Robust estimation of local affine maps and its applications to image matching

Mariano Rodríguez, Gabriele Facciolo, Rafael Grompone von Gioi, Pablo Muse, Julie Delon

Robust estimation of local affine maps and its applications to image matching

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

† CMLA, ENS Paris-Saclay, 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. This defines a zero-order approximation of the mapping between two images, determined by corresponding point coordinates. 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. In this work, we propose a Local Affine Transform Estimator (LATE) method based on neural networks. We show that LATE drastically improves the accuracy of local geometry estimation between images when compared to the state of the art. The second contribution of this paper consists of two modifications to the RANSAC algorithm, that use LATE to improve the homography estimation between a pair of images. Our experiments show that these approaches outperform RANSAC for different choices of image descriptors and image datasets, and permit to increase the number of correctly matched image pairs in challenging matching databases.

1 Introduction

A physical object with smooth or piecewise smooth boundary captured by real cameras at different positions undergoes smooth apparent deformations. These regular deformations are locally well approximated by affine transforms of the image plane; indeed, for any smooth deformation, its first order Taylor approximation is an affine map. By focusing on local image regions, or patches, the perspective changes of objects can therefore be modeled by affine image deformations.

This observation has motivated the development of comparison methods based on local descriptors that are as affine invariant as possible. The problem of constructing affine invariant image descriptors by using an affine Gaussian scale space, which is equivalent to simulating affine distortions followed by the heat equation, has a long history starting with [9, 3, 11, 12]. The idea of affine shape adaptation was used as a basis for the work on affine invariant interest points and affine invariant matching in [12, 2, 15, 16, 36, 35, 34]. Finally, the detectors MSER (Maximally Stable Extremal Region) [14] and LLD (Level Line Descriptor) [23, 24, 4] both rely on image level lines.

Yet, the affine invariance of these descriptors in images acquired with real cameras is limited by the fact that optical blur and affine transforms do not commute, as shown in [22]. To overcome this limitation, the authors of [22] propose to optically simulate affine transformations. This idea was also exploited in [25, 18, 29, 31] and more recently by the SIFT-AID method [32], which combines SIFT keypoints with a CNN-based patch descriptor trained to capture affine invariance. Another recent possibility to obtain affine invariance is by learning affine-covariant region representations [19], where a patch is normalized before description. The latter method is the state-of-the-art in image matching under strong viewpoint changes.

CNN-based geometric matching between images has also been tested for the case of affine and homography transformations [28, 5]. In [28], the POOL4 layer of the VGG-16 network [33] was used for acquiring features from images and correlation maps fed to a regression network that outputs the best affine transform that fits the query to the target image. In a direct approach, the authors of [5] trained a network to estimate the homography relating the query to the target image. Both [28, 5] were trained on synthetically generated images, but neither of them took into account the blur caused by camera zoom-out or tilt.

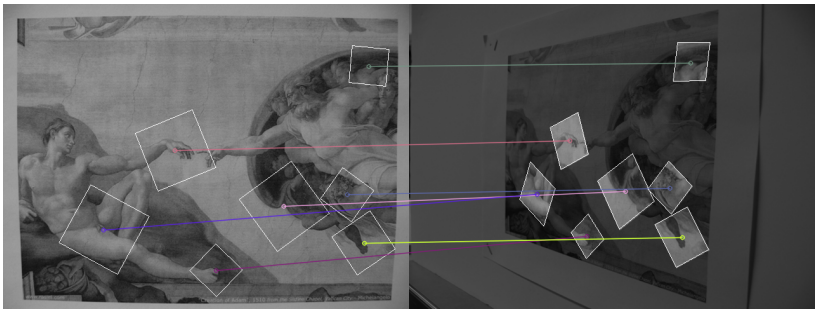

Figure 1: Some correspondences together with local affine maps estimated by the proposed LATE network. Patches on the target are warped versions of their corresponding query patch.

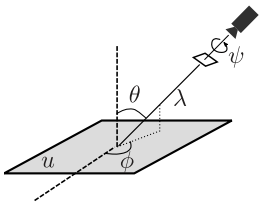

Figure 2: Geometric interpretation of equation (1).

In this paper we propose a Local Affine Transform Estimator (LATE) based on neural networks. The network architecture, which is derived from the one in [5], takes two 60×60 patches as input and estimates both the direct and inverse affine maps relating the two patches. The two-way estimation leads to an increase in the robustness of the method. While a set of correspondences represents a zero-order approximation of the geometry deformation relating two images, when local affine maps are provided a first-order approximation (i.e. tangent planes) is obtained; see Figure 1. Undoubtedly, the higher the order in the approximation, the more accurate the representation, therefore leading to better results.

Robust geometry estimation is one of those areas that can exploit the first-order approximation proposed by the LATE method. We present two modifications of the RANSAC (RAndom SAmple Consensus) algorithm [6] that improve the discrimination power in homography estimation from a set of SIFT-like matches. Briefly, those modifications consist in: first, a new fitting of homographies from local affine maps where only 2 matches are needed instead of 4 for the zero-order approximation; second, a reformulation of the consensus set (inliers) relying on the local geometry.

The rest of this paper is organized as follows. Section 2 summarizes a formal methodology for simulating local viewpoint changes provoked by real cameras, as required for training our network. The LATE method is introduced in Section 3. Section 4 presents our modified RANSAC algorithm. The use of the proposed methods is illustrated with experiments in Section 5. Finally, Section 6 present our concluding remarks.

2 On Affine Maps

As stated in [22, 30], a digital image $\mathbf{u}$ obtained by any camera at infinity is modeled as $\mathbf{u} = \mathbf{S}_1 \mathbb{G}_1 A u$, 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 by δ-sampling. 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 $\mathcal{A}^+ = \{L + v \in \mathcal{A} | \det(L) > 0\}$ 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 was pointed out in [22], 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 and $T_t = \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix}$ 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:

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

where I is the Shannon-Whittaker interpolator and the superscript z indicates that the operator takes place only in the z-direction. We denote 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 a loss of information. We refer to [32] for an example where no optical transformation $\mathbb{A}$ is found between two views. With this in mind, we adopt the same data generation scheme proposed for training the affine invariant descriptors in [32]. That is, given an image $\mathbf{u}$ and a pair of random affine transformations $\mathbb{A}_1$ and $\mathbb{A}_2$, we simulate affine views $\mathbf{u}_1 = \mathbb{A}_1(\mathbf{u})$ and $\mathbf{u}_2 = \mathbb{A}_2(\mathbf{u})$. Simulations involve maximal viewpoint angles of 75° with respect to $\mathbf{u}$. As for [32], the MS-COCO [10] dataset will provide instances of $\mathbf{u}$ in training and validation. Patch pairs seeing the same scene from $\mathbf{u}_1$ and $\mathbf{u}_2$ are said to belong to the same *class* and will be used to train the networks.

2.1 Local affine approximation of homographies

Let $H = (h_{ij})_{i,j=1,\cdots,3}$ 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 determine the local affine map. We denote by $y = (y_1, y_2) = \left(\frac{(H\mathbf{x})_1}{(H\mathbf{x})_3}, \frac{(H\mathbf{x})_2}{(H\mathbf{x})_3} \right) = \eta(x)$ the image of x by the homography η.

The first order Taylor approximation of η at x leads to

$$\eta(x + z) = v + L(x + z) + o(\|z\|), \tag{3}$$

where a brief computation shows that the vector v and the matrix L are determined through the following system of equations:

$$\begin{bmatrix} h_{11} - y_1 h_{31} & h_{12} - y_1 h_{32} \\ h_{21} - y_2 h_{31} & h_{22} - y_2 h_{32} \end{bmatrix} = (h_{31} x_1 + h_{32} x_2 + h_{33}) L, \tag{4}$$

$$v = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - Lx. \tag{5}$$

3 Local Affine Transform Estimator

In this section we present the Local Affine Transform Estimator (LATE) network whose architecture is inspired by [5]. Unfortunately, the network as it is used in [5] often incurs in wrong geometry estimations in the presence of strong blur or tilt, even when trained for this task. To address this issue, LATE estimates geometry in both directions (direct and inverse) and aggregates all the information afterwards. As will be shown in Section 5, this small architectural modification significantly improves the network performance.

The LATE architecture, see Figure 3, consists of 4 blocks of two convolutional layers each followed by batch normalization and ReLU activations. The first block receives as input two patches belonging to the same class in the form of a two channel image. Between each block a max-pooling layer is introduced. A 2D spatial dropout with a probability 0.5 is applied after the last convolutional layer. Finally, two fully connected layers are in charge of the final regression steps. The last layer outputs a vector of dimension 16, corresponding to the coordinates of eight points, the four transformed patch corners in both directions.

As argued in [32], the affine approximation holds locally, which suggests the use of small patch sizes; on the other hand, small patches entail less information, leading to insufficient geometry anchors. As a compromise, we set the patch size to 60×60, which provides a good balance between locality and sufficient viewpoint information.

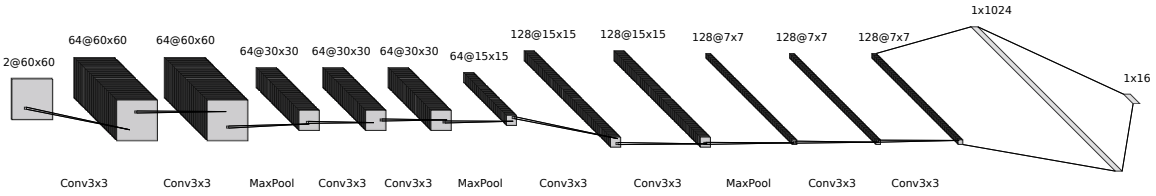

Figure 3: The proposed LATE network architecture. The last two layers are fully connected.

3.1 Training

The LATE network is trained, as in [5], with pairs of patches belonging to the same class and involving small differences in translation, rotation and zoom (but not in tilt). In this way, the networks will specialize in the task of estimating affine maps between patches that ideally incur in high similarity scores for a perfect matching method based on SIFT-like keypoints. Both networks are trained from scratch until reaching a *plateau* for the loss in training and validation. While training we simulate contrast changes on all input patches.

Let A_1, A_2 denote two affine maps and $\mathbb{A}_1$, $\mathbb{A}_2$ their respective optical simulations. We assume A_1, A_2 involve small perturbations in terms of similarity transformations. Let P_1 and P_2 be two square 60×60-patches centered at the origin of $\mathbb{A}_1(\mathbf{u})$ and $\mathbb{A}_2(\mathbf{u})$, respectively. Let $X = [x_1, x_2, x_3, x_4]$, where x_i are the 2D coordinates of the four corners of a patch following a fixed order. We also define 4- and 8-point ground truth parameterizations respectively for the network [5] and the LATE network,

$$\begin{array}{rcl} X^4 & := & A_1^{-1} A_2(X), \\ X^8 & := & \left[A_1^{-1} A_2(X),\, A_2^{-1} A_1(X)\right], \end{array} \tag{6}$$

where $[\cdot,\cdot]$ denotes the concatenation of both vectors. Let $\mathcal{N}^k$ be one of the presented networks with k-point parameterization. Then the loss is defined by

$$\sum_{i=1}^{k} \|\mathcal{N}^k(P_1, P_2)_i - X_i^k\|_{L_2}, \tag{7}$$

where the sub-index i represents the i-th element of the vector.

3.2 From patches in the Gaussian pyramid to local affine maps

The training process described above allows the networks to be coupled with matching methods like SIFT [13], RootSIFT [1], SIFT-AID [32] among many others. Indeed, a SIFT like patch is simply the square crop at the origin of some similarity transformation (translation, rotation and zoom) of the original image; additionally, patches corresponding to matched keypoints should suffer small similarity deformations but possibly strong tilts.

Let P_q and P_t be two square 60×60-patches coming respectively from the Gaussian pyramid of the query and target images. Let c_q and c_t be their centers expressed in image coordinates. Let also A_q and A_t be the affine maps that convert, respectively, from query and target coordinates to patch coordinates (known from keypoint information). The missing piece to express the local affine transformation between query and target images centered at c_q, is the affine map between P_q and P_t.

When fully trained, the presented networks are expected to predict the movements of patch corners. Let $(x_i^q \leftrightarrow x_i^t)_{i=1,\dots,k}$ be a set of correspondences produced by one of the networks $\mathcal{N}^k$, where x_i^q and x_i^t denote query and target patch-coordinates, respectively, and k-point determines the point parameterization. We call A a solution of the linear least squares problem

$$\min_A \sum_{i=1}^{k} \left\| A x_i^q - x_i^t \right\|_{L_2}^2, \tag{8}$$

the affine map estimated from the correspondences predicted by the network $\mathcal{N}^k$. Finally, the local affine map transforming the query into the target (in image coordinates) around c_q is

$$A_{q \to t} = A_t A A_q^{-1}. \tag{9}$$

We call LATE method, the method returning $A_{q \to t}$ from the LATE network.

4 Robust Homography Estimation

The standard RANSAC (RAndom SAmple Consensus) algorithm [6] 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 [20, 21, 26, 27], but the core idea behind them remains.

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

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

2. (Consensus) Count the number of matches from M_T that are within a distance threshold of κ (i.e. counting inliers).

Notice that steps 1-2 only take into account point coordinates. From now on, we call this method *RANSAC*. With eight degrees of freedom for a homography matrix and each match defining two equations, this implies $s = 4$. The following subsections support the claim that adding the affine information provided by the LATE network can further improve the performance of the RANSAC algorithm.

4.1 Homography fitting from local affine maps

From Section 2.1 we know how to locally approximate a homography by an affine map. Conversely, in this section we address the problem of determining a homography from a set of approximate affine maps at different locations. Let $x \leftrightarrow y$ be a match and $L = (l_{ij})_{i,j=1,2}$ the linear map in Equation 3. Then, according to Equation 4, the unknown homography η must satisfy

$$E_{6\times 9} \cdot \vec{h} = \vec{0}, \tag{10}$$

where

$$E_{6\times 9} = \begin{bmatrix} 1 & & & & & & -y_1 - l_{11}x_1 & -l_{11}x_2 & -l_{11} \\ & 1 & & & & & -l_{12}x_1 & -y_1 - l_{12}x_2 & -l_{12} \\ & & & 1 & & & -y_2 - l_{21}x_1 & -l_{21}x_2 & -l_{21} \\ & & & & 1 & & -l_{22}x_1 & -y_2 - l_{22}x_2 & -l_{22} \\ x_1 & x_2 & 1 & & & & -y_1x_1 & -y_1x_2 & -y_1 \\ & & & x_1 & x_2 & 1 & -y_2x_1 & -y_2x_2 & -y_2 \end{bmatrix}, \tag{11}$$

and $\vec{h}^t = \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{21} & h_{22} & h_{23} & h_{31} & h_{32} & h_{33} \end{bmatrix}$ is a vectorized version of the matrix H associated to η. The first four rows of $E_{6\times 9}$ are determined by Equation 4 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 can over-determine the homography matrix. Indeed, putting everything together provides with 12 equations,

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

where E_i denotes the matrix E appearing in Equation 10 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 [7] for more details.

We call *RANSAC*$_{2pts}$, a RANSAC version in which we modify the classic homography fitting of step 1 by the homography fitting of this section together with the LATE estimator. Remark that RANSAC$_{2\text{pts}}$ only needs two samples at each iteration ($s = 2$).

4.2 Affine consensus for RANSAC homography

In many practical applications the probability of a false match having coordinates agreeing with the testing homography just by chance is not necessarily small. The classic consensus in RANSAC could be dramatically affected if this situation occurs. Using the geometric information around each match allows us to further reduce the aforementioned probability, thus improving the performance of the RANSAC algorithm.

In this section we use the affine information to redefine the consensus set of a model. Inliers are now defined as those matches whose added affine information go along with the testing homography. Let A_E and A_H be, respectively, the estimated affine map by the LATE method and the testing affine map computed from the testing homography (using Equation 4). Let also $[\lambda_E, \psi_E, t_E, \phi_E]$ and $[\lambda_H, \psi_H, t_H, \phi_H]$ be, respectively, the corresponding affine decomposition (as in Equation 1) of A_E and A_H. We define the α-vector between A_E and A_H as:

$$\alpha\left(A_E, A_H\right) = \left[\max\left(\frac{\lambda_E}{\lambda_H}, \frac{\lambda_H}{\lambda_E}\right), \angle\left(\psi_E, \psi_H\right), \max\left(\frac{t_E}{t_H}\frac{t_H}{t_E}\right), \angle\left(\phi_E, \phi_H\right)\right], \tag{13}$$

where $\angle(\cdot,\cdot)$ denotes the angular difference.

Finally, we propose to add four more thresholds (based on the α-vector) to the classic one on the Euclidean distance. A perfect match involves an α-vector equal to $[1, 0, 1, 0]$. These four thresholds correspond to the four dimensions of the α-vector. If we assume independence on each dimension, the resulting probability of a match passing all thresholds is the multiplication of probabilities. With this in mind, we claim that rough thresholds are sufficient to obtain good performances. Thus, we propose to further refine inliers by accepting only those matches also satisfying

$$\alpha\left(A_E, A_H\right) < \left[2, \frac{\pi}{4}, 2, \frac{\pi}{8}\right], \tag{14}$$

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

We call $RANSAC_{affine}$ the version of $\text{RANSAC}_{2\text{pts}}$ that includes the affine consensus presented in this section.

5 Experiments

The network [5] is not precise enough in capturing local point movements in the presence of strong blur or tilt; Figure 4 visually shows geometric errors incurred by the network [5] (4 points) and ours (LATE). Four random patch pairs from the validation dataset (synthetic data) start showing the Achilles heel of network [5]: zoom and translation. This visualization already justifies the use of the inverse information in the LATE method.

Up until now, the LATE network has only seen optically simulated input patches. The passage from affine cameras to real cameras is a big gap to fill by both networks. We expect them to generalize the affine world to all sorts of geometry as long as the Taylor approximation holds. Let us now test the performance of the LATE network on real data.

As a first evaluation of the precision in a realistic environment we used the viewpoint dataset from SIFT-AID [32], consisting of five pairs of images with their groundtruth homographies and 3352 true matches. Notice that Equations 4-5 allow us to compute groundtruth local affine maps around each match. Figure 5 shows the accuracy of the LATE and [5] networks represented by error density functions with respect to the affine decomposition appearing in Equation 1. Ideally, we expect a Dirac delta function for a perfect method. Please note the resemblance in the case of the LATE network. This experiment illustrates the failure of the network [5] in predicting zoom and translation (as shown in Figure 4). This confirms the choice in LATE of also tracking points movements incurred by the inverse affine map.

In the previous paragraphs we have shown the capacity of the LATE method to identify local affine maps. We now highlight the benefit of local geometry in estimating homographies. The following experiment was conducted on four well known datasets for homography estimation. All datasets include groundtruth homographies that were used to verify accuracy. First, local features were detected and matched by RootSIFT [1] with matching ratio set to 0.8 and SIFT-AID [32] with matching threshold set to 4000. Then, each homography estimator method (RANSAC, $\text{RANSAC}_{2\text{pts}}$ and $\text{RANSAC}_{\text{affine}}$) was 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. 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. Results on this experiment can be found in Table 1. Both $\text{RANSAC}_{2\text{pts}}$ and $\text{RANSAC}_{\text{affine}}$ methods outperform RANSAC in the number of successes and identified image pairs for both RootSIFT [1] and SIFT-AID [32] in all datasets. This proves that the affine information, which is the only difference with respect to the baseline RANSAC, systematically improves the homography estimation. Even if Table 1 is not about comparing matching methods,

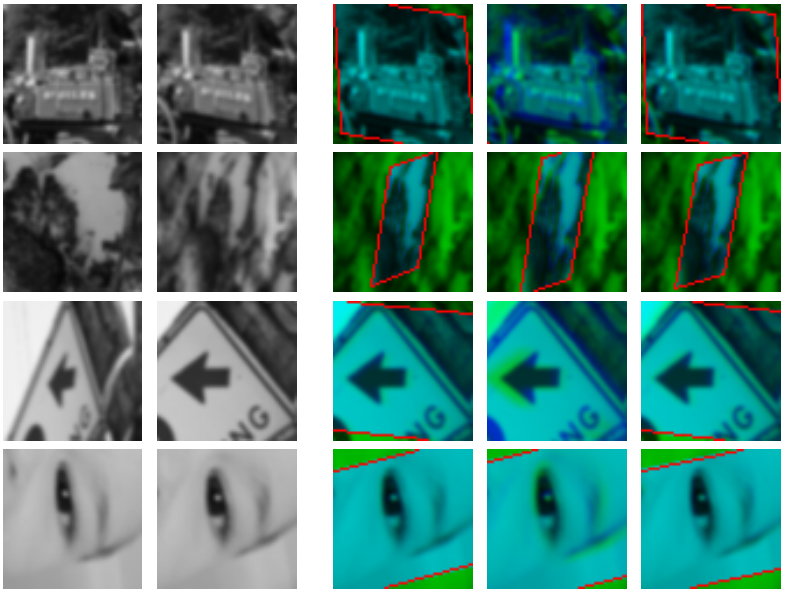

Figure 4: Four pairs of patches selected at random from the validation dataset and used as query and target input patches (columns 1-2). The three last columns show the drift error depicted by intense blue or intense green colors. Light blue means no error. Blue and green channels correspond to the target patch and a warped version of the corresponding query patch (the red line delimits its borders); The red channel is filled with zeros. 3rd column: groundtruth; 4th column: network in [5] (4 points); 5th column: LATE network. Input patches are shown without contrast difference for clear visualization.

	EF [37]				EVD [18]				OxAff [17]				SymB [8]			
Method	S	33	inl.	AvE	S	15	inl.	AvE	S	40	inl.	AvE	S	46	inl.	AvE
RootSIFT+RANSAC	2403	26	51	3.2	0	0	0	-	3806	39	580	1.2	2693	31	102	2.8
RootSIFT+RANSAC$_{2pts}$	2633	28	46	3.7	0	0	0	-	3893	39	566	1.2	3219	34	84	3.3
RootSIFT+RANSAC$_{affine}$	2805	30	28	3.4	0	0	0	-	3899	39	404	1.1	3297	36	54	3.4
SIFT-AID+RANSAC	879	23	78	6.6	82	1	40	7.8	3600	39	1477	4.8	1014	19	450	6.8
SIFT-AID+RANSAC$_{2pts}$	1829	27	84	6.1	99	1	72	6.3	3917	40	1459	4.5	1867	30	327	6.5
SIFT-AID+RANSAC$_{affine}$	1996	30	39	5.8	166	5	37	8.2	3939	40	852	4.0	2341	38	138	6.6

Table 1: Homography estimation performances for RANSAC, RANSAC$_{2pts}$ and RANSAC$_{affine}$. Each RANSAC ran for 1000 iterations. Each pair of images was fed 100 times to all RANSACs to obtain results. Legend: S - the number of successes; the number of correctly matched image pairs; inl. - the average number of correct inliers; AvE - the average pixel error. The numbers of image pairs in a dataset are boxed.

we observe that RANSAC$_{affine}$, and indirectly the LATE method, raises the performance of SIFT-AID to achieve comparable (and in two datasets better) results than AdHesAffNet [19], reported to have 33, 4, 40 and 37 correctly matched image pairs in the datasets EF, EVD, OxAff and SymB, respectively.

6 Conclusions

We proposed a CNN based method to locally estimate affine maps between images. Our experiments show that the LATE method provides accurate first-order approximations in various geometric scenarios. This information proved to be valuable in the case of homography estimation, for which we presented two RANSAC versions that systematically improved results in four well known datasets [37, 18, 17, 8]. The proposed method is generic and its applications to stereo matching, as well as to guided matching without any global geometry assumption, will be explored in future work.

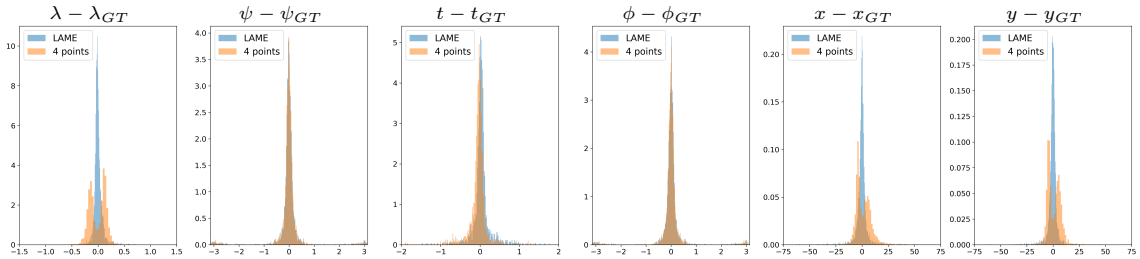

Figure 5: Affine error prediction in terms of the affine decomposition (Equation 1), for the network [5] (4 points) in orange and the proposed LATE method in blue. The used dataset [32] contains 3352 patch pairs with corresponding groundtruth. The sub-index GT means groundtruth, conversely, no sub-index stands for estimated parameters.

References

[1] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, pages 2911–2918, 2012. ISBN 9781467312264. doi: 10.1109/CVPR.2012.6248018.

[2] A. Baumberg. Reliable feature matching across widely separated views. *CVPR*, 1:774–781, 2000.

[3] J. Blom. Topological and Geometrical Aspects of Image Structure. *University of Utrecht*, 1992.

[4] F. Cao, J.-L. Lisani, J.-M. Morel, P. Musé, and F. Sur. *A Theory of Shape Identification*. Springer Verlag, 2008.

[5] D. DeTone, T. Malisiewicz, and A. Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016.

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

[7] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[8] D. C. Hauagge and N. Snavely. Image matching using local symmetry features. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 206–213. IEEE, 2012.

[9] T. Iijima. Basic equation of figure and and observational transformation. *Systems, Computers, Controls*, 2(4):70–77, 1971.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[11] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Royal Institute of Technology, Stockholm, Sweden, 1993.

[12] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D brightness structure. *ECCV*, pages 389–400, 1994.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[14] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *IVC*, 22(10):761–767, 2004.

[15] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. *ECCV*, 1:128–142, 2002.

[16] K. Mikolajczyk and C. Schmid. Scale and Affine Invariant Interest Point Detectors. *IJCV*, 60 (1):63–86, 2004.

[17] 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*, 65(1-2):43–72, 2005.

[18] D. Mishkin, J. Matas, and M. Perdoch. MODS: Fast and robust method for two-view matching. *CVIU*, 141:81–93, 2015. URL http://dblp.uni-trier.de/db/journals/cviu/cviu141.html#MishkinMP15;http://dx.doi.org/10.1016/j.cviu.2015.08.005.

[19] 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)*, pages 284–300, 2018.

[20] L. Moisan, P. Moulon, and P. Monasse. Automatic Homographic Registration of a Pair of Images, with A Contrario Elimination of Outliers. *IPOL*, 2:56–73, 2012. ISSN 2105-1232. doi: 10.5201/ipol.2012.mmm-oh. URL http://www.ipol.im/pub/art/2012/mmm-oh/.

[21] L. Moisan, P. Moulon, and P. Monasse. Fundamental Matrix of a Stereo Pair, with A Contrario Elimination of Outliers. *IPOL*, 6:89–113, 2016.

[22] J.-M. Morel and G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

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

[24] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.-M. Morel. An A Contrario Decision Method for Shape Element Recognition. *IJCV*, 69(3):295–315, 2006.

[25] Y. Pang, W. Li, Y. Yuan, and J. Pan. Fully affine invariant SURF for image matching. *Neurocomputing*, 85:6–10, 2012. ISSN 09252312. doi: 10.1016/j.neucom.2011.12.006.

[26] 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(8):2022–2038, 2013. https://doi.org/10.1109/TPAMI.2012.257.

[27] M. Rais, G. Facciolo, E. Meinhardt-Llopis, M. J.-M., B. A., and C. B. Accurate motion estimation through random sample aggregated consensus. *CoRR*, abs/1701.05268, 2017. URL http://arxiv.org/abs/1701.05268.

[28] I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional neural network architecture for geometric matching. *TPAMI*, 2018.

[29] M. Rodriguez and R. Grompone von Gioi. Affine invariant image comparison under repetitive structures. In *ICIP*, pages 1203–1207, Oct 2018. doi: 10.1109/ICIP.2018.8451060. URL https://rdguez-mariano.github.io/pages/acdesc.

[30] M. Rodriguez, J. Delon, and M. J.-M. Covering the space of tilts. application to affine invariant image comparison. *SIIMS*, 11(2):1230–1267, 2018. URL https://rdguez-mariano.github.io/pages/imas.

[31] M. Rodriguez, J. Delon, and J.-M. Morel. Fast affine invariant image matching. *IPOL*, 8:251–281, 2018. doi: 10.5201/ipol.2018.225. URL https://rdguez-mariano.github.io/pages/hyperdescriptors.

[32] 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. Preprint, Feb. 2019. URL https://hal.archives-ouvertes.fr/hal-02016010.

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. *BMVC*, pages 412–425, 2000.

[35] T. Tuytelaars and L. Van Gool. Matching Widely Separated Views Based on Affine Invariant Regions. *IJCV*, 59(1):61–85, 2004.

[36] T. Tuytelaars, L. Van Gool, and Others. Content-based image retrieval based on local affinely invariant regions. *Int. Conf. on Visual Information Systems*, pages 493–500, 1999.

[37] C. L. Zitnick and K. Ramnath. Edge foci interest points. In *2011 International Conference on Computer Vision*, pages 359–366. IEEE, 2011.