# 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

# 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, Universit Paris-Saclay, CNRS, France
§ IIE, Universidad de la República, Uruguay
‡ MAP5, Université Paris Descartes, France

**ABSTRACT**

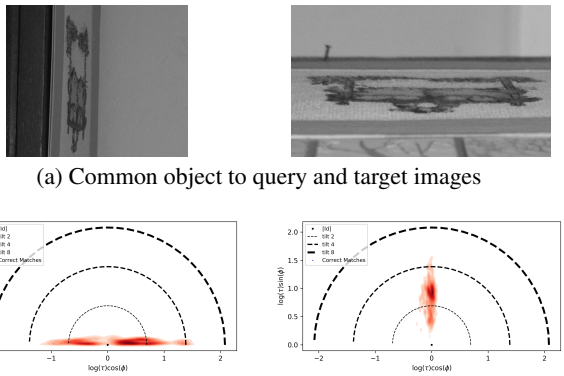
The classic approach to image matching consists in the detection, description and matching of keypoints. In the description, the local information surrounding the keypoint is encoded. This locality enables affine invariant methods. Indeed, smooth deformations caused by viewpoint changes are well approximated 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 $60^\circ$). 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.



(a) Common object to query and target images

(b) Kernel density estimations.

**Fig. 1**: Kernel density estimations in the Space of Tilts of affine maps extracted by Affnet [16] for both images in the 'cat' pair from the EVD [13] dataset.

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 $75^\circ$. 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-

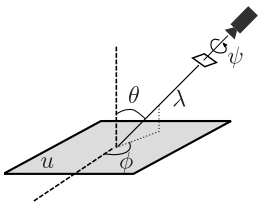



**Fig. 2**: Geometric interpretation of equation (1).

sion, 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 $90^\circ$.

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 single affine normalizing map. We prefer to compute a finite set of possible normalizing representations for each patch based on all the affine information extracted by Affnet [16]. In practice, Affnet [16] predictions will be used to select convenient affine transformations to be tested in IMAS methods. This leads to a substantial boost in IMAS speed without sacrificing performance.

The rest of this paper is organized as follows. Section 2 summarizes a formal methodology for handling local viewpoint changes induced by real cameras. Two adaptive coverings based on Affnet [16] are introduced in Section 3. They will make way for adaptive IMAS 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 $\mathbf{u}$ 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), $\mathbb{G}_\delta$ 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 $\mathbb{G}_1$ 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. Finally, we define the set $\mathcal{A}^+_* = \mathcal{A}^+ \setminus \mathcal{S}$, where we exclude pure similarities. As it was pointed out in [9], every $A \in \mathcal{A}^+_*$ is uniquely decomposed as

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

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 $\theta = \arccos \frac{1}{\tau}$ 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:

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

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

$$\mathbb{A} := \mathbf{S}_1 A \mathbb{G}^\phi_{\sqrt{\tau^2-1}} \mathbb{G}_{\sqrt{\lambda^2-1}} I\,. \tag{3}$$

If $\lambda > 1$ or $\tau > 1$, the operator $\mathbb{A}$ is not invertible and therefore incurs in information loss. This means that it is not be possible to recover the frontal image from a slanted view. Instead, IMAS methods try to simulate common slanted views where descriptors can match.

**The Space of Tilts.** The Space of Tilts, denoted by $\Omega$ and formally introduced in [10], is a quotient space where each class represents a set of affine maps with equal tilt and tilt direction (parameters $\tau$ and $\phi$ from Equation 1) and includes all possible camera spins and zooms (parameters $\psi$ and $\lambda$ from Equation 1). This space focuses on the last part $T_\tau R_2$ of the decomposition (1) because it is the one that is imperfectly dealt with by most SIIM methods. Image descriptors like those proposed in the SIFT method are invariant to similarities (translations, rotations and zooms), which in terms of the camera position interpretation (see Figure 2) correspond to a fronto-parallel motion of the camera, a spin of the camera and to an optical zoom.

We say that two classes $[A]$ and $[B]$ in the Space of Tilts are equal if and only if $T_{\tau(A)} R_{\phi(A)} = T_{\tau(B)} R_{\phi(B)}$, where each side in this equation represents the last part of the decomposition of Equation 1 for $A$ and $B$. Clearly, the Space of Tilts can be parametrized by picking representative affine maps (of the form $T_\tau R_\phi$) from each class as

$$\Omega = [Id] \bigcup \left\{ \bigcup_{(\tau,\phi) \in ]1,\infty[ \times [0,\pi[} [T_\tau R_\phi] \right\}.$$

As demonstrated in [10], the function

$$d : \begin{cases} \Omega \times \Omega & \to & \mathbb{R}_+ \\ ([A], [B]) & \mapsto & \log\left(\tau\left(BA^{-1}\right)\right) \end{cases}, \tag{4}$$

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.

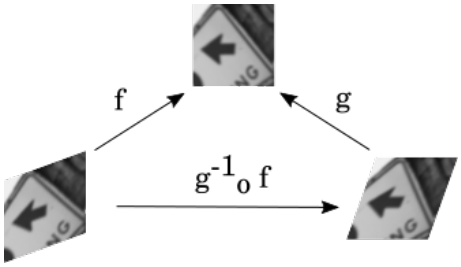
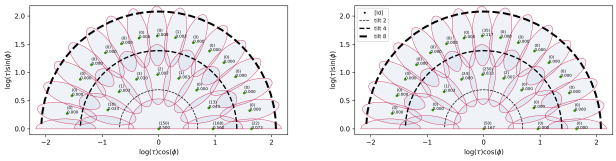
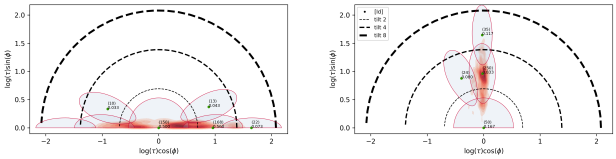


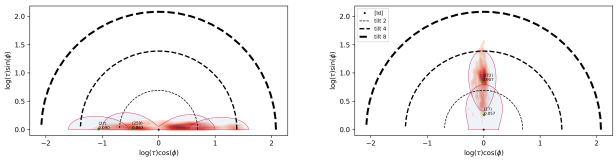

**Fig. 3**: Sketch of an ideal normalization procedure. $f, g$ two normalizing affine maps.



(a) Optimal Affine-RootSIFT [14]. Execution time **10.14s**$^\star$. Corresponding to 25 affine simulations each.



(b) Adaptive-ARootSIFT. Execution time **2.49s**$^\star$. Corresponding to 5 and 4 affine simulations.



(c) Greedy-ARootSIFT. Execution time **1.28s**$^\star$. Corresponding to 2 and 2 affine simulations.

**Fig. 4**: Proposed affine simulations for the 'cat' image pair from the EVD [13] dataset. $\star$ OpenMP parallelization was deactivated to truly measure complexity.

## 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 $\mathcal{S}_{1.7}$) 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].

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 $\mathcal{S}_{1.7}$ 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 $\mathcal{S}_{1.7}$. 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 $\mathcal{S}_{1.7}$ 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

**input:**
$\mathcal{A}$ - Set of normalizing affine maps provided by Affnet [16] from all patches of an image.
**parameters:**
$r$ - Tilt radius (default to 1.7).
$S_r$ - Set of optimal affine simulations (default to $S_{1.7}$).
$\alpha$ - Cover threshold (default to 0.01).
**start:**
$\mathcal{S}_{FT} = \emptyset$. // initialization
**foreach** $S \in \mathcal{S}_r$ **do**

$$p = \frac{\sum_{A \in \mathcal{A}} \mathbb{1}_{d([A],[S]) \leq \log r}}{|\mathcal{A}|}. \quad (5)$$

**if** $p \geq \alpha$ **then**
$\mathcal{S}_{FT} = \mathcal{S}_{FT} \bigcup \{S\}$.
**return** $\mathcal{S}_{FT}$

**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.

| Matching method | SIFT-AID dataset [19] | | | | | | EVD dataset [13] | | | | | | OxAff dataset [23] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | S | 5 | inl. | $N_q$ | $N_t$ | ET | S | 15 | inl. | $N_q$ | $N_t$ | ET | S | 40 | inl. | $N_q$ | $N_t$ | ET |
| SIFT-AID [19] ⋆ | **500** | **5** | 476 | 1.0 | 1.0 | 4.48 | 100 | 1 | 159 | 1.0 | 1.0 | 4.32 | 3794 | 38 | 1539 | 1.0 | 1.0 | 7.96 |
| RootSIFT [5] | 400 | 4 | 243 | 1.0 | 1.0 | 1.27 | - | - | - | - | - | - | 3900 | 39 | 1119 | 1.0 | 1.0 | 1.56 |
| HesAffNet [16] ⋆ | 491 | **5** | 241 | 1.0 | 1.0 | 1.05 | 228 | 4 | 50 | 1.0 | 1.0 | 1.45 | **4000** | **40** | 576 | 1.0 | 1.0 | 1.20 |
| ASIFT [11] | 400 | 4 | 551 | 41.0 | 41.0 | 33.04 | 751 | **9** | 129 | 41.0 | 41.0 | 25.54 | **4000** | **40** | 5697 | 41.0 | 41.0 | 48.68 |
| Optimal Affine-RootSIFT [14] | **500** | **5** | 685 | 25.0 | 25.0 | 5.66 | **768** | **9** | 186 | 25.0 | 25.0 | 4.96 | **4000** | **40** | 2794 | 25.0 | 25.0 | 8.12 |
| Adaptive-ARootSIFT ⋆ | **500** | **5** | 382 | 5.8 | 5.6 | 2.07 | 664 | 8 | 115 | 6.5 | 6.3 | 2.66 | **4000** | **40** | 1711 | 5.4 | 5.0 | 2.67 |
| Greedy-ARootSIFT ⋆ | 438 | **5** | 315 | 2.6 | 2.4 | 1.82 | 419 | 5 | 117 | 3.1 | 3.1 | 2.36 | **4000** | **40** | 1099 | 2.5 | 2.1 | 2.28 |

**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 $100\times$ number ); the number of correctly matched image pairs; inl. - the average number of correct inliers; The numbers of image pairs in a dataset are boxed; $N_q, N_t$ - the average number of simulated affine maps on query and target; 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.

**Algorithm 2:** Greedy Selection

**input:**
$\mathcal{A}$ - Set of normalizing affine maps provided by Affnet [16] from all patches of an image.
**parameters:**
$r$ - Tilt radius (default to 1.7).
$\alpha$ - Cover threshold (default to 0.05).
**start:**
$\tilde{\mathcal{A}} = \mathcal{A}, \mathcal{S}_G = \emptyset$. // initialization
**while** $|\tilde{\mathcal{A}}| \geq \alpha|\mathcal{A}|$ **do**

$$S = \arg\max_{S \in \tilde{\mathcal{A}}} \sum_{A \in \tilde{\mathcal{A}}} \mathbb{1}_{d([A],[S]) \leq \log r}. \quad (6)$$

$\mathcal{S}_G = \mathcal{S}_G \bigcup \{S\}$.
$\tilde{\mathcal{A}} = \tilde{\mathcal{A}} \setminus \{\, [A] \in \Omega \mid d([A],[S]) \leq \log r \,\}$.
**return** $\mathcal{S}_G$

## 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 RANSAC-USAC [24] is applied and we declared a success if at least $80\%$ of inliers (in consensus with the estimated homography) were in consensus with the groundtruth homography. RANSAC-USAC [24] was run 100 times to measure the probability of success in retrieving the corresponding ground truth homographies. Six metrics are reported: the number of successes; the number of correctly matched image pairs; the average number of correct inliers; the average number of affine simulations for query and target; and the average elapsed time in seconds. A perfect method would achieve the maximum number of successes in retrieving the groundtruth homography while being as fast as possible; where this maximum number of successes equals the number of images in the dataset times a hundred. A large number of matches is not an indicator of a method's good performance but can be used as tiebreaker measure if two methods are equally good in identifying geometric models.

As was been pointed out in [19], IMAS methods benefit from lots of keypoints that come exclusively from simulated versions of the input images. Indeed, SIIM detectors themselves are not affine invariant. Therefore, the more affine simulations in an IMAS method, the larger amount of matches it will possibly recognize. Notice in Table 1, for the OxAff dataset [23], that Optimal Affine-RootSIFT [14] has far fewer matches on average than ASIFT [11]. However, as previously stated, the number of matches might be misleading about the method's true performance. Table 1 points out that Optimal Affine-RootSIFT [14] performs better than ASIFT [11] in two datasets; indeed, the former method has more successes in retrieving groundtruth homographies (i.e. larger probability of success) with even one more identified pairs of images in the SIFT-AID dataset [19]. With this in mind, we can declare Optimal Affine-RootSIFT [14] to be state of the art in viewpoint invariant image matching. On the other hand, execution times of Optimal Affine-RootSIFT [14] are higher than non-simulating methods but still considerably faster than ASIFT [11].

Table 1 shows that adaptive IMAS methods provide a good compromise between performance and speed. Adaptive-ARootSIFT attains the same level of performance of Optimal Affine-RootSIFT [14] (best in all three datasets) in successfully identifying groundtruth homographies while reducing by half the average computing time with best case scenario reduced by four. Even if not as fast as Affnet [16], Adaptive-ARootSIFT provides a remarkable boost in successes and identified image pairs with respect to the former method, highlighted in the EVD [13] dataset. HessAffnet [16] was forced to detect 2000 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 which will be used in future work to create IMAS methods that gradually incorporate simulations on demand and stop as soon as a significant geometric model has been identified.

## 6. REFERENCES


[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[2] J. M. Morel and G. Yu, "On the consistency of the SIFT Method," Tech. Rep. Prepublication, to appear in Inverse Problems and Imaging (IPI), CMLA, ENS Cachan, 2008.

[3] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," *CVPR*, vol. 2, pp. 506–513, 2004.

[4] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *ECCV*, vol. 1, pp. 404–417, 2006.

[5] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, 2012, pp. 2911–2918.

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *IVC*, vol. 22, no. 10, pp. 761–767, 2004.

[7] P. Musé, F. Sur, F. Cao, and Y. Gousseau, "Unsupervised thresholds for shape matching," *ICIP*, 2003.

[8] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *TPAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[9] J.-M. Morel and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[10] M. Rodriguez, J. Delon, and J.-M. Morel, "Covering the space of tilts. application to affine invariant image comparison," *SIIMS*, vol. 11, no. 2, pp. 1230–1267, 2018.

[11] G. Yu and J.-M. Morel, "ASIFT: An Algorithm for Fully Affine Invariant Comparison," *IPOL*, vol. 1, pp. 1–28, 2011.

[12] Y. Pang, W. Li, Y. Yuan, and J. Pan, "Fully affine invariant SURF for image matching," *Neurocomputing*, vol. 85, pp. 6–10, 2012.

[13] D. Mishkin, J. Matas, and M. Perdoch, "MODS: Fast and robust method for two-view matching.," *CVIU*, vol. 141, pp. 81–93, 2015.

[14] M. Rodriguez, J. Delon, and J.-M. Morel, "Fast affine invariant image matching," *IPOL*, vol. 8, pp. 251–281, 2018.

[15] M. Rodriguez and R. Grompone von Gioi, "Affine invariant image comparison under repetitive structures," in *ICIP*, Oct 2018, pp. 1203–1207.

[16] D. Mishkin, F. Radenovic, and J. Matas, "Repeatability is not enough: Learning affine regions via discriminability," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 284–300.

[17] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *CVPR*, 2015, pp. 4353–4361.

[18] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *JMLR*, vol. 17, no. 1-32, pp. 2, 2016.

[19] M. Rodriguez, 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 *ICIP*, Sep 2019.

[20] 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.

[21] M. Rodriguez, G. Facciolo, R. Grompone von Gioi, P. Musé, and J. Delon, "Robust estimation of local affine maps and its applications to image matching," in *WACV*, 2020.

[22] M. Karpushin, *Local features for RGBD image matching under viewpoint changes*, Ph.D. thesis, 2016.

[23] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, R. Kadir, and L. Van Gool, "A comparison of affine region detectors," *International journal of computer vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[24] 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*, vol. 35, no. 8, pp. 2022–2038, 2013.