Smart Automatic Mixed Reality-Based Construction Inspection Framework

Boan Tao[1], Jiajun Li[1] and Frédéric Bosché [1]

[1]School of Engineering, University of Edinburgh, Edinburgh, Scotland, UK

boan.tao@ed.ac.uk, Jiajun.Li@ed.ac.uk, f.bosche@ed.ac.uk,

Abstract -

With increasingly complex construction projects, improving inspection efficiency and accuracy is an important challenge. This paper proposes a novel MR-based construction inspection framework that integrates BIM, MR, and AI technologies to achieve automatic inspection tasks. The framework comprises object detection, 2D to 3D projection, and digital twin-based object recognition and MR-based visualisation to provide an efficient inspection process. The framework is evaluated in an indoor construction environment with common elements like electrical sockets and switches as a typical example to validate our approach in real-world applications.

Keywords -

Mixed reality; BIM; Digital Twin; Construction Inspection; Camera Projection; Object detection; Deep Learning

1 Introduction

In the construction industry, the importance of efficiency and precision in construction inspection processes cannot be overstated. Traditional inspection approaches, predominantly manual and reliant on 2D drawings and physical presence, are increasingly challenged by the complexity and scale of modern construction projects [1]. Therefore, exploration of digital technologies for enhancing efficiency in these steps is necessary.

Building Information Modeling (BIM), and now Digital Twinning (DT), have emerged as a foundational element in the evolution of construction technologies, offering detailed 3D representations and facilitating effective planning and management. Mixed Reality (MR) blends digital information with the physical environment, which offers an immersive platform that enhances the visualisation of BIM model directly on construction sites. Concurrently, computer Vision (CV) technologies leveraging Artificial Intelligence (AI) are emerging as transformative tools for automating the detection and analysis of site elements and anomalies. Thus, an integrated approach that synergises the detailed visualisation of BIM, the immersive experience of MR, and the analytical capabilities of CV could address the current limitations of traditional inspection methods, including issues with accuracy, efficiency, and safety.

This paper proposes a smart and automatic construction inspection framework that integrates the strengths of BIM, MR, and AI. In the framework, construction inspectors use MR glasses that autonomously perform inspection tasks based on the inspector's location. This system is uniquely optimised to work in a automatic way and with computational efficiency, ensuring effective performance with minimal power consumption during site inspections.

The rest of the paper is organised as follows. Section 2 reviews the relevant literature on BIM, MR and AI in the context of construction inspection applications. Section 3 introduces our automatic inspection framework, detailing its design specifically for integration with BIM, MR and AI technologies. Section 4 illustrates and evaluates the performance of this framework. Section 5 discusses performance and limitations of our method. Section 6 proposes current challenges and future developments. The paper concludes in section 7.

2 Related work

The potential of combining BIM and MR for real-time data processing in construction site inspections is exemplified in Feng and Chen [2]. They propose a system combining BIM and MR, specifically using the head-mounted MR device HoloLens. This system allows construction engineers to visualise the BIM model overlaid at the actual construction site, facilitating real-time comparison between planned and actual work, and enabling efficient inspection. Riedlinger et al. [3] demonstrate the potential benefits of the combination of BIM and MR for bridge inspection, including increased precision in locating damages and time-saving potential in damage recording. Ammari and Hammad [4] further extend this integration to multisource facilities information, BIM models, and feature-based tracking in an MR-based setting to enhance collaboration and visual communication between field workers and managers. Similarly, Nguyen et al. [5] design a MR-based system for bridge inspection and maintenance. The system is designed to overlay relevant data and information directly onto the physical bridge structure as viewed through MR devices. This feature enables inspectors to see and assess real-time

information about the bridge's condition, maintenance requirements, and other critical data in situ.

The incorporation of AI into MR marks a significant step towards automating inspection processes. Karaaslan et al. [6] and Zakaria et al. [7] discuss the integration of MR and real-time machine learning to enhance structural inspections, particularly for concrete infrastructures like bridges. They use deep learning models that can localise and quantify concrete defects in real-time using MR device. These studies underscore AI's role in analysing BIM data to detect defects and predict maintenance needs, showcasing the potential for more intelligent and proactive construction management.

The existing research primarily focuses on the pairwise combination of these technologies, such as BIM with MR or MR with AI, without fully harnessing the synergistic potential of combining all three. Moreover, current systems still largely depend on manual user input for tasks like locating specific areas or activating the system, which undermines efficiency. There is a need to develop a more autonomous MR system, empowered by BIM and AI, that can independently, automatically and passively identify and process construction site data without extensive user intervention.

3 Method

3.1 Method overview

Our proposed system architecture encompasses two primary components: MR device, specifically chosen as the Hololens2 (HL2), and a Computation Centre (CC), which can be either a local computer or a cloud-based platform. This framework is notably effective in two key use cases within a fully developed BIM context: Facilities Management (FM) inventory and construction project progress and quality monitoring. Firstly, for FM inventory, it enables dynamic interaction with the facility's digital twin, allowing managers to visualise, track, and manage assets efficiently. Secondly, for project progress and quality monitoring, it provides a real-time inspection tool for ensuring construction adheres to planned works. This aids in identifying and rectifying deviations, thus maintaining project integrity and facilitating quality control.

The comprehensive workflow of our proposed framework is depicted in fig. 1. In operation, users equipped with HL2 navigate the construction site. The HL2 (red rectangle) maintains real-time communication with the computation centre, continuously transmitting spatial data regarding the user's position and orientation. Upon receiving this spatial data, the CC (blue rectangle) initiates a series of processes, and send result back to the HL2. Key stages include:

1. Detection zone analysis. The system first evaluates whether the user is situated within a specially predefined detection zone for each element in the BIM model that needs to be controlled, thereby facilitating a focused and efficient inspection process. The design of the zone is discussed in Section 3.2. It is completed in an offline setting, with the zones stored in the database of CC.

2. Camera activation and data acquisition. If the user is within the detection zone, the computation centre sends an activation command to the HL2, which then starts capturing video frames in real-time and transmits them and the camera's intrinsic parameters back to the centre.

3. Object detection. The object detector runs in real-time on camera frames to detect target objects (e.g. building components or defects) within that detection zone.

4. Orientation validation. The system ensures that user faces the target objects and incident angles between user orientation and wall are within acceptable thresholds, to increase the accuracy of the subsequent camera projection and matching calculations (see next step).

5. 2D to 3D Projection. Utilising the 2D detection boxes coordinates, the system computes their projected coordinates in the BIM model (or Digital Twin), through 3D projection using the pinhole camera model.

6. Deviation assessment. The projected 3D coordinates are then compared against the as-planned object positions. Compliance is determined based on predefined deviation thresholds, and the results recorded and linked to the project BIM model.

7. Visualisation. The inspection results are simultaneously reported to the user visually, highlighting non-compliant from compliant objects, providing an intuitive and immediate visual cue for inspection outcomes.

The following sub-sections provide mode details about the whole process.

3.2 Detection zone analysis

The detection zones are created to focus on specific areas that need inspection or monitoring. When setting up these zones for electrical elements like switches and sockets (which are the focus of the validation presented later), walls are used as primary reference points, with the zones defined as bounding boxes extending from the walls. Parameters for each detection zone are established based on the inspection requirements. Here, the

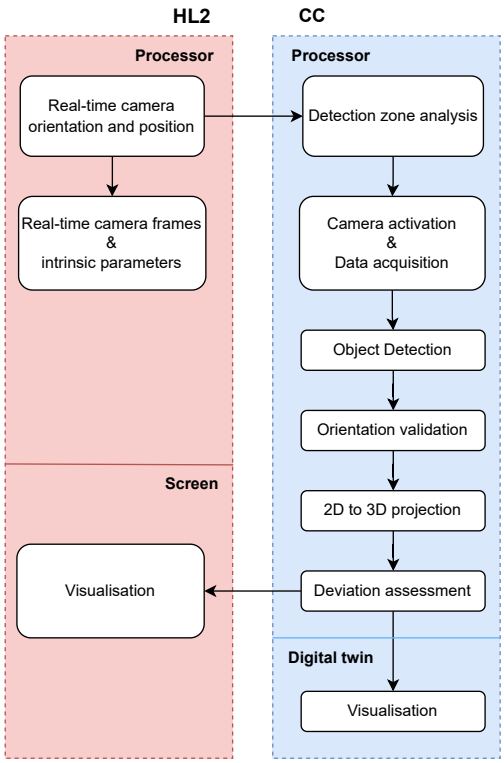

Figure 1. Real time workflow of the system

primary parameter is the distance between the inspector and the target element. We set the maximum distance at 2 meters, aligning with the optimal range of the HL2 front camera. This distance ensures that the camera captures images of sufficient quality for the computer vision algorithm to perform reliable object detection.

Each detection zone stores their essential information, including: the precise locations and categories of target elements and relevant geometrical data, such as targets' surface normal lines. By pre-storing this data, the system can rapidly process and analyse the images captured by the inspector, significantly speeding up the inspection process.

3.3 Camera activation and data acquisition

HL2 is equipped with an array of sensors that capture spatial and visual data [8]. This includes a depth sensor, an RGB camera, and sensors dedicated to tracking head, hand, and eye movements. Spatial sensors capture spatial information like position, orientation, and movement of the user's head and hands. The front-facing RGB camera captures conventional colour imagery. This can be used for applications requiring visual data from the user's perspective.

Zaccardi et al. [9] provides insights into using Unity's Barracuda on HoloLens 2 for real-time medical AR systems. They found that simpler models like Lenet5 can achieve over 30 fps. In contrast, more complex models like EfficientNetB0 result in a much lower frame rate, highlighting the balance between model complexity and performance. Therefore, in theory, the computational capabilities of current MR hardware are sufficient to support the execution of deep learning models, including the projection of 3D objects. However, for more effective communication with digital twins and to assess the framework's performance more accurately, we perform both the detection and projection processes in CC. Dibene and Dunn [10] propose a HL2 server application to facilitate the real-time streaming of sensor data over TCP (Transmission Control Protocol). This protocol ensures reliable, ordered, and error-checked delivery of a stream of bytes between applications running on hosts communicating via an IP network. In this project, we implement a multiprocessing approach to efficiently direct the streams of front camera and spatial input data towards a centralised computational hub. This approach facilitates the concurrent processing of diverse data inputs, enhancing the overall efficiency and throughput of the system.

3.4 Object detection

In this study, the overall system is illustrated using the inspection of sockets and switches as an example. But, the method is naturally adaptable to other objects (e.g. fire safety equipment [11]). To detect sockets and switches in images captured by the HL2 camera, a deep learning model is developed, based on YOLOv5m [12], noted for its rapid and precise performance. The pre-trained YOLOv5m model is then retrained (transfer learning) using a dataset comprising 2,026 indoor images featuring sockets and switches, enhanced through various augmentation techniques such as rotation, shearing, and mosaic effects to mimic lens distortion and complex indoor scenarios. The evaluation of the system involved the analysis of 73 images, incorporating 163 instances, and yielded a precision rate of 95% and a recall rate of 86.6%. The system has an inference time of 8.4 milliseconds, and a Non-Maximum Suppression (NMS) time of 2.5 milliseconds per image for an image dimension of (32, 3, 640, 640). This processing speed is particularly advantageous for real-time applications in construction inspection, highlighting the system's capability in both accuracy and efficiency in object detection tasks.

3.5 Real-time position and orientation

In the HL2, image and video streams undergo distortion correction within the image-processing framework prior to application accessibility [13]. Thus we assume that the transmitted image frames conform to a perfect pinhole camera model without distortion. It satisfies the

perspective projection equation [14]:

$$p_i = K[\mathbf{R}|\mathbf{t}]P_i \tag{1}$$

The value of camera's intrinsic matrix K, which encapsulates the camera's focal length and the principal point offset, is computed in real-time by the HL2 auto focus-length system and communicated to the computation centre. The extrinsic matrix $\mathbf{E} = [\mathbf{R}|\mathbf{t}]$, encapsulating the rotation and translation vectors of the camera, represents the camera's pose relative to the world coordinates. It undergoes real-time updates to reflect the changes in the camera's position and orientation as the user navigates through the site.

3.5.1 Initialisation

The camera's initial pose $\mathbf{E_0} = [\mathbf{R}_0|\mathbf{t}_0]$ can be measured by various methods, including QR code scanning [15], or visual analysis of recognisable structures or features [16]. In this study, the initialisation of the camera's pose is conducted through the scanning of a QR code, strategically affixed to a predetermined location (a wall in the case of the experiments reported below).

The QR code is identified, and the coordinates of its corners are extracted, denoted as q_i in the image coordinates. Their corresponding 3D coordinates in a local world coordinate system, designated as Q_i, are known from the pose of the matching twin QR code in the BIM model.

Using the 2D-3D point correspondences (q_i and Q_i), the rotation vector ($\mathbf{R}$) and translation vector ($\mathbf{t}$) of the camera coordinate relative to the world coordinate is calculated. This computation is grounded in the principles outlined in eq. (1).

3.5.2 Real-time updating

HL2 transmits real-time orientation ($\Delta\mathbf{R}$) and position ($\Delta\mathbf{T}$) changes relative to the initial pose. This data is used to update the user's pose and the camera's extrinsic matrix.

Rotation update: The new orientation matrix $\mathbf{R}_{\text{new}}$ is computed by multiplying the initial orientation $\mathbf{R}_0$ with the change in orientation $\Delta\mathbf{R}$:

$$\mathbf{R}_{\text{new}} = \mathbf{R}_0 \cdot \Delta\mathbf{R} \tag{2}$$

Position update: The new position vector $\mathbf{P}_{\text{new}}$ is updated by applying the change in position $\Delta\mathbf{T}$ relative to the initial orientation $\mathbf{R}_0$, and adding it to the initial position $\mathbf{P}_0$:

$$\mathbf{P}_{\text{new}} = \mathbf{R}_0 \cdot \Delta\mathbf{T} + \mathbf{P}_0 \tag{3}$$

Extrinsic matrix update: The extrinsic matrix $\mathbf{E}_{\text{new}}$ of the camera, which transforms points from the world coordinates to the camera coordinates, is updated using the new orientation and position:

$$\mathbf{E}_{\text{new}} = \left[\mathbf{R}_{\text{new}} \;\middle|\; -\mathbf{R}_{\text{new}} \cdot \mathbf{P}_{\text{new}}\right] \tag{4}$$

3.6 2D to 3D projection

Given the 2D image coordinates set (u, v) of the vertices of the bounding box enclosing the detected object from section 3.4, the first step is to normalise these coordinates to the camera's coordinate system. The normalised camera coordinates (x, y) are obtained by:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = K^{-1} \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}, \tag{7}$$

In each frame, the detection target is identified based on the camera-object angle, defined as the angle formed between the camera's line of sight and the normal to the object's surface. This process involves measuring the camera-object angle for every object within the designated detection zone. The object exhibiting the smallest such angle is then selected as the primary detection target for that specific frame. The orthogonal distance, represented as d, between this selected object and the camera, is effectively the z-coordinate value of the object within the camera's coordinate system.

Subsequently, the camera coordinates are transformed by applying a scaling factor equal to d. This step translates the 2D coordinates into 3D camera coordinates (X_c, Y_c, Z_c):

$$X_c = x \cdot d, \quad Y_c = y \cdot d, \quad Z_c = d. \tag{8}$$

The final step involves transforming these camera coordinates into 3D world coordinates. This transformation is accomplished using the camera's extrinsic matrix $\mathbf{E}_{\text{new}}$ obtained in section 3.5.2:

$$\begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = \mathbf{E}_{\text{new}} \cdot \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}. \tag{9}$$

3.7 Deviation assessment and Visualisation

Section 3.6 calculates in real-time the projection of 3D bounding boxes that captures the 'as-is' location of elements within the detection zone. For each detected 'as-is' element, we compute the centroid of its 3D bounding box. This centroid serves as a representative point for comparing the 'as-is' element with corresponding 'as-designed' elements of the same category within the detection zone. The comparison process involves identifying the 'as-designed' element whose centroid is closest to that of the 'as-is' element. This proximity-based selection aims to match each 'as-is' element to the most relevant 'as-designed' counterpart.

Given the dynamic and continuous operation of the camera, multiple 3D bounding boxes are projected for the same target over time. These projections may exhibit variations due to factors such as noise, distortion,

and limitations inherent to the sensing equipment. To account for these variations, we compute an average centroid for the 'as-is' element across all captured frames. This averaged centroid is then compared to the centroid of the closest 'as-designed' element.

The spatial deviation between the averaged 'as-is' centroid and the 'as-designed' centroid is quantitatively assessed against a predefined threshold. This assessment determines whether the 'as-is' element conforms to the planned design specifications.

The detection and conformance checking results are recorded in the Digital Twin as the average of the projected bounding boxes.

Finally, the result is sent back to the HL2 where the detected bounding box are shown coloured in:

- *green*, if the element is matched and found conforming;
- *red*, if the element is matched and found non-conforming;
- *grey*, if the element is not matched.

4 Experimental result

4.1 Result visualisation

Figure 2 and fig. 3 show the digital twin as updated in real time in the CC. The grey mesh is the BIM of room. Within this virtual representation of the room, four different coloured squares are observable; these are designated as detection zones. The HL2 in its current pose (updated in real time) is shown in black. As introduced in section 3.1, the front camera on the HL2 is only activated when the HL2 is situated within these coloured detection zones. If the target object is detected and checked as conforming, the target object is shown with a small green sphere, representing the 'as-is' position. In the digital twin screenshot in fig. 2, three green spheres can be seen on the wall next to the blue detection zone, representing three detected and conforming objects.

The HL2 screen interface, shown in fig. 4 and fig. 5, reports essential information to the user during the inspection process. It reports when the user enters a detection zone and the designated targets for inspection. Objects that align with the as-planned design are explicitly listed on the screen, and for enhanced visual clarity, these compliant objects are highlighted within green bounding boxes. Conversely, objects detected but found to deviate from the as-planned design are enclosed within grey boxes, indicating that their projected 3D positions do not match any element's as-planned position.

4.2 Performance analysis

4.2.1 Initialisation

Using scanning QR codes for determining camera position and orientation is a cost-effective and accessible method. However, this approach has its limitations. The accuracy can be significantly affected by factors such as poor lighting, low camera resolution, and environmental interference. To enhance the accuracy of the initialisation of the camera's position and orientation, we continuous scan the QR code for a duration of 5 seconds while remaining stationary. Then we calculate the mean value of the position and orientation collected during this period. Therefore, transient errors caused by sudden changes in the environment or by the initial positioning of the camera can be averaged out.

In our experiment, a comparative evaluation is conducted between the computed camera position derived from the pin hole model and the position obtained through manual measurements. This comparison revealed that the average position deviation in this initialisation step is approximately 3.49 cm.

This discrepancy can be attributed to two significant factors. Firstly, lens distortion, particularly in the form of radial and tangential distortions, can alter the perceived geometry of the scanned QR code, leading to inaccuracies in the calculation of the camera's position and orientation. Secondly, during the process of breathing, subtle but impactful body movements occur, which can inadvertently shift the camera's position, albeit slightly.

4.3 Real-time projection

During 2D to 3D projection, the method casts rays from the camera's origin through the image plane and into the 3D world. The precision of the projection process is subject to variation due to several factors, including the camera-object angle, the distance between the camera and the object, and the camera's incidence angle, which is defined as the angle between the camera's optical axis and the normal of the surface. To elucidate the correlation between these factors and projection errors, we conducted an experimental study using a single socket target. The experiment is initialised by scanning QR code and then detection and projection are performed at varying angles and distances

We define deviation as the spatial distance calculated from the centre point of the 'as-designed' socket to the centroid of the 3D projected bounding box. In total 12,507 data points are acquired for analysis. In the analysis, the controlled variable method is utilised to ensure rigour and accuracy in the interpretation of the data.

Initially, we fix the camera incidence angles at 0° or 5°, given that the majority of the data fall within this range. Additionally, these angles are chosen due to their

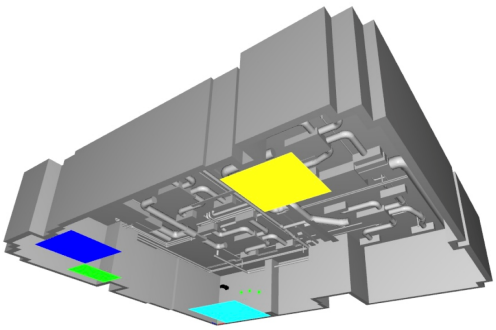

Figure 2. Screenshot of the Digital Twin (switch#1, socket#1, socket#2)

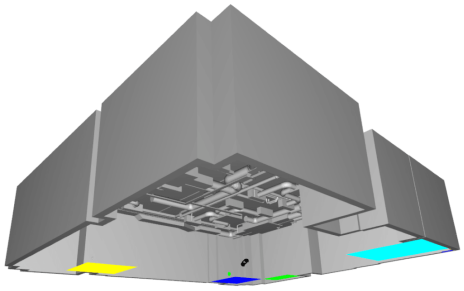

Figure 3. Screenshot of the Digital Twin 2 (socket#4)

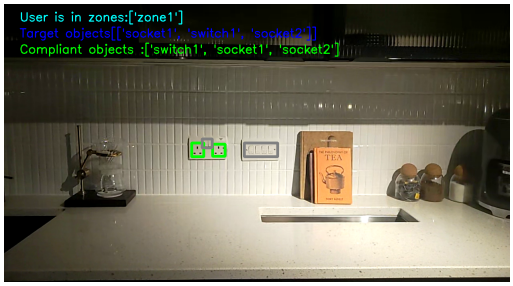

Figure 4. HL2 Screen Interface 1 (switch#1, socket#1)

Figure 5. HL2 Screen Interface 2 (socket#4)

minimal distortion impact on the projection, ensuring they did not significantly affect the analysis of other parameters. Employing the set parameter of camera incidence angle to select the test subdataset(4,586 data), we analyse the relationship between the camera-object distance and the observed deviations. The results are summarised in the 2D scatter plot shown in fig. 6. Our findings indicate that the deviation maintains a consistent level of stability, remaining below 0.25 m, up to a camera-object distance of 1.1 m. Beyond this threshold, the deviation increases significantly and becomes more erratic. This phenomenon can be attributed primarily to two factors: (1) the amplification of errors in preceding stages, such as sensor measurement or QR code initialisation, due to longer distances; and (2) the inherent limitations of the camera's capabilities adversely affecting detection at extended ranges.

Setting the specified range, where the camera-object distance is less than 1.1 m, result in minimal deviation, as evidenced by prior findings. Then, we investigate the relationship between camera-object incidence angle and deviation, as illustrated in Figure 7. The analysis demonstrates that there is a direct correlation between the deviation and camera-object incidence angle within a range of less than 10° . As this angle surpasses 10°, we observe that the deviation becomes both unstable and significantly higher. Within the angle range of 0° to 5°, deviation remains below 0.22 m, with an average

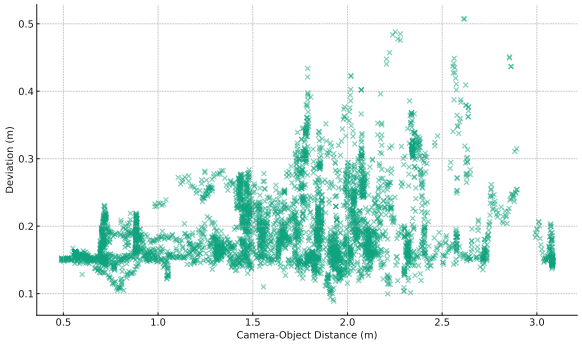

Figure 6. Relationship between the camera-object distance and 3D projection deviation.

deviation of 0.16 m. This can be attributed to two possible reasons: (1) camera-object incidence angle affects image distortion and perspective projection, leading to greater deviations at wider angles; and (2) the Inertial Measurement Unit (IMU) sensor measurement inside HL2 is not accurate and stable and thus accumulates errors during calculations.

5 Discussion

In light of the aforementioned findings, it can be deduced that optimal system performance is attained when the camera-object incidence angle is less than 5° and camera-object distance is under 1.1 m. Under these spe-

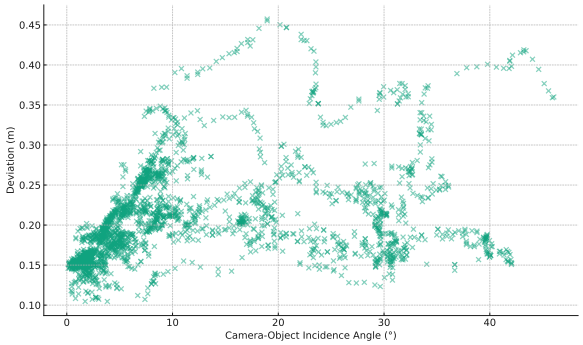

Figure 7. Relationship between the camera-object incidence angle and 3D projection deviation.

cific conditions, the system demonstrates enhanced efficacy, as evidenced by a mean deviation of approximately 16 cm. That deviation can be ascribed to the following factors.

First, there is an inherent error in the process of initialising the camera's location and orientation using QR code scanning. As discussed in section 4.2.1, this error results in a positional deviation of approximately 3.5 cm. Additionally, a deviation in orientation has been identified, further investigation into which is considered for future research endeavours.

Second, several types of distortions can affect the outcome. Firstly, perspective and lens distortions impact how the sizes and shapes in an image are seen, which can lead to errors in the final 3D model. Then, the way lighting and shadows appear in the image can also change how accurately objects are detected and represented. Additionally, sensor errors, particularly from devices like inertial measurement units (IMUs), introduce further errors. These sensors sometimes struggle to track the exact position and movement of the camera, especially during quick motions.

Considering the various challenges inherent in the process of 2D to 3D projection, and the technological capability of HL2, it appears that using that system, construction positioning conformance can only be confirmed with a threshold of 16 cm. To improve the accuracy of our object detection and projection, two main strategies can be employed. First, we can train our object detection model with images taken in different lighting conditions. This approach would make the model more versatile and accurate in varying lighting environments. Second, we can use additional tools like external sensors to support and enhance the initialisation of camera position and orientation.

6 Future development

It is important to note that our methodology currently assumes a singular detection object per frame. In scenarios involving multiple objects, the projection outcomes for objects other than the primary target are prone to deviations. To address this issue, our future research will develop and integrate an algorithm capable of filtering outliers and averaging projection results.

In construction site management, accurately identifying complex elements like multifunctional media sockets is challenging due to their diverse designs and the need to distinguish their specific types and orientations. A strategy to address this would be to utilise sophisticated object detection technologies, trained on an extensive array of socket designs and configurations.

Besides, construction sites often involve situations where materials and equipment that partially occlude crucial elements. The compact placement of items on sites complicates the identification process. To overcome these obstacles, applying data augmentation methods such as cutout and mosaic in the training phase can enhance the model's ability to handle occlusions. Additionally, enhancing the network design with attention mechanisms enables the model to pinpoint more nuanced features, boosting its detection performance.

The proposed system is designed to automate the process of (progress and) quality control in construction projects, ensuring that all installed components, such as sockets, switches, and structural elements, adhere to the project's specifications and quality standards. This application can significantly reduce human error and increase the efficiency of the inspection process. The system holds potential for other applications, such as monitoring and ensuring compliance with safety regulations on construction sites. By detecting potential hazards or non-compliance with safety standards (e.g., improper installation of safety equipment, obstruction of emergency exits), the system can contribute to a safer work environment.

7 Conclusion

This paper presents a novel MR-based construction inspection framework. The framework integrates AI-based object detection with 2D to 3D projection techniques and matching against the facility's DT to achieve automatic and passive inspection work, facilitated by the communication system between the MR device and computation centre. The results are stored in the DT and can be reviewed in an interactive, and user-friendly way by the MR user on site. The framework's practicality and effectiveness were evaluated in an indoor construction environment. The results from these tests demonstrate the system's feasibility in real-world inspection processes, albeit with limitations on the quality of the results that can reasonably be achieved.

References

[1] Y. Li and P. Gu. Free-form surface inspection techniques state of the art review. *Computer-Aided Design*, 36(13):1395–1417, 2004. doi:10.1016/j.cad.2004.02.009.

[2] C.-W. Feng and C.-W. Chen. Using bim and mr to improve the process of job site construction and inspection. *WIT Transactions on the Built Environment*, 192:21–32, 2019. doi:10.2495/BIM190031.

[3] U. Riedlinger, F. Klein, M. Hill, C. Lambracht, S. Nieborowski, R. Holst, S. Bahlau, and L. Oppermann. Evaluation of mixed reality support for bridge inspectors using bim data: Digital prototype for a manual task with a long-lasting tradition. *i-com*, 21(2):253–267, 2022. doi:10.1515/icom-2022-0019.

[4] K. El Ammari and A. Hammad. Remote interactive collaboration in facilities management using bim-based mixed reality. *Automation in Construction*, 107:102940, 2019. doi:10.1016/j.autcon.2019.102940.

[5] D. C. Nguyen, R. Jin, and C. H. Jeon. Developing a mixed-reality based application for bridge inspection and maintenance. In *The 20th International Conference on Construction Applications of Virtual Reality (CONVR 2020)*, 2020. doi:10.1108/CI-04-2021-0069.

[6] E. Karaaslan, M. Zakaria, and F. N. Catbas. Mixed reality-assisted smart bridge inspection for future smart cities. In *The Rise of Smart Cities*, pages 261–280. 2022. doi:10.1016/B978-0-12-817784-6.00002-3.

[7] M. Zakaria, E. Karaaslan, and F. N. Catbas. Real-time ai-based bridge inspection using mixed reality platform. In *Structures Congress 2023*, pages 120–131, 2023. doi:10.1061/9780784484777.012.

[8] D. Ungureanu, F. Bogo, S. Galliani, P. Sama, X. Duan, C. Meekhof, J. Stühmer, T. J. Cashman, B. Tekin, J. L. Schönberger, et al. Hololens 2 research mode as a tool for computer vision research. *arXiv preprint arXiv:2008.11239*, 2020. doi:10.48550/arXiv.2008.11239.

[9] S. Zaccardi, T. Frantz, D. Beckwée, E. Swinnen, and B. Jansen. On-device execution of deep learning models on hololens2 for real-time augmented reality medical applications. *Sensors*, 23 (21):8698, 2023. doi:10.3390/s23218698.

[10] J. C. Dibene and E. Dunn. Hololens 2 sensor streaming. *arXiv preprint arXiv:2211.02648*, 2022. doi:10.48550/arXiv.2211.02648.

[11] A. Corneli, B. Naticchia, M. Vaccarini, F. Bosché, and A. Carbonari. Training of yolo neural network for the detection of fire emergency assets. In *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, volume 37, pages 836–843, 2020. doi:10.22260/ISARC2020/0115.

[12] G. Jocher, A. Stoken, J. Borovec, L. Changyu, A. Hogan, L. Diaconu, J. Poznanski, L. Yu, P. Rai, R. Ferriday, et al. ultralytics/yolov5: v3. 0. *Zenodo*, 2020. doi:10.5281/zenodo.3983579.

[13] Microsoft. Locatable camera in mixed reality. `https://learn.microsoft.com/en-us/windows/mixed-reality/develop/advanced-concepts/locatable-camera-overview`, 2023. Accessed: 2023-12-01.

[14] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 666–673 vol.1, 1999. doi:10.1109/ICCV.1999.791289.

[15] J.-I. Kim, H.-S. Gang, J.-Y. Pyun, and G.-R. Kwon. Implementation of qr code recognition technology using smartphone camera for indoor positioning. *Energies*, 14(10):2759, 2021. doi:10.3390/en14102759.

[16] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl, and T. Sattler. Back to the feature: Learning robust camera localization from pixels to pose. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3246–3256, 2021. doi:10.1109/CVPR46437.2021.00326.