UNVEILING LARGE-SCALE HISTORICAL CONTENTS WITH V-SLAM AND MARKERLESS MOBILE AR SOLUTIONS

A. Torresani[1,2], S. Rigon[1], E. M. Farella[1], F. Menna[1], F. Remondino[1]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
E-mail: <atorresani><srigon><elifarella><fmenna>< remondino>@fbk.eu
Web: http://3dom.fbk.eu

[2] Department of Information Engineering and Computer Science (DISI), University of Trento, Trento, Italy

KEY WORDS: Markerless AR, Mobile AR, visual SLAM, photogrammetry, 3D modelling, historical photos, heritage.

ABSTRACT:
Augmented Reality (AR) is already transforming many fields, from medical applications to industry, entertainment and heritage. In its most common form, AR expands reality with virtual 3D elements, providing users with an enhanced and enriched experience of the surroundings. Until now, most of the research work focused on techniques based on markers or on GNSS/INS positioning. These approaches require either the preparation of the scene or a strong satellite signal to work properly. In this paper, we investigate the use of visual-based methods, i.e., methods that exploit distinctive features of the scene estimated with Visual Simultaneous Localization and Mapping (V-SLAM) algorithms, to determine and track the user position and attitude. The detected features, which encode the visual appearance of the scene, can be saved and later used to track the user in successive AR sessions. Existing AR frameworks like Google ARCore, Apple ARKit and Unity AR Foundation recently introduced visual-based localization in their frameworks, but they target mainly small scenarios. We propose a new Mobile Augmented Reality (MAR) methodology that exploits OPEN-V-SLAM to extend the application range of Unity AR Foundation and better handle large-scale environments. The proposed methodology is successfully tested in both controlled and real-case large heritage scenarios. Results are available also in this video: https://youtu.be/Q7VybmiWIuI.

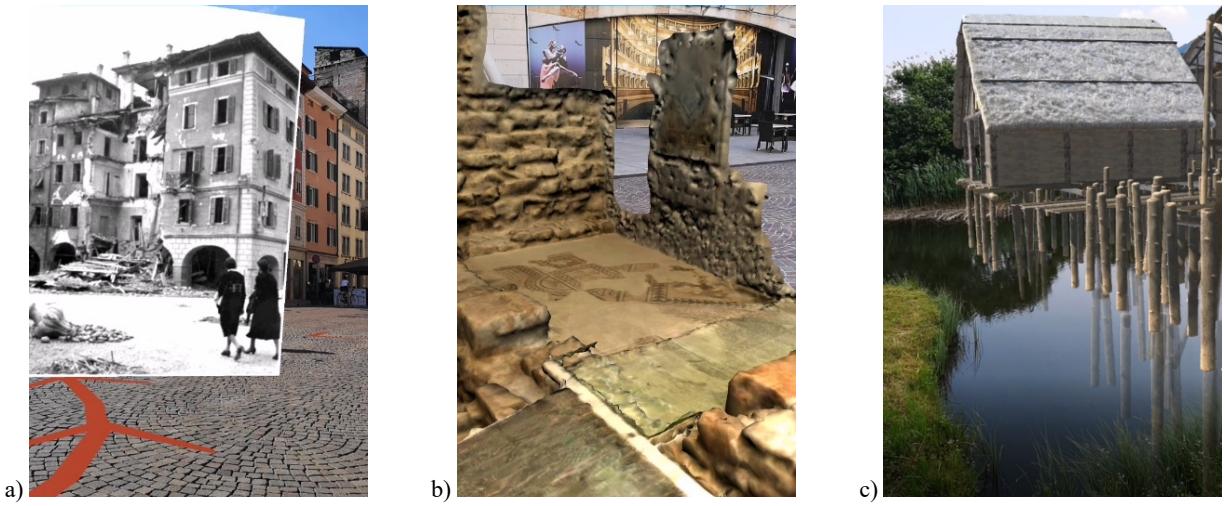

Figure 1. The three large-scale scenarios used in the paper and the AR results based on markerless smartphone solution: (a) historical photographs of the city of Trento, (b) the remains of the underground roman city in Trento and (c) the pile dwelling site of Fiavè.

1. INTRODUCTION

Augmented Reality (AR) aims to blend the real and the digital worlds, enabling immersive and interactive experiences where the reality is fused with virtual content perceivable by humans as visual, auditory or haptic stimuli (Bekele and Champion, 2019). Visual augmentation is probably the most popular and allows users to see the real world populated with virtual 3D content with mobile devices (Mobile Augmented Reality - MAR), either through smart glasses or Holographic Head Mounted Display - HHMD (like the Microsoft HoloLens, Google Glasses, etc.) or smartphone / tablet screens (Chatzopoulos et al., 2017). AR has enormous potential and is already influencing many fields ranging from industry and aerospace to medical and heritage, also pushed by well documented, free and relatively easy-to-use AR frameworks (Nowacki and Woda, 2019). In the cultural heritage context, AR has been broadly exploited in museums or heritage sites to enhance the visitor experience with extended and interactive information, both in indoor and outdoor.

To properly overlay digital 3D elements on the real scene, the relative user viewpoint and the viewing direction must be known (six degrees of freedom of the camera – 6DoF). This process is called tracking or localization and can be carried out in different ways (Table 1). Marker-based tracking works by identifying, in the acquired images, known physical targets or objects previously placed in the scene. This approach has been widely applied, thanks to its robustness and low computational requirements. Nevertheless, it is not always practical or even possible to place markers in the environment. On the other hand, markerless or location-based tracking solutions localize the user with respect to known global (for example, from a Global Navigation Satellite System - GNSS) (Pagani et al., 2016;

Nakamura et al., 2018) or local coordinate systems. In the latter case, the tracking can exploit feature maps of the environment, also known as sparse point clouds, typically estimated with Visual Simultaneous Localization and Mapping (V-SLAM) algorithms (Cadena et al., 2016; Sualeh and Kim, 2019). In these approaches, the world itself is converted into a big set of distinctive visual features, against which the camera position and attitude is tracked. The nice property of these algorithms is that they can theoretically work in both outdoor and indoor environments without the need of marker installation in the scene or having GNSS coverage. The main drawback is that, by relying on scene contents, they are sensitive to scene changes. Real environments constantly change (daily and seasonal illumination, weather, moving objects, vegetation, structural changes) and how to enable, in these cases, long-term and robust tracking is still an ongoing research topic (Saputra et al., 2018; Wald et al., 2020).
Google ARCore[1], Apple ARKit[2] and Unity AR Foundation[3] are unquestionably among the most used MAR frameworks nowadays. Recently, these libraries introduced the support to persistent and multiuser marker-less AR experiences. AR sessions can now persist after their termination, allowing different users to resume them successively, and visualize the original augmented content. Without requiring markers, these solutions take advantage of the 3D visual features of the scene estimated with the build-in V-SLAM algorithms during the creation of the AR session. Unfortunately, these functionalities are not yet designed to work in environments bigger than a common room (Feigl et al., 2020). When the size of the scene increases, their use becomes impractical, and the tracking accuracy rapidly degenerates over time. One of the most relevant shortcomings is the limited size of the feature maps, which significantly reduces the area where the user can be accurately localized and tracked.
In this paper, by combining OPEN-V-SLAM (Sumikura et al., 2019) with Unity AR Foundation, we present how to enable accurate and markerless MAR experiences in large-scale outdoor scenarios. More specifically, the proposed solution aims to overcome two of the biggest limitations of current MAR frameworks, i.e. (i) the limited size of the localization areas and (ii) significant tracking drift in large environments. The proposed method was tested and demonstrated in three different large-scale heritage case studies:

- Historical images in the city centre of Trento, Italy (Figure 1a). Users are guided to discover, with their smartphone, archival photos acquired in the city in the 20th century overlaying them to the actual city. This application is part of the TOTEM[4] project and helps to valorize historical archives showing urban changes.
- Archaeological remains of the underground roman city in Trento, Italy (Figure 1b). The MAR application allows users to discover this hidden treasure by virtually visiting the underlying structures while walking in the above square. This solution gives the possibility to overcome some access limitations, especially in the actual pandemic period.
- The UNESCO pile-dwelling site in Fiavé, Trento, Italy (Figure 1c). As part of the JUDIT[5] project, some attractive ICT solutions have been designed to better valorize and communicate history to the young generations. Using a smartphone, 3D reconstructions of the ancient pile-dwelling structures are superimposed to the wooden remains, partially emerging from the lake water surface, to show how such houses were formed and located on the lake.

[1] https://developers.google.com/ar
[2] https://developer.apple.com/augmented-reality/arkit/
[3] https://unity.com/unity/features/arfoundation

1.1 Main contributions

The main contributions of the paper are:

- A new methodology, built upon OPEN-V-SLAM and Unity AR Foundation (Section 3);
- Empowering of MAR applications in dynamic scenarios without using markers or GNSS/INS data;
- Enabling large feature maps to work in large spaces;
- Reduced tracking drift while localizing the device in large environments (Section 4.1);
- Application and evaluation in three challenging and large-scale heritage scenarios (Section 4.2).

MARKER-BASED	
Pros	Low computational cost, robust, fast, indoor and outdoor, not affected by changes in the scenes
Cons	Invasive, small localization area, target could not be placeable
GNSS/INS-BASED	
Pros	No target, low computational costs, not affected by changes in the scenes
Cons	Only outdoor, problems in narrow streets or forested areas, accuracy related to satellite coverage
MARKERLESS (visual-based)	
Pros	No target, indoor and outdoor
Cons	Affected by changes in the scenes, computationally intensive

Table 1. Summary of tracking & localization methods for MAR.

2. RELATED WORKS

2.1 Visual SLAM

The goal of a SLAM algorithm is twofold: *localization* – the estimation of the ego-motion of an agent in an initially unknown environment, and *mapping* – the estimation of the representation of the environment where the agent is moving. The fundamental property of SLAM is that it achieves its goals solely using sensor data captured by the agent. Cameras have received enormous interest for their weight, cost, power consumption and ubiquity, and they have been extensively used as the main sensor to perceive the reality. In these cases, SLAM is technically called Visual SLAM and it is probably the most common branch of SLAM today, finding application in robotics, Augmented Reality (AR), 3D mapping and autonomous navigation. The first V-SLAM approaches were based on filtering techniques and MONO-SLAM (Davison et al., 2007) is probably the most known algorithm of this category. PTAM (Klein and Murray, 2007) represented an important paradigm change, proposing to split localization and mapping in two different concurrent operations, and replacing probabilistic state estimation with a least square optimization (bundle adjustment or pose graph optimization). They proposed to update the state of the map only on selected frames, called keyframes, reducing the overall system weight and allowing the V-SLAM algorithm to keep a real-time behaviour in bigger scenarios. While direct methods can exploit more information of the image, produce dense or semi-dense point clouds, and work more easily in non-collaborative scenarios, undirect methods achieve in general higher accuracies in both localization and mapping and are more robust to illumination changes. DTAM (Newcombe et al., 2011) and LSD-SLAM (Engel et al., 2014) are well known direct algorithms,

[4] https://totem.fbk.eu/
[5] https://judit.fbk.eu/

with the latter capable of achieving real-time performances without using GPU acceleration. Among the most representative undirect algorithms, there are, in addition to PTAM: ORB-SLAM (Mur-Artal et al., 2015), which extends the concepts of PTAM with place recognition and the ability to work in larger scenarios; ORB-SLAM2 (Mur-Artal et al., 2017), which supports stereo and RGB-D cameras; ORB-SLAM3 (Campos et al., 2021) which supports IMU and multiple maps; OPEN-V-SLAM which is based on ORB-SLAM2 and allows an easy map reuse; KIMERA (Rosinol et al., 2021), notable for producing semantically annotated meshes in real-time.

2.2 Mobile Augmented Reality in Cultural Heritage

Mobile Augmented Reality (MAR) has also been exploited to valorize cultural heritage assets (De Carolis et al., 2018; Boboc et al., 2019; Carrozzino et al., 2019). Yin et al. (2021) presented heritage tourists' needs and involvement in mobile AR solutions, providing a theoretical framework for designing mobile AR heritage applications. Marto and Gonçalves (2019) presented an evaluation of a MAR application in the Conimbriga archaeological site. In Puyuelo (2013), marker-based solutions were explored to display 3D models of the UNESCO heritage site of "La Lonja" in Valencia. In Duguleana et al. (2016), the authors used a GNSS-based mobile AR application to visualize the appearance of the leaning tower in Pisa in different time ages. Buana and Meily (2021) used AR to support the visit of the Taman Ayun Temple. An AR heritage guide of Brno (Czech Republic) is presented in Střelák et al. (2019), where the authors also compared GNSS and visual localization. Kolivand et al. (2018) reviewed markerless AR approaches to support tourist experiences in cultural heritage locations. In Palma et al. (2019), authors presented an ARKit-based MAR solution to create augmented views of baroque atria.
Head Mounted Holographic Displays (HMHD), such as Hololens, Google Glasses, etc. start to be used as mobile devices for AR experiences in heritage scenarios (Teruggi et al., 2021), but their applicability is still limited with respect to marker-, GNSS-based and markerless smartphone/tablet solutions.

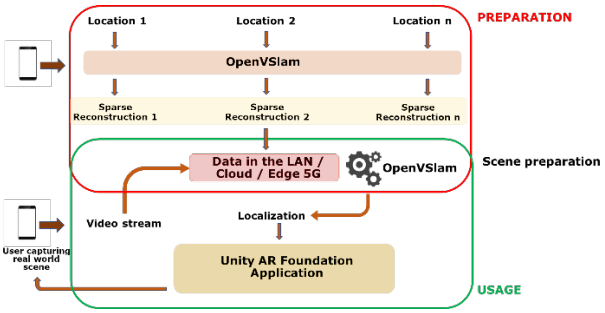

Figure 2. The proposed methodology. The area is first surveyed, creating a sparse reconstruction with OPEN-V-SLAM. The sparse reconstruction is then used to prepare the AR scene, to localize the user and correct the camera poses of the Unity AR Foundation session.

3. PROPOSED SOLUTION

The proposed methodology, based on OPEN-V-SLAM and Unity AR Foundation, is summarized in Figure 2. Compared with other algorithms, OPEN-V-SLAM enables an easy management of the estimated sparse reconstructions, which can be re-imported and used for localizing the camera in successive executions. Unity AR Foundation contains various powerful features for the AR development and supports multiple mobile and wearable devices. The implemented solution is logically divided in two macro phases: *preparation* – i.e. the steps required to prepare the AR session and *usage* - where the user localization and AR experience take place.

3.1 Preparation

3.1.1 Sparse reconstruction

The area of interest is firstly surveyed, acquiring different videos of the scene with a smartphone. Videos are then processed offline by OPEN-V-SLAM to extract keyframes and estimate a 3D sparse reconstruction of the area. The sparse reconstruction contains both the visual features of the scene and a bag-of-words (Gálvez-López and Tardos, 2012) representation of the selected keyframes. Their combination allows a fast and precise localization of a device/user in successive runs (Mur-Artal and Tardós, 2014).

3.1.2 AR scene preparation

The obtained 3D sparse reconstructions are then exploited to place the virtual objects in the desired positions. However, some preliminary steps are initially required. First, the sparse reconstructions need to be scaled, since the scale is arbitrary (monocular estimation). To this aim, a dense 3D reconstruction is computed using a Multi-View-Stereo algorithm and then the scale is computed exploiting ground truth data (e.g. using an ICP algorithm with scale factor). With the scale correctly computed, the virtual objects can be placed in the desired positions, also taking advantage of the dense point clouds to perform an accurate positioning of the virtual objects. The final rigid body transformation T, encoding the scale and eventual rotation changes, is saved for the usage phase (Section 3.2) to handle the change of coordinate systems between OPEN-V-SLAM and Unity AR Foundation.

3.2 Usage

3.2.1 Localization

To start the AR session, the smartphone position and orientation must be firstly determined. The mobile application needs to send the live stream of the smartphone camera to the OPEN-V-SLAM algorithm, which is running either in a node of the local area network or remotely. The bi-directional communications with web sockets are managed using Socket.IO. Exploiting the previously computed sparse reconstruction of the scene (Section 3.1.1), the V-SLAM algorithm then estimates the camera pose of each received image I with the EPnP algorithm, using RANSAC to discard outliers and restricting the set of considered 3D points only to those being observed by the most similar keyframes of I (according to the bag-of-words representation). If a pose P with sufficient inliers is found, then P is further refined with pose graph optimization and returned, with Socket.IO, to the mobile application.

3.2.2 Unity AR Foundation session and correction

The mobile application waits for the first camera pose to initialize the AR session and displays the augmented content. When the AR application receives a valid pose P, the origin of the AR scene moves to a new position Φ, obtained multiplying T (section 3.1.2) with P. The AR session is then initialized and the tracking is performed internally by the AR Foundation framework. During the session, the mobile application sends, at regular time intervals, new images to OPEN-V-SLAM, in order to get corrected camera poses and, eventually, reduce the tracking drift accumulated by the device. At time t, when a new image is sent to OPEN-V-SLAM, the application stores the current pose of the

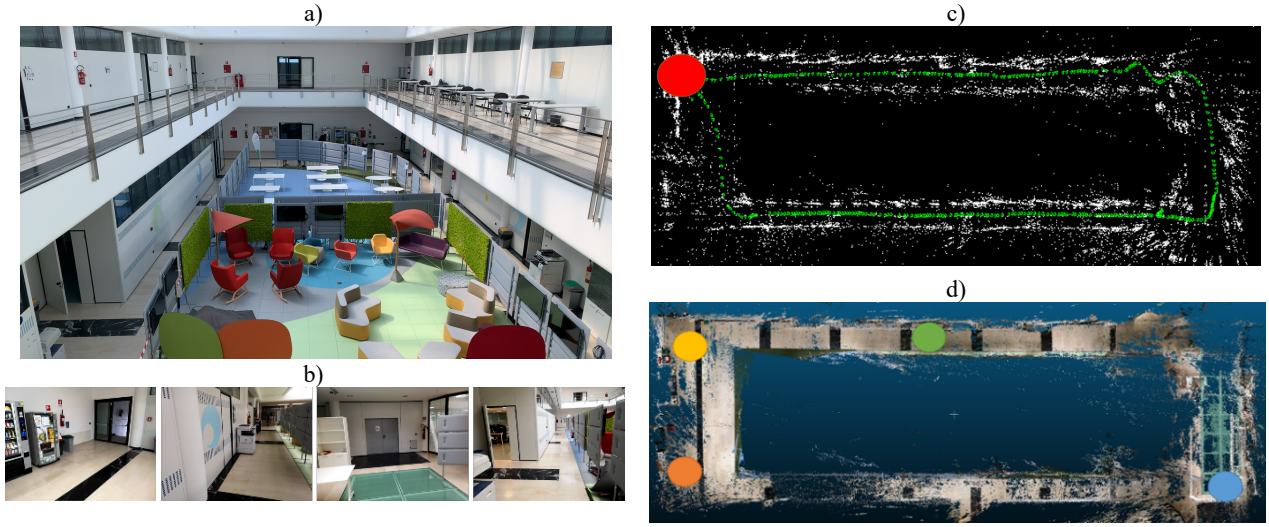

Figure 3. The area for the accuracy test (a) and some images of the scene (b). The obtained camera poses and sparse reconstruction of the area (c). The generated dense point cloud used to position the virtual elements (c) with colored dots showing locations of virtual objects (Fig. 4a, c).

camera P_t. At time $t+n$, when the corrected pose P'_t is received, the current pose P_{t+n} is updated as follows:

$$P_{t+n} = \Pi\, P_{t+n} \quad (1)$$

where Π is the relative transformation between the original - not corrected pose P_t and the corrected pose P'_t.
As shown in Section 4.1, the sparse reconstruction of OPEN-V-SLAM is in general more accurate, mainly thanks to the loop closure correction, than that estimated internally by AR Foundation, and therefore it can be used as a correction tool.

4. EXPERIMENTS

4.1 Accuracy test

The tracking accuracy of Unity AR Foundation in a large indoor scenario (around 100 meters trajectory) is firstly evaluated by comparing the achieved performances when the correction step (Section 3.2.2) is or is not applied. The considered area is the open space of the FBK North building (Figure 3a). The place was first surveyed with a Huawei P20Pro following a rectangular trajectory along the perimeter of the space. An in-house application was used to acquire images at 5Hz and 640x480 px resolution, obtaining some 1283 images. Before processing the images within OPEN-V-SLAM, the P20Pro device was geometrically calibrated, as V-SLAM approaches generally do not estimate, for real-time constraints, the intrinsic camera parameters during the bundle adjustment optimizations. The Zhang method (Zhang, 2000) implemented in OpenCV was employed, using a 22x15 chessboard calibration pattern printed on a rigid plate. Figure 3b shows the sparse reconstruction of the area obtained with OPEN-V-SLAM. Acquisitions started and ended from the same position (red dot in Figure 3c) and the V-SLAM algorithm correctly detected the loop. This loop enforced a global bundle adjustment optimization that significantly corrected the accumulated drift and produced a globally consistent map of the area. The selected keyframes and their corresponding poses were then imported in Agisoft Metashape to generate a dense point cloud of the scene (Figure 3d). The scale was applied by measuring some control points in the area. Finally, close to the starting point, four virtual arrows were placed over the corresponding real objects, namely an extinguisher (yellow), a printer (green), a door (blue) and a notice board (orange).
To qualitatively check the correctness of the OPEN-V-SLAM poses and 3D reconstruction, four different localizations nearby the selected positions were tested. Figure 4a shows the screen capture of the mobile application immediately after the OPEN-V-SLAM localization. The discrepancy between the real objects and the pointing arrows are very small, validating the the correctness of both OPEN-V-SLAM sparse reconstruction and proposed methodology.
In addition to this visual evaluation, the accuracy of the camera poses estimated internally by AR Foundation was also assessed, taking the poses of OPEN-V-SLAM as ground truth. We executed two consecutive AR sessions, few minutes apart from each other and following the same rectangular trajectory along the open space perimeter. In the first session, no correction step (Section 3.2.2) was applied, with OPEN-V-SLAM being used only to estimate the first pose of the camera. In the second session, we enabled the correction step at intervals of 10 seconds. Both sessions, started and ended from the printer (green circle), following a clockwise direction. During the sessions, we logged the camera pose estimated by AR Foundation and saved the corresponding images. The images were then oriented in OPEN-V-SLAM with respect to the same reference sparse reconstruction. Finally, a L2 norm of the difference of the camera centers of OPEN-V-SLAM (after applying T - section 3.1.2) and AR Foundation was computed. Figure 4b plots the obtained results. As shown, the correction step significantly decreased the tracking drift of the application, reducing the average L2 norm by around 51%. In particular, the first session accumulated a significant drift after the 400-th image, producing a very visible mismatch (Figure 4c – left) between the real and virtual objects. On the other hand, when the correction is applied, the drift was bounded in a more acceptable range and, despite some occasional spikes in the plot, the AR experience and results were much more accurate (Figure 4c – right).

4.2 Case studies in heritage scenarios

The proposed methodology was tested in three challenging and real-world cases, characterized by significant changes of the scene appearance (illumination variations, vegetation growth, moving/new/missing objects, etc.).

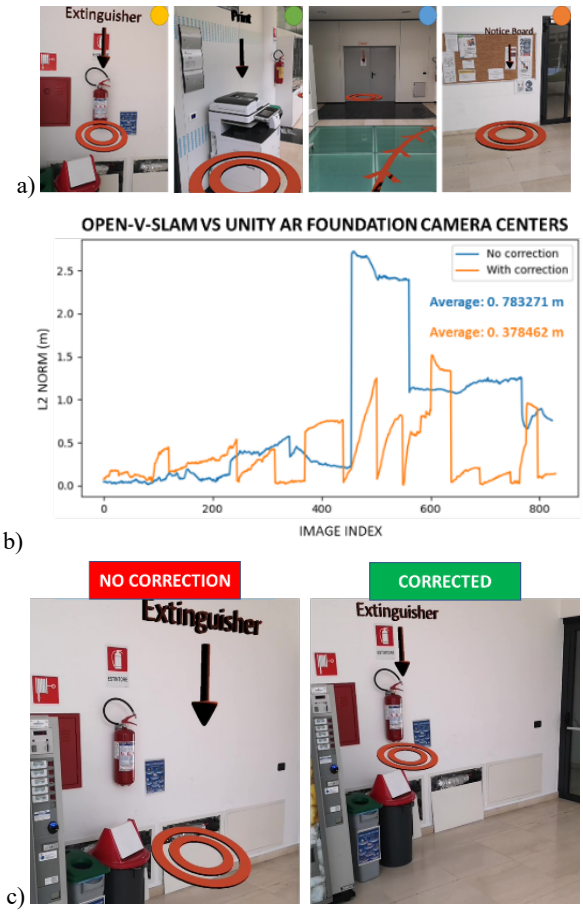

Figure 4. Visual impressions of the OPEN-V-SLAM accuracy localization (a). Plot of the L2 NORM of the difference of the camera centers with/without correction (b). Effects of the proposed correction step (c).

4.2.1 Historical images of Trento, Italy

The AR application allows users to visualize some historical archival images depicting the "Duomo" square (Figure 5a) in the 20th century and overlaid them to the actual situation. An initial sparse reconstruction of the area (Figure 5b) was used to (i) localize the user and (ii) orient the historical images in the AR scene, using some manually picked points and the DLT algorithm. The performances of the localization were verified from several positions L_i (Figure 5c). All the localizations were performed some days after the creation of the reference sparse point cloud, which was obtained from images recorded in the early afternoon. L1, L2, L3 happened approximately at the same afternoon hour, while L4 was carried out in the morning.

4.2.2 Underground roman city of Trento, Italy

The MAR experience allows users to explore the remains of the hidden Roman city, preserved below the modern square (Figure 6a). The ancient site was digitally reconstructed in 3D with photogrammetric techniques. The reference sparse reconstruction (Figure 6b) was achieved using early afternoon video acquisition, while the localizations (Figure 6c) were performed in the mornings of the forthcoming days.

4.2.1 UNESCO pile dwelling site in Fiavé, Trento, Italy

The UNESCO pile dwelling site of Fiavè (Figure 7a) features a MAR application which allows users to display the 3D reconstructive models of the pile dwelling buildings over the original remains (Figure 7d,e). The area is very challenging for vision algorithms because the prehistoric wooden poles are immersed in a broad natural scenario, surrounded by water and continuously changing vegetation. Luckily, the wooden walkway close to the poles could, despite its repetitiveness, provide quite stable feature points for the user localization. In this case, the sparse reconstruction (Figure 7b) and the localizations happened the same day, during the morning. Three localizations (Figure 7c) were tried along the planned visitor path.

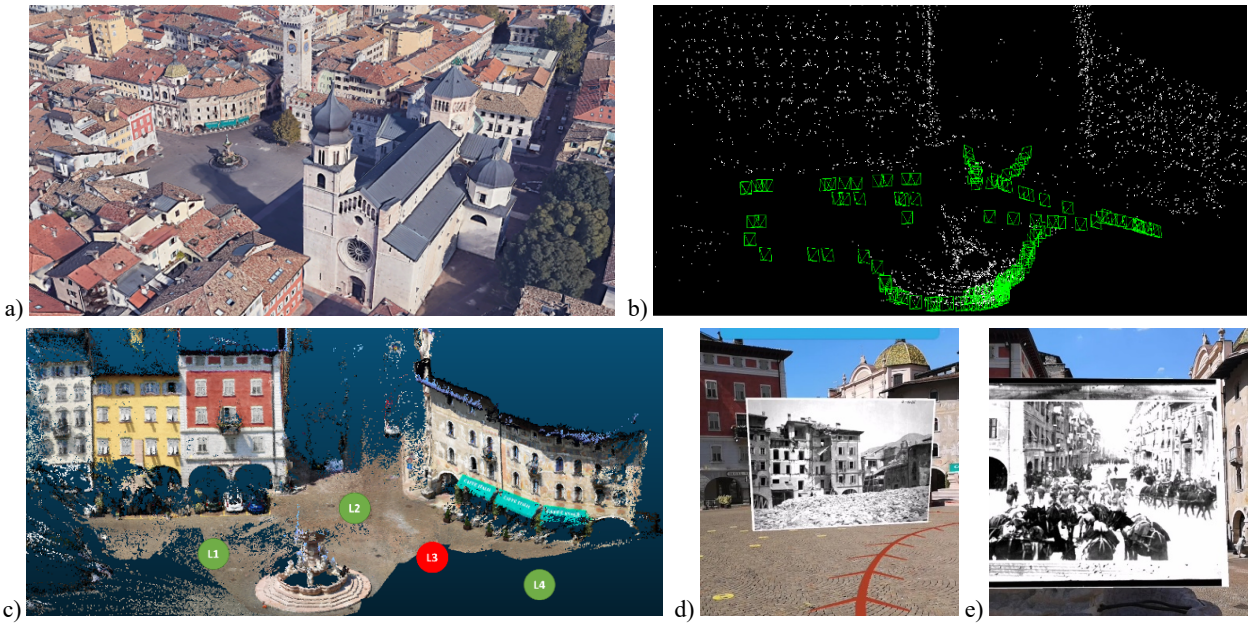

Figure 5. The “Duomo” square area as seen in Google Earth (a). Camera poses and reference sparse reconstruction generated in the preparation phase (b). The dense point cloud of the area with the positions Li for the localization tests reported in Section 4.2.2 (c). Examples of the MAR application showing the historical images blended with the reality while a mobile device is moved in the square (d, e).

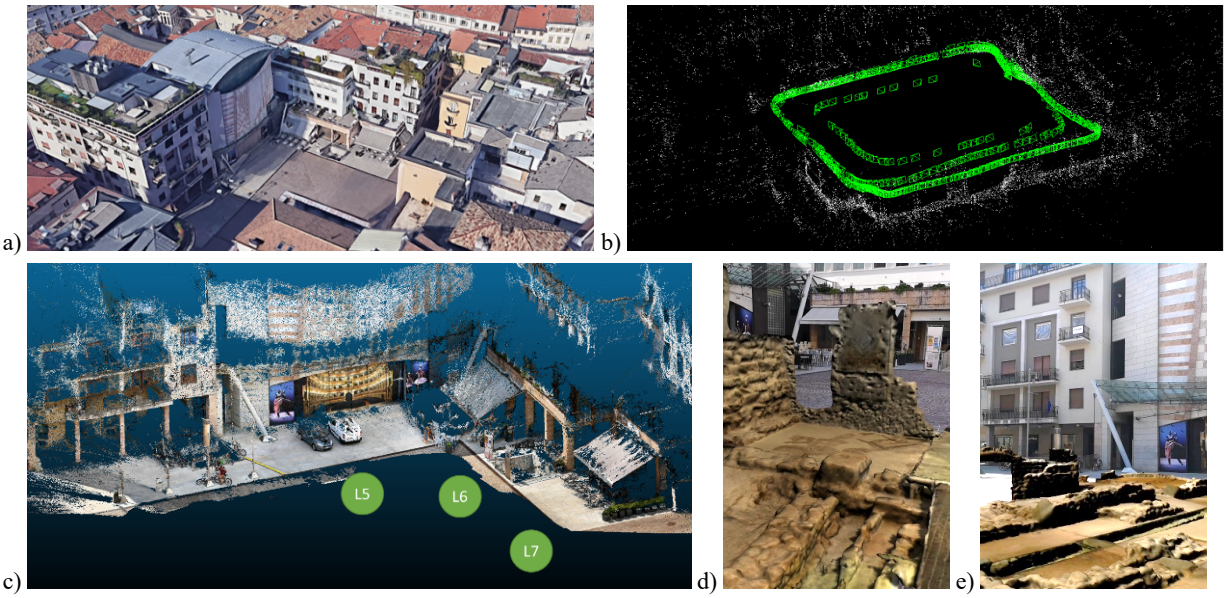

Figure 6. The square above the Roman city in Trento as seen in Google Earth (a) and the sparse reconstruction from smartphone video sequences (b). Part of the dense point cloud of the area and the positions L_i for the localization tests (c). Some views of the MAR application overlaying the 3D digital model of the roman remains to the modern square (d, e).

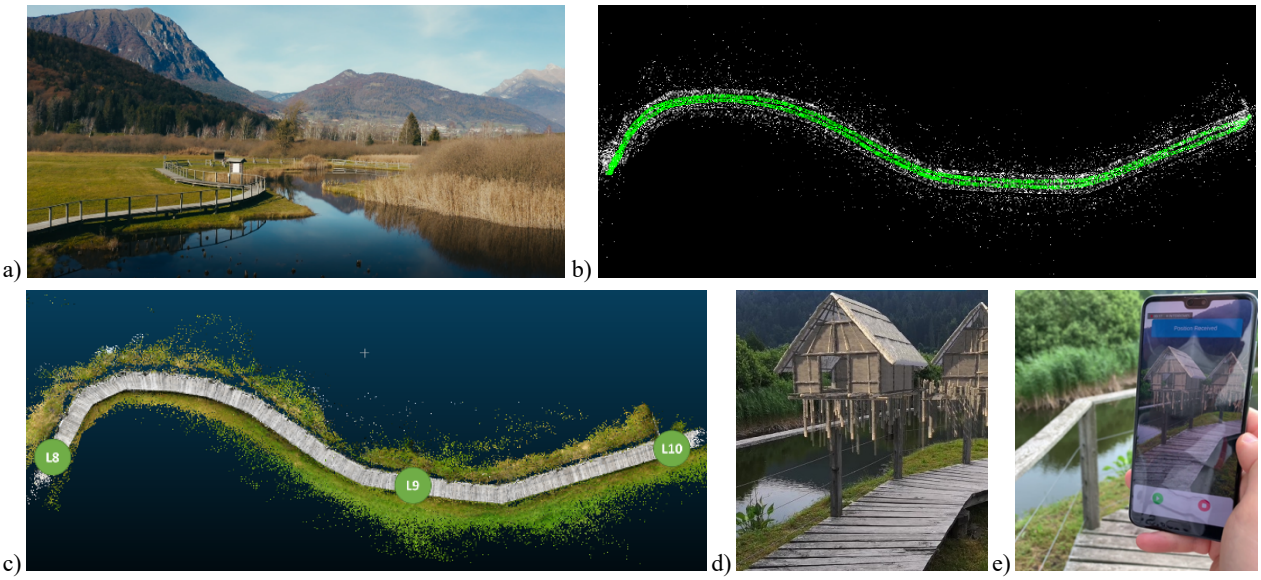

Figure 7. The UNESCO pile dwelling of Fiavé (a). The reference sparse reconstruction of the area (b). The dense point cloud of the area and the L_i positions tested for the localization (c). The reconstructive models of the pile-dwellings structures overlapped to the remaining wooden poles, visible through the AR application (d, e).

4.2.2 Localization results

The presented localization tests are evaluated in terms of completion time and, in the more problematic cases, by analyzing some inner metrics of the OPEN-V-SLAM localization procedure. Most of the localization tests were completed in less than 1.5 seconds from the start of the AR application (Table 2). The most problematic tests are L4, L6 (long completion time) and L9 (failed). In these cases, the localization was tempered by both wrong localization candidates and/or an insufficient number of feature matches (Figure 8). The fact that these problems occurred in areas well covered by the sparse reconstructions suggests that both the ORB matching and the bag-of-word candidates suffer when the scene changes significantly (illumination and scene objects) and/or it presents many repetitive patterns.

Dataset	Localization id	Time (s)
Historical images	L1	1.30
Historical images	L2	0.793
Historical images	L3	FAIL
Historical images	L4	1.329
Roman city	L5	1.094
Roman city	L6	7.330
Roman city	L7	0.774
Pile dwelling	L8	0.696
Pile dwelling	L9	8.287
Pile dwelling	L10	0.518

Table 2. The different case studies and the time (sec) required to localize the user from the start of the AR application.

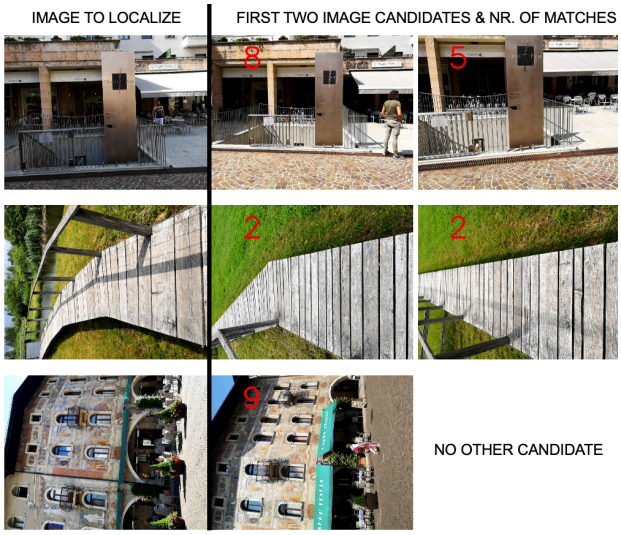

Figure 8. Examples of localization problems due to scene changes. The algorithm can retrieve wrong localization candidates (middle) and/or candidates with an insufficient number of matches (top, bottom).

5. DISCUSSION AND CONCLUSIONS

This work presented a new methodology for enabling MAR experiences in large-scale scenarios. The proposed solution exploits OPEN-V-SLAM to extend the capabilities of Unity AR Foundation, allowing it to be more effectively used in large environments. In particular, the sparse reconstruction of OPEN-V-SLAM is used both to broaden the area where the users can be localized and to correct the tracking drift of the AR Foundation framework. A controlled indoor test was used to demonstrate the improvements brought by our solution (51% tracking drift reduction). The presented methodology was also used to enable MAR experiences in three real case scenarios for the valorisation of heritage contents. Exploiting the more static elements of the scene, the device/user was successfully localized in most of the experiments.

The presented experiments highlighted both potentialities and limitations of the proposed V-SLAM markerless MAR solution. Compared to marker-based approaches, the presented methodology does not require an invasive scene preparation. Compared to GNSS/INS-based approaches, our method does not require a strong satellite signal to accurately localize the user, allowing the application to work also in urban canyons, mountainous / forested areas, underground or indoor. Compared to existing MAR libraries (such as ARCore, ARKit and AR Foundation), the presented procedure improves the AR experience in large-scale scenarios, with larger localization areas and reduced tracking drift.

The weakness of the method is mainly related to the localization procedure as its outcome depends on multiple factors:

- The initial sparse reconstruction of the area must have abundant feature points spread across the scene. Real-time image detector and descriptors, such as the ORB (Rublee et al., 2011) employed in OPEN-V-SLAM, are generally not as robust as other "offline" methods (e.g. SIFT) to viewpoint variations.
- Relevant scene variations between the pre-recorded step and the device/user localization (illumination variations, object changes, moving people, etc.) can invalidate many of the estimated features and affect the validity of the sparse reconstruction.
- Bag-of-word approaches can inherit the weakness of the associated image descriptor, and they might be unable to discriminate well when the scene changes.
- Non-collaborative surfaces (e.g. white walls) or repetitive texture patterns could limit the number of possible features or make difficult and error-prone their matching.

The key step is thus to improve the scene understanding and the feature matching techniques, so that the V-SLAM algorithm can successfully recognize the same place even in presence of important changes. Another possible solution is to estimate different sparse reconstructions of the area covering various illumination and scene appearance conditions, but this approach is hardly maintainable and very inefficient.

Beside this, future works will investigate the possibility to update the surveyed scene using the images acquired by the mobile platform, also using collaborative approaches (Nocerino et al, 2017). Moreover, we will explore the use of different feature detectors/descriptors, with particular attention to learned approaches (Remondino et al., 2021). Recently they showed promising improvements over hand-crafted features, especially regarding the re-localization performance in challenging environments. We would also like to test the proposed methodology in other scenarios, with particular interest in industry, where AR can be a valuable tool to support the chain work. Finally, we will replace Socket.IO with native WebSocket API, in order to improve the real-time communication between Unity AR Foundation and OPEN-V-SLAM. We believe that with future improvements of photogrammetric and computer vision algorithms, vision-based localization techniques could improve their performance in difficult dynamic environments and become an even more valuable tool in the upcoming MAR applications.

ACKNOWLEDGEMENTS

Part of the presented work was realized within the research activities of the JUDIT (https://judit.fbk.eu/) and TOTEM (https://totem.fbk.eu/) projects funded by Fondazione CARITRO.

REFERENCES

Bekele, M.K. and Champion, E., 2019. A comparison of immersive realities and interaction methods: cultural learning in virtual heritage. *Frontiers in Robotics and AI*, 6, p.91.

Buana, W. and Meily, S.O., 2021. Augmented Reality Application using Dynamic Location-Based Tracking of Taman Ayun Temple. Lontar Komputer: *Jurnal Ilmiah Teknologi Informasi*, 12(1), pp.24-32.

Boboc, R.G., Duguleană, M., Voinea, G.D., Postelnicu, C.C., Popovici, D.M. and Carrozzino, M., 2019. Mobile augmented reality for cultural heritage: Following the footsteps of Ovid among different locations in Europe. *Sustainability*, 11(4), p.1167.

Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I. and Leonard, J.J., 2016. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE *Transactions on Robotics*, 32(6), pp.1309-1332.

Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M. and Tardós, J.D., 2021. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. IEEE *Transactions on Robotics*.

Carrozzino, M., Voinea, G.-D., Duguleană, M., Boboc, R. G., and Bergamasco, M., 2019. Comparing innovative XR systems in cultural heritage. A case study. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W11, 373-378.

Chatzopoulos, D., Bermejo, C., Huang, Z. and Hui, P., 2017. Mobile augmented reality survey: From where we are to where we go. IEEE *Access*, 5, pp.6917-6950.

Davison, A.J., Reid, I.D., Molton, N.D. and Stasse, O., 2007. MonoSLAM: Real-time single camera SLAM. IEEE *Trans. PAMI*, 29(6), pp.1052-1067.

De Carolis, B. N., Gena, C., Kuflik, T., Origlia, A., & Raptis, G.E., 2018. AVI-CH 2018. Advanced Visual Interfaces for Cultural Heritage. In Proc. *ACM Int. Conf. on Advanced Visual Interfaces*.

Duguleana, M., Brodi, R., Girbacia, F., Postelnicu, C., Machidon, O. and Carrozzino, M., 2016. Time-travelling with mobile augmented reality: A case study on the piazza dei miracoli. *EuroMed Conference*, pp. 902-912.

Engel, J., Schöps, T. and Cremers, D., 2014. LSD-SLAM: Large-scale direct monocular SLAM. Proc. *ECCV*, pp. 834-849.

Feigl, T., Porada, A., Steiner, S., Löffler, C., Mutschler, C. and Philippsen, M., 2020. Localization Limitations of ARCore, ARKit, and Hololens in Dynamic Large-scale Industry Environments. In *VISIGRAPP* (1: GRAPP), pp. 307-318.

Gálvez-López, D. and Tardos, J.D., 2012. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5), pp.1188-1197.

Klein, G. and Murray, D., 2007. Parallel tracking and mapping for small AR workspaces. Proc. 6th IEEE and ACM International *Symposium on Mixed and Augmented Reality*, pp. 225-234.

Kolivand, H., El Rhalibi A., Abdulazeez S. and Praiwattana P., 2018. Cultural Heritage in Marker-Less Augmented Reality: A Survey. In: Turcanu-Carutiu, D. (ed.), *Advanced Methods and New Materials for Cultural Heritage Preservation*. Intechopen.

Marto, A., Gonçalves, A., 2019. Mobile AR: User Evaluation in a Cultural Heritage Context. *Appl. Sci*., 9, 5454.

Mur-Artal, R. and Tardós, J.D., 2014, May. Fast relocalisation and loop closing in keyframe-based SLAM. In 2014 IEEE International Conference on Robotics and Automation (ICRA) (pp. 846-853). IEEE.

Mur-Artal, R., Montiel, J.M.M. and Tardos, J.D., 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE *Transactions on Robotics*, 31(5), pp.1147-1163.

Mur-Artal, R. and Tardós, J.D., 2017. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE *Transactions on Robotics*, 33(5), pp.1255-1262.

Nakamura, H., Sonobe, T., Toda, H., Fujisawa, N., Taketomi, T., Plopski, A., Sandor, C., Kato, H., 2018. Fusion of VSLAM/GNSS/INS for Augmented Reality Navigation in Ports. Proc. 31st *Int. ION GNSS*+, pp. 1765-1775.

Newcombe, R.A., Lovegrove, S.J. and Davison, A.J., 2011. DTAM: Dense tracking and mapping in real-time. In IEEE *ICCV*, pp. 2320-2327.

Nocerino, E., Poiesi, F., Locher, A., Tefera, Y.T.; Remondino, F.. Chippendale, P.; Van Gool, L., 2017. 3D Reconstruction with a Collaborative Approach Based on Smartphones and a Cloud-Based Server. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*, XLII-2/W8, pp. 187–194.

Nowacki, P. and Woda, M., 2019. Capabilities of arcore and arkit platforms for ar/vr applications. In *Int. Conference on Dependability and Complex Systems,* pp. 358-370.

Palma, V., Spallone, R., and Vitali, M., 2019: Augmented Turin baroque atria: AR experiences for enhancing cultural heritage. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLII-2/W9, pp. 557-564.

Pagani, A., Henriques, J., and Stricker, D., 2016. Sensors for location-based augmented reality the example of Galileo and EGNOS. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLI-B1, 1173-1177.

Puyuelo, M., Higón, J.L., Merino, L. and Contero, M., 2013. Experiencing augmented reality as an accessibility resource in the UNESCO heritage site called "la lonja", Valencia. *Procedia Computer Science*, 25, pp.171-178.

Remondino, F., Menna, F., Morelli, L., 2021. Evaluating hand-crafted and learning-based features for photogrammetric applications. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLIII-B2-2021, 549–556.

Rosinol, A., Abate, M., Chang, Y. and Carlone, L., 2020. Kimera: an open-source library for real-time metric-semantic localization and mapping. Proc. *IEEE ICRA*, pp. 1689-1696.

Rublee, E., Rabaud, V., Konolige, K. and Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. Proc IEEE *ICCV*, pp. 2564-2571.

Saputra, M.R.U., Markham, A. and Trigoni, N., 2018. Visual SLAM and structure from motion in dynamic environments: A survey. ACM *Computing Surveys* (CSUR), 51(2), pp.1-36.

Strasdat, H., Montiel, J.M. and Davison, A.J., 2012. Visual SLAM: why filter? *Image and Vision Computing*, 30(2), pp.65-77.

Střelák, D., Škola, F. and Liarokapis, F., 2016. Examining user experiences in a mobile augmented reality tourist guide. Proc. 9th *ACM PETRA Conference*, pp. 1-8.

Sualeh, M. and Kim, G.W., 2019. Simultaneous localization and mapping in the epoch of semantics: a survey. *International Journal of Control, Automation and Systems*, 17(3), pp.729-742.

Sumikura, S., Shibuya, M. and Sakurada, K., 2019. OpenVSLAM: A versatile visual SLAM framework. Proc. 27th *ACM International Conference on Multimedia*, pp. 2292-2295.

Teruggi, S., Grilli, E., Fassi, F., Remondino, F., 2021. 3D surveying, semantic enrichment and virtual access of large cultural heritage. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,*in press.

Yin, C.Z.Y., Jung, T., tom Dieck, M.C., Lee, M.Y., 2021. Mobile Augmented Reality Heritage Applications: Meeting the Needs of Heritage Tourists. *Sustainability*, 13, 2523.

Wald, J., Sattler, T., Golodetz, S., Cavallari, T. and Tombari, F., 2020. Beyond controlled environments: 3D camera re-localization in changing indoor scenes. Proc. *ECCV*, pp. 467-487.

Zhang, Z., 2000. A flexible new technique for camera calibration. IEEE *Trans. PAMI*, 22(11), pp.1330-1334.