Published in Image Processing On Line on 2023–03–21.
Submitted on 2021–05–05, accepted on 2022–10–31.
ISSN 2105–1232 © 2023 IPOL & the authors CC–BY–NC–SA
This article is available online with supplementary materials,
software, datasets and online demo at
https://doi.org/10.5201/ipol.2023.356

Robust Homography Estimation from Local Affine Maps

Mariano Rodríguez, Gabriele Facciolo, and Jean-Michel Morel

Université Paris-Saclay, ENS Paris-Saclay, Centre Borelli, Gif-sur-Yvette, France
{mariano.rodriguez, gabriele.facciolo, jean-michel.morel}@ens-paris-saclay.fr

Communicated by Gregory Randall *Demo edited by* Mariano Rodríguez

Abstract

The corresponding point coordinates determined by classic image matching approaches define local zero-order approximations of the global mapping between two images. But the patches around keypoints typically contain more information, which may be exploited to obtain a first-order approximation of the mapping, incorporating local affine maps between corresponding keypoints. Several methods have been proposed in the literature to compute this first-order approximation. In this paper we present several modifications of the RANSAC (RANdom SAmple Consensus) algorithm, that uses affine approximations and a-contrario procedures to improve the homography estimation between a pair of images. The a-contrario methodology provides a definition of the soundness of an estimation and allows for adaptive thresholds for inlier/outlier discrimination. These approaches outperform the state-of-the-art for different choices of image descriptors and image datasets, and permit to increase the probability of success in identifying image pairs in challenging matching databases.

Source Code

The reviewed and documented source code and an online demo are available at the web page of this article[1]. Compilation and usage instructions are included in the README.md file of the archive.

Keywords: homography; image comparison; image matching; robust estimation; RANSAC; affine invariance; scale invariance; local descriptors; affine normalization; SIFT; convolutional neural networks

---

[1] https://doi.org/10.5201/ipol.2023.356

Mariano Rodríguez, Gabriele Facciolo, and Jean-Michel Morel, *Robust Homography Estimation from Local Affine Maps*, Image Processing On Line, 12 (2023), pp. 65–89. https://doi.org/10.5201/ipol.2023.356

1 Introduction

Image matching consists in establishing correspondences between different images. This problem is recognized as difficult, especially under severe viewpoint changes between images. This is a fundamental step in many computer vision and image processing applications such as scene recognition [73, 7, 66, 12, 44, 65, 74, 21, 79, 45] and detection [18, 51], object tracking [81], robot localization [67, 72, 48, 5, 52], image stitching [2, 6], image registration [78, 30] and retrieval [22, 20], 3D modeling and reconstruction [15, 19, 75, 1, 57], motion estimation [76], photo management [68, 77, 27, 9], symmetry detection [32] or even image forgeries detection [10].

State-of-the-art image matching algorithms usually consist of three parts: *detector*, *descriptor* and *matching step*. They first detect points of interest in the compared images and select a region around each point of interest, and then associate an invariant descriptor or feature to each region. Correspondences may thus be established by matching the descriptors. Detectors and descriptors should be as invariant as possible.

Local image detectors can be classified by their incremental invariance properties. All of them are translation invariant. The Harris point detector [23] is also rotation invariant. The Harris-Laplace, Hessian-Laplace and the DoG (Difference-of-Gaussian) region detectors [35, 37, 31, 16] are invariant to rotations and changes of scale. Based on the AGAST [33] corner score, BRISK [28] performs a 3D nonmaxima suppression and a series of quadratic interpolations to extract the BRISK keypoints; both detections aim at quickly providing rotation and scale invariances. Some moment-based region detectors [29, 4] including the Harris-Affine and Hessian-Affine region detectors [36, 37], an edge-based region detector [69, 71], an intensity-based region detector [70, 71], an entropy-based region detector [26], and two level line-based region detectors MSER ("maximally stable extremal region") [34] and LLD ("level line descriptor") [49, 50, 8] are designed to be invariant to affine transforms. MSER, in particular, has been demonstrated to have often better performance than other affine invariant detectors, followed by Hessian-Affine and Harris-Affine [38]. In his pivotal paper [31], Lowe proposed a scale-invariant feature transform (SIFT) that is invariant to image scaling and rotation and partially invariant to illumination and viewpoint changes. The SIFT method combines the DoG region detector that is rotation, translation and scale invariant (a mathematical proof of its scale invariance is given in [46]) with a descriptor based on the gradient orientation distribution in the region, which is partially illumination and viewpoint invariant [31]. These two stages of the SIFT method will be called respectively *SIFT detector* and *SIFT descriptor*. The SIFT detector is *a priori* less invariant to affine transforms than the Hessian-Affine and the Harris-Affine detectors [35, 37].

The apparent deformations of objects caused by changes of the camera position can be locally approximated by affine maps, which explains why robust affine invariant methods allow to capture strong homography deformations. A possible way to obtain affine invariance is through the recently proposed Affnet [41]. Affnet is a Convolutional Neural Network (CNN) that was first conceived for improving the normalized representations of Hessian-Affine [37]. As proposed in [41], these normalized representations are then described and matched by HardNet [39]. More recently the AID descriptor [61], a CNN-based patch descriptor trained to capture affine invariance, is able to cope directly with strong viewpoint deformations. Still, it seems that more classical Image Matching by Affine Simulation (IMAS) methods [47, 53, 40, 64, 60] provide the best affine invariances of them all [63]. Therefore, IMAS methods might be more suited for very strong viewpoint differences, although the price to pay is a heavier computational load.

First-order approximations of the local geometry, or simply affine approximations (see Figure 1), can be easily obtained from affine detectors like MSER [34], Harris-Affine[36] or Hessian-Affine [37]. Similarly, the SIFT detector can also be armed with local affine approximations [62]. When estimating homographies from sets of correspondences with the RANSAC algorithm [17], the use of first-order approximations allows to increase the performance in homography estimation. This has already been proposed in [56] by composing normalized affine maps provided by the Hessian Laplace detector. This information can be replaced with the one provided by Affnet [41] or LOCATE [62] since they have been shown to produce more accurate affine maps. In addition, a modification in the RANSAC consensus step has been proposed in [62], encouraging geometry consistency. Instead of defining inliers only through location agreement, the authors also consider the agreement in tilt, rotations and scale of the local affine maps.

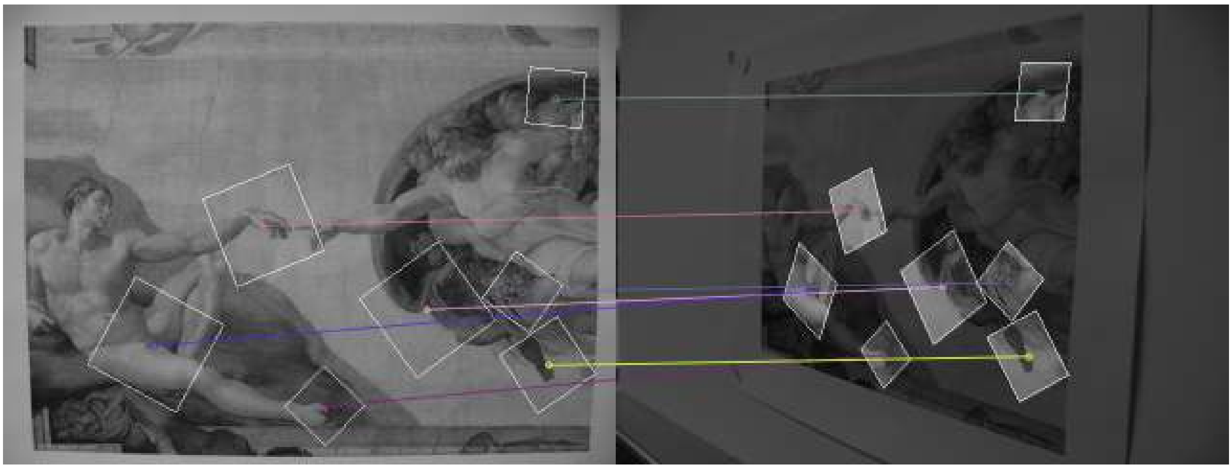

Figure 1: (From [62]) Some correspondences together with local affine approximations of local geometry. Patches on the target are warped versions of their corresponding query patch.

A well established way of automatically estimating the 2D homography relating two images, see [24] p.123, is:

1. **Detection and description.** Compute and describe interest points in each image.
2. **Putative correspondences.** Compute a set of interest point matches based on similarity between their descriptors.
3. **RANSAC robust estimation.** Choose the homography η with the largest number of inliers.
4. **Optimal estimation.** Re-estimate η from all correspondences classified as inliers, by minimizing the Maximum likelihood cost function, see [24] p.95.
5. **Guided matching.** Further interest point correspondences can be determined using the estimated η or the affine approximation around each match.

In this paper, we focus on some available improvements for the step 3. In particular, we highlight the advantages of three local affine approximations in improving the robustness of RANSAC homography estimations. They also enable RANSAC to reduce the amount of iterations by generating homography candidates with only two matching pairs instead of four. Furthermore, incorporating an *a-contrario* methodology [11] will, as in ORSA [42], result in: first, a threshold for inlier/outlier discrimination that is adaptive; second, a measure of the soundness of an estimation. In order to validate and measure the impact of these improvements, all modifications to RANSAC will be tested under several choices of detectors, descriptors and matchers which determine the first two steps.

The last two steps (optimal estimation and guided matching) could be seen as optional, and can be iterated until the number of correspondences is stable. Note that the local affine information could also be exploited in these steps. In the case of guided matching, the affine information has already been used to predict location and camera parameters from affine invariant detectors [13, 14] and also from local geometry estimators [62].

The rest of this paper is organized as follows. Section 2 summarizes a formal methodology for approximating locally the viewpoint changes induced by motion of real cameras. Three methods for computing those local approximations are introduced in Section 3. Section 4 presents the modified RANSAC model. The proposed methods are illustrated with experiments in Section 5. Concluding remarks are presented in Section 7.

2 Affine Maps and Homographies

As stated in [47, 59, 62], a digital image $\mathbf{u}$ obtained by any camera at infinity is modeled as $\mathbf{u} = \mathbf{S}_1\mathbb{G}_1 Au$, where $\mathbf{S}_1$ is the image sampling operator (on a unitary grid), A is an affine map, u is a continuous image and $\mathbb{G}_\delta$ denotes the convolution by a Gaussian kernel broad enough to ensure no aliasing is induced by the δ-sampling. This model takes into account the blur incurred when tilting or zooming a view. Note that $\mathbb{G}_1$ and A generally do not commute.

Let $\mathcal{A}$ denote the set of affine maps and define $Au(x) = u(Ax)$ for $A \in \mathcal{A}$, where x is a 2D vector and Ax denotes function evaluation, $A\,(x)$. We define the set of invertible orientation preserving affinities

$$\mathcal{A}^+ = \{L + v \in \mathcal{A} |\, \det(L) > 0\},$$

where L is a linear map and v a translation vector. We call $\mathcal{S}$ the set of similarity transformations, which are any combination of translations, rotations and zooms. Lastly, we define the set

$$\mathcal{A}^+_* = \mathcal{A}^+ \setminus \mathcal{S},$$

where we exclude pure similarities. As it was pointed out in [47], every $A \in \mathcal{A}^+_*$ is uniquely decomposed as

$$A = \lambda R_1(\psi) T_t R_2(\phi), \tag{1}$$

where R_1, R_2 are rotations, $T_t = \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix}$ with $t > 1$, $\lambda > 0$, $\phi \in [0, \pi)$ and $\psi \in [0, 2\pi)$. Furthermore, the above decomposition comes with a geometric interpretation (see Figure 2) where the longitude ϕ and latitude $\theta = \arccos \frac{1}{t}$ characterize the camera's viewpoint angles (or tilt), ψ parameterizes the camera roll and λ corresponds to the camera zoom. The so-called optical affine maps involving a tilt t in the z-direction and zoom λ are formally simulated by:

$$\mathbf{u} \mapsto \mathbf{S}_1 A \mathbb{G}^z_{\sqrt{t^2-1}} \mathbb{G}_{\sqrt{\lambda^2-1}} I \mathbf{u},$$

where I is the Shannon-Whittaker interpolator and the superscript z indicates that the operator takes place only in the z-direction. We denote the corresponding operator by

$$\mathbb{A} := \mathbf{S}_1 A \mathbb{G}^z_{\sqrt{t^2-1}} \mathbb{G}_{\sqrt{\lambda^2-1}} I\,.$$

The operator $\mathbb{A}$ is not always invertible and therefore its application might incur into a loss of information. We refer to [61] for an example where no optical transformation $\mathbb{A}$ is found between two views.

2.1 Local Affine Approximation of Homographies

Let $H = (h_{ij})_{i,j=1,\ldots,3}$ be the 3×3 matrix associated to the homography $\eta\,(\cdot)$. Let $\mathbf{x}$ be the homogeneous coordinates vector associated to the image point $x = (x_1, x_2)$ around which we want to

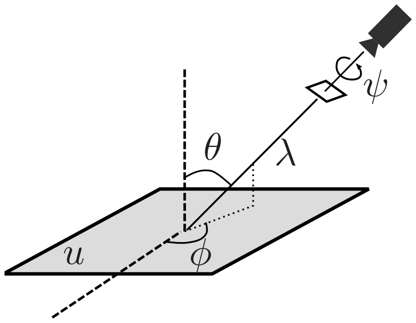

Figure 2: (From [62]) Geometric interpretation of Equation (1).

determine the local affine map. We denote by

$$y = (y_1, y_2) = \left(\frac{(H\mathbf{x})_1}{(H\mathbf{x})_3}, \frac{(H\mathbf{x})_2}{(H\mathbf{x})_3} \right) = \eta\left(x\right),$$

the image of x by the homography η.

The first order Taylor approximation of η at x leads to

$$\eta\left(x+z\right) = v + L\left(x+z\right) + o\left(\|z\|\right). \tag{2}$$

More specifically, if $x = (0,0)$, we know that

$$y_i\left(z_1, z_2\right) \quad = \quad \left(h_{i1}z_1 + h_{i2}z_2 + h_{i3}\right)\left(\tfrac{1}{h_{33}} - \tfrac{h_{31}}{h_{33}^2}z_1 + -\tfrac{h_{32}}{h_{33}^2}z_2 + o\left(\|z\|\right)\right), \quad i = 1, 2.$$

Then, by polynomial identification in the Taylor formula

$$v + L\left(z\right) \quad = \quad \tfrac{1}{h_{33}}\begin{pmatrix} h_{13} \\ h_{23} \end{pmatrix} + \left[\tfrac{1}{h_{33}}\begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix} - \tfrac{1}{h_{33}^2}\begin{pmatrix} h_{13} \\ h_{23} \end{pmatrix}\begin{pmatrix} h_{31} & h_{32} \end{pmatrix}\right]\begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

where

$$\tfrac{1}{h_{33}}\begin{pmatrix} h_{13} \\ h_{23} \end{pmatrix} \quad = \quad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

If $x \neq (0,0)$, a simple change of variables $z \to z + x$ would lead us back to the case $x = (0,0)$. Notice that the resulting homography,

$$\tilde{\eta}\left(z\right) = \eta\left(z + x\right),$$

has an associated matrix determined by columns,

$$H_{\tilde{\eta}} = \left[\begin{array}{ccc} H\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} & H\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} & H\begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix} \end{array}\right].$$

This brief computation shows that the vector v and the matrix L are given by

$$L = \begin{bmatrix} \frac{h_{11} - y_1 h_{31}}{h_{31}x_1 + h_{32}x_2 + h_{33}} & \frac{h_{12} - y_1 h_{32}}{h_{31}x_1 + h_{32}x_2 + h_{33}} \\ \frac{h_{21} - y_2 h_{31}}{h_{31}x_1 + h_{32}x_2 + h_{33}} & \frac{h_{22} - y_2 h_{32}}{h_{31}x_1 + h_{32}x_2 + h_{33}} \end{bmatrix}, \tag{3}$$

$$v = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - Lx. \tag{4}$$

This derivation allows us to compute the exact local affine approximation for a given homography. This will be useful for Section 4.2-4.3 and to assess the accuracy of our method when using annotated datasets.

3 Computing Local Affine Approximations

A SIFT-like patch is simply the square crop at the origin of some similarity transformation (translation, rotation and zoom) of the original image. This also stands true for affine invariant image matching methods, for which a patch can be considered as the square crop at the origin of a tilted version of the original image followed by some similarity transformation. If two patches are a match (i.e. their descriptors are similar), then, through the assumption of locality and Taylor's formula, we can approximate the geometry transformation by an affine map.

Consider two square patches, P_q and P_t, coming, for example, from the Gaussian pyramid of the query and target images, respectively. Let c_q and c_t be their centers expressed in image coordinates. Let also A_q be the affine map that converts from the query image domain to patch coordinates; likewise A_t converts from target to patch coordinates. Note that, in the case of SIFT-like patches, the affinities A_q and A_t are pure similarities, combining just the translation, rotation and zoom corresponding to the location, orientation and scale associated to SIFT-like keypoints. Lastly, in order to locally approximate the transformation between query and target images (centered at c_q and c_t), we only need the affine map relating P_q and P_t, denoted by A. Figure 3 illustrates the affine differences between two patches and how the query is transformed into the target. All in all, around c_q, the local affine map transforming the query into the target (in image coordinates) is written as

$$A_{q\to t} = A_t^{-1} A A_q. \tag{5}$$

The same procedure described above can be derived for Harris-Affine and Hessian-Affine region detectors [36, 37]. Only that A_q, A_t are not restrained to similarities but to affine maps in general.

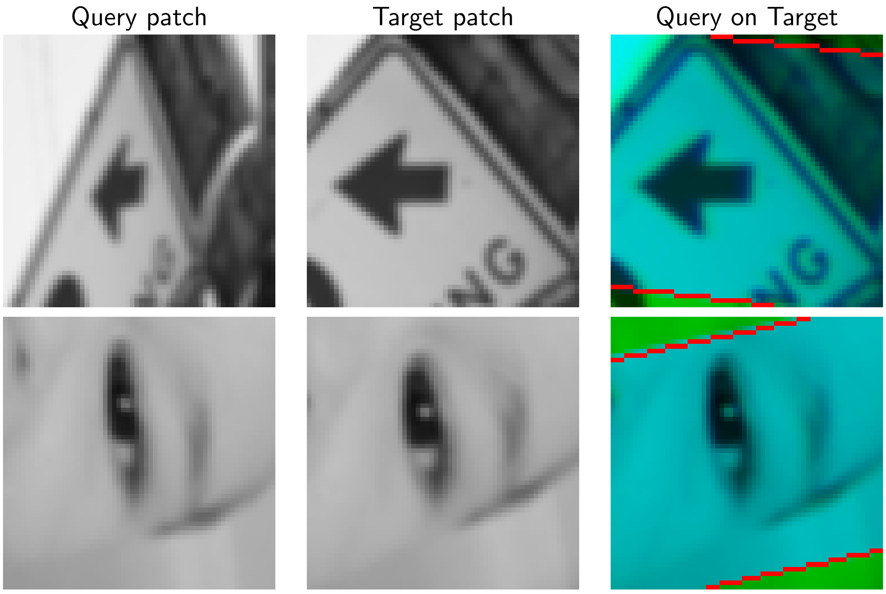

Figure 3: (From [62]). Two pairs of patches used as query and target input patches (columns 1-2). Each pair differs only by an affine map. Blue and green channels in the 3rd column correspond to the target patch and a warped version of the query patch (the red line delimits its borders); the red channel is filled with zeros.

3.1 Affine Connections between Patches

We now list different choices of methods for computing A, the affine map locally connecting the query patch to the target patch.

3.1.1 The Naive Method

If for some reason no strong tilt deformations are expected between query and target patches, then, a fair assumption would be for A to be the identity in all cases. Unfortunately, this is rarely the case for real life images. Nevertheless, it is a simplification worth trying because it entails no additional computational complexity. Formally, we set

$$A := Id. \tag{6}$$

3.1.2 The Affnet Method

The Affnet [41] method was conceived to predict normalizing ellipse shapes for single patches based on a 3-variable parametrization. Figure 4 depicts the passage from Affnet affine maps to the affine map transforming query into target (i.e. A). The connection provided by two Affnet-normalizing affine maps for the query and target patches is richer than each normalizing transformation. Indeed, for different choices of $A_1 = T_1R_1$ and $A_2 = T_2R_2$ one would need the four parameters (zoom, camera rotation, tilt and tilt direction) in Equation (1) in order to express $A_2A_1^{-1}$. However, Affnet does not estimate translations. Formally, we set

$$A := A_2\left(A_1^{-1}\mathbf{x} - A_1^{-1}\mathbf{c}\right) + \mathbf{c}, \tag{7}$$

where $\mathbf{c}$ denotes the center of patch domain and A_i are the estimated affine maps by Affnet.

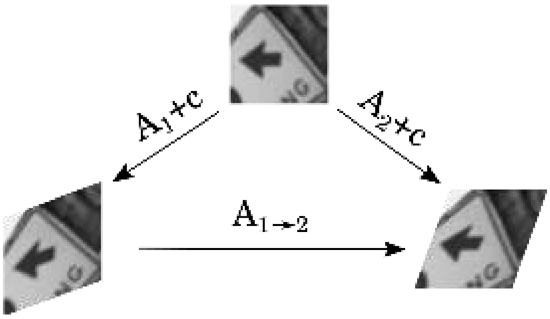

Figure 4: Passage from Affnet affine maps (A_1, A_2) to the connecting mapping $A_{1\to2}$. The center of the normalized patch (on top) corresponds to the origin in normalized coordinates.

3.1.3 The LOCATE Method

The LOCal Affine Transform Estimator (LOCATE) network presented in [62] directly estimates the affine transform A_{LOCATE} that maps the query patch into the target patch. LOCATE simultaneously tracks the direct and inverse maps which significantly improves the network performance in predicting local approximating affine maps. This network was trained exclusively with simulated patches from an affine camera model. We expect it to generalize the affine world to all sorts of geometry as long as the Taylor approximation holds. LOCATE is able to estimate all six parameters composing the affine map. Formally, we set

$$A := A_{\text{LOCATE}}. \tag{8}$$

3.2 Precision

In order to measure the precision in a realistic environment of these three methods (Naive, Affnet and LOCATE) we used the viewpoint dataset presented in [61], consisting of five pairs of images with their ground truth homographies and 3352 true matches. Notice that Equations (3)-(4) allow us to compute ground truth local affine maps around each match. Figure 5 shows the accuracy of Naive (Identity), Affnet [41] and LOCATE, represented by error density functions with respect to the affine decomposition appearing in Equation (1). Ideally, we expect a Dirac delta function centered at 0 for a perfect method. This is approximately true for the LOCATE method. Note in Figure 5 that translations from the Affnet [41] method do not quite match those from the Identity method; this difference can be explained by the connecting mapping itself (see Equation (7)) which is different from $A_2A_1^{-1}\mathbf{x}$. As expected, LOCATE is more precise than Affnet [41]. Indeed, Affnet analyzes one patch at a time, whereas LOCATE has access to both patches simultaneously. However, in practice, using Affnet involves fewer computations. This trade-off must be resolved depending on the application.

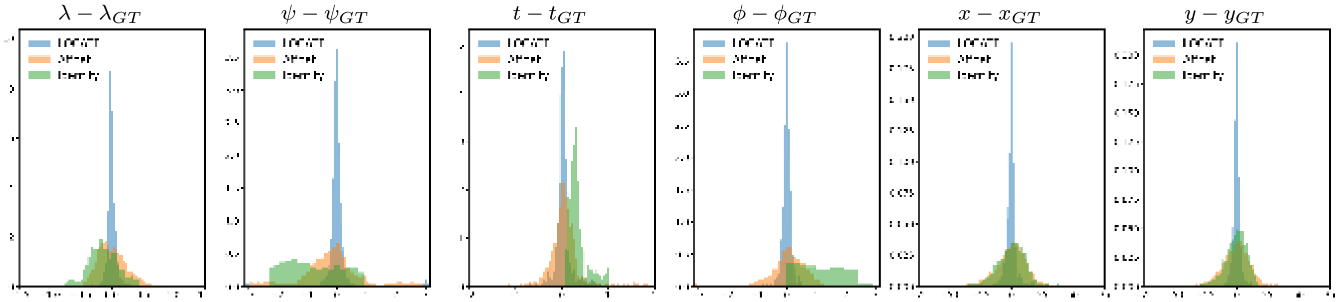

Figure 5: (From [62]) Affine error prediction in terms of the affine decomposition of Equation (1) (namely zoom λ, camera rotation ψ, tilt t, tilt direction ϕ, and translation x, y), for the LOCATE method, the Affnet method [41] and the identity map method. The [61] dataset is used; it contains 3352 patch pairs with corresponding ground truth. The sub-index GT means ground truth, conversely, no sub-index stands for estimated parameters.

4 Robust Homography Estimation

The standard RANSAC algorithm computes the parameters fitting a mathematical model from observed data in the presence of outliers. Numerous improvements have been proposed in the literature for RANSAC, see [42, 43, 54, 55], but the core idea remains the same.

In the case of homography estimation, the classic RANSAC algorithm returns the homography η_j computed at iteration j having the largest consensus of inliers among all iterations. The j-iteration can be described in two steps:

1. (Fitting) Randomly select s matches $(x^i \leftrightarrow y^i)_{i=1,\ldots,s}$ from the set of all matches (M_T) and compute the homography η_j that yields the best fit.

2. (Consensus) Determine, with respect to an error function ξ_{η_j}, the matches from M_T in consensus with η_j.

In this paper, two error functions ξ_η are used. First, the symmetric transfer error,

$$\xi^4_\eta\,(x \leftrightarrow y) := \left\| \begin{pmatrix} \eta_j(x) - y \\ x - \eta_j^{-1}(y) \end{pmatrix}_{4\times 1} \right\|_{l_2}. \tag{9}$$

Note that the 4 dimensional vector inside the norm, corresponding to the concatenation of two vectors of dimension two. The second proposal is a function measuring the classic symmetric transfer error as well as the affine coherence,

$$\xi_\eta^8(x \leftrightarrow y) := \left\| \begin{pmatrix} \eta_j(x) - y \\ x - \eta_j^{-1}(y) \\ \alpha\left(A_E^{(x\leftrightarrow y)}, A_H^{(x\leftrightarrow y)}\right) - \left(\begin{smallmatrix} 0 \\ 1 \\ 0 \\ 1 \end{smallmatrix}\right) \end{pmatrix}_{8\times 1} \right\|_{l_2}, \tag{10}$$

where the α-vector was introduced in [62] and will be revisited here in the following sections. The vector inside the norm has eight dimensions, corresponding to the concatenation of two vectors of dimension two and a vector of dimension four.

A classic way of determining inliers in step 2 of the RANSAC algorithm is by thresholding the function ξ_η. Another way is through the Number of False Alarms (NFA), based on the *a-contrario* procedure [11]. Let $\varepsilon_i = \xi_\eta(m_{l_i})$ be the ordered errors of matches in M_T with respect to the current testing homography η, i.e.,

$$\varepsilon_1 \leq \varepsilon_2 \leq \cdots \leq \varepsilon_{|M_T|}.$$

As explained in [42], the NFA of the testing homography η and its k smallest error matches from M_T is defined as,

$$NFA(\eta) = (|M_T| - s)\begin{pmatrix} |M_T| \\ k \end{pmatrix}\begin{pmatrix} k \\ s \end{pmatrix} \mathrm{P}\left(\xi_\eta(\mathbf{m}) \leq \varepsilon_k\right)^{(k-s)}, \tag{11}$$

where $\mathrm{P}\left(\xi_\eta(\mathbf{m}) \leq \varepsilon_k\right)^{(k-s)}$ corresponds to the probability of k matches to fall within an error threshold ε_k with respect to ξ_η. This probability expression is explained through independence among matches and the power $k - s$ accounts for the fact that s matches (the ones used for estimating η) are automatically inliers with respect to η. If there are k inliers, potentially all s out of them yield the correct configuration; explaining the term $\begin{pmatrix} k \\ s \end{pmatrix}$. Also, there are $\begin{pmatrix} |M_T| \\ k \end{pmatrix}$ possible subsets of M_T with k elements. The number k of inliers is usually not known in advance, so all values of k are tested (from $s+1$ to $|M_T|$), which explains the factor $(|M_T| - s)$ in Equation (11). In practice, the errors (measured by ξ_η) of all data terms are collected and sorted: ε_1 to $\varepsilon_{|M_T|}$. For each possible k, we compute the NFA as in Equation (11), and keep only the minimum of them all, provided it is below some threshold, usually set to 1 in the a-contrario methodology.

4.1 The Benchmark RANSAC

Usually the steps 1-2 only take into account point coordinates. If no further improvements are applied, this defines a base RANSAC, that we denote by $RANSAC_{base}$. Since the homography matrix has eight degrees of freedom and each match defines two equations, then the number of matches must be at least $s = 4$.

The a-contrario $\mathrm{RANSAC}_{\mathrm{base}}$ is equivalent to the ORSA Homography estimation method presented in [42]. The probability term in Equation (11) does not need to be exact for the NFA to work, and can be approximated by assuming a uniform distribution in $\mathbb{R}^4$,

$$\mathrm{P}\left(\xi_\eta^4(\mathbf{m}) \leq \varepsilon\right) \approx \frac{\varepsilon^4 \frac{\pi^2}{2}}{w_q h_q w_t h_t}, \tag{12}$$

where we find in the numerator the volume of a sphere of radius ε in $\mathbb{R}^4$ and in the denominator the volume of the set with all possible coordinates for the query (of size $[w_q, h_q]$) and target (of size $[w_t, h_t]$) images.

Algorithm 1 details the operations taking place in $\mathrm{RANSAC}_{\mathrm{base}}$.

Algorithm 1 RANSAC$_{\text{base}}$

```
input:
$M_T$ - set of all matches.
parameters:
$N_{\text{iters}}$ - Number of iterations.
$\kappa$ - Spatial inliers threshold.
useNFA - A bool stating if NFA measure should be used.
output:
$M_I$ - Set of inliers;    $\eta_{j^\star}$ - Best fitting homography.
start:
foreach $j \in \{1, \cdots, N_{iters}\}$ do
    Randomly select $m_1 = (x^1 \leftrightarrow y^1), \cdots, m_4 = (x^4 \leftrightarrow y^4)$ from $M_T$
    // Homography fitting by the Direct Linear Transformation (DLT) algorithm    ▷ see [24] p.88
    $E_i = \begin{bmatrix} x_1^i & x_2^i & 1 & & & & -y_1^i x_1^i & -y_1^i x_2^i & -y_1^i \\ & & & x_1^i & x_2^i & 1 & -y_2^i x_1^i & -y_2^i x_2^i & -y_2^i \end{bmatrix}_{2\times 9}$,    ▷ where $z^i = (z_1^i, z_2^i)$
    $E = \begin{bmatrix} E_1 \\ \vdots \\ E_4 \end{bmatrix}_{8\times 9}$
    $\vec{h}_j$ is the unit singular vector corresponding to the smallest singular value of $E$.    ▷ where $\vec{h}_j$ is a vectorized version of the matrix $H_j$ associated to the homography $\eta_j$.
    // Selection of inliers with respect to the symmetric transfer error
    $I_j = \left\{(x \leftrightarrow y) \in M_T \,|\, \xi_\eta^4\left((x \leftrightarrow y)\right) < \kappa\right\}$
    if useNFA then
        $\varepsilon_{j,\cdot} = \textsc{Sort}\left(\left\{\xi_\eta^4\left((x \leftrightarrow y)\right)\right\}_{(x \leftrightarrow y) \in I_j}\right)$
        foreach $k \in \{5, \cdots, |I_j|\}$ do
            $NFA_k^j = (|M_T| - 4)\binom{|M_T|}{k}\binom{k}{4}\left(\frac{\varepsilon_{j,k}^4 \frac{\pi^2}{2}}{w_q h_q w_t h_t}\right)^{k-4}$
if useNFA then
    $j^\star, k^\star = \arg\min_{j,k} NFA_k^j$
    if $NFA_{k^\star}^{j^\star} < 1$ then
        $M_I$ is the subset of $I^{j^\star}$ achieving the first $k^\star$ smaller values in $\varepsilon_{j^\star,\cdot}$.
    else
        $M_I = \emptyset$
else
    $j^\star = \arg\max_j |I_j|$
    $M_I = I_{j^\star}$ if $|I_{j^\star}| > 4$ else $\emptyset$
return $M_I$, $\eta_{j^\star}$
```

4.2 Homography Fitting from Local Affine Maps in RANSAC

From Section 2.1 we know how to locally approximate a homography by an affine map. Conversely, the problem of determining a homography from a set of affine maps at different locations was addressed in [3, 56]. Let $x \leftrightarrow y$ be a match and $L = (l_{ij})_{i,j=1,2}$ the linear map in Equation (2). Then the unknown homography η must satisfy

$$E_{6\times 9} \cdot \vec{h} = \vec{0}, \tag{13}$$

where $E_{6\times 9}$ is the matrix

$$\begin{bmatrix} 1 & & & & & & -y_1 - l_{11}x_1 & -l_{11}x_2 & -l_{11} \\ & 1 & & & & & -l_{12}x_1 & -y_1 - l_{12}x_2 & -l_{12} \\ & & & 1 & & & -y_2 - l_{21}x_1 & -l_{21}x_2 & -l_{21} \\ & & & & 1 & & -l_{22}x_1 & -y_2 - l_{22}x_2 & -l_{22} \\ x_1 & x_2 & 1 & & & & -y_1x_1 & -y_1x_2 & -y_1 \\ & & & x_1 & x_2 & 1 & -y_2x_1 & -y_2x_2 & -y_2 \end{bmatrix}, \tag{14}$$

and $\vec{h} = [h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{33}]^T$ is a vectorized version of the matrix H associated to η. The first four rows of $E_{6\times 9}$ are determined by Equation (3) and the last two are the classic equations derived from rewriting $\eta(x) = y$ in terms of $H\mathbf{x} = \mathbf{y}$.

Clearly, two matches with their corresponding local affine maps already over-determine the homography matrix. Indeed, putting those equations together provides us with 12 equations

$$\begin{bmatrix} E_1 \\ E_2 \end{bmatrix}_{12\times 9} \cdot \vec{h} = \vec{0},$$

where E_i denotes the matrix E appearing in Equation (13) for each match. To avoid the solution $\vec{h} = \vec{0}$ we look for a unitary vector $\vec{h}$ minimizing

$$\left\| \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} \cdot \vec{h} \right\|,$$

see [24] (algorithm 4.1, p.91) for computational details.

We call $RANSAC_{2pts}$ a $\text{RANSAC}_{\text{base}}$ version in which the classic homography fitting of step 1 is replaced by the homography fitting of this section. Note that $\text{RANSAC}_{\text{2pts}}$ only needs two samples at each iteration ($s = 2$). These samples carrying additional information consisting in affine approximations of local geometry deformations around them. The affine approximations can be provided by one of the methods introduced in Section 3.

The a-contrario $\text{RANSAC}_{\text{2pts}}$ differs little from the a-contrario $\text{RANSAC}_{\text{base}}$. Indeed, the only difference being $s = 2$ in Equation (11). Algorithm 2 details all operations taking place in $\text{RANSAC}_{\text{2pts}}$.

4.3 Affine Consensus for RANSAC Homography

When matching two image patches, the transformation that relates them may not be consistent with the global transformation of the scene. This can be due to the presence of symmetric objects or even to failures in the matching process. For instance, suppose that two patches centered at the same scene location but with incoherent rotations are identified by a matching method. The symmetry issue is easy to address as usually we should have encountered as many keypoints as degrees of symmetry around the center; so at least two rotations will correspond. However, aberrant matches are not treated by the matching method nor by RANSAC. This problem can be circumvented by imposing

Algorithm 2 RANSAC$_{\text{2pts}}$

```
input:
$M_T$ - set of all matches.
$A_T$ - Local affine approximations associated to each match in $M_T$.
parameters:
$N_{\text{iters}}$ - Number of iterations.
$\kappa$ - Spatial inliers threshold.
useNFA - A bool stating if NFA measure should be used.
output:
$M_I$ - Set of inliers;    $\eta_{j^\star}$ - Best fitting homography.
start:
foreach $j \in \{1, \cdots, N_{iters}\}$ do
    Randomly select $m_1 = (x^1 \leftrightarrow y^1)$ and $m_2 = (x^2 \leftrightarrow y^2)$ from $M_T$. Let also $L_1 + v_1, L_2 + v_2 \in A_T$ be their associated affine maps.
    // Homography fitting by the Direct Linear Transformation (DLT) algorithm    ▷ see [24] p.88
    $E_i = \begin{bmatrix} 1 & & & & & & -y_1^i - l_{11}^i x_1^i & -l_{11}^i x_2^i & -l_{11}^i \\ & 1 & & & & & -l_{12}^i x_1^i & -y_1^i - l_{12}^i x_2^i & -l_{12}^i \\ & & & 1 & & & -y_2^i - l_{21}^i x_1^i & -l_{21}^i x_2^i & -l_{21}^i \\ & & & & 1 & & -l_{22}^i x_1^i & -y_2^i - l_{22}^i x_2^i & -l_{22}^i \\ x_1^i & x_2^i & 1 & & & & -y_1^i x_1^i & -y_1^i x_2^i & -y_1^i \\ & & & x_1^i & x_2^i & 1 & -y_2^i x_1^i & -y_2^i x_2^i & -y_2^i \end{bmatrix}_{6\times 9}$,    ▷ where $z^i = \left(z_1^i, z_2^i\right)$, $L_i = (l_{lk}^i)_{l,k=1,2}$.
    $E = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}_{12\times 9}$
    $\vec{h}_j$ is the unit singular vector corresponding to the smallest singular value of $E$.    ▷ where $\vec{h}_j$ is a vectorized version of the matrix $H_j$ associated to the homography $\eta_j$.
    // Selection of inliers with respect to the symmetric transfer error
    $I_j = \left\{(x \leftrightarrow y) \in M_T \mid \xi_\eta^4\left((x \leftrightarrow y)\right) < \kappa\right\}$
    if useNFA then
        $\varepsilon_{j,\cdot} = \text{SORT}\left(\left\{\xi_\eta^4\left((x \leftrightarrow y)\right)\right\}_{(x \leftrightarrow y) \in I_j}\right)$
        foreach $k \in \{3, \cdots, |I_j|\}$ do
            $NFA_k^j = (|M_T| - 2)\begin{pmatrix} |M_T| \\ k \end{pmatrix}\begin{pmatrix} k \\ 2 \end{pmatrix}\left(\frac{\varepsilon_{j,k}^4 \frac{\pi^2}{2}}{w_q h_q w_t h_t}\right)^{k-2}$
if useNFA then
    $j^\star, k^\star = \arg\min_{j,k} NFA_k^j$
    if $NFA_{k^\star}^{j^\star} < 1$ then
        $M_I$ is the subset of $I^{j^\star}$ achieving the first $k^\star$ smaller values in $\varepsilon_{j^\star,\cdot}$.
    else
        $M_I = \emptyset$
else
    $j^\star = \arg\max_j |I_j|$
    $M_I = I_{j^\star}$ if $|I_{j^\star}| > 2$ else $\emptyset$
return $M_I$, $\eta_{j^\star}$
```

consistency between the local approximating affine maps and the testing homography proposed by RANSAC.

To impose local geometry consistency, most existing works [80, 40] propose to measure the incurred error in mapping keypoints of a match $x \leftrightarrow y$, e.g. $\|y - A(x)\| + \|x - A^{-1}(y)\|$. Unlike them, [62] proposes to enforce geometry consistency directly on the transformations parameters given by Equation (1). Finally, we redefine the consensus set of the RANSAC model by imposing geometry consistency as in [62].

Inliers are now defined as follows. Let A_E and A_H be, respectively, the estimated affine map by one of the methods introduced in Section 3 and the testing affine map computed from the testing homography (using Equation (3)). Let also $[\lambda_E, \psi_E, t_E, \phi_E]$ and $[\lambda_H, \psi_H, t_H, \phi_H]$ be, respectively, the affine parameters of A_E and A_H. We define the α-vector between A_E and A_H as

$$\alpha\left(A_E, A_H\right) \quad = \quad \left[\max\left(\tfrac{\lambda_E}{\lambda_H}, \tfrac{\lambda_H}{\lambda_E}\right), \angle\left(\psi_E, \psi_H\right), \max\left(\tfrac{t_E}{t_H}\,\tfrac{t_H}{t_E}\right), \angle\left(\phi_E, \phi_H\right)\right], \tag{15}$$

where $\angle(\cdot, \cdot)$ denotes the angular difference. To test consistency between A_E and A_H we add to the classic threshold on the Euclidean distance, four more thresholds on the α-vector. A perfect match would result in an α-vector equal to $[1, 0, 1, 0]$. If we assume independence on each dimension, the resulting probability of a match passing all thresholds is the multiplication of individual probabilities. With this in mind, we claim that rough thresholds are enough to obtain good performances and that there is no need to optimize them. Thus, we propose to further refine inliers by accepting only those matches also satisfying

$$\alpha\left(A_E, A_H\right) < \left[2, \frac{\pi}{4}, 2, \frac{\pi}{8}\right], \tag{16}$$

where the above logical operation is true if and only if it holds true for each dimension.

We call *RANSAC*

Algorithm 3 RANSAC$_{\text{affine}}$

```
input:
$M_T$ - set of all matches.
$A_T$ - Local affine approximations associated to each match in $M_T$.
parameters:
$N_{\text{iters}}$ - Number of iterations;    $\vec{\alpha}_{\max}$ - Affine inliers thresholds (a 4-sized vector).
$\kappa$ - Spatial inliers threshold;    useNFA - A bool stating if NFA measure should be used.
output:
$M_I$ - Set of inliers;    $\eta_{j^\star}$ - Best fitting homography.
start:
foreach $j \in \{1, \cdots, N_{iters}\}$ do
    Randomly select $m_1 = (x^1 \leftrightarrow y^1)$ and $m_2 = (x^2 \leftrightarrow y^2)$ from $M_T$. Let also $L_1 + v_1, L_2 + v_2 \in A_T$ be their associated affine maps.
    // Homography fitting by the Direct Linear Transformation (DLT) algorithm    ▷ see [24] p.88
    $E_i = \begin{bmatrix} 1 & & & & & & -y_1^i - l_{11}^i x_1^i & -l_{11}^i x_2^i & -l_{11}^i \\ & 1 & & & & & -l_{12}^i x_1^i & -y_1^i - l_{12}^i x_2^i & -l_{12}^i \\ & & & 1 & & & -y_2^i - l_{21}^i x_1^i & -l_{21}^i x_2^i & -l_{21}^i \\ & & & & 1 & & -l_{22}^i x_1^i & -y_2^i - l_{22}^i x_2^i & -l_{22}^i \\ x_1^i & x_2^i & 1 & & & & -y_1^i x_1^i & -y_1^i x_2^i & -y_1^i \\ & & & x_1^i & x_2^i & 1 & -y_2^i x_1^i & -y_2^i x_2^i & -y_2^i \end{bmatrix}_{6\times 9}$,    ▷ where $z^i = \left(z_1^i, z_2^i\right)$, $L_i = (l_{lk}^i)_{l,k=1,2}$.
    $E = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}_{12\times 9}$
    $\vec{h}_j$ is the unit singular vector corresponding to the smallest singular value of $E$.    ▷ where $\vec{h}_j$ is a vectorized version of the matrix $H_j$ associated to the homography $\eta_j$.
    // Selection of inliers with respect to the symmetric transfer error and the affine information
    $S_j = \left\{(x \leftrightarrow y) \in M_T \,|\, \xi_\eta^4\left((x \leftrightarrow y)\right) < \kappa\right\}$
    $T_j = \left\{(x \leftrightarrow y) \in M_T \,|\, \alpha\left(A_E^{(x\leftrightarrow y)}, A_H^{(x\leftrightarrow y)}\right) < \vec{\alpha}_{\max}\right\}$
        ▷ $A_E^{(x\leftrightarrow y)}$ is the associated affine map to $(x \leftrightarrow y)$ in $A_T$; and $A_H^{(x\leftrightarrow y)}$ is the best approximating affine map at $(x \leftrightarrow y)$ computed from the testing homography $\eta_j$ using Equation (3).
    if useNFA then
        $\varepsilon_{j,\cdot} = \textsc{Sort}\left(\left\{\xi_\eta^8\left((x \leftrightarrow y)\right)\right\}_{(x\leftrightarrow y)\in S_j}\right)$
        foreach $k \in \{3, \cdots, |S_j|\}$ do
            $NFA_k^j = (|M_T| - 2)\begin{pmatrix} |M_T| \\ k \end{pmatrix}\begin{pmatrix} k \\ 2 \end{pmatrix}\left(\frac{\varepsilon_{j,k}^8 \frac{\pi^4}{4!}}{w_q h_q w_t h_t 12^2 \pi^2}\right)^{k-2}$
if useNFA then
    $j^\star, k^\star = \arg\min_{j,k} NFA_k^j$
    if $NFA_{k^\star}^{j^\star} < 1$ then
        $M_I$ is the subset of $S^{j^\star}$ achieving the first $k^\star$ smaller values in $\varepsilon_{j^\star,\cdot}$.
    else
        $M_I = \emptyset$
else
    $j^\star = \arg\max_j |S_j \bigcap T_j|$
    $M_I = S_{j^\star} \bigcap T_{j^\star}$ if $|S_{j^\star} \bigcap T_{j^\star}| > 2$ else $\emptyset$
return $M_I$, $\eta_{j^\star}$
```

present several pairs incurring in strong viewpoint differences. In particular, viewpoint differences between pairs in the EVD dataset are extreme and most matching methods will struggle to find correct matches. Lastly, the SymB dataset consists of painting-to-photo pairs, which will be challenging for detectors, descriptors and local geometry estimators that where not intended to be used under these circumstances. All datasets include ground truth homographies that were used to verify the accuracy.

Four combinations of state-of-the-art detectors and descriptors were used as starting point for all RANSACs. These choices are: SIFT [31] + AID [61]; SIFT [31] + HardNet [39]; HessianAffine [37] + AID [61]; HessianAffine [37] + HardNet [39]. Once local features were detected and matched, then each homography estimation method was applied and we declared a success if at least 80% of inliers (in consensus with the estimated homography) were in consensus with the ground truth homography. Four metrics are reported: the number of successes; the number of correctly matched image pairs; the average number of correct inliers; and the average pixel error. Notice that all these metric indicators are computed by first thresholding the symmetric transfer error ($\xi^4_{\eta_{gt}}(\cdot) \leq \kappa$) with respect to the ground truth homography η_{gt} of all matches in consensus with the estimated homography. These last two elements are nothing more than the output: M_I, $\eta_{j^\star}$. The two steps of RANSAC (fitting and consensus) are iterated 1000 times for each of the RANSAC variants, except for ORSA that is iterated 10000 times and USAC that adapts the number of iterations at each execution.

In Section 4 we have introduced, for the proposed RANSACs, a procedure allowing to order and validate estimations with respect to a measure of statistical significance, the NFA. However, the typical thresholding for determining inliers is less costly and might be a preferred option depending on the application. With this in mind, experiments are separated in two (Subsection 5.1 and Subsection 5.2) and are presented in form of tables. Each table should be analyzed by blocks. Each block consists of a fixed matching method and dataset. The reader should pay special attention to the first two columns: total number of successes and total number of image pairs identified at least once. The best method should have these two indicators as high as possible, for most choices of matching method and dataset (i.e. blocks). In all cases, the gap between RANSAC$_{\text{base}}$ and RANSAC$_{\text{2pts}}$/RANSAC$_{\text{affine}}$ will give a measure of the improvement provided by the local affine approximations.

5.1 Fixed Thresholds for Inlier Discrimination

Table 1 shows a comparison of homography estimation methods using fixed thresholds for inlier/outlier discrimination. The performance of each RANSAC (combined with the affine approximations of Section 3.1) is tested on the four aforementioned datasets for the four choices of detectors and descriptors presented above. The RANSAC USAC [54] method is added to Table 1 in order to compare the proposed RANSACs against a well established state-of-the-art method.

The USAC [54] method shows better performances than RANSAC$_{\text{base}}$ when equipped with the HardNet [39] descriptor. This situation is inverted for the AID [61] descriptor. A plausible explanation is the higher rate of false positive matches for the AID [61] descriptor with respect to the HardNet [39] descriptor, which might be harming the performance of some steps (like the $T_{d,d}$ and Bail-Out tests, among others) present in USAC [54].

The reader will also note the performance drop of the LOCATE method in estimating affine approximations between image pairs in the case of the SymB [25] dataset. Indeed, LOCATE misreads the information when analyzing the painting-to-photo patch pairs, an invariance for which LOCATE was not trained. The Affnet [41] method is less affected by these painting-to-photo image pairs as it analyzes separately each patch, so it is the structure that provides the invariance. However, the Affnet and Naive methods show similar performances under the SymB [25] dataset.

Incorporating the affine information to the homography fitting step allowed to boost the total number of successes in retrieving ground truth homographies for almost any configuration of detector

Detector + Descriptor	Affine maps	Homography Estimator	EF dataset [82]				EVD dataset [40]				OxAff dataset [38]				SymB dataset [25]			
			S	$\boxed{33}$	inl.	AvE	S	$\boxed{15}$	inl.	AvE	S	$\boxed{40}$	inl.	AvE	S	$\boxed{46}$	inl.	AvE
SIFT + AID	None	RANSAC$_{\text{base}}$	182	19	85	6.7	17	1	53	7.1	714	39	1651	4.8	220	22	484	6.6
		USAC	176	10	144	5.3	0	0	0	-	659	35	1887	4.6	200	11	767	6.6
	LOCATE	RANSAC$_{\text{2pts}}$	367	24	91	6.7	**21**	**2**	71	6.9	783	**40**	1603	5.0	382	31	348	7.0
		RANSAC$_{\text{affine}}$	**400**	**28**	44	6.3	**21**	**2**	58	6.4	**794**	**40**	939	4.6	417	35	168	7.0
	Affnet	RANSAC$_{\text{2pts}}$	341	26	71	6.5	19	1	42	8.1	737	39	1217	4.7	365	32	252	6.9
		RANSAC$_{\text{affine}}$	350	**28**	31	6.1	19	1	28	6.9	765	**40**	539	4.3	393	35	120	6.8
	Naive	RANSAC$_{\text{2pts}}$	355	26	80	6.5	14	1	31	8.6	756	**40**	1299	4.8	402	33	297	7.0
		RANSAC$_{\text{affine}}$	381	27	49	6.5	19	3	11	9.1	721	38	816	4.7	**450**	**37**	151	7.0
SIFT + HardNet	None	RANSAC$_{\text{base}}$	500	25	48	3.3	0	0	0	-	760	38	784	2.2	580	29	114	2.9
		USAC	500	25	48	3.3	0	0	0	-	**780**	**39**	763	2.2	580	29	113	2.9
	LOCATE	RANSAC$_{\text{2pts}}$	560	28	43	3.6	0	0	0	-	**780**	**39**	764	2.3	**620**	**31**	106	3.0
		RANSAC$_{\text{affine}}$	560	28	31	3.7	0	0	0	-	**780**	**39**	532	2.2	**620**	**31**	71	3.3
	Affnet	RANSAC$_{\text{2pts}}$	560	28	43	3.5	0	0	0	-	**780**	**39**	712	2.1	**620**	**31**	94	2.9
		RANSAC$_{\text{affine}}$	560	28	26	3.4	0	0	0	-	**780**	**39**	370	2.0	**620**	**31**	59	3.1
	Naive	RANSAC$_{\text{2pts}}$	560	28	43	3.6	0	0	0	-	**780**	**39**	717	2.2	**620**	**31**	105	3.0
		RANSAC$_{\text{affine}}$	**580**	**29**	38	3.4	0	0	0	-	729	38	553	2.2	**620**	**31**	80	2.8
HessAff + AID	None	RANSAC$_{\text{base}}$	74	8	33	4.8	6	**2**	13	5.9	582	34	161	2.4	92	12	87	3.6
		USAC	63	4	68	4.9	0	0	0	-	580	29	170	2.3	67	4	135	3.3
	LOCATE	RANSAC$_{\text{2pts}}$	214	20	30	5.3	**25**	**2**	12	7.2	**693**	**38**	142	2.5	**227**	22	49	4.2
		RANSAC$_{\text{affine}}$	215	**21**	14	4.3	24	**2**	7	5.4	689	**38**	84	2.1	222	**24**	33	3.7
	Affnet	RANSAC$_{\text{2pts}}$	202	17	25	5.0	7	**2**	9	7.9	659	35	123	2.5	225	22	42	4.2
		RANSAC$_{\text{affine}}$	**216**	19	10	4.4	20	**2**	5	6.9	668	37	57	2.2	224	**24**	22	3.7
	Naive	RANSAC$_{\text{2pts}}$	185	17	27	5.0	3	1	8	7.5	660	36	126	2.5	219	22	39	4.3
		RANSAC$_{\text{affine}}$	178	17	13	4.2	13	**2**	4	6.9	657	37	58	2.2	198	20	22	3.7
HessAff + HardNet	None	RANSAC$_{\text{base}}$	418	21	27	3.8	0	0	0	-	721	37	134	1.6	384	22	36	3.6
		USAC	431	23	26	3.7	0	0	0	-	720	37	134	1.5	402	22	35	3.6
	LOCATE	RANSAC$_{\text{2pts}}$	535	27	22	3.6	0	0	0	-	**760**	**38**	127	1.6	517	26	27	3.5
		RANSAC$_{\text{affine}}$	535	27	13	3.2	0	0	0	-	**760**	**38**	78	1.5	555	28	18	2.9
	Affnet	RANSAC$_{\text{2pts}}$	516	26	21	3.5	0	0	0	-	**760**	**38**	121	1.5	460	24	28	3.3
		RANSAC$_{\text{affine}}$	543	**29**	11	3.4	**20**	**1**	2	1.9	**760**	**38**	61	1.4	**563**	**30**	13	3.5
	Naive	RANSAC$_{\text{2pts}}$	517	26	21	3.2	0	0	0	-	**760**	**38**	123	1.5	499	25	25	3.7
		RANSAC$_{\text{affine}}$	**554**	**29**	11	3.4	**20**	**1**	2	1.9	**760**	**38**	61	1.4	560	29	13	3.8

Table 1: Homography estimation performances for RANSAC USAC [54], RANSAC$_{\text{base}}$, RANSAC$_{\text{2pts}}$ and RANSAC$_{\text{affine}}$. Four combinations of detectors and descriptors are used: SIFT [31] + AID [61]; SIFT [31] + HardNet [39]; HessianAffine [37] + AID [61]; HessianAffine [37] + HardNet [39]. Affine approximations are provided by each method presented in Section 3.1; 'None' states that no affine information is provided. Each RANSAC runs for 1000 internal iterations, except for USAC that adapts the number of iterations at each execution. To measure probability of success, all RANSACs were run 20 times on resulting matches from each pair of images. Legend: S - the number of successes (bounded by 20×$\boxed{\text{number}}$); the number of correctly matched image pairs; inl. - the average number of correct inliers; AvE - the average pixel error. All these metric indicators are computed by first thresholding the symmetric transfer error with respect to the ground truth homography of all matches in consensus with the estimated homography. The $\boxed{\text{numbers}}$ of image pairs in a dataset are boxed.

and descriptor. Indeed, having decreased the sample size (from 4 to 2) has increased the probability that at least one of the 1000 random samples that were drawn while iterating is free from outliers. This implies that the homography fitting step is more likely to capture the true homography in fewer iterations. Therefore, the processing time spent in computing local approximating affine maps could be compensated later on by decreasing the number of internal iterations. Furthermore, we have observed that, in general, even if RANSAC$_{\text{affine}}$ produces less apparent inliers, the quality of those matches yields a higher probability of success for the same number of internal iterations. Moreover, the affine approximations provided by all methods presented in Section 3.1 often resulted in an added value.

5.2 Adaptive Thresholds for Inlier Discrimination

A comparison of the proposed a-contrario RANSACs is provided in Table 2. ORSA [42] is used for benchmarking. Note that ORSA runs for 10000 iterations whereas the proposed methods runs for 1000 iterations. Nevertheless, $\text{RANSAC}_{\text{2pts}}$ and $\text{RANSAC}_{\text{affine}}$ attain comparable (sometimes better) results with respect to ORSA [42]. This points out that the affine information is making up for the 9000 never-done iterations in the proposed a-contrario RANSACs.

The reader will notice in Table 1 that HessianAffine [37] + HardNet [39] combined with two of the proposed methods systematically exhibit 2 correct matches out of 3 inliers of the estimated homography; whereas in Table 2 the a-contrario methods did not validate those 3 matches to be of statistical significance.

Detector + Descriptor	Affine maps	A-contrario Homography Estimator	EF dataset [82]				EVD dataset [40]				OxAff dataset [38]				SymB dataset [25]			
			S	$\boxed{33}$	inl.	AvE	S	$\boxed{15}$	inl.	AvE	S	$\boxed{40}$	inl.	AvE	S	$\boxed{46}$	inl.	AvE
SIFT + AID	None	$\text{RANSAC}_{\text{base}}$	201	19	80	6.4	13	1	53	7.2	712	39	1539	4.4	221	21	489	6.5
		ORSA	279	21	119	6.2	19	1	172	6.8	**783**	**40**	1519	4.7	**408**	30	469	7.0
	LOCATE	$\text{RANSAC}_{\text{2pts}}$	363	26	89	6.5	21	2	80	6.9	780	**40**	1520	4.8	388	**33**	337	7.0
		$\text{RANSAC}_{\text{affine}}$	**377**	**27**	66	5.9	**22**	**2**	63	7.5	780	**40**	1208	3.9	380	31	272	6.5
SIFT + HardNet	None	$\text{RANSAC}_{\text{base}}$	460	23	48	3.2	0	0	0	-	740	37	702	1.6	500	25	114	2.6
		ORSA	**506**	**26**	44	3.0	0	0	0	-	**760**	**38**	686	1.6	579	**29**	99	2.6
	LOCATE	$\text{RANSAC}_{\text{2pts}}$	500	25	46	3.2	0	0	0	-	**760**	**38**	726	1.9	**580**	**29**	104	2.8
		$\text{RANSAC}_{\text{affine}}$	500	25	43	3.2	0	0	0	-	**760**	**38**	705	1.9	**580**	**29**	96	2.7
HessAff + AID	None	$\text{RANSAC}_{\text{base}}$	85	10	32	4.5	4	1	15	7.4	583	35	139	1.9	95	12	78	3.5
		ORSA	138	12	39	4.3	15	1	41	6.2	651	35	130	1.5	174	16	58	2.7
	LOCATE	$\text{RANSAC}_{\text{2pts}}$	**209**	**19**	30	4.9	**19**	**2**	12	6.3	**687**	**38**	133	2.1	**221**	**21**	47	4.0
		$\text{RANSAC}_{\text{affine}}$	187	18	26	4.1	10	1	10	4.6	678	37	118	1.8	215	20	41	3.4
HessAff + HardNet	None	$\text{RANSAC}_{\text{base}}$	405	21	25	3.2	0	0	0	-	716	36	123	1.3	363	20	35	3.0
		ORSA	432	22	25	3.2	0	0	0	-	716	36	123	1.3	402	**21**	32	2.9
	LOCATE	$\text{RANSAC}_{\text{2pts}}$	**453**	**23**	24	3.6	0	0	0	-	**740**	**37**	124	1.3	**415**	**21**	32	3.5
		$\text{RANSAC}_{\text{affine}}$	430	22	23	3.2	0	0	0	-	**740**	**37**	117	1.3	380	19	31	3.2

Table 2: A-contrario homography estimation performances for ORSA [42], $\text{RANSAC}_{\text{base}}$, $\text{RANSAC}_{\text{2pts}}$ and $\text{RANSAC}_{\text{affine}}$. Four combinations of detectors and descriptors are used: SIFT [31] + AID [61]; SIFT [31] + HardNet [39]; HessianAffine [37] + AID [61]; HessianAffine [37] + HardNet [39]. Affine approximations are provided by the LOCATE method presented in Section 3.1; 'None' states that no affine information is provided. Each RANSAC runs for **1000** internal iterations, except for ORSA that runs for **10000** iterations. To measure probability of success, all RANSACs were run 20 times on resulting matches from each pair of images. Legend: S - the number of successes (bounded by 20× $\boxed{\text{number}}$); the number of correctly matched image pairs; inl. - the average number of correct inliers; AvE - the average pixel error. All these metric indicators are computed by first thresholding the symmetric transfer error with respect to the ground truth homography of all matches in consensus with the estimated homography. The $\boxed{\text{numbers}}$ of image pairs in a dataset are boxed.

6 Online Demo

The online demo is equipped with two detectors and two descriptors from the state-of-the-art. They will provide the matches that will be passed as input to the RANSAC methods introduced in this paper. These are not to be considered as suggestions, but merely as the means to test RANSAC. The choice of the matching method and of the local affine estimator is left to the reader, and will depend mostly on a trade-off between speed and performance. We refer to [58] for a more detailed analysis on this trade-off.

The default parameters in this online demo are: $\kappa = 24$, $N_{iters} = 1000$, $\alpha_{max} = \left[2, \frac{\pi}{4}, 2, \frac{\pi}{8}\right]$, useNFA=*true*, detector=*SIFT*, descriptor=*AID* and LOCATE as the method computing local affine approximations. The demo receives as input two images, and a set of correspondences are computed from them with the selected matching method. Then local approximations are computed around each match. Matches and their local geometry approximations are then passed to RANSAC. If RANSAC outputs a positive amount of matches then the two images are said to match; otherwise the images are said to not match. An example of an execution of this demo is shown in Figure 6.

7 Conclusions

In this paper we reviewed three methods for estimating local affine maps between images. They provide first-order approximations of local geometry. This information is proved to be beneficial for homography estimation, for which we presented several RANSAC versions that systematically improved results in four well known datasets [82, 40, 38, 25]. The proposed RANSACs regularly improved the number of successes in retrieving the ground truth homographies with respect to a baseline RANSAC, and, in a minor degree, with respect to well-established homography estimation methods like USAC [54] and ORSA [42]. The Number of False Alarms (NFA) from the *a-contrario* procedure [11] helps us measure the soundness of estimated homographies and allows for adaptive thresholds for inlier/outlier discrimination. The computations needed for estimating local affine maps around each match can be compensated later on by reducing the number of internal iterations of these RANSAC algorithms.

Image Credits

- Images appearing in Figure 1 belong to the Extreme View Dataset[2].
- Figure 3 original images belong to the MS-COCO Dataset[3].
- Images appearing in Figure 6 © Gabriele Facciolo.

---

[2] `http://cmp.felk.cvut.cz/wbs`

[3] `https://cocodataset.org/`

(a) Resulting matches from the SIFT-AID [61] method.

(b) Transforming query into target with the best scoring homography found by the a-contrario RANSAC$_{\text{affine}}$ equipped with the LOCATE method.

(c) Matches in consensus with the above homography.

(d) 10 random matches from (c) with their estimated affine approximations.

Figure 6: Visual results associated with the online demo.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S.M. Seitz, and R. Szeliski, *Building Rome in a day*, Communications of the ACM, 54 (2011), pp. 105–112. https://doi.org/10.1145/2001269.2001293.

[2] A. Agarwala, M. Agrawala, M. Cohen, D. Salesin, and R. Szeliski, *Photographing long scenes with multi-viewpoint panoramas*, International Conference on Computer Graphics and Interactive Techniques, (2006), pp. 853–861. https://doi.org/10.1145/1141911.1141966.

[3] D. Barath and L. Hajder, *Novel ways to estimate homography from local affine transformations*, SciTePress, (2016).

[4] A. Baumberg, *Reliable feature matching across widely separated views*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1 (2000), pp. 774–781. https://doi.org/10.1109/CVPR.2000.855899.

[5] M. Bennewitz, C. Stachniss, W. Burgard, and S. Behnke, *Metric Localization with Scale-Invariant Visual Features Using a Single Perspective Camera*, European Robotics Symposium, (2006), pp. 195–209. https://doi.org/10.1007/11681120_16.

[6] M. Brown and D. Lowe, *Recognising panoramas*, in IEEE International Conference on Computer Vision (ICCV), 2003, pp. 1218–1225. https://doi.org/10.1109/ICCV.2003.1238630.

[7] M. Brown and S. Süsstrunk, *Multi-spectral SIFT for scene category recognition*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 177–184. https://doi.org/10.1109/CVPR.2011.5995637.

[8] F. Cao, J.-L. Lisani, J.-M. Morel, P. Musé, and F. Sur, *A Theory of Shape Identification*, Springer Verlag, 2008. ISBN 978-3-540-68481-7.

[9] E.Y. Chang, *EXTENT: fusing context, content, and semantic ontology for photo annotation*, International Workshop on Computer Vision Meets Databases, (2005), pp. 5–11. https://doi.org/10.1145/1160939.1160945.

[10] D. Cozzolino, G. Poggi, and L. Verdoliva, *Efficient dense-field copy–move forgery detection*, IEEE Transactions on Information Forensics and Security, 10 (2015), pp. 2284–2297. https://doi.org/10.1109/TIFS.2015.2455334.

[11] A. Desolneux, L. Moisan, and J.-M. Morel, *From Gestalt Theory to Image Analysis*, Springer, 2008. ISBN 978-0-387-74378-3.

[12] Q. Fan, K. Barnard, A. Amir, A. Efrat, and M. Lin, *Matching slides to presentation videos using SIFT and scene background matching*, ACM International Workshop on Multimedia Information Retrieval, (2006), pp. 239–248. https://dl.acm.org/doi/10.1145/1178677.1178710.

[13] E. Farhan and R. Hagege, *Geometric expansion for local feature analysis and matching*, SIAM Journal on Imaging Sciences, 8 (2015), pp. 2771–2813. https://doi.org/10.1137/140997671.

[14] E. Farhan, E. Meir, and R. Hagege, *Local Region Expansion: a Method for Analyzing and Refining Image Matches*, Image Processing On Line, 7 (2017), pp. 386–398. https://doi.org/10.5201/ipol.2017.154.

[15] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press, 1993. ISBN 978-0262061582.

[16] L Février, *A wide-baseline matching library for Zeno*, 2007. Internship report.

[17] 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 (1981), pp. 381–395. https://doi.org/10.1145/358669.358692.

[18] G. Fritz, C. Seifert, M. Kumar, and L. Paletta, *Building detection from mobile imagery using informative SIFT descriptors*, Lecture Notes in Computer Science, (2005), pp. 629–638. https://doi.org/10.1007/11499145_64.

[19] A. Geiger, J. Ziegler, and C. Stiller, *Stereoscan: Dense 3d reconstruction in real-time*, in Intelligent Vehicles Symposium (IV), IEEE, 2011, pp. 963–968. https://doi.org/10.1109/IVS.2011.5940405.

[20] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, *Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (2013), pp. 2916–2929. https://doi.org/10.1109/TPAMI.2012.193.

[21] I. Gordon and D.G. Lowe, *What and Where: 3D Object Recognition with Accurate Pose*, Lecture Notes in Computer Science, 4170 (2006), p. 67. https://doi.org/10.1007/11957959_4.

[22] J.S. Hare and P.H. Lewis, *Salient regions for query by image content*, International Conference on Image and Video Retrieval, (2004), pp. 317–325. https://doi.org/10.1007/978-3-540-27814-6_39.

[23] C. Harris and M. Stephens, *A combined corner and edge detector*, Alvey Vision Conference, 15 (1988), p. 50. https://doi.org/10.5244/C.2.23.

[24] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2003. ISBN 9780511811685.

[25] D. C. Hauagge and N. Snavely, *Image matching using local symmetry features*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 206–213. http://dx.doi.org/10.1109/CVPR.2012.6247677.

[26] T. Kadir, A. Zisserman, and M. Brady, *An Affine Invariant Salient Region Detector*, in European Conference on Computer Vision (ECCV), 2004, pp. 228–241. https://doi.org/10.1007/978-3-540-24670-1_18.

[27] B.N. Lee, W.Y. Chen, and E.Y. Chang, *Fotofiti: web service for photo management*, ACM International Conference on Multimedia, (2006), pp. 485–486. https://doi.org/10.1145/1180639.1180737.

[28] S. Leutenegger, M. Chli, and R.Y. Siegwart, *BRISK: Binary Robust invariant scalable keypoints*, in IEEE International Conference on Computer Vision (ICCV), 2011, pp. 2548–2555. https://doi.org/10.1109/ICCV.2011.6126542,.

[29] T. Lindeberg and J. Garding, *Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D brightness structure*, European Conference on Computer Vision (ECCV), (1994), pp. 389–400. https://doi.org/10.1007/3-540-57956-7_42.

[30] C. Liu, J. Yuen, and A. Torralba, *SIFT flow: Dense correspondence across scenes and its applications*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 33 (2011), pp. 978–994. https://doi.org/10.1109/TPAMI.2010.147.

[31] D. Lowe, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision (IJCV), 60 (2004), pp. 91–110.

[32] G. Loy and J.O. Eklundh, *Detecting symmetry and symmetric constellations of features*, European Conference on Computer Vision (ECCV), 2 (2006), pp. 508–521. https://doi.org/10.1007/11744047_39.

[33] E. Mair, G.D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, *Adaptive and generic corner detection based on the accelerated segment test*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6312 LNCS, 2010, pp. 183–196. https://doi.org/10.1007/978-3-642-15552-9_14.

[34] J. Matas, O. Chum, M. Urban, and T. Pajdla, *Robust wide-baseline stereo from maximally stable extremal regions*, Image and Vision Computing, 22 (2004), pp. 761–767. https://doi.org/10.1016/j.imavis.2004.02.006.

[35] K. Mikolajczyk and C. Schmid, *Indexing based on scale invariant interest points*, IEEE International Conference on Computer Vision (ICCV), 1 (2001), pp. 525–531. https://doi.org/10.1109/ICCV.2001.937561.

[36] ——, *An affine invariant interest point detector*, European Conference on Computer Vision (ECCV), 1 (2002), pp. 128–142. https://doi.org/10.1007/3-540-47969-4_9.

[37] ——, *Scale and Affine Invariant Interest Point Detectors*, International Journal of Computer Vision (IJCV), 60 (2004), pp. 63–86.

[38] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, *A Comparison of Affine Region Detectors*, International Journal of Computer Vision, 65 (2005), pp. 43–72. https://doi.org/10.1007/s11263-005-3848-x.

[39] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, *Working hard to know your neighbor's margins: Local descriptor learning loss*, in Advances in Neural Information Processing Systems, 2017, pp. 4826–4837.

[40] D. Mishkin, J. Matas, and M. Perdoch, *MODS: Fast and robust method for two-view matching*, Computer Vision and Image Understanding, 141 (2015), pp. 81–93. https://doi.org/10.1016/j.cviu.2015.08.005.

[41] D. Mishkin, F. Radenovic, and J. Matas, *Repeatability is not enough: Learning affine regions via discriminability*, in European Conference on Computer Vision (ECCV), 2018, pp. 284–300. https://doi.org/10.1007/978-3-030-01240-3_18.

[42] L. Moisan, P. Moulon, and P. Monasse, *Automatic Homographic Registration of a Pair of Images, with A Contrario Elimination of Outliers*, Image Processing On Line, 2 (2012), pp. 56–73. https://doi.org/10.5201/ipol.2012.mmm-oh.

[43] ——, *Fundamental Matrix of a Stereo Pair, with A Contrario Elimination of Outliers*, Image Processing On Line, 6 (2016), pp. 89–113. https://doi.org/10.5201/ipol.2016.147.

[44] P. Moreels and P. Perona, *Common-frame model for object recognition*, Neural Information Processing Systems, 17 (2004). https://resolver.caltech.edu/CaltechAUTHORS:20160929-122218111.

[45] ——, *Evaluation of Features Detectors and Descriptors based on 3D Objects*, International Journal of Computer Vision, 73 (2007), pp. 263–284. https://doi.org/10.1007/s11263-006-9967-1.

[46] J.M. Morel and G. Yu, *Is SIFT scale invariant?*, Inverse Problems and Imaging, 5 (2011). https://doi.org/10.3934/ipi.2011.5.115.

[47] J.-M. Morel and G. Yu, *ASIFT: A new framework for fully affine invariant image comparison*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 438–469. https://doi.org/10.1137/080732730.

[48] A. Murarka, J. Modayil, and B. Kuipers, *Building Local Safety Maps for a Wheelchair Robot using Vision and Lasers*, in Canadian Conference on Computer and Robot Vision, IEEE Computer Society Washington, DC, USA, 2006, pp. 25–25. https://doi.org/10.1109/CRV.2006.20.

[49] P. Musé, F. Sur, F. Cao, and Y. Gousseau, *Unsupervised thresholds for shape matching*, IEEE International Conference on Image Processing (ICIP), 2 (2003), pp. II–647. https://doi.org/10.1109/ICIP.2003.1246763.

[50] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.-M. Morel, *An A Contrario Decision Method for Shape Element Recognition*, International Journal of Computer Vision, 69 (2006), pp. 295–315. https://doi.org/10.1007/s11263-006-7546-0.

[51] A. Negre, H. Tran, N. Gourier, D. Hall, A. Lux, and J.L. Crowley, *Comparative study of People Detection in Surveillance Scenes*, Structural, Syntactic and Statistical Pattern Recognition, Proceedings Lecture Notes in Computer Science, 4109 (2006), pp. 100–108. https://doi.org/10.1007/11815921_10.

[52] D. Nister and H. Stewenius, *Scalable recognition with a vocabulary tree*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), (2006), pp. 2161–2168. https://doi.org/10.1109/CVPR.2006.264.

[53] Y. Pang, W. Li, Y. Yuan, and J. Pan, *Fully affine invariant SURF for image matching*, Neurocomputing, 85 (2012), pp. 6–10. https://doi.org/10.1016/j.neucom.2011.12.006.

[54] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J-M. Frahm, *USAC: a universal framework for random sample consensus*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (2013), pp. 2022–2038. https://doi.org/10.1109/TPAMI.2012.257.

[55] M. Rais, G. Facciolo, E. Meinhardt-Llopis, Morel J.-M., Buades A., and Coll B., *Accurate motion estimation through random sample aggregated consensus*, CoRR, abs/1701.05268 (2017). https://doi.org/10.48550/arXiv.1701.05268.

[56] C. Raposo and J.P. Barreto, *Theory and practice of structure-from-motion using affine correspondences*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5470–5478. https://doi.org/10.1109/CVPR.2016.590.

[57] F. Riggi, M. Toews, and T. Arbel, *Fundamental Matrix Estimation via TIP-Transfer of Invariant Parameters*, International Conference on Pattern Recognition (ICPR), (2006), pp. 21–24. https://doi.org/10.1109/ICPR.2006.588.

[58] Mariano Rodríguez, *Affine invariant image comparison*, PhD Thesis, École Normale Supérieure Paris-Saclay, Jul 2020. https://hal.archives-ouvertes.fr/tel-02954027v1.

[59] M. Rodríguez, J. Delon, and J.-M. Morel, *Covering the space of tilts. application to affine invariant image comparison*, SIAM Journal on Imaging Sciences, 11 (2018), pp. 1230–1267. https://epubs.siam.org/doi/10.1137/17M1140509.

[60] ——, *Fast affine invariant image matching*, IPOL, 8 (2018), pp. 251–281. https://doi.org/10.5201/ipol.2018.225.

[61] M. Rodríguez, G. Facciolo, R. Grompone von Gioi, P. Musé, J.-M. Morel, and J. Delon, *SIFT-AID: boosting SIFT with an affine invariant descriptor based on convolutional neural networks*, in IEEE International Conference on Image Processing (ICIP), Sep 2019. https://doi.org/10.1109/ICIP.2019.8803425.

[62] M. Rodríguez, G. Facciolo, R. Grompone von Gioi, P. Musé, and J. Delon, *Robust estimation of local affine maps and its applications to image matching*, in IEEE Winter Conference on Applications of Computer Vision (WACV), 2020. https://doi.org/10.1109/WACV45572.2020.9093646.

[63] M. Rodríguez, G. Facciolo, R. Grompone von Gioi, P. Musé, J. Delon, and J-M. Morel, *CNN-assisted coverings in the Space of Tilts: best affine invariant performances with the speed of CNNs*, in IEEE International Conference on Image Processing (ICIP), 2020. https://doi.org/10.1109/ICIP40778.2020.9191245.

[64] M. Rodríguez and R. Grompone von Gioi, *Affine invariant image comparison under repetitive structures*, in IEEE International Conference on Image Processing (ICIP), Oct 2018, pp. 1203–1207. https://doi.org/10.1109/ICIP.2018.8451060,.

[65] J. Ruiz-del Solar, P. Loncomilla, and C. Devia, *A New Approach for Fingerprint Verification Based on Wide Baseline Matching Using Local Interest Points and Descriptors*, Lecture Notes in Computer Science, 4872 (2007), p. 586. https://doi.org/10.1007/978-3-540-77129-6_51.

[66] P. Scovanner, S. Ali, and M. Shah, *A 3-dimensional SIFT descriptor and its application to action recognition*, in International Conference on Multimedia, ACM, 2007, pp. 357–360. https://doi.org/10.1145/1291233.1291311.

[67] S. Se, D. Lowe, and J. Little, *Vision-based mobile robot localization and mapping using scale-invariant features*, IEEE International Conference on Robotics and Automation (ICRA), 2 (2001). https://doi.org/10.1109/ROBOT.2001.932909.

[68] N. Snavely, S.M. Seitz, and R. Szeliski, *Photo tourism: exploring photo collections in 3D*, ACM Transactions on Graphics (TOG), 25 (2006), pp. 835–846. https://doi.org/10.1145/1141911.1141964.

[69] T. Tuytelaars and L. Van Gool, *Content-based image retrieval based on local affinely invariant regions*, International Conference on Advances in Visual Information Systems, (1999), pp. 493–500. https://doi.org/10.1007/3-540-48762-X_61.

[70] ——, *Wide baseline stereo matching based on local, affinely invariant regions*, in British Machine Vision Conference, 2000, pp. 412–425. https://doi.org/10.5244/C.14.38.

[71] ——, *Matching Widely Separated Views Based on Affine Invariant Regions*, International Journal of Computer Vision, 59 (2004), pp. 61–85.

[72] C. Valgren and A.J. Lilienthal, *SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments*, Robotics and Autonomous Systems, 58 (2010), pp. 149–156. https://doi.org/10.1016/j.robot.2009.09.010.

[73] K. Van De Sande, T. Gevers, and C. Snoek, *Evaluating color descriptors for object and scene recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32 (2010), pp. 1582–1596. https://doi.org/10.1109/TPAMI.2009.154.

[74] M. Veloso, F. Von Hundelshausen, and Paul E Rybski, *Learning visual object definitions by observing human activities*, in IEEE-RAS International Conference on Humanoid Robots, IEEE, 2005, pp. 148–153. http://dx.doi.org/10.1109/ICHR.2005.1573560.

[75] M. Vergauwen and L. Van Gool, *Web-based 3D Reconstruction Service*, Machine Vision and Applications, 17 (2005), pp. 411–426. https://doi.org/10.1007/s00138-006-0027-1.

[76] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, *Deepflow: Large displacement optical flow with deep matching*, in IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1385–1392. https://doi.org/10.1109/ICCV.2013.175.

[77] K. Yanai, *Image collector III: a web image-gathering system with bag-of-keypoints*, International Conference on World Wide Web, (2007), pp. 1295–1296. https://doi.org/10.1145/1242572.1242816.

[78] G. Yang, C.V. Stewart, M. Sofka, and C-L. Tsai, *Registration of challenging image pairs: Initialization, estimation, and decision*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 23 (2007), pp. 1973–1989. https://doi.org/10.1109/TPAMI.2007.1116.

[79] J. Yao and W.K. Cham, *Robust multi-view feature matching from multiple unordered views*, Pattern Recognition, 40 (2007), pp. 3081–3099. https://doi.org/10.1016/j.patcog.2007.02.011.

[80] Y. Zheng and D. Doermann, *Robust point matching for nonrigid shapes by preserving local neighborhood structures*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (2006), pp. 643–649. https://doi.org/10.1109/TPAMI.2006.81.

[81] H. Zhou, Y. Yuan, and C. Shi, *Object tracking using SIFT features and mean shift*, Computer Vision and Image Understanding, 113 (2009), pp. 345–352. https://doi.org/10.1016/j.cviu.2008.08.006.

[82] C. L. Zitnick and K. Ramnath, *Edge foci interest points*, in International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 359–366. https://doi.org/10.1109/ICCV.2011.6126263.