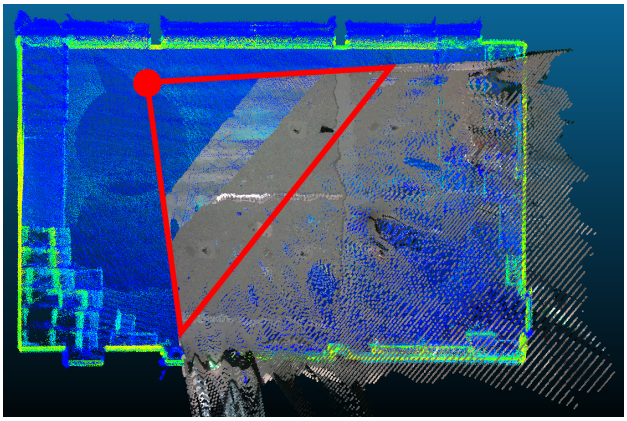

Figure 2. The reference point cloud seen from above with a camera point cloud. The red triangle illustrates the horizontal viewing angle and the distance of 5 meters in the negative z-direction in camera coordinates. Notice the distortion of the camera point cloud compared to the reference point cloud close to the wall.

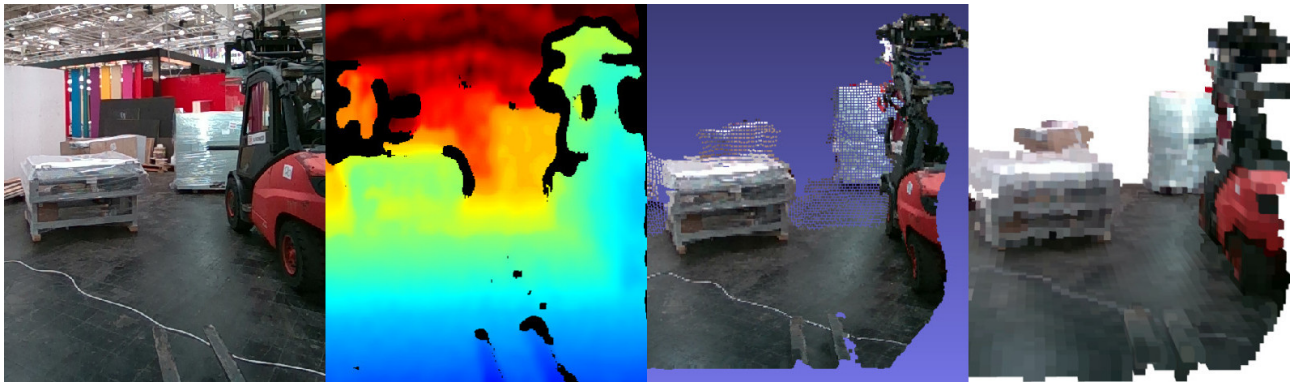

Figure 3. Color (left) and depth (center left) streams of one of the stereo cameras mounted to a forklift together with the produced point cloud (center right) and the voxel representation (right).

C. Data Set B

The second data set covers a small time span of the build-up of a large industry fair. More specifically, it shows the on-going build-up work towards the end within multiple fair halls. The main difference to data set A is that cameras are not a fixed at a location.

For capturing the static state, our lidar was placed on a push cart and slowly moved through the halls. This implies that the observed scenario involves highly dynamic information, e.g., moving workers and forklifts. This data, however, can still be used as a reference for evaluating localization algorithms based on point clouds. In addition to its primary purpose, the data obtained from the highly dynamic environment serves an additional role in the testing and evaluation of future algorithms, in particular those algorithms that are designed to store and update database data of moving or temporary objects. The mapped space is much larger than in data set A and far more complex in geometry, which makes it suitable for evaluating the handling of blind spots in sensor data as coverage is far from being ubiquitous.

Camera recordings were done by mounting three stereo cameras to a forklift as shown in Figure 4. All mounts were designed to be installed next to the lamp bases of the forklift and oriented towards the direction of illumination. This offers the advantage that the observed geometry is well illuminated in darker environments, which is important for stereo vision. Two of the cameras were placed at the front side and one at the rear side of the vehicle. The cameras were oriented vertically to get a better coverage of taller objects as the camera field of view (FoV) is wide. A 3D-printed mount for horizontal orientation is in development as we expect it to be applicable in more scenarios. The recordings show the forklift driving in the fair halls starting from an entry point. In Figure 3 a moment of the video stream of one of the mounted cameras is shown together with the reconstructed point cloud and voxel representation. The collected data set covers a time span of an half hour.

Figure 4. A stereo camera mounted to a forklift.

Like for data set A, no 5G localization was available. We therefore also recorded trajectory data via the IMUs of the cameras. A major drawback of IMUs is their inherent sensor noise, which can become increasingly problematic over time. Despite this drawback, IMU data plays a role in bridging time spans between localization via 5G or based on point clouds. This is necessary because there is no guarantee of successful sensor localization in every situation. By combining IMU data with additional localization methods like 5G or SLAM, the overall system receives valuable reinforcement, leading to improved accuracy and reliability. This integration ensures continuity of data, maintaining a consistent understanding of spatial position and orientation, even when other localization methods are absent or subject to inconsistencies. In this regard, the inclusion of IMU data enables testing of point cloud localization in conjunction with IMU data, even when 5G localization is not yet available.

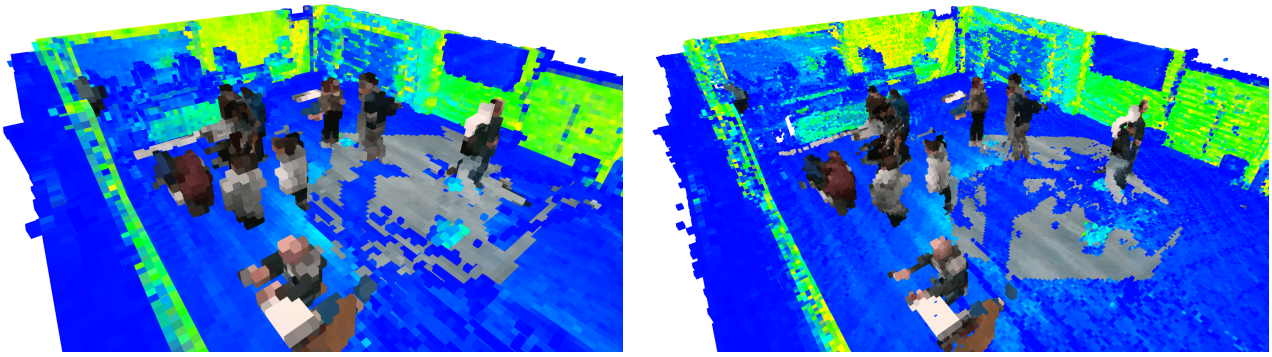

Figure 5. A voxelized situation from data set A. The images show students visiting the exhibition, represented as voxels with edge length of 10 (left) and 5 (right) centimeters.

IV. Data Processing

Up to this point, we presented the preparation of data for voxelization. The cameras producing point cloud streams in camera coordinates are virtually placed in the reference map according to the manual localization. The orientation is set manually as well by comparing the image stream with the reference point cloud. We intend to do this automatically by localizing the camera point clouds in the reference map [6]. The conversion step from point cloud to voxels is done using Open3D [14]. Figure 5 illustrates the result of combining the camera streams with the reference map as voxels.

The voxel representation shows that the overall approach yields the expected result. Space occupancy is successfully monitored from multiple directions. However, distortions still occur and suggest to increase depth resolution. Also, the result reveals how necessary an appropriate coverage is. One aspect for further development is the intelligent integration of camera streams when they compete for writing a voxel.

Doing the same with a camera frame in data set B yields a similar result, which can be seen in Figure 3. However, as no accurate localization method is available, especially as the observed scenario is highly dynamic and covers a large area with long distances, integrating the camera frames is very challenging. Without any initial localization in the large reference map to progress from using e.g. IMU data, real-time localization becomes unrealistic. Given a 5G localization at presumably centimeter accuracy [1] we expect a similar outcome as in Figure 5.

The provided data sets serve as test data for the development of the database pipeline, which is responsible for updating the voxelized point cloud database. The database stores the data in the form of voxels, which are obtained through a combination of stereo cameras and LiDAR sensors. To capture the area, we mount the cameras on common vehicles such as forklift trucks. To optimize the merging of camera streams and eliminate the need for costly computations, we calibrate the cameras with one another. This calibration establishes a local coordinate system that positions each camera in relation to the others.

Once the data is recorded, it is transmitted via 5G to either the voxel server or another intermediate computer system for further processing. The data is recorded in the form of RGB-D (RGB and depth) streams, which are then employed by the object recognition algorithms. The object recognition algorithms exclusively rely on RGB images to detect and classify various objects, such as cars, people, walls, streets, greenery, and more. The data of these objects is then linked and stored with the related generated voxels, resulting in voxels enriched by metadata such as space occupancy, color, and object type.

Additionally, we leverage the point clouds generated from the RGB-D images to localize ourselves in the world. Localization in space is crucial for the project, considering that general availability of 5G localization, which is studied by one of our project partners, is currently limited [15]. Therefore, the development of alternative localization methods becomes necessary. While 5G localization provides a broad estimation, we rely on the precise localization achieved through point cloud analysis. By determining our exact position, we can query the database for the corresponding bounding box area and map our new data into that specific area. This enables us to calculate a change-delta, which represents the differences in the voxel world. We transmit this change-delta back to the database, updating only the relevant changes in the specified voxel area. This localization process is important for both the usage of the database in AR/VR applications and for updating the data within the database.

Efficiency is a critical aspect of our pipeline, given the large volume of data involved and the need for optimal server performance with multiple participants. Therefore, the development of optimizations in the data loading and storage processes is crucial. These data sets can be used to explore improvements and streamline the pipeline within the existing framework. Figure 6 provides an overview of the entire pipeline, illustrating the flow of data and the various components involved in the process.

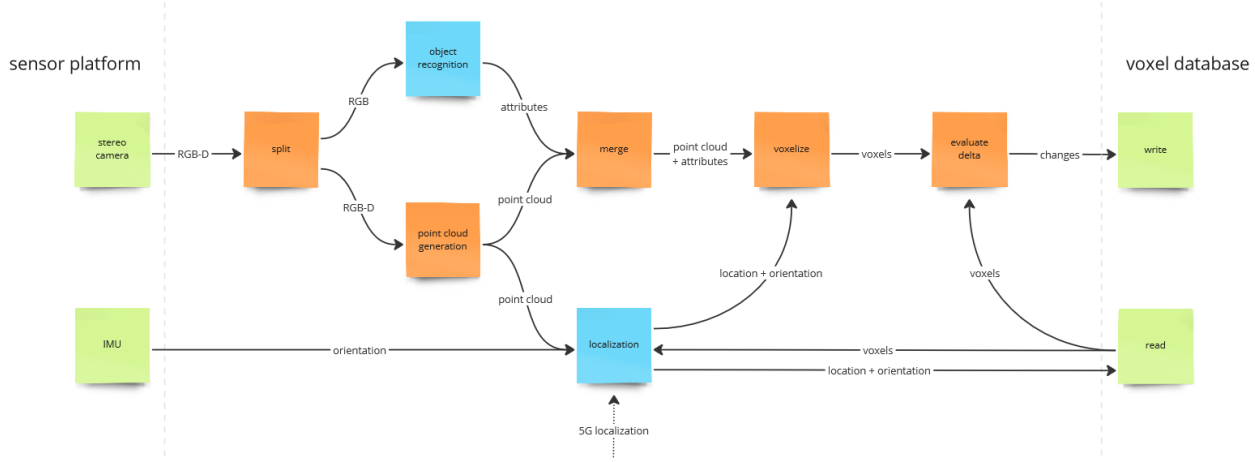

Figure 6. Simplified scheme of the database pipeline for updating data in the voxel database, which stores voxels enriched by metadata such as space occupancy, color, and object type 5G localization, object recognition as well as database design and implementation are tasks that are handled by our project partners, details of which are out of the scope of this paper.

V. Conclusion

We have presented an approach for 3D data acquisition in indoor scenarios and the captured spatio-temporal data from two specific instances. Both data sets will be used for development and testing of the data pipeline and algorithms for voxel-based digital twins. This includes the development of sensor localization and streaming over 5G, discretization of the data and its storage, object detection, and state forecasting for different application scenarios.

Our data acquisition approach employs low-cost sensors mounted on regularly moving vehicles such as forklift trucks in fairground and storage halls. Future work will focus on improved sensor localization and sensor data fusion.

Acknowledgment

We thank Arief Pratama for the sensor market survey, Jonathan Misslisch for designing and printing the sensor mounts, and Christoph von Viebahn as well as our project partners for valuable discussions and support.

References

[1] J. A. del Peral-Rosado, R. Raulefs, J. A. López-Salcedo, and G. Seco-Granados, "Survey of cellular mobile radio localization methods: From 1G to 5G," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 1124–1148, 2018.

[2] V. Stojanovic, H. Shoushtari, C. Askar, A. Scheider, C. Schuldt, N. Hellweg, and H. Sternberg, "A conceptual digital twin for 5G indoor navigation," in *MOBILITY 2021: The Eleventh International Conference on Mobile Services, Resources, and Users*, 2021, pp. 5–14.

[3] D. M. Botín-Sanabria, A.-S. Mihaita, R. E. Peimbert-García, M. A. Ramírez-Moreno, R. A. Ramírez-Mendoza, and J. d. J. Lozoya-Santos, "Digital twin technology challenges and applications: A comprehensive review," *Remote Sensing*, vol. 14, no. 6, p. 1335, 2022.

[4] N. Kshetri, "The economics of digital twins," *IEEE Computer*, vol. 54, no. 4, pp. 86–90, 2021.

[5] M. Minos-Stensrud, O. H. Haakstad, O. Sakseid, B. Westby, and A. Alcocer, "Towards automated 3D reconstruction in SME factories and digital twin model generation," in *2018 18th International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2018, pp. 1777–1781.

[6] F. S. Mortazavi, O. Shkedova, U. Feuerhake, C. Brenner, and M. Sester, "Voxel-based point cloud localization for smart spaces management," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-1/W1-2023, pp. 325–332, 2023.

[7] K. Akiyama, K. Azuma, R. Shinkuma, and J. Shiomi, "Real-time adaptive data transmission against various traffic load in multi-LIDAR sensor network for indoor monitoring," *IEEE Sensors Journal*, vol. X (early access), pp. 1–15, 2023.

[8] K. Suzuki, R. Shinkuma, N. Nakamura, and G. Trovato, "Spatial model for capturing size and shape of object from point cloud data for robot vision system with LIDAR sensors," in *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, 2023, pp. 493–496.

[9] T. Taurino and A. Villa, "A method for applying industry 4.0 in small enterprises," *IFAC-PapersOnLine (9th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2019)*, vol. 52, no. 13, pp. 439–444, 2019.

[10] K. Shi, K. Qian, and H. Yu, "Visual human localization and safety monitoring in a digital twin of workspace," in *2022 41st Chinese Control Conference (CCC)*, 2022, pp. 6117–6121.

[11] Kitware, "LidarView," https://lidarview.kitware.com/.

[12] D. Girardeau-Montaut, "CloudCompare," https://www.danielgm.net/cc/.

[13] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia, "MeshLab: An open-source mesh processing tool," in *Eurographics Italian Chapter Conference 2008*. The Eurographics Association, 2008, pp. 129–136.

[14] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," 2018, arXiv:1801.09847.

[15] K. Shamaei and Z. M. Kassas, "Receiver design and time of arrival estimation for opportunistic localization with 5G signals," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4716–4731, 2021.