3D RECONSTRUCTION WITH A COLLABORATIVE APPROACH BASED ON SMARTPHONES AND A CLOUD-BASED SERVER

E. Nocerino [a], F. Poiesi [b], A. Locher [c], Y. T. Tefera [b], F. Remondino [a], P. Chippendale [b], L. Van Gool [c]

[a] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Email: (nocerino, remondino)@fbk.eu
[b] Technologies of Vision (TeV) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Email: (poiesi, tefera, chippendale)@fbk.eu
[c] Computer Vision Lab, ETH Zurich, Switzerland - Email: (alocher, vangool)@vision.ee.ethz.ch

Commission II

KEY WORDS: smartphone, 3D reconstruction, low-cost, collaborative, SfM, photogrammetry, dense image matching.

ABSTRACT:
The paper presents a collaborative image-based 3D reconstruction pipeline to perform image acquisition with a smartphone and geometric 3D reconstruction on a server during concurrent or disjoint acquisition sessions. Images are selected from the video feed of the smartphone's camera based on their quality and novelty. The smartphone's app provides on-the-fly reconstruction feedback to users co-involved in the acquisitions. The server is composed of an incremental SfM algorithm that processes the received images by seamlessly merging them into a single sparse point cloud using bundle adjustment. Dense image matching algorithm can be lunched to derive denser point clouds. The reconstruction details, experiments and performance evaluation are presented and discussed.

1. INTRODUCTION

Image-based approaches have become viral for 3D digitization in the last years. Requirements and needs of digital replica significantly change according to the application field, steering the choice of equipment, as well as software tools. In industrial metrology, accuracy and reliability are crucial factors, which imply the adoption of high-cost, professional-grade camera and lens systems, coupled with software applications fully manageable only by expert operators. In the geospatial domain, completeness of the results, accuracy of georeferencing, handling of huge amount of data, reliability and speed of automatic procedures, integration and homogenization of data from different sources are key topics. Researches and studies in the cultural heritage field specifically focus, among other topics, on colour fidelity, geometric level of details, handling, visualization and sharing of 3D models.
Today, a range of economic activities, whose origin can be traced back to the beginning of the new millennium, is driving the digital economy all around the word, i.e. the creative industries (EY, 2015). Also referred to as 'creative and cultural industries' or 'creative and digital industries', they embrace thirteen sub-sectors: advertising, architecture, arts and antiques market; crafts; design; designer fashion; film and video; music; performing arts; publishing; interactive leisure and software; software and computer services; television and radio (Skillset, 2013). People working in the creative economy rely on their individual creativity, skill and talent, to produce economic values.
To answer the needs of this growing community, technologies and tools are rapidly developing and changing. Emblematic is the progress of 3D printers, more and more used to realise fully-operational, market-ready products rather than quick and cheap prototypes (The Economist, 2011). Similarly, we are witnessing a 'democratization' and massive spread of 3D digitization techniques (Alderton, 2016; Nancarrow, 2016; Santos et al., 2017), with an increasing demand for hardware and software solutions economically accessible, easily understandable and manageable by almost anyone wills to express his or her creativity through 3D digital products.

The work described in this paper arises in this context and presents a collaborative image-based 3D digitization pipeline. Different users acquire – simultaneously or in separate sessions – images with their smartphones and images are then 3D processed via a cloud-based server. A smartphone's app provides on-the-fly visual feedback about the 3D reconstruction to users co-involved in the digitization process. The idea is to (i) guide users during the image acquisitions and (ii) combine images from multiple devices from concurrent or disjoint acquisition sessions. The developed approach (Poiesi et al. 2017) and the achieved results, produced in real-world scenarios (i.e. a cultural heritage site and a city square), are compared against reference data, produced employing a professional-grade reflex camera and state-of-the-art image processing software solutions.

2. RELATED WORKS AND MAIN INNOVATIONS

Image-based 3D reconstruction methods using mobile devices have been pioneered in the research domain (Tanskanen et al., 2013; Kolev et al., 2014; Muratov et al., 2016), and are starting to appear on app stores for smart devices (e.g., ItSeez3D[1], TRNIO[2]). These methods implement very similar workflows, relying on Structure from Motion (SfM) and dense image matching (DIM) or Multi View Stereo (MVS) algorithms, run either on the phone or on a server. Being the 3D reconstruction procedure computationally intensive, a feasible solution is to split the process between the mobile device and the cloud-based server (Untzelmann et al., 2013; Locher et al., 2016c). In this case, the smartphone is used as imaging device to capture images of the scene of interest, whereas the SfM and DIM steps are performed on the server. Current 3D reconstruction solutions running on smartphones only offer feedback to single users during image acquisitions, and do not yet seamlessly include collaborative approaches with simultaneous feedback to the multiple. The most common solution for collaborative mapping, based either on Simultaneous Localization and Mapping (SLAM) or SfM approaches, is to produce separate maps that are finally fused together (Forster et al., 2013; Untzelmann et al., 2013; Morrison et al., 2016; Schmuck, 2017).

[1] https://itseez3d.com, last accessed: Oct 2017.

[2] http://www.trnio.com, last accessed: Oct 2017.

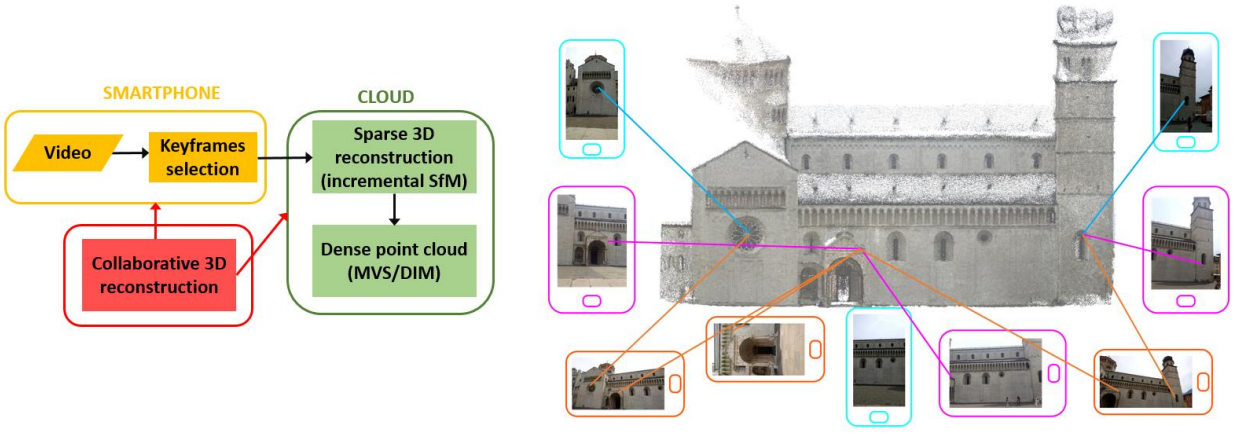

Figure 1: Part of the entire REPLICATE workflow (from Nocerino et al, 2017) jointly performed on smart devices and a cloud-based server (left) and the collaborative aspect of 3D digitization procedures presented in this article (right).

The procedure presented in this paper is based on an incremental SfM approach (Schonberger and Frahm, 2016), which updates and augments the global sparse 3D point cloud when a new image is uploaded. From video acquired by different smartphones, only significant frames are selected, sent to the server and process to increment the sparse 3D reconstruction. The updated results provide the user with visual feedback during the acquisition process and are accessible both on the mobile app and on a web-based visualization service developed on the server.
While the sparse reconstruction of the scene is computed on the server and constantly updated when new images are sent via the SfM procedure, the DIM step produces dense point clouds, made available to the users on a web-based visualization window.

3. THE PROPOSED PIPELINE

The implemented approach is part of an image-based 3D reconstruction workflow under development within the EU funded H2020 project REPLICATE[3] (Nocerino et al., 2017, Fig. 1). A smartphone app allows the image acquisition phase, whereas the processing procedure is jointly performed on the smartphone as well as on a server (Locher et al., 2016c).

3.1 Image acquisition app and device-server communications

Each user running the smartphone app must first be authenticated by the cloud service. A unique smartphone identifier (ID) is assigned based on the user's account credentials, the device's manufacturer, its model and operating system. The smartphone app is used to acquire the video stream, extract the best frames (Section 3.2) and send them to the server for the 3D reconstruction procedure (Section 3.3) Accelerometer measurements from the device's Inertial Measurement Unit (IMU) are also transmitted together with the images to aid pose estimation and object reconstruction. Smartphone vibration is implemented as haptic feedback to help the user to understand whether the images are acquired correctly (i.e. the device motion is not to fast).
Network communication between the reconstruction server and device is bidirectional and asynchronous. The app offers a user the option to start a new acquisition session or to update past acquisitions with new images in case of collaborative approaches (Section 3.4). To visualize updated point clouds as feedback, the smartphone sends periodic requests to the server.

The remote server handles user authentication, processes the images and generates updated results visualized by the device app and web-based interface. The web page enables users to see estimated camera positions and interact with the dense point cloud. The user can share the reconstruction job via an email option with other users, who become contributors. Contributors can then increment the reconstruction of an object by uploading more images of new acquisitions.

3.2 Image selection from smartphone's video stream

Images are selected from the smartphone's app based on both their quality and on their novelty. The selection is based on the computation of a frame's sharpness and the number of new features present (Sieberth et al., 2016). Hence, a 'content rich' frame should be sharp (i.e. in focus and with no motion blur) and it should contain new visual information about the object. Newness is quantified by comparing current feature points with those extracted from previous frames. The quantification of the overlap is calculated for pairs of frames and by using ORB keypoints (Rublee et al., 2011). The image overlap is inferred by matching descriptors among adjacent frames based on the Hamming distance. If no frames were selected for a certain minimum interval of time, a frame is transmitted anyway.

3.3 Orientation and 3D reconstruction

The 3D reconstruction server adopts an incremental SfM algorithm followed by the DIM step, using multiple threads to process independent and asynchronous uploads of images from different users. Two pipelines are under testing: the first, described in Poiesi et al. (2017) and Nocerino et al. (2017) is based on approaches proposed by Sweeney et al. (2015), Schonberger et al. (2016), Locher et al. (2016a) and Locher et al. (2016b). The second procedure, hereafter presented, follows the SfM/DIM pipeline presented by Schonberger and Frahm (2016).

3.4 Collaborative approach

The developed method includes also a collaborative 3D reconstruction which allows the processing of images coming from multiple smartphone devices during concurrent or disjoint acquisition sessions.
For each new image uploaded to the server, the algorithm matches news computed features to those from a subset of images acquired within the same acquisition job. This subset is composed of images already stored in the database featuring high

[3] http://www.replicateproject.eu, last accessed: Oct 2017.

similarities in image content with the new one (Poiesi et al., 2017). Relative image orientation is initially estimated via 2D-3D correspondences using feature points extracted on all the images, regardless of which smartphone they were captured from. If available, the nominal values of the interior orientation parameters are derived from EXIF metadata or extracted from the database containing already registered devices. The essential matrix is then estimated using a five-point algorithm (Nistér, 2004). When the camera parameters are not available, the fundamental matrix is estimated using an eight-point algorithm and, subsequently, the essential matrix is inferred (Nistér and Stewenius, 2006).

Successively a Bundle Adjustment (BA) is applied. We are currently evaluating two approaches to efficiently handle video frames acquired by different devices and progressively process them on the cloud-based server.

The first implementation, used in this paper (Section 4) entails an image-variant self-calibrating BA, i.e., for each image, a set of interior orientation parameters, comprising the principal distance, principal points coordinates and two radial distortion parameters, could be estimated.

The second approach, presented in Poiesi el al. (2017), is based on a two-step procedure, where the interior and exterior orientation parameters are refined as follows. Images acquired in the same session and using same device are forced to share the same camera calibration parameters in the adjustment procedure. A local bounded BA refines only newly uploaded images with their associated points. Once the reconstruction has sufficiently grown, a full BA over all images and points is performed, taking into account the separate camera calibration groups. The implemented two-stage BA saves computation time and increases the stability of the BA optimization.

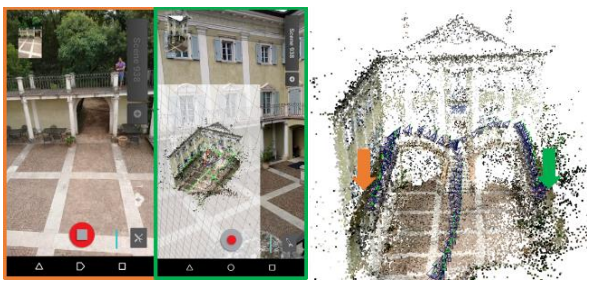

Figure 2: Example of on-the-fly visual feedback inside the smartphone's app (left) during a collaborative digitization process (here two users involved) or on the web browser (right).

3.5 3D reconstruction preview and visualisation

All users involved in a collaborative acquisition can visualize their (simultaneous or joint) 3D reconstruction progresses via a dedicated preview window in the smartphone's app as well as interact with the reconstruction session via a web page.

The preview model in the app shows to the user, while he/she is acquiring images, the sparse point cloud with image positions from all concurrent users (Fig. 2). The preview window runs on a separate thread that periodically sends requests to the server to check and, in case, display the updated scene reconstruction. When the user terminates the acquisition and all images are uploaded, the 3D reconstruction process is completed on the server.

In a web browser, users can visualize the oriented images and the sparse point cloud, download the estimated camera parameters and access intermediate reconstructions.

4. EXPERIMENTS AND VALIDATION

The following section reports three experiments, performed in real case scenarios, to showcase the capabilities of the proposed pipeline. The collected datasets (4.1 – Table 1) and reference data (4.2) are described, and the collaborative reconstruction results are shown together with a quality assessment (4.3). All experiments, acquired with different smartphones, were afterwards processed with 20 cores on an Intel Xeon 2.30GHz computer with 128 GB of RAM.

4.1 Datasets

The experiments entail the acquisition of video streams collected using six different off-the-shelf Android smartphones in three different locations (Table 1). To the authors knowledge, currently there are no datasets that involve multiple and different smartphones recording buildings or objects from different viewpoints. For this reason, our datasets are available for research purposes at the url http://tev.fbk.eu/collaborative3D.

	Seq.	*Device model*	*Resolution (px)*	*No. selected frames*	*Device orientation*
Saranta Kolones	1	**Huawei P9**	1920x1080	152	L/P
	2		1920x1080	154	L/P
	3		1920x1080	210	L/P
	4	**OnePlus One**	1920x1080	117	P
	5		1920x1080	105	P
	6		1920x1080	59	P
	7		1920x1080	44	P
	8	**Samsung S6**	3840x20160	84	L
	9			54	L
	10			56	P
Piazza Duomo	1	**LG Nexus 5X**	1920x1080	64	L/P
	2	**Samsung Galaxy Alpha**	640x480	91	L/P
	3	**SonyZ5**	1920x1080	74	L
Caffe Italia	1	**LG Nexus 5X**	1920x1080	175	L/P
	2	**Samsung Galaxy Alpha**	640x480	218	L/P
	3	**SonyZ5**	1920x1080	107	L/P

Table 1. Main characteristics of the employed datasets. L stands for landscape and P for portrait.

The first dataset (*Saranta Kolones*) features the 'Saranta Kolones' monument within the Pafos archaeological area in Cyprus. The site is ca 16x16x5m. Ten videos (at 30Hz) were recorded by three different smartphones in different orientations (landscape and portrait). Due to network connection limitations on the Saranta Kolones's site, the dataset was recorded using the video mode of the smartphones and post-processed later by the image selection algorithm (Section 3.2). A collaborative acquisition approach was simulated by stirring in and transmitting to the cloud-based server the extracted frames from the different devices.

Other two datasets were acquired using the smartphone's app in the cathedral square of Trento, Italy: from the smartphone's camera video feed, frames were selected by the image selector algorithm during the acquisition (Section 3.2) and directly uploaded to the reconstruction server. The *Piazza Duomo* dataset features the north facing facade of the cathedral (ca 100m wide and 30m tall). The third dataset (*Caffe Italia*) focuses on the south facing facade of a painted building in the same square. The facade is 30m wide/long and 15m tall.

Figures from 3 to 6 show the results of the implemented 3D reconstruction procedure for the three datasets.

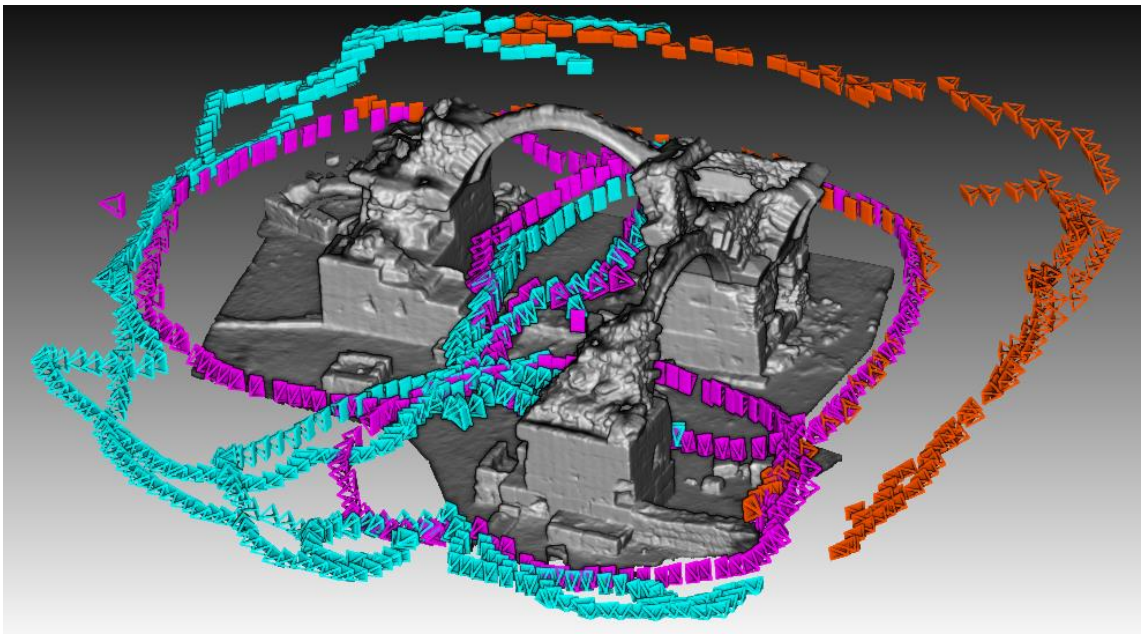

Figure 3: The shaded mesh model of the surveyed *Saranta Kolones* monument. The position and orientation of the extracted frames are shown as pyramids with colours indicating the three employed devices (Table 1).

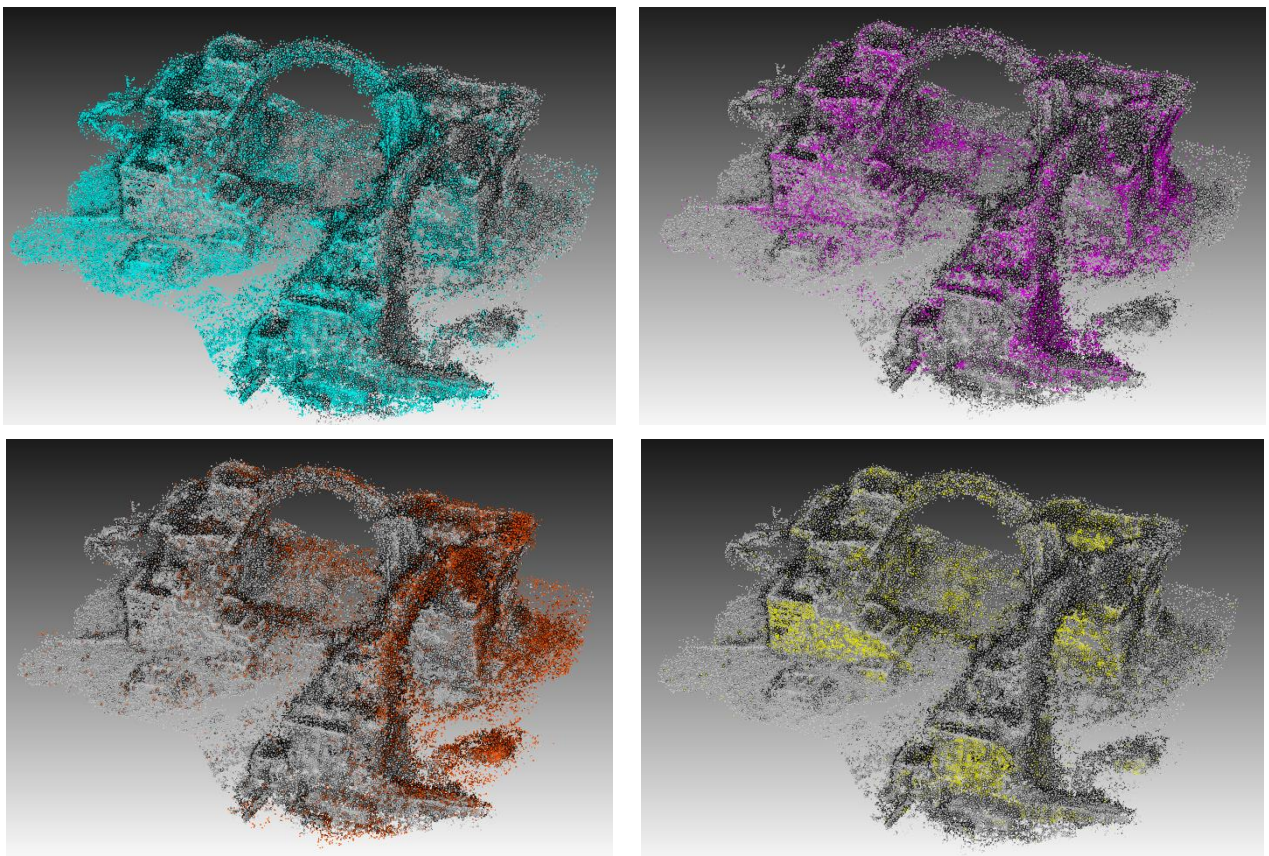

Figure 4: The sparse point cloud for the *Saranta Kolones* dataset. The points are coloured based on the smartphone they are triangulated from (Table 1); in grey the entire point cloud; in yellow the points triangulated from images belonging to multiple devices.

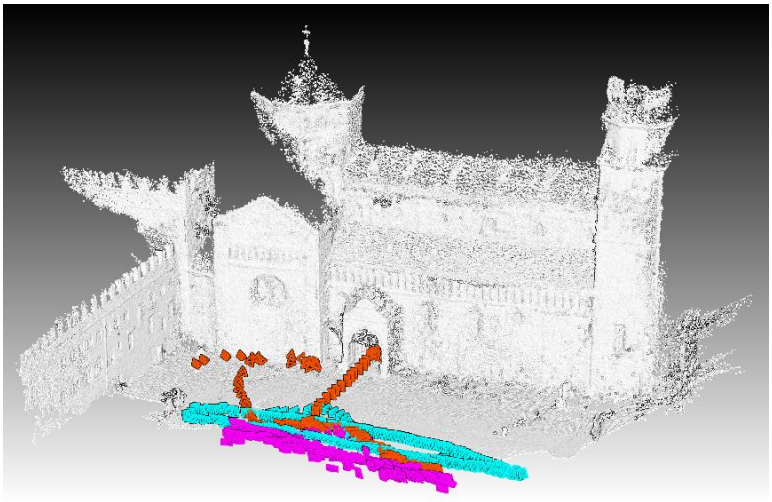

Figure 5: The shaded dense point cloud of the *Piazza Duomo* dataset. The position and orientation of the extracted frames are shown in different colours to indicate the device they were acquired from (Table 1).

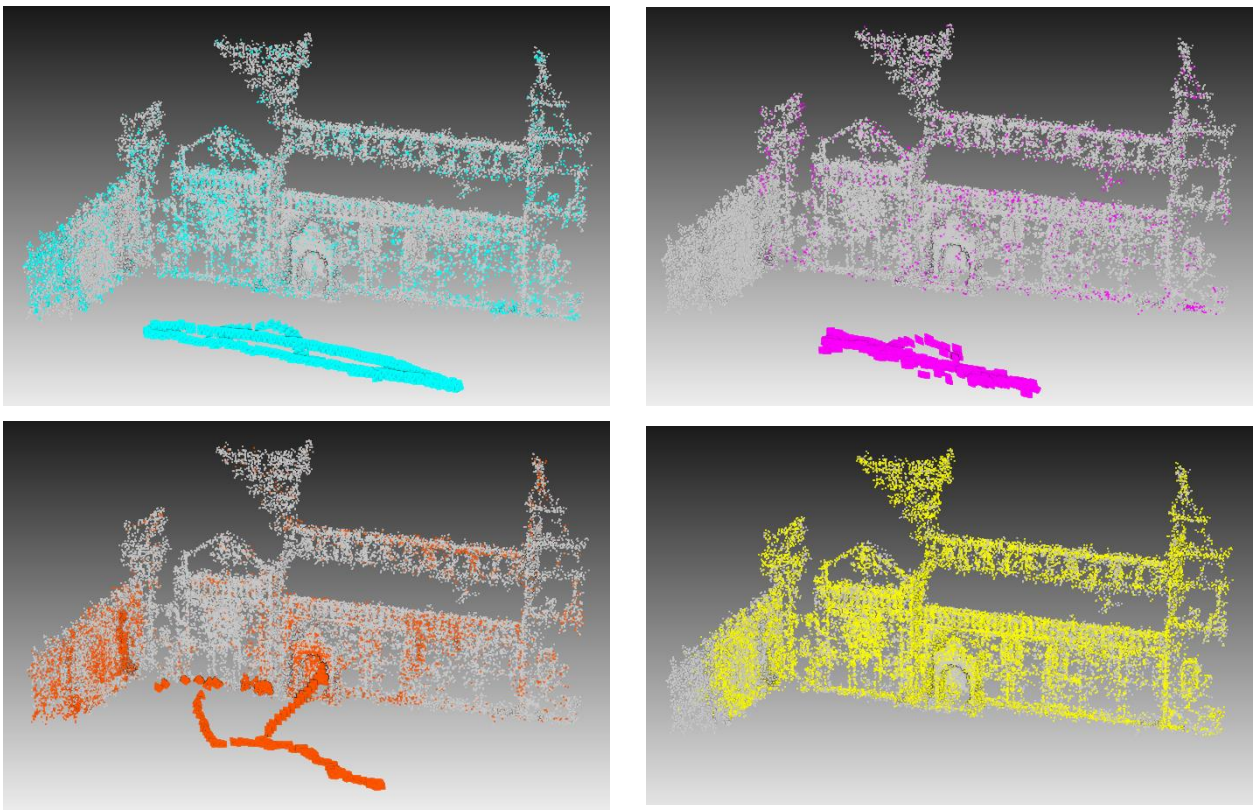

Figure 6: The sparse point cloud of the *Piazza Duomo* dataset. The points are coloured according to the smartphone they are triangulated from (Table 1); in yellow the points triangulated from images belonging to multiple devices.

4.2 Reference photogrammetric models

The reference (ground truth) datasets were acquired with a professional-grade digital single lens reflex (DSLR) camera, processed using state-of-the-art commercial software application and evaluated by computing the root mean square error (RMSE) on check points, measured through classing topographic surveying. For the *SarantaKolones* dataset, the Nikon D3X was equipped with a Nikkor 28 mm fixed focal length lens, 176 images were acquired and 20 points were used as check, providing a RMSE better than 5 mm.

The photogrammetric survey of the entire Trento cathedral square, comprising 359 images, was realised with the Nikon D3X camera coupled with two prime lenses, a Nikkor 35 mm and a Nikkor 50 mm. The RMSE on 18 check points resulted better than 10 mm.

4.3 Evaluation

To evaluate the metric potentialities of the implemented collaborative reconstruction pipeline, the dense point clouds of the three datasets acquired with different smartphones (section 4.1) are compared against the ground truth dense point clouds obtained using a standard photogrammetric procedure (section 4.2).

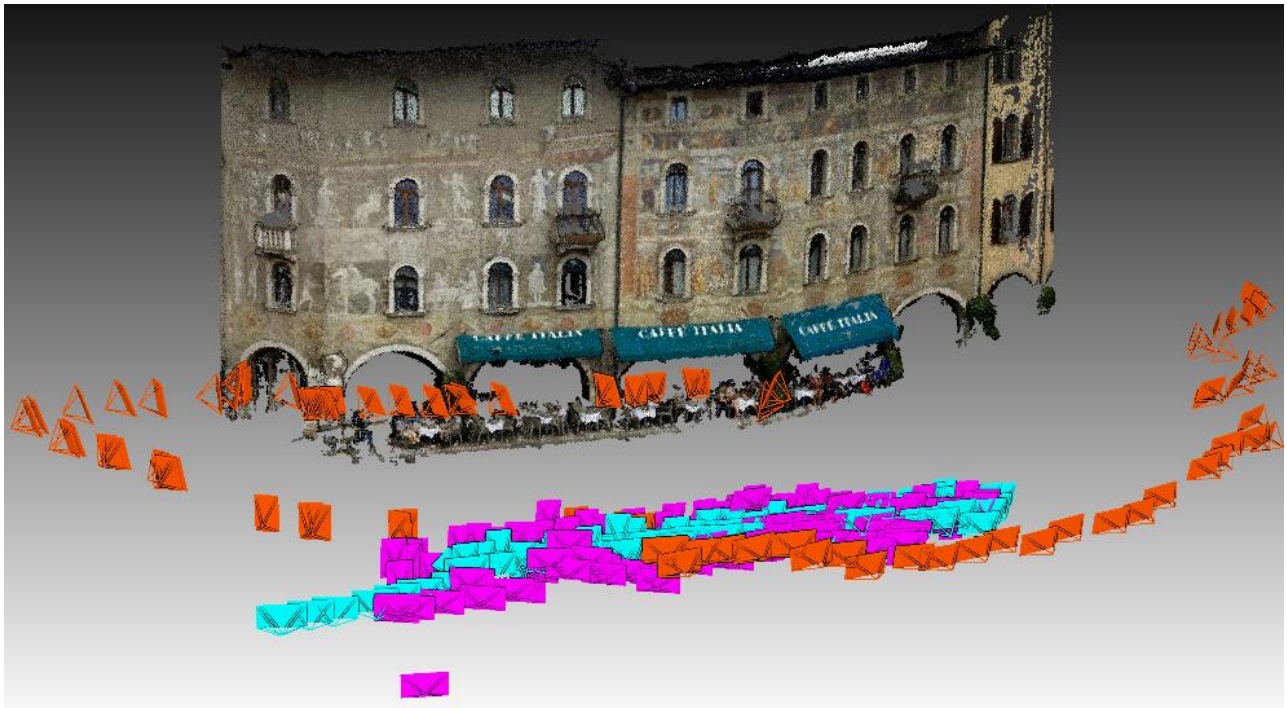

Figure 6: The RGB dense point cloud of the *Caffe Italia* dataset. The position and orientation of the extracted frames acquired with three smartphones are shown using different colours according to the employed device (Table 1).

The point clouds are first cleaned from noisy elements, then aligned in a local coordinate reference system to the reference data by means of the iterative closest point (ICP) with scale factor registration method, implemented in the open source software application CloudCompare. The signed distances between the corresponding dense point clouds are then computed, using the CloudCompare M3C2 plugin, which implements the Multiscale Model to Model Cloud Comparison method (Lague et al., 2013, Figure 7). The evaluation analyses show that the greater differences, up to 50 cm, are localised on the edges of the structures, where the poorest image quality of the camera embedded in the smartphones is most evident. However, the global geometry of the structures in all the three case studies features deviations up ten times the average point cloud resolution.

5. CONCLUSIONS

The paper presented a 3D acquisition and reconstruction pipeline where multiple users can collaboratively acquire images of a scene of interest to produce a 3D dense point cloud.

The pipeline entails an app running on smartphones that automatically selects the best frames out of a video stream and uploads them to a cloud-based server. Here the images are processed through a SfM and DIM procedures. The users can concurrently visualize the camera poses and joint 3D point cloud coming from other users / smartphones, either on the device or on a web-server page.

The proposed procedure was evaluated through comparisons with reference data produced employing a standard photogrammetric acquisition and processing workflow. The analyses showed that the achieved results may suffice for the purposes of people involved in the creative industries.

Future works will involve the implementation of Augmented Reality-based guidance for the user during image acquisition, based on device pose tracking and 3D reconstruction algorithms running on the smartphone. Moreover, a semi-automatic editing procedure to improve the dense point cloud quality is under development.

ACKNOWLEDGMENTS

The paper is part of the EU project REPLICATE (http://www.replicateproject.eu) which has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement number 687757. The authors thank Dr. Fabio Menna (3DOM-FBK) for the acquisition of the *Saranta Kolones* reference dataset during the CIPA 2017 Summer School (http://cipa.icomos.org/activities/summer_schools) and his support in the measurement and processing of check points with classic surveying techniques.

REFERENCES

Alderton, M., 2016. Digitization, Documentation, and Democratization: 3D Scanning and the Future of Museums. https://www.autodesk.com/redshift/digitization-future-of-museums/

EY, 2015. Cultural times. The first global map of cultural and creative industries. http://www.worldcreative.org/wp-content/uploads/2015/12/EYCulturalTimes2015_Download.pdf

Forster, C., Lynen, S., Kneip, L. and Scaramuzza, D., 2013, November. Collaborative monocular slam with multiple micro aerial vehicles. Proc. IEEE/RSJ Int. Conf. on *Intelligent Robots and Systems (IROS)*, pp. 3962-3970.

Hosseininaveh, A., Sargeant, B., Erfani, T., Robson, S., Shortis, M., Hess, M. and Boehm, J., 2014. Towards fully automatic reliable 3D acquisition: From designing imaging network to a

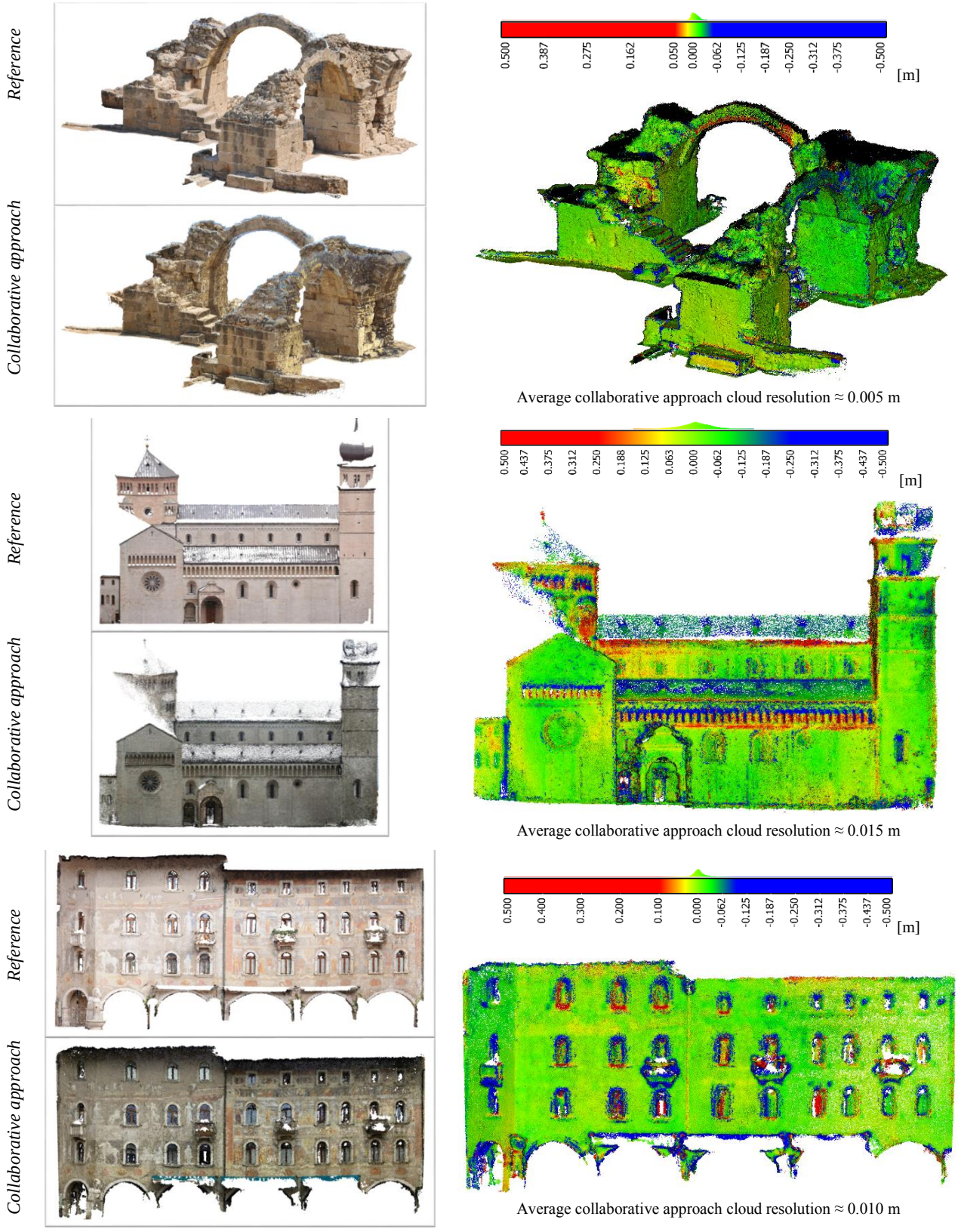

Figure 7: Quality evaluation of the proposed collaborative approach. On the left column, for each dataset the dense point clouds from the reference 3D reconstruction and the collaborative approach are shown. The right column shows the colour-coded map of the signed distances computed between the reference and the collaborative dense point clouds. The given differences are in meters.

complete and accurate point cloud. *Robotics and Autonomous Systems*, 62(8), pp.1197-1207.

Kolev, K., Tanskanen, P., Speciale, P. and Pollefeys, M., 2014. Turning mobile phones into 3D scanners. Proc. *IEEE CVPR*, pp. 3946-3953.

Lague, D., Brodu, N. and Leroux, J., 2013. Accurate 3D comparison of complex topography with terrestrial laser scanner: Application to the Rangitikei canyon (NZ). *ISPRS Journal of Photogrammetry and Remote Sensing*, 82, pp.10-26.

Locher, A., Perdoch, M., Van Gool, L., 2016a. Progressive Prioritized Multi-view Stereo. Proc. *IEEE CVPR*, pp. 3244-3252.
Locher, A., Havlena, M., Van Gool, L., 2016b. Progressive 3D Modeling All the Way. Proc. *3D Vision (3DV)*, pp. 11-18.

Locher, A., Perdoch, M., Riemenschneider, H. and Van Gool, L., 2016c. Mobile phone and cloud—A dream team for 3D reconstruction. Proc. IEEE Winter Conference on *Applications of Computer Vision (WACV)*, pp. 1-8.

Morrison, J.G., Gálvez-López, D. and Sibley, G., 2016. MOARSLAM: Multiple operator augmented RSLAM. In *Distributed Autonomous Robotic Systems*, pp. 119-132.

Nancarrow, J.H., 2016. Democratizing the Digital Collection: New Players and New Pedagogies. *Three-Dimensional Cultural Heritage. Museum Worlds*, 4(1), pp.63-77.

Nistér, D., 2004. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6), pp.756-770.

Nistér, D. and Stewenius, H., 2006. Scalable recognition with a vocabulary tree. Proc. *IEEE CVPR*, pp. 2161-2168.

Nocerino, E., Lago, F., Morabito, D., Remondino, F., Porzi, L., Poiesi, F., Bulo, S.R., Chippendale, P., Locher, A., Havlena, M. and Van Gool, L., 2017. A smartphone-based 3D pipeline for the creative industry–the REPLICATE EU project. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp.535-541.

Poiesi, F., Locher, A., Chippendale, P., Nocerino, E., Remondino, F. and Van Gool, L., 2017. Cloud-based collaborative 3D reconstruction using smartphones. Proc. *14th European Conference on Visual Media Production* (CVMP 2017), 9 pages, https://doi.org/10.1145/3150165.3150166.

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

Santos, P., Ritz, M., Fuhrmann, C. and Fellner, D., 2017. 3D mass digitization: a milestone for archaeological documentation. In: *Virtual Archaeology Review*, *8*(1).

Schmuck, P., 2017. Multi-UAV collaborative monocular SLAM. In *Robotics and Automation* (ICRA), pp. 3863-3870.

Schonberger, J.L. and Frahm, J.M., 2016. Structure-from-motion revisited. Proc. *IEEE CVPR*, pp. 4104-4113.

Sieberth, T., Wackrow, R., Chandler, J.H., 2016: Automatic detection of blurred images in UAV image sets. *Journal of Archaeological Science*, Vol. 122, pp. 1-16

Skillset, C., 2013. Classifying and measuring the creative industries. Department for Culture Media & Sport.

Sweeney, C., Sattler, T., Hollerer, T., Turk, M. and Pollefeys, M., 2015. Optimizing the viewing graph for structure-from-motion. Proc. *IEEE ICCV*, pp. 801-809.

Tanskanen, P., Kolev, K., Meier, L., Camposeco, F., Saurer, O. and Pollefeys, M., 2013. Live metric 3D reconstruction on mobile phones. Proc. *IEEE ICCV*, pp. 65-72.

The Economist, 2011. 3D printing. The printed world. http://www.economist.com/node/18114221

Untzelmann, O., Sattler, T., Middelberg, S. and Kobbelt, L., 2013. A scalable collaborative online system for city reconstruction. Proc. *IEEE ICCV*, pp. 644-651.