Keypoints dictionary learning for fast and robust alignment

Aitor Artola, Yannis Kolodziej, Jean-Michel Morel, Thibaud Ehret

KEYPOINTS DICTIONARY LEARNING FOR FAST AND ROBUST ALIGNMENT

Aitor Artola[1,2] *Yannis Kolodziej*[2] *Jean-Michel Morel*[1,3] *Thibaud Ehret*[1]

[1]Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, France
[2]Visionairy
[3]City University of Hong Kong, Department of Mathematics, Kowloon, Hong Kong
aitor.artola@ens-paris-saclay.fr

ABSTRACT

Sparse keypoints based methods allow to match two images in an efficient manner. However, even though they are sparse, not all generated keypoints are necessary. This uselessly increases the computational cost during the matching step and can even add uncertainty when these keypoints are not discriminatory enough, thus leading to imprecise, or even wrong, alignment. In this paper, we address the important case where the alignment deals with the same scene or the same type of object. This enables a preliminary learning of optimal keypoints, in terms of efficiency and robustness. Our fully unsupervised selection method is based on a statistical *a contrario* test on a small set of training images to build without any supervision a dictionary of the most relevant points for the alignment. We show the usefulness of the proposed method on two applications, the stabilization of video surveillance sequences and the fast alignment of industrial objects containing repeated patterns. Our experiments demonstrate an acceleration of the method by 20 factor and significant accuracy gain.

Index Terms— Keypoints, SIFT, a contrario detection, RANSAC, image alignment

1. INTRODUCTION

Keypoint methods seek to detect points of interest in a scene. This detection must be robust to changing viewpoints and transformations. Traditionally, keypoints are detected on well localized structures such as corners and blobs, and computed on the Laplacian of the scale space of the image [1]. Each keypoint is associated a descriptor acting as a comparison key. It can be matched to other descriptors in other views of the same scene. The most notable keypoint and descriptor generator is the Scale Invariant Feature Transform (SIFT) [1, 2], which has been followed by a flurry of variants such as SURF, KAZE, AKAZE, ORB and BRISK, reviewed in [3]. A new category even uses deep learning to learn this task [4, 5].

A performance comparison of keypoint and descriptor generators can be found in several reviews [3, 6, 7], and the creation of evaluation databases such as HPAtches [8]. The main comparison criteria are the distinctiveness and repeatability of such descriptors. The first criterion guarantees the uniqueness of the descriptor while the second considers its robustness to perturbations.

In this article we fully reconsider these criteria in the particular but important case where the images to be compared are numerous and correspond either to repeated views of a same scene (like in video surveillance), or successive snapshots of similar industrial objects (for quality assessment purposes).

We propose to learn the most distinctive and repeatable keypoint-descriptors pairs on each given scene or object, with the obvious goal of speeding up video stabilization, or object registration.

The L_2 distance is the most common choice to compare descriptors, but any L_p metric can be used [2] and more sophisticated metrics have been proposed in [9, 10]. The classic SIFT matching criterion has the drawback that it discards descriptors that happen to repeat in the target image. This is a big limitation, particularly for industrial parts that often have repetitive patterns. This question is addressed in [10] through a statistical test that is not disturbed by the presence of repetitive structures.

Once keypoints have been matched through a comparison of their descriptors, the second step of image matching is to find the underlying transformation between two images. This is universally done by RANdom SAmple Consensus (RANSAC) [11], a method to eliminate outliers by iterative random testing. Alternatives and variants accelerating the method, or setting an automatic decision threshold, are given in [12, 13].

The related video stabilization problem [14–19] is often handled in the general case of a camera in motion. It aims at selecting the right keypoints and accelerating their tracking. In the case of static camera, the problem is to identify the background keypoints that can be used reliably for fast background matching.

The focus on descriptors and descriptor matching is fully changed by the assumption that we dispose of several images and that "matching will be repeated many times". Another issue that requires reconsidering the "generic image matching procedure" is the frequent occurrence in industrial control of objects with many repeated patterns, where the descriptors might be distinctive in general, but not in that particular scene. None of the generic image matching methods listed above have that focus.

Our problem here is find a short cut that learns the most unique and repeated keypoint and descriptors pairs in a given scene, so that by their use RANSAC is trivialized and image matching becomes extremely fast.

2. LEARNING A DICTIONARY FOR FAST AND ROBUST ALIGNMENT

The major computational bottleneck when trying to align images with sparse descriptor methods like SIFT [1] is caused by the number of keypoints that needs to be processed. Picking a random small subset of reference keypoints is not an option, as many may represent regions of the image that repeat themselves, or that move (such as a pedestrian in the case of a surveillance footage).

Work supported by a CIFRE scholarship of the French Ministry for Higher Studies, Research and Innovation.

We assume that we dispose of a learning set of N images of a scene, representative of the future images we want to align, and moderately large (about 20 in our experimental setup). These images may for example be samples of industrial parts that we want to register for anomaly detection, or the N first frames of a surveillance video. Our goal is to align quickly all future images of the scene to one of them, e.g. to the first one. To this aim, we want to learn a small dictionary of the most robust and distinctive keypoints from the N sample images, thus ensuring fast and robust matching. With such keypoints, the improvement is twofold: fewer keypoints will be used during the matching step, and the RANSAC step is made trivial by using the most distinctive keypoints.

Our selection method proceeds as follows. We start by computing the keypoints and their respective descriptors by any classic method like SIFT. For each image I_i, i from 1 to N, this generates a list of L_i keypoints. These keypoints are characterized by their positions $\mathbf{x}_{i,j}$ and descriptors $\mathbf{d}_{i,j}$ for j in $\{1,\dots,L_i\}$. We then align the remaining $N-1$ images onto the reference image using standard keypoint matching procedure, followed by RANSAC process. For the matching, we use the adaptive threshold, defined by the *a contrario* descriptor matching method from Rabin *et al.* [10]. Indeed, the SIFT relative threshold (obtained by comparing the distance between a matched descriptor to its best matching candidate with the distance to second closest descriptor [2]) does not authorize multiple matching. Similarly, we apply an *a contrario* RANSAC [20, 21] to estimate the transformation between two images, thus avoiding the use of fixed distance thresholds. This step allows us to project all $\mathbf{x}$ onto the reference image so that their spatial positions can be compared. We will assume that $\mathbf{x}$ corresponds to the projected position in the following.

Finally, we apply exactly the same matching procedure to all remaining image pairs.

Connection Graph. The matching of keypoints can be modeled as a graph (V, E), with V the set of vertices and E the set of edges, where the vertices v are the keypoints, *i.e.* $(\mathbf{x}_{i,j}, \mathbf{d}_{i,j})$ for $i \in \{1,\dots,N\}$ and $j \in \{1,\dots,L_i\}$, and the edges represent successful matches, *i.e.* $e = (v_1, v_2)$ if and only if there is a match, not necessarily verified by RANSAC, between the two keypoints represented by v_1 and v_2. For simplicity, we drop vertices v with fewer than the average number of verified matches.

To create relevant groups of keypoints, we associate all keypoints that are at a distance smaller than δ to a given seed keypoint. Consider now a group of points at a given position. Such a group of points can be expressed as a subgraph (V', E') of (V, E) with its internal connectivity, *i.e.* edges from V' to V' (this corresponds to E'), and its external connectivity, *i.e.* edges from V' to $V \setminus V'$.

For a given group of points, modeled by its subgraph (V', E'), we define its number of internal connections $n_{in} = |E'|$ and its number of external connections

$$n_{out} = \left|\left\{e = (v_1, v_2) \mid v_1 \in V' \text{ or } v_2 \in V'\right\} \setminus E'\right|.$$

The goal is now to select the most useful groups of points, namely those that correspond to a single image feature that matches in most images and that is as unambiguous as possible, namely does not match to other positions in the targets.

Once a subgraph (V', E') has been selected, we estimate a template $\tilde{v} = (\tilde{\mathbf{x}}, \tilde{\mathbf{d}})$ representative of the keypoints in the group by taking the element-wise median of its coordinates and descriptors. This is the keypoint that will be saved in the dictionary.

***A contrario* point selection.** As indicated, we want to select the most useful groups of points for our dictionary based on a connectivity criterion. Firstly, its internal connectivity should be as large as possible. This means that matching keypoints are often found at this position in the training images. Hence, a new image will have a high probability of having a keypoint matching with this group (this is the repeatability criterion). Secondly, its external connectivity should be as small as possible (this is the distinctiveness criterion). Indeed, groups of keypoints with high number of external connections are likely to be on repetitive structures and therefore may lure RANSAC into false or partial alignments.

The decision thresholds for these two criteria are defined using *a contrario* detection models [22]. Consider the background stochastic model where keypoint matching is assumed to be uniformly ran-

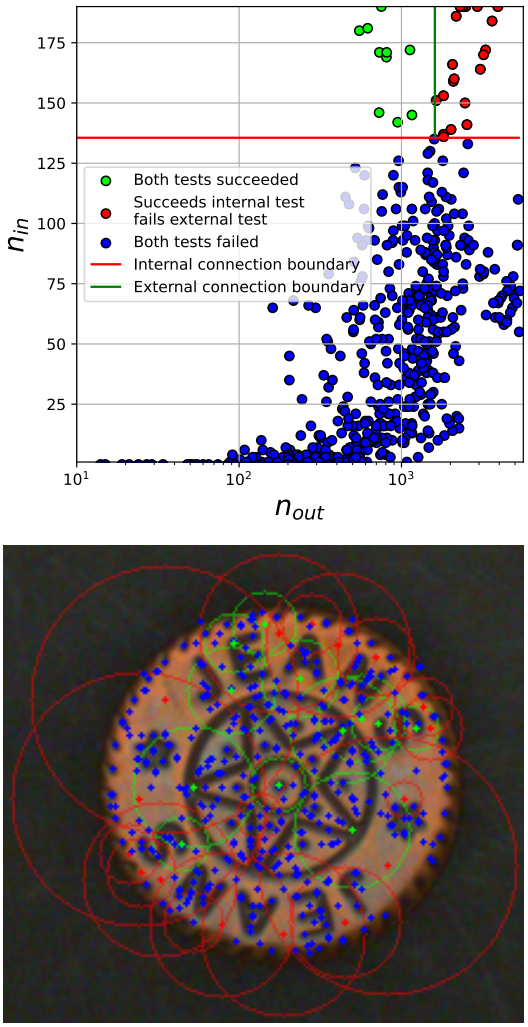

Fig. 1. Distribution of points of interest validated by the *a contrario* test. The top figure is the distribution of groups of points in terms of internal and external connectivity. The bottom figure shows the location of the keypoints in the dictionary onto the reference image. The blue marks correspond to groups that do not pass the statistical test on the number of internal connections and the reds correspond to groups that do not pass the statistical test on the number of external connections. The greens are those that passed both tests and are included in the dictionary.

dom, *i.e*. matching occurred purely by chance and is therefore independent from the keypoints' positions in the images. In that model, the number of internal connections N_{in} of a subgraph (V', E') corresponding to a group of keypoints presented previously is a binomial variable of parameters $(|V'|(|V'|-1)/2, p_{in})$. The probability $p_{in} = \bar{n}_{ver}/(N-1)$ corresponds to the probability that the link between two points have been verified at least in one direction, with $\bar{n}_{ver}$ the mean number of verified links per point.

Given a subgraph of keypoints as above, we therefore evaluate the probability that it could happen just by chance as $\mathbb{P}(N_{in} \geqslant n_{in})$. The number of tests that are being made is the total number of tested groups of aligned keypoints, *i.e*. $N_{test} = |V|$. By the Bonferroni correction, we can therefore ensure that the expected number of false alarms for our subgraph is smaller than ε by imposing

$$N_{test}\mathbb{P}(N_{in} \geqslant n_{in}) \leqslant \varepsilon. \tag{1}$$

This means that to pass the test, the number of false alarm (NFA) of the group of keypoints has to be lower than ε. The probability in (1) can be bounded from above using the Hoeffding approximation alongside the success ratio $r_{in} = 2n_{in}/(N(N-1))$ so that

$$\frac{2}{N(N-1)} \log \mathbb{P}(N_{in} \geqslant n_{in}) \leqslant -r_{in} \log \frac{r_{in}}{p_{in}} - (1-r_{in}) \log \frac{1-r_{in}}{1-p_{in}}. \tag{2}$$

Defining the success ratio using N, the number of images *i.e*. the optimal number of points in a group, instead of $|V'|$, the actual number of points in the group, favors larger groups since it considers the $N - |V'|$ missing points as matching failure.

The case of external connection is very similar. We use the same *a contrario* model defined previously. The only difference is that instead of minimizing the number of external connections, we maximize the number of "non connections" N_{out} so that

$$N_{test}\mathbb{P}(N_{out} \geqslant (|V| - |V'|)|V'| - n_{out}) \leqslant \varepsilon. \tag{3}$$

This allows us to use Hoeffding's upper bound similarly to the internal connections case.

We remarked that the number of external connections is also a good indicator to guide RANSAC tests. A point with a high number of external connections is likely to be on a repetitive structure in the image and therefore may propose false or partial alignments. For a more optimal matching, we order RANSAC tests according to the number of external connections found while building the dictionary. The lowest external connectivity is tested first, then the second lowest is tested in second and so on. Given that RANSAC uses multiple keypoints to estimate an alignment, dictionary tuples are ordered by the product of their number of external connections.

3. EXPERIMENTS

We compared the proposed method on two types of applications, namely industrial part alignment and surveillance video stabilization. In both cases, being fast and robust is crucial. Indeed, both applications require to be real time to be deployed in practice and need to be robust to potential changes (either potential anomalies when tracking industrial parts or scene changes such as pedestrians for video surveillance). Moreover, alignment is a crucial step for downstream tasks such as anomaly detection for industrial parts [23] or change detection [24].

For video surveillance stabilization, we use the "Camera jitter" category of the change detection database CDNET 2014 [25]. It illustrates the case when the camera is fixed a non-perfectly rigid stand such as a lamppost in windy condition or next to a busy road. For the case of industrial parts alignment, we created our own data set comprised of seven models of buttons. One of these is shown in Figure 1. We acquired, in industrial like conditions, between 50 and 150 images for each model. We chose these objects because they contain many repetitive patterns and as such might trap methods that are not robust enough into non-optimal alignments.

For all experiments, we used $N = 20$ images to build the dictionary and a false alarm threshold $\varepsilon = 10^{-2}$ to be conservative. Figure 1 shows the connectivity of each group of keypoints. The groups in the upper left corner of the figure, in green, are the best because they have a high internal connectivity and a low external connectivity, this means that they are both robust and distinctive. This figure shows that the two statistical tests allow us to isolate groups of keypoints with a strong inner connectivity and a weak outer one. The location of these keypoints is consistent with our expectation: they are located on singular regions of the object. For example even though the writing "JEANS" is repeated twice, its interaction with the central star is unique and this is where the majority of the points have been selected.

Table 1. Comparison of industrial parts alignment using either directly the keypoints of the first image of the set (im) or the learned dictionary (dict) in terms of RMSE, matching+RANSAC computation time and number of alignment tests in RANSAC (N_{iter}, see eq 4). Tests were performed with four different versions of RANSAC, (R1) RANSAC with an error threshold of 3 pixels and an inlier threshold of 10%, (R2) AC-RANSAC, (R3) AC-RANSAC with an error threshold of 3, (R4) AC-RANSAC with an error threshold of 3 and a norm difference threshold of 3 (unique scale prior).

		RANSAC			
	Ref	R1	R2	R3	R4
RMSE	im	37.3	72.1	33.5	29.9
	dict	**25.8**	**38.2**	**26.0**	**25.3**
time (in ms)	im	115.9	108.8	115.9	113.9
	dict	**6.2**	**6.2**	**6.3**	**6.4**
N_{iter}	im	99.9	11.2	81.1	42.3
	dict	**5.3**	**2.9**	**5.4**	**3.9**

We compare the proposed alignment scheme to the classic SIFT+RANSAC combo in Table 1 for RMSE after alignment, computation time and number of iterations of RANSAC. The number of iterations N_{iter} of RANSAC, defined by Moisan *et al*. [21], corresponds to

$$N_{iter} = \left\lceil \frac{\log \beta}{\log\left(1 - \frac{n_{inlier}(n_{inlier}-1)}{n_{pts}^2}\right)} \right\rceil. \tag{4}$$

It is defined by the accepted failure rate β, here fixed at 10^{-2}, and the number of points n_{inlier} that verifies the best model among the n_{pts} total points. This shows that the selection of the right keypoints allowed to trivialize RANSAC, the number of tests performed being very low. We also show in Table 1 that the outcome does not depend on the type of RANSAC used. Overall, the proposed method achieves a much better alignment for a fraction of about 5% of the cost of SIFT+RANSAC. We also looked into the order to do RANSAC tests, either randomly like it is usually done or in

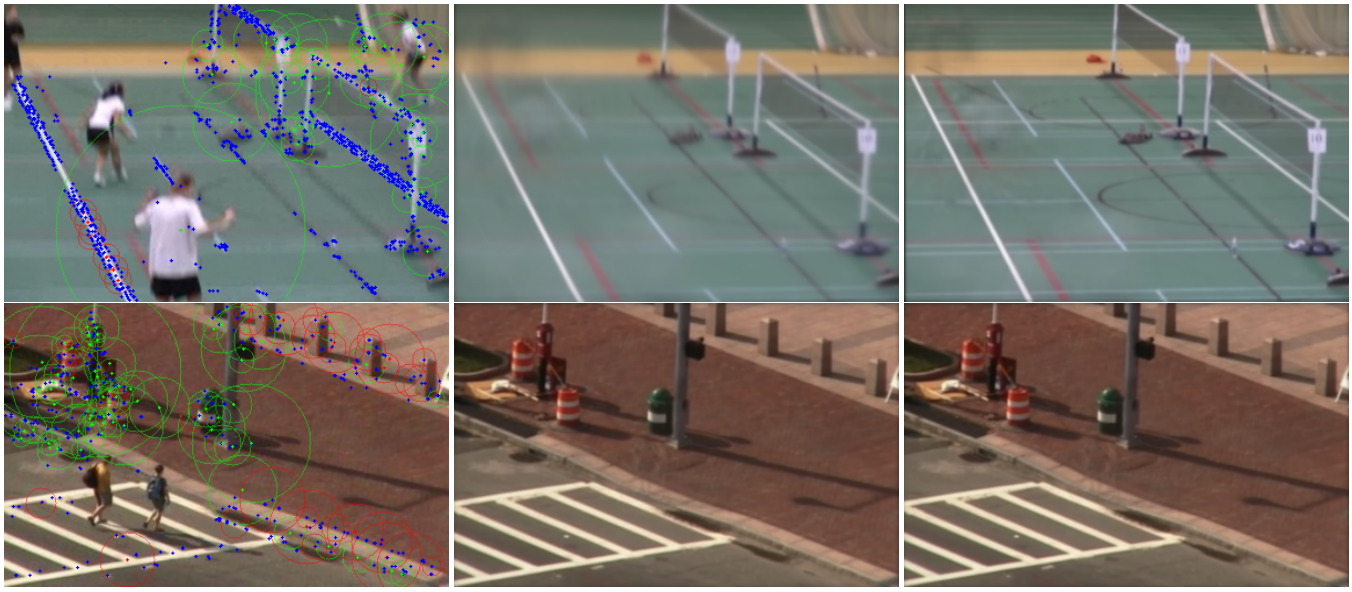

Fig. 2. Stabilization of a video using either SIFT with respect to the reference frame (center) or the dictionary (right). The results are shown in terms of average of the aligned frames. The blue keypoints correspond to graphs that do not pass the statistical test on the number of internal connections and the red correspond to graphs that do not pass the statistical test on the number of external connections. The greens are those that passed both tests and are included in the dictionary.

increasing order of the number of external connections like we proposed. We found out that the proposed order is slightly better since it increases the robustness of the alignment for the same number of computations.

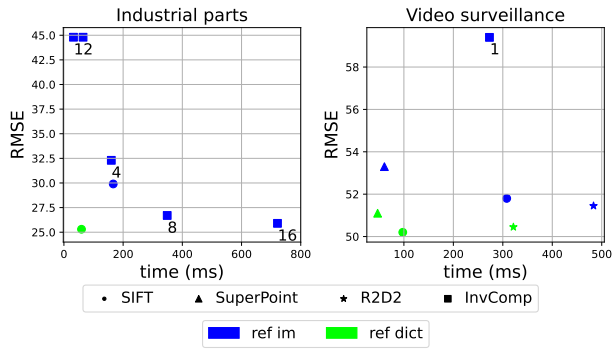

Fig. 3. Results on industrial parts alignment (left) and surveillance video stabilization (right). We compare SIFT [2], SuperPoint [4] and R2D2 [5] with either a reference image or a reference dictionary. Keypoints based methods are also compared with Inverse compositional [26] (the number next to the square indicates the number of initializations, see text for details). Note that deep learning methods [4, 5] did not work on the buttons dataset.

We also compared the performance with other methods in Figure 3. In particular, we compare the performance (in terms of quality vs computation time) of SIFT [2], SuperPoint [4] and R2D2 [5], three different keypoints based methods. Contrary to SIFT, SuperPoint and R2D2 produce keypoints using deep neural networks. Note that these deep learning methods did not work on the industrial parts as they produced keypoints that did not match with each other. We also compared with the Inverse Compositional algorithm [26, 27], a pixel-wise image registration method that optimizes the RMSE between the warped source and the target. Since this method is sensitive to initialization, we provide results with different numbers of initialization. For a given number of initialization k, we compute the alignment using Inverse Compositional for all rotated images using an angle $2l\pi/k$ for l from 1 to k. The reported RMSE corresponds to the best RMSE found after all these alignments. Its poor performance on the video surveillance data can be explained by the large number of outliers between the images (such as, for example, pedestrians or passing cars). This shows that this method is not robust enough for this type of application. On the contrary, the proposed method is both faster and produces a better alignment than all other methods.

Figure 2 shows two examples of video surveillance alignment using the average of the aligned sequence. For the first example, the average frame of the sequence computed using the dictionary has much more details, such as the lines on the ground, compared to the average frame computed with only SIFT+RANSAC.

4. CONCLUSION

We proposed a method to learn the most efficient keypoints for the alignment of images of a scene or of an object, without excluding the use of keypoints with repeated descriptors in the object or in the scene. The method is fully generic and can be plugged to any keypoint based alignment algorithm. The dictionary learning only requires about 20 sample images. Our experiments prove a systematic speed up of a factor of about 20 with respect to a generic matching procedure. This is explained by the reduced size of the dictionary and by the fact that RANSAC is nearly trivialized by the choice of the descriptors.

5. REFERENCES

[1] D.G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 1150–1157 vol.2.

[2] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[3] Shaharyar Ahmed Khan Tareen and Zahra Saleem, "A comparative analysis of sift, surf, kaze, akaze, orb, and brisk," in *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2018, pp. 1–10.

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[5] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel, "R2d2: Reliable and repeatable detector and descriptor," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[6] Krystian Mikolajczyk and Cordelia Schmid, "A performance evaluation of local descriptors," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA*. 2003, pp. 257–263, IEEE Computer Society.

[7] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 23–79, 2021.

[8] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 2017, pp. 3852–3861, IEEE Computer Society.

[9] Jianguo Zhang, Marcin Marszalek, Svetlana Lazebnik, and Cordelia Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2006, New York, NY, USA, 17-22 June, 2006*. 2006, p. 13, IEEE Computer Society.

[10] Julien Rabin, Julie Delon, and Yann Gousseau, "A statistical approach to the matching of local features," *SIAM J. Imaging Sci.*, vol. 2, no. 3, pp. 931–958, 2009.

[11] Martin A. Fischler and Robert C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[12] Lionel Moisan and Bérenger Stival, "A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix," *Int. J. Comput. Vis.*, vol. 57, no. 3, pp. 201–218, 2004.

[13] Rahul Raguram, Ondrej Chum, Marc Pollefeys, Jiri Matas, and Jan-Michael Frahm, "Usac: A universal framework for random sample consensus," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 2022–2038, 2013.

[14] Sebastiano Battiato, Giovanni Gallo, Giovanni Puglisi, and Salvatore Scellato, "SIFT features tracking for video stabilization," in *14th International Conference on Image Analysis and Processing (ICIAP 2007), 10-14 September 2007, Modena, Italy*, Rita Cucchiara, Ed. 2007, pp. 825–830, IEEE Computer Society.

[15] Ken-Yi Lee, Yung-Yu Chuang, Bing-Yu Chen, and Ming Ouhyoung, "Video stabilization using robust feature trajectories," in *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*. 2009, pp. 1397–1404, IEEE Computer Society.

[16] Jing Dong and Haibo Liu, "Video stabilization for strict real-time applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 716–724, 2017.

[17] Rong Hu, Rongjie Shi, I-Fan Shen, and Wenbin Chen, "Video stabilization using scale-invariant features," in *11th International Conference on Information Visualisation, IV 2007, 2-6 July 2007, Zürich, Switzerland*. 2007, pp. 871–877, IEEE Computer Society.

[18] Wilko Guilluy, Laurent Oudre, and Azeddine Beghdadi, "Video stabilization: Overview, challenges and perspectives," *Signal Process. Image Commun.*, vol. 90, pp. 116015, 2021.

[19] Yao Shen, Parthasarathy Guturu, Thyagaraju R. Damarla, Bill P. Buckles, and Kamesh Namuduri, "Video stabilization using principal component analysis and scale invariant feature transform in particle filter framework," *IEEE Trans. Consumer Electron.*, vol. 55, no. 3, pp. 1714–1721, 2009.

[20] 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*, vol. 2, pp. 56–73, 2012.

[21] Lionel Moisan, Pierre Moulon, and Pascal Monasse, "Fundamental Matrix of a Stereo Pair, with A Contrario Elimination of Outliers," *Image Processing On Line*, vol. 6, pp. 89–113, 2016.

[22] Agnes Desolneux, Lionel Moisan, and Jean-Michel Morel, *From gestalt theory to image analysis: a probabilistic approach*, vol. 34, Springer Science & Business Media, 2007.

[23] Aitor Artola, Yannis Kolodziej, Jean-Michel Morel, and Thibaud Ehret, "Glad: A global-to-local anomaly detector," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5501–5510.

[24] Olivier Barnich and Marc Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image processing*, vol. 20, no. 6, pp. 1709–1724, 2010.

[25] Yi Wang, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar, "Cdnet 2014: An expanded change detection benchmark dataset," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 393–400.

[26] Thibaud Briand, Gabriele Facciolo, and Javier Sánchez, "Improvements of the Inverse Compositional Algorithm for Parametric Motion Estimation," *Image Processing On Line*, vol. 8, pp. 435–464, 2018.

[27] Simon Baker and Iain Matthews, "Equivalence and efficiency of image alignment algorithms," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. IEEE, 2001, vol. 1, pp. I–I.