Comparative Assessment of Image Processing Algorithms for the Pose Estimation of Uncooperative Spacecraft

Pasqualetto Cassinis, Lorenzo; Fonod, R.; Gill, Eberhard; Ahrns, Ingo; Gil Fernandez, Jesus

Publication date
2019

Document Version
Final published version

Published in
International Workshop on Satellite Constellations and Formation Flying

Citation (APA)
Pasqualetto Cassinis, L., Fonod, R., Gill, E., Ahrns, I., & Gil Fernandez, J. (2019). Comparative Assessment of Image Processing Algorithms for the Pose Estimation of Uncooperative Spacecraft. In *International Workshop on Satellite Constellations and Formation Flying: 16-19 July, Glasgow, Uk* Article IWSCFF 19-43.

COMPARATIVE ASSESSMENT OF IMAGE PROCESSING ALGORITHMS FOR THE POSE ESTIMATION OF UNCOOPERATIVE SPACECRAFT

Lorenzo Pasqualetto Cassinis[*], Robert Fonod[†], Eberhard Gill[‡], Ingo Ahrns[§], and Jesus Gil Fernandez[¶]

This paper reports on a comparative assessment of Image Processing (IP) techniques for the relative pose estimation of uncooperative spacecraft with a monocular camera. Currently, keypoints-based algorithms suffer from partial occlusion of the target, as well as from the different illumination conditions between the required offline database and the query space image. Besides, algorithms based on corners/edges detection are highly sensitive to adverse illumination conditions in orbit. An evaluation of the critical aspects of these two methods is provided with the aim of comparing their performance under changing illumination conditions and varying views between the camera and the target. Five different keypoints-based methods are compared to assess the robustness of feature matching. Furthermore, a method based on corners extraction from the lines detected by the Hough Transform is proposed and evaluated. Finally, a novel method, based on an hourglass Convolutional Neural Network (CNN) architecture, is proposed to improve the robustness of the IP during partial occlusion of the target as well as during feature tracking. It is expected that the results of this work will help assessing the robustness of keypoints-based, corners/edges-based, and CNN-based algorithms within the IP prior to the relative pose estimation.

INTRODUCTION

In the past years, advancements in the field of Distributed Space Systems have been made to cope with the increasing demand for robust and reliable engineering solutions in challenging scenarios for Guidance, Navigation, and Control (GNC), such as Formation Flying missions, In-Orbit Servicing (IOS), and Active Debris Removal (ADR).

The estimation of the position and attitude (pose) of an inactive target by an active servicer spacecraft is a critical task in the design of current and planned space missions, due to its relevance for close-proximity operations, i.e. ADR and/or IOS. Relative pose estimation systems based solely on a monocular camera are recently becoming an attractive alternative to systems

[*]PhD candidate, Department of Space Engineering, Delft University of Technology, Kluyverveg 1 2629 HS, Delft, The Netherlands, *L.C.PasqualettoCassinis@tudelft.nl.*

[†]Assistant Professor, Department of Space Engineering, Delft University of Technology, Kluyverveg 1 2629 HS, Delft, The Netherlands, *R.Fonod@tudelft.nl.*

[‡]Professor, Department of Space Engineering, Delft University of Technology, Kluyverveg 1 2629 HS, Delft, The Netherlands, *E.K.A.Gill@tudelft.nl.*

[§]Engineer, Space Robotics Projects, Airbus DS GmbH, Airbusallee 1 28199, Bremen, Germany, *ingo.ahrns@airbus.com.*

[¶]Engineer, TEC-ECN, ESTEC, Keplerlaan 1 2201 AZ, Noordwijk, The Netherlands, *Jesus.Gil.Fernandez@esa.int.*

based on active sensors or stereo cameras, due to their reduced mass, power consumption and system complexity.

Monocular pose estimation consists in estimating the relative pose of a target spacecraft with respect to the servicer spacecraft by only using 2D images, either taken by a monocular camera or fused from more monocular cameras, as measurements. More specifically, *model-based* monocular pose estimation methods receive as input a 2D image and match it with an existing wireframe 3D model of the target spacecraft to estimate the pose of such target with respect to the servicer camera. Key features are extracted from the 2D image and matched to the corresponding elements of the 3D model.

The Image Processing (IP) system is a fundamental step for feature-based pose estimation, and several methods exist in literature to extract and detect target features from a monocular 2D image, based on the specific application. Interested readers are referred to Ref. 1 for a detailed overview of current IP algorithms. From a high-level perspective, the target features to be detected can be divided into corners, edges, keypoints (or interest points), and depth maps.

Generally, corners and edges are preferred features over keypoints when there are challenges in extracting textured features from the image due to noise and/or poorly textured targets. Reference 2 selected the Canny edge detection algorithm to detect edges in the image, and opted for a subsequent Hough Transform (HT)[3] to extract the detected lines. Several tests were conducted to assess the robustness of the IP with respect to image noise at different variance levels. However, a limitation of their method was that it focused on the extraction of rectangular structures on a large target spacecraft. The same feature detection and extraction methods were used in Ref. 4 in combination with a Low-Pass Filter. Their method was tested with the PRISMA image dataset and proved to be flexible with respect to the spacecraft shape, but it lacked of robustness to illumination and background conditions. Furthermore, the method lacked of robustness with respect to the spacecraft symmetry. More recently, a novel technique to eliminate the background of images was proposed in Ref. 5, called Weak Gradient Elimination (WGE). In this method, thresholding is performed based on the weak gradient intensities corresponding to the Earth in the background. In the next step, the Sobel algorithm and the HT (S/HT) are used to extract and detect features. By creating two parallel processing flows, the method proved to be able to extract main body features as well as particular structures such as antennas, and thus to solve the symmetry ambiguity which characterized other IP schemes. Furthermore, the implementation of the WGE method returned a much higher robustness with respect to Earth in the background compared to the other methods. However, scenarios in which the Earth horizon is present in the background represented a challenge for the IP. Furthermore, the robustness of the method against adverse illumination conditions was not fully proven. Alternatively, Ref. 6 introduced a new IP scheme in which three different parallel processing streams, which use the Shi-Tommasi (ST) corners detector, the HT, and the Line Segment Detector (LSD), are exploited in order to filter three sets of points and improve the robustness of the feature detection. This was performed in order to overcome the different drawbacks of each single method. Feature fusion was then used to synthesise the detected points into polylines which resemble parts of the spacecraft body. By including a background removal step similar to the WGE,[5] which makes use of a Prewitt operator in combination with a gradient filter, the authors could demonstrate the robustness of their IP with respect to the Earth's horizon in the background. However, despite the validation of the algorithm with actual space images from the PRISMA mission as well as adverse illumination conditions, the performance of feature extraction methods in terms of accuracy, ro-

bustness and sensitivity was not fully proven. Moreover, the repeatability of the detected features through a sequence of images, which is essential for feature tracking, was not assessed.

If keypoint features are preferred over corners and edges, several methods exist to perform feature detection and extraction. A Scale Invariant Feature Transform (SIFT),[7] in combination with the Binary Robust Independent Elementary Features (BRIEF) method,[8] was implemented in Ref. 9 to extract the target interest points. Reference 10 investigated the performance of several keypoint detectors applied to VIS/TIR synthetic images. In their work, the combination of the Fast-Hessian feature detector with the Binary Robust Invariant Scalable Keypoints (BRISK) descriptor proved to have comparable performance in both spectra, resulting in a promising option when reduced memory usage represent a key requirement. Furthermore, a Fast Retina Keypoint (FREAK) descriptor was adopted in Ref. 11 in combination with the Edge Drawing Lines (EDL) detector to extract keypoints, corners, and edges to find the correspondence between features. In their method, a depth mapping was further performed which aided the features extraction. The limitation of keypoint-based methods is that they require an offline database for image matching. In this context, an extensive analysis on the impact of different camera views as well as different illumination conditions between the offline database and the 2D image acquired in orbit is still not present.

Recently, some authors investigated the capability of Convolutional Neural Networks (CNN) to perform keypoint localization prior to the actual pose estimation.[12,13] The output of these networks is a set of so called *heatmaps* around pre-trained features. The coordinates of the heatmap's peak intensity characterize the predicted feature location, with the intensity indicating the network confidence of locating the corresponding keypoint at the correct location.[12] Notably, this statistical information is not included in the other feature method, and it can be used to threshold the detected features based on the detection accuracy which characterises each single feature. Additionally, due to the fact that the trainable features can be selected offline prior to the training, the matching of the extracted feature points with the features of the 3D wireframe model can be performed without the need of a large search space for the image-model correspondences, which usually characterizes most of the edges/corners-based methods.[4] However, there is yet no implementation of CNNs for feature extraction from space imagery.

As a general remark, IP algorithms based on keypoint features detectors present some advantages compared to algorithms based on edge and corner detectors, given their invariance to perspective, scale, and (to some extent) illumination changes.[7,14] However, they could still be sensitive to extreme illumination scenarios as well as to partial occlusion of some of the target features. On the other hand, despite a lower performance under adverse illumination, edges and corners detectors are retained to be more robust than features detectors in case of partial occlusion of the target, especially during tracking.[15]

In this framework, the aim of this paper is to provide a comparative assessment of past and novel IP algorithms for the relative pose estimation of an uncooperative spacecraft. Special focus is given to the robustness of the selected methods against adverse illumination conditions, partial occlusions of the target, and feature tracking. In the context of edges/corners detection and extraction, a novel method is assessed in which the edges extracted by the HT are post-processed by identifying the intersections between two close lines and associating corner points to such intersections. The performance of such method during feature tracking is assessed. Besides, several keypoint detectors and descriptors are compared by assessing their performance during varying illumination scenarios. Finally, a CNN architecture, based on the work conducted

in Ref. 12, is proposed. The performance of this CNN-based method is assessed during feature tracking, partial occlusions of the target and adverse illuminations. Results are compared to the other IP methods. The comparative assessment is performed on representative synthetic images of the European Space Agency (ESA)'s Envisat spacecraft, rendered in the Cinema 4D software.

The paper is organized as follows. Firstly, a detailed overview of the IP methods is provided together with a description of the framework adopted for the comparative assessment. Preliminary results are then reported for each method separately. Finally, the main conclusions and recommendations are provided.

REVIEW OF IP METHODS

Edges/corners Detectors

From a high-level perspective, the detection of edges and corners can be summarized as follows:

1. Image segmentation
2. Feature detection
3. Feature extraction

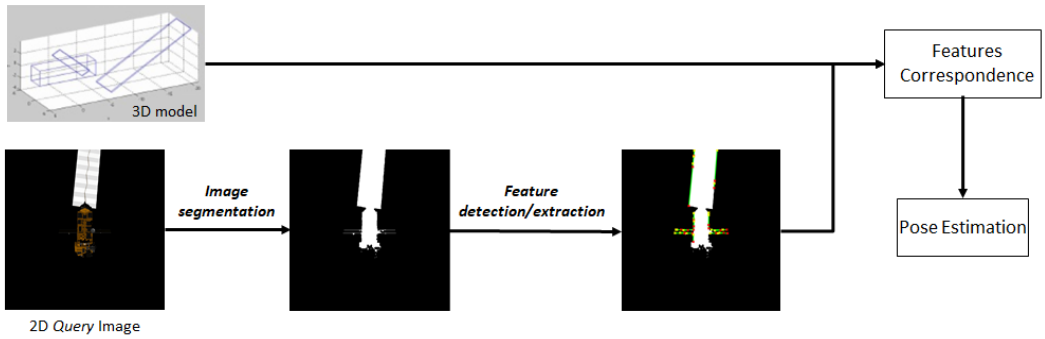


Figure 1. Generic schematic of a feature detection/extraction flow.

Figure 1 shows a high-level representation of such scheme. Image segmentation aims at simplifying the representation of an image to be handled by the edges and corners detectors. The segmentation process can be divided into two main steps. Firstly, thresholding and binarization are applied to convert the pixel image into a binary image. Secondly, dilation and erosion are used to fill holes and recover the image contour. Generally, WGE is further considered in the thresholding step to filter the image background in case of Earth-in-the-background geometries. The benefits of image segmentation stand in an improved detection of edges and corners. After segmentation, features are detected by corner and/or line detectors. Finally, feature extraction is perform to extract the location of the detected features. The Canny algorithm and the HT are considered in this paper for the detection and extraction of edges from the target image, respectively. The input to the Canny edge detector is a monocular image after the image segmentation process. The algorithm computes the intensity gradients of the image and detects

horizontal, vertical and diagonal edges. This information is then used by the HT which identifies and extracts lines from the image. In the proposed method, the detected lines are finally post-processed to extract corners from their intersection. More specifically, two Hough lines are replaced by a corner feature if the minimum relative distance between their ending points is less than a predefined threshold, herewith set to be 5 pixels.

Keypoints Detectors and Descriptors

Compared to edges/corners detectors, keypoint detectors identify points whose nearby region in the image is rich in terms of textures. As such, keypoints are usually stable under illumination/brightness variations. As opposed to other methods based on edges/corners, the application of keypoint detectors prior to pose estimation can be summarized in the following main steps:

1. Keypoints detection
2. Keypoints description and matching
3. Outliers removal with RANdom SAmple Consensus (RANSAC)[16]

Figure 2 shows a high-level representation of the feature matching process.

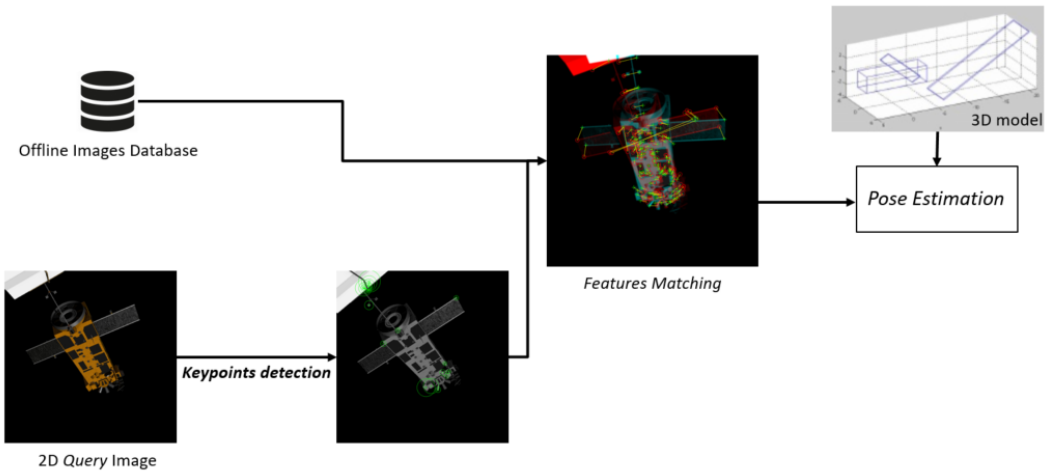


Figure 2. Generic schematic of a feature matching flow.

The keypoints detected in a so called *query* image, taken in orbit, are matched with keypoints extracted offline from an image database by comparing their associated descriptors. These descriptors are functions which describe a small patch around each keypoint. This matching step is not present in edges/corners detectors, and it represents a key task prior to the pose estimation. In fact, the offline database should generally cover as many camera views of the target as possible, in order to guarantee a reliable match, while at the same time it should contain as less images as possible, in order not to make the matching computationally expensive. Once the query features are matched with the offline features, the RANSAC algorithm is used to remove potential outliers in the matching. By using the intuition that outliers will lead to a wrong transformation

between two images, this iterative method randomly selects a minimum set of matches which can return a transformation between two images, and returns a set of inliers based on the transformation error that follows from the chosen set. The size of the minimum set of matches used to initialize RANSAC depends on the type of method used to compute the error. In this paper, the computation of the fundamental matrix between two images is adopted. This transformation is preferred over i.e. homography due to the non-planar motion of the target between two images, which usually occurs in close-proximity operations around uncooperative spacecraft.

For each match between point x in image 1 and point x' in image 2, the fundamental matrix F is the matrix which is related to the so called *epipolar constraint*:

$$x^T F x' = 0 \rightarrow Af = 0, \tag{1}$$

where A is built from the minimum set of matches and f is a vector which contains all the elements of F. More specifically, Eqn. 1 exploits the fact that, for a moving camera, a point in image 2 has its corresponding point in image 1 which lies on a so called epipolar line, whose equation is represented by $l = Fx'$. A minimum of eight matches are needed to solve the epipolar constraint. Equation 1 can be solved by the 8-points algorithm.[17]

CNN Hourglass Architecture

CNNs are currently emerging as a promising features extraction method, mostly due to the capability of their convolutional layers to extract high-level features of objects with improved robustness against image noise and illumination condition. In order to optimize CNNs for the features extraction process, a stacked hourglass architecture has been proposed in Ref. 12, and other architectures such as the U-net, or variants of the hourglass, were tested in recent years. Compared to the network proposed in Ref. 12, the proposed *single hourglass architecture* is composed of only one encoder/decoder block, constituting a single hourglass module. The encoder includes six blocks, each including a convolutional layer, a batch normalization module and max pooling, whereas the six decoder blocks accomodate an up-sampling block in spite of the max pooling one. Each convolutional layer is formed by a fixed number of filter kernels of size 3x3. In the current analysis, 128 kernels are considered per convolutional layer. Figure 3 shows the high-level architecture of the network layers, together with the corresponding input and output.

As already mentioned in the Introduction, the output of the network is a set of heatmaps around the selected features. These heatmaps are extracted by associating them to Gaussian-like profiles. Ideally, the heatmap's peak intensity associated to a wrong detection should be relatively small compared to the correctly detected features, highlighting that the network is not confident about that particular wrongly-detected feature. At the same time, the heatmap's amplitude should provide an additional insight into the confidence level of each detection, a large amplitude being related to large uncertainty about the detection. As such, thresholding based on the relative peak intensity and amplitude could be performed to eliminate the wrong features and improve the accuracy of the detection.

The network is trained with the x- and y- image coordinates of the feature points, computed offline based on the intrinsic camera parameters as well as on the feature coordinates in the target body frame, which were extracted from the 3D CAD model prior to the training. During the training, the network is optimized to locate 16 features of the Envisat spacecraft, consisting of

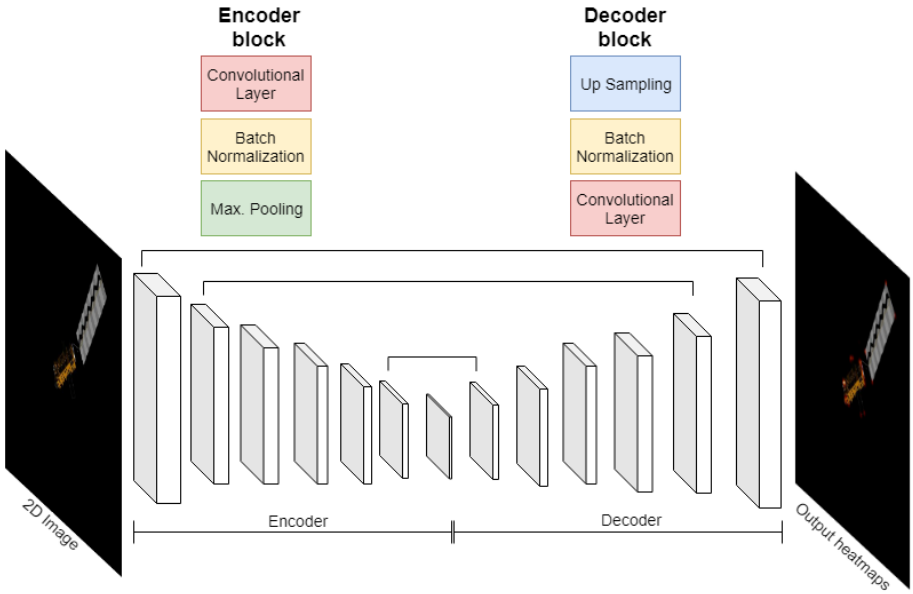


Figure 3. Overview of the complex hourglass architecture. Downsampling is performed in the *encoder* stage, in which the image size is decrease after each block, whereas upsampling occurs in the *decoder* stage. The output of the network consists of heatmap responses, and is used for keypoints localization.

the corners of the main body, the Synthetic-Aperture Radar (SAR) antenna, and the solar panel, respectively. Figure 4 illustrates the ground truth features location for a specific target pose.

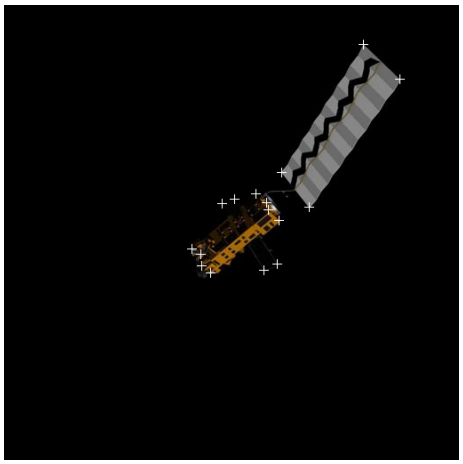


Figure 4. Example of ground truth of features coordinates.

FRAMEWORK OF COMPARATIVE ASSESSMENT

Synthetic images of the Envisat spacecraft were rendered in the Cinema 4D software. Table 1 lists the main camera parameters adopted in the rendering. Firstly, a yaw rotation of the target of 2 deg/s is rendered to assess the tracking performance of each method under varying illumination conditions, assuming a constant distance between the servicer and the target. In this

Table 1. Parameters of the camera used to generate the synthetic images in Cinema 4D.

Parameter	Value	Unit
Image resolution	512×512	pixels
Focal length	$3.9{\cdot}10^{-3}$	m
Pixel size	$1.1{\cdot}10^{-5}$	m

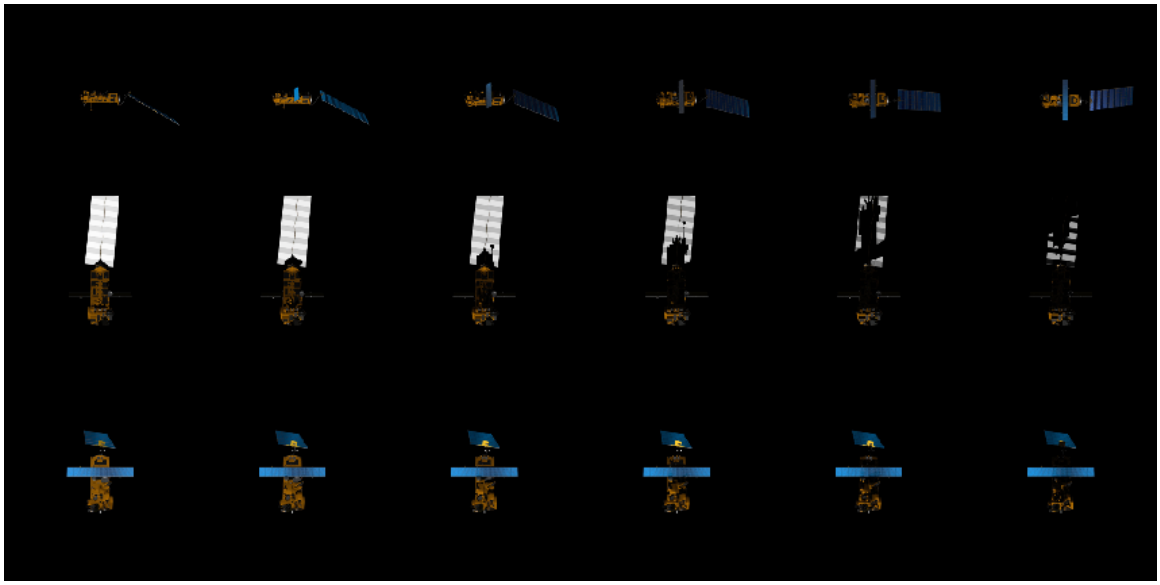


Figure 5. Target spacecraft simulated in Cinema 4D under varying illumination conditions - Yaw (top), illumination 1 (middle), and Illumination 2 (bottom) scenarios. The yaw sequence is shown every 10 frames in order to visualize the target rotation.

case, the different illumination conditions between two subsequent frames are a result of the changing orientation of the target assembly with respect to the Sun. Secondly, an analysis is conducted to assess the impact of varying the Sun position with respect to the target on feature extraction. To accomplish that, two different relative pose of the target spacecraft with respect to the camera are rendered in Cinema 4D under multiple Sun-target geometries. A total of 1,296 images are rendered for each of the two scenarios, by varying the Sun elevation and azimuth by 10 degrees each to span a full sphere. The initial sequences of both the yaw scenario and the two illumination scenarios considered in this paper are represented in Figure 5.

Edges/Corners Detectors

The feature tracking performance of the proposed HT corner detector is assessed by computing the number of tracked features between two subsequent frames for all three scenarios. More specifically, a feature is considered to have been tracked from image 1 to image 2 if the minimum distance between such feature in image 1 and all the features in image 2 is less than a threshold, herewith arbitrary chosen to be five pixels. This is done since the maximum distance which can

Table 2. Feature matching methods and their detectors/descriptors.

Method	Detector	Descriptor	Descriptor Type	# Descriptor Elements
ORB	oFAST	rBRIEF	Binary	256
F-Hessian+SURF	F-Hessian	SURF	Floating Point	64
FAST+FREAK	FAST	FREAK	Binary	512
F-Hessian+FREAK	F-Hessian	FREAK	Binary	512
BRISK	-	-	Binary	512

occur between points of a match pair is associated to a small inter-frame rotation (yaw scenario) or to no inter-frame rotation (illumination scenarios).

Feature Matching

The analysis is conducted by using the 8-points algorithm to detect inliers between two different images. Different feature detectors and descriptors are considered, in order to compare the performance of different matching methods. Table 2 lists each method. Aside from some of the algorithms mentioned in the Introduction, the Oriented FAST, Rotated BRIEF (ORB)[18] and the Speeded Up Robust Features (SURF)[14] algorithms are also considered.

For each match, the True Positive (TP) inliers are extracted from the total number of inliers by computing the 2-norm between the matched points and by checking if it is below a predefined threshold, selected in the same fashion as for the edges/corners detectors. In the illumination scenarios, feature matching is performed between the first image and each subsequent ones, whereas in the tracking scenario the matching is performed between two subsequent frames. In both scenarios, the performance is assessed in terms of matching precision, defined as the ratio of TP inliers to the total number of inliers, which also accounts for False Positives (FP),

$$\text{precision} = \frac{TP}{(TP + FP)} \tag{2}$$

CNN Hourglass Network

For the training, validation, and test databases, constant Sun elevation and azimuth angles of 30 degrees were chosen in order to recreate favourable as well as adverse illumination conditions. Relative distances between camera and target were chosen in the interval 90 m - 180 m. Relative attitudes were generated by discretizing the yaw, pitch, and roll angles of the target with respect to the camera by 10 degrees each. The resulting database was then shuffled to randomize the images, and was ultimately split into training (40,000 images), validation (4,000 images), and test (~10,000 images) databases. Figure 6 shows some of the images included in the training dataset.

During training, the validation dataset is used beside the training one to compute the validation losses and avoid overfitting. A learning rate of 0.1 is selected together with a batch size of 20 im-

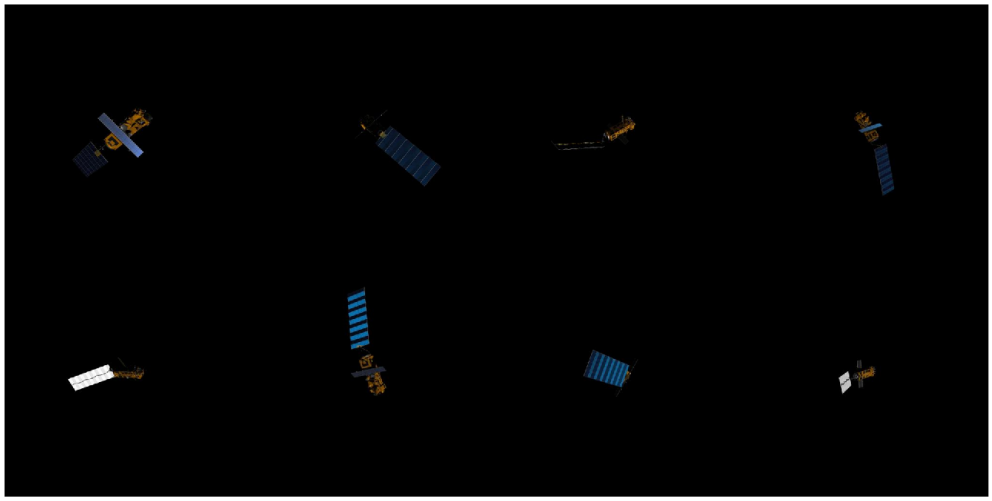


Figure 6. A montage of eight synthetic images selected from the training set.

ages and a total number of 20 epochs. Finally, the network performance after training is assessed with the test database.

The performance of the hourglass network during test is assessed in terms of Root Mean Squared Error (RMSE) between the ground truth (GT) and the x, y coordinates of the extracted features, which is computed as:

$$E_{\text{RMSE}} = \sqrt{\frac{\sum_{i=1}^{n_{tot}} \left[(x_{GT,i} - x_i)^2 + (y_{GT,i} - y_i)^2 \right]}{n_{tot}}} \tag{3}$$

where $n_{tot} = 16$ represents the total amount of features that need to be detected per image. The predicted feature coordinates are obtained from each heatmap's peak location. Besides, the network performance during the yaw and illumination scenarios is assessed to verify the capability of the network to perform feature tracking.

PRELIMINARY RESULTS

Edges/Corners Detector

Figure 7 shows the performance of the proposed corner detector during the initial sequences of the yaw rotation and the two illumination scenarios, respectively. This qualitatively shows the challenges of tracking the Envisat corners for the above scenarios, as the detected corners seem quite unstable and highly sensitive to the rotation of the target as well as to changing illumination conditions. Besides, Figure 8 illustrates the number of tracked features between two subsequent frames for the full yaw scenario and for the initial sequence of the illumination scenarios. As can be seen, the number of tracked features varies considerably during the sequences, suggesting

the lack of robustness of the proposed method against inter-frame rotation of the target and illumination changes.

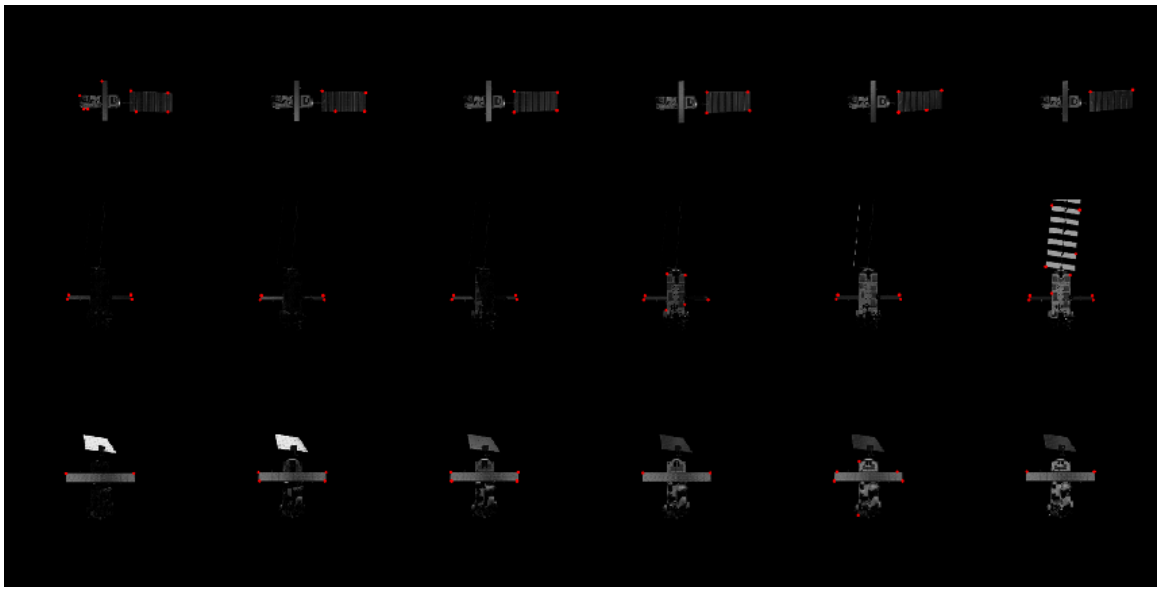


Figure 7. Qualitative feature tracking performance for the proposed corner detector.

Keypoints Detectors and Descriptors

Figure 9 shows the number of detected TP inliers for the yaw scenario. As can be seen, the ORB algorithm outperforms the other methods for the entire query sequence. Its large number of inliers is, to some extent, to be associated to the method itself, which, for any feature matching between two images, always aims at detecting up to 500 matches (see Ref. 18). Besides, the small inter-frame motion between two subsequent image does not seem to severely affect the number of TP inliers for all the methods. This suggests that all the methods can return reliable matches, provided that the camera view does not change considerably between two frames.

Conversely, Figures 10 and 11 illustrate the impact of changing illumination conditions on the number of detected TP inliers and the associated precision, respectively, for the initial sequence of the two recreated illumination scenarios. As can be seen in Figure 10, due to large illumination changes, in the first illumination scenario all methods fail in returning inliers already from the 10th image of the initial sequence. On the other hand, in the second scenario, some of the methods manage to output a constant number of TP inliers throughout the whole initial sequence, as a result of the less adverse illumination changes. In particular, SURF and BRISK seem to guarantee nonzero inliers for almost the whole sequence. However, the precision percentages reported in Figure 11 indicate that the TP inliers detected by SURF are generally less than the FP inliers, whereas the precision for BRISK maintains well above 50% throughout the whole sequence.

The analysis is then extended to cover the all the illumination changes. Table 3 reports the matching performance for the entire sequences of 1,295 images. The percentage of valid matches, defined as the scenarios in which a method manages to return nonzero TP inliers, are reported together with the mean TP inliers and the mean precision.

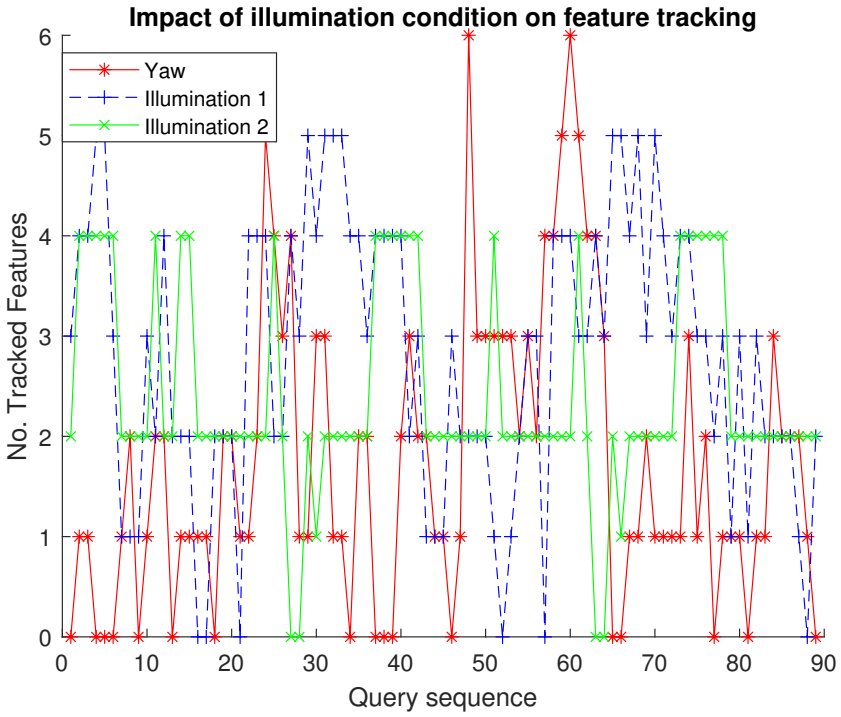


Figure 8. Feature tracking performance for the proposed corner detector.

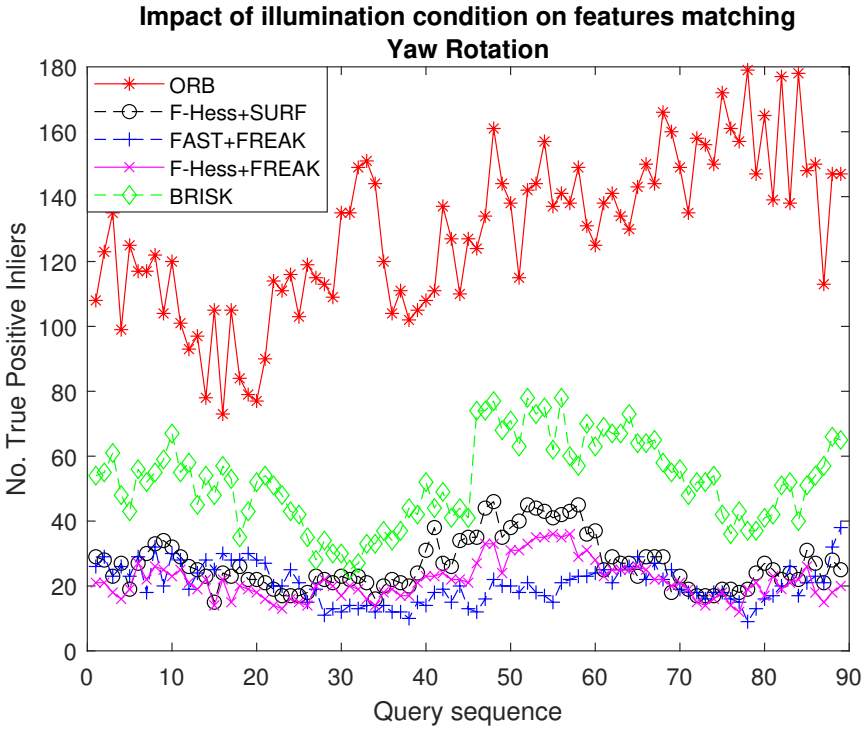


Figure 9. Impact of target rotation on features matching.

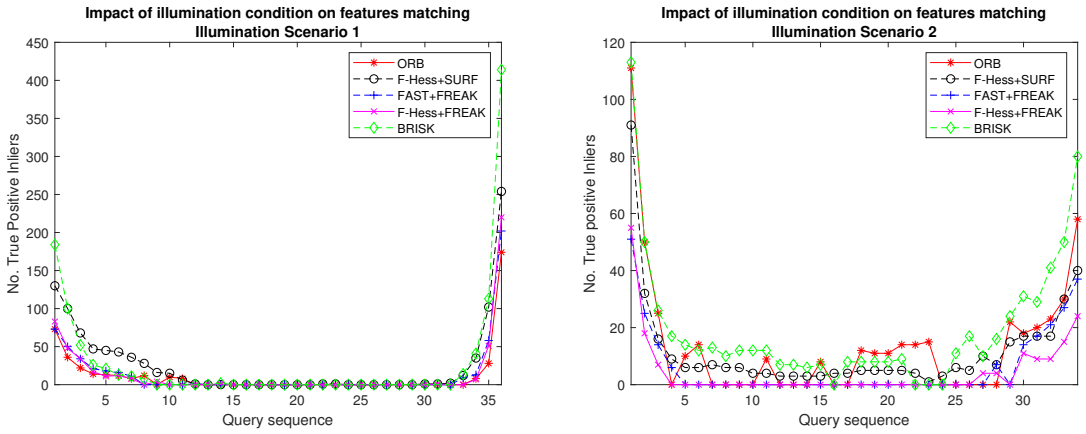


Figure 10. Impact of varying Sun illumination on feature matching performance in terms of detected TP inliers.

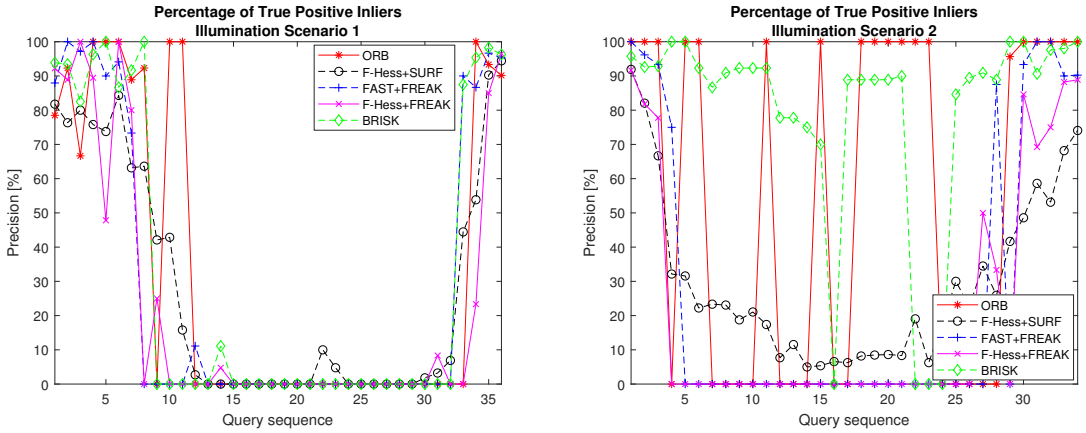
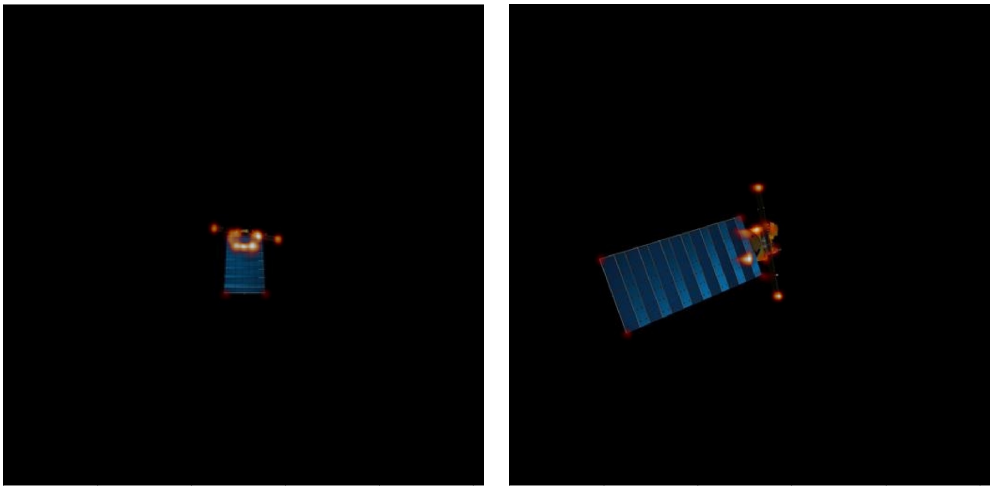


Figure 11. Impact of varying Sun illumination on feature matching performance in terms of precision.

Table 3. Feature matching results for illumination scenario 1/2.

Method	# Valid Matches (%)	Mean # TP Inliers	Mean Precision [%]
ORB	28/51	10/11	28/49
F-Hess+SURF	60/100	18/10	22/30
FAST+FREAK	23/20	6/3	21/18
F-Hessian+FREAK	32/48	9/3	20/19
BRISK	26/96	13/17	25/87



(a) Occulted spacecraft body (b) Non visibility of SAR antenna

Figure 12. Robustness of feature detection with CNNs. The network has been trained to recognize the features pattern, and can correctly locate the body features which are not visible, i.e. parts occulted by the solar panel and corners of the SAR antenna.

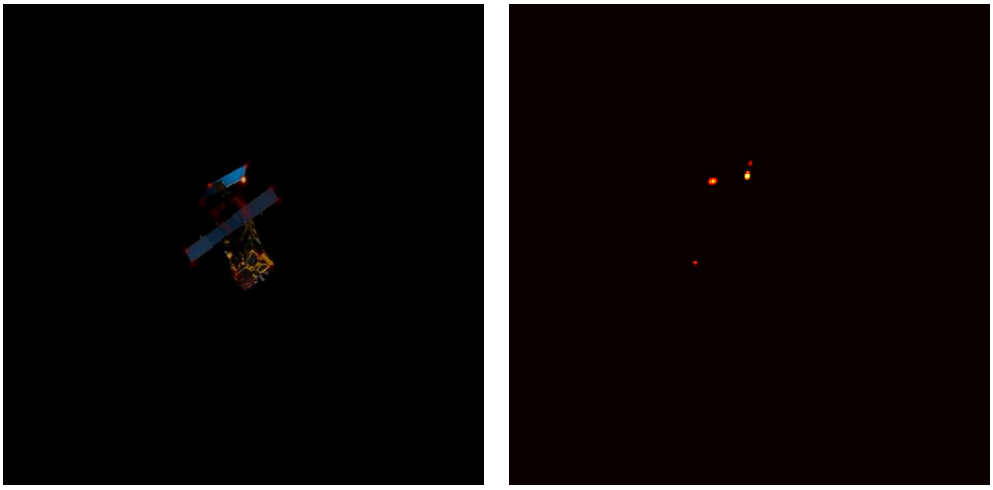
Despite the limited simulated scenarios considered for keypoints detection, the above analysis indicates that a careful selection of the offline database is required for a reliable feature matching with a query image taken in orbit. The images in the database should cover enough camera views of the target while at the same time span different illumination conditions.

Hourglass Network

Preliminary results for the CNN are firstly reported for the test database. The key advantages of relying on CNNs for feature detection is that the network is generally capable of learning the relative position between features under a variety of illumination conditions and relative poses present in the training. As a result, both features which are not visible due to adverse illumination and features occulted by other parts of the target can be detected. Figure 12 illustrates some of these scenarios.

Besides, there are at least two aspect associated to the specific selection of the trainable features, which affects the overall network performance over the test database. First of all, since all the selected features represent highly symmetrical points of the spacecraft, such as corners of the solar panel, SAR antenna or main body, the network is in some scenarios unable to distinguish between similar features, and returns multiple heatmaps for a single feature output. This is illustrated in the right-hand side of Figure 13, where it can be seen that the heatmap response of one of the corners of the solar panel includes also the responses of the other similar corners. As a result, the x, y coordinates extracted from the peak of the heatmap can be associated to the wrong feature (i.e. one of the other corners of the solar panel), and lead to a large error. Secondly, besides confusing similar feature, the network has also learned to detect generic corners. As previously mentioned, this is a result of the features selection prior to the training. In other words, in some scenario the network is detecting unwanted corners of the target instead of correctly identifying the wanted features, due to the fact that all the trainable features are corners.

Figure 14 illustrates four different test scenarios which summarize the network performance



(a) Total heatmap obtained by overlapping each single heatmap

(b) Heatmap related to a specific feature - solar panel corner

Figure 13. Example of multiple features detection. A large RMSE errors would result from the difficulties of the network in distinguishing similar features, such as the corners of the SAR antenna, the solar panel, and main body.

observed so far. For each test case, the total heatmap is shown together with the RMSE between each detected feature and the ground truth,

$$E_{\text{RMSE},i} = \sqrt{(x_{GT,i} - x_i)^2 + (y_{GT,i} - y_i)^2} \tag{4}$$

In the upper-left figure, a good features detection, characterized by a total RMSE of 1.37 pixels, is presented. As can be seen, each detected feature deviates from its ground truth by less than 2 pixels. The upper-right figure also present a good detection scenario, in which the total RMSE is even at subpixel level. In the lower-left image, a scenario is presented in which the RMSE relates to an acceptable detection. Finally, the lower-right image shows a scenario in which in the network is unable to correctly detect the last feature, thus returning a very large RMSE. In the specific, due to its corner detector behavior, the network has detected a corner in the spacecraft antenna rather than one of the corners of the solar panel.

Notably, the last scenario suggests that, generally, large RMSE might be associated to the incorrect detection of just one (or a few) features, rather than a completely wrong detection. Figure 15 extends this finding by showing the histogram associated to the number of features detected with a RMSE error above 50 pixels for the whole test database. As can be seen, in most of the cases only a few features are wrongly detected. Only 5% of the output heatmaps contain more than 3 features with a very large RMSE error.

Ideally, the heatmap's peak intensity associated to a wrong detected feature should be relatively small compared to the other detected features, highlighting that the network is not confident about that particular wrongly-detected feature. At the same time, the heatmap's amplitude should provide an additional insight into the confidence level of each detection, a large amplitude being related to large uncertainty about the detection. As such, thresholding based on the relative peak intensity and amplitude should eliminate the wrong features and improve the RMSE behaviour over the whole test dataset.

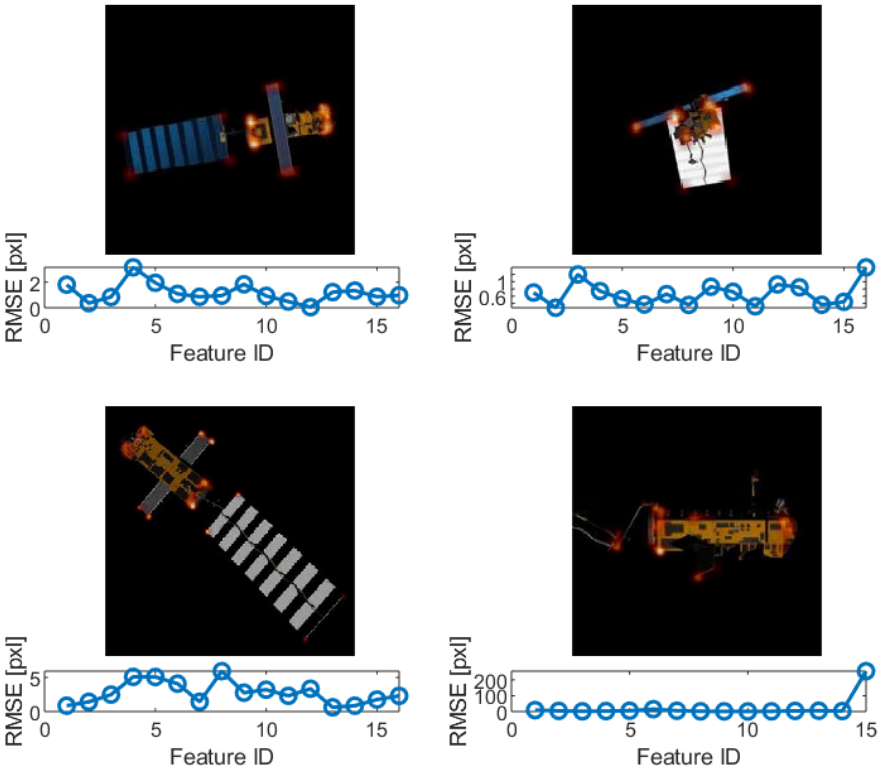


Figure 14. Four different test scenarios for feature detection.

As a starting point for the investigation of the inclusion of the additional heatmap's information in the RMSE computation, the left-hand side of Figure 16 reports a comparison of the RMSE with the Gaussian-like profile parameters extracted from each feature's heatmap for the low-right scenario in Figure 14. For clarity, the values are shown for all the features except the last one, due to its very large error relative to the others. As can be seen, the largest RMSE is associated to a very weak total brightness and peak intensity, whereas the smallest RMSE relates to a rather large peak intensity and low amplitude. To extend the reasoning, it can be derived that a very good detection is generally associated to a Gaussian-like profile with a very large peak and a very small standard deviation, whereas a wrong detection is associated to a very small peak and a large standard deviation. An acceptable detection would lie somewhere in between, that is to say, it could be characterized by both large peak and large standard deviation. The right-hand side of Figure 14 also reports the Gaussian-like profiles associated to the above-mentioned detections.

Clearly, threshold values for both the peak intensity and the amplitude of the heatmaps have to be carefully tuned beforehand. In general, very severe threshold values would sensibly decrease the RMSE and discard many non-optimal features, although they will at the same time decrease the number of detected features per image. As this latter aspect can severely affect the accuracy of the relative pose estimation, due to the fact that pose estimation solvers usually require at least 3-to-6 feature locations, a trade-off is required for the threshold selection, in order to find a compromise between RMSE reduction and reduction of the total number of available features.

By referring to the scenario depicted in 16, some preliminary threshold values can be selected to be equal to the median of the peak intensities and amplitudes of the detected features, which

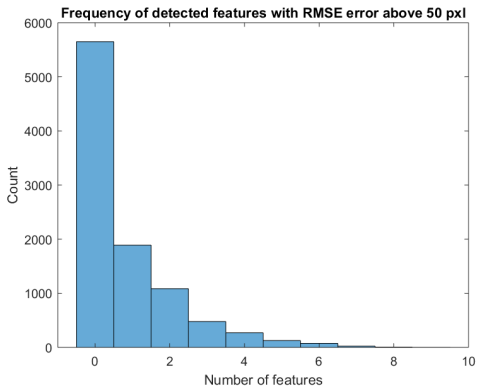
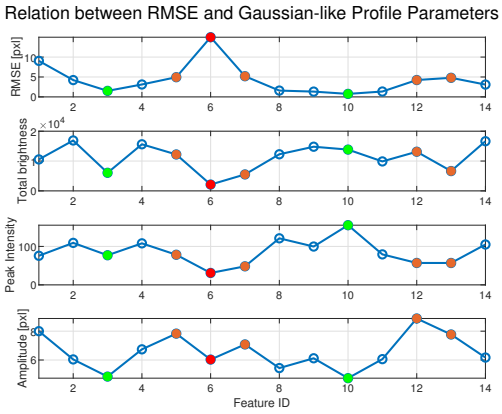
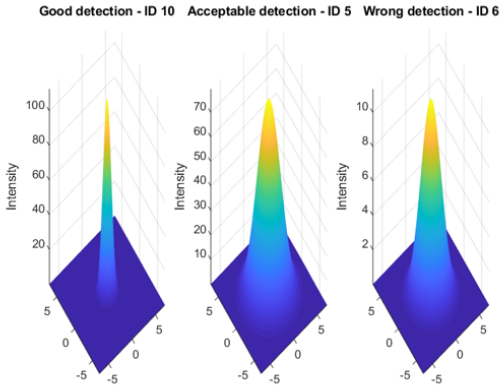


Figure 15. Histogram of number of features detected with RMSE error above 50 pixels.



(a) Comparison of RMSE with Gaussian-like profile parameters



(b) Gaussian-like profiles derived from the heatmaps

Figure 16. Relation between RMSE and Gaussian-like parameters. Some of the good, acceptable and bad detections are reported in green, orange, and red, respectively. As can be seen, a large peak intensity and a small amplitude are associated to the lowest RMSE. As such, thresholding can discard those features whose amplitude and signal are too large and/or too weak, respectively.

amounted to 78 and 6, respectively. A heatmap detected by the CNN is discarded if its peak intensity and amplitude are smaller/larger than their threshold values. Figure 17 illustrates the histograms of the RMSE and total number of detected features, before and after applying the threshold. As can be seen in the left-hand side, thresholding greatly helps reducing the RMSE over the whole test database. However, the effect on the total number of detected features is a much more stretched-out distribution. As a consequence of that, some scenarios can be found in which the number of detected features are too few to expect an accurate relative pose estimation. As such, it might be worth selecting less stringent threshold values.

It has to be mentioned that, even in case no thresholding is performed, it is still expected that the RANSAC algorithm, embedded in most of the pose estimation solvers, would return a pose which is not affected by the outlier features. Also, beside RANSAC, the intensity of the feature peak can be accomodated in a weight matrix which accounts for the uncertainty of the feature.[12]

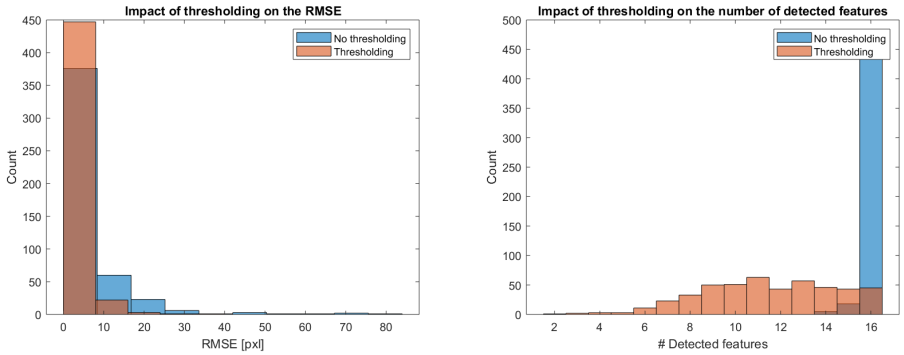


Figure 17. Impact of applying a threshold to the feature detection on the RMSE (left) and the total number of features detected per image (right).

Finally, the performance of the hourglass network is assessed for the yaw rotation and the two illumination scenarios. Figure 18 illustrates the number of tracked features using the threshold values previously described. When compared to the results in Figure 8, these results indicate that the CNN outperforms the HT-based method in regards to features tracking for the entire yaw sequence. However, a performance drop can be observed at the start as well as towards the end of the sequence. This is a consequence of having selected highly symmetrical features, as already anticipated for the test database. More specifically, the network is confusing between similar features for some of the camera views of the target. As such, the improvements in the network performance resulting from selecting asymmetrical features shall be assessed. Besides, the performance in the two illumination scenarios is still characterized by less than four tracked features for most of the two sequences. Since this is a consequence of the fixed Sun azimuth and elevation adopted in the training phase, more illumination conditions shall be recreated in the offline database in order to return a robust detection in changing illumination scenarios.

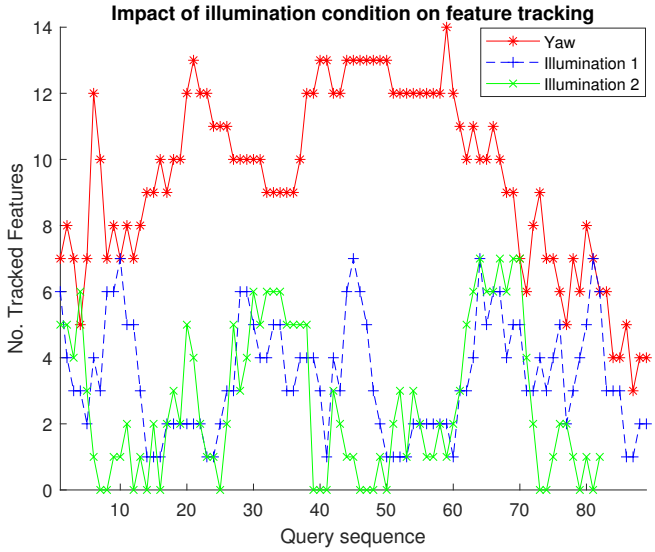


Figure 18. Feature tracking performance for the proposed CNN-based detector.

CONCLUSIONS

In this paper, a comparative assessment of IP algorithms has been conducted to evaluate their performance with respect to illumination variations, partial occlusion, and inter-frame rotation, during close-proximity operations around a target spacecraft.

A novel corner detection method, based on the HT, was implemented and evaluated in terms of performance during yaw rotation of the target and illumination changes due to varying Sun-target geometries. Despite the potentials compared to a simple Hough detection and extraction algorithm, results showed a certain level of instability of the detected corner in both scenarios. Specifically, the feature tracking performance resulted to be affected by both small target rotation and illumination changes.

Besides, the performance of five feature detectors/descriptors was assessed in the context of feature matching. Results highlighted the challenges of feature matching when illumination conditions between two images vary during the orbit, suggesting to carefully select the illumination conditions to be recreated in an offline database, when matching this latter with a query image taken in orbit. Future works will extend these finding by assessing the number of camera views as well as the number of different Sun azimuth and elevation which are required in an offline database to guarantee a robust and reliable feature matching.

The CNN performance results obtained for the simple hourglass architecture already illustrated the key advantages of relying on CNNs for feature extraction. Above all, the capability of detecting non-visible and occulted features, together with the additional statistical information returned by the detected heatmaps, represent a considerable advantage with respect to standard methods. Furthermore, the corner detection during the selected tracking scenario proved to be more robust than the HT-based method. Nevertheless, there are still some aspects which would require further analyses. First of all, due to the fact that the training was performed with a fixed Solar azimuth and elevation, it is not guaranteed that the network will be able to perform a reliable detection in case the illumination conditions in orbit differ from the offline one. Moreover, the impact of selecting asymmetrical features during the training phase on the detection performance is still unclear. Secondly, the absence of any object in the image background during the training phase might lead to a completely wrong feature detection, in case of Earth-in-the-background geometries in orbit. Moreover, a *stacked hourglass architecture*, in which two or more hourglass modules are stacked together in order to refine the heatmap response, has not been assessed yet. Future analyses will focus on these aspects in order to improve the robustness of CNN-based feature extraction methods for space applications.

Finally, future works will validate all the methods with realistic images acquired in a laboratory environment to assess the performance of IP on space-like imagery. Software-In-the-Loop tests are planned to also assess the capability of implementing such methods in space-representative processors.

ACKNOWLEDGMENT

This study is funded and supported by the European Space Agency and Airbus Defence and Space under Network/Partnering Initiative (NPI) program with grant number NPI 577-2017.

REFERENCES

[1] L. Pasqualetto Cassinis, R. Fonod, and E. Gill, "Review of the Robustness and Applicability of Monocular Pose Estimation Systems for Relative Navigation with an Uncooperative Spacecraft," *Progress in Aerospace Sciences*, In Press, Available online June 2019, https://doi.org/10.1016/j.paerosci.2019.05.008.

[2] X. Du, B. Liang, W. Xu, and Y. Qiu, "Pose measurement of large non-cooperative satellite based on collaborative cameras," *Acta Astronautica*, Vol. 68, No. 11&12, 2011, pp. 2047–2065, 10.1016/j.actaastro.2010.10.021.

[3] R. Duda and P. Hart, "Use of the Hough Transformation To Detect Lines and Curves in Pictures," *Communications of the ACM*, Vol. 15, No. 1, 1972, pp. 11–15, 10.1145/361237.361242.

[4] S. D'Amico, M. Benn, and J. Jorgensen, "Pose Estimation of an Uncooperative Spacecraft from Actual Space Imagery," *International Journal of Space Science and Engineering*, Vol. 2, No. 2, 2014, pp. 171–189, 10.1504/IJSPACESE.2014.060600.

[5] S. Sharma, J. Ventura, and S. D'Amico, "Robust Model-Based Monocular Pose Initialization for Noncooperative Spacecraft Rendezvous," *Journal of Spacecraft and Rockets*, Vol. 55, No. 6, 2018, pp. 1–16, 10.2514/1.A34124.

[6] V. Capuano, S. Alimo, A. Ho, and S.-J. Chung, "Robust Features Extraction for On-board Monocular-based Spacecraft Pose Acquisition," *AIAA Scitech 2019 Forum*, San Diego, CA, USA, 2019, 10.2514/6.2019-2005.

[7] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, Vol. 60, No. 2, 2004, pp. 91–110, 10.1023/B:VISI.0000029664.99615.94.

[8] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," *European Conference on Computer Vision*, 2010, pp. 778–792, 10.1007/978-3-642-15561-156.

[9] J. Shi, S. Ulrich, and S. Ruel, "Spacecraft Pose Estimation using a Monocular Camera," *67th International Astronautical Congress*, Guadalajara, Mexico, 2016.

[10] D. Rondao, N. Aouf, and O. Dubois-Matra, "Multispectral Image Processing for Navigation Using Low Performance Computing," *69th International Astronautical Congress*, Bremen, Germany, 2018.

[11] D. Rondao and N. Aouf, "Multi-View Monocular Pose Estimation for Spacecraft Relative Navigation," *2018 AIAA Guidance, Navigation, and Control Conference*, Kissimmee, FL, USA, 2018, 10.2514/6.2018-2100.

[12] G. Pavlakos, X. Zhou, A. Chan, K. Derpanis, and K. Daniilidis, "6-DoF Object Pose from Semantic Keypoints," *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[13] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," *European Conference on Computer Vision*, 2016, pp. 483–499.

[14] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, Vol. 110, No. 3, 2008, pp. 346–359, 10.1016/j.cviu.2007.09.014.

[15] V. Lepetit and P. Fua, "Monocular Model-Based 3D Tracking of Rigid Objects: A Survey," *Foundations and Trends in Computer Graphics and Vision*, Vol. 1, No. 1, 2005, pp. 1–89, 10.1561/0600000001.

[16] M. A. Fischer and R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, Vol. 24, No. 6, 1981, pp. 381–395, 10.1145/358669.358692.

[17] R. Hartley, "In defense of the eight-point algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 6, 1997, pp. 580–593, 10.1109/34.601246.

[18] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," *International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 2564–2571, 10.1109/ICCV.2011.6126544.