# DEPTH CAMERAS ON UAVs: A FIRST APPROACH

A. Deris[a], I. Trigonis[a], A. Aravanis[b], E.K. Stathopoulou[c, *]

[a]Olyzon Consulting, Kolokotroni 13 Trikala, Greece - info@olyzon.gr
[b]TopoDrone.gr - Karavela & Gazi, 31100 Lefkada, Greece - aravanisandreas@gmail.com
[c] National Technical University of Athens, School of Rural and Surveying Engineering, Lab. of Photogrammetry
Zografou Campus, Heroon Polytechniou 9 st., 15780, Zografou, Athens, Greece - elliestath@central.ntua.gr

**Commission II**

**KEY WORDS:** depth cameras, 3D reconstruction, Structure from Motion, SLAM, depth, UAV

**ABSTRACT:**

Accurate depth information retrieval of a scene is a field under investigation in the research areas of photogrammetry, computer vision and robotics. Various technologies, active, as well as passive, are used to serve this purpose such as laser scanning, photogrammetry and depth sensors, with the latter being a promising innovative approach for fast and accurate 3D object reconstruction using a broad variety of measuring principles including stereo vision, infrared light or laser beams. In this study we investigate the use of the newly designed Stereolab's ZED depth camera based on passive stereo depth calculation, mounted on an Unmanned Aerial Vehicle with an ad-hoc setup, specially designed for outdoor scene applications. Towards this direction, the results of its depth calculations and scene reconstruction generated by Simultaneous Localization and Mapping (SLAM) algorithms are compared and evaluated based on qualitative and quantitative criteria with respect to the ones derived by a typical Structure from Motion (SfM) and Multiple View Stereo (MVS) pipeline for a challenging cultural heritage application.

## 1. INTRODUCTION

Depth cameras are systems able to capture simultaneously color and depth information of every pixel of the scene, resulting dense point clouds or triangulated meshes. In the latest years, this kind of sensors -also known as RGB-D or range cameras- has increased popularity in the photogrammetry, computer vision and robotics communities, although its origins are in home entertainment and gaming industry. Numerous such system configurations, available in the market or custom made solutions according the needs of every project, can be categorized with respect to their working principle as Time of Flight cameras (ToF), active or passive vision cameras. The latter are based on well-known stereo view principles that calculate the object depth through triangulation, while the active cameras work similarly to structured light scanners by projecting infrared (IR) light on the object and computing the deformations (light-coding). ToF sensors are very similar to commercial ToF scanners that convert the computed time delay into depth information. Along with their cost effectiveness, ease of usage and light-weight shape, such sensors have gained popularity among the 3D reconstruction researchers and professionals mainly due to their effectiveness in retrieving a detailed scene in almost real time also for dynamic scenes. However, due to their noise sensitivity and resolution limitations the scientific community is still cautious regarding the usage of RGB-D cameras on projects that require outcomes of high accuracy, e.g. monitoring, geometric recording and modeling of cultural heritage assets and sites (Gonzalez-Jorge et al., 2013; Lachat et al., 2015). Indeed, depth cameras have been mostly used for indoor scene projects due to their sensitivity in noise, minimum and maximum distance to object constraints, restricted field of view and connectivity limitations.

In mapping studies with depth cameras, algorithms known as Simultaneous Localization and Mapping (SLAM) are generally used for frame registration, i.e. camera orientations and sparse 3D reconstruction. Originally dedicated to robot navigation, the main insight of SLAM algorithms family is to calculate the sensor movement with simultaneous reconstruction of the 3D scene points and was firstly introduced by Leonard et al., 1991) based on earlier work by Smith et al. (1990). The so-called environment features (landmarks) are used to calculate the position of the sensor in real time using EKF (extended Kalman filter). A usual SLAM pipeline can be summarized as: landmark extraction, data association, state estimation, state and landmark update.

Since Microsoft Kinect's emergence some years ago, great examples of such cameras are continuously made available to the public such as Asus Xtion Pro Live and Intel RealSense, which work basically by projecting a known IR pattern on the object and calculating thus the scene depth from the deformations. The recent release of Kinect 2.0 is based on ToF technology, while other cameras such as Bumblebee and Stereolab ZED make use of the stereo vision to estimate the depth.

Stereolab's CUDA-based ZED camera, used in this project, is a low-cost, lightweight and handy camera system configuration based on passive stereo vision. It consists of two optical sensors on a fixed base distance of 120 mm (Figure 1) with a maximum sensor to object distance of 20 m. It outputs either a real-time reconstructed mesh of the scene or a high resolution side-by-side video in a custom video format .svo that, being processed in the accompanying software, calculates the depth of the scene using visual odometry and SLAM. It was chosen as a solution

* Corresponding author

against other depth sensors as it is proven to be more efficient while capturing outdoor scenes -even under direct sunlight- due to the fact that the depth computations do not use infrared light. On the other hand, such passive sensors depend on environment light and cannot function properly in low illuminated or dark scenes.


Figure 1. ZED stereo camera by https://www.stereolabs.com/

The paper is structured as follows: Section 2 presents the related work regarding UAV and depth cameras usage existing in the literature, while Section 3 describes the data acquisition details followed by processing steps and result evaluation is Section 4. Conclusions and future work are outlined in the last section.

## 2. RELATED WORK

Unmanned Aerial Systems have become common practise in collecting visual information such as images and videos in fields of application, such as mapping, structure monitoring and cultural heritage documentation. Such systems can be equipped with optic (RGB) or thermal (IR) cameras either built-in or custom made. Especially for cultural heritage mapping, various studies exist in the literature, evaluating platform's performance and mapping accuracy (Remondino et al., 2011; Nocerino et al., 2013; Brutto et al., 2014; Georgopoulos et al., 2016).

In the robotics community, while depth cameras are being used for indoor mapping applications frequently in the recent years (Henry et al, 2010; Endres et al., 2012; Kerl et al., 2013), there is also a variety of research projects testing SLAM techniques for UAV equipped with laser scanners (Zhang and Singh, 2014) or depth cameras (Bachrach et al., 2014; Loianno et al., 2015; Karrer et al., 2016; Huang et al., 2017). On the other hand, although depth cameras have been tested for their performance in cultural heritage applications (Wenzel et al., 2012; Cappelletto et al., 2016), up to our knowledge, there is no much published work on using depth cameras mounted on UAV for cultural heritage applications.

In this paper we present an ad-hoc configuration setup having a Stereolab ZED camera mounted on an Unmanned Aerial System (UAV), namely a DJI Phantom 2 (Figure 2a). Such setups on UAVs are relatively novel and currently being tested for their performance (Karrer et al., 2016). As a case study for our experiments, a 16th century *turbe* (tomb) located in Trikala, central Greece was chosen (Figure 2b and 2c). This particular structure is an octagonal masonry building of about 12.5m height and diameter of 10m with two arc windows on six out of its eight facades, while there is one entrance on the seventh and the last one has no opening. The 3D reconstruction results provided by ZED camera are compared with the ones achieved using the typical Structure from Motion (SfM) and Multiple View Stereo (MVS) pipeline with images acquired by a similar platform, DJI Phantom 3 Professional and its built-in RGB camera. A network of normally distributed Ground Control Points (GCPs) measured with conventional surveying techniques are also used for the evaluation of the 3D products, scaling and georeferencing.

(a)

(b) (c)

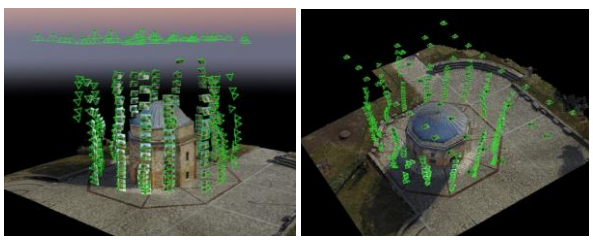
Figure 2. (a) Our custom-made configuration with a ZED camera mounted on a DJI Phantom 2
(b) and (c) the turbe in Trikala, Greece.

## 3. DATA ACQUISITION

Detailed flight plans for both image acquisition techniques were scheduled in advance considering manual control of the platforms and taking into account the complexity of the structure to be recorded, the obstacles of the environment and other flight path constraints such as flight height, battery life etc.

Custom made round markers of 7cm diameter were used as GCPs, sixty four in total, distributed equally on all the eight facades of the building (Figure 3a). The coordinates of the points were calculated using conventional surveying methods (total station) and georeferenced using GPS.

(a)



(b) (c)

Figure 3. (a) GCPs distribution over the object (b) and (c) image network for SfM pipeline.

For the complete image-based reconstruction of the object of interest, a DJI Phantom 3 was used as a platform with its built-in camera. A total of 277 images were captured with an overlap

that reached 90% in most of the cases (Figure 3b and 3c). The baseline between sequential views was 4m in average while the distance from the object varied from 5 up 8 meters, giving an average GSD below 1cm.

A ZED camera was mounted on a DJI Phantom 2 platform using an ad-hoc customization designed by the authors (Figure 2a) managing to keep it fixed as much as possible, as up to our experience this plays an important role to the quality of the final result. It has to be underlined that ZED stereo system calculates the depth maps from real time side-by-side video, allowing the user the possibility to check for possible gaps and occlusions and on-the-job flight plan reschedule, if needed. However, in our case video frames recording in .svo format was chosen instead of real-time reconstruction, as it enables raw data extraction, post-processing and possible reuse. Thus, a video mode of 720p with 30fps was chosen to keep balance between file size and adequate output resolution - namely 2560*720 pixel for the two cameras (side-by-side). After various experiments, the realization of sequential smaller flights was decided, rather than flying over the entire structure at once, allowing for small data size and easier video manipulation (see next Section). Indeed, 16 flights were planned; 8 along the center of every facade and 8 along its edges (Figure 4). However, in order to ensure full coverage and data redundancy, each of the 16 flights was executed twice; once starting from the ground and ending to the roof and one reverse, trying to uphold a constant sensor-object distance and keeping constant and low velocity. It was a relatively fast acquisition, as each flight duration was below two minutes. The raw data of every flight were saved as .svo files with each file size varying from 4.5 to 5.5 GB. During acquisition, ZED explorer software was used to control the camera movement with real time visualization of the video.

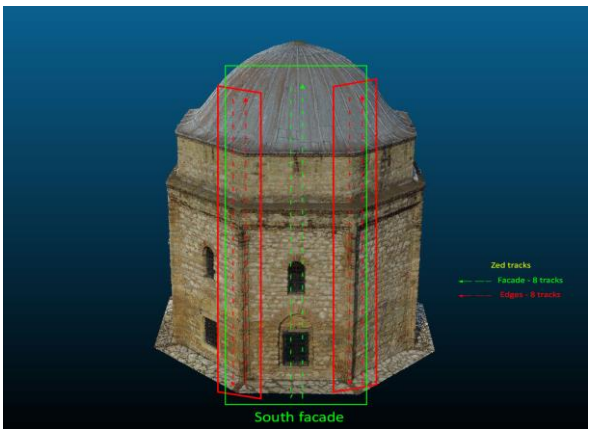



Figure 4. The flight path of ZED camera for every facade. Green trace indicates the double path (upwards and downwards) for the central part of the facade, while the red ones are the flights executed to capture the edges.

## 4. DATA PROCESSING AND EVALUATION OF THE RESULTS

### 4.1 Image-based modelling

SfM and MVS are well-established semi-automatic solutions in the field of image-based 3D reconstruction. Similarly to SLAM, they basically rely on accurate detection, description and matching of image feature points to estimate the camera pose and the sparse 3D structure of the scene at the same time. MVS procedure allows for further densification of this point cloud, as almost every pixel of the scene is reconstructed in the 3D space, resulting in a detailed point cloud. Various software implementations of these algorithms are currently available.

The optic images captured with DJI Phantom 3 were processed using state of the art Structure from Motion (SfM) and Multiple View Stereo (MVS) algorithms, as implemented by Pix4Dmapper Pro software suite. Current algorithms have been proven robust enough to reconstruct the 3D structure of objects even by unordered data with harsh variations in scale, viewing angles and illumination conditions. However, when a high level of accuracy is required such as in the case of cultural heritage mapping, strict camera network plans, pre-calibrated cameras, wise combination of camera settings and the use of GCPs is highly recommended (Wenzel et al., 2013; Nocerino et al., 2014). In our case, no further customization of the camera settings or pre-calibration could be applied due to equipment limitations, the flight plans were however designed in detail in order to avoid information gaps and get the best possible resolution.

After several variable combinations in the available software, a sparse cloud of 2M points, a dense cloud of 11.5M points and a triangulated mesh of 1M faces were generated (Figure 5).

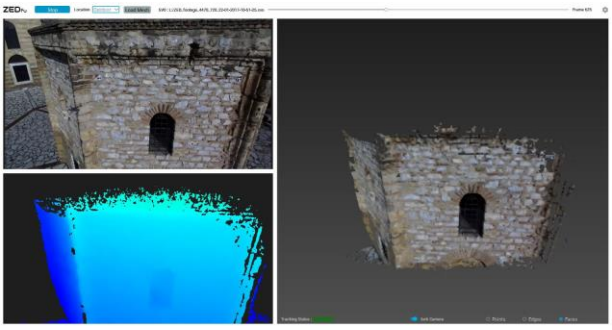
(a) (b)

(c)

Figure 5. (a) Dense point cloud (b) mesh (c) textured mesh generated in Pix4Dmapper Pro.

### 4.2 ZED camera data processing

The .svo files coming from the ZED camera were visualized and processed in ZED Fu software. In this software, the reconstruction of the scene is carried out while the user is able to select the frames to be reconstructed. 16 textured meshes (exported in .obj and .ply) were reconstructed, after a quick selection of the most stable parts of the flights (Figure 6a). Approximately 40K faces were reconstructed for every façade (Figure 6b). Subsequently, the individual meshes were co-registered in the same coordinate system based on the GCPs measurements in CloudCompare, resulting in a final complete mesh.



(a)

Figure 6. (a) SLAM reconstruction of .svo files in ZED Fu (b) mesh output of one .svo file.

Moreover, for testing purposes, frames were extracted from the video files and reconstructed using SfM/MVS pipeline as implemented in Pix4Dmapper Pro (Figure 7), keeping one every 20 frames and using approximately 25 frame sets for one façade (south façade).

Figure 7: SfM/MVS reconstruction of ZED video frames using Pix4Dmapper Pro.

### 4.3 Evaluation

The general visual appearance evaluation of the results, being examined in the state of the generated triangulated mesh (as the direct product of the reconstruction with the ZED camera is), shows that the results are of comparable quality (color and texture homogeneity and precision, gaps, noise), although the number of generated triangles differ significantly with the SfM/MVS pipeline producing almost twice as much triangles as SLAM. However, in certain areas, the visual result produced by ZED camera pipeline is inferior to the typical SfM/MVS and that is mainly due to the existence of some blurry frames within the videos on top of the light sensitivity of the setup. The number of generated points in the dense point cloud state cannot be compared directly, as face to node conversion has to be applied to the SLAM generated mesh.

The 3D products of the two approaches, being in the same reference system (as the same GCPs were used in both) are compared among them, making the convention to keep the SfM/MVS one as "ground truth" dataset using the open source CloudCompare software. As shown in Figure 8, the two clouds (after the conversion of the SLAM mesh to a node point cloud to enable cloud to cloud comparison) do not vary significantly, with most of the points varying some cm (<5). The majority of the problems for ZED occur, as expected, on the edges of the structure and the more dark areas, where the error can be up to three times bigger.

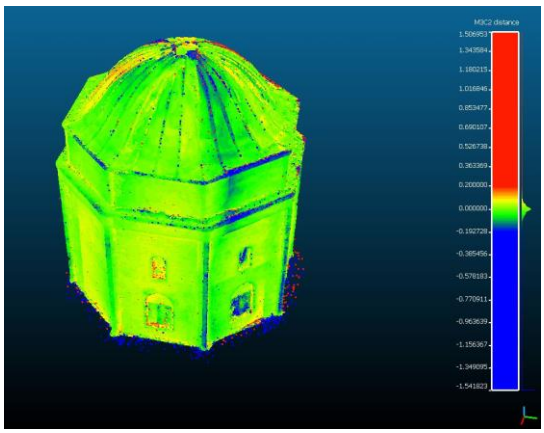
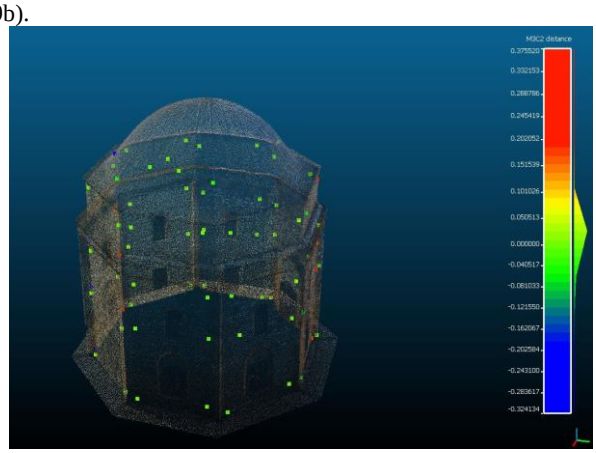
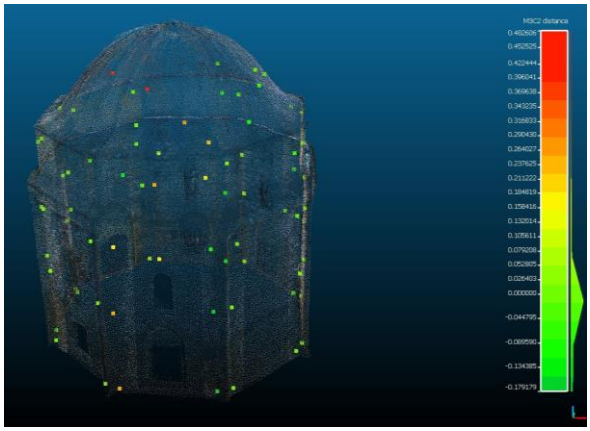


Figure 8: The accuracy of the cloud generated after post-processing the SLAM mesh with respect to the dense cloud of the SfM/MVS pipeline.

Furthermore, considering the comparison of both resulting point clouds keeping the sparse "grid" of GCPs as ground truth information, the SfM/MVS cloud error is in most of the cases below 3cm, except some roof and edges areas that are considered as outliers (Figure 9a). On the other hand, the comparison of the SLAM cloud with the GCPs outputs more areas with large deviations (even 3 to 4 times larger in some cases), while still the majority of them is below 3 cm (Figure 9b).



(a)



(b)

Figure 9: (a) The SfM reconstructed dense point cloud (here subsampled for visualization purposes) and (b) ZED mesh nodes error with respect to the GCPs (in m).

For the one façade that SfM reconstruction was performed on the ZED video frames (south façade), a relative comparison was performed between this result and the SLAM cloud. For the façade itself (central part of the images), the deviation is in the order of magnitude of some cm (<6) yet, as expected, this error increases towards the edges and gets large values moving towards the adjacent façade parts that are reconstructed (Figure 10).

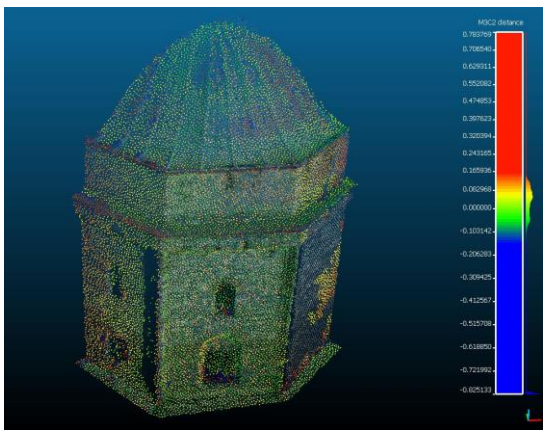


Figure 10: the relative deviation between SLAM cloud and SfM/MVS cloud generated using the ZED frames.

Regarding the processing time, as expected, the reconstruction of the model using the .svo files in ZED Fu provides a much faster solution (approx. 4 min for every facade) than the typical SfM/MVS pipeline implemented in Pix4Dmapper Pro software. This is due to the fact that by definition SLAM provides real-time reconstruction of the scene. As in our approach we followed individual scene reconstruction for every facade, some extra time was needed for the co-registration of the datasets, being still faster than SfM/MVS.

## 5. CONCLUSIONS AND FUTURE WORK

This work studies the use of a Stereolab's ZED stereo camera mounted on a UAV for capturing the depth of outdoor scenes. In our tests, we applied a custom made setup for the stabilization of the camera on a DJI Phantom 2 and used an octagonal structure as a case study. Such passive stereo systems usually implement Simultaneous Localization and Mapping (SLAM) for real-time depth estimation. These 3D reconstruction results are compared and evaluated with respect to the ones generated by a typical image-based reconstruction using Structure from Motion (SfM) and Multiple View Stereo (MVS) algorithms using also Ground Control Points (GCPs) for scaling, georeferencing and control.

ZED stereo camera is a promising setup that could provide comparable results with the SfM/MVS pipeline under certain conditions. Up to now and according to our experiments, it is proven to be adequate for mapping projects with lower resolution and accuracy requirements, yet cannot be implemented as such for high accuracy studies such as structure monitoring or documentation of cultural heritage objects. Our approach with careful flight planning, shorter flight duration and recording of .svo files instead of real-time mesh generation is considered to be helpful in increasing the accuracy of the final product as it enables post-processing and best frame pre-selection by the user, avoiding in this way the blurry parts. Meanwhile, as implemented in our approach the UAV system should be carefully customized, the camera should be fixed on it in the best possible way and the flight velocity should be kept constant to avoid blurry frames. However, uncontrolled parameters, such as wind or insufficient lighting of some areas, may decrease the efficiency of the application and should be taken into consideration. As this study is considered as a first approach to the topic, there are many open issues and challenges to optimize the pipeline and achieve better results e.g. by using a more adequate balance between the sensor-object distance and the processing time and file sizes. Future work will be focused on the customization of the ZED SDK to optimize the results by adding and customizing more variables as well as parameter values changes in the SLAM procedure.

## ACKNOWLEDGEMENTS

The authors would like to thank the Ephorate of Antiquities in Trikala, Hellenic Ministry of Culture and Sports for providing the permission to use this monument as a case study.

## REFERENCES

Bachrach, A., Prentice, S., He, R., Henry, P., Huang, A. S., Krainin, M., Maturana, D., Fox, D., Roy, N., 2012. Estimation, planning, and mapping for autonomous flight using an RGB-D camera in GPS-denied environments. *The International Journal of Robotics Research*, 31(11), pp. 1320-1343.

Brutto, M. L., Garraffa, A. & Meli, P., 2014. UAV platforms for cultural heritage survey: first results. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(5), pp. 227.

Cappelletto, E., Zanuttigh, P. & Cortelazzo, G. M., 2016. 3D scanning of cultural heritage with consumer depth cameras. *Multimedia Tools and Applications*, 75(7), pp. 3631-3654.

Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D. & Burgard, W. 2012. An evaluation of the RGB-D SLAM system. In: *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pp. 1691-1696.

Georgopoulos, A., Oikonomou, C., Adamopoulos, E. & Stathopoulou, E. K., 2016. Evaluating Unmanned Aerial

Platforms for Cultural Heritage Large Scale Mapping. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* XLI-B5, pp. 355-362.

Gonzalez-Jorge, H., Riveiro, B., Vazquez-Fernandez, E., Martínez-Sánchez, J. and Arias, P., 2013. Metrological evaluation of microsoft kinect and asus xtion sensors. *Measurement,* 46(6), pp. 1800-1806.

Henry, P., Krainin, M., Herbst, E., Ren, X. and Fox, D., 2010. RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. *Proc. 12th International Symposium on Experimental Robotics (ISER)*, 20, pp. 22-25.

Huang, A.S., Bachrach, A., Henry, P., Krainin, M., Maturana, D., Fox, D. and Roy, N., 2017. Visual odometry and mapping for autonomous flight using an RGB-D camera. In: *Robotics Research.* Springer, New York, pp. 235-252.

Karrer, M., Kamel, M., Siegwart, R. & Chli, M., 2016. Real-time dense surface reconstruction for aerial manipulation. In: *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pp. 1601-1608.

Kerl, C., Sturm, J. & Cremers, D., 2013. Dense visual SLAM for RGB-D cameras. In: *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pp. 2100-2106.

Lachat, E., Macher, H., Landes, T. & Grussenmeyer, P., 2015. Assessment and calibration of a RGB-D camera (kinect v2 sensor) towards a potential use for close-range 3D modeling. *Remote Sensing,* 7(10), pp. 13070-13097.

Leonard, J. J. & Durrant-Whyte, H. F., 1991. Mobile robot localization by tracking geometric beacons. *IEEE transactions on robotics and automation,* 7(3), pp. 376-382.

Loianno, G., Thomas, J. & Kumar, V., 2015. Cooperative localization and mapping of MAVs using RGB-D sensors. In: *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pp. 4021-4028.

Nocerino, E., Menna, F. & Remondino, F., 2014. Accuracy of typical photogrammetric networks in cultural heritage 3D modeling projects. In: *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* 40(5), pp. 465.

Nocerino, E., Menna, F., Remondino, F. & Saleri, R., 2013. Accuracy and block deformation analysis in automatic UAV and terrestrial photogrammetry-Lesson learnt. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* 2, pp. 5.

Remondino, F., Barazzetti, L., Nex, F., Scaioni, M. and Sarazzi, D., 2011. UAV photogrammetry for mapping and 3d modeling--current status and future perspectives. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* 38(1), pp. C22.

Smith, R., Self, M. & Cheeseman, P., 1990. Estimating uncertain spatial relationships in robotics. In: *Autonomous robot vehicles.* Springer, New York, pp. 167-193.

Wenzel, K., Abdel-Wahab, M., Cefalu, A. & Fritsch, D., 2012. High-resolution surface reconstruction from imagery for close range cultural Heritage applications. In: *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* 39, pp. B5.

Wenzel, K., Rothermel, M., Fritsch, D. & Haala, N., 2013. Image acquisition and model selection for multi-view stereo. In: *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* 40, pp. 251-258.

Zhang, J. & Singh, S., 2014. LOAM: Lidar Odometry and Mapping in Real-time. In: *Robotics: Science and Systems Conference (RSS)*, 2.

https://www.stereolabs.com/ (accessed on 15th January 2017)