IMAGE ORIENTATION WITH A HYBRID PIPELINE ROBUST TO ROTATIONS AND WIDE-BASELINES

F. Bellavia[a], L. Morelli[b,c], F. Menna[b], F. Remondino[b]

[a]Dept. of Mathematics and Computer Science (DMI), University of Palermo, Italy
Web: https://sites.google.com/view/fbellavia – Email: bellavia.fabio@gmail.com

[b]3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Web: http://3dom.fbk.eu – Email: {lmorelli,fmenna,remondino}@fbk.eu

[c]Dept. of Civil, Environmental and Mechanical Engineering (DICAM), University of Trento, Italy

Commission II

KEY WORDS: keypoints, detectors, descriptors, deep learning, image matching, Structure-from-Motion, photogrammetry.

ABSTRACT:

The extraction of reliable and repeatable interest points among images is a fundamental step for automatic image orientation (Structure-From-Motion). Despite recent progresses, open issues in challenging conditions - such as wide baselines and strong light variations - are still present. Over the years, traditional hand-crafted methods have been paired by learning-based approaches, progressively updating the state-of-the-art according to recent benchmarks. Notwithstanding these advancements, learning-based methods are often not suitable for real photogrammetric surveys due to their lack of rotation invariance, a fundamental requirement for these specific applications. This paper proposes a novel hybrid image matching pipeline which employs both hand-crafted and deep-based components, to extract reliable rotational invariant keypoints optimized for wide-baseline scenarios. The proposed hybrid pipeline was compared with other hand-crafted and learning-based state-of-the-art approaches on some photogrammetric datasets using metric ground-truth data. Results show that the proposed hybrid matching pipeline has high accuracy and appeared to be the only method among the evaluated ones able to register images in the most challenging wide-baseline scenarios.

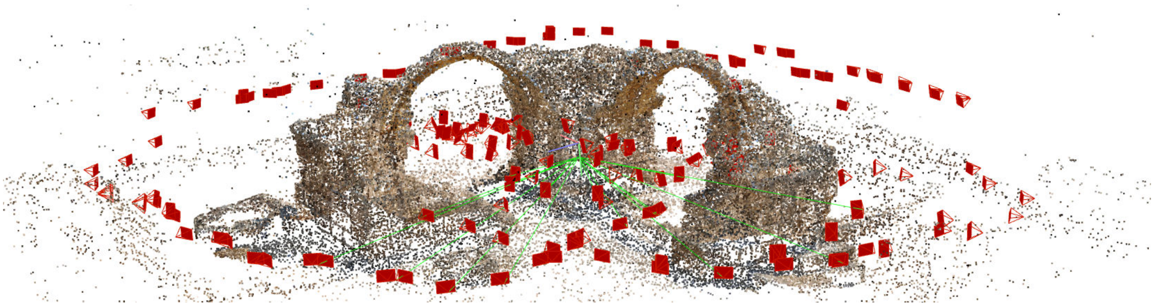


Figure 1. A typical large image network featuring convergent and rotated images as well as scale changes to survey the ancient arches at Saranta Kolones (Cyprus). The recovered camera network was obtained with the automated image orientation procedure presented in this paper, which combines hand-crafted and learning-based image matching methods and outperforms state-of-the-art solutions considering metrics in object space.

1. INTRODUCTION

Photogrammetry has become a valuable, powerful, automated and cheap alternative to active sensors for the generation of textured 3D models (Remondino et al., 2017). The typical photogrammetric workflow consists of the identification of image correspondences via sparse image matching, the estimation of unknown camera parameters and 3D object coordinates with a Bundle Adjustment (BA) method (normally called Structure-from-Motion - SfM), and dense image matching (or Multi-View Stereo - MVS) for the generation of dense point clouds. Sparse image matching, traditionally based on hand-crafted keypoint detectors and descriptors (Lowe, 2004; Bay et al., 2008; Alcantarilla et al., 2013), are based on a priori knowledge inspired by professional knowledge and intuitive experience (Yao et al., 2021). More recently, encouraged by deep-learning advancements, novel deep-based solutions have been proposed aiming to overcome the limitations of current hand-crafted methods in case of wide-baseline images or strong illumination changes (Verdie et al., 2015; Yao et al., 2021; Jin et al., 2020). While the first attempts in this research direction focused on the different steps of the image matching pipeline separately, more recent solutions provide end-to-end deep networks that jointly optimize the whole pipeline steps: LIFT (Yi et al., 2016), LF-Net (Ono et al., 2018), SuperPoint (DeTone et al., 2018), R2D2 (Revaud et al., 2019), D2-Net (Dusmanu et al., 2019), ASLFeat (Luo et al., 2020), etc. This last design choice increases both the keypoint repeatability and reliability and, consequently, the image matching success rate, proving beneficial for the final pose estimation accuracy. Nevertheless, current end-to-end deep architectures can be not suitable for general-purpose photogrammetric applications due to their limitation in handling large image rotations (Remondino et al., 2021). This specific design choice is made to maximize the discriminative ability of the matching process in more common general-user application scenarios with all images roughly up-

https://doi.org/10.5194/isprs-archives-XLVI-2-W1-2022-73-2022

right (Pautrat et al., 2020).

In the analysis and evaluation of different image matching pipelines for SfM applications, both Schönberger et al. (2017) and Jin et al. (2020) highlighted the importance of performing the evaluation on real scenarios with challenging conditions, including uncalibrated cameras, unordered images acquired with different sensors, strong light variations and viewpoint changes. Commonly adopted evaluation criteria rely on SfM output statistics, such as the mean reprojection error and the mean track length (Schönberger et al., 2017), or pseudo ground-truth (Jin et al., 2020) obtained from a superset of the input images employed for the evaluation. Nevertheless, Remondino et al. (2021) showed that the above evaluation criteria generally disagree with accurate metric ground-truth data provided by topographic surveys in terms of Ground Control Points/Check Points (GCPs/CPs). More recently, the SimLocMatch challenge of IMW2021 employed synthetic rendered scenes in order to have available ground-truth data known by construction. In any case, this last solution is not completely satisfactory as synthetic scenes are generally unable to fully simulate real world scenarios.

1.1 Aim of the paper

In order to take advantage of the recent deep-solutions while maintaining the rotation invariance property, in this work we build upon the recent hybrid image matching pipeline by Bellavia and Mishkin (2021) - which resulted among the best in the recent Image Matching Workshop (IMW2021) contest[1] - and present a complete Hybrid Pipeline (HP) suitable for photogrammetric applications (see Fig. 1 as an example). The pipeline is based on the classic detect-then-describe approach and integrates both hand-crafted and deep-based state-of-the-art methods. Moreover, a novel module, named Keypoint Filtering by Coverage (KFC), is designed, added and evaluated in order to improve the final BA accuracy. The proposed method is detailed in Sec. 2, while its performance is tested and compared in Sec. 3 using several state-of-the-art feature extractors, both hand-crafted and learning-based. All the evaluations are performed downstreaming the SfM pipeline and the BA of COLMAP (Schönberger and Frahm, 2016). Unlike recent comparative evaluations (Jin et al., 2020), our evaluation uses a set of well distributed Check Points (CPs) in order to provide reliable quantitative measures in object space (Remondino et al., 2021).

2. THE PROPOSED HYBRID PIPELINE (HP)

The proposed image matching Hybrid Pipeline (HP) is composed of the following modules: HarrisZ$^+$ (Bellavia and Mishkin, 2021) for the keypoint extraction; OriNet and AffNet (Mishkin et al., 2018) for patch normalization; HardNet8 (Pultar, 2020) as keypoint descriptor; blob matching and Delaunay Triangulation Matching (DTM - Bellavia (2021)) for descriptor matching and local spatial filtering; Degenerate SAmple Consensus (DegenSAC - Chum et al. (2005)) for model-based final correspondence assignment and outlier rejection; the Keypoint Filtering by Coverage (KFC) module to exclude image pairs from the BA according to their keypoint coverage. With the exception of OriNet, AffNet and HardNet8, which are deep-based, the other steps are hand-crafted. At the moment, HP is implemented in Matlab with the exception of the deep-based components and DegenSAC which are available through the Kornia library (Riba et al., 2020). The code is freely available[2].

HarrisZ$^+$ is an update of the HarrisZ (Bellavia et al., 2011) corner detector optimized and tuned to take advantage of the recent progress in the other image matching pipeline steps. The original HarrisZ makes use of a sort of “attention mask” to enhance the input image derivatives in order to both discard non-relevant image regions as well as to improve the adaptive filter response to corners. With respect to HarrisZ, HarrisZ$^+$ outputs more keypoints, better localized and distributed over the image, improving keypoint repeatability while maintaining a high level of discriminability.

OriNet and AffNet are two state-of-the-art deep networks respectively employed for patch orientation estimation and affine-shape adaptation. Together, they perform the patch normalization which is required to prepare the keypoint patch to the descriptor extraction.

HardNet8 is the latest version of the state-of-the-art learning-based HardNet descriptor (Mishchuk et al., 2017). HardNet8 introduces several changes in the training process and datasets as well in the network architecture that make it able to surpass the original HardNet version.

Blob matching and DTM are employed respectively for selecting matches on the basis of the descriptor similarity and local spatial information. The former extends the Nearest Neighbor Ratio (NNR - Lowe (2004)) selection to be symmetric and to include many-to-many matches, while the latter introduces an iterative correspondence pruning according to the local keypoint neighborhoods, also avoiding to set user-based threshold in NNR.

DegenSAC is an extension of the Random SAmple Consensus (RANSAC - Fischler and Bolles (1981)) which applies model-constraint match filtering according to epipolar geometry. With respect to RANSAC, DegenSAC better handles the presence of dominant planes which leads to configurations close to the degenerate ones.

KFC is designed to exclude from the BA correspondences belonging to image pairs where their keypoint patches covers less than 35%, unless their removal breaks the image connection graph. KFC arises from the observation that it is reasonably expected that a pipeline more robust to perspective distortions, due to convergent images, is also able to retain matches less accurate in terms of keypoint localization, hence decreasing the final BA accuracy. Similar ideas have been already used in the literature, such as the keypoint convex hull area at the base of the hierarchical BA approach of Toldo et al. (2015). More in detail, given the n DegenSAC keypoint correspondences (k_{iz}, k_{jz}) with $z = 1, \cdots, n$ between the images I_i and I_j, an overlap mask M_{ij} for I_i is computed by marking the 31×31 px square blocks centered at each k_{iz} on I_i and likewise a mask M_{ji} for I_j is computed. The square 31×31 patch is chosen as an approximation of the 35×35 circular descriptor patch at the minimum scale associated with the HarrisZ$^+$ keypoint. The image overlap graph $G = (I, E)$ is defined so that each image I_i represents a node and there is an edge $E_{ij} = \min(M_{ij}, M_{ji})/s$ between nodes I_i and I_j, where s is image size in px, only if $M_{ij} \neq 0$ and $i \neq j$. The Minimum Spanning Tree (MST) of the complementary graph $G' = (I, E')$ with edges $E'_{ij} = \max(E) - E_{ij}$ is then computed and edges in $MST(G')$ are removed to get a further graph G''. The final image pairs involved in the BA are only those corresponding to the nodes linked by edges in $\{E_{ij} > t_{ov}\} \cup MST(G') \cup MST(G'')$, with the image overlap threshold t_{ov} experimentally set to 35%.

[1] https://www.cs.ubc.ca/research/image-matching-challenge/current/

[2] https://sites.google.com/view/fbellavia/research

3. DATA AND EVALUATION

3.1 Dataset

The proposed pipeline was evaluated on two datasets representing typical conditions of photogrammetric surveys, in terms of scene and acquisition network. These datasets feature common and challenging acquisition setup in order to analyze a broad spectrum of photogrammetric applications (Nocerino et al., 2014; Remondino et al., 2017, 2021). All datasets have CPs for accuracy analyses in object space. To be noted that the derived Ground Sampling Distance (GSD) is reported with respect to the current working dimensions of the images, which are downscaled as discussed in Sec. 3.2.

Ventimiglia Theatre dataset. This dataset consists of three sub-datasets acquired with a Nikon D3X (24 MP, full frame sensor, 50 mm focal length) of the Ventimiglia Theatre (Italy) on an Unmanned Aerial Vehicle (UAV): the first one (*Ventimiglia Theatre Nadiral* - blue cameras in Fig. 2) contains only nadiral acquisitions arranged in two parallel strips, a typical situation in professional photogrammetric surveys with a weak image network; the second sub-dataset (*Ventimiglia Theatre Nadiral+Oblique*) adds to the previous set two overlooking oblique strips (red cameras in Fig. 2), i.e. a network designed to improve self-calibration, with the aim of testing the keypoint detector capability and evaluate matching performances on strong converging images; the third sub-dataset (*Ventimiglia Theatre Oblique*) consists of two oblique convergent cross-strips (the blue and red cameras in Fig. 3), a particularly challenging situation due to the significant viewpoint changes. The average GSD in all Ventimiglia images is 11 mm, while 9 CPs are available for the accuracy analyses.

Paestum Wall dataset. This dataset depicts part of the old defensive walls of the archaeological site of Paestum (Italy), for a length of about 80 m. Besides a first sub-dataset (*Paestum Wall Normal*) of normal images arranged in two strips (blue cameras in Fig. 4), the complete dataset (*Paestum Wall Full*) adds convergent images (red cameras in Fig. 4). A third sub-dataset (*Paestum Wall Oblique*) contains only a sub-set of oblique images from the complete dataset with low overlap (see Fig. 5). The particular acquisition setup of Peastum Wall sub-datasets provides a clear insight of the keypoint localization accuracy, since the error perpendicular to the wall tends to visibly distort the straight wall profile as there is no loop closure detection or an adequate camera network to mitigate the errors propagated along the wall. All Paestum images were acquired with the same Nikon D3X camera and have a GSD between 4-9 mm. For the metric evaluations, 22 CPs are available.

Worth to note that all sparse point clouds shown in Figs. 2-5 are obtained using the proposed HP.

3.2 Evaluation setup

The evaluation setup was carefully designed in order to obtain a fair comparison and the evaluation scripts are freely available on GitHub[3]. The proposed HP was compared against state-of-the-art methods already analyzed and evaluated in Remondino et al. (2021): SIFT (Lowe, 2004), RootSIFT (Arandjelović and Zisserman, 2012), SURF (Bay et al., 2008) and AKAZE (Alcantarilla et al., 2013) as hand-crafted methods and ASLFeat, R2D2, Key.Net (Barroso-Laguna et al., 2019) plus HardNet (KN+HN), LF-Net and SuperPoint among the deep learning methods. Moreover, Agisoft Metashape[4] was also included in the evaluation, being generally the reference commercial tool used in close-range applications.

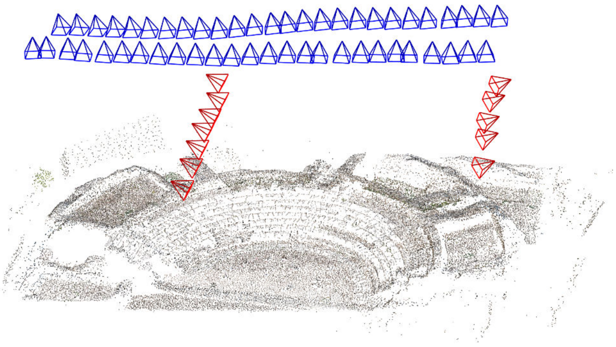


Figure 2. The UAV dataset *Ventimiglia Theatre Nadiral+Oblique* with 52 nadiral (blue) and 12 oblique (red) images.

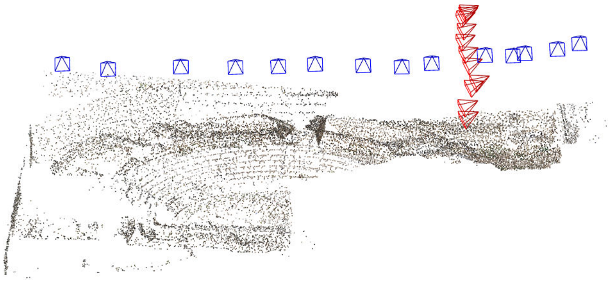


Figure 3. The UAV dataset *Ventimiglia Theatre Oblique* with two orthogonal image strips (blue, red) with convergent orientation. Only the proposed HP method could correctly orient all images.

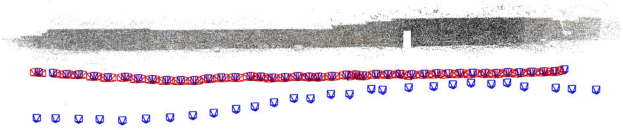


Figure 4. The terrestrial dataset *Paestum Wall Full* mixing normal (blue) and convergent (red) images.

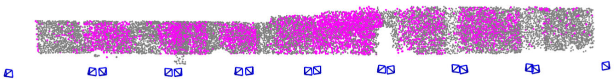


Figure 5. The terrestrial dataset *Paestum Wall Oblique* with only a strip of wide-baseline oblique images. The sparse model obtained ignoring two view tracks is highlighted in magenta.

The COLMAP camera model is shared for all the input images of the same dataset, but the specific camera model used may change between the different datasets as the best camera model strongly depends on the camera network configuration. Camera models with more parameters are reasonably preferable when a suitable image network is available. In our evaluation, the employed camera models were RADIAL (same focal length in the horizontal and vertical directions and two radial distortion coefficients) and OPENCV (two focal length parameters and two coefficients for both radial and tangential distortions), with/without applying a Principal Point (PP) refinement within the final global BA. Moreover, during the triangulation step of the SfM pipeline, two view tracks (i.e. tie points visible only in

[3] `https://github.com/3DOM-FBK/COLMAP_scripts`

[4] `https://www.agisoft.com/`

two images) were generally ignored (i.e. the minimum Track Length (mTL) was set to 3) with the exception of the *Paestum Wall Oblique* dataset for which otherwise a too sparse point distribution is obtained (see Fig. 5), leading to worse results in the BA. For completeness, the preliminary analysis for the selection of the most appropriate camera model setup for each dataset is reported in the Appendix.

Images were also downsized to 1500×1000 px (1/4 of the original input size) due to the computational constraints imposed by the deep architectures included in the evaluation. Moreover, to allow a fair comparison and similarly to recent benchmarks and works (Jin et al. (2020), Remondino et al. (2021)), the number of extracted keypoints for each image was limited to 8000 whereas the employed NNR thresholds were set to: 0.80 for SIFT, RoofSIFT and ASLFeat; 0.85 for KH+HN; 0.90 for SURF, AKAZE and SuperPoint; 0.95 for R2D2 and LF-net. Since COLMAP default RootSIFT only imposes a very soft constraint on the number of keypoints, it was replaced by the OpenCV implementation. Anyways, no relevant differences were noted during the experimentation. Finally, with the exception of our hybrid pipeline which employs DegenSAC, COLMAP internal RANSAC (with default parameters) was applied before the BA.

Results are reported in terms of the Root Mean Square Error (RMSE), i.e. the difference between ground truth (CPs) and computed 3D coordinates (CPs are not included in the BA). For a thorough analysis, results achieved by thresholding the keypoint reprojection error of the BA to 4 px (COLMAP default) and 1 px are both included. For completeness, reported BA statistics - as percentages of the SIFT values considered as reference - include: the number of Registered Images (RI, i.e. oriented images), the number of computed 3D points in the sparse cloud, the Mean Track Length (MTL) and the Mean Reprojection Error (MRE). For Metashape, MRE is replaced by the root mean square error of the reprojection errors, since this is the only measurement provided and it gives at least an indicative metric.

3.3 Results

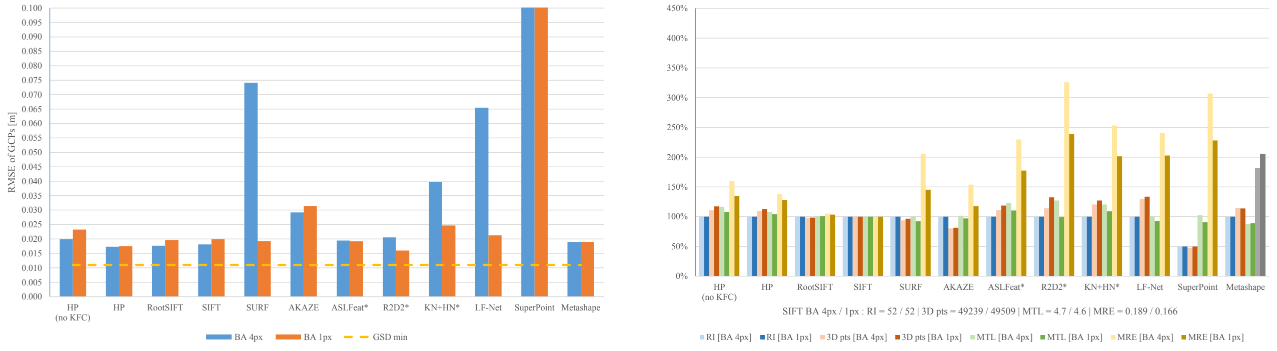
Ventimiglia Theatre dataset. All compared methods for the orientation of the *Ventimiglia Theatre Nadiral* dataset reached similar RMSEs (see Fig. 6(a)) and were able to orient all images, with the exception of SuperPoint (see the RI value reported together with other BA statistics in Fig. 6(b); notice that Metashape root mean square reprojection errors are reported as gray histogram bars instead of the MRE ones in the figures), which failed to orient the whole dataset. This is probably due to the low number of keypoints that normally SuperPoint detects: using only the cross-check to define matches by skipping the NNR check, also SuperPoint was able to register all the images but with very high RMSE. It should be remarked that fully deep pipelines - with the exception of LF-Net - have required to manually rotate the input images (indicated by the * superscript in the legend of the figures). R2D2 and HP achieved the best RMSEs while AKAZE and KN+HN obtain less favorable results. KFC is effective to improve the results of the proposed pipeline and limiting the BA reprojection error from 4 px to 1 px is only beneficial for some methods, such as R2D2, KN+HN and LF-Net. No method was able to reach accuracy better than the GSD value, highlighting the challenging condition of this dataset. The MTL, the MRE and the number of 3D points, reported in Fig. 6(b), seem unable to provide alternatives to the metric ground-truth provided by the RMSEs of the CPs. In particular, R2D2 can outperform SIFT in terms of RMSE of CPs, although obtaining a higher MRE.

The inclusion of oblique images, added into the *Ventimiglia Theatre Nadiral+Oblique* dataset, improves the BA accuracy provided that matches can be robustly established. As shown in Fig. 7(a), the RMSE for most methods can achieve an accuracy close or better than the GSD measurement. However, it must be noted that both SuperPoint and RootSIFT failed to register all the images (see the RI value reported in Fig. 7(b)). While for SuperPoint the previous observations hold, for RootSIFT it was found that one of the two oblique strips has not enough matches to be correctly registered. Likely, RootSIFT would be able to orient all images using a different camera model than that selected in the preliminary analysis reported in the Appendix. This underlines the strong dependency of the final achievable results on the whole pipeline configuration and will be investigated in future works. KN+HN and HP provide the best RMSEs, while LF-Net, RD2D and AKAZE reach the worst. Again, KN+HN, as other fully deep methods with the exception of LF-Net and SuperPoint, required the manual rotation of the images. No relevant differences are found when including KFC in the proposed pipeline, while a BA reprojection error of 1 px can slightly degrade the model accuracy according to the RMSE, with the exception of SURF that can reach in the worst case the highest error, truncated in Fig. 7(a) to 0.1 m for visualization purposes. Also in this case the BA statistics (shown in Fig. 7(b)) seem unable to correlate with the derived RMSEs of the CPs. While in terms of RMSE the general performance is quite similar among the compared methods, HP performs significantly better in terms of valid matches, as suggested by the highest number of valid matches found by HP in the nadiral-oblique image pair with the larger overlap of the *Ventimiglia Theatre Nadiral+Oblique* dataset, reported in Fig. 8. The corresponding matches found using the proposed HP and other methods are shown in Fig. 9 for a visual qualitative evaluation.

Finally, with the *Ventimiglia Theatre Oblique* dataset, only HP, even without KFC, was able to register all the input images while all other compared methods, including both RootSIFT and Agisoft Metashape, failed in this task, creating a separate model for each of the two strips (see Fig. 3). This highlighs the ability of the proposed pipeline to orient images at very different viewpoints.

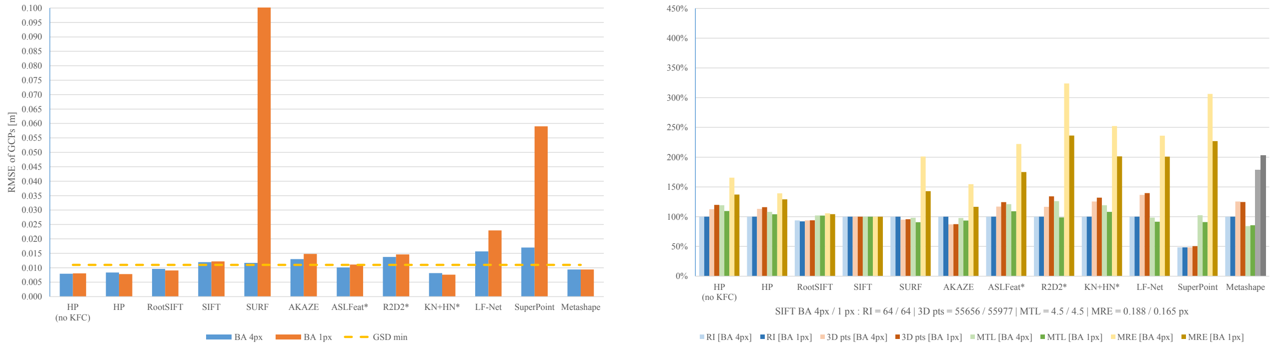
Paestum Wall dataset. For the purpose of matching correspondences, the *Paestum Wall Normal* dataset (blue images in Fig. 4) is simpler than the other datasets since the scene is planar and not exposed to relevant viewpoint changes, but it presents scale variations given by the two image strips. The RMSE histograms for the chosen camera setup are reported in Fig. 10(a), with histogram bars for SURF truncated to 0.1 m in the worst case for visualization purposes. Both RootSIFT and Metashape are on par with the GSD lower limit, and moving the BA reprojection error from 4 px to 1 px is in general very beneficial. Other methods, including HP with or without KFC, get results close to the GSD upper limit, except for SURF, R2D2, KN+HN and LFNet which show significantly higher errors. For this particular dataset, the KFC module does not seem to be strongly effective. Notice also that, according to Table 3 reported in the Appendix, the gap between HP and the top methods would be further reduced with a mTL of 2. BA model statistics generally employed as pseudo ground-truth, reported in Fig. 11(a), does not seem to be strongly correlated with the metric ground-truth provided by the CPs.

The *Paestum Wall Full* (blue and red images in Fig. 4) enriches the camera network of the previous *Paestum Wall Normal* dataset with some highly convergent images which introduce perspective distortions of the wall plane but generally help



(a) (b)

Figure 6. *Ventimiglia Theatre Nadiral* dataset: (a) RMSEs of the CPs and (b) BA statistics as percentages with respect to SIFT values (OPENCV camera model with no PP post-refinement, mTL=3). The min GSD is reported with respect to the downsampled images. The * superscript indicates that input images were manually rotated as the considered learning-based method is not invariant to camera rotation.



(a) (b)

Figure 7. *Ventimiglia Theatre Nadiral+Oblique* dataset: (a) RMSEs of the CPs and (b) BA statistics as percentages with respect to SIFT values (RADIAL camera model with PP post-refinement, mTL=3). The min GSD is reported with respect to the downsampled images. The * superscript indicates that input images were manually rotated as the considered learning-based method is not invariant to camera rotation.

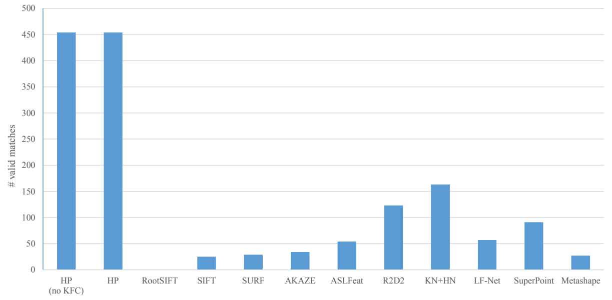


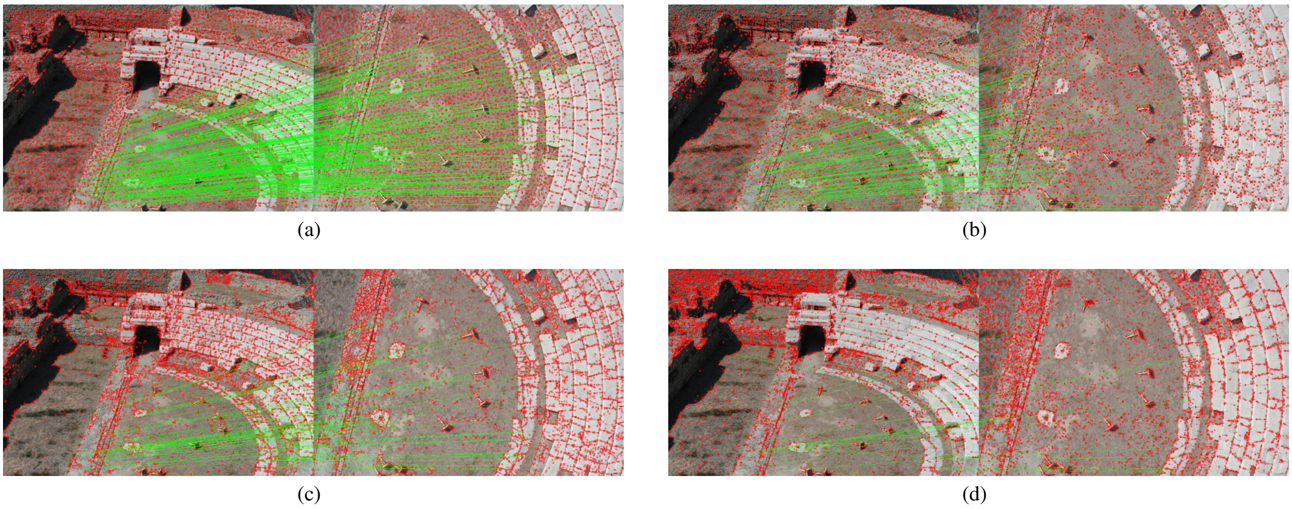
Figure 8. Number of valid RANSAC/DegenSAC matches for the nadiral-oblique image pair with the largest overlap in the *Ventimiglia Theatre Nadiral+Oblique* dataset.

in decreasing the RMSE of the CPs (see Fig. 11(a)). Only for RootSIFT and Metashape the introduction of the oblique images causes a slight increase of the RMSE, which is however in the GSD trusted range. With the exception of SuperPoint and R2D2, decreasing the BA reprojection error from 4 px to 1 px still improves the final accuracy. The best results are achieved by HP, on par with the GSD lower limit. HP without KFC, SIFT, RootSIFT, ASLFeat, LF-Net and Metashape are also effective, providing results in the GDS range. As for the previous case, BA model statistics reported in Fig. 11(b) do not provide an effective error model measurement.

The *Paestum Wall Oblique* dataset (see Fig. 4) is the most challenging since it features a sparse camera network with quite convergent images, which required to use a mTL of 2 during COLMAP triangulation step to derive a sufficient keypoint coverage for the BA. RMSEs and BA statistics are reported in Figs. 12(a)-12(b), respectively. SURF, AKAZE and SuperPoint failed to register all images (see the RI value in Fig. 12(b)) while HP and ASLFeat achieved the best results in terms of RMSE, followed by SIFT, RootSIFT, Metashape, SuperPoint, SURF and HP without KFC. With the exception of R2D2 and AKAZE, BA with 1 px reprojection error improved the model accuracy, as well as the use of the KFC module in HP. Also for this test, BA statistics (see Fig. 12(b)) do not have strong relations with the accuracy values provided by RMSEs.

4. CONCLUSIONS

This paper analyzes recent image matching algorithms for SfM applications in photogrammetry, proposing a hybrid pipeline which combines both hand-crafted and learning-based approaches. Recent literature, in particular for end-to-end deep network architectures, tends to assume that the input images have almost the same orientation, neglecting the problem of rotation invariance, which is - on the contrary - crucial in photogrammetric surveys. Unlike these approaches, the proposed HP method integrates modern state-of-the-art algorithms, pre-

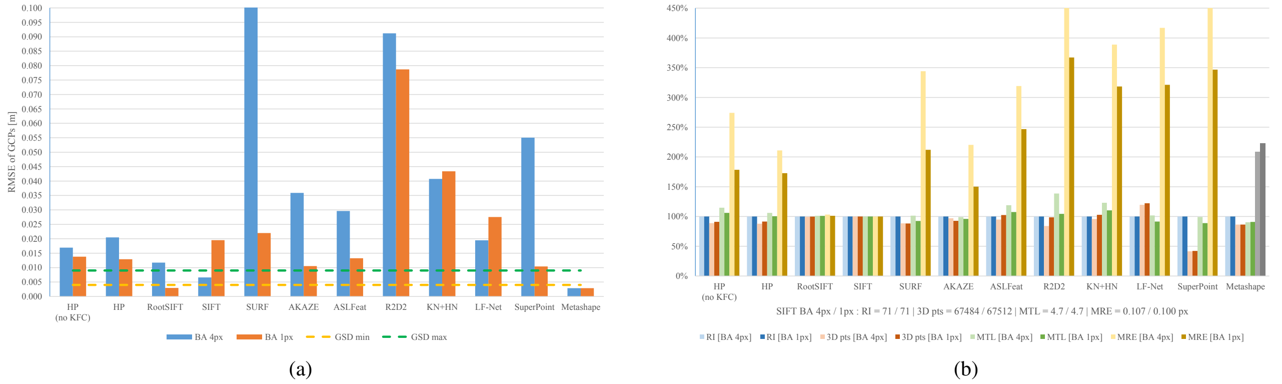

(a)

(b)

(c)

(d)

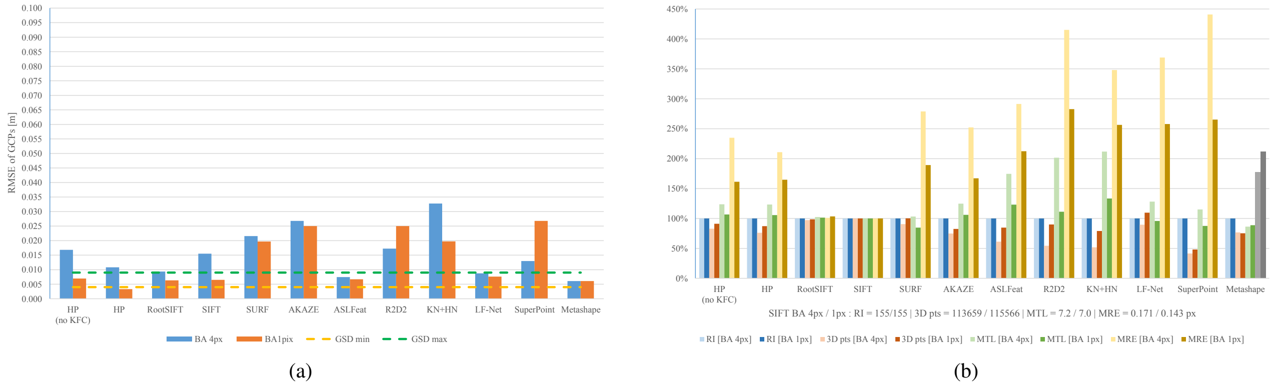
Figure 9. *Ventimiglia Theatre Nadiral+Oblique* dataset: matches comparison in a nadiral-oblique image pair with the largest overlap for (a) the proposed HP, (b) R2D2, (c) SuperPoint and (d) SIFT.



(a)

(b)

Figure 10. *Paestum Wall Normal* dataset: (a) RMSEs of the CPs and (b) BA statistics as percentages with respect to SIFT values (OPENCV camera model with PP post-refinement, mTL=3). The min e max GSDs are reported with respect to the downsampled images.

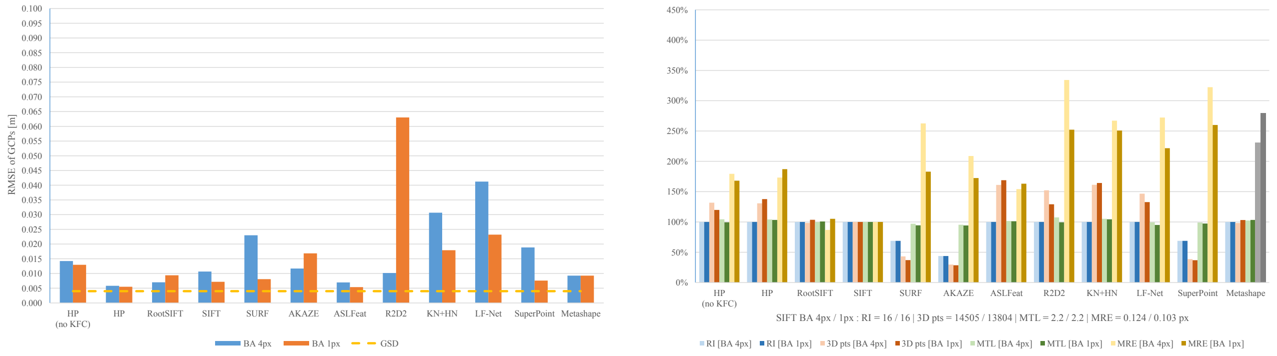


(a)

(b)

Figure 11. *Paestum Wall Full* dataset: (a) RMSEs of the CPs and (b) BA statistics as percentages with respect to SIFT values (OPENCV camera model with PP post-refinement, mTL=3). The min and max GSDs are reported with respect to the downsampled images.

serving the rotation invariance. This design makes our hybrid pipeline modular, flexible and capable of exploiting new deep-learning descriptors which are not natively rotational invariant. HP also embeds the novel KFC module as final step before the BA which is generally able to improve the final accuracy by filtering weak connections in the camera network. HP performed better or on a par with standard SIFT-based pipelines in common acquisition scenarios, with quite stable results in terms of RMSE accuracy. The achieved results are in line with Nocerino et al. (2014), although the processing reported in this paper was performed on much smaller images due to the computational limits imposed by deep architectures. Furthermore, HP was the only method able to handle challenging conditions with strong perspective distortions found e.g. in the *Ventimiglia Theatre Ob-*



(a)

(b)

Figure 12. *Paestum Wall Oblique* dataset: (a) RMSEs of the CPs and (b) BA statistics as percentages with respect to SIFT values (RADIAL camera model with PP post-refinement, mTL=2). The min and max GSDs are reported with respect to the downsampled images.

lique dataset: HP provided a complete registration of the images and was able to extract significantly more valid matches on the highly distorted images (see Figs. 3, 8 and 9).

The provided evaluation analysis also underlines the need of a careful setup of the full pipeline to evaluate its real performances, not only limited to the image matching part, as in previous comparisons, but also including the BA. The selection of the camera model and BA parameters strongly depend on the camera network and the kind of images, thus affecting the final model accuracy. As photogrammetry surveys provide GCPs/CPs, these must be used to find the optimal setup. Finally, according to aforementioned analysis, common BA statistics such as MRE, MTL and the number of the 3D points of the final model, are unable to provide a good approximation of the model accuracy which can be metrically estimated through GCPs used as CPs.

Future work will involve the extension of the above analysis with the introduction of further datasets and the inclusion of more recent image matching approaches. Additionally, an exhaustive evaluation of the different BA configuration setup will be investigated, as well as scalability issues, both in terms of running times and computational resource requirements.

ACKNOWLEDGEMENTS

F. Bellavia is funded by the Italian Ministry of Education and Research (MIUR) under the program PON Ricerca e Innovazione 2014-2020, cofunded by the European Social Fund (ESF), CUP B74I18000220006, id. proposta AIM 1875400, linea di attività 2, Area Cultural Heritage.

References

Alcantarilla, P., Nuevo, J., Bartoli, A., 2013. Fast explicit diffusion for accelerated features in nonlinear scale space. *Proceedings of the British Machine Vision Conference (BMVC)*.

Arandjelović, R., Zisserman, A., 2012. Three things everyone should know to improve object retrieval. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2911–2918.

Barroso-Laguna, A., Riba, E., Ponsa, D., Mikolajczyk, K., 2019. Key.Net: Keypoint detection by handcrafted and learned CNN filters. *Proceedings of the International Conference on Computer Vision (ICCV)*.

Bay, H., Ess, A., Tuytelaars, T., Gool, L. V., 2008. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3), 346-359.

Bellavia, F., 2021. SIFT Matching by Context Exposed. *arXiv ePrint 2106.09584*.

Bellavia, F., Mishkin, D., 2021. HarrisZ$^+$: Harris Corner Selection for Next-Gen Image Matching Pipelines. *arXiv ePrint 2109.12925*.

Bellavia, F., Tegolo, D., Valenti, C., 2011. Improving Harris corner selection strategy. *IET Computer Vision*, 5(2), 86–96.

Chum, O., Werner, T., Matas, J., 2005. Two-view geometry estimation unaffected by a dominant plane. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T., 2019. D2-Net: A trainable CNN for joint detection and description of local features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fischler, M., Bolles, R., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6), 381-395.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K., Trulls, E., 2020. Image Matching across Wide Baselines: From Paper to Practice. *International Journal of Computer Vision*, 129, 517-547.

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

Luo, Z., Zhou, L., Bai, X., Chen, H., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L., 2020. ASLFeat: Learning local features of accurate shape and localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J., 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 4829–4840.

Mishkin, D., Radenovic, F., Matas, J., 2018. Repeatability is not enough: Learning affine regions via discriminability. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Nocerino, E., Menna, F., Remondino, F., 2014. Accuracy of typical photogrammetric networks in cultural heritage 3d modeling projects. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40, 465–472.

Ono, Y., Trulls, E., Fua, P., Yi, K. M., 2018. LF-Net: Learning local features from images. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Pautrat, R., Larsson, V., Oswald, M., Pollefeys, M., 2020. Online invariance selection for local feature descriptors. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Pultar, M., 2020. Improving the HardNet Descriptor. *arXiv ePrint 2007.09699*.

Remondino, F., Menna, F., Morelli, L., 2021. Evaluating hand-crafted and learning-based features for photogrammetric applications. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43-B2, 549–556.

Remondino, F., Nocerino, E., Toschi, I., Menna, F., 2017. A critical review of automated photogrammetric processing of large datasets. *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 591–599.

Revaud, J., Weinzaepfel, P., de Souza, C., Humenberger, M., 2019. R2D2: Repeatable and reliable detector and descriptor. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Riba, E., Mishkin, D., Ponsa, D., Rublee, E., Bradski, G., 2020. Kornia: an open source differentiable computer vision library for pytorch. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Schönberger, J., Frahm, J., 2016. Structure-from-Motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schönberger, J., Hardmeier, H., Sattler, T., Pollefeys, M., 2017. Comparative evaluation of hand-crafted and learned local features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Toldo, R., Gherardi, R., Farenzena, M., Fusiello, A., 2015. Hierarchical structure-and-motion recovery from uncalibrated images. *Computer Vision and Image Understanding*, 140, 127-143.

Verdie, Y., Yi, K., Fua, P., Lepetit, V., 2015. TILDE: A temporally invariant learned detector. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5279–5288.

Yao, G., Yilmaz, A., Meng, F., Zhang, L., 2021. Review of Wide-Baseline Stereo Image Matching Based on Deep Learning. *Remote Sensing*, 13(16).

Yi, K., Trulls, E., Lepetit, V., Fua, P., 2016. LIFT: Learned invariant feature transform. *Proceedings of the European Conference on Computer Vision (ECCV)*.

APPENDIX

Tables 1-5 show the preliminary analysis for selecting the most appropriate camera model (RADIAL or OPENCV) on each dataset according to the image network. Only a meaningful subset of the image matching methods was employed: HP without KFC, SIFT, Metashape and R2D2 (which often provides worse results). For all methods the maximum reprojection error in the BA was set to 4 px (COLMAP default) and 1 px. The column corresponding to the selected camera setup is highlighted in bold, while the best setup for each method in blue. It can be clearly noted that the choice of the camera setup is not trivial for some datasets.

Camera model	RADIAL	RADIAL	**OPENCV**	OPENCV
PP refinement	no	yes	**no**	yes
mTL	3	3	**3**	3
HP (no KFC) - 4px BA	0.029	0.036	**0.020**	0.056
HP (no KFC) - 1px BA	0.031	0.037	**0.023**	0.032
SIFT - 4px BA	0.022	0.027	**0.018**	0.064
SIFT - 1px BA	0.024	0.029	**0.020**	0.043
R2D2 - 4px BA	0.019	0.017	**0.021**	0.110
R2D2 - 1px BA	0.022	0.025	**0.016**	0.039
Metashape	0.025	0.030	**0.019**	0.046

Table 1. RMSEs [m] for different camera setups of the *Ventimiglia Theatre Nadiral* dataset.

Camera model	RADIAL	**RADIAL**	OPENCV	OPENCV
PP refinement	no	**yes**	no	yes
mTL	3	**3**	3	3
HP (no KFC) - 4px BA	0.009	**0.008**	0.011	0.009
HP (no KFC) - 1px BA	0.011	**0.008**	0.011	0.010
SIFT - 4px BA	0.017	**0.012**	0.013	0.012
SIFT - 1px BA	0.012	**0.012**	0.013	0.013
R2D2 - 4px BA	0.026	**0.014**	0.016	0.014
R2D2 - 1px BA	0.027	**0.015**	0.030	0.028
Metashape	0.016	**0.009**	0.014	0.019

Table 2. RMSEs [m] for different camera setups of the *Ventimiglia Theatre Nadiral+Oblique* dataset.

Camera model	RADIAL	RADIAL	OPENCV	**OPENCV**	OPENCV
PP refinement	no	yes	no	**yes**	yes
mTL	3	3	3	**3**	2
HP (no KFC) - 4px BA	0.058	0.028	0.032	**0.017**	0.050
HP (no KFC) - 1px BA	0.019	0.018	0.017	**0.014**	0.006
SIFT - 4px BA	0.013	0.016	0.011	**0.007**	0.096
SIFT - 1px BA	0.012	0.017	0.045	**0.019**	0.004
R2D2 - 4px BA	0.372	0.071	0.379	**0.091**	0.034
R2D2 - 1px BA	0.083	0.079	0.083	**0.079**	0.101
Metashape	0.014	0.018	0.004	**0.003**	0.007

Table 3. RMSEs [m] for different camera setups of the *Paestum Wall Normal* dataset.

Camera model	RADIAL	RADIAL	OPENCV	**OPENCV**
PP refinement	no	yes	no	**yes**
mTL	3	3	3	**3**
HP (no KFC) - 4px BA	0.036	0.013	0.011	**0.017**
HP (no KFC) - 1px BA	0.034	0.015	0.011	**0.007**
SIFT - 4px BA	0.055	0.016	0.015	**0.015**
SIFT - 1px BA	0.037	0.016	0.012	**0.006**
R2D2 - 4px BA	0.028	0.021	0.013	**0.017**
R2D2 - 1px BA	0.026	0.020	0.019	**0.025**
Metashape	0.031	0.019	0.009	**0.006**

Table 4. RMSEs [m] for different camera setups of the *Paestum Wall Full* dataset.

Camera model	RADIAL	**RADIAL**	OPENCV	OPENCV
PP refinement	no	**yes**	no	yes
mTL	2	**2**	2	2
HP (no KFC) - 4px BA	0.013	**0.014**	0.016	0.057
HP (no KFC) - 1px BA	0.012	**0.013**	0.072	0.076
SIFT - 4px BA	0.011	**0.011**	0.034	0.019
SIFT - 1px BA	0.009	**0.007**	0.007	0.006
R2D2 - 4px BA	0.012	**0.010**	0.055	0.036
R2D2 - 1px BA	0.060	**0.063**	0.298	0.295
Metashape	0.0181	**0.009**	0.009	0.013

Table 5. RMSEs [m] for different camera setups of the *Paestum Wall Oblique* dataset.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVI-2-W1-2022-73-2022