Challenges in image matching for cultural heritage: an overview and perspective

F. Bellavia[1][0000−0002−1688−8476], C. Colombo[2][0000−0001−9234−537X], L. Morelli[3,4][0000−0001−7180−2279], and F. Remondino[3][0000−0001−6097−5342]

[1] University of Palermo, Palermo, Italy fabio.bellavia@unipa.it
[2] University of Florence, Florence, Italy carlo.colombo@unifi.it
[3] Bruno Kessler Foundation (FBK), Trento, Italy {lmorelli,remondino}@fbk.eu
[4] University of Trento, Italy

Abstract. Image matching, as the task of finding correspondences in images, is the upstream component of vision and photogrammetric applications aiming at the reconstruction of 3D scenes, their understanding and comparison. Such applications are of special importance in the context of cultural heritage, as they can support archaeologists to digitally preserve, restore and analyze antiquities, but also to compare their changes over time. The success of deep learning, now firmly established, paired with the evolution of computer hardware, has led to many advances in image processing, including image matching. Despite this progress, image matching still offers challenges, in terms of the matching process itself but also on other practical and technical aspects. This paper gives an overview of the current status of the research in image matching with a particular focus on cultural heritage, presenting both strengths and weaknesses of the most recent approaches by means of visual comparisons on exemplar challenging image pairs. Besides assisting researchers and practitioners in the choice of the most suitable solution for a given task, this analysis also suggests lines of research worth to be investigated by the community in the near future.

Keywords: image matching· cultural heritage· SIFT· deep learning· SfM.

1 Overview

1.1 Introduction

Image matching [33] plays a key role in the design of reliable and effective vision and photogrammetric methods, which represent nowadays an essential resource in several fields of human knowledge and technology dealing with digital images. In particular, the preservation and valorization of cultural heritage can greatly benefit from images, which are often the unique source of data for retrieving valuable information [13,37]. The image matching task can be summarized as the detection of correct correspondences between two or more images of the same 3D scene, taken under different viewpoints, acquisition conditions or times. Image matching can be restricted to a sparse set of well characterized points extracted with traditional handcrafted methods or more recent learning-based approaches [25]. These sparse correspondences are normally used within

the Structure-from-Motion (SfM) image orientation process, where they get refined by exploiting globally inherent geometric constraints in an optimization scheme known as bundle adjustment. The usual SfM output consists in the camera network configuration, i.e. the camera poses and calibration parameters, as well as a sparse 3D point cloud of the surveyed scene [31]. The recovered camera network configuration is then employed to obtain a finer and more complete 3D description of the scene by applying dense image matching methods, either pairwise or exploiting Multi View Stereo (MVS) [33,26]. Popular open source processing pipelines are COLMAP [29], OpenMVG+OpenMVS [23,24] and Meshroom [1], actively updated and extended by the research community. Several good commercial tools for professional use exist too.

1.2 Traditional image matching

Until recently, sparse image matching for SfM has been characterized by the following steps: (1) detecting keypoints, (2) localizing meaningful and salient regions of the image, (3) extracting these regions as patches, generally normalized in order to achieve invariance to image transformations, (4) computing the descriptor vectors associated to keypoints, whose distance is used to establish the candidate correspondences, (5) filtering the correspondences according to descriptor statistics, for instance using the best and second best overall distances, and (6) filtering the surviving matches by means of spatial global or local constraints, as those provided by epipolar geometry [14] exploited through RANdom SAmple Consensus (RANSAC) [12].

Scale Invariant Feature Transform (SIFT) [18] has dominated sparse image matching for nearly two decades. SIFT matching relies on blob-like keypoints, whose associated patches are normalized to become invariant to scale and rotation changes. For each patch, the SIFT descriptor is computed as the histogram of the gradient orientation, correspondences are then assigned according to the Euclidean distance between descriptors, and the Nearest Neighbor Ratio (NNR) strategy, often followed by RANSAC, is employed to rank them. SIFT provides a handcrafted, highly engineered and optimized matching approach, still valid today and able to obtain robust results. Indeed, all the previous mentioned SfM pipelines are based on SIFT (actually on RootSIFT [3], which introduces a slight variation in the descriptor computation). That said, further alternatives or extensions to the standard SIFT matching pipeline have been proposed with mixed fortunes during the years. The interested reader may refer to [2,15,19,33] for some recent and comprehensive reviews and comparisons.

1.3 Learning-based image matching

As in other computer vision areas, the advent of deep learning has represented a breakthrough for image matching. Besides handcrafted approaches designed on the basis of human intuition and expertise, machine learning techniques have been employed with encouraging results mainly in the design of more robust and efficient keypoint descriptors, often referred to as data-driven descriptors [19]. A remarkable turning point was undoubtedly the L2-Net deep descriptor [35],

which outperformed SIFT in many challenging scenarios. L2-Net is at the basis of the architecture for the HardNet descriptor [21], currently the state-of-the-art standalone descriptor according to several recent benchmarks [16], and employed successfully in many hybrid image matching pipelines [7].

Deep networks have progressively replaced the components of the image matching pipeline, moving from hybrid pipelines to full end-to-end deep architectures. For instance, besides descriptors, deep design has been successfully applied for the patch normalization, providing invariance to rotations and to more general affine transformations. OriNet and AffNet [22] are respectively two examples in this sense. As an additional step towards a full image matching deep architecture, end-to-end networks also integrated keypoint extraction (save for the last matching step). SuperPoint [10] can be considered a cornerstone in this evolution process, as it provides an effective way to train the whole network using synthetic images and homographic adaptations, thus without depending on handcrafted training data as with previous solutions [38]. Another successful strategy to overcome learning difficulties in end-to-end learning was demonstrated by DIScrete Keypoints (DISK) [36], that uses reinforcement learning. With the introduction of deep learning, alternatives to the detect-then-describe paradigm which binds the descriptor characterization to the keypoint definition, typical of the classic image matching pipelines, were also proposed. The detect-and-describe D2-Net [11] and the describe-to-detect D2D network [34], respectively treats as equal or gives more priority to the descriptor optimization than to the keypoint extractor. Another aspect that emerged is the gradual shift in the design of the network architecture and of the loss function towards solutions that strongly resemble their handcrafted counterparts, of which the respective deep equivalents appear as differentiable versions to be optimized on training data.

More recently, the final steps of the image matching pipeline have been absorbed into deep architectures too. This adaptation has started with the introduction of context normalization [39], which made it possible to effectively filter correspondences according to spatial constraints, and has gone further with the Order Aware Network (OANet) [40] up to the more recent SuperGlue [28]. This last state-of-the-art deep network is a full end-to-end deep architecture based on SuperPoint [10], able to associate and discard candidate matches on the basis of spatial information and descriptor statistics, relying for this last aspect on attentional graph neural networks. The Local Feature TRansformer (LoFTR) [32] futher added a coarse-to-fine schema to obtain semi-dense correspondences. Finally, it is worth mentioning the use of deep architectures such as in [17] for further refining the bundle adjustment 2D input and 3D output point localization, even if not directly involved in the matching process.

2 Analysis and evaluation

2.1 Rationale

For the comparison of image matching pipelines in photogrammetry and computer vision applications, two criteria are generally and reasonably set out. On the one hand, there is the ability to establish matches in the case of severe image

transformations, and on the other hand, the localization accuracy of the established matches. Often these two criteria tend towards opposite goals, as a better matching ability implies to include correspondences localized less accurately.

Several comparisons of image matching pipelines specifically designed for SfM have been proposed through the years. In this respect, the main problem researchers had to face was the unavailability of reliable Ground-Truth (GT) data. In order to circumvent this problem, [30] proposed to rank the matching methods according to specific statistics of the final SfM reconstructed 3D model, such as the number of register images, the mean reprojected error of the 3D points in the images, the track length and the point cloud size. As shown in [8], this solution does not correlate well with an accurate GT. A better approach is proposed in [16], which builds a pseudo GT by running a SfM pipeline on a rich set of images of the scene with a good coverage, and then verifies the matching pipelines indirectly according to the pose error obtained using a restricted and more challenging subset on the initial images. Another solution, explored by SimLocMatch[5], relies instead on synthetic rendered scenes as effective GT data. Finally, the approach employed in [8] makes use of accurate metric GT data provided by topological surveys in terms of ground control points. These points are used to establish check points upon which to measure pose errors very accurately, hence again providing an indirect evaluation of the image matching pipeline. Pseudo GTs generally offer a reasonable rough estimation, in particular in terms of matching ability, giving rise to an indirect evaluation that gets a method ranking very close to those obtained through a direct comparison on synthetic datasets. However, when it comes to analyze matching pipelines with high and similar levels of matching accuracy, the metric GT approach on real scenes leads to a better evaluation, even if the relative datasets are more difficult to obtain. It should also be taken into account that the level of scene complexity achieved by synthetic rendered images is inferior to that of real images and, as it was noted in [8], also the bundle adjustment setting, besides the image matching pipeline setup, may assume a critical role in the final pose estimation accuracy.

Among the most recent benchmark comparisons, the Image Matching Challenge (IMC)[6] has become an annual appointment to test the latest developments. Although currently only relying on pseudo GTs following [16], and not properly focused on cultural heritage, it has represented a good starting point for any successive investigation aiming at a realistic snapshot of the current situation. In the lastest IMC (2021) the fully end-to-end networks SuperGlue, LoFTR and DISK, and the Hybrid Pipeline (HP [8]) based on HarrisZ$^+$ [7] obtained the best results, a ranking that was confirmed also by SimLocMatch. Some of these methods were included in other evaluations more focused on cultural heritage applications, where their superiority over other approaches was generally confirmed. In particular, [8] addressed the analysis of the metric accuracy of several matching methods on modern image datasets, while [20] employed the IMW benchmark configuration on historical images ranging from 1860 to today.

[5] https://simlocmatch.com/ (currently offline)
[6] https://www.cs.ubc.ca/research/image-matching-challenge/current/

2.2 Results and discussion

According to [8], in terms of SfM pose estimation accuracy, SIFT pipeline is still competitive and among the best with respect to the recent approaches when the camera network is robust and provides a good coverage of the scene, i.e. with "close" images having a high overlap and low relative distortions in terms of both viewpoint and illumination changes. However, when these assumptions are not satisfied, the ability to robustly match in the presence of strong image deformations, even if less accurately, becomes essential. In fact, this helps keeping image connections in the camera network and avoids failures in registering some images, which would affect the performance of the whole SfM pipeline. For this reason, the following discussion will not address the keypoint localization accuracy (and hence the camera pose accuracy), extensively covered by previous literature (see Sec. 2.1), but will focus instead on the ability of each pipeline to provide more or less precise correspondences in challenging scenarios that are likely to be found in cultural heritage: it is not-so-infrequent to have to register images acquired from different viewpoints (e.g. aerial and terrestrial images), by different cameras, with different illumination conditions and at different times (e.g. multitemporal images) with the aim to detect the occurred changes.

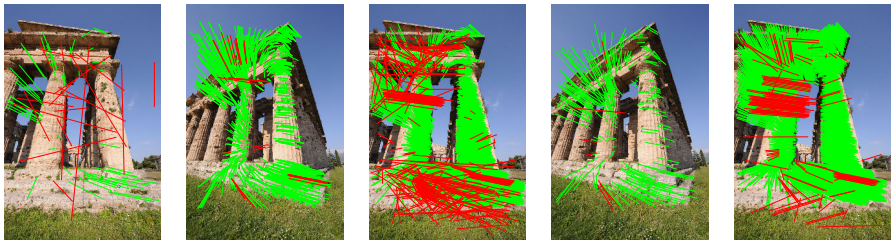
The chosen evaluation protocol that defines matches as correct is the one proposed in [6] and extended in [4] using hand-taken correspondences. These hand-taken matches are employed to obtain the epipolar geometry of the scene, so as to filter candidate matches on the basis of the epipolar error, and to compute a rough interpolated optical flow over the images to further refine filtered matches, since epipolar error only cannot be sufficient to disambiguate them. This protocol is reasonable for the qualitative evaluation through visual inspection presented hereafter, with a minimal probability of obtaining a wrong GT estimate. While providing a quantitative evaluation for challenging scenarios is in most cases unfeasible due to major difficulty to obtain an accurate metric GT, the proposed qualitative evaluation is sufficient to describe the potential and the limits of the compared methods. Besides, this setup does not employ synthetic images, and provides a direct evaluation on the true unconstrained matching ability of the evaluated methods.

The image matching pipelines included in this comparison are SIFT (used as reference), HP, DISK, SuperGlue and LoFTR. For SIFT, the VLFeat implementation is used[7], for HP and DISK the code available by the respective authors is employed, and for SuperGlue and LoFTR their respective Kornia implementations [27]. The matching ability on six different and challenging image pairs, with resolution from about 1024×768 to 1500×1000, of interest for cultural heritage is analyzed. As a good recommended practice [16], at the end of each matching pipeline DegenSAC [9] is executed to filter matches. Since the image pairs are quite challenging, the corresponding DegenSAC epipolar error threshold is set to 3 px, which is relatively high for precise photogrammetry applications, but can provide a better insight into the rough matching ability. Moreover, since this analysis concerns with a visual qualitative localization of the correspondences,

[7] https://www.vlfeat.org/

error thresholds for evaluation are set to 40 px. Due to lack of space, only the most relevant matching results are shown. The reader is strongly invited to inspect the complete report, available as additional material together with the evaluation code and data [8].

Figure 1 shows the optical flow of the matches superimposed on one of the two images of the pair. The input images (shown alternated for the sake of clarity) represent the front side of the Temple of Neptune in Paestum (Italy), and were acquired by the same camera with fixed illumination conditions. The scene is approximately planar but presents a relatively high viewpoint distortion. SIFT was barely able to find a sufficient number of correspondences, while the other methods worked with no critical issues. Note that HP and SuperGlue found less correspondences but also less wrong matches (which can become an issue in photogrammetry applications as they can invalidate the bundle adjustment estimation) than DISK and LoFTR. Specifically, SuperGlue, HP, DISK and LoFTR found matches in increasing order. Wrong matches were relevant for LoFTR, and even more for DISK.



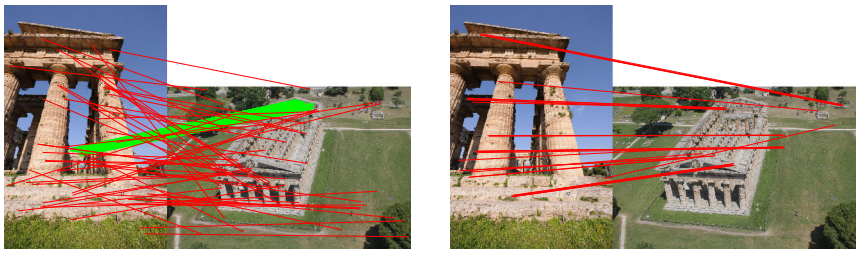
(a) SIFT (b) HP (c) DISK (d) SuperGlue (e) LoFTR

Fig. 1: Matching results in terms of the optical flow for an image pair of the Temple of Neptune (Paestum, Italy), displayed alternating the two images. Correct and wrong matches are shown in green and red, respectively.

Figure 2 still refers to the Temple of Neptune, but the image pair includes a terrestrial and an aerial image taken from an Unmanned Autonomous Vehicle (UAV), presenting large scale and illumination variations. Only DISK was able to provide some correct matches, while the other methods failed (of these, only LoFTR is reported in Fig. 2). Nevertheless, DISK was not able to correctly handle the scale variation but it was the only method to correctly localize the small image portion of corresponding regions at a similar scale.

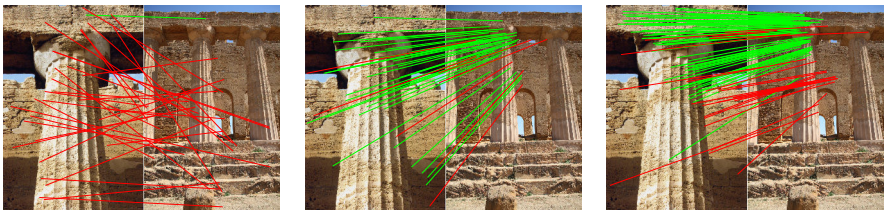
Figure 3 includes an image pair from the Temple of Concordia (Agrigento, Italy), taken with two different cameras within a time interval of about fifteen years, before and after the restoration process to which the temple underwent. Also in this case, the images present a relevant scale variation. SuperGlue and LoFTR were the only methods able to detect correct matches, with SuperGlue providing less wrong matches and a better distribution of the correct ones. In general, it seems that the coarse-to-fine approach of LoFTR, once it has found

[8] https://drive.google.com/drive/folders/1ws1SvRnym3FPh1J6K4lomTIqEsxR5k49



(a) DISK (b) LoFTR

Fig. 2: Correct and wrong matching results on a challenging image pair of the Temple of Neptune (Paestum, Italy), shown respectively in green and red.



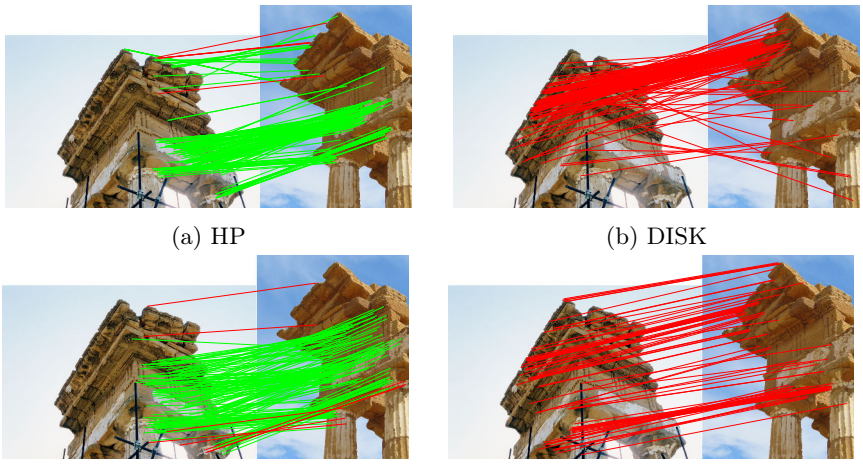
(a) DISK (b) SuperGlue (c) LoFTR

Fig. 3: Correct and wrong matching results on the image pair of the Temple of Concordia (Agrigento, Italy), shown respectively in green and red.

putative matching areas at the coarser scale, is quite indulgent in discarding matches at the finer scale. Differently from the previous image pair, DISK was not able to obtain correct matches, as well as SIFT and HP (not shown). Moreover, concerning HP, only an exiguous number of keypoints was detected in the matching region of the second image, due to a relative low global contrast of this area with respect to the whole image.

Figure 4 shows two images from the Temple of the Dioscuri (Agrigento, Italy), taken in the same conditions and time interval of those reported for the Temple of Concordia. Only HP and SuperGlue were able to correctly find correspondences. DISK and LoFTR failed to provide matches, probably due to their inability to handle both the middle level image rotations and the repeating patterns that occur within the overlapped regions. Notice also that HP was able to find matches in the upper part, while SuperGlue was more effective in the in-between part. Moreover, DISK and LoFTR produced a large number of wrong matches. Conversely, SIFT (not shown), although unable to find correct correspondences, found a very low number of wrong matches.

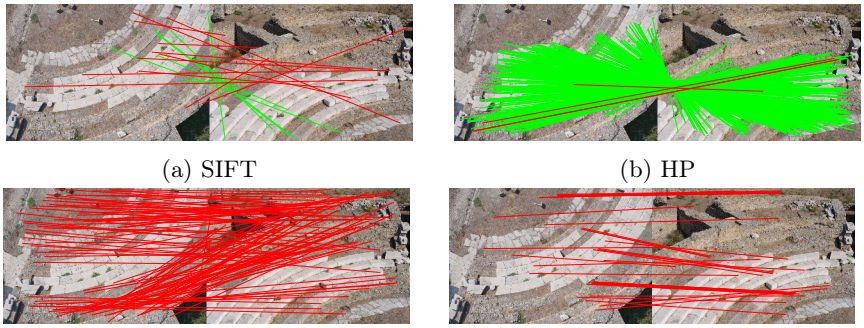
Figure 5 reports the matching results on two aerial images taken from different UAV strips, presenting relevant perspective distortions and a significant relative rotation, a common situation in UAV or Autonomous Underwater Vehicle (AUV) surveys. Only HP was able to fully assign matches. The other methods failed: DISK, LoFTR and SuperGlue (the last one is not shown) due to their invariance to even moderate rotations, while SIFT since it is unable to tolerate



(a) HP (b) DISK

(c) SuperGlue (d) LoFTR

Fig. 4: Correct and wrong matching results on the image pair of the Temple of the Dioscuri (Agrigento, Italy), shown respectively in green and red.



(a) SIFT (b) HP

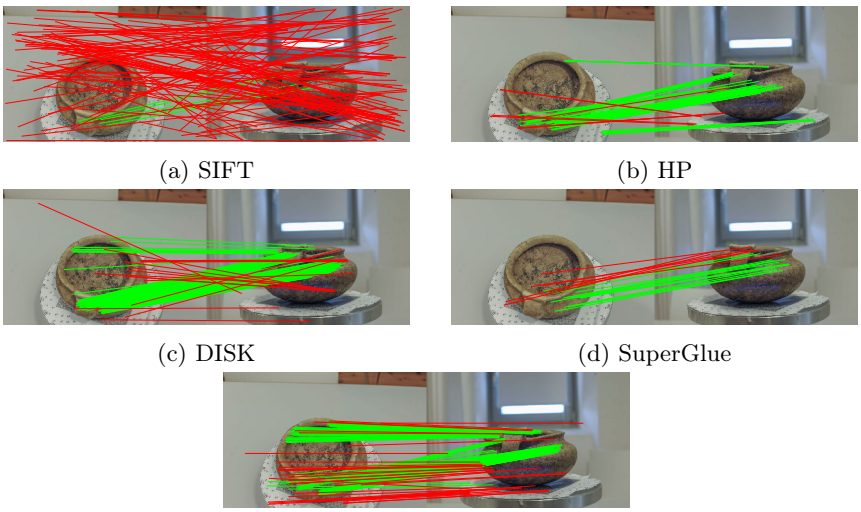
(c) DISK (d) LoFTR

Fig. 5: Correct and wrong matching results on the image pair of the Ventimiglia Theatre (Ventimiglia, Italy), shown respectively in green and red.

the viewpoint distortion, although being invariant to rotations (as suggested by the presence of few correct matches). Note again that DISK, and to a minor extent LoFTR, provided more wrong matches than SuperGlue or SIFT.

Finally, Fig. 6 presents the matching results on two views of an ancient vase from the archaeological area of Fiavé (Trento, Italy), subjected to a strong relative tilt change. All the methods succeeded in finding correct matches but they also produced spurious correspondences. Specifically, SIFT, SuperGlue, HP, DISK and LoFTR provided in order an increasing number of correct matches. Moreover, HP and SuperGlue obtained the lowest number of wrong matches, followed by DISK, LoFTR and SIFT. Note also that HP was the only method able to trace correspondences at the base plane, yet it missed the matches on the left upper part, that were detected instead by DISK and LoFTR. It is also worth

mentioning that at the original input images resolution of 6048×4032 instead of the current processing resolution of 1500×1000, none of the methods was able to find correct matches, a fact that highlights the criticality of the detection scale.



(a) SIFT (b) HP

(c) DISK (d) SuperGlue

(e) LoFTR

Fig. 6: Correct and wrong matching results on the image pair of an ancient vase (Fiavé, Italy), shown respectively in green and red.

3 Conclusions and perspective

The presented investigation showed that recent learning-based image matching techniques provide an unquestionable advance with respect to traditional handcrafted methods. However, there are still issues and not a clear winner, leaving large margins for improvement.

End-to-end LoFTR, DISK and SuperGlue are unable to handle average and high degrees of image rotation. This issue can be circumvented by rotating one of the input image (at least eight times for a tolerance to 45°) but this would be computationally expensive, efficient solutions in this sense for handcrafted methods have been investigated [5]. Both DISK and LoFTR output a high number of matches, including wrong correspondences. In the case of failure, this can represent a critical issue that can mislead the next steps of the SfM pipeline, but also in case of success this can create problems due to computational requirements for data management. Furthermore, LoFTR outputs discrete keypoint locations for the first image and sub-pixel keypoint localizations for the second one, thus requiring some engineering in order to handle multiple images. On the other hand, the lower number of correspondences extracted by SuperGlue, inherited from Superpoint, can be limiting in some cases [25].

HP is rotation invariant, provides a reasonably high number of correct matches, outputs a low number of wrong correspondences in case of failure and is robust to viewpoint distortions similarly to end-to-end architectures. Nevertheless, it

is less tolerant to strong scale variations, and can have issues related with the global image contrast. Being a modular pipeline, HP can be more easily adapted for specific tasks. For instance, by removing OriNet from its steps can improve its matching ability when there are no relevant rotations in the input images. Finally, the retraining of the single deep modules of HP requires less computational efforts than with a fully end-to-end network.

That said, SIFT still offers advantages in common-user non-challenging scenarios, which makes it still irreplaceable in commercial applications: SIFT is less computational expensive than its competitors with inferior hardware requirements, and provides state-of-the-art pose accuracy estimation in case of robust camera network input setups [25]. Moreover, it works without efforts with high resolution images (e.g. aerial image datasets), while its competitors cannot be launched using a high-end consumer-grade system configuration (with the exception of HP that runs much slower than SIFT with these images anyways). Multi-scale tiling can be devised in this case, which should also provide a solution in case of relevant scale changes, but again at the expense of increasing the computational cost, and probably decreasing the matching accuracy as the global overview of the scene is somewhat lost.

Although recent image matching approaches can be useful for research activities dealing with challenging SfM applications in cultural heritage, these methods are not yet fully ready and mature for common user applications. Effective solutions are still an open question, which offers new research opportunities and challenges to the scientific community, not only for the matching process itself but also in providing efficient and scalable solutions.

Acknowledgements The authors would like to thank the Archaeological and Landscape Park of the Valley of the Temples of Agrigento (Italy), the Cultural Heritage Directorate of the Autonomous Province of Trento (Italy) and the Superintendence of the Imperia and Savona provinces (Italy) for providing some of the images used in this work. 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

1. AliceVision: Meshroom: A 3D reconstruction software (2018), https://github.com/alicevision/meshroom
2. Apollonio, F., Ballabeni, A., Gaiani, M., Remondino, F.: Evaluation of feature-based methods for automated network orientation. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. **XL-5**, 47–54 (2014)
3. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
4. Bellavia, F.: SIFT matching by context exposed. IEEE Trans. Pattern Anal. Mach. Intell. **(early access)** (2022)

5. Bellavia, F., Colombo, C.: Rethinking the sGLOH descriptor. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 931–944 (2018)
6. Bellavia, F., Colombo, C.: Is there anything new to say about SIFT matching? Int. J. Comput. Vis. **128**(7), 1847–1866 (2020)
7. Bellavia, F., Mishkin, D.: HarrisZ$^+$: Harris corner selection for next-gen image matching pipelines. arXiv ePrint 2109.12925 (2021)
8. Bellavia, F., Morelli, L., Menna, F., Remondino, F.: Image orientation with a hybrid pipeline robust to rotations and wide-baselines. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. **XLVI-2/W1-2022**, 73–80 (2022)
9. Chum, O., Werner, T., Matas, J.: Two-View Geometry Estimation Unaffected by a Dominant Plane. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2005)
10. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
11. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
12. Fischler, M., Bolles, R.: Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
13. Gruen, A., Remondino, F., Zhang, L.: Photogrammetric reconstruction of the Great Buddha of Bamiyan, Afghanistan. Photogramm. Rec. **19**(107), 177–199 (2004)
14. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2000)
15. Hartmann, W., Havlena, M., Schindler, K.: Recent developments in large-scale tie-point matching. ISPRS J. Photogramm. Remote Sens. **115**, 47–62 (2016)
16. Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E.: Image matching across wide baselines: From paper to practice. Int. J. Comput. Vis. **129**(2), 517–547 (2021)
17. Lindenberger, P., Sarlin, P., Larsson, V., Pollefeys, M.: Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)
18. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
19. Ma, J., Jiang, J., Fan, A., Jiang, J., Yan, J.: Image matching from handcrafted to deep features: A survey. Int. J. Comput. Vis. **129**(7), 23–79 (2021)
20. Maiwald, F., Lehmann, C., Lazariv, T.: Fully automated pose estimation of historical images in the context of 4d geographic information systems utilizing machine learning methods. ISPRS Int. J. Geoinf. **10**(11) (2021)
21. Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J.: Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss. In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) (2017)
22. Mishkin, D., Radenovic, F., Matas, J.: Repeatability is Not Enough: Learning Affine Regions via Discriminability. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
23. Moulon, P., Monasse, P., Perrot, R., Marlet, R.: OpenMVG: Open multiple view geometry. In: Proc. Int. Workshop on Reproducible Research in Pattern Recognition (2016), https://github.com/openMVG/openMVG

24. OpenMVS: open Multi-View Stereo reconstruction library (2022), https://github.com/cdcseacave/openMVS
25. Remondino, F., Menna, F., Morelli, L.: Evaluating hand-crafted and learning-based features for photogrammetric applications. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. **XLIII-B2-2021**, 549–556 (2021)
26. Remondino, F., Spera, M., Nocerino, E., Menna, F., Nex, F.: State of the art in high density image matching. Photogramm. Rec. **29**(146), 144–166 (2014)
27. Riba, E., Mishkin, D., Ponsa, D., Rublee, E., Bradski, G.: Kornia: an open source differentiable computer vision library for pytorch. In: Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2020)
28. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
29. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), https://colmap.github.io/
30. Schönberger, J., Hardmeier, H., Sattler, T., Pollefeys, M.: Comparative evaluation of hand-crafted and learned local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
31. Stathopoulou, E., Welponer, M., Remondino, F.: Open-source image-based 3D reconstruction pipeline: review, comparison and evaluation. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. **XLII-2/W17**, 331–338 (2019)
32. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
33. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer-Verlag, 2st edn. (2022)
34. Tian, Y., Balntas, V., Ng, T., Laguna, A.B., Demiris, Y., Mikolajczyk, K.: D2D: Keypoint extraction with describe to detect approach. In: Proceedings of the 15th Asian Conference on Computer Vision (ACCV) (2020)
35. Tian, Y., Fan, B., Wu, F.: L2-Net: deep learning of discriminative patch descriptor in euclidean space. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6128–6136 (2017)
36. Tyszkiewicz, M.J., Fua, P., Trulls, E.: DISK: Learning local features with policy gradient. In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) (2020)
37. Vincent, M., Coughenour, C., Remondino, F., Gutierrez, M., Bendicho, V.L.M., Fritsch, D.: Rekrei: A public platform for digitally preserving lost heritage. In: Proceedings of the 44th CAA Conference (2016)
38. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
39. Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
40. Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., Chen, Y., Liao, H., Quan, L.: Learning two-view correspondences and geometry using order-aware network. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5844–5853 (2019)