2022 The 22st International Conference on Control, Automation and Systems (ICCAS 2022)
BEXCO, Busan, Korea, Nov. 27∼Dec. 01, 2022

Performance Enhancement Step for Motion Estimation via Feature-based Image Matching

Keita Miyaura*, Armagan Elibol and Nak Young Chong

School of Information Science, Japan Advanced Institute of Science and Technology,
Ishikawa, 923-1292, Japan ({s2110165, aelibol,nakyoung}@jaist.ac.jp) * Corresponding author

Abstract: Most of the complicated and sophisticated tasks in visual robotics applications usually build upon the image matching step as matching images of the same scene can provide important information (e.g., camera motion). Image matching is generally done via extracting and matching some distinctive points via their feature vectors. This procedure generates some mismatched points due to imperfections. Mismatched points are called outliers and identified via probabilistic methods. Since the probabilistic methods work iteratively, they generally occupy a large portion of the computational cost of the whole image matching pipeline. In this paper, we present a simple yet efficient algorithm that is employed for eliminating the outliers aiming at reducing the total number of iterations needed in the probabilistic methods. Our method is motivated by the common way of visualizing the established matches among images. We tile images together and search for parallel lines connecting correspondences. We present extensive computational and comparative experiments using both simulated data involving along with real images and using a real dataset.

Keywords: image matching, motion estimation, outlier rejection

1. INTRODUCTION

Thanks to the advancements in sensing technology, camera-carrying mobile platforms have become more and more accessible and available for a vast variety of engineering and science disciplines. This also leads to scientific developments in mainly computer vision, image processing, and machine learning areas. Without a doubt, image matching (or registration) has secured its position and necessity at the core of many more sophisticated tasks. It is defined as a procedure of overlaying two images that have some overlapping area. In order to overlay images, the coordinate transformation (motion) between their coordinate frames (top-left corner as origin) needs to be calculated. The success of several different high-level methods in computer vision and robotics (*e.g.*, mapping, 3D reconstruction, localization, and similar others) relies on image matching. Image registration methods are mainly categorized in three categories [22]; Optical Flow [11, 19], Fourier Transform based [16], and Feature-based. Over last two decades, developments on feature point detection and description (usually Scale invariant feature transform (SIFT) [12] and Speeded up robust features (SURF) [2]) made it possible to compute the transformation between images even under extreme cases (*e.g.*, various scale and viewpoints changes). These advancements direct researchers to use Feature-based methods more. Different deep-learning-based methods have been also proposed for feature detection, and matching (*e.g.*, [13, 21]) and comparative benchmarking was presented in [1]. The D2-Net framework for joint detection and description of local features was proposed in [6]. Its performance in localization tasks outperforms the other methods while in image matching, it has some limitations.

Feature-based image matching starts with detecting some salient points (regarded as features) in images. These feature points are represented with a vector of scalars obtained using pixel values in their neighborhood (*e.g.,* set of histograms of orientation gradients) and this vector is called a descriptor and this process is called feature description. Descriptors are matched by comparing the Euclidean distances between descriptor vectors. During this matching procedure, some feature points usually are not correctly matched. These mismatches are referred to as outliers. Probabilistic methods (*e.g.*, Random sample consensus (RANSAC) [9], Least median of squares regression (LMeds) [18] and similar others) have been employed in order to remove outliers and compute the transformation (or motion). Over the years, there have been several improvements proposed and presented for RANSAC, which is based on random sampling and using a threshold to identify inliers and outliers, in two directions namely, improving sampling methods and automatic threshold selection. PROSAC (Progressive Sample Consensus) [4] was proposed with an enhanced sampling algorithm based on ranked features according to similarity scores of their descriptors and samples were drawn through their ranking. In terms of decisive threshold selection for inlier-outlier separation, *automatic* RANSAC methods have been proposed [5, 15, 17]. Although they have improved the overall performance, due to their iterative procedure nature, their computational cost can be still high. Some recent works have also considered some improvements using geometric relations [7, 8]. As stated by the authors, the approach in [7] requires a good calibration of threshold values for the dataset used. In [8], the usage of geometric invariants has been shown its efficiency, but it might still include a computational cost while extracting the geometric invariants stably from images.

In this paper, we present a simple yet efficient prefiltering step in order to improve the performance of the robust estimation method by reducing the total number

This work was supported by the U.S. Air Force Office of Scientific Research under AFOSR/AOARD FA2386-20-1-4019 grant.

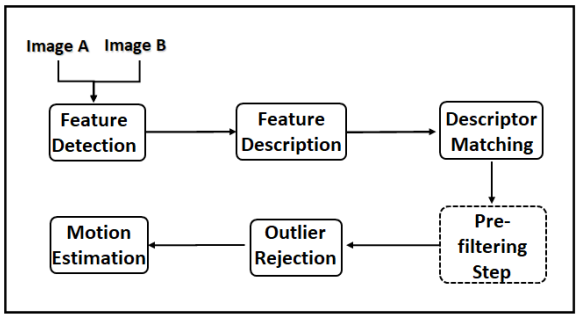


Fig. 1.: Image Matching Pipeline with the proposed pre-filtering step.

of outliers. We make use of angle histograms of the lines connecting correspondences when images are tiled together as inliers commonly form parallel lines. We present extensive computational experiments using both simulated and real data. Since our pre-filtering step aims to reduce the total number of outliers thus increasing the inlier ratio, its usage is beneficial for all types of robust estimation methods.

2. PRE-FILTERING STEP FOR OUTLIER REDUCTION METHOD

Correspondences are usually visualized via tiling overlapping images together and drawing a line connecting each correspondence. Inliers mostly form a group of parallel lines in such visualization since differences in slope are preserved [3, 10] based on the geometrical invariants (specifically for certain type(s) of transformations). Motivated by this, we present a pre-filtering step to remove the outliers before applying a robust estimation method for outlier rejection and motion estimation. Our proposed approach applies a translation transformation to the feature point positions of the second image so that their positions with respect to the first image coordinate frame are obtained when they are tiled together. Then we compute the angles of the lines connecting correspondences. Once angles are computed, we find the peak group in the histogram created using the angle values. The lines, thus feature points forming those lines are considered as inliers and kept for further processing via robust estimation methods. Algorithmic representation of the proposed pre-filtering step is provided in Alg. 1 while image matching pipeline with embedded our proposed step is illustrated in Fig. 1.

An example of applying the proposed filtering step tiling horizontally and vertically can be seen in Figs. 2 and 3. Before applying the filtering step, a total of 14 correspondences were established and only 5 inliers. After applying the filtering and unification step, a total of 7 correspondences remained including 5 inliers and 2 outliers.

Algorithm 1: Algorithm for filtering correspondences via angle grouping

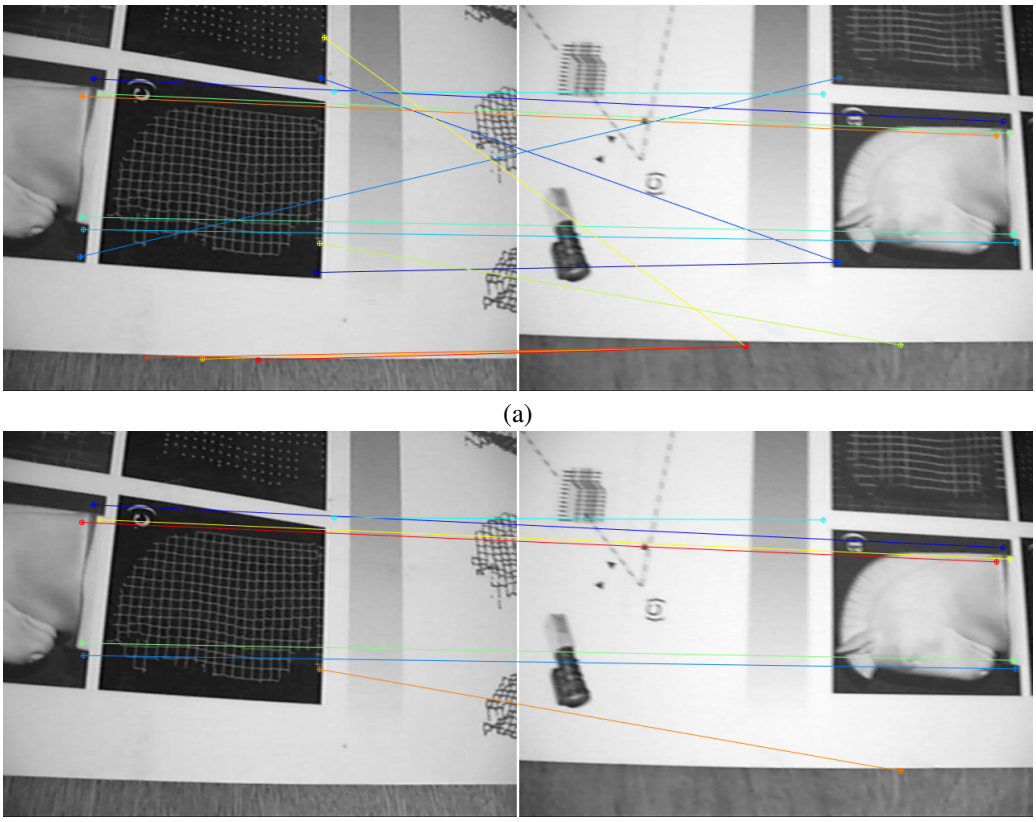
Input: Matched feature positions in the local image coordinate frames
$\mathbf{P} = (\mathbf{x_i}, \mathbf{y_i}) \quad \mathbf{i} = \mathbf{1}, \mathbf{2}, \cdots, \mathbf{n}$ and
$\mathbf{M} = (\mathbf{x_i}, \mathbf{y_i}) \quad \mathbf{i} = \mathbf{1}, \mathbf{2}, \cdots, \mathbf{n}$,
Image size (u, v),
Bin-width for angle grouping, w
Output: a set of correspondences indices to be kept, Ind
foreach *feature correspondences* p *and* m **do**
  $\mathbf{A1} \leftarrow$ Compute the slope using $p = (x_p, y_p)$ and translated coordinate $m = (x_m + u, y_m)$
  $\mathbf{A2} \leftarrow$ Compute the slope using $p = (x_p, y_p)$ and translated coordinate $m = (x_m, y_m + v)$
Compute histograms over values in $\mathbf{A1}$ and $\mathbf{A2}$ using bin-width w
Keep correspondence indices in the bin width the maximum number of elements
$Ind \leftarrow$ Unify indices coming from two histograms

3. EXPERIMENTAL RESULTS

We have tested our proposed pre-filtering step via both extensive simulations using 35 different homography matrices used in [14] covering most of the challenging situations and real images in the *graffiti* dataset [14] (available at `https://www.robots.ox.ac.uk/~vgg/data/affine/`) as it provides various viewpoints between images and ground-truth homographies. For simulation experiments, We used SURF [2] to extract features from an image and used homographies to generate their correspondences as ground truth. From this ground-truth data, for each simulation parameter summarized in Table 1, we randomly generated a set of correspondences that is composed of both inliers and outliers correspondingly and run M-estimator Sample Consensus (MSAC) [20] with the proposed pre-filtering step and without it. Moreover, we also corrupted correspondences positions by adding a zero-mean noise with different levels of standard deviations. If the error computed as in Eq. 1 is less than the distance threshold of 5 pixels, the estimated homography is considered correct and such a trial is counted as successful. The error is computed as a mean of distance between the total number (n) of correspondences $((\mathbf{p}, \mathbf{m}))$ when they are mapped with the estimated homography $\mathbf{H}$.

$$\mu = \frac{\sum_{i=1}^{n} ||\mathbf{p_i} - \mathbf{H} \times \mathbf{m_i}||_\mathbf{2}}{n} \tag{1}$$

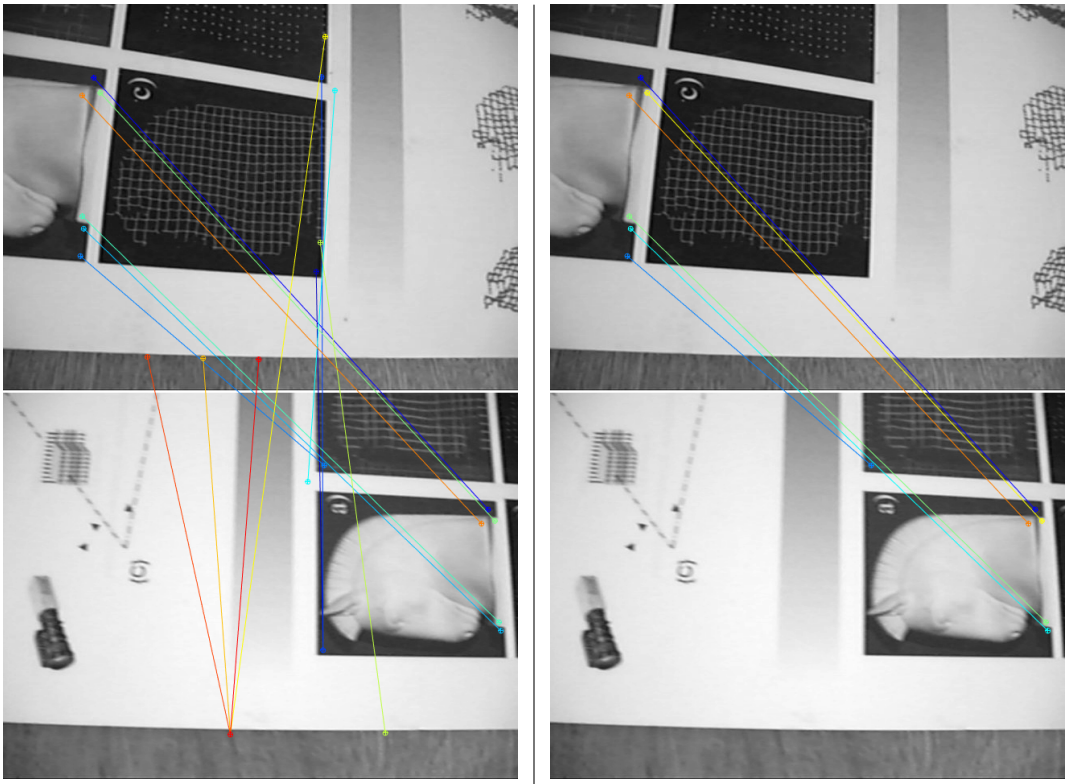
A threshold of 5 pixel is determined as the maximum error obtained by using ground-truth transformations with noise corrupted correspondences during our simulations. For each homography, we have $3 \times 5 \times 5$ different simulation parameter configurations and we repeated each configuration $1,000$ times leading to a total of $2,625,000 (35 \times 75 \times 1,000)$ trials.



(a)

(b)

Fig. 2.: (a) Established Correspondences. A total of 14 correspondences and only 5 inliers. (b) Remaining Correspondences after applying filtering step. A total of 7 correspondences including 2 outlier and 5 inliers.



(a)

(b)

Fig. 3.: (a) Established correspondences. A total of 14 correspondences and only 5 inliers. (b) Remaining correspondences after applying filtering step. A total of 6 correspondences including 1 outlier and 5 inliers.

Table 1.: Parameters used in Experiments

Number of Transformations	35
Total Number of Correspondences (inliers + outliers)	$[100, 250, 500]$
Outlier Ratios	$[0.5, 0.6, 0.7, 0.8, 0.9]$
Noise Standard Deviations	$[0.0, 0.5, 1.0, 1.5, 2.0]$
Maximum Number of Iterations in RANSAC	$2,500$
Distance Threshold	5 pixels

We present the obtained results in the format of confusion matrix in Table 2. In total, the proposed pre-filtering step was able to improve the result by approximately $\%10$. Moreover, using pre-filtering step helped to reduce the number of iterations needed during robust estimation. This is mainly due to the fact that it reduces the total number of correspondences while increasing the inlier ratio and both are favorable for reducing the number of iterations. Statistically summarized results on the number of random trials in RANSAC are given in Table 3.

Since our pre-filtering step relies on counting parallel lines connecting correspondences when they are tiled together, a total number of inliers plays a more important role than the inlier ratio. In other words, our proposed filtering step performs better in the case of an outlier ratio of 0.9 and the total number of correspondences of 500 than in the case of 100 total number of correspondences with the same outlier ratio. For the total cases $(275,029)$ where our filtering step failed to provide accurate homography, we plotted the histogram for each tested number of correspondences with respect to inlier ratios used, and is depicted in Fig. 4. As can be seen, the total number of failure cases using a total number of 100 correspondences is larger than 250 and 500 for each inlier (or outlier) ratio used in experimental tests.

We also present the total number of cases in which the proposed pre-filtering step was successful to increase the initial inlier ratio and mean inlier ratio values in Table 4. The column *greater* denotes the number of cases where the inlier ratio has increased, while the column *less* provides the total number of cases where the inlier ratio has decreased after applying the pre-filtering step. Mean inlier ratios for both cases as well as overall were presented in the last three columns. It can be noted that there is a discrepancy between the total numbers presented in Table 2 and Table 4. This is due to the fact that in some cases although the pre-filtering step has decreased the inlier ratio, the robust estimation method was still able to obtain the correct motion and vice-versa.

We tested our pre-filtering step on the *graffiti* image set. We followed the standard SIFT-based pipeline to obtain correspondences and estimated the projective transformations both with and without using the proposed pre-filtering step. We used an error threshold of 7.5pixels and a maximum of $5,000$ sampling iterations in RANSAC. The obtained results are summarized in Table 5. Since the ground truth transformations are available, we also computed the number of inliers using them and reported at the last two columns for comparison and providing insights about the established correspondences quality using the same error threshold. From the table, it can be seen that the proposed pre-filtering step was able to improve the image matching pipeline overall. For the image pair 1-5, the estimated homography using the pre-filtering step was at a similar accuracy level to the ground truth while in the case of without using the pre-filtering step, the transformation estimated was not correct. In the case of image pairs 1-6, both methods failed to estimate correctly, and this was mostly due to the selected error threshold and the established correspondences, which can be considered as a direct consequence of the feature detection and matching method used.

4. CONCLUSIONS AND FUTURE WORK

Image matching is one of the most crucial and fundamental steps in many sophisticated computer vision and robotics tasks. Although there have been tremendous efforts and breathtaking advancements in feature extraction, description, and matching steps, especially via using deep learning methods, outliers do occur, and robust estimation is still needed to remove them. They are considered as one of the steps requiring the most computational time due to their iterative procedure. At this point, in order to reduce this burden, we present a simple yet efficient pre-filtering method, our pre-filtering method aims to improve the inlier ratio for a given set of correspondences via grouping them by the tangent of the line connecting them when two images are tiled together. We present its efficiency using extensive simulations and real data. It has not only improved the success rate of image matching but also helped to reduce the number of iterations done during the robust estimation since it reduces the total number of correspondences and increases the overall inlier ratio. In future work, we will focus on extending our pre-filtering step to be able to cope with the multiple motion hypothesis since it currently assumes that outliers do neither come from nor obey a single motion.

REFERENCES

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 2017.

[2] H. Bay, T. Tuytelaars, and L. J. Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417, Graz, Austria, May 2006.

Table 2.: Summary of Obtained Results in Confusion Matrix Format

		Without Pre-filtering		
		Unsuccessful	Successful	Total
With Pre-filtering	Unsuccessful	0.066 (173,987)	0.038 (101,042)	0.105 (275,029)
	Successful	0.137 (359,405)	0.758 (1,990,566)	0.895 (2,349,971)
	Total	0.203 (533,392)	0.797 (2,091,608)	1.000 (2,625,000)

Table 3.: Number of Iterations in Robust Estimation

	min.	max.	mean	std.
With (overall)	3	2,500	644.88	908.69
Without (overall)	57	2,500	1,301.31	1,061.55
With (succ.)	3	2,500	440.99	709.54
With (unsucc.)	21	2,500	2,386.97	434.94
Without (succ.)	57	2,500	995.79	977.06
Without (unsucc.)	219	2,500	2,499.34	32.20

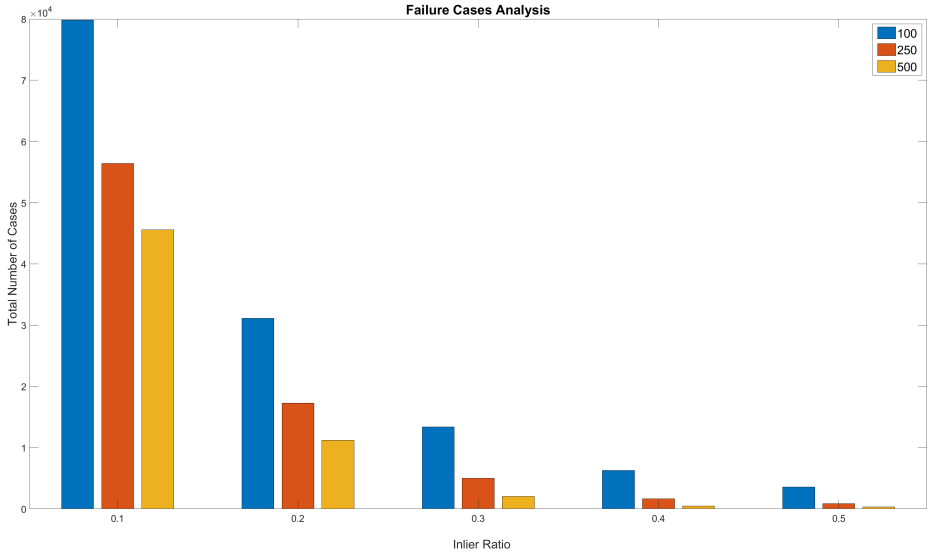


Fig. 4.: Histogram of the total failure cases using pre-filtering step with respect to inlier ratios. The standard image matching procedure also failed in %63.26 of the total cases reported in this figure. In the case of a total of 100 correspondences has a larger number of failures in each inlier ratio tested leading to conclude that a total number of inliers plays a more important role than the ratio.

Table 4.: Total Number of Cases Iterations in Robust Estimation

Inlier Ratio	Total Number of Trials			Mean Ratio		
	Greater	Less	Equal	Greater	Less	Overall
0.1	448,928	73,222	2,850	0.230	0.067	0.207
0.2	461,613	61,145	2,242	0.403	0.151	0.373
0.3	469,820	54,032	1,148	0.534	0.242	0.504
0.4	477,120	46,402	1,478	0.638	0.340	0.611
0.5	482,220	39,337	3,443	0.723	0.442	0.701
Total	2,339,701	274,138	11,161			

[3] Michael Bolt, Timothy Ferdinands, and Landon Kavlie. The most general planar transformations that map parabolas into parabolas. *Involve: A Journal of Mathematics*, 2(1):79 – 88, 2009.

[4] Ondrej Chum and Jiri Matas. Matching with prosac-progressive sample consensus. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 220–226. IEEE, 2005.

[5] Andrea Cohen and Christopher Zach. The likelihood-ratio test and efficient robust estimation. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2282–2290, 2015.

[6] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla,

Table 5.: Experimental Results with Graffiti Dataset Images

Image Pairs	Method	Number of Corresp.	RANSAC Iterations	Number of Inliers	Inlier Ratio	With Ground Truth Number of Inliers	With Ground Truth Inlier Ratio
1-2	Without	181	14	141	0.779	141	0.779
	With pre-filtering	153	8	141	0.922		
1-3	Without	136	61	82	0.603	85	0.625
	With pre-filtering	86	29	84	0.977		
1-4	Without	71	1400	17	0.239	17	0.239
	With pre-filtering	39	127	19	0.487		
1-5	Without	52	5001	5	0.096	8	0.154
	With pre-filtering	22	263	8	0.364		
1-6	Without	49	5001	5	0.102	2	0.041
	With pre-filtering	19	959	5	0.263		

Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[7] A. Elashry, B. Sluis, and C. Toth. Improving ransac feature matching based on geometric relation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2021:321–327, 2021.

[8] Armagan Elibol and Nak Young Chong. Efficient image registration for underwater optical mapping using geometric invariants. *Journal of Marine Science and Engineering*, 7(6):178, 2019.

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

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Harlow, UK, second edition, 2004.

[11] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

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

[13] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021.

[14] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005.

[15] Lionel Moisan, Pierre Moulon, and Pascal Monasse. Automatic Homographic Registration of a Pair of Images, with A Contrario Elimination of Outliers. *Image Processing On Line*, 2:56–73, 2012.

[16] B Srinivasa Reddy and Biswanath N Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE transactions on image processing*, 5(8):1266–1271, 1996.

[17] Clément Riu, Vincent Nozick, and Pascal Monasse. Automatic RANSAC by Likelihood Maximization. *Image Processing On Line*, 12:27–49, 2022.

[18] Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

[19] Deqing Sun, Stefan Roth, and Michael J. Black. Secrets of optical flow estimation and their principles. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2432–2439, 2010.

[20] P.H.S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156, 2000.

[21] X.Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3279–3286, 2015.

[22] B. Zitová and J. Flusser. Image registration methods: A survey. *Image and Vision Computing*, 21(11):977–1000, 2003.