CNN-ASSISTED COVERINGS IN THE SPACE OF TILTS: BEST AFFINE INvariant PERFORMANCES WITH THE SPEED OF CNNS
Mariano Rodríguez, Gabriele Facciolo, Rafael Grompone von Gioi, Pablo Muse, Julie Delon, Jean-Michel Morel

To cite this version:
Mariano Rodríguez, Gabriele Facciolo, Rafael Grompone von Gioi, Pablo Muse, Julie Delon, et al.. CNN-ASSISTED COVERINGS IN THE SPACE OF TILTS: BEST AFFINE INvariant PERFORMANCES WITH THE SPEED OF CNNS. 2020. hal-02494121

HAL Id: hal-02494121
https://hal.archives-ouvertes.fr/hal-02494121
Preprint submitted on 28 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
CNN-ASSISTED COVERINGS IN THE SPACE OF TILTS:
BEST AFFINE INVARIANT PERFORMANCES WITH THE SPEED OF CNNS

M. Rodríguez, † G. Facciolo, † R. Grompone von Gioi, † P. Musé,§ J. Delon, † and J.-M. Morel †

† Centre Borelli, ENS Paris-Saclay, Université Paris-Saclay, CNRS, France
§ IIE, Universidad de la República, Uruguay
‡ MAP5, Université Paris Descartes, France

ABSTRACT

The classic approach to image matching consists in the detection, description and matching of keypoints. In the description, the local information surrounding the keypoint is encoded. This locality enables affine invariant methods. Indeed, smooth deformations caused by viewpoint changes are well approximated by affine maps. Despite numerous efforts, affine invariant descriptors have remained elusive. This has led to the development of IMAS (Image Matching by Affine Simulation) methods that simulate viewpoint changes to attain the desired invariance. Yet, recent CNN-based methods seem to provide a way to learn affine invariant descriptors. Still, as a first contribution, we show that current CNN-based methods are far from the state-of-the-art performance provided by IMAS. This confirms that there is still room for improvement for learned methods. Second, we show that recent advances in affine patch normalization can be used to create adaptive IMAS methods that select their affine simulations depending on query and target images. The proposed methods are shown to attain a good compromise: on the one hand, they reach the performance of state-of-the-art IMAS methods but are faster; on the other hand, they perform significantly better than non-simulating methods, including recent ones. Source codes are available at https://rdguez-mariano.github.io/pages/adimas.

Index Terms — image comparison, affine invariance, IMAS, SIFT, RootSIFT, convolutional neural networks.

1. INTRODUCTION

Image matching, which consists in deciding whether or not several images represent some common or similar objects, is a problem recognized as difficult, especially because of the viewpoint changes between images. The classic approach to image matching consists in three steps: detection, description and matching [1]. First, keypoints are detected in both images. Second, regions around these points are described by local descriptors. Finally, all these descriptors are compared and possibly matched. Both the detection and description steps are usually designed to ensure some invariance to various geometric or radiometric changes. A benefit of local descriptors is that viewpoint deformations are well approximated by affine maps. Indeed, for any smooth deformation, its first order Taylor approximation is an affine map. This observation has motivated the development of comparison methods based on local descriptors that are as affine invariant as possible.

The best established image comparison method is SIFT [1]. This method was shown in [2] to be invariant to image rotations, translations, and camera zoom-outs. SIFT has inspired numerous variations over the past 15 years [3, 4, 5]. In this paper, we refer to these methods as Scale Invariant Image Matching (SIIM). Several attempts have also been made to create local image descriptors invariant to affine transformations [6, 7, 8]. Yet the affine invariance of these SIIM methods in images acquired with real cameras is limited by the fact that optical blur and affine transforms do not commute, as shown in [9]. Thus, none of the previously mentioned descriptors can be considered fully affine invariant. In [10], RootSIFT [5] was reported to be the robustest descriptor to affine viewpoint changes (up to 60°). To overcome this limitation, several Image Matching by Affine Simulation (IMAS) solutions have been proposed: ASIFT [11], FAIR-SURF [12], MODS [13], Optimal Affine-RootSIFT [14], Affine-AC-W [15]. From them, Optimal Affine-RootSIFT was proven to be the best choice in terms of performance. The downside of simulation-based methods is the added computations.

The recent advances in deep-learning have also contributed to the development of local descriptors. Mimicking the classic process of image matching, they learn a similarity measure between image patches [17, 18]. In particular, affine invariance is currently being learned from data [19, 16]. The SIFT-AID method [19] combines SIFT keypoints with a CNN-based patch descriptor trained to capture affine invariance up to 75°. The Affnet method [16], conceived to predict normalizing ellipse shapes for single patches based on a 3-variable parametrization, was used with HardNet [20] (a CNN-based SIIM method) to create affine invariant descriptions; its authors called this method HesAffNet. The information provided by Affnet [16] can be obtained quickly but comes with a cost in preci-
profession, see [21] for more details. Still, this information concentrates in the Space of Tilts even if Affnet [16] was not trained for this task. Figure 1 shows kernel density estimations in the Space of Tilts (formally introduced in [10]) for query and target images in the ‘cat’ pair from the EVD [13] dataset. Notice the concentration around orthogonal directions in the Space of Tilts of affine maps provided by Affnet [16] from query and target images. Just by looking at those densities one can already infer that the common object to both images was seen from camera positions that differ by 90°.

As usual in matching methods involving normalization, each patch in HessAffnet [16] is normalized to a single and possibly unprecise and/or even erroneous representation. Instead, in this paper we propose not to rely on the precision nor on the existence of a precise and/or even erroneous representation. Instead, in this paper we propose to simulate common slanted views where descriptors can match. Two adaptive coverings based on Affnet [16] are introduced in Section 3. They will make way for adaptive IMAS (Affine Maps) methods. The performance of the proposed methods is illustrated with experiments in Section 4. Finally, Section 5 presents our concluding remarks.

2. AFFINE MAPS AND THE SPACE OF TILTS

Affine Maps. As stated in [9, 10], a digital image \( u \) obtained by any camera at infinity is modeled as \( u = S_i G_i A_i u \), where \( S_i \) is the image sampling operator (on a unitary grid), \( G_i \) denotes the convolution by a Gaussian kernel broad enough to ensure no aliasing by \( \delta \)-sampling, \( A \) is an affine map and \( u \) is a continuous image. This model takes into account the blur incurred when tilting or zooming a view. Notice that \( G_i \) and \( A \) generally do not commute.

Let \( A \) denote the set of affine maps and define \( A u(x) = u(Ax) \) for \( A \in A \), where \( x \) is a 2D vector and \( A x \) denotes function evaluation, \( A (x) \). We define the set of invertible orientation preserving affinities \( A^+ = \{ L + v \in A | \det(L) > 0 \} \) where \( L \) is a linear map and \( v \) a translation vector. We call \( S \) the set of similarity transformations, which are any combination of translations, rotations and zooms. Finally, we define the set \( A^+ = A^+ \setminus S \), where we exclude pure similarities. As it was pointed out in [9], every \( A \in A^+ \) is uniquely decomposed as

\[
A = \lambda R_1(\psi) T_\tau R_2(\phi),
\]

where \( R_1, R_2 \) are rotations and \( T_\tau = \begin{bmatrix} \tau & 0 \\ 0 & 1 \end{bmatrix} \) with \( \tau > 1, \lambda > 0 \), \( \phi \in [0, \pi] \) and \( \psi \in [0, 2\pi] \). Furthermore, the above decomposition comes with a geometric interpretation (see Figure 2) where the longitude \( \phi \) and latitude \( \theta = \arccos \frac{1}{2} \) characterize the camera’s viewpoint angles (or tilt). \( \psi \) parameterizes the camera spin and \( \lambda \) corresponds to the camera zoom.

The so-called optical affine maps involving a tilt \( \tau \) in the \( \phi \)-direction and zoom \( \lambda \) are formally simulated by

\[
u \mapsto S_1 A G^\phi \sqrt{\tau^2 - 1} G \sqrt{\lambda^2 - 1} I u,
\]

where \( I \) is the Shannon-Whittaker interpolator and the superscript \( \phi \) indicates that the convolution operator is 1D and has its tilt applied in the \( \phi \)-direction. We denote by

\[
\mathcal{A} := S_1 A G^\phi \sqrt{\tau^2 - 1} G \sqrt{\lambda^2 - 1} I.
\]

As demonstrated in [10], the function

\[
d : \Omega \times \Omega \rightarrow \mathbb{R}^+
\]

is a metric acting on the Space of Tilts that measures the affine distortion from a fixed affine viewpoint to surrounding affine viewpoints. These distortions affect the performance of all SIIM methods [10, 22] but most of them are able to successfully identify affine viewpoint distortions under log 1.7 for image sizes around \( 700 \times 550 \).

In the context of image matching by affine simulation (IMAS), one crucial question to answer is: What is the best set of affine transforms to apply to each image to gain full practical affine invariance? For example, green points in Figure 4-(a) represent the affine maps to be simulated on query and target images in the case of Optimal Affine-RootSIFT. Disks represent the set of affine maps that are distorted by no more than \( \log 1.7 \) (in terms of the distance in Equation 4) from the center. Notice in Figure 4-(a) that a whole zone of classes with distortions up to \( \log 4 \sqrt{2} \) is covered by the union of disks. This means that any distortion in that zone is reduced to less than \( \log 1.7 \) from at least one of the centers. This idea of reduction is the key to the success in IMAS methods, as it ensures that any strong deformation between images can be reasonably reverted so as the matching method in question is able to cope with it.
To determine an appropriate set of affine simulations to be used by the affine simulations. This motivates the use of Affnet [16] in order to mathematically determine the set of simulations presented in Optimal Affine-RootSIFT [14] corresponding to 25 affine simulations each.

However, density estimations like those in Figure 1-(b) are time consuming and would dramatically slow down the matching process. Instead, we propose to quickly analyze the affine information and then determine two reasonable sets of affine maps (for query and target) to be simulated by an IMAS method. We now present two methodologies for building meaningful small sets of optical affine simulations for IMAS methods.

Fixed tilts selection. Here we want to determine a small (if not the smallest) subset of \( S_1, \tau \) whose elements will be used to generate the simulations for the adaptive IMAS methods. This set should be such that the performance of the resulting adaptive IMAS methods is comparable to simulating the entire set \( S_1, \tau \). Algorithm 1 receives as input the information extracted by Affnet [16] from a set of patches. Then, indirectly, each of these patches will vote for a transform in \( S_1, \tau \) and return the set of affine maps to be simulated by an IMAS method.

We call Adaptive-ARootSIFT the adaptive IMAS method whose simulations are selected by Algorithm 1 and RootSIFT is used to describe patches.

Algorithm 1: Fixed Tilts Selection

\[
\begin{align*}
\text{input:} & \quad \mathcal{A} - \text{Set of normalizing affine maps provided by Affnet [16] from all patches of an image.} \\
\text{parameters:} & \quad r - \text{Tilt radius (default to 1.7),} \\
& \quad S_r - \text{Set of optimal affine simulations (default to } S_1, \tau \text{).} \\
& \quad \alpha - \text{Cover threshold (default to 0.01).} \\
\text{start:} & \quad S_{FT} = \emptyset. \quad \text{// initialization} \\
\text{foreach } S \in S_r, & \quad \text{do} \\
& \quad p = \frac{\sum_{A \in \mathcal{A}} \mathbb{1}_{d([A],[S]) \leq \log r}}{|\mathcal{A}|}. \\
& \quad \text{if } p \geq \alpha \text{ then} \\
& \quad \quad S_{FT} = S_{FT} \cup \{S\}. \\
\text{return } S_{FT}
\end{align*}
\]

Greedy selection. We can also determine the set of simulations in a greedy iterative way until some criterion is satisfied. Algorithm 2 presents the formal procedure. Notice that \( S \) in Equation 6 is the current affine map in \( \mathcal{A} \) with more close neighbors than any other. We call Greedy-ARootSIFT the adaptive IMAS method whose simulations are selected by Algorithm 2 and RootSIFT is used to describe patches.

Figure 4-(b,c) illustrates the selected simulations by Adaptive-ARootSIFT and Greedy-ARootSIFT for the cat image pair in the EVD [13] dataset. Notice that, when no OpenMP parallelization is used, both proposed methods run respectively 4 and 7 times faster than the Optimal Affine-RootSIFT [14] method. As it will be seen in our experiments, Optimal Affine-RootSIFT is still the state of the art in viewpoint performance.

3. ADAPTIVE COVERINGS

The Affnet method [16] is trained to predict affine-covariant region representations, where a patch is normalized before description, see Figure 3. The advantage of this approach is that the normalization can be obtained quickly, but at the expense of precision [21]. On the other hand, methods like ASIFT [11] optically simulate affine distortions to both query and target images in order to match them. The set of simulations presented in Optimal Affine-RootSIFT [14] correspond to an optimal \( \log 1.7 \)-covering (denoted by \( S_1, \tau \)) appearing in Figure 4-(a). When Optimal Affine-RootSIFT is applied, it has been observed that most matches come from a small subset of all the affine simulations. This motivates the use of Affnet [16] in order to determine an appropriate set of affine simulations to be used by IMAS methods. We call this general procedure the Adaptive IMAS method. As in the case of IMAS methods [10], to mathematically ensure that Adaptive IMAS works one needs to:

1. Dilate query and target density estimations in the Space of Tilts by a factor of \( \sqrt{r} \), where \( r \) is the radius corresponding to the maximal viewpoint tolerance of the SIIM method (we assume \( r = 1.7 \) for RootSIFT);

2. Find two sets of affine maps covering both dilated regions in step 1. We assume that the dilation in step 1 is already taking place thanks to the already jittered information provided by Affnet [16].
Table 1: Image matching performances on three viewpoint datasets. After matching each image pair, RANSAC-USAC [24] is run 100 times to measure its probability of success in retrieving corresponding ground truth homographies. Legend: S - the number of successes (bounded by $100 \times \text{number}$); the number of correctly matched image pairs; inl. - the average number of correct inliers; ET - the average elapsed time in seconds. Hardware settings: (CPU) Intel i7-6700HQ 2.60GHz; (GPU) NVIDIA Quadro M5000M. OpenMP parallelization with 8 threads. + Uses GPU.

### 4. EXPERIMENTS

We now focus on the evaluation of the adaptive IMAS methods. Table 1 shows performances on three known datasets for homography estimation in the presence of viewpoint changes. All datasets include groundtruth homographies that were used to verify accuracy. First, correspondences from a matching method are obtained, then Adaptive-ARootSIFT [14] is applied and we declared a success if at least 80% of keypoints and, as in [16], incorporates the HardNet [20] descriptor. The average number of simulations in Greedy-ARootSIFT has halved with respect to Adaptive-ARootSIFT. This last fact is not quite perceived in execution times of Table 1 due to parallelism but is best appreciated in Figure 4 where parallelism was deactivated.

### 5. CONCLUSION

In this paper we show that Image matching by affine simulation (IMAS) methods are still the state of the art in matching images involving strong viewpoint differences. We observe that the information provided by Affnet [16] is valuable in determining convenient simulations to be used in IMAS methods. The resulting adaptive IMAS methods yield a substantial acceleration with respect to classic IMAS methods without sacrificing performance. Also, Equation 5 provides a natural order to simulations appearing in Optimal Affine-RootSIFT [14] and Greedy-ARootSIFT [14] are higher than non-simulating methods but still considerably faster than ASIFT [11].
6. REFERENCES


