HANDCRAFTED AND LEARNING-BASED TIE POINT FEATURES – COMPARISON USING THE EUROSDR RPAS BENCHMARK DATASET

M. V. Peppa [1], L. Morelli [2], J. P. Mills [1]*, N. T. Penna [1], F. Remondino [2]

[1] School of Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom
– (maria-valasia.peppa, jon.mills, nigel.penna)@newcastle.ac.uk
[2] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Web: http://3dom.fbk.eu – Email: (lmorelli, remondino)@fbk.eu

Commission II, EuroSDR Theme Session

KEY WORDS: Aerial triangulation, Benchmark, CNN, Deep learning, EuroSDR, RPAS, SfM photogrammetry, Tie points

ABSTRACT:

The identification of accurate and reliable image correspondences is fundamental for Structure-from-Motion (SfM) photogrammetry. Alongside handcrafted detectors and descriptors, recent machine learning-based approaches have shown promising results for tie point extraction, demonstrating matching success under strong perspective and illumination changes, and a general increase of tie point multiplicity. Recently, several methods based on convolutional neural networks (CNN) have been proposed, but few tests have yet been performed under real photogrammetric applications and, in particular, on full resolution aerial and RPAS image blocks that require rotationally invariant features. The research reported here compares two handcrafted (Metashape local features and RootSIFT) and two learning-based methods (LFNet and Key.Net) using the previously unused EuroSDR RPAS benchmark datasets. Analysis is conducted with DJI Zenmuse P1 imagery acquired at Wards Hill quarry in Northumberland, UK. The research firstly extracts keypoints using the aforementioned methods, before importing them into COLMAP for incremental reconstruction. The image coordinates of signalised ground control points (GCPs) and independent checkpoints (CPs) are automatically detected using an OpenCV algorithm, and then triangulated for comparison with accurate geometric ground-truth. The tests showed that learning-based local features are capable of outperforming traditional methods in terms of geometric accuracy, but several issues remain: few deep learning local features are trained to be rotation invariant, significant computational resources are required for large format imagery, and poor performance emerged in cases of repetitive patterns.

1. INTRODUCTION

1.1 EuroSDR RPAS Benchmark

A desirable development in any application of airborne photogrammetry is the freeing of processing pipelines from the need for supporting information in terms of ground control points (GCPs) and/or local Global Navigation Satellite System (GNSS) base stations. Reported experiences and professional opinions vary as to what extent this is feasible for high quality geometric survey, but it is clear that its ultimate achievement would bring significant time and cost benefits to National Mapping and Cadastral Agencies (NMCAs). As a result, a EuroSDR benchmark was initiated in 2021 with the aim of evaluating the true geometric quality of real-world survey data generated from Remotely Piloted Aircraft System (RPAS) photogrammetry and LiDAR under different control configurations, focussing primarily on the geometric quality of data generated in the absence of ground control and local GNSS base station information. Further details on the EuroSDR RPAS Benchmark can be found at Geospatial.github (2022).

1.2 Handcrafted versus learning-based feature detection

The requirement for precise, repeatable and stable tie points in order to derive camera pose and sparse 3D representation of a surveyed scene is well understood in Structure-from-Motion (SfM) photogrammetry. However, the rigorous identification of tie points in large image datasets is still an open research topic in the photogrammetric and computer vision communities. Tie points may be established by extracting keypoints using handcrafted feature detector and descriptor methods (e.g. Lowe, 2004; Bay et al., 2006; Bellavia et al., 2021). The scale-invariant feature transform (SIFT) constitutes one of the most well established feature detector and descriptor operators. It detects points of interest in images at the local extremes created by the difference-of-Gaussians smoothing function (Lowe, 2004). Similar to SIFT, various handcrafted feature detectors have been implemented in SfM photogrammetric software packages over the years (e.g. Snavely et al, 2008; Metashape, 2011; Schonberger and Frahm, 2016). Despite the ease and success of their implementation, especially in “black-box” SfM software, the SIFT-like feature detector operates in isolation, disconnected from the entire SfM self-calibrating bundle adjustment pipeline, focusing purely on tie point extraction during the initial step of descriptor matching (Bellavia et al., 2021; Remondino et al., 2021). With recent advances in deep learning technology, solutions based on convolutional neural network (CNN) methods have been proposed (Bellavia et al., 2022a) that jointly train the detectors and descriptor to increase reliability and matching success rate (e.g. DeTone et al., 2018; Dusmanu et al., 2019; Ono et al., 2019; Revaud, 2019; Luo et al. 2020). Other approaches have combined the strengths of state-of-the-art neural networks and traditional handcrafted algorithms (e.g. Schonberger et al., 2017, Jin et al., 2021; Bellavia et al., 2022b).

In parallel, Remondino et al. (2021) compared state-of-the-art handcrafted and learning-based methods for the establishment of tie points in various image datasets. The investigation highlighted the practical challenges for feature matching and evaluated

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-1183-2022 | © Author(s) 2022. CC BY 4.0 License.

selected methods under different acquisition conditions and scene characteristics. Local features were extracted on down sampled images (1500 x 1000 pixels) because of CNN computational constraints. In particular, the performance of eight different methods was evaluated in relation to the number of tie points extracted, time required and root mean square errors (RMSEs) against surveyed independent check points (CPs) utilising RPAS images capture in various network configurations (i.e. nadir / oblique). Among others, the Local Feature Network (LFNet) (Ono et al., 2018), Key.Net (Barroso-Laguna et al., 2019) and SIFT-like variants implemented in a SfM self-calibrating bundle adjustment such as COLMAP (Schonberger and Frahm, 2016) were evaluated. Remondino et al. (2021) reported that learning-based methods provided similar RMSEs to those of the handcrafted solutions, especially in cases of strong imaging network configurations.

The research reported in this paper builds upon the previous investigations and rigorously evaluates the different approaches to extract image correspondences to apply an aerial triangulation using the EuroSDR RPAS benchmark. It is known that RPAS image blocks have many characteristics that can negatively influence learning-based methods, e.g. high resolution, camera rotations, scale changes, etc. This research analyses such characteristics using imagery collected with a top-end DJI RPAS sensor / platform combination, the DJI Zenmuse P1 (DJI P1, 2022), for which performance with handcrafted / learning-based algorithms has not previously been investigated. Therefore, this research demonstrates for the first time the potential and limitations of state-of-the-art deep learning neural networks implemented with the latest DJI RPAS datasets.

2. STUDY SITE AND RPAS DATASET ACQUISITION

2.1 Study site

Guided by a task force of NMCA experts and academics, in August 2021 a coordinated test field of GCPs, CPs, test objects and profiles was established at Wards Hill quarry in Northumberland, UK. The quarry was actively producing limestone in the 1920s. Nowadays, the site is privately owned and primarily used by the owners for livestock grazing. The study site (Figure 1) has an extent of 350 x 250 m and a c. 40 m ground lowering where the limestone was quarried. The site is mainly vegetated with coarse grass as well as occasional shrubs and small trees.

The established test field consisted of 51 CP and eight GCP targets. An overview plan of the test field layout can be seen in Figure 1. The CP test field approximates to a 6 (labelled A-F) x 8 (labelled 1-8) array, with CPs randomly placed in each grid square and identified using an alphanumeric code according to their position in the field (e.g. A4). Three supplementary targets were added to densify the test field in the base of the quarry and were labelled S1, S2 and S3. The eight GCPs were labelled with a prefix of G, followed by their position in the 6 x 8 array - GA1, GA5, GA8, GC6, GD4, GF1, GF4 and GF8.

All targets were circular Perspex disks of 300 mm diameter with a 100 mm diameter black centre to aid pointing. CP targets were fabricated in white Perspex and GCP targets in yellow (Figure 2). In order to aid identification in RPAS LiDAR datasets (not used herein), and to keep clear of low-lying vegetation, targets were mounted in the field on wooden stakes that were driven into the ground and secured with a single screw through the centre of the target. A spirit level was used to approximately level the surface of the circular targets during setting out. The height of the target planes above ground level ranged from c. 0.15 m to c. 0.45 m. Figure 3 shows an example of a target (CP A7) set-up in the field. It should be noted that targets were marked with both a white A4 clipboard ID and a red / white ranging rod, which may have been vertical or horizontal at the time of data acquisition, to support target identification during processing and location in the field.

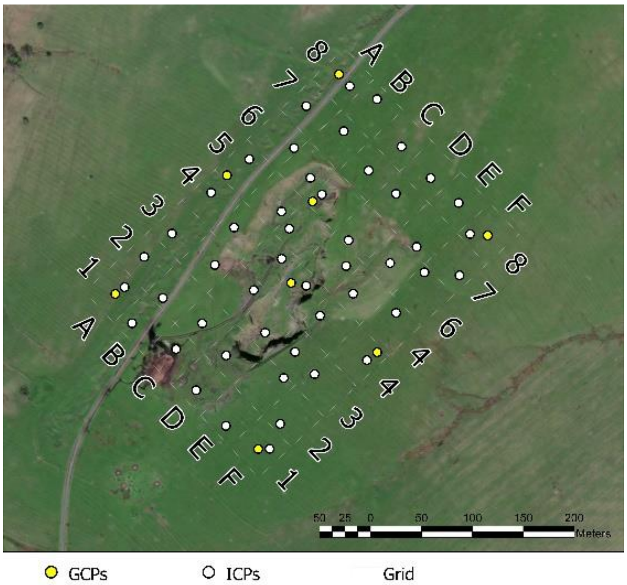


Figure 1. Test field target distribution at Wards Hill quarry, Northumberland, UK, Map data: Google Earth 2021.

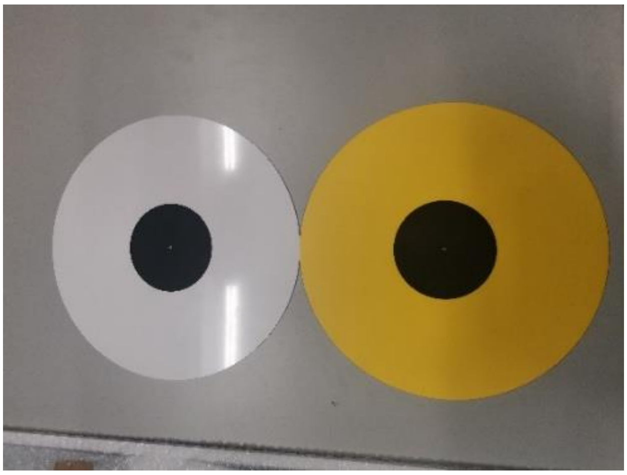


Figure 2. 300 mm diameter targets - CPs (l) and GCPs (r).

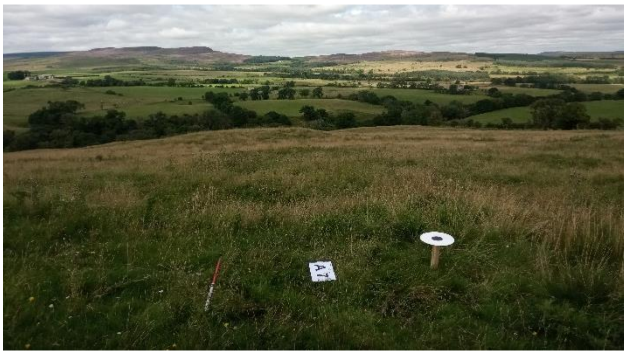


Figure 3. CP A7 located in the field and marked with a clipboard and ranging rod.

Four Global Navigation Satellite System (GNSS) reference stations were established at the study site and surveyed during the EuroSDR RPAS Benchmark field campaign (22nd to 26th August 2021, inclusive). Stations were observed in GNSS static mode, delivering sub-cm level 3D accuracy relative to a local base station in Ordnance Survey Great Britain 36 (OSGB36). Details on ETRS89 to OSGB36 coordinate transformation can be found in Ordnance Survey (2020). The GNSS observations from the reference stations were used to calculate the coordinates of the 59 targets. Three Leica GS18 GNSS receivers in "Static and kinematic" mode were used to collect data during four separate occupations. The three antennas were mounted on bipods with all antenna heights set to 1.800 m. Three minutes of 'static' data were collected at targets in the east and west of the study site, with 5 minutes of 'static' data collected at targets in the central area, which were generally in the quarry itself and hence subject to reduced sky visibility. The receivers / antennas were held upright when moving between the different stations, with kinematic data logging continuously at 1 Hz throughout in order to ensure long data arcs were recorded. This enabled improved ambiguity resolution rather than only recording GNSS data when set up for 3-5 minutes on the targets alone. All GNSS processing was undertaken using Leica Infinity 3.5.0. The average standard deviations of all calculated targets' coordinates across the four occupations were 5.3 mm in Easting, 6.1 mm in Northing and 3.4 mm in height.

2.2 RPAS dataset acquisition

The study site was simultaneously surveyed using a number of different RPAS mounted instruments, each limited to a single survey flight to represent "real-world" operation. The research reported in this paper utilised a DJI Zenmuse P1 (DJI P1, 2022) dataset acquired using a DJI Matrice 300 RPAS platform (Heliguy, 2022), as shown in Figure 4. The DJI Zenmuse P1 carries a 45 megapixel full-frame sensor (35.9 x 24 mm), with 4.4 µm nominal pixel size and the DJI DL F2.8 LS ASPH lens with a 35 mm nominal focal length (DJI P1, 2022). The DJI Zenmuse P1 sensor is mounted on a 3-axis stabilised gimbaled system, which is fixed on the DJI Matrice 300 RTK RPAS platform (DJI M300, 2022).

Figure 4. DJI Zenmuse P1 M300 RPAS data acquisition.

The DJI Zenmuse P1 flight was conducted by Heliguy (2022) on 23rd August 2022, at a 50 m height above ground level at an aircraft speed of 5 m/s. The DJI Zenmuse P1 sensor was set up with an automatic camera exposure and a continuous focus, capturing images with an 80% forward and 70% lateral overlap. It should be noted that the RTK link was disabled, as this dataset was used for Phase 1 of the EuroSDR RPAS Benchmark, where no GCPs or GNSS base station information were made available (Geospatial.github, 2022). A total of 974 nadir-looking images were collected from 24 parallel flight lines and 25 oblique images were captured with a 45º off-nadir angle. To reduce the computational power while maintaining a high image overlap (Figure 5), a subset of the full DJI P1 image network, corresponding to 423 images, was utilised in the analysis presented here. From the full image network, one nadir-looking image was selected every two images along the flight lines and all oblique images and those at the edges of the study site were disregarded. The resulting ground sampling distance (GSD) of the photogrammetric block was 7 mm.

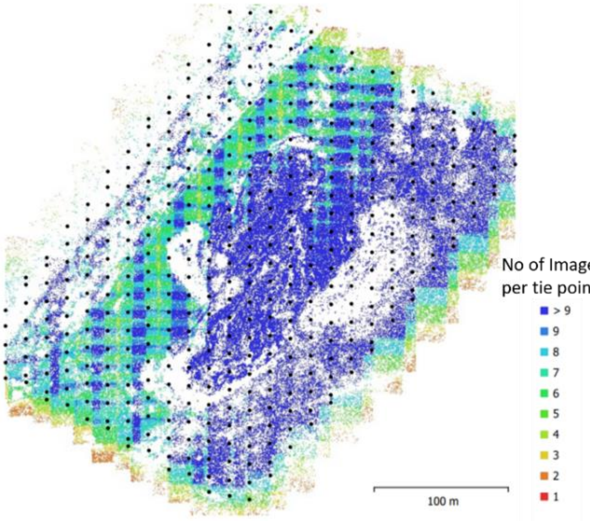


Figure 5. Image overlap for 423 DJI P1 images with 507,580 tie points as estimated in Metashape (see Table 2).

3. METHODOLOGY

The methodological workflow consists of two main stages. In stage 1, a circular black and white (b/w) target detection algorithm was developed to automatically estimate the centre of the circular targets in image space, which were then used as input in the following stage. Stage 2 evaluates two handcrafted, Metashape SIFT and RootSIFT (Arandjelović and Zisserman, 2012), and two learning-based methods, LFNet and Key.Net, estimating cameras interior orientation (IOP) and exterior orientation (EOP) parameters as well as RMSEs on targets.

3.1 Stage 1: Circular b/w target detection

In stage 1 the automated circular b/w target detection comprises established image processing and computer vision (OpenCV, 2015) routines developed in Python (Burger and Burge, 2009). Every RPAS image was converted to grayscale, then a Gaussian Blur was applied using a 7 x 7 convolution kernel. A Canny edge detector (Canny, 1986) was applied with minimum and maximum threshold intensity values of 253 and 255, respectively. After image processing, contour extraction was applied to the binary Canny image. To remove noise at this point of the workflow, detected contours of any shape with areas of 20 pixels or less were filtered and removed. The area bounded by the closed contour was calculated in OpenCV according to Green's formula (Green, 2008). For the remaining contours detected in each image, a best fitting circle was used as an approximation to the contour shape and its radius was calculated. Only circles with radii within the range 1-100 pixels were used, while other circles were removed. The aforementioned settings and criteria were specified after fine-tuning each step via a "trial and error" procedure.

To check that the detected circles from the contour function correspond to the actual circular b/w targets, a Hough circle transform was also applied with four different sets of parameters, as listed in Table 1.

Hough circle transform setups	Parameter 1	Parameter 2	Min radius [pix]	Max radius [pix]
1	1	10	1	10
2	1	20	1	20
3	1	30	1	30
4	20	30	20	30

Table 1. Parameterisation setup for Hough circle transform in OpenCV after a "trial and error" procedure.

Parameters 1 and 2 refer to threshold limits for edge detection and centre of the circle estimation, respectively. The minimum and maximum radii correspond to the range of the circular target size. The parameters in Table 1 were set up after a "trial and error" procedure using a small subset of the 423 RPAS images. To specify the parameters, the 7 mm GSD that was estimated via the SfM pipeline in Metashape (Table 3), and the minimum 100 mm and maximum 300 mm diameter of the target's physical size (Figure 2) were considered. The settings shown in Table 1 were found to provide optimal results, detecting either the inner black and / or the outer white circle of the signalised targets. A final criterion compared the coordinates of the centres detected using contours and those using a Hough circle transform. If differences of the centre coordinates in the x and y-axis in image space were greater than 2 pixels, then the detected circles were disregarded. This condition allowed for filtering out erroneous detections such as rounded shapes with similar brightness to the targets. The final coordinates of each target's centre were extracted from the fitted circle that met the aforementioned 2- pixel condition.

To assess the accuracy of the automated circular b/w target detection algorithm, the centre of the targets was also manually marked in Metashape, following common practice often adopted in the SfM photogrammetric pipeline. After resolving an initial relative orientation of the image block, epipolar lines per stereo image pair supported the manual identification of each target centre. Target labels were also manually set in Metashape. The x, y coordinates in image space of each target's centre visible in all images were extracted from Metashape and compared against those estimated from the automated target detection algorithm. To ensure a consistent comparison, 0.5 pixels were added to the x and y image coordinates of the OpenCV algorithm to align with the interior coordinate system of Metashape. The coordinate system in Metashape has its origin in the middle of the top-left pixel with (x, y) equal to (0.5, 0.5) (Metashape, 2022).

3.2 Stage 2: Aerial triangulation with handcrafted and learning-based feature detection

In stage 2, aerial triangulation was undertaken using tie points obtained by traditional handcrafted features as well as with learning-based methods. The limitations of handcrafted local features in dealing with strong illumination and perspective changes have been addressed in recent years by new detectors and descriptors based on CNNs. Among several approaches, end-to-end methods jointly train the detector and the descriptor, sharing the computations either partially or completely. The basic idea is to minimise a cost function, maximising the discriminability of the descriptor where the network extracts the keypoints. However, these methods are often not designed to handle rotations, a property that is currently little investigated in the computer vision community since many of the datasets used for training do not contain such images. On the contrary, in aerial photogrammetry, rotation invariance is fundamental in order to handle the 180-degree inversion of the photographic sensor between nadir strips. Therefore, despite the abundance of learning-based methods, those that are rotation invariant are few in number.

Among the end-to-end methods available, it was chosen to test LFNet, where the architecture respects the classical pipeline based on detection, patch orientation estimation, and patch description, but the training process is unique. Superpoint (DeTone et al., 2018) has some rotation invariance, but extracts a limited number of keypoints which makes it unsuitable for aerial datasets (Remondino et al., 2021). RoRD (Parihar et al., 2021) is one of the few rotationally invariant end-to-end methods for which source code is available, and it is planned to extend the analyses reported in this paper to it in the future. No other learning-based end-to-end methods are believed to currently be available for testing.

Another approach is the detect-then-describe method that can combine different detectors and descriptors, both learning-based and handcrafted. This study tests Key.Net+AffNet+HardNet8, in the following reported simply as Key.Net, that has performed well in previous evaluations. Key.Net is the detector that is almost completely learning-based, apart from the first few layers that are handcrafted, while HardNet8, an evolution of the original HardNet (Mishchuk et al., 2017), is a learning-based descriptor for which invariance to rotation is guaranteed by AffNet (Mishkin et al., 2018).

In addition to rotation invariance, the other major limitation of learning-based methods is the high computational demand that does not permit processing of large-format images. To overcome this limitation, it was decided to tile the images at full resolution (8192 x 5460 pixels) into tiles of 2500 x 2500 pixels, from which keypoints and descriptors are first extracted using the methods described above. Images are then reassembled to form a unique database with the keypoints appropriately translated in a unique 2D reference system, together with their associated descriptors. To further optimize computation times, a maximum of 10200 keypoints per image were extracted, without significantly affecting the final accuracy of the model, as demonstrated in the results section.

While COLMAP was used for incremental reconstruction, for the matching step it was not possible to use the matcher integrated in the software, optimized for the use of the GPU, since the LFNet and Key.Net descriptors have a size of 256 and 128 respectively, while COLMAP only manages SIFT-like descriptors with 128 parameters. Therefore, the OpenCV-Python matcher was adopted following a brute-force approach with a matching filter that uses cross-check and ratio-test. The values of the ratio-test thresholds were 0.80, 0.85, and 1.00 for RootSIFT, HarNet, and LFNet, respectively. The values chosen are those proposed in the literature, apart from LFNet which, using the value 0.95 (Jin et al., 2020), loses most of the matches. The matches were imported into COLMAP for geometric verification and image orientation. After assessing different camera models, the "OPENCV" camera model was found to be the most appropriate for the lens used: focal length (fx, fy), principal point (cx, cy), two radial distortion (k1, k2), and two tangential distortion (p1, p2) parameters.

Knowing the orientation parameters of the cameras, it was possible to triangulate forward the targets with the "point_triangulator" COLMAP API. Finally, the model was scaled, rotated and translated through a Helmert transformation using all available targets. This operating mode was chosen since COLMAP does not allow users to use points as control points, i.e. by adding constraints in the bundle adjustment. To validate the comparisons made in COLMAP and for double-checking the results obtained, independent processing was carried out with Metashape, whose local features approach is unknown.

4. RESULTS

4.1 Stage 1: Circular b/w target detection

Out of all 423 RPAS images, 331 images included targets. In those 331 images, a total number of 461 correct target detections (i.e. true positives) were extracted from the automated OpenCV algorithm, whereas the manual target identification in Metashape resulted in 542 target detections (Table 2).

Target detection method	Total detections	True positives	False positives	False negatives
Metashape (manual)	542	542	0	0
OpenCV (automated)	1379	461	81	837

Table 2. Output of the automated circular target detection OpenCV algorithm with comparison against Metashape output.

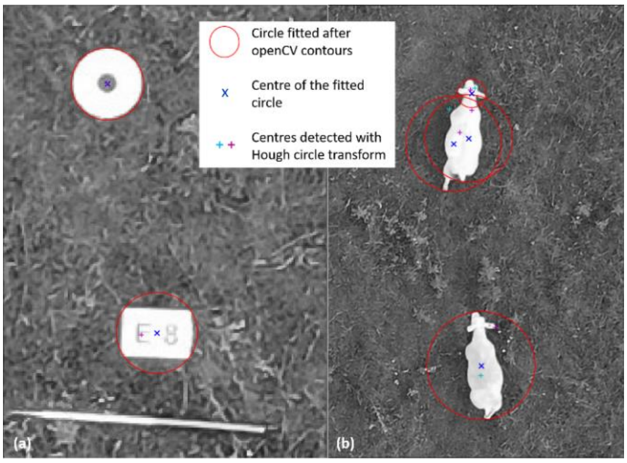

Circle fitted after openCV contours
x Centre of the fitted circle
++ Centres detected with Hough circle transform
E8
(a)
(b)

Figure 6. Examples of correct and erroneous target detections, the latter removed from the final outputs via the developed automated OpenCV workflow.

Figure 6a (top left) shows an example of a true positive, as detected using the OpenCV contour formula with its fitted circle in red and its centre marked in blue “x” symbol. The same target was detected with the Hough circle transform and the estimated location of its centre coincided with the one identified with OpenCV contours, hence passed the criterion of the 2 pixel coordinate difference. As evidenced in Figure 6a and b, OpenCV contours erroneously identified shapes other than circles. Since detections from the Hough circle transform at those locations had their centre locations further than 2 pixels from the centres extracted with OpenCV contours, these were automatically disregarded.

As reported in Table 2, the 81 false positives obtained with the OpenCV workflow were missed possibly due to that fact that the Hough circle transform algorithm was finely tuned to recognise targets primarily viewed from above. There were few targets located at the edges of the images that were seen from an oblique angle as depicted in Figure 7, which could not be located either with OpenCV contours or with the Hough circle transform. Whilst the automated OpenCV algorithm missed only 81 targets from the 542 total detections, it produced a relatively high noise with 837 false negatives (Table 2). The resulting noise is attributed to the fact that both the OpenCV contours and Hough circle transform algorithm were finely tuned to identify as many circles as possible, therefore the specified settings (e.g. Table 1) were possibly too sensitive to noise.

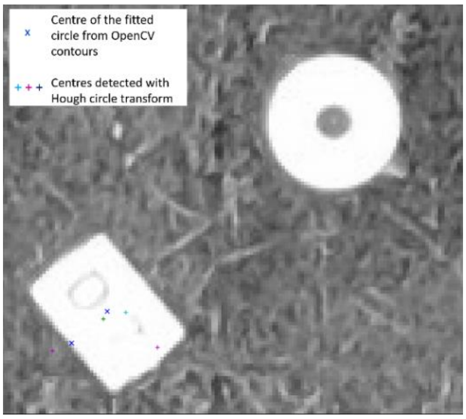

x Centre of the fitted circle from OpenCV contours
+++ Centres detected with Hough circle transform

Figure 7. Example of false positives mis-detected with the OpenCV algorithm due to a relatively oblique viewing angle.

4.2 Stage 2: Aerial triangulation with handcrafted and learning-based feature detection

The quantitative comparisons between the handcrafted and learning-based local features are based on several statistics that COLMAP (Schönberger et al., 2016) provides downstream from the SfM pipeline, and the RMSE on the residuals calculated on all available targets obtained from the Helmert transformation, as reported in Table 3. The COLMAP statistics include the number of points in the sparse cloud, the mean track length (MTL), and the mean reprojection error (MRE). The MTL in particular provides the redundancy of the observations that enter the bundle adjustment, while the MRE, although important to be minimized during the adjustment, is not able to provide an estimate of the geometric accuracy of the model alone (Remondino et al., 2021).

In this comparison the reference model is the RootSIFT model run in COLMAP, its results reported in row (e) of Table 3. To verify the quality of this reference, it was compared with a model obtained from Metashape setting the alignment parameters similarly to those of COLMAP: high accuracy (i.e. keypoints extracted on full size images), the key point limit set to 10200, the same camera parameters of the OpenCV camera model (fx, fy, cx, cy, k1, k2, p1, p2), and not using the targets as constraints in the bundle adjustment (i.e. no camera model optimization). The resulting RMSE at the CPs for COLMAP is 4.7 cm, quite similar to the 4.2 cm reported by Metashape in row (d), demonstrating that the COLMAP pipeline is equivalent to that of Metashape under the same operating conditions.

Since the RMSE appeared to be quite high when considering the GSD of 7 mm, this value was found to be related to the lack of radial distortion factor k3 in the camera model. In fact, if parameter k3 is added in Metashape (keeping all other parameters of model (d) equal), an RMSE equal to 1.8 cm is obtained (model (c) in Table 3). Unfortunately, COLMAP does not allow the addition of the k3 parameter, so comparison with the learning-based methods was conducted with the OpenCV camera model.
In (b) the photogrammetric model is still not optimized and has been scaled with only six targets used to optimize the camera model in (a). The optimized model (a) has been reported for completeness, as it represents the procedure in aerial photogrammetry where greater accuracy is achieved when a high quality ground survey is available in support, as in this case, with standard deviations of a few millimetres. Finally, note how between (b) and (a) the number of extracted keypoints increases from 10200 to 50000 without a significant increase in accuracy.

With the RootSIFT model cross-validated, it is now possible to analyse the results of the learning-based methods. LFNet, model (f), could not identify sufficient valid tie points to be able to register the entire block of images correctly, registering only 213/423 images. However, Key.Net, model (g), managed to orientate the entire block, obtaining an RMSE of 3.7 cm, slightly better than that of both RootSIFT and Metashape. It is also interesting to note the 50% increase in the MTL of Key.Net over the other methods (4.5 versus a MTL of 3 for others). Finally, it should be noted that there is no direct relationship between the values for MRE and those of the RMSE, as already previously reported in Remondino et al. (2021) and Bellavia et al. (2022).

It is also worthwhile to dwell on why LFNet fails to record all images. First of all, it should be taken into account that the value recommended by Jin et al. (2020) for Key.Net + HardNet is 0.95, based on the descriptor distribution. With this threshold, for certain images, LFNet could not even hold a match. The reason is that the images in this dataset are challenging with patches very similar to each other (grass fields), while the LFNet descriptor is not very discriminative, an assertion demonstrated both by the high ratio threshold value of 0.95 and by the test reported in Figure 9. In Figure 9 it is possible to compare the matches before the geometric verification, and it is clear how LFNet contains a high number of outliers probably linked to the low discriminability of its descriptor.

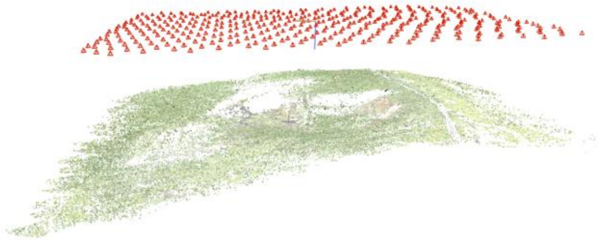


Figure 8. Sparse cloud calculated with COLMAP and Key.Net local features.

5. CONCLUSIONS AND FUTURE WORK

The paper has introduced the EuroSDR RPAS benchmark datasets and its activities related to the image orientation based on learning-based tie points. In stage 1, the automated OpenCV algorithm has provided an independent method of target detection that does not rely on the “black-box” SfM photogrammetric software packages such as Metashape, while eliminating the time consuming and labour intensive task of manual inspection. However, to reduce the high number of false negatives and remove the noisy results, investigations to apply and incorporate RANSAC filtering within the automated OpenCV algorithm is under development.

In stage 2, the investigation has focused on the rigorous evaluation of object space accuracy arising from triangulation using tie points derived from handcrafted and learning-based feature extraction methods for aerial photogrammetric surveys. First of all, this work has set itself the goal of pushing previously reported evaluations to more extensive datasets, at full resolution, focusing on RPAS images that require local features able to manage the rotations of the sensor during flights. The large format images have been processed by dividing the original images into tiles computationally manageable by the learning-based local features. Rotation invariance can only be managed by networks trained ad-hoc for this specific task. It is emphasized that learning-based methods which are invariant to rotation are currently lacking in the literature, a very limited category if compared to the numerous methods proposed in recent years that are not invariant to rotations. Furthermore, all these methods remain slow in terms of the time taken for feature extraction, while demanding high computational resources.

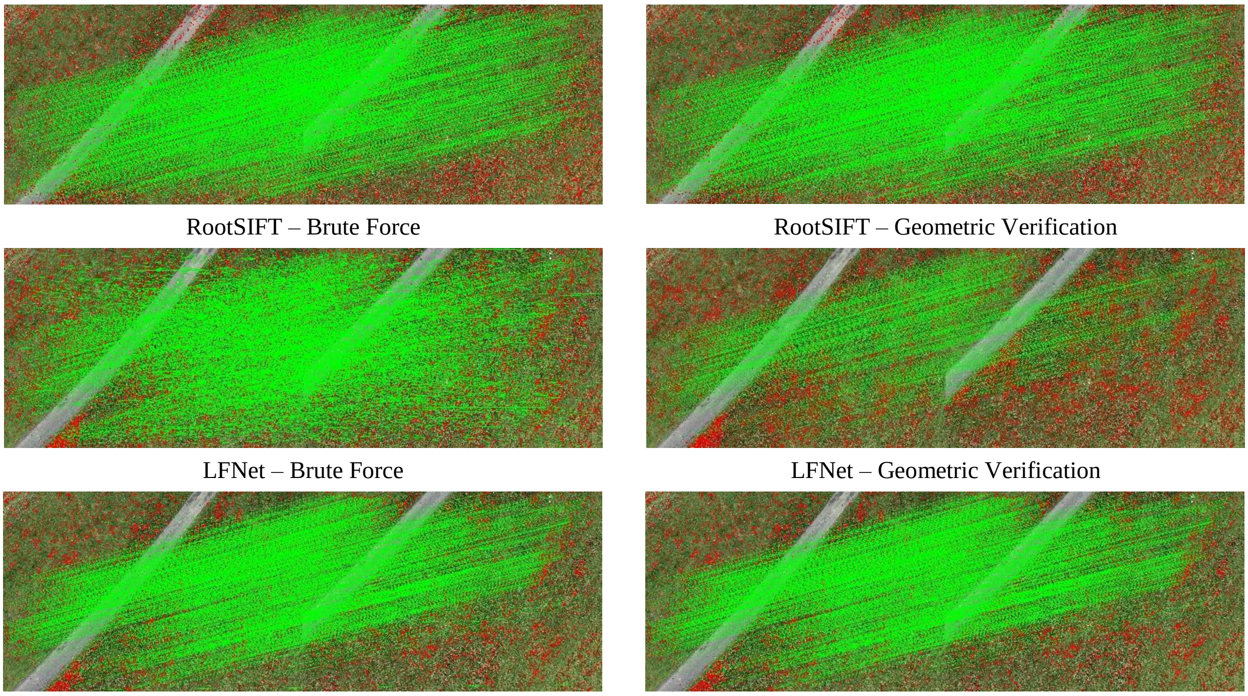
Among the tested methods, LFNet, which represents the family of end-to-end methods proven not to have very discriminative descriptors, a factor which in this dataset is amplified by the presence of repetitive textures. On the contrary, Key.Net demonstrated slightly better results than RootSIFT and Metashape, demonstrating the ability of these new methods to compete with well-established handcrafted methods such as RootSIFT. Furthermore, Key.Net displayed a significantly higher MTL than the other methods, an interesting feature for more difficult datasets, for example with little overlap between images, or where there are strong variations in the image content and in the radiometric distribution, such as imagery from multi-temporal datasets. In the future, extensive tests on multi-temporal aerial datasets will be conducted in order to take full advantage of the robustness of the learning-based descriptors on imagery with significant changes in radiometric content.

ACKNOWLEDGEMENTS

The authors acknowledge EuroSDR for supporting and funding the RPAS Benchmark. Items of equipment used by Newcastle University researchers in the benchmark field campaign were funded by UKCRIC - UK Collaboratorium for Research in Infrastructure & Cities: Newcastle Laboratories (EPSRC award EP/R010102/1). Thanks are also due to both DJI and Heliguy for the provision of the DJI P1 RPAS Benchmark dataset used in the study reported in this paper. Many thanks to James Goodyear, Nathan Harrap, Less-lee Davidson and Craig Robson for fieldwork assistance.

	Local features	Optimization Camera Model	# kpts # tie pts	RMSE [m] on GCPs	RMSE [m] on CPs	# GCPs	# CPs	#3D points	MTL	MRE [pix]
(a)	Metashape	Optimized OpenCV + k3	50000 no limits	0.006	0.014	7	49	507580	3.1	0.20
(b)	Metashape	Not optimized OpenCV + k3	10200 10200	0.019	0.019	7	49	223399	3.0	0.31
(c)	Metashape	Not optimized OpenCV + k3	10200 10200	-	0.018	0	56	223399	3.0	0.31
(d)	Metashape	Not optimized OpenCV	10200 10200	-	0.042	0	56	223557	3.0	0.33
(e)	RootSIFT	Not optimized OpenCV	10200 10200	-	0.047	0	56	187815	3.0	0.57
(f)	LFNet	Not optimized OpenCV	10200 10200	-	incomplete registration	0	56	29065	3.1	0.36
(g)	Key.Net	Not optimized OpenCV	10200 10200	-	0.037	0	56	208803	4.5	0.81

Table 3. Overall comparison between handcrafted and learning-based tie point extraction methods. In each self-calibrating bundle adjustment the adopted cameral model followed the OpenCV convention (fx, fy, cx, cy, k1, k2, p1, p2).



RootSIFT – Brute Force

RootSIFT – Geometric Verification

LFNet – Brute Force

LFNet – Geometric Verification

KN+HN – Brute Force

KN+HN – Geometric Verification

Figure 9. Image matching comparison before and after geometric verification.

REFERENCES

Arandjelovi´ c, R., Zisserman, A., 2012. Three things everyone should know to improve object retrieval. Proc. *IEEE CVPR*, 2911–2918.

Barroso-Laguna, A., Riba, E., Ponsa, D., Mikolajczyk, K., 2019 Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. Proc. *ICCV*.

Bay, H., Tuytelaars, T., Gool, L.V., 2006. SURF: Speeded-Up Robust Features. Proc. *ECCV*, pp. 404-417.

Bellavia, F., 2021. SIFT matching by context exposed. *arXiv*: 2106.09584.

Bellavia, F., Colombo, C., Morelli, L., Remondino, F., 2022a. Challenges in image matching for cultural heritage: an overview and perspective. Proc. 2nd Int. workshop on Fine Art Pattern Extraction and Recognition (FAPER2022), Springer LNCS, in press.

Bellavia, F., Morelli, L., Menna, F. and Remondino, F., 2022b. 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.

Burger, W., Burge, M.J., 2009. Principles of Digital Image Processing: Core Algorithms, Undergraduate Topics in Computer Science. Springer-Verlag, London, pp. 5-26.

Canny, J., 1986. A computational approach to edge detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679-698.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. SuperPoint: Self-Supervised Interest Point Detection and Description. Proc. *IEEE CVPR.*

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A. and Sattler, T., 2019. D2-net: A trainable CNN for joint description and detection of local features. Proc. *IEEE CVPR*, pp. 8092-8101.

DJI M300, 2022. https://www.dji.com/uk/matrice-300/specs, last accessed 20/03/22.

DJI P1, 2022. https://www.dji.com/uk/zenmuse-p1, last accessed 14/01/22.

Geospatial.github, 2022. https://geospatialncl.github.io/eurosdr-rpas-benchmark/, last accessed 14/01/22.

Green G., 2008. An Essay on the Application of mathematical Analysis to the theories of Electricity and Magnetism. *arXiv: 0807.0088v1*.

Heliguy, 2022. https://www.heliguy.com/blogs/posts/eurosdr-drone-mapping-survey-accuracy-project, last accessed 17/01/2022.

Jin, Y., Mishkin, D., Mishchuk, A. et al., 2020. Image Matching Across Wide Baselines: From Paper to Practice. *Int Journal of Computer Vision*, Vol. 129, pp. 517-547.

Lowe, D.G., 2004. Distinctive image features from scale invariant keypoints. *Int. Journal of Computer Vision*, 60(2).

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.

Metashape, 2011. Algorithms used in Agisoft Metashape, https://www.agisoft.com/forum/index.php?topic=89.msg323#msg323, last accessed 15/03/2022.

Metashape, 2022. Agisoft Metashape User Manual, https://www.agisoft.com/pdf/metashape-pro_1_8_en.pdf, last accessed 15/03/2022.

Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J., 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. Proc. *NeurIPS*, 48294840.

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

Ono, Y., Trulls, E., Fua, P., Yi, K.M., 2019. LFNet: Learning local features from images. Proc. *NeurIPS*.

OpenCV, 2015. Open Source Computer Vision Library. https://Opencv.org/, last accessed 01/12/2022.

Ordnance Survey, 2020. A guide to Coordinate Systems in Great Britain. https://www.ordnancesurvey.co.uk/documents/resources/guide-coordinate-systems-great-britain.pdf, last accessed 01/12/2022.

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. *IEEE/RSJ IROS*, pp. 1593-1600.

Remondino, F., Menna, F., and Morelli, L., 2021. Evaluating handcrafted and learning-based features for photogrammetric applications. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., *XLIII-B2-2021*, 549–556.

Revaud, J., Weinzaepfel, P., de Souza, C.R., Humenberger, M., 2019. R2D2: Repeatable and Reliable Detector and Descriptor. Proc. *NIPS*.

Schonberger, J. L. and Frahm, J. M., 2016. Structure-from-Motion revisited. Proc. *IEEE CVPR*.

Schonberger, J. L., Hardmeier, H., Sattler, T., Pollefeys, M., 2017. Comparative evaluation of handcrafted and learned local features. Proc. *IEEE CVPR*, pp. 1482-1491.

Snavely, N., Seitz, S.M. and Szeliski, R., 2008. 'Modeling the world from Internet photo collections', *International Journal of Computer Vision*, 80(2), pp. 189-210.