VIDEOGRAMMETRY VS PHOTOGRAMMETRY FOR HERITAGE 3D RECONSTRUCTION

A. Torresani [1,2], F. Remondino [1]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
[2] Università degli studi di Trento, Italy
Email: <atorresani><remondino>@fbk.eu
http://3dom.fbk.eu

Commission II, WG II/8

KEY WORDS: Visual SLAM, Key-frame selection, Structure-from-motion, 3D reconstruction.

ABSTRACT:

In the last years we are witnessing an increasing quality (and quantity) of video streams and a growing capability of SLAM-based methods to derive 3D data from videos. Video sequences can be easily acquired by non-expert surveyors and possibly used for 3D documentation purposes. The aim of the paper is to evaluate the possibility to perform 3D reconstructions of heritage scenarios using videos (“videogrammetry”), e.g. acquired with smartphones. Video frames are extracted from the sequence using a fixed-time interval and two advanced methods. Frames are then processed applying automated image orientation / Structure from Motion (SfM) and dense image matching / Multi-View Stereo (MVS) methods. Obtained 3D dense point clouds are then visually validated as well as compared with photogrammetric ground truth achieved acquiring images with a reflex camera or analysing 3D data’s noise on flat surfaces.

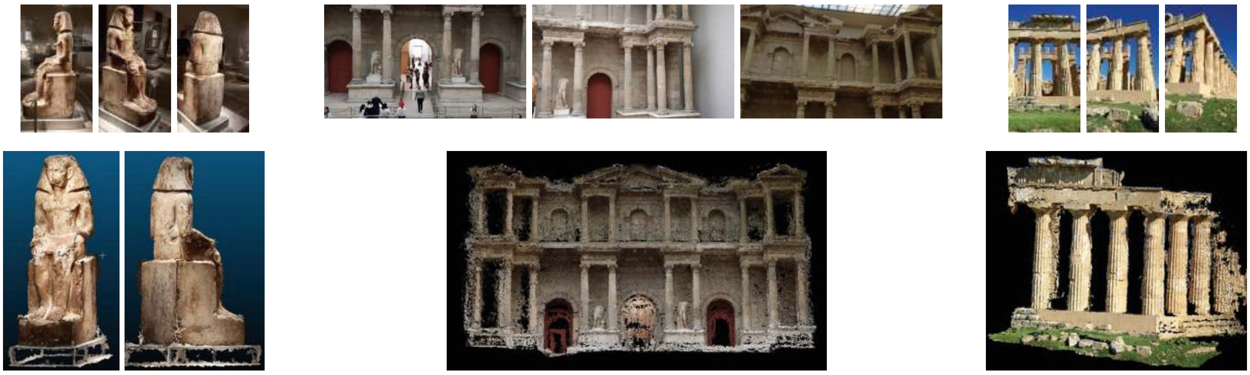


Figure 1: Heritage 3D documentation using videos from smartphone devices: examples of indoor (Egyptian museum’s statue in Torino, Italy and Ishtar Gate at the Pergamon museum in Berlin, Germany) and outdoor (Greek temple in Selinunte, Italy) scenarios.

1. INTRODUCTION

During the last 2-3 decades, a emergent set of 3D imaging sensors and tools started to be released and used for cultural heritage 3D documentation, providing an indispensable support to the (digital) preservation, archival, analysis and valorization of heritage assets (Remondino, 2011). Image-based solutions (Remondino and El-Hakim, 2006), offered by photogrammetric and computer vision methods, are among the most interesting as they allow the simultaneous retrieval of shapes and colours from high-resolution image or even from archival data (Wiedemann et al., 2000; Gruen et al., 2004; Condorelli and Rinaudo, 2018), videos (Pollefeyes et al., 2002; Sung and Lin, 2017) or smartphones (Kolev et al., 2014; Nocerino et al., 2017).

In photogrammetric 3D reconstruction tasks, the image network geometry is a crucial and tricky phase (Fraser, 1984). This is getting more and more important also in the heritage community, where the use of automated processing tools allows any user to take some randomly acquired images, blindly load them into a package, push a button and enjoy the obtained 3D model (Remondino et al., 2017). One of the key aspects for successful 3D reconstructions is the image scale, the image overlap, the viewing angle as well as the baseline between the images. It is known that too narrow baselines are not optimal for the triangulation of tie points whereas very wide baselines complicate the matching of detected keypoints. Is therefore clear that, assuming good knowledge of photography, image-based 3D reconstructions require expertise in acquiring the images. Lack of photography and photogrammetry knowledge as well as human mistakes could prevent precise and detailed 3D reconstructions. Nevertheless, the use of videos could a be an important support and step toward easier and less error-prone on-site acquisitions. This could be also favourable in case of existing video footage of lost heritage (Vincent et al., 2015). The use of videos for scenes 3D reconstruction purposes is nowadays seeing a revival after the initial breakthrough experiments of some 15 years ago (Sato et al., 2002; Pollefeyes et al., 2007). Indeed, there is an increasing radiometric and geometric quality of video footages and, moreover, users can exploit the ever-growing capabilities of Simultaneous Localisation and Mapping (SLAM) methods to process video streams and derive 3D information (Taketomi et al., 2017).

Videogrammetry, i.e. the processing of video streams for retrieving metric 3D information, was originally used for industry-based applications, such as motion capture, crash tests analyses, biomechanics, mobile mapping, etc. (Gruen, 1997). Few interest points (normally coded targets) in the images were tracked and matched in order to triangulate them and derive sparse 3D point clouds. Nowadays, thanks to the progresses in hardware (CPU/GPU) performances, recent developments in the robotics community and the availability of fast and reliable Visual-SLAM methods (Mur-Artal et al., 2015), videos acquired

Name	Type	Features	Point cloud	Global opt	Opt algorithm	Loop closure	Reference
MonoSLAM	Feature-based	FAST	Sparse	No	Kalman filter	No	(Davison et al., 2007)
PTAM	Feature-based	FAST	Sparse	Yes	Bundle Adj.	No	(Klein et al., 2009)
ORB-SLAM	Feature-based	ORB	Sparse	Yes	Bundle Adj.	Yes	(Mur-Artal et al., 2015)
LSD-SLAM	Direct	-	Semi-dense	Yes	Pose graph opt.	Yes	(Engel et al., 2014)

Table 1: Brief summary of some V-SLAM algorithms.

with off-the-shelf cameras or even smartphones could also be used for dense 3D reconstruction purposes. Actually, V-SLAM and the image orientation (also called Structure from Motion – SfM) step are facing very similar problems: deriving a set of 3D point starting from image correspondences by means of triangulation. The main difference is that V-SLAM is designed to work in real-time on video streams, i.e. it expects very short baselines and delivers camera poses and sparse 3D points as soon as the video frames are processed. On the other hand, traditional image orientation algorithms and SfM solutions require a suitable image network with good baselines, providing more precise 3D results but being more computationally expensive.

Therefore, to fully exploit videos for 3D reconstruction purposes, applying automated image orientation / SfM (Ozyesil et al., 2017) and dense image matching / Multi-View Stereo (MVS - Remondino et al., 2014; Furukawa and Hernandez, 2015) methods, some frames should be carefully selected (the so called *keyframes*) before running the 3D reconstruction pipeline.

1.1 Aim of the paper

The aim of the work is to evaluate the potentials and limitations of video-based 3D reconstructions of heritage scenarios (Fig. 1). More specifically, the paper reports how keyframes extracted from video sequences could be exploited for dense 3D heritage reconstruction. We evaluate the following keyframe selection methods (Section 3):

- Keyframe selection at fixed-time intervals;
- 2D-feature-based approach implemented;
- 3D-based approach of ORB-SLAM (Mur-Artal et al., 2015).

The selected keyframes are then used within a 3D reconstruction pipeline to generate 3D dense point clouds (Section 4). Three heritage datasets are considered for the experiments (Section 4.1): a façade of Trento's cathedral (Italy), the gate with lion of the Trento's cathedral (Italy) and the Arches Castle in Paphos, (Cyprus). A geometric evaluation of the derived 3D dense point clouds is also carried out (Section 4.2) using plan fitting error evaluation and ground truth data produced with reflex camera-based photogrammetric survey.

2. SIMULTANEOUS LOCALISATION AND MAPPING

Simultaneous Localisation and Mapping (SLAM) algorithms are used to simultaneously retrieve, in real time, the 3D structure of the environment and the positions (trajectory) of the imaging or scanning sensor. When the device is a camera, SLAM is called Visual SLAM (V-SLAM) and the formulation of the problem is very similar to the photogrammetric pipeline where different images are used to reconstruct a surveyed scene in 3D. The main difference is that V-SLAM is designed to work on densely sampled frames of a video stream and perform operations in real time. V-SLAM algorithms are divided in two main families:

- feature-based algorithms (Davison et al., 2007; Mur-Artal et al., 2015): they employ detector/descriptors algorithms (e.g. ORB - Rublee et al., 2011; FAST – Rosten and Drummond, 2006; etc.) to quickly find and track image correspondences, sequentially bundle images minimizing the reprojection error and, eventually, refine the entire trajectory.
- direct algorithms (Engel et al., 2014): they work directly on pixels intensities by keeping a depth map estimation for high gradient pixels and estimating the camera positions through the minimisation of the photometric error.

Table 1 summaries briefly the current V-SLAM panorama.

3. KEYFRAME SELECTION APPROACHES

The keyframe selection phase is a mandatory prerequisite to derive 3D data from videos, either with V-SLAM methods or using a more complex image orientation / SfM approach.

In the literature different techniques have been proposed to select keyframes from videos. They can be categorized in trivial (random way or at fixed-time intervals frame extraction), clustering-based (some global image features, such as color histograms, is chosen and then frames lying close to the centroids of the clusters are selected – Girgensohn and Boreczky, 2000), 2D-feature-based approaches (Guan et al., 2013; Nocerino et al., 2017) and 3D-based methods (Resch et al., 2015). In the following sections, these last two methods, considered the most valued, are described in detail.

3.1 2D-feature-based selection

2D-feature-based approaches (Guan et al., 2013; Nocerino et al., 2017) use image keypoints, such as SIFT (Lowe, 2004) or ORB, to measure the newness of each frame and decide whether to select it or not. The newness is measured by comparing the 2D features of the current frame with a pool of features extracted from the previous frames. Given a sequence of video frames, the i-th frame f_i is selected to be a keyframe if the following properties hold:

- f_i is sharp (not blurred);
- the percentage of keypoints in f_i that matches with a pool of keypoints in the previous frame f_{i-1} is below a certain threshold.

This approach ensures that blurred and redundant frames, having probably small baselines between each other, are discarded because of the high percentage of matched keypoints. The main limitation of this approach is that it is not exploiting the current 3D knowledge of the scene.

3.2 3D-based selection

3D-based methods (Resch et al., 2015) take advantage of the current geometric properties of the reconstructed scene, such as the geometry of the point cloud and the viewing angles of the video frames, to decide when to select a new keyframe out of the analysed video. The main advantage of those methods is that they exploit the geometric knowledge of the scene e.g. to control the baseline of the selected frames. Modern V-SLAM algorithms, such as ORB-SLAM (Mur-Artal et al., 2015), lie in this category as they geometrically select the subset of frames on which the scene 3D geometry is computed. ORB-SLAM was designed to have quite relaxed keyframe selection policies and stronger checks during the keyframe culling phase. In this way the point cloud is rapidly updated with the selected keyframes but their geometric relevance as well as their redundancy is rigorously checked during the culling phase.

Dataset name	**Trento's cathedral façade – Trento-1**	**Trento's cathedral lion – Trento-2**	**Arches Castle – Cyprus**
Example of frames			
Video Resolution	1920 x 1080 px	1920 x 1080 px	1920 x 1080 px
Smartphone	Samsung S9 plus	Samsung S9 plus	Samsung S6
Validation	RMSE on plane fitting	Visual inspection of derived 3D dense point cloud	Cloud2Cloud distance and profiles from a photogrammetric point cloud generated from 175 Nikon D3X images (6048 x 4032 px);
Video length	~ 2 min: 59 sec	~ 2 min: 45 sec	~ 4 min: 20 sec
# frames	5374	4953	7826

Table 2: Main characteristics of the datasets used for the paper experiments.

Given a frame f, the algorithm selects f as a keyframe if the number of frames since the last selected keyframe is greater than a certain threshold, and if the number of keypoints of f that match with the keypoints of the last selected keyframe is less than 90%. A keyframe k is removed if at least 90% of the keypoints of k are seen by at least other two previously selected keyframes. In this way ORB-SLAM keeps only informative keyframes and, at the same time, it bounds the time required to perform all the background optimisations.

4. METHODOLOGY AND RESULTS

In order to evaluate the potential and usability of videos acquired with consumer devices (e.g. smartphones) for the 3D documentation of heritage scenarios, various video sequences were acquired, trying to image the entire scene from multiple views. Then, frames were selected/extracted using the aforementioned selection procedures (Fig. 2-3): Fixed-time interval (FTI), 2D-feature-based (2D) and 3D-based method (3D).

```
keyframes = []

PROCEDURE TimeSelection(video, time)
   FOR frame in video
      IF time(frame) mod time == 0
          keyframes.add(frame)

PROCEDURE 2DSelection(video, newnessTh)
   poolOfDesc = []
   FOR frame in video
      IF isBlurred(frame) CONTINUE
      frameDesc = computeORB(frame)
      IF match(poolOfDesc, frameDesc) > newnessTh
          keyframes.add(frame)
      poolOfDesc.add(frameDesc)

PROCEDURE 3DSelection (video)
   selectedKeyframes, pointCloud = ORBSLAM(video)
   FOR keyframe in selectedKeyframes
      keyframes.add(keyframe)
```

Figure 2: Pseudo-code of the three considered keyframe selection methods.

Finally, the extracted keyframes were processed using the COLMAP (Schönberger et al., 2016a-b) workflow (version 3.5) in order to derive 3D dense point clouds. A self-calibration was performed to retrieve both interior and exterior camera parameters while dense 3D reconstructions were achieved with the MVS patch-based approach.

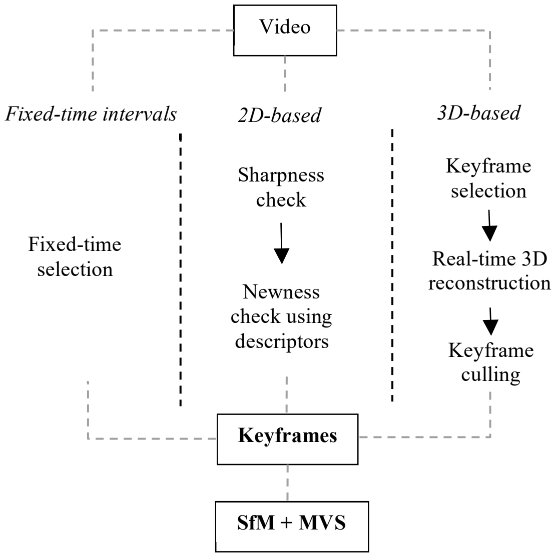


Figure 3. The main steps of the tested selection methods which extract keyframes from the given video sequence in order to process them into a photogrammetric pipeline.

4.1 Datasets

Three datasets (Table 2) are used for the evaluation: two of the Cathedral in Trento (Italy) and one of the Arches Castle in Paphos (Cyprus). Table 3 shows some information of the extracted/selected frames per dataset, based on the selection method.

Dataset	**# frames**
Trento-1_FTI	180
Trento-1_2D	409
Trento-1_3D	189
Trento-2_FTI	165
Trento-2_2D	311
Trento-2_3D	641
Cyprus_FTI	262
Cyprus_2D	700
Cyprus_3D	149

Table 3. Number of extracted keyframes for each dataset and applied method (FTI: fixed-time interval, 1 frame every 30 frames; 2D: 2D-feature-based; 3D: 3D-based).

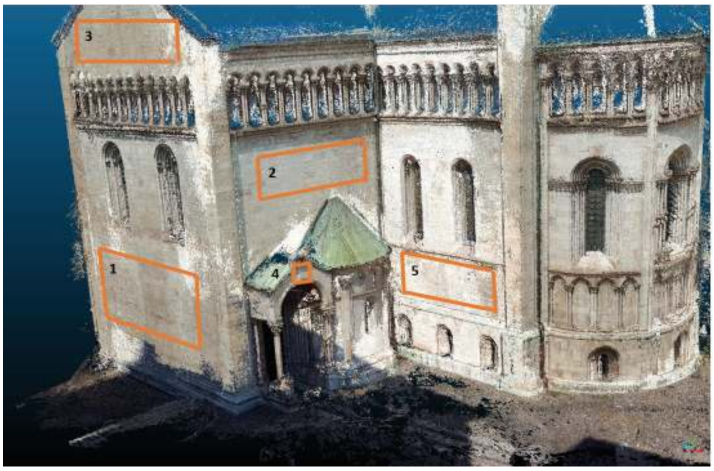


Figure 4. Locations of the five areas used to calculate the plane fitting RMSE useful to evaluate geometric performances of the video-based 3D reconstruction.

Areas	KS	RMSE (m)
1	FTI	**0.007**
	2D	0.01
	3D	0.037
2	FTI	**0.013**
	2D	0.025
	3D	0.016
3	FTI	0.026
	2D	0.035
	3D	**0.025**
4	FTI	0.008
	2D	0.017
	3D	**0.008**
5	FTI	0.053
	2D	**0.033**
	3D	0.044

Table 4: RMSEs for the 5 planar areas of Figure 5 depending on the keyframe selection method (KS).

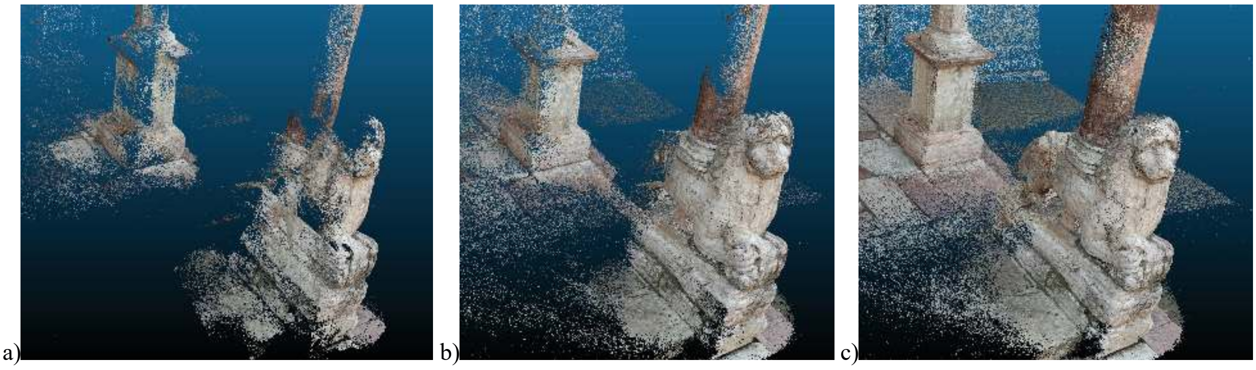


Figure 5: Qualitative comparison of the videogrammetric point clouds of the Trento-2 dataset: fixed-time selection (a), 2D-feature-based selection (b) and 3D-based selection (c). The latter approach is delivering the most complete and dense 3D reconstruction of the imaged scene.

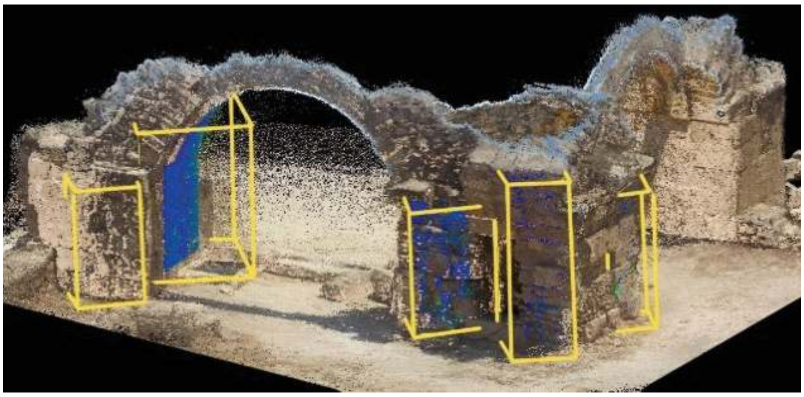


Figure 6: Location of the five areas (from left to right: 1,2,3,4 and 5) used to perform the cloud-to-cloud distances between the videogrammetry-based and ground truth point clouds (Table 5).

Area	KS	C2C Mean (m)	C2C std. dev. (m)
1	FTI	0.015	0.008
	2D	0.011	0.009
	3D	**0.01**	**0.005**
2	FTI	**0.025**	**0.029**
	2D	0.044	0.042
	3D	0.028	0.064
3	FTI	0.134	0.033
	2D	0.091	0.030
	3D	**0.008**	**0.006**
4	FTI	0.015	0.010
	2D	0.021	0.022
	3D	**0.014**	**0.009**
5	FTI	0.070	0.039
	2D	0.143	0.070
	3D	**0.16**	**0.014**

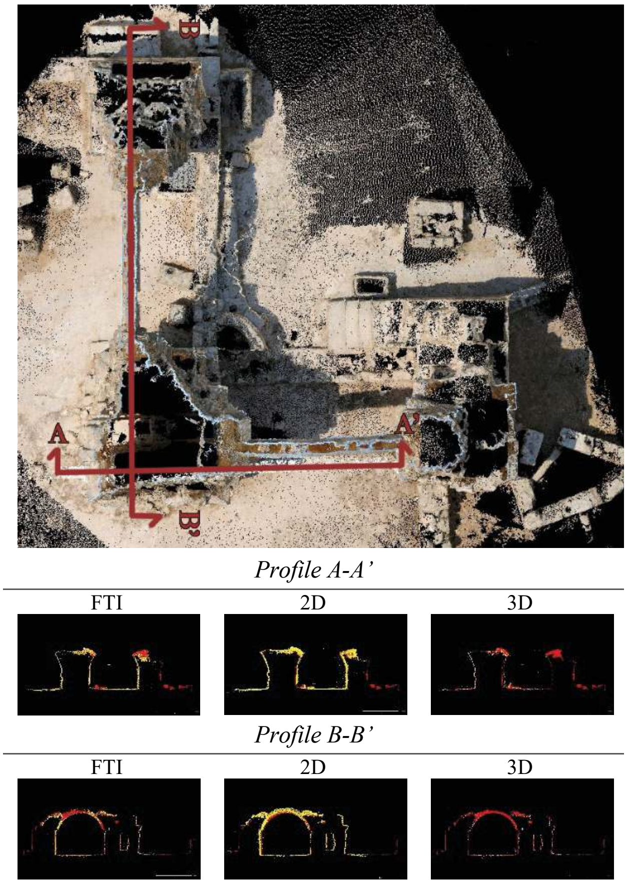
Table 5: Cloud to cloud distances between the photogrammetric cloud of the Arches Temple (ground truth) and the videogrammetric clouds computed with the different keyframe selection methods (KS).

4.2 Results

Trento's Cathedral façade (Trento-1). Table 4 reports the RMSE distances on five selected planes of the cathedral façade which was geometrically reconstructed with the videogrammetry approach (Fig. 5). The 3D-based frame selection method is providing, in general, better results, although not outperforming the other selection methods.

Trento's Cathedral lion. The dataset features a quite complex scenario, with an entrance gate, columns, statues and several architectural elements. The derived point clouds (Fig. 5) clearly highlight how fixed-time interval and 2D-feature-based frame extraction are not suitable to provide a sufficient and correct number of frames to completely reconstruct the scene.

Arches Castle. Dense point cloud derived from every frame selection method were aligned using ICP with the available ground truth (photogrammetric dense point cloud, computed at ½ image resolution). Then Cloud-to-cloud distances were computed (Table 5), in five specific areas (Fig. 6). Also in this case, a 3D-based frame selection approach provides for the most accurate results. Furthermore, Figure 7 shows a graphical comparison of two profiles on the estimated points clouds. The keyframes selected with the 3D-based approach provide for the closest point cloud to the photogrammetric ground truth.



Profile A-A'

FTI 2D 3D

Profile B-B'

FTI 2D 3D

Figure 7. Visual comparison of profiles (A-A' and B-B') extracted on the videogrammetry (red) and ground truth (yellow) point clouds for the employed frame selection methods.

4.3 Discussion

Experiments and results show that in all considered datasets the photogrammetric reconstruction is greatly helped by the geometric keyframe selection of the V-SLAM approach. Regarding the Cyprus dataset, both cloud-to-cloud distance analysis (Table 5) and profile analysis (Figure 7) show that 3D documentation performed with frames extracted with the 3D-based method can almost match the quality of point clouds created using images acquired with reflex cameras. Moreover, the efficacy of the 3D-based approach in selecting good reconstruction keyframes is clearly visible in Figure 5: the 3D-based dense point cloud is far more complete than the point clouds obtained by the other two frame selection methods. It is evident how the time-based selection can be completely inefficient when the camera moves around complex objects as discontinuities changes are not considered during the selection procedure. On the other hand, the RMSE analyses of the planes (Table 4) does not show a clear winner: it must be noticed that in this case the camera movement and the distance to the cathedral was constant during the acquisition. Results show that in these circumstances the three methods almost present the same performances. Finally, another strength of the 3D-based selection is that it does not require any prior knowledge of the scene to perform the tuning of the parameters: the inner structure of the V-SLAM algorithm is designed to work in many different scenarios (fast or slow camera movements, different distances to the object to reconstruct, etc.) without the need of using scenario-related settings.

5. CONCLUSIONS

The paper presented how videogrammetry, i.e. the use of videos for 3D reconstruction purposes, can be a valuable source for 3D documentation in the heritage community. Video frames can be extracted with various approaches, such as: time-based selection, 2D-feature-based selection and 3D-based selection (e.g. based on the ORB-SLAM algorithm). Experiments show that a videogrammetric approach can deliver comparable 3D results to a photogrammetric solution based on a reflex camera images, in particular when the reconstruction frames are selected considering the surveyed geometry.

3D documentation of heritage scenarios with high-end high-resolution digital cameras will be never surpass. Nevertheless video-based 3D reconstruction, if coupled with advanced frame selection algorithms and a precise 3D processing methodology, could be a valuable alternative, being easier and less prone to errors, especially when operators are lacking photogrammetric knowledge required to acquire images with good reconstruction and geometric properties.

REFERENCES

Condorelli., F., Rinaudo, F., 2018. Cultural Heritage reconstruction from historical images and videos. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. 42(2), pp. 259-265.

Davison, A.J., Reid, I.D., Molton, N.D. and Stasse, O., 2007. MonoSLAM: Real-time single camera SLAM. IEEE Transactions on Pattern Analysis & Machine Intelligence, (6), pp.1052-1067.

Engel, J., Schöps, T. and Cremers, D., 2014, September. LSD-SLAM: Large-scale direct monocular SLAM. In European conference on computer vision (pp. 834-849). Springer, Cham.

Fraser, C., 1984. Network design consideration for non-topographic photogrammetry. Photogrammetric Engeneering and Remote Sensing, Vol. 50(8), pp. 1115-1126.

Furukawa, Y. and Hernandez, C., 2015. Multi-view stereo: a tutorial. Foundations and Trends in Computer Graphics and Vision, Vol. 9(1-2).

Girgensohn, A. and Boreczky, J., 2000. Time-constrained keyframe selection technique. Multimedia Tools and Applications, 11(3), pp.347-358.

Gruen, A., 1997. Fundamental of videogrammetry – a review. Human Movement Science, Vol. 16(2-3), pp. 155-187.

Gruen, A., Remondino, F., Zhang, L., 2004. Photogrammetric Reconstruction of the Great Buddha of Bamiyan, Afghanistan. The Photogrammetric Record, Vol.19(107), pp. 177-199.

Guan, G., Wang, Z., Lu, S., Da Deng, J. and Feng, D.D., 2013. Keypoint-based keyframe selection. IEEE Transactions on circuits and systems for video technology, 23(4), pp.729-734.

Klein, G. and Murray, D., 2009, October. Parallel tracking and mapping on a camera phone. In 2009 8th IEEE International Symposium on Mixed and Augmented Reality, pp. 83-86.

Kolev, K., Tanskanen, P., Speciale, P., Pollefeys, M., 2014. Turning Mobile Phones into 3D Scanners. Proc. CVPR.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, Vol. 60(2), pp.91-110.

Mur-Artal, R., Montiel J.M.M., Tardos, J.D., 2015. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Trans. on Robotics, vol. 31, no. 5, pp. 1147-1163.

Nocerino, E., Lago, F., Morabito, D., Remondino, F., Porzi, L., Poiesi, F., Rota Bulo, S., Chippendale, P., Locher, A., Havlena, M., Van Gool, L., Eder, M., Fötschl, A., Hilsmann, A., Kausch, L., Eisert, P., 2017. A smartphone-based pipeline for the creative industry - The REPLICATE project. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLII-2-W3, pp. 535-541.

Ozyesil, O., Voroninski, V., Basri, R., Singer, A., 2017. A Survey of Structure from Motion. Acta Numerica, Vol. 26, pp. 305-364.

Pollefeys, M., van Gool, L., Vergauwen, M., Cornelis, K., Verbiest, F., Tops, J., 2002. Video-to-3D. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol.34(B3), pp. 252-257.

Pollefeys, M., Nister, D., Frahm, J.-M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S.-J., Merrell, P., SalmiC., Sinha, S., Talton, B., Wang, L., Yang, Q., 2007. Detailed real-time urban 3D reconstruction from video. Int. Journal of Computer Vision, Vol. 78(1-2), pp. 143-167.

Remondino, F., Nocerino, E., Toschi, I., Menna, F., 2017. A critical review of automated photogrammetric processing of large datasets. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLII-2/W5, pp. 591-599.

Resch, B., Lensch, H., Wang, O., Pollefeys, M. and Sorkine-Hornung, A., 2015. Scalable structure from motion for densely sampled videos. Proc. CVPR, pp. 3936-3944.

Remondino, F., El-Hakim, S., 2006. Image-based 3D modelling: a review. The Photogrammetric Record, Vol.21(115), pp. 269-291.

Remondino, F., 2011: Heritage recording and 3D modeling with photogrammetry and 3D scanning. Remote Sensing, 3(6), pp. 1104-1138.

Remondino, F., Spera, M.G., Nocerino, E., Menna, F., Nex, F., 2014. State of the art in high density image matching. The Photogrammetric Record, Vol. 29(146), pp. 144-166.

Remondino, F., Nocerino, E., Toschi, I., Menna, F., 2017. A critical review of automated photogrammetric processing of large datasets. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLII-2/W5, pp. 591-599.

Rosten, E. and Drummond, T., 2006. Machine learning for high-speed corner detection. Proc. ECCV, pp. 430-443, Springer, Berlin, Heidelberg.

Rublee, E., Rabaud, V., Konolige, K. and Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. Proc. ICCV.

Sato, T., Kanbara, M., Yokoya, N., Takemura, H., 2002. Dense 3-D reconstruction of an outdoor scene by hundreds-baseline stereo using a hand-held video camera. Int. Journal of Computer Vision, Vol. 47(1-3), pp. 119-129.

Schönberger, J.L. and Frahm, J.M., 2016a. Structure-from-motion revisited. Proc. CVPR, pp. 4104-4113.

Schönberger, J.L., Zheng, E., Frahm, J.M. and Pollefeys, M., 2016b. Pixelwise view selection for unstructured multi-view stereo. Proc. ECCV, pp. 501-518.

Sung, B.-Y., and Lin, C.-H., 2017. A fast 3D scene reconstructing method using continuous video. EURASIP Journal on Image and Video Processing, 18.

Taketomi, T., Uchiyama, H., Ikeda, S., 2017. Visual SLAM algorithms: a survey from 2010 to 2016. Transactions on Computer Vision and Applications, Vol. 9(16).

Vincent, M. L., Coughenour, C., Remondino, F., Flores Gutierrez, M., Lopez-Menchero Bendicho, V. M., Frtisch, D., 2015: Crowd-sourcing the 3D digital reconstructions of lost cultural heritage. Proc. IEEE Digital Heritage, Vol. 1, pp. 171-172.

Wiedemann, A., Hemmleb, M., Albertz, J., 2000. Reconstruction of historical buildings based on images from Meydenbauer archives. In: International Archives of Photogrammetry and Remote Sensing. Vol. 33(B5).