INITIAL EVALUATION OF THE POTENTIAL OF SMARTPHONE STEREO-VISION IN MUSEUM VISITS

A. Masiero[a, *], G. Tucci[b], A. Conti[b], L. Fiorini[b], A. Vettore[a]

[a] Interdepartmental Research Center of Geomatics (CIRGEO), University of Padova,
Viale dell'Università 16, Legnaro (PD) 35020, Italy - (masiero@dei.unipd.it, antonio.vettore@unipd.it)
[b] GeCO Lab, Department of Civil and Environmental Engineering, University of Florence,
Via Micheli 8 , 50121 Florence, Italy - (grazia.tucci@unifi.it, arch.a.conti@gmail.com, lidia.fiorini@gmail.com)

KEY WORDS: Smartphone, Stereo-vision, Mobile mapping, Recognition, Laser Scanning, Augmented Reality Museum

ABSTRACT:

The recent introduction of new technologies such as augmented reality, machine learning and the worldwide spread of mobile devices provided with imaging, navigation sensors and high computational power can be exploited in order to drammatically change the museum visit experience. Differently from the traditional use of museum docents or audio guides, the introduction of digital technologies already proved to be useful in order to improve the interest of the visitor thanks to the increased interaction and involvement, reached also by means of visual effects and animations. Actually, the availability of 3D representations, augmented reality and navigation abilities directly on the visitor's device can lead to a personalized visit, enabling the visitor to have an experience tailored on his/her needs. In this framework, this paper aims at investigating the potentialities of smartphone stereo-vision to improve the geometric information about the artworks available on the visitor's device. More specifically, in this work smartphone stereo-vision will used as a 3D model generation tool in a 3D artwork recognition system based on a neural network classifier.

1. INTRODUCTION

During the last decade, the worldwide smartphone market quickly increased arriving to more than 1.4 billion units sold per year. Despite a slight slowdown in 2017, the market is expected to grow further in the next years thanks to the introduction of 5G devices (Richter, 2018), which will probably play a fundamental role in driving the world customers to the new capabilities of Internet of Things (IoT) devices.

According to this scenario, smart interactions between people and objects should be enabled in the next years by the development of a number of applications exploiting artificial intelligence in order to provide in real-time ad hoc services and information tailored on the customers' needs.

Such newly developed technologies will also have consequences on cultural heritage applications. Actually, several museums have already started exploiting certain new technologies, in particular for what concerns virtual/augmented reality, e.g. hundreds of museums are currently available for virtual tours on Google Arts & Culture (Google, n.d.). Despite certain museums also started the introduction of certain interactive applications for their visitors, the availability of smart interaction applications is currently quite limited.

Motivated by the above considerations, this paper aims at investigating certain potentialities of newly developed technologies in cultural heritage applications, and, in particular, for the improvement of visitor museum experience, which is a key factor for the future success of museums (Museums+Heritage Advisor, 2015).

The following observations are at the basis of this work:

- The recent development of artificial intelligence, and in particular convolutional neural networks and deep learning, allowed to obtain very good results in image-based classification and recognition applications.
- Nowadays, smart mobile devices, such as smartphones, are part of the everyday life of most of the world population. Such devices are embedded with several sensors, and, more specifically, they are usually provided of navigation and visual sensors: for what concerns the latter, most of the smartphone producers are currently including multiple rear (and front) cameras in their devices.
- Portable mobile mapping systems, such as backpacks and trolleys, can be very attractive solutions in oder to easily and quickly produce spatial models of indoor environments, such as museums.

According to the above observations, it is clear that the sensors embedded in recent smartphones shall open new perspectives on the possible visitor museum experience. In particular, artificial intelligence shall ensure automatic recognition of artworks, and, consequently, enabling the real-time supply of information and services as requested by the visitors, such as augmented reality directly on the visitor's device.

More specifically, this work aims at investigating the potential of smartphone stereo-vision capabilities in museum visits. The rationale is that of exploiting 3D information provided by the embedded dual-camera in order to enhance the smartphone artwork recognition ability.

Despite a wide number of image-based recognition solutions have already been developed, the recognition of objects based on 3D is still an open research field (Guo et al., 2014, Li et al., 2015, Zanuttigh and Minto, 2017), in particular in the case of 3D information provided by smartphone dual-camera. This work aims at investigating the role that 3D information obtained with smartphone stereo-vision can play to accomplish artwork recognition. The Galleria dell'Accademia di Firenze is used as case study in this paper.

The paper is organized as follows: first, Section 2. provides an overview of the system. Section 3. describes the mobile mapping

*Corresponding author.

system used to acquire the reference model used in the recognition procedure. Then, Section 4. describes the smartphone 3D reconstruction procedure and Section 5.. Finally, some conclusions are drawn in Section 6..

2. OVERVIEW

The recent spread of deep learning, convolution neural networks and other artificial intelligence based recognition methods have tremendously improved the image classification ability of the automatic recognition tools (LeCun et al., 2015). Despite such tools currently allow to obtain good classification results in most of the applications, their use in object recognition is still an open research problem. Many shape descriptors have been proposed in order to extend the use of 2D feature descriptors to 3D data, for instance enabling 2D-3D matches (Geppert et al., 2018): the proposed approaches usually consider either the use of shape descriptors or view-based descriptors (exploiting how object look from a certain point of view) (Su et al., 2015). Then, machine learning or convolutional neural networks based classifiers are typically used in order to perform the object recognition step.

Nowadays, several smartphones embedded with multiple rear cameras are available on the market opening the possibility of quickly producing 3D models just by exploiting the smartphone stereovision capabilities. Despite such capabilities should be soon available to the users (e.g. depth computation module introduced since Android 9), actually up to now few research investigations are available in this field, probably mostly due to the restrictions of most of the mobile device producers, which limited the access to the multiple rear camera resources to developers. Similarly to (Masiero et al., 2018), in this work a smartphone LG G6, Fig. 1, has been considered to cope with this issue. Indeed, the access to LG dual camera is allowed through the LG Mobile Developer Dual Camera module. Despite the application used in this work has been developed specifically for such device, its use can be extended to any other device if the multiple rear cameras are fully accessible.

(a) (b)

Figure 1. Smartphone LG G6: front (a) and rear view (b).

The dual camera availability can be exploited in order to obtain 3D information from images acquired (almost) simultaneously by such to cameras, which hence can be considered as a stereovision system, despite the short baseline clearly limits its depth estimation capability. In particular, 3D reconstruction should be considered only for objects quite close to the cameras.

The rationale of this work is that of exploiting smartphone stereovision in order to produce a rough 3D reconstruction of the artwork of interest to the user (i.e. to a potential museum visitor). An accurate 3D model of the museum artworks is assumed to be available, and hence it is used as reference in a 3D recognition and matching process.

Actually, information from the smartphone navigation system is also assumed to be available, in terms of a rough position estimates inside of the museum and of camera orientation during the image acquisition: such information can be exploited in order to reduce the number of artworks which should be considered in the recognition process. It is worth to notice that, despite the use of more complex navigation strategies (for instance introducing the use of sensors not currently embedded in smartphones (Dabove et al., 2018, Sakr et al., 2018)) allows very good indoor positioning performance, in this work just a rough position information is assumed to be available (Lukianto and Sternberg, 2011, Saeedi et al., 2014, Masiero et al., 2014), hence not requiring the installation of any specific sensor/infrastructure in the museum nor on the user device.

3. REFERENCE MODEL

A commercial indoor mapping solution, namely the Trimble Indoor Mobile Mapping Solution (TIMMS), has been used to produce a 3D model of the Galleria dell'Accademia di Firenze, which is used as case study. Such model is then used as a reference for the recognition procedure.

The TIMMS system integrates data from different sensors in order to provide a reliable 3D model: a FARO light detection and ranging (LiDAR) system, six cameras, an inertial measurement unit. The system weight is 49.5 kg, the allowed moving speed is 1 m/s and the nominal accuracy of the system (related to a local reference system) is 2–4 cm, overall. Details about the TIMMS LiDAR and camera sensors are provided in the following table (specifications taken from the TIMMS data sheet).

Table 1. TIMMS characteristics

	LiDAR
max range	$\approx$ 130 m
error	2 mm
resolution@10 m	$\approx$ 5 mm
resolution@25 m	$\approx$ 12 mm
	Cameras
number	6
sensor resolution	2 Mpixel
focal length	3.3 mm

Despite TIMMS has been used to produce a 3D model of all the Galleria dell'Accademia di Firenze, actually this work focuses only in the analysis of the room shown in Fig. 2, and, more specifically, ten statues inside such room have been considered in the analysis reported in the following sections. Indeed, such number of artworks is quite compatible with the assumption that a rough estimate of the position and orientation of the user's device is available, and hence the artwork recognition procedure can be restricted to only those object in the neighborhood of the user (and in a location compatible with the camera orientation).

The size of the considered room is approximately 36 m $\times$ 9 m. All the considered artworks are within the altitude (4 m) of the point cloud shown in Fig. 2, which is composed by 52 million points.

The ten artworks considered in this work are shown in Fig. 3 and 4. Since they significantly differ in size, the cardinality of their point clouds is quite different, ranging from 65 to 500 thousand points. The smallest and largest ones are shown in Fig. 3(a)

Figure 2. View of the bottom part of the room inside of the Galleria dell'Accademia di Firenze considered as case study in this work.

and , respectively, and their size is 0.50 m×0.50 m×0.85 m and 1.85 m×0.65 m×0.95 m, approximately.

It is also worth to mention that, despite in this work the reference model has been obtained by means of a TIMMS system, different solutions, such as other mobile mapping systems or terrestrial laser scanning surveys (Tucci et al., 2018), can also be considered in order to produce a reference 3D model for the considered approach.

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 3. Statues considered as case studies.

4. SMARTPHONE STEREO-VISION

As already shown in several works in the literature (Poiesi et al., 2017), the standard smartphone camera can be conveniently used as an imaging sensor to obtain 3D photogrammetric models of the objects of interest. However, certain external information should be added in order to obtain a metric reconstruction (Alsubaie et al., 2017). However, since nowadays a double rear camera is embedded in most of the recent smartphones, smartphones can potentially be used for stereo-vision 3D reconstruction.

(a) (b)

Figure 4. Couple of stereo images of a bacchante, Dirce, (Lorenzo Bartolini, Galleria dell'Accademia di Firenze) acquired with a smartphone LG G6.

This work has been deployed by exploiting the stereo-vision system of a smartphone LG G6, whose camera characteristics are reported in Table 2.

Table 2. LG G6 characteristics

sensor resolution	4160 pix × 3120 pix
pixel physical side size	1.12 μm
standard camera focal length	4.03 mm
wide-angle camera focal length	2.01 mm
baseline between cameras	$\approx$ 1.8 mm

Since the two cameras embedded in the LG G6 have a quite different focal length, then the overlapping between them is limited to the central part of the image taken by the wide-angle camera, as shown in Fig. 5. Clearly this limits the resolution of the models produced by such stereo system. Furthermore, the small baseline between the smartphone cameras imposes restrictions on the distance of the object to be reconstructed: the error on the estimated depths increases with the camera-object distance, typically being acceptable only if the distance is within a couple of meters.

In principle, images provided by smartphone dual-cameras can be processed as those of a standard stereo vision system. However, there is a slight synchronization error (typically 10–80 milliseconds) between the stereo-images acquired by the smartphone, which should be taken into account.

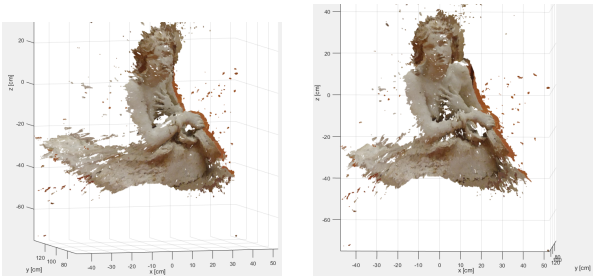
Since cameras have been pre-calibrated, first, images have been undistorted by exploiting the available camera models. Then feature matching (Brown and Lowe, 2007, Bay et al., 2008, Lingua et al., 2009), RANSAC outlier rejection and bundle adjustment are used in order to properly estimate the variation of the relative camera pose with respect to the one calibrated in a static environment. According to our experiments, human hand vibration/instability during the two image acquisitions leads to up to 1–2 degrees of orientation variation with respect to the nominal (pre-calibrated) pose between the stereo-cameras.

After motion compensation, dense matches between the rectified images (an example is shown in Fig. 5 are computed by using a semi global matching-like approach (Hirschmuller, 2008), producing quite detailed models of the photographed artworks (see for instance Fig. 6).

It is worth to notice that, due to estimation errors, the implemented procedure allows to compensate only partially the camera motion (and in particular the orientation change) during the

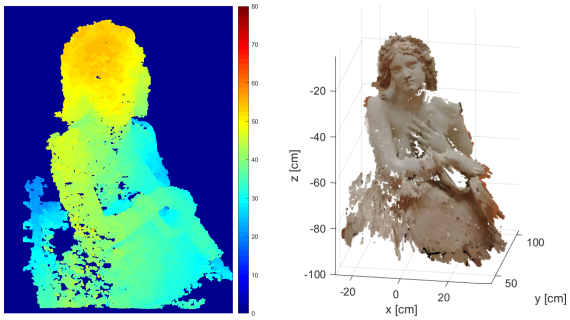
two image acquisitions. Furthermore, certain pixel matching errors can also be present, and the generated 3D point cloud might include certain parts of the environment in the neighborhood to the photographed artwork. Consequently, a point removal step has been implemented in order to remove noisy points and areas which are probably not related to the artwork: such procedure is based on the computation of connected regions in the point cloud and on the removal of areas either too small and disconnected from the other ones, or at a quite different distance from the cameras with respect to the other regions. Fig. 7 shows an example of the obtained results (disparity map and 3D point cloud).

Figure 5. Rectified stereo images (gray-scale).



(a) (b)

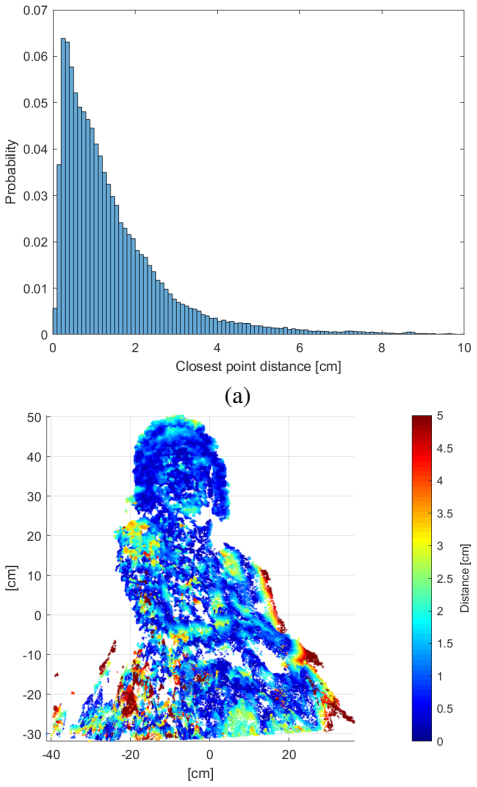
Figure 6. Two views of the 3D reconstruction obtained from the couple of stereo images shown in Fig. 4.



(a) (b)

Figure 7. (a) Disparity map and (b) 3D model obtained from the stereo images shown in Fig. 4 after removing undesired points.

Finally, Fig. 8 compares the 3D model obtained with smartphone stereo-vision with the reference one for the artwork shown in Fig. 4. The two point clouds have been finely aligned by using the Iterative Closest Point algorithm (Besl and McKay, 1992, Chen and Medioni, 1992). From Fig. 8 it is quite clear that most of the unreliable points are in the extreme parts of the model, whereas the point to point distance on most of the Bacchante's face is lower than 1 cm (average point distance in the reference model is 0.3 cm in this area).



(a)

(b)

Figure 8. Comparison between smartphone stereo-vision 3D model and the reference one for the artwork shown in Fig. 4. (a) Point to point distance distribution. (b) Distance mapped on the 3D model.

5. RECOGNITION

Several methods have been recently considered in the literature in order to achieve 3D shape recognition (Su et al., 2015).

Since in this work the reference model has been provided by a mobile mapping system, differently from other recognition systems in this case classification is not based on the collection of a set of 2D descriptors on images of the objects of interest. Furthermore, despite in principle 3D shape descriptors (for instance Normal Aligned Radial Feature (NARF) (Steder et al., 2010)) can be used here, in practice the presence of noisy points and of gaps in the smartphone 3D model make such approach not so attractive.

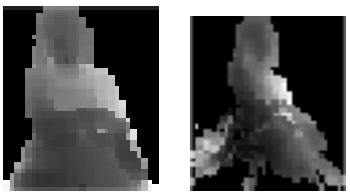
Instead, in this work the recognition step take advantage of the device orientation estimated by means of the inertial sensors embedded in the smartphone. Such information is used in order to approximately determining the observation direction of the smartphone with respect to artworks in its neighborhood.

Then, a low resolution depth map is obtained from the smartphone 3D model by projecting the points along the observation direction (Fig. 9 shows an example of statue point cloud obtained with the TIMMS and Fig. 10(b) shows the obtained depth map). Similarly, each of the point clouds of the considered artworks is

also projected along the approximate observation direction of the device, obtaining reference depth maps (Fig. 10(a)) to be compared with the one obtained from the smartphone reconstruction.

(a) (b)

Figure 9. Two views of the 3D model obtained by means of TIMMS of the statue shown in Fig. 4.



(a) (b)

Figure 10. Comparison between depth computed from the reference model (a) and the one derived from smartphone stereo-vision 3D model (b) for the artwork shown in Fig. 4.

Actually, since the orientation information obtained by the inertial sensors is only an estimation of the real observation direction, multiple depth maps can computed for each artwork in order to partially take into account of the error on the observation direction. For instance, in this work a set of 48 rotations uniformly spaced of 3 degrees along both the vertical and horizontal direction have been considered. Then, the Euclidean distance between points in the depth maps produced by the reference and by the smartphone model are computed: since several depth maps are available for the reference model, only the one associated to the minimum distance from the smartphone depth map is considered in the following.

Finally, the artwork classification is made by using a (properly pre-trained) multi-class SVM classifier (in the one-against-one version) (Hsu and Lin, 2002): in particular, the 60th-75th-90th percentiles (along with the RMS) of the Euclidian distance between the depth maps is used fed as input of the SVM classifier in order to reduce the influence of noisy reconstructed points.

In the considered ten case studies the presented approach allowed to obtain approximately a 70% correct recognition performance: such performance is probably highly affected by the quite limited in size and noisy 3D reconstructions obtained with the smartphone stereo-vision.

6. CONCLUSIONS

The introduction of recently developed methods exploiting 3D information, such as augmented and virtual reality, are becoming quite popular in many cultural heritage applications (Bekele et al., 2018), and can play an important for improving the visitor experience in museums.

In this paper, smartphone stereo-vision has been used in order to obtain 3D models of artworks, and then initially evaluating the potentiality of such generated 3D information for object recognition, using the Galleria dell'Accademia di Firenze as case study.

A commercial indoor mapping solution, namely the Trimble Indoor Mobile Mapping Solution (TIMMS), has been used to produce a 3D model, which is used as a reference for the recognition procedure. The use of an easy portable mobile mapping system is a particularly convenient solution to such aim: indeed, after quick surveys, it can generate relatively accurate 3D models of indoor environments. Despite the point cloud generated by the mobile mapping system is sufficient in order to implement the approach described in this work, clearly the use of a higher level of 3D information (e.g. BIMs (Banfi, 2017)) can be considered as well.

A stereo-image dataset has been acquired with a smartphone LG G6 by means of an ad hoc Android application, implemented in order to (almost) simultaneously acquire images by the two rear cameras of such device. Assuming the cameras to be pre-calibrated, the acquired stereo-images can be rectified, a depth map can be obtained after computing a disparity map, and a 3D model can be obtained via triangulation (check for instance Fig. 4, 5 and 6).

Then, this paper provided an initial evaluation of the recognition results which can be obtained by comparing the (typically quite noisy) depth maps obtained by smartphone stereo vision with those generated by the TIMMS model (Fig. 9). To this aim, the artwork point clouds have properly been re-oriented according the approximate observation direction of the smartphone, obtained by the inertial system embedded in the device.

The results obtained with the considered multi-class SVM classifier are decent, but not so exciting. Actually, the classification performance is probably affected by several factors, such as the device motion, that has been only partially compensated, the short baseline of the smartphone dual camera system (and the different focal length of the two cameras), which ensured the generation of quite good 3D information only for the parts of the objects close to the smartphone, and the availability of smartphone 3D information only for certain areas of the considered objects, which might be insufficient when considering an object not so small, or when the photographed area does not contain any specific detail.

However, since 3D reconstruction from smartphone is reliable only for quite close objects, the aim of this work is rather that of supporting a more general recognition system than substituting image based classification. Nevertheless, improvement to the recognition procedure and more detailed investigations on the system performance will be considered in our future work.

REFERENCES

Alsubaie, N. M., Youssef, A. A. and El-Sheimy, N., 2017. Improving the accuracy of direct geo-referencing of smartphone-based mobile mapping systems using relative orientation and scene geometric constraints. Sensors 17(10), pp. 2237.

Banfi, F., 2017. BIM orientation: Grades of generation and information for different type of analysis and management process. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-2/W5, pp. 57–64.

Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L., 2008. Speeded-up robust features (SURF). Computer vision and image understanding 110(3), pp. 346–359.

Bekele, M. K., Pierdicca, R., Frontoni, E., Malinverni, E. S. and Gain, J., 2018. A survey of augmented, virtual, and mixed reality for cultural heritage. Journal on Computing and Cultural Heritage (JOCCH) 11(2), pp. 7.

Besl, P. and McKay, N., 1992. A method for registration of 3-D shapes. IEEE Transactions on Pattern Analisys and Machine Intelligence 4(2), pp. 239–256.

Brown, M. and Lowe, D., 2007. Automatic panoramic image stitching using invariant features. International Journal of Computer Vision 74(1), pp. 59–73.

Chen, Y. and Medioni, G., 1992. Object modelling by registration of multiple range images. Image and Vision Computing 10(3), pp. 145–155.

Dabove, P., Di Pietra, V., Piras, M., Jabbar, A. A. and Kazim, S. A., 2018. Indoor positioning using ultra-wide band (uwb) technologies: Positioning accuracies and sensors' performances. In: 2018 IEEE/ION Position, Location and Navigation Symposium (PLANS), IEEE, pp. 175–184.

Geppert, M., Liu, P., Cui, Z., Pollefeys, M. and Sattler, T., 2018. Efficient 2d-3d matching for multi-camera visual localization. arXiv preprint arXiv:1809.06445.

Google, n.d. Arts & culture. `https://artsandculture.google.com/`. Online; accessed 19 January 2018.

Guo, Y., Bennamoun, M., Sohel, F., Lu, M. and Wan, J., 2014. 3d object recognition in cluttered scenes with local surface features: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(11), pp. 2270–2287.

Hirschmuller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence 30(2), pp. 328–341.

Hsu, C.-W. and Lin, C.-J., 2002. A comparison of methods for multiclass support vector machines. IEEE transactions on Neural Networks 13(2), pp. 415–425.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature 521(7553), pp. 436.

Li, B., Lu, Y., Li, C., Godil, A., Schreck, T., Aono, M., Burtscher, M., Chen, Q., Chowdhury, N. K., Fang, B. et al., 2015. A comparison of 3d shape retrieval methods based on a large-scale benchmark supporting multimodal queries. Computer Vision and Image Understanding 131, pp. 1–27.

Lingua, A., Marenchino, D. and Nex, F., 2009. Performance analysis of the SIFT operator for automatic feature extraction and matching in photogrammetric applications. Sensors 9(5), pp. 3745–3766.

Lukianto, C. and Sternberg, H., 2011. Stepping – smartphone-based portable pedestrian indoor navigation. Archives of photogrammetry, cartography and remote sensing 22, pp. 311–323.

Masiero, A., Fissore, F., Piragnolo, M., Guarnieri, A., Pirotti, F. and Vettore, A., 2018. Initial evaluation of 3D reconstruction of close objects with smartphone stereo vision. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-1, pp. 289–293.

Masiero, A., Guarnieri, A., Pirotti, F. and Vettore, A., 2014. A particle filter for smartphone-based indoor pedestrian navigation. Micromachines 5(4), pp. 1012–1033.

Museums+Heritage Advisor, 2015. Insight: Visitor experience is key to museum and heritage resilience. Online; accessed 19 January 2018.

Poiesi, F., Locher, A., Chippendale, P., Nocerino, E., Remondino, F. and Van Gool, L., 2017. Cloud-based collaborative 3d reconstruction using smartphones. In: Proceedings of the 14th European Conference on Visual Media Production (CVMP 2017), ACM, p. 1.

Richter, F., 2018. The smartphone market is in a pre-5g slump. `https://www.statista.com/chart/14070/global-smartphone-shipment-forecast/`. Online; accessed 19 January 2018.

Saeedi, S., Moussa, A. and El-Sheimy, N., 2014. Context-aware personal navigation using embedded sensor fusion in smartphones. Sensors 14(4), pp. 5742–5767.

Sakr, M., Masiero, A. and El-Sheimy, N., 2018. Evaluation of dynamic ad-hoc UWB indoor positioning system. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-1, pp. 379–385.

Steder, B., Rusu, R. B., Konolige, K. and Burgard, W., 2010. NARF: 3D range image features for object recognition. In: Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), Vol. 44.

Su, H., Maji, S., Kalogerakis, E. and Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision, pp. 945–953.

Tucci, G., Visintini, D., Bonora, V. and Parisi, E. I., 2018. Examination of indoor mobile mapping systems in a diversified internal/external test field. Applied Sciences 8(3), pp. 401.

Zanuttigh, P. and Minto, L., 2017. Deep learning for 3d shape classification from multiple depth maps. In: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, pp. 3615–3619.