DIGITAL HERITAGE (2025)
S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, and F. Remondino (Editors)

Spatial random access to explore heritage site using spherical video

El Mustapha Mouaddib[1], Guillaume Caron[1,2] and Arsalane Zarghili[3]

[1]University of Picardie Jules Verne, MIS lab, Amiens, France
[2]CNRS-AIST JRL (Joint Robotics Laboratory), IRL, Tsukuba, Japan
[3]Université Sidi Mohamed Ben Abdellah, Laboratoire Systèmes Intelligents et Application, FST Fez, Morocco

Abstract

Spherical images are particularly adapted to develop virtual tours of heritage monuments. While Lidar scanning, photogrammetry and rotating camera systems can lead to produce spherical images, compact dual-fisheye cameras are accurate enough for many uses, such as virtual tour, and easy to use by a non-expert. Classical methods require the images to be placed manually on a blueprint of the environment. Then, the user navigation in the virtual monument is limited to generated transitions between a few locations.

Instead of a few pictures, this paper considers a spherical video recorded while walking within a monument and illustrates how to explore spatial dimensions of the environment where the video was taken, beyond the time scroll-bar. To this end, spherical visual Simultaneous Localization And Mapping (SLAM) and the alignment of the resulting map to an architectural blueprint are combined to create a spatio-temporal virtual tour from a video.

The concept is demonstrated on the al-Qarawiyyin Mosque in Fez, Morocco, with access to the virtual tour interface at this [link].

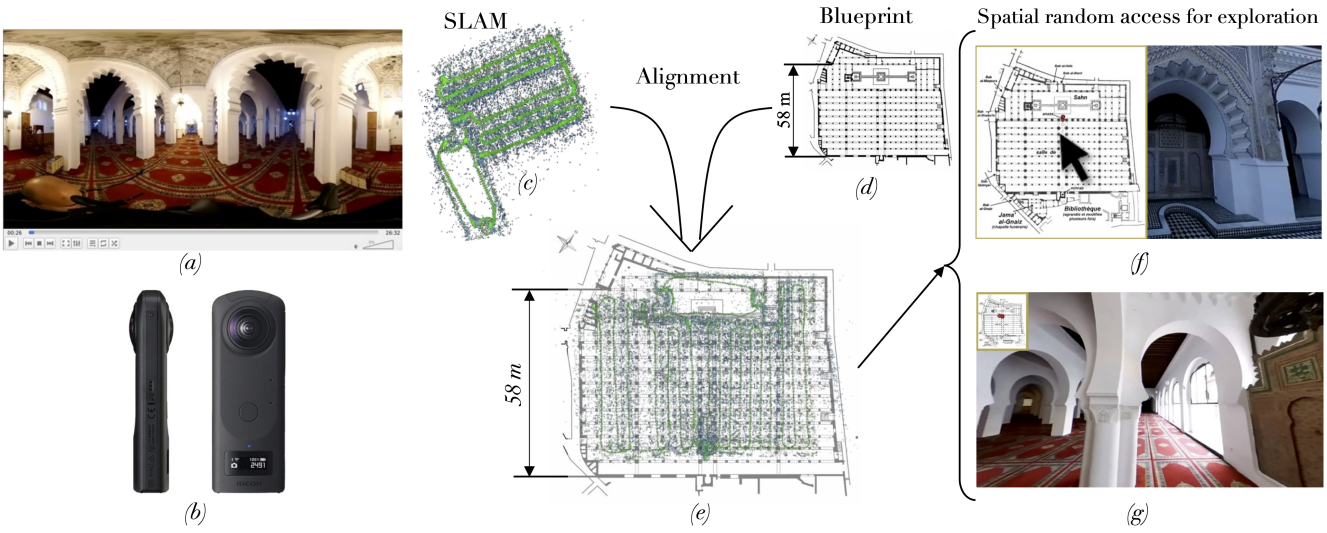

Figure 1: *Pipeline overview from spherical video (left) to spatial navigation within the video (right): (a) spherical video as equirectangular format; (b) the dual-fisheye Ricoh Theta Z1 used in our experiments; (c) excerpt of SLAM results (green: camera locations; dark blue: natural landmarks); (d) blueprint of the al-Qarawiyyin Mosque. Aligning (c) to (d) allows to get the camera location for the whole video in the blueprint (e). From outdoor (f), one reaches the time of the video at the other side of the door with just a click on the blueprint (black mouse icon), hence displaying the video at the desired spatial position (g) whereas the camera trajectory was very different.*

DOI: 10.2312/dh.20253164

1. Introduction

Terrestrial Lidar Scanning and Photogrammetry are the gold standards for capturing the reality of heritage buildings in geometry and appearance. The digital twins thus produced, have shown their interest for archiving, study, restoration, the enhancement and interpretation of architectural heritage. However, despite their relative democratization, making digital twins still requires dedicated knowledge hindering their use by non-experts. Furthermore, their cost for making a virtual tour matching the geometry of the heritage building oftentimes exceeds available resources. In parallel, it has been a decade that compact cameras capable of a 360-degree field-of-view are on the general consumer market and their video quality today is remarkable, easily reaching an 8K resolution at 30 high quality images per second even indoor, for less than 1% the cost of a Lidar scanner.

360-degree panoramas (360-pano) tour or 360-Virtual Reality showed interest as a travel substitution tool during the COVID-19 pandemics. [SNE21] investigates the ability of virtual reality to replace physical travel and discuss the acceptance and perception of this alternative. In [RCM23], the authors propose a complete review and evaluation of the tools and services to produce virtual tour websites. They also propose a framework for creating 360-pano tours and apply it to explore a Historic Hotel (Subiaco Hotel). However, to our knowledge, there is no framework for exploiting 360-degree videos.

In this article, we propose an original solution for virtual tours based on the use of videos made by a person who systematically walks through the environment and films with a spherical camera. This systematic exploration, combined with spherical vision, allows to approach the complete coverage of the environment and can be used by non-experts.

The exploitation, a posteriori, of these videos can be done directly but does not allow to understand the global space because it is sequential. In addition, the video sequence follows the path chosen by the maker of the video. Thus, in repetitive environments subject to perceptual aliasing, users who do not know the filmed place encounter difficulties in spatial representation of the filmed place. Our idea is to improve this spatial perception by linking the images of the video to a global blueprint and to make accessible the panoramic image of the video corresponding to the chosen position in real time. The link between images and the environment map is done by Simultaneous Localization And Mapping (SLAM) software that computes, only from the (images of) the video, the trajectory achieved by the camera and a sparse tridimensional model of the surroundings (not useful in this work). We applied this approach in the highly challenging environment that is the al-Qarawiyyin Mosque in Fez, Morocco. Indeed, the highly repetitive visual appearance among the approximately 200 spans and the strong difference of luminosity between indoor and outdoor parts of the 3600 square meters site gathers the hardest cases for computer vision that are perceptual aliasing and dynamic range. The final alignment of the trajectory built from the camera images and the architectural blueprint straightforwardly aligns the 50 thousand images captured in 30 minutes to each location at which they were captured, computed even faster on a conventional laptop computer. As a result, the virtual tour web interface using that data offers people throughout the world the experience of exploring the fine architectural details of the UNESCO recorded major heritage building.

The remainder of this article thus reviews related works on SLAM (Sec. 2) before developing our spherical SLAM-to-blueprint alignment method (Sec. 3). Then, the results section (Sec. 4) applies the concept of spatial random access within a spherical video to the al-Qarawiyyin mosque with demonstration of the new virtual tour software produced, before conclusion (Sec. 5).

2. Related works on SLAM

Visual Simultaneous Localization And Mapping is the technique of computing the motion of a camera in 3D space together with a model of the environment in which the camera is displaced, both from the images themselves captured by the camera [SHS*23]. SLAM appeared in the robotics research community [DWRN96] with the underlying principle of making the computing system to reconstruct the environment and estimate the camera pose progressively while the camera is moving, in real-time. Similarly, the computer vision community developed the Structure From Motion (SFM) technique [LH81] that is itself very similar to one of the core aspects of the photogrammetry [Lin13], the latter being complemented by Multi-View Stereoscopic reconstruction to compute a dense model of the environment [SR23], without the real-time constraint.

Indeed, the use of all the pixels captured within many images represents large amounts of data to process that can lead to accurate reconstruction but prevents real-time processing despite using costly hardware. That is why among the large variety of SLAM works, the most efficient ones for real-time applications build their software architecture with visual features that are the most reliable natural landmarks of the scenes where the camera is moved. The image processing that extracts such visual features at the earliest stage of the SLAM pipeline allows lightweight processing in the further so-called stages of tracking, mapping and loop closure [SHS*23]. As a result, a sparse map of the reconstructed feature locations represents the environment [MAT17].

For both the SFM and SLAM techniques to work, they require a significant overlapping of the scene content within pairs of images captured, generally recommended at 80% of each image content for photogrammetry [Lin13]. Hence, in practical scenarios, the larger the camera Field-of-View (FoV), the easier to reach such a content overlapping ratio, furthermore decreasing the amount of images to capture. Since the advent of compact dual-fisheye cameras [Li06], the full spherical FoV is available at the level of compact consumer grade cameras (*e.g.* the Ricoh Theta camera series, see Fig. 1b). Interestingly, the SLAM community has recently considered the equirectangular representation (Fig. 1a) of the spherical FoV [SSS19] furthermore with the StellaVSLAM software made available open-source[†]. Increasing the camera FoV for images of the same pixel definition degrades the angular resolution, thus possibly locally decreasing the accuracy of estimations. But with this type of images, the success rate of SLAM gets close to 100%, contrary to using cameras of narrower FoV [CCK*21], making way

[†] https://github.com/stella-cv/stella_vslam

easier its practical use for environment mapping, in addition to the shorter data capture duration.

Finally, whatever the camera used, any SLAM system drifts significantly over the traveled distance if the camera does not loop over its past trajectory with the SLAM system able to detect such loops and optimize the map built until then. Using a SLAM system with such a critical loop closure capability, e.g. StellaVSLAM [SSS19], one must consider looping on the camera trajectory to ensure the correctness of the overall map and reconstructed camera poses.

3. Spherical SLAM-to-blueprint alignment

SLAM, applied to a spherical video, allows for the creation of a sparse cloud of the environment and the trajectory taken by the camera. A link must then be created between the camera trajectory and the site blueprint chosen for interactive navigation. This link is achieved by registering the blueprint and the SLAM result (sparse point cloud or trajectory). There are several possibilities for registering, and the literature on the subject is abundant. We will only use the one based on points because it is sufficient, as we will see in the results section (Sec. 4). If we transpose this method to our problem, it seems more natural to align the sparse point cloud with the blueprint because they both represent the environment elements (walls, pillars, etc.). However, the representation of the reconstructed environment can be very different from that shown on the blueprint. On the other hand, the sparse reconstruction is exact but not very precise, as shown in Figure 1, which makes the choice of correspondences difficult and can lead to significant discrepancies between the trajectory actually taken and the free spaces. We therefore chose to register the reconstructed trajectory on the blueprint by choosing manually a set of way points on this trajectory and their equivalent on the blueprint.

The blueprint set of points is $\mathbf{q} = (q_1, q_2, ..., q_n)^t$, where q_i represents the two dimensional coordinates. The trajectory set of points is $\mathbf{p} = (p_1, p_2, ..., p_n)^t$, where p_i represents the two dimensional coordinates. We suppose that the link between these two sets of points can be modeled by a similarity transformation:

$$\mathbf{p} = \mathbf{s} \cdot \mathbf{R} \cdot \mathbf{q} + \mathbf{T}, \tag{1}$$

where $\mathbf{s}$ is the scale vector (different scale for each axis); $\mathbf{R}$ is the rotation matrix and $\mathbf{T}$ the translation vector.

The alignment between the blueprint and the trajectory, consists in simultaneously estimating the scale, the rotation matrix and the translation vector by solving the equation system constructed by stacking the equations (1) for all pairs $\mathbf{q}$ and $\mathbf{p}$.

The selection of used points was done manually to guarantee the consistency of the registration. Moreover this operation is not time-consuming because about ten points are enough. The advantage of this approach is to force the trajectory to be on the free spaces actually used by the camera and to guarantee the consistency of their superposition for the comfort during navigation. This alignment makes it possible to superimpose the trajectory on the blueprint (Fig. 1.e).

4. Results

SLAM applied to a spherical video gives the camera pose and time for each frame (trajectory) and the reconstructed sparse 3D model of the environment surrounding the trajectory of the camera. Thanks to the alignment, we have a direct relation between a chosen location on the blueprint, its corresponding spatial pose on the trajectory, its corresponding time and finally the spherical image. We developed an interface, based on Javascript language, that displays the blueprint and gives the possibility to the user to chose interactively the location for which he wishes to view the spherical image. It is also possible to modify the view direction in the spherical image.

To validate our approach, we chose an important and famous architectural and spiritual Moroccan monument: al-Qarawiyyin mosque. This Mosque is one of the oldest universities in the world, and one of the most prominent historical landmarks that the city of Fez prides itself on as the scientific and spiritual capital of Morocco. al-Qarawiyyin University was founded in the year 859 *AD*. The radiance of this institution has continued for nearly 12 centuries, as it has remained a hub of active scientific and jurisprudential movements, and a breeding ground for the emergence of many Arab and Western scholars as well. al-Qarawiyyin Mosque is situated in the western sector of the city of Fez. It was transformed from a small mosque into a massive multi-facility complex thanks to numerous expansions and renovations. The initial goal was to build a mosque limited to worship, but it evolved thanks to the care of the successive states that ruled Morocco to become a scientific institute providing education in various fields. The entrances of al-Qarawiyyin University are distributed across 17 gates. It is characterized by its green domes, its ancient wooden mihrab, and its circular fountain that occupies the center of its courtyard. Its roof is divided between a covered section and an open section to the sky. This site is particularly relevant for illustrating the benefits of spherical video and virtual tours. In January 2024, we were given permission to carry out a two-hour scanning campaign. Since this time was not sufficient to carry out 3D lasergrammetry surveys of the entire mosque or even to take enough photographs to carry out photogrammetry, we opted to produce spherical videos with a Ricoh Theta camera (Fig. 1.b). Each aisle of the mosque was covered, filming at $8K$ resolution. The trajectory of The camera is materialized by the green dots visible in Figure 1. The total duration of the video is approximately 30 minutes.

The interface of our software (Fig. 2) is composed of two areas. The top left area shows an interactive blueprint and the right one, which is also interactive, is dedicated to the spherical image visualisation and exploration. When the user moves the mouse to the blueprint area, its image is zoomed to facilitate the choice of the point to visualize on the blueprint. The software converts the acquired position (pixel coordinates) in its corresponding metric position on the trajectory and its equivalent time. This time is then used to reach the good spherical image directly in the video. In fact, the video is used directly, which avoids breaking it down into images, thus saving memory.

5. Conclusion

We presented a solution which permits to access directly to a spherical image by entering its corresponding position on a blueprint. This solution makes the virtual exploration of buildings possible by using only a spherical video as the only input data. This solution is suitable when access to a monument is very limited or when its 3D survey is not possible or not allowed. In addition, making a video allows to cover more space and faster than taking static spherical images. The pipeline was applied to the case of a famous mosque for which access, outside of prayer times, is very limited. This solution improves accessibility to the monument for all people prevented of access, whatever the reason. Our pipeline could be also interesting for archaeologists or art historians who could use a spherical camera to quickly record the visited sites.

References

[CCK*21] CHAPPELLET K., CARON G., KANEHIRO F., SAKURADA K., KHEDDAR A.: Benchmarking Cameras for Open VSLAM Indoors. In *Int. Conf. on Pattern Recognition* (2021), pp. 4857–4864. 2

[DWRN96] DURRANT-WHYTE H., RYE D., NEBOT E.: Localization of autonomous guided vehicles. In *Robotics Research* (London, 1996), Giralt G., Hirzinger G., (Eds.), Springer London, pp. 613–625. 2

[LH81] LONGUET-HIGGINS H.: A computer algorithm for reconstructing a scene from two projections. *Nature 293* (1981), 133–135. 2

[Li06] LI S.: Full-view spherical image camera. In *Int. Conf. on Pattern Recognition* (2006), vol. 4, pp. 386–390. 2

[Lin13] LINDER W.: *Digital Photogrammetry: Theory and Applications*. 06 2013. 2

[MAT17] MUR-ARTAL R., TARDÓS J. D.: ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Trans. on Robotics 33*, 5 (2017), 1255–1262. 2

[RCM23] RAHAMAN H., CHAMPION E., MCMEEKIN D.: Outside inn: Exploring the heritage of a historic hotel through 360-panoramas. *Heritage 6*, 5 (2023), 4380–4410. 2

[SHS*23] SAHILI A. R., HASSAN S., SAKHRIEH S. M., MOUNSEF J., MAALOUF N., ARAIN B., TAHA T.: A Survey of Visual SLAM Methods. *IEEE Access 11* (2023), 139643–139677. 2

[SNE21] SARKADY D., NEUBURGER L., EGGER R.: *Virtual Reality as a Travel Substitution Tool During COVID-19*. 01 2021, pp. 452–463. 2

[SR23] STATHOPOULOU E. K., REMONDINO F.: A survey on conventional and learning-based methods for multi-view stereo. *The Photogrammetric Record 38*, 183 (2023), 374–407. 2

[SSS19] SUMIKURA S., SHIBUYA M., SAKURADA K.: OpenVSLAM: A versatile visual SLAM framework. In *ACM Int. Conf. on Multimedia* (2019), pp. 2292–2295. 2, 3

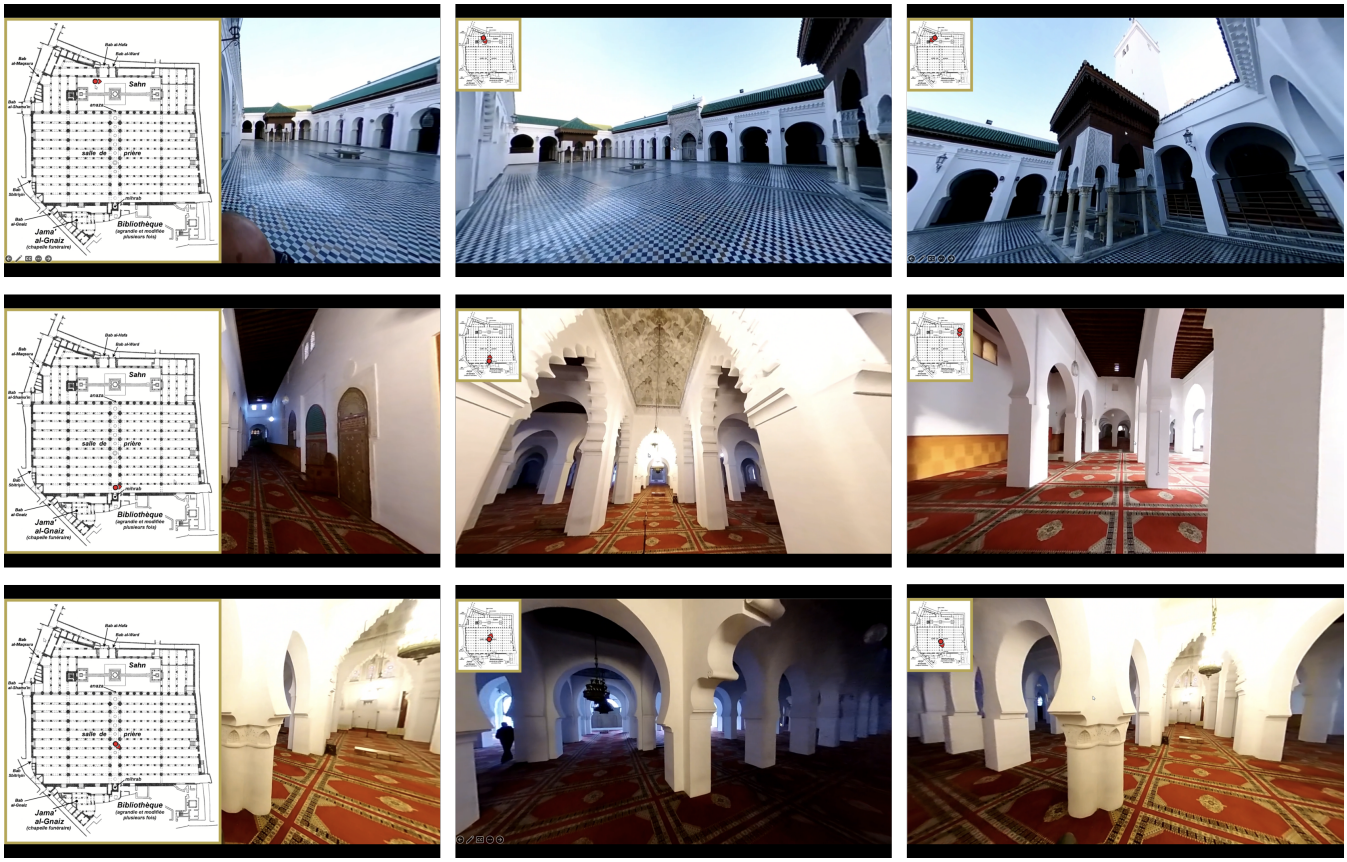

Figure 2: *Three navigation and visualisation samples. Each row shows different orientations of the view at the same position. The left column shows the zoomed blueprint and a view of a partial view of the spherical image. The chosen position and the orientation of the view are visible by the red arrow. The middle and the right columns correspond to different orientations at the same position.*