PHOTOGRAMMETRY NOW AND THEN
- FROM HAND-CRAFTED TO DEEP-LEARNING TIE POINTS -

Luca Morelli [a,b], Fabio Bellavia [c], Fabio Menna [a], Fabio Remondino [a]

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

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

[c] Dept. of Mathematics and Computer Science, University of Palermo, Palermo, Italy - fabio.bellavia@unipa.it

Commission II

KEY WORDS: Historical images, cultural heritage, tie points, image matching, deep learning, local features, RANSAC.

ABSTRACT:
Historical images provide a valuable source of information exploited by several kinds of applications, such as the monitoring of cities and territories, the reconstruction of destroyed buildings, and are increasingly being shared for cultural promotion projects through virtual reality or augmented reality applications. Finding reliable and accurate matches between historical and present images is a fundamental step for such tasks since they require to co-register the present 3D scene with the past one. Classical image matching solutions are sensitive to strong radiometric variations within the images, which are particularly relevant in these multi-temporal contexts due to different types of sensitive media (film/sensors) employed for the image acquisitions, different lighting conditions and viewpoint angles. In this work, we investigate the actual improvement provided by recent deep learning approaches to match historical and nowadays images. As learning-based methods have been trained to find reliable matches in challenging scenarios, including large viewpoint and illumination changes, they could overcome the limitations of classic hand-crafted methods such as SIFT and ORB. The most relevant approaches proposed by the research community in the last years are analyzed and compared using pairs of multi-temporal images.

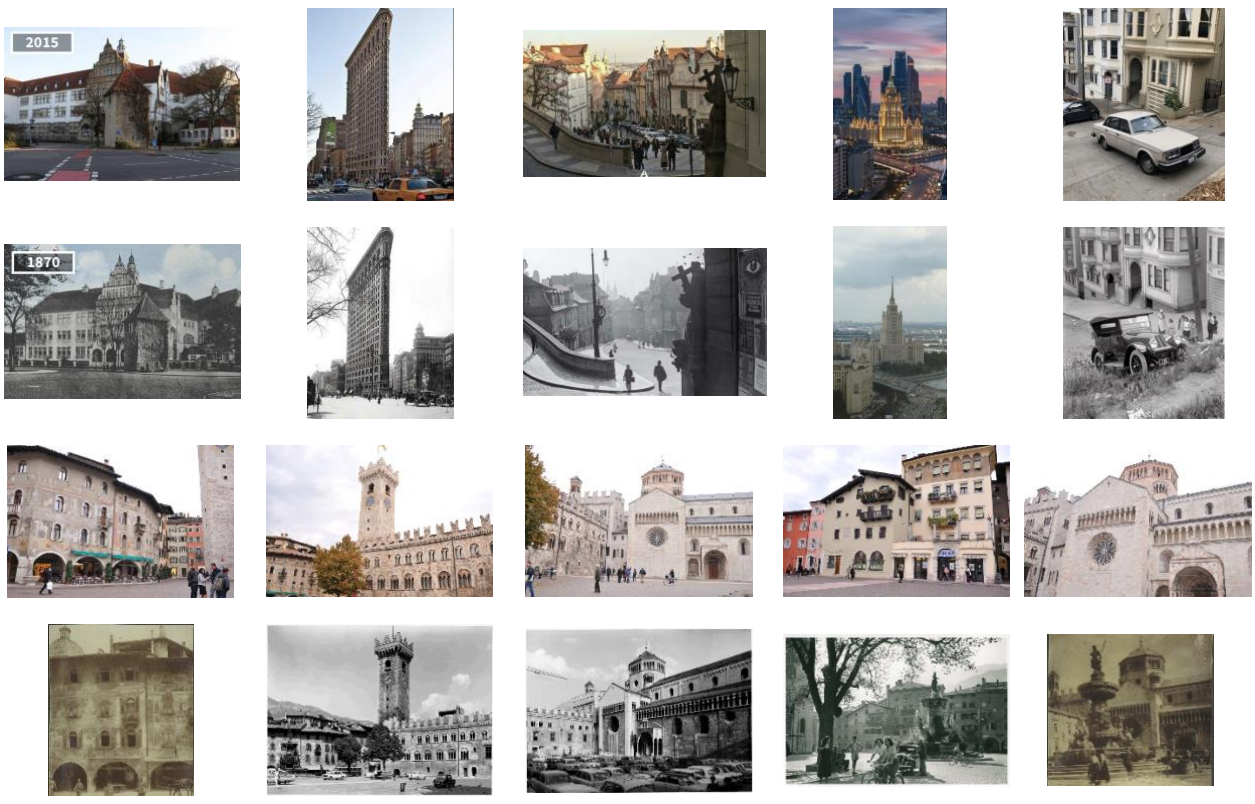

Figure 1: Multi-temporal images employed in this work for the evaluation of hand-crafted and learning-based tie points. “now” and “then” image pairs are shown in the odd and even rows, respectively. The first two rows, from left to right depict scenes from Osnabrück[1], New York[2], Prague[1], Moscow[2] and San Francisco[2], the last two rows refer to the main square in Trento (Italy).

[1] https://www.boredpanda.com/
[2] https://www.demilked.com/

1. INTRODUCTION

Automatic image matching is a fundamental task in several engineering applications, and it has been traditionally performed with hand-crafted approaches like the Scale Invariant Feature Transform (SIFT, Lowe, 2004), the Oriented FAST and Rotated BRIEF (ORB, Rublee et al., 2011) and the Speeded Up Robust Features (SURF, Bay et al., 2006). Nevertheless, it was widely experimentally verified that these methods do not perform well in challenging conditions such as strong changes in radiometric content and angle of views. Indeed, more recent deep-learning methods appear to provide promising and encouraging results (Remondino et al., 2021; Chen et al., 2021; Bellavia et al., 2022a).

Multi-temporal datasets are particularly challenging for image matching, since they present several critical situations: large viewpoint and radiometric changes, blurry and noisy areas, or artefacts (Maiwald, 2019a). These conditions are particularly significant when matching old historical images with nowadays images. Moreover, the printing process from the original film and the digitization of the prints or of the original film itself by using a flatbed scanner or by photographing the hardcopies is likely to introduce additional geometric deformations with complex mathematical modelling (Nocerino et al., 2012a). Multi-temporal image co-registrations can be useful for multiple purposes including, but not limited to, Augmented and Virtual Reality (AR/VR) applications (Torresani et al., 2021; Maiwald et al., 2019b), the valorisation of archival photos (Nocerino et al., 2012b), 3D reconstruction of destroyed building facades (Brauer-Burchardt and Voss, 2002) or statues (Gruen et al., 2004), and environmental and climate changes monitoring (Holmlund, 2021).

When dealing with multi-temporal datasets, the approaches are different depending on the number of images per epoch. If the number of images at each epoch is enough to orient a photogrammetric block, the approach of Zhang et al. (2020; 2021) can be used. This approach first seeks matches among images of the same epoch, and then among images of different epochs. Zhang et al. (2020) increased the number of extracted SIFT keypoint avoiding the use of the Nearest Neighbour Ratio (NNR) threshold and heavily relying on geometric constraints for the detection of outliers. Moreover, in Zhang et al. (2021) deep-based local features were used on Digital Terrain Models (DTMs).

Classic hand-crafted local features can sometimes be used when it is not possible to rely on the reconstructed scene of single epochs (Ali and Whitehead, 2014; Maiwald, 2019a), but the extracted correspondences are usually insufficient, and a high number of outliers is present. In this case traditional descriptors generally fail (Farella et al., 2022), and tie points must be collected manually (Holmlund, 2021; Torresani et al., 2021).

Recently, Maiwald (2019a) proposed a benchmark for historical images, analyzing popular hand-crafted local features such as ORB+SURF, the Maximally Stable Extremal Regions (MSER, Matas et al., 2004) in conjuction with SURF, and the Radiation-Invariant Feature Transform (RIFT, Li et al., 2018). The results underline the complexity of the historical images in automatic image matching due to the wide viewpoint differences and the very dissimilar radiometric distributions.

To overcome the limitations due to strong radiometric changes in multi-temporal image matching, several new local features based on Deep Neural Networks (DNNs) have been proposed, starting from the Temporally Invariant Learned DEtector (TILDE, Verdie et al., 2015), where keypoints were extracted by a neural network trained on multi-temporal and multi-seasonal images of static cameras. More recently, hybrid pipelines or end-to-end DNNs for image matching have been proposed (Remondino et al., 2021; Chen et al., 2021; Jin et al., 2021; Bellavia and Mishkin, 2022c). Although these DNNs may lack strong scale and rotation invariance, the research is very active. Better and more robust DNNs are progressively appearing to improve scale, rotation, and radiometric invariances as well as the localization accuracy, the repeatability and the reliability of the keypoints.

Several of these image matching DNNs have already been successfully applied and tested in satellite and aerial multi-temporal datasets (Ghuffar et al., 2022; Zhang et al., 2021; Farella et al., 2022) and terrestrial multi-temporal historical image pairs (Maiwald et al., 2021).

1.1. Paper aims

The aim of the paper is the evaluation of hand-crafted and deep-learning local features for the computation of image correspondences between current ("now") and historical ("then") images. To the best of the authors' knowledge, this is the first contribution that provides an extensive evaluation on such dilated temporal range. Furthermore, an alternative evaluation metric, previously proposed in (Bellavia, 2022d), that exploits rough optical flow estimation and the epipolar error, is employed.

Although the reported evaluations employ only image pairs, the obtained results are useful also in the case of one-to-many registration of multi-temporal images, as those employed in AR/VR applications and in multi-view historical photogrammetric applications (Maiwald et al., 2021), since Structure-from-Motion (SfM) relies on the single image pair matching as base step.

2. DATASETS AND METHODOLOGY

2.1 Datasets

The evaluation dataset includes ten multi-temporal image pairs, depicting city scenes, see Fig. 1. The historical photos are black and white or sepia and are dated around the first half of the twentieth century. On one hand, the first five image pairs of Fig. 1 are from two different internet collections[1,2] and depict different urban environment around the world. These image pairs present relatively low resolution: 692x824 px for San Francisco, 692x916 px for the Flatiron Building (New York), 1384x846 px for Prague, 688x1164 px for Moscow, and 1384x910 px for the Osnabrück. Although these image pairs show limited viewpoint variations, they are challenging due to the presence of strong radiometric changes, noise due to the acquisition stage and due to the acquisition support employed. Moreover, the scene modifications due to the ages cause to have less than fifty per cent of consistent image area between the images, i.e. of scene objects not altered across the time. The limited scene consistency can be observed for instance in Fig. 3, focusing on the overall support region of the correct matches within the matching images.

The remaining five images of Fig. 1 depict the main square of Trento (Italy). The historical scans are from the "TOTEM" project (Torresani et al., 2021), and have a limited image resolution (972x1364 px, 679x512 px, 713x512 px, 1024x689 px, and 984x1230 px), while the recent images have been collected for this work and have both a resolution of 1500x1000 px (actually, they have been down-sampled from the original acquisition resolution for computational reasons). In addition to the first image pairs, these ones contain also relevant scale and viewpoint variations.

All the images are oriented upright, so that rotation issues affecting many end-to-end matching networks (Remondino et al., 2021; Su et al., 2022) can be neglected.

2.2 Compared methods

To represent traditional local features, we included SIFT, which is still the state-of-the art among hand-crafted approaches, and ORB, widely used for real-time applications, such as SLAM (Simultaneous Localization And Mapping). The combined methods MSER+SURF and ORB+SURF evaluated in Maiwald (2019a) were also tested, but the results were not included because of their very poor performance. We included SIFT in its VLFeat[3] implementation as well as its extension RootSIFT (Arandjelović and Zisserman, 2012) available through COLMAP (Schonberger and Frahm, 2016).
Concerning deep image matching methods, we included the current state-of-the-art according to recent evaluation (Remondino et al., 2021; Chen et al., 2021; Jin et al., 2021; Ma et al., 2021; Bellavia et al., 2022b): the DIScrete Keypoints (DISK, Tyszkiewicz et al., 2020), the Hybrid Pipeline (HP, Bellavia et al., 2022c) also without rotational invariance provided by OriNet (Mishkin et al., 2018) (denoted as HP_upright), the Local Feature TRansformer (LoFTR, Sun et al., 2021) and its rotation invariant extension SE2-LoFTR (Bökman et al., 2022), SuperPoint+SuperGlue (Sarlin et al., 2020), the Accurate Shape and Localization Features (ASLFeat, Luo et al., 2020), the Repeatable Detector and Descriptor (R2D2, Revaud et al., 2019) and the Local Feature Network (LF-Net, Ono et al., 2018). Two further deep-learning methods providing quite interesting results were also added in this evaluation: the Rotation-Robust Descriptors (RoRD, Parihar et al., 2021) for its rotation invariance and the Accurate and Lightweight Keypoint Detection and Descriptor Extraction (ALIKE, Zhao et al., 2022) for its ability to run in real-time.

2.3 Evaluation metrics

A commonly adopted evaluation metric for image matching assumes that matches are correct on the basis of their epipolar error with respect to a ground-truth fundamental matrix, estimated manually or computed by SfM. However, it frequently happens that wrong matches lie along the epipolar lines. Indirect evaluation on the pose can be also employed to bypass this situation, but in the case of complex scene and high pose errors, these are unable to effectively discriminate which method is better.
To overcome these limitations, Maiwald (2019a) employed the trifocal tensor, since its dataset was arranged in triplets instead of pairs. As our dataset is composed of image pairs taken from close viewpoints, the trifocal tensor is not usable. Moreover, in this setup a match appearing only in two images cannot be evaluated, hence distorting the recall of the evaluated method.
To avoid these issues, the alternative approach defined in Bellavia (2022d) and successfully applied in Bellavia et al. (2022b) is used. The approach starts by considering the epipolar error to decide whether a match is correct, but it further removes ambiguities by imposing that the optical flow of correct matches must be consistent with respect to the sparse optical flow of a local neighbourhood of hand-taken correspondences. When feasible, local homographies are employed to refine the sparse optical flow to improve the accuracy. The hand-taken correspondences are used twice: to get the fundamental matrix ground truth and to obtain a sparse optical flow.
Figure 3b shows in yellow the manually taken tie points, while in green and blue the matches that are selected as correct according to the sparse flow and its local homography, respectively. In contrast, Fig. 3c shows in cyan and magenta wrong matches satisfying the epipolar error constraints correctly discarded by the proposed approach on the basis of the sparse optical flow and the local homographies, respectively.

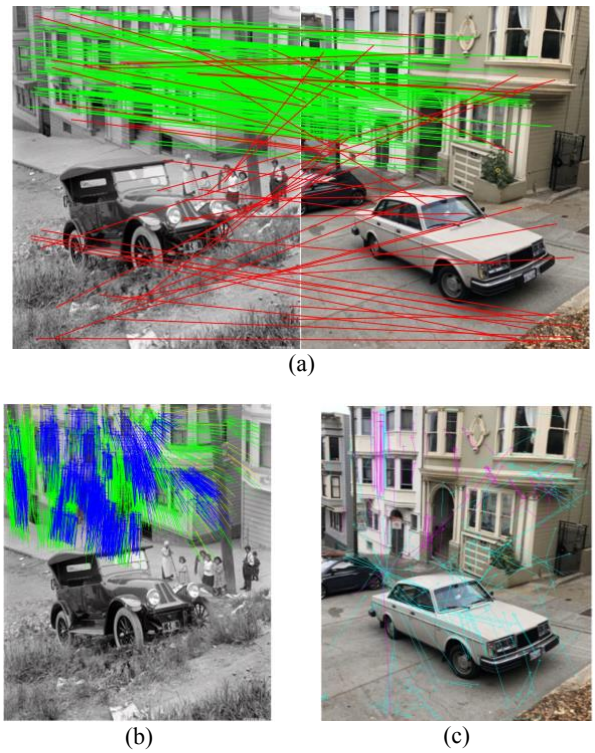

Figure 3: Evaluation on the San Francisco image pair. (a) an example of valid (green) and invalid (red) matches, in the specific case for DISK with DEGENSAC; (b) sparse optical flow for hand-taken ground-truth matches (yellow) and detected correct matches (green, blue); (c) detected wrong matches (cyan, magenta) that also satisfy the epipolar error constraints.

More in detail, the epipolar error threshold was set to 15 px, as well as the threshold to define the optical flow consistency, while local homographies reprojection error was fixed to 5 px. Although these thresholds are relatively high, these aspects should be considered: (i) no camera distortion correction can be trustworthy applied, (ii) sub-pixel accuracy for hand-taken matches is very unlikely and (iii) in the analysis we are primarily interested in the ability to localise corresponding image areas more than the localization accuracy. In Remondino et al. (2021) and Bellavia et al. (2022a) evaluations on the accuracy of keypoint localization with deep-learning local features are presented.
Besides the match ratio, i.e. the number of correct matches with respect to the detected matches, results are also given in terms of absolute correct matches, and match coverage, i.e. how well the matches are distributed over the images. Specifically, the coverage is defined according to the Keypoint Filtering by Coverage (KFC) mask (Bellavia et al., 2022a). The KFC mask is computed by expanding keypoints of correct matches on each image as 31x31 px blocks, so that the final mask is defined as the union of the blocks on each image. The KFC coverage is defined as the maximum ratio over a pair of images between the corresponding masked area and the image area. Notice that the block radius is 15 px as the optical flow and the epipolar errors. The KFC coverage provides a sort of recall accounting for the

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

spatial distribution of the retrieved matches over the images: dense but small clusters of matches, which basically correspond to a same match, weigh less than large but less dense clusters of matches covering a wider area. Further analyses (not reported) show that the KFC coverage is strongly correlated with similar measures that define the mask using more computationally complex area definitions, such that of the convex hull or of the alpha shape of the correct keypoints.

3. RESULTS

Results are reported in Table 1 in terms of KFC coverage, absolute correct matches and match ratio, respectively without or with applying DEGENSAC (Chum et al., 2005) as final step. The aim of employing DEGENSAC is to further refine matches according to global geometric epipolar constraints. This is a widely adopted and suggested practice in the literature (Jin et al., 2021). Using DEGENSAC, it is reasonable to get an increase of the matching ratio with a negligible reduction of the matching coverage. The DEGENSAC threshold was set to 4 px and it is relatively high for the same motivation given for the epipolar error threshold setup in Sec. 2.3. Furthermore, with the exception of LoFTR, SuperGlue, DISK and HP that use ad-hoc matching assignment, NNR ratio is used and reported with different thresholds, or alternatively replaced by the simpler Nearest Neighbour (NN) matching strategy. For the reference SIFT the standard VLFeat implementation, which sets the NNR threshold to 1/1.2=0.85, is used. To further ease the analysis, Figure 4 also reports the results in terms of scatter plots according to the match ratio and KFC coverage, providing a sort of precision-recall plot, with absolute correct matches indicated by the radius of the scatter points.

According to the results, LoFTR, SE2-LoFTR, and SuperGlue provided overall the best-balanced results without or with DEGENSAC. SuperGlue is pushing more on the matching precision while LoFTR on the matching coverage and the amount of the absolute number of correct matches. For all the three methods, there are no relevant differences in using or not DEGENSAC, which implies a robust computation of the matches. Moreover, by inspecting LoFTR results with respect to its rotation invariant version SE2-LoFTR, it is evident that the a-priori knowledge of image orientation can provide a relevant boost to the results, since possible match ambiguities are clearly removed. This is also confirmed by the better results obtained by HP_upright with respect to HP.

Image matching method	KFC coverage without DEGENSAC	KFC coverage with DEGENSAC	correct matches without DEGENSAC	correct matches with DEGENSAC	match ratio % without DEGENSAC	match ratio % with DEGENSAC
SuperGlue	0.108	0.073	131	75	0.71	0.71
LoFTR	0.207	0.175	843	568	0.66	0.63
SE2-LoFTR	0.138	0.106	407	268	0.53	0.50
DISK	0.123	0.062	407	108	0.33	0.33
HP	0.033	0.020	70	23	0.34	0.35
HP_upright	0.053	0.036	141	51	0.48	0.48
VLSIFT	0.007	0.002	6	1	0.05	0.07
ORB - NN	0.028	0.000	34	0	0.02	0.00
ORB - NNR 0.90	0.012	0.000	11	0	0.04	0.00
ORB - NNR 0.80	0.005	0.000	3	0	0.07	0.00
ORB - NNR 0.70	0.001	0.000	1	0	0.12	0.00
RootSIFT - NN	0.079	0.000	107	0	0.04	0.00
RootSIFT - NNR 0.90	0.034	0.000	39	0	0.15	0.00
RootSIFT - NNR 0.80	0.015	0.000	13	0	0.37	0.00
RootSIFT - NNR 0.70	0.004	0.000	4	0	0.79	0.00
ASLFeat - NN	0.116	0.053	244	115	0.27	0.38
ASLFeat - NNR 0.90	0.025	0.011	28	13	0.71	0.33
ASLFeat - NNR 0.80	0.003	0.001	3	1	0.55	0.10
ASLFeat - NNR 0.70	0.001	0.000	0	0	0.20	0.00
ALIKE - NN	0.115	0.068	221	134	0.15	0.29
ALIKE - NNR 0.90	0.021	0.018	24	19	0.37	0.35
ALIKE - NNR 0.80	0.005	0.004	4	3	0.48	0.18
ALIKE - NNR 0.70	0.001	0.000	1	0	0.20	0.00
LFNet - NN	0.021	0.016	19	15	0.03	0.14
LFNet - NNR 0.90	0.007	0.006	6	5	0.13	0.21
LFNet - NNR 0.80	0.002	0.002	2	2	0.24	0.18
LFNet - NNR 0.70	0.001	0.000	1	0	0.20	0.00
R2D2 - NN	0.072	0.030	105	43	0.23	0.36
R2D2 - NNR 0.90	0.013	0.006	11	6	0.64	0.26
R2D2 - NNR 0.80	0.001	0.000	1	0	0.40	0.00
R2D2 - NNR 0.70	0.000	0.000	0	0	0.00	0.00
RoRD - NN	0.059	0.039	93	61	0.09	0.34
RoRD - NNR 0.90	0.007	0.006	7	6	0.34	0.24
RoRD - NNR 0.80	0.001	0.000	1	0	0.10	0.00
RoRD - NNR 0.70	0.000	0.000	0	0	0.00	0.00

Table 1: KFC coverage, absolute correct matches and match ratio for the evaluated image matching methods. The results are averaged on the whole dataset. Increasing better results are highlighted with darker colors.

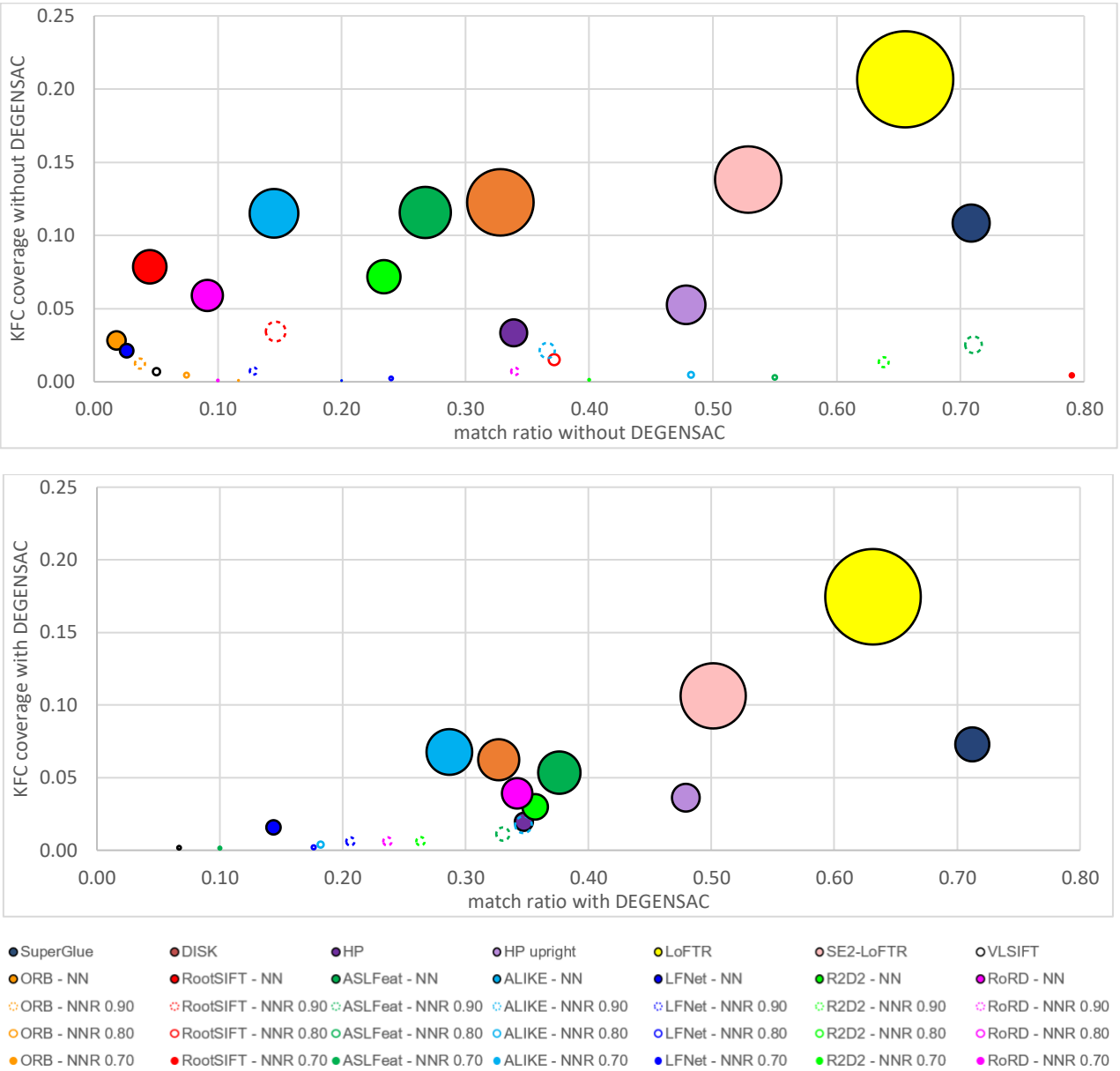

Figure 4: Scatter plots of the evaluated image matching approaches in terms of their match ratio and KFC coverage without (top row) and with (bottom row) DEGENSAC. The radius of each scatter point indicates the number of correct matches for the associated matching method.

For DISK, HP, HP_upright, LoFTR, SE2-LoFTR, and SuperGlue the match ratio remains stable with or without DEGENSAC while the absolute number of correct matches and the KFC coverage decrease. For the remaining methods, refining the initial NN matches by NNR at any threshold level drastically reduces the absolute number of matches and the KFC coverage, but increases the match ratio. Nevertheless, the number of matches becomes too exiguous to be used by DEGENSAC, causing a final decrease of the match ratio. These observations imply that the matched features for these specific image pairs are not so easily disambiguated.

VLSIFT, ORB and RootSIFT, the handcrafted methods included in this evaluation, perform poorly according to any metric employed. With the exception of LF-Net, the remaining deep approaches RoRD, R2D2, ALIKE and ASLFeat with the NN matching strategy followed by DEGENSAC obtain in order better increasing results than the handcrafted methods. For these deep methods avoiding to initially filter matches according to NNR and relying on geometric constraints through DEGENSAC for the match refinement seems the best solution, maybe due to their specific design. HP without DEGENSAC can be grouped with these last deep methods, while DISK and HP_upright, also without DEGENSAC, provide slightly better results with respect to these deep methods when focusing respectively on the KFC coverage and the match ratio.

Finally, Figure 5 shows the matching results of RootSIFT (handcrafted), HP_upright (hybrid), SuperGlue and LoFTR (both deep) on the dataset image pairs, indicating correct and wrong matches respectively in green and red.

4. CONCLUSIONS AND FUTURE WORKS

This work presented an analysis of recent and state-of-the-art image matching approaches to retrieve image correspondences in challenging multi-temporal images. The evaluation was based on the consistency of the epipolar geometry and of a sparse optical flow, both computed according to manually-measured matches. In this way, correct matches can be discriminated correctly and robustly, providing an analysis in terms of the number of correct matches as well as their image coverage.

The results show that recent deep learning-based image matching methods are sufficiently robust to strong radiometric variations due to different sensors and illumination conditions, as well as viewpoint changes. According to the reported evaluations, these methods can be employed on challenging multi-temporal datasets in which traditional image matching methods such as SIFT typically fails, yet the results are still far to equal those obtained in common application scenarios.
As future works, more challenging image pairs will be included in the dataset and alternative evaluation metrics and image matching methods will be investigated and tested.

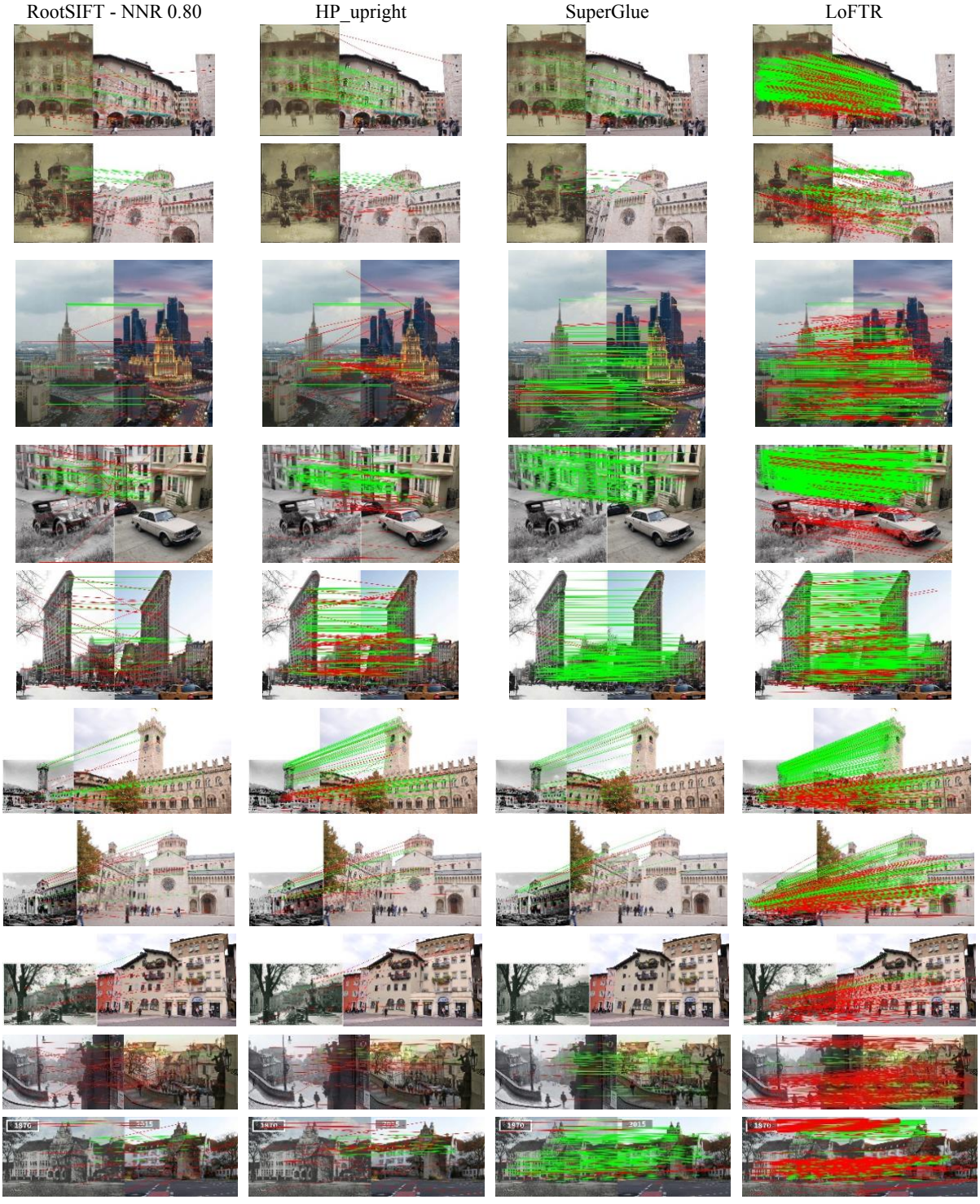

Figure 5: Correct (green) and wrong (red) matches for RootSIFT, HP_upright, SuperGlue and LoFTR on the employed image pairs.

ACKNOLEDGMENTS

This work has been partly supported by the project "AI@TN" funded by the Autonomous Province of Trento, Italy. Authors are also thankful to Fondazione Museo Storico di Trento, Municipality of Trento and Superintendence for Cultural Heritage of the Autonomous Province of Trento for providing the historical images of Trento used in the paper investigations.

REFERENCES

Ali, H.K. and Whitehead, A., 2014. Modern to Historic Image Matching: ORB/SURF an Effective Matching Technique. *Computers and their Applications*.

Arandjelović, R. and Zisserman, A., 2012, June. Three things everyone should know to improve object retrieval. *Proc. IEEE CVPR,* pp. 2911-2918.

Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded Up Robust Features. In *European conference on computer vision*, pp. 404-417. Springer, Berlin, Heidelberg, 2006.

Bellavia, F., Morelli, L., Menna, F. and Remondino, F., 2022a. Image orientation with a hybrid pipeline robust to rotations and wide-baselines. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *46*, pp.73-80.

Bellavia, F., Colombo, C., Morelli, L. and Remondino, F., 2022b. Challenges in image matching for cultural heritage: an overview and perspective. *Proc. FAPER2022, Springer LNCS*.

Bellavia, F. and Mishkin, D., 2022c. HarrisZ+: Harris corner selection for next-gen image matching pipelines. *Pattern Recognition Letters*, *158*, pp.141-147.

Bellavia, F., 2022d. SIFT matching by context exposed. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bökman, G. and Kahl, F., 2022. A case for using rotation invariant features in state of the art feature matchers. *Proc. IEEE CVPR*, pp. 5110-5119.

Brauer-Burchardt, C. and Voss, K., 2002. Facade reconstruction of destroyed buildings using historical photographs. *The International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, *34*(5/C7), pp.543-550.

Chen, Lin, Franz Rottensteiner, and Christian Heipke. "Feature detection and description for image matching: from hand-crafted design to deep learning." *Geo-spatial Information Science* 24, no. 1 (2021): 58-74.

Chum, O., Werner, T. and Matas, J., 2005, June. Two-view geometry estimation unaffected by a dominant plane. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 772-779). IEEE.

DeTone, D., Malisiewicz, T. and Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 224-236).

Farella, E.M., Morelli, L., Remondino, F., Mills, J.P., Haala, N. and Crompvoets, J., 2022. The EUROSDR TIME benchmark for historical aerial images. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*, pp.1175-1182.

Fischler, M.A. and Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), pp.381-395.

Ghuffar, S., Bolch, T., Rupnik, E. and Bhattacharya, A., 2022. A pipeline for automated processing of Corona KH-4 (1962-1972) stereo imagery. *arXiv preprint arXiv:2201.07756*.

Gruen, A., Remondino, F. and Zhang, L., 2004. Photogrammetric reconstruction of the great Buddha of Bamiyan, Afghanistan. *The Photogrammetric Record*, *19*(107), pp.177-199.

Holmlund, E.S., 2021. Aldegondabreen glacier change since 1910 from structure-from-motion photogrammetry of archived terrestrial and aerial photographs: utility of a historic archive to obtain century-scale Svalbard glacier mass losses. *Journal of glaciology*, *67*(261), pp.107-116.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M. and Trulls, E., 2021. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, *129*(2), pp.517-547.

Li, J., Hu, Q. and Ai, M., 2018. RIFT: Multi-modal image matching based on radiation-invariant feature transform. *arXiv preprint arXiv:1804.09493*.

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

Luo, Z., Zhou, L., Bai, X., Chen, H., Zhang, J., Yao, Y., Li, S., Fang, T. and Quan, L., 2020. ASLFeat: Learning local features of accurate shape and localization. *Proc. IEEE CVPR,* pp. 6589-6598.

Ma, J., Jiang, X., Fan, A., Jiang, J. and Yan, J., 2021. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, *129*(1), pp.23-79.

Maiwald, F., 2019a. Generation of a benchmark dataset using historical photographs for an automated evaluation of different feature matching methods. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*, pp.87-94.

Maiwald, F., Bruschke, J., Lehmann, C. and Niebling, F., 2019b. A 4D information system for the exploration of multitemporal images and maps using photogrammetry, web technologies and VR/AR. *Virtual Archaeology Review*, *10*(21), pp.1-13.

Maiwald, F., Lehmann, C. and Lazariv, T., 2021. Fully automated pose estimation of historical images in the context of 4D geographic information systems utilizing machine learning methods. *ISPRS International Journal of Geo-Information*, *10*(11), p.748.

Matas, J., Chum, O., Urban, M. and Pajdla, T., 2004. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, *22*(10), pp.761-767.

Mishkin, D., Radenovic, F. and Matas, J., 2018. Repeatability is not enough: Learning affine regions via discriminability. In *Proc. ECCV,* pp. 284-300.

Nocerino, E., Menna, F. and Remondino, F., 2012a. Multi-temporal analysis of landscapes and urban areas. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *39*, p.B4.

Nocerino, E., Menna, F., Remondino, F., 2012b. GNSS/INS aided precise re-photographing. *Proc. 18th IEEE Intern. Conference on Virtual Systems and MultiMedia* (VSMM), pp. 235-242.

Ono, Y., Trulls, E., Fua, P. and Yi, K.M., 2018. LF-Net: Learning local features from images. *Advances in Neural Information Processing Systems*, *31*.

Parihar, U.S., Gujarathi, A., Mehta, K., Tourani, S., Garg, S., Milford, M. and Krishna, K.M., 2021. RoRD: Rotation-robust descriptors and orthographic views for local feature matching. Proc. *IEEE IROS,* pp. 1593-1600.

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

Revaud, J., Weinzaepfel, P., De Souza, C., Pion, N., Csurka, G., Cabon, Y. and Humenberger, M., 2019. R2D2: repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*.

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

Sarlin, P.E., DeTone, D., Malisiewicz, T. and Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. Proc. *IEEE CVPR*, pp. 4938-4947.

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J. and Kahl, F., 2018. Benchmarking 6DOF outdoor visual localization in changing conditions. Proc. *IEEE CVPR*, pp. 8601-8610.

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

Su, S., Zhao, Z., Fei, Y., Li, S., Chen, Q. and Fan, R., 2022. SIM2E: Benchmarking the Group Equivariant Capability of Correspondence Matching Algorithms. *arXiv preprint arXiv:2208.09896*.

Sun, J., Shen, Z., Wang, Y., Bao, H. and Zhou, X., 2021. LoFTR: Detector-free local feature matching with transformers. Proc. *IEEE CVPR*, pp. 8922-8931.

Torresani, A., Rigon, S., Farella, E.M., Menna, F., Remondino, F., 2021. Unveiling large-scale historical contents with V-SLAM and markerless mobile AR solutions. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *46*, pp.761-768.

Tyszkiewicz, M., Fua, P. and Trulls, E., 2020. DISK: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, *33*, pp.14254-14265.

Verdie, Y., Yi, K., Fua, P. and Lepetit, V., 2015. TILDE: A temporally invariant learned detector. Proc. *IEEE CVPR*, pp. 5279-5288.

Zhang, L., Rupnik, E., Pierrot-Deseilligny, M., 2020. Guided feature matching for multi-epoch historical image blocks pose estimation. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*

Zhang, L., Rupnik, E. and Pierrot-Deseilligny, M., 2021. Feature matching for multi-epoch historical aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *182*, pp.176-189.

Zhao, X., Wu, X., Miao, J., Chen, W., Chen, P.C. and Li, Z., 2022. ALIKE: Accurate and Lightweight Keypoint Detection and Descriptor Extraction. *IEEE Transactions on Multimedia*.