Optimal Multi-view Correction of Local Affine Frames

Ivan Eichhardt
ivan.eichhardt@sztaki.mta.hu
Machine Perception Research Laboratory, MTA SZTAKI, Budapest, Hungary

Daniel Barath
barath.daniel@sztaki.mta.hu
Centre for Machine Perception, Department of Cybernetics Czech Technical University, Prague, Czech Republic

Abstract

A method is proposed for correcting the parameters of a sequence of detected local affine frames through multiple views. The technique requires the epipolar geometry to be pre-estimated between each image pair. It exploits the constraints which the camera movement implies, in order to apply a closed-form correction to the parameters of the input affinities. Also, it is shown that the rotations and scales obtained by partially affine-covariant detectors, *e.g.* AKAZE or SIFT, can be upgraded to be full affine frames by the proposed algorithm. It is validated both in synthetic experiments and on publicly available real-world datasets that the method almost always improves the output of the evaluated affine-covariant feature detectors. As a by-product, these detectors are compared and the ones obtaining the most accurate affine frames are reported. To demonstrate the applicability in real-world scenarios, we show that the proposed technique improves the accuracy of pose estimation for a camera rig, surface normal and homography estimation.

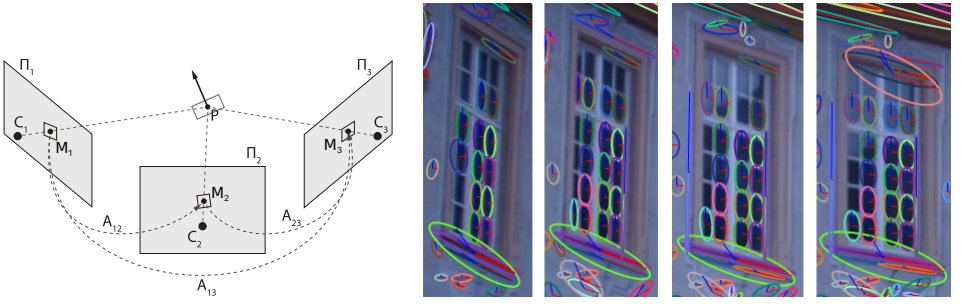
The source code is available at github.com/eivan/multiview-LAFs-correction.

1 Introduction

A method is proposed for estimating local affine frames [26] (LAFs) accurately in a rigid[1] scene observed by multiple cameras. In particular, we are interested in finding the affine mappings which are the closest in the least-squares sense to the detected ones and, also, for which the constraints implied by the camera movement hold. The method takes a sequence of affine features detected by an affine-covariant feature detector (*e.g.*, Affine-SIFT [23]) and returns the affinities corrected by the proposed closed-form procedure. Also, the method is applicable when a not fully affine-covariant detector is used, *e.g.* AKAZE [1] or SIFT [17] which estimate solely parts of the corresponding LAFs, *e.g.* scales and orientations. The proposed method returns the underlying affine frames consistent with the camera movement.

Nowadays, a number of algorithms have been proposed for solving various computer vision problems by exploiting affine correspondences. For instance, Perdoch *et al.* [27] proposed techniques for approximating the epipolar geometry between two images by generating point correspondences from the affine features. Bentolila and Francos [10] showed a method to estimate the fundamental matrix using three correspondences. Raposo *et al.* [31]

[1] The generalisation to multiple rigid motions each satisfying a different constraint is straightforward.



(a) (b)

Figure 1: **(a)** Three cameras ($\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$) observing point $\mathbf{P}$. The shape of the region induced by the plane on which $\mathbf{P}$ lies in the ith image is described by local affine frame $\mathbf{M}_i$ (LAF). The LAFs around the projected points between the ith and jth views are related by local affine transformation $\mathbf{A}_{ij}$. **(b)** Example multi-view region correspondences represented by oriented ellipses (LAFs) across multiple views. Corresponding ellipses are denoted by colour.

proposed a solution for essential matrix estimation using two feature pairs. Eichhardt and Chetverikov [11] proposed a generalisation of the approach considering arbitrary central projection. Baráth *et al*. [7] proved that even the semi-calibrated case (*i.e*., when the objective is to find the essential matrix and a common focal length) is solvable from two correspondences. Homographies can also be estimated from two features [15] without any a priori knowledge about the camera movement. In the case of known epipolar geometry, a single affine correspondence is sufficient for estimating a homography [2]. Affine correspondences were successfully used in multi-homography estimation [3]. Also, affine frames contain information about the surface normals [22]. Therefore, if the cameras are calibrated, the normal can be estimated from a single correspondence [15]. Multi-view surface normal estimation [8, 12] is also possible. Pritts *et al*. [28, 29] showed that the radial distortion parameters can be retrieved, as well, using affine frames.

Affine correspondences encode higher-order information about the underlying scene geometry. This is what makes the listed algorithms able to estimate geometric models, *e.g*., homographies and fundamental matrices, using significantly fewer correspondences than point-based methods. Being more complex than 2D points, the accurate estimation of affine frames is a more complicated task. The estimation is, in practice, done by applying an affine- or partially affine-covariant feature detector which simultaneously recovers points and the corresponding affine frames. Some methods investigate the shapes of corresponding image regions (*e.g*., MSER [18], TMBR [35]). Other techniques generate synthetic views by transforming the input images by affine transformations (*e.g*., ASIFT [23], MODS [20]), whilst some of them optimise each detected feature by minimising a photo-consistency-based cost function [19]. However, affine correspondences are significantly more noisy than points even when applying state-of-the-art feature detectors.

Barath *et al*. [5] proposed two constraints describing the relationship of stereo epipolar geometry and affine correspondences. The constraints are built on the fact that a geometrically valid affine frame must transform the normals of the corresponding epipolar lines into each other. Also, the scaling factor along the normal direction is determined by the epipolar geometry and, thus, can be calculated from the fundamental matrix. Exploiting these constraints, the EG-L_2-Optimal algorithm is proposed in [5] to make an input affine corre-

spondence consistent with the fundamental matrix by an efficient closed-form approach.

In this paper, we extend the EG-L_2-Optimal technique by generalising the constraints to multiple views. The proposed method is applicable when a sequence of corresponding affine frames is given through multiple images (see Fig. 1). It is efficient due to being solved by a closed-form approach. It is validated both on synthetic experiments and on a number of real-world datasets that the method always improves the output of state-of-the-art affine- and partially affine-covariant feature detectors. As a by-product, these detectors are compared and the best ones, in terms of finding the most geometrically accurate affine frames, are reported. As possible applications, it is shown that the proposed method improves homography and surface normal estimation. Also, using the corrected affine frames makes the relative motion estimation of a camera rig more accurate.

2 Epipolar Constraints on Affine Features

In this section, first, the required theoretical background is discussed. Then we show the constraints which a pair of affine frames imply on the two-view epipolar geometry.

Notation and Preliminaries. A *local affine frame* (LAF) is a pair $(\mathbf{x}, \mathbf{M})$ of a point $\mathbf{x} = [u,\ v]^\mathrm{T}$ and a 2×2 linear transformation $\mathbf{M} \in \mathbb{R}^{2\times 2}$. Matrix $\mathbf{M}$ is defined by the partial derivatives, *w.r.t.* the image directions, of the projection function [2]. An *affine correspondence* $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{A})$ is a triplet, where $\mathbf{x}_1 = [u_1\ v_1]^\mathrm{T}$ and $\mathbf{x}_2 = [u_2\ v_2]^\mathrm{T}$ is a corresponding pair of points in two images and $\mathbf{A}$ is a 2×2 linear transformation which is called *local affine transformation* and defined as $\mathbf{A} = \mathbf{M}_2\mathbf{M}_1^{-1}$, where $\mathbf{M}_i$ is the matrix from the corresponding LAF in the ith image, $i \in \{1, 2\}$.

The fundamental ($\mathbf{F}$) and essential ($\mathbf{E}$) matrices ensure the epipolar constraint as $\tilde{\mathbf{x}}_2^\mathrm{T}\mathbf{F}\tilde{\mathbf{x}}_1 = \tilde{\mathbf{x}}_2^\mathrm{T}\mathbf{K}_2^{-\mathrm{T}}\mathbf{E}\mathbf{K}_1^{-1}\tilde{\mathbf{x}}_1 = 0$, where $\mathbf{K}_i$ is the intrinsic calibration matrix of the ith camera and $\tilde{\mathbf{x}}_i$ is the homogeneous form of point $\mathbf{x}_i$.

Constraints on affine correspondences. Suppose that we are given an affine correspondence $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{A})$ constructed from two LAFs $(\mathbf{x}_1, \mathbf{M}_1)$ and $(\mathbf{x}_2, \mathbf{M}_2)$ such that

$$\mathbf{A} = \mathbf{M}_2\mathbf{M}_1^{-1}. \tag{1}$$

In case of pinhole cameras, the following constraint [11] holds.

$$\mathbf{A}^\mathrm{T}\underbrace{\mathbf{I}_{2\times 3}\mathbf{F}\tilde{\mathbf{x}}_1}_{\mathbf{a}} + \underbrace{\mathbf{I}_{2\times 3}\mathbf{F}^\mathrm{T}\tilde{\mathbf{x}}_2}_{\mathbf{b}} = \mathbf{0}, \tag{2}$$

where $\mathbf{I}_{2\times 3}$ is a 2×3 identity matrix and $\mathbf{F}$ is the fundamental matrix. Note that, in case of arbitrary central projection, $\mathbf{a} = \nabla q_2^\mathrm{T}\mathbf{E}q_1$ and $\mathbf{b} = \nabla q_1^\mathrm{T}\mathbf{E}^\mathrm{T}q_2$, where q_i is the bearing vector corresponding to $\mathbf{x}_i$ and ∇q_i is its gradient *w.r.t.* $\mathbf{x}_i$. This relationship is described in [11] in depth. A compact form of the expression is $\mathbf{A}^\mathrm{T}\mathbf{a} + \mathbf{b} = \mathbf{0}$.

Constraints on local affine frames. In order to define how a pair of LAFs is constrained by the epipolar geometry, we plug formula (1) into (2). The obtained equation is as follows: $\mathbf{A}^\mathrm{T}\mathbf{a} + \mathbf{b} = \left(\mathbf{M}_2\mathbf{M}_1^{-1}\right)^\mathrm{T}\mathbf{a} + \mathbf{b} = \mathbf{M}_1^{-\mathrm{T}}\mathbf{M}_2^\mathrm{T}\mathbf{a} + \mathbf{b} = \mathbf{0}$. After left-multiplying the expression by $\mathbf{M}^{-\mathrm{T}}$, the following epipolar constraint on a pair of LAFs is given:

$$\mathbf{M}_2^\mathrm{T}\mathbf{a} + \mathbf{M}_1^\mathrm{T}\mathbf{b} = \mathbf{0}. \tag{3}$$

3 Multi-view EG-L_2-Optimal Correction

Let $\mathcal{V}$ be the set of views in a multi-view correspondence, *i.e.* $\mathbf{x}_k$ ($\forall k \in \mathcal{V}$) are projections of the same point in space where $\left(\mathbf{x}_k, \hat{\mathbf{M}}_k\right)$ is the respective LAF. The set of pairwise correspondences is $\mathcal{C} \subseteq \mathcal{V} \times \mathcal{V}$. The objective is to find all $\mathbf{M}_k$, such that

$$\min_{\mathbf{M}_k} \sum_{k \in \mathcal{V}} \left\| \mathbf{M}_k^{\mathrm{T}} - \hat{\mathbf{M}}_k^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \qquad \text{s.t.} \qquad \forall (i,j) \in \mathcal{C} : \mathbf{M}_j^{\mathrm{T}} \mathbf{a}_{ij} + \mathbf{M}_i^{\mathrm{T}} \mathbf{b}_{ij} = \mathbf{0}, \tag{4}$$

where $\mathbf{a}_{ij}$ and $\mathbf{b}_{ij}$ are as defined above, *e.g.* $\mathbf{a}_{ij} = \nabla q_j^{\mathrm{T}} \mathbf{E} q_i$ and $\mathbf{b}_{ij} = \nabla q_i^{\mathrm{T}} \mathbf{E}^{\mathrm{T}} q_j$ for the pair (i,j) of views. An equivalent form of (4) using Lagrange multipliers $\lambda_{ij} \in \mathbb{R}^2$ is as follows:

$$\min_{\mathbf{M}_k, \lambda_{ij}} \sum_{k \in \mathcal{V}} \frac{1}{2} \left\| \mathbf{M}_k^{\mathrm{T}} - \hat{\mathbf{M}}_k^{\mathrm{T}} \right\|_{\mathrm{F}}^2 - \sum_{(i,j) \in \mathcal{C}} \lambda_{ij}^{\mathrm{T}} \left(\mathbf{M}_j^{\mathrm{T}} \mathbf{a}_{ij} + \mathbf{M}_i^{\mathrm{T}} \mathbf{b}_{ij} \right). \tag{5}$$

Optimality conditions. To find the globally optimal solution, the 1st-order optimality conditions have to be investigated. For each $k \in \mathcal{V}$, the gradient $\nabla_{\mathbf{M}_k^{\mathrm{T}}}$ of the expression in (5) is

$$\mathbf{M}_k^{\mathrm{T}} - \sum_{(i,k) \in \mathcal{C}} \lambda_{ik} \mathbf{a}_{ik}^{\mathrm{T}} - \sum_{(k,j) \in \mathcal{C}} \lambda_{kj} \mathbf{b}_{kj}^{\mathrm{T}} = \hat{\mathbf{M}}_k^{\mathrm{T}}, \tag{6}$$

The gradient $\nabla_{\lambda_{mn}}$ of (5) corresponding to the Lagrange multiplier λ_{mn} gives an expression resembling the epipolar constraints in (3) as follows:

$$\mathbf{M}_n^{\mathrm{T}} \mathbf{a}_{mn} + \mathbf{M}_m^{\mathrm{T}} \mathbf{b}_{mn} = \mathbf{0}. \tag{7}$$

Given all the 1st-order optimality conditions, an equivalent form can be constructed as a single linear system as follows:

$$\begin{bmatrix} \mathbf{I}_{2|\mathcal{V}| \times 2|\mathcal{V}|} & \mathbf{B} \\ \mathbf{B}^{\mathrm{T}} & \mathbf{0}_{|\mathcal{C}| \times |\mathcal{C}|} \end{bmatrix} \begin{bmatrix} \Omega \\ \Lambda \end{bmatrix} = \begin{bmatrix} \hat{\Omega} \\ \mathbf{0}_{|\mathcal{C}| \times 2} \end{bmatrix}, \tag{8}$$

where $\Omega = \begin{bmatrix} \mathbf{M}_1^{\mathrm{T}} & \dots & \mathbf{M}_{|\mathcal{V}|}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$, $\hat{\Omega} = \begin{bmatrix} \hat{\mathbf{M}}_1^{\mathrm{T}} & \dots & \hat{\mathbf{M}}_{|\mathcal{V}|}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ and $\Lambda = \begin{bmatrix} \dots & \lambda_{ij} & \dots \end{bmatrix}^{\mathrm{T}}$.

Note that $\begin{bmatrix} \mathbf{I}_{2|\mathcal{V}| \times 2|\mathcal{V}|} & \mathbf{B} \end{bmatrix}$ encodes the optimality conditions in (6), and $\mathbf{B}^{\mathrm{T}} \in \mathbb{R}^{|\mathcal{C}| \times 2|\mathcal{V}|}$ holds the optimality conditions of (7). Each line of $\mathbf{B}^{\mathrm{T}}$ holds $\mathbf{a}_{ij}$ and $\mathbf{b}_{ij}$ needed for an epipolar constraint: $\mathbf{B}^{\mathrm{T}}\Omega$ is zero, if Ω stores LAFs consistent with the epipolar geometry.

Efficient solution to the linear system. Due to the block matrix structure of (8), formula $\Omega = \hat{\Omega} - \mathbf{B}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}\hat{\Omega}$ can be used to compute the optimal solution, where $\mathbf{B}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}$ is a projection matrix into the column space of $\mathbf{B}$. To avoid numerical instability, the direct computation of the inverse is not preferred. To our experiments, the most stable solution is given by the column-pivoting Householder QR decomposition of $\mathbf{B}$, in case $\mathbf{B}$ is noise-free, *i.e.*, when the point coordinates and the epipolar geometries are consistent with the reconstruction. In a Structure-from-Motion (SfM) system, it can be guaranteed that $\mathbf{B}$ contains no noise by deriving the essential matrices and bearing vectors with their gradients from the camera poses and reconstructed 3D points.

In other cases, when only pairwise epipolar geometries are known, we propose to apply the following approach using singular value decomposition (SVD). It is evident that due to $\mathbf{B}^{\mathrm{T}}\Omega = \mathbf{0}$, the left-nullspace of $\mathbf{B}$ is expected to be non-empty. If the null-space is at least two-dimensional, it can contain Ω, however, the structure of $\mathbf{B}$ suggests it is three-dimensional. Thus, we propose to use formula $\Omega = \hat{\Omega} - \mathbf{U}_{(:,1\ldots2|\mathcal{V}|-3)}\mathbf{U}^{\mathrm{T}}_{(:,1\ldots2|\mathcal{V}|-3)}\hat{\Omega}$, where $\mathbf{USV}^{\mathrm{T}} = \mathbf{B}$ is the SVD of $\mathbf{B}$ and $\mathbf{U}_{(:,1\ldots2|\mathcal{V}|-3)}$ is the matrix consisting of the left $2|\mathcal{V}|-3$ columns of $\mathbf{U}$.

Refinement of partially affine-covariant regions. When a scale- and orientation-covariant detector is applied, *e.g.* AKAZE [1] or SIFT [17], only a part of the affine frames are obtained, *e.g.*, the orientation and scale. In this case, the affine frames can be approximated as $\hat{\mathbf{M}} \sim \sigma\mathbf{R}$, where $\sigma \in \mathbb{R}^{+}$ is the scale of the local frame, while $\mathbf{R} \in \mathbb{R}^{2\times2}$ encodes the dominant orientation of the underlying region. Thus, *with no special treatment* of the partially affine-covariant regions, the proposed method can be applied.

4 Experimental results

In this section, the proposed method for correcting LAFs is tested both in synthetic experiments and on publicly available real-world datasets. First, we show how the proposed method improves the accuracy of detected LAFs. Finally, it is demonstrated on a number of real-world problems, *i.e.* homography, surface normal and motion estimation of a camera rig, that using the proposed method leads to superior results.

Synthetic experiments. To test the proposed method in a fully controlled environment, N cameras were generated by their projection matrices looking towards the origin, each located in a random surface point on a sphere of radius 5. Then, a random 3D oriented point, at most one unit away from the origin and with random normal, was projected into the cameras. The ground truth LAF in each image was calculated from the projection matrix and the surface normal as in [6]. Zero-mean Gaussian noise with σ standard deviation was added to both the point locations and affine parameters. Each reported result is averaged over $1,000$ runs. The processing time of 5 views is ≈ 0.03ms.

In Fig. 2(a), the errors of the noisy LAFs, *i.e.* the input without the correction, are plotted as the function of the noise level σ (horizontal axis; in pixels) and view number (vertical). In Fig. 2(b), the errors of the corrected frames are shown when using the ground truth fundamental matrices for the correction. In Fig. 2(c), the errors are shown when the $\mathbf{F}$s are estimated from the noisy point coordinates applying the normalised 8-point algorithm [14]. It can be seen that the proposed method is consistent, *i.e.*, the more views are given, the more accurate the results are. Also, Fig. 2(c) shows that the method *significantly improves the input LAFs* even if the estimated epipolar geometries are noisy. More detailed evaluation is provided in the supplementary material.

Comparing feature extractors. In this section, commonly used feature extractors are applied to images of the Strecha dataset [32] and their outputs are corrected by the proposed method. The dataset[2] consists of six image sequences of size 3072×2048 of buildings. Both the intrinsic and extrinsic parameters are given for all images. To obtain ground truth LAFs in each image sequence, we first applied an SfM pipeline [24] with the known camera

[2] Available at http://cvlabwww.epfl.ch/data/multiview/denseMVS.html

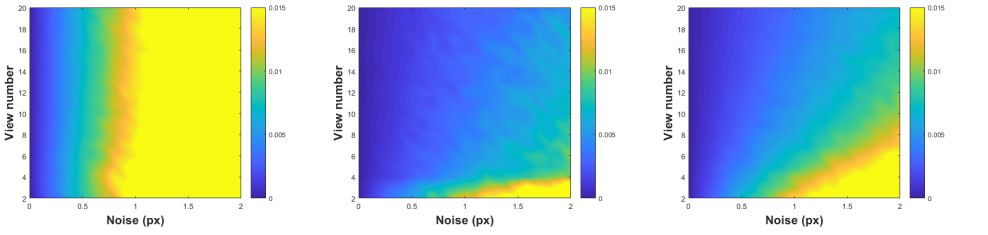


Figure 2: *Accuracy of the proposed method.* The error of the noisy (a) and corrected (b–c) LAFs are plotted as the function of the noise level σ (horizontal axis; in pixels) and view number (vertical axis). Plot (a) shows the error of the input. For (b), the ground truth fundamental matrix was used. For (c), $\mathbf{F}$ was estimated from the noisy points. The error is $(1/K)\sum_{i=1}^{K}\left\|\mathbf{I}-\mathbf{M}_{i,\text{gt}}^{-1}\mathbf{M}_{i,\text{est}}\right\|_{\text{F}}$, where K is the number of views, $\mathbf{I}$ is a 3×3 identity matrix, $\mathbf{M}_{i,\text{gt}}$ and $\mathbf{M}_{i,\text{est}}$ are, respectively, the ground truth and estimated LAFs in the ith view.

parameters obtaining a number of points along the images. Then, the points were manually assigned to dominant planes. Since each plane defines a homography between every view pair, the ground truth affine correspondences between the view pairs were calculated from the homography parameters as described in [2]. The evaluated extractors can be divided into four groups: (i) scale and rotation-covariant ones, like SIFT [17], AKAZE [1], Hessian [33], Difference of Gaussians (DoG) [33], and Harris-Laplace (Harris) [33]. (ii) Affine-covariant extractors using the Baumberg-iteration [9] such as Hessian-Aff, DoG-Aff and Harris-Aff, and (iii) methods using simulated views, such as ASIFT [23], AAKAZE, etc. (iv) Also, we tested the recently published Hes-Aff-Net [21] which obtains affine regions by running CNN-based shape regression on Hessian keypoints. In the experiments, the VlFeat library [33] provides the Hessian, DoG and Harris extractors, and their covariant counterparts: Hessian-Aff, DoG-Aff and Harris-Aff using its built-in version of the shape adaptation procedure (*i.e.*, the Baumberg iteration). We used the SIFT and AKAZE implementations included in OpenMVG [25]. For AAKAZE and ASIFT, the view-simulation of [23] is used, feeding warped versions of the input images to the detectors.

For the experiments, we used a modified version of OpenMVG [25] which, together with the point coordinates, stores the LAFs throughout the reconstruction. For each detector, we performed feature extraction, then established multi-view correspondences. The Global SfM pipeline [24] of OpenMVG estimated the camera motion and created a 3D point cloud of the scene. A robust triangulation procedure then established multi-view tracks of LAFs, with geometrically consistent centroids. Finally, the corrected LAFs were obtained by the proposed method using the estimated poses.

The results are in Table 1. After the header, the odd rows report the accuracy of the extracted LAFs. The even rows show the quality of the corrected ones. Pairs of rows show the results of a particular detector. The sequences of the Strecha dataset are from the 3rd to 8th columns. The last two columns show the mean and median errors on the entire dataset. It can be seen that the proposed method almost *always improved* the input LAFs. The most accurate detector is AAKAZE with the proposed correction. Also, it can be seen that the proposed technique significantly improves partially affine-covariant detectors, *e.g.* SIFT, as well. We were surprised that SIFT, without the correction, obtains more accurate LAFs than

ASIFT on average. The reason is however simple. ASIFT extracts, on average, ten times more correspondences which greatly influences its mean error. However, the median error of ASIFT is 0.19 while that of SIFT is 0.20. Other detectors are shown in Fig. 3(a).

The processing time – calculated from all real-world experiments – of the proposed technique is reported in Fig. 3(b). It can be seen that, on average, it runs for less than 0.1 ms when having 4-5 views. The runtime of the algorithm increases in a quadratic trend as more views are added, as expected from the structure of the matrix in (8).

detector	LAF type	(a)	(b)	(c)	(d)	(e)	(f)	mean	median
SIFT	Extracted	0.22	0.22	0.23	0.26	0.31	0.29	0.26	0.20
	Corrected	**0.14**	**0.12**	**0.13**	**0.18**	**0.18**	**0.21**	**0.16**	**0.11**
Hessian	Extracted	0.25	0.25	0.26	0.26	0.33	0.29	0.27	0.22
	Corrected	**0.14**	**0.14**	**0.12**	**0.16**	**0.20**	**0.20**	**0.16**	**0.11**
Hessian-Aff	Extracted	0.29	0.29	0.29	0.37	0.41	0.35	0.33	0.25
	Corrected	**0.13**	**0.12**	**0.13**	**0.23**	**0.16**	**0.18**	**0.16**	**0.10**
DoG-Aff	Extracted	0.25	0.25	**0.29**	0.27	0.43	0.39	0.31	0.19
	Corrected	**0.08**	**0.08**	0.40	**0.16**	**0.27**	**0.23**	**0.20**	**0.07**
AAKAZE	Extracted	0.26	0.25	0.31	0.32	0.30	0.28	0.29	0.22
	Corrected	**0.11**	**0.10**	**0.13**	**0.18**	**0.12**	**0.13**	**0.13**	**0.08**
ASIFT	Extracted	0.24	0.24	0.25	0.28	0.31	0.30	0.27	0.19
	Corrected	**0.11**	**0.11**	**0.12**	**0.17**	**0.14**	**0.16**	**0.14**	**0.08**
Hess-Aff-Net	Extracted	0.25	0.28	0.26	0.27	0.37	0.29	0.29	0.23
	Corrected	**0.12**	**0.13**	**0.12**	**0.16**	**0.17**	**0.18**	**0.15**	**0.10**

Table 1: *Comparison of feature detectors* in terms of the accuracy of the obtained LAFs. The accuracy (same metric as in Fig. 2) of the extracted and corrected (by the proposed method) LAFs are put in the odd and even rows, respectively. The scenes (columns) of the Strecha dataset: (a) `castle-P19`, (b) `castle-P30`, (c) `entry-P10`, (d) `fountain-P11`, (e) `herz-jesus-P25` and (f) `herz-jesus-P8` were fed into the [25] SfM pipeline. The proposed method almost *always improves* the extracted LAFs.

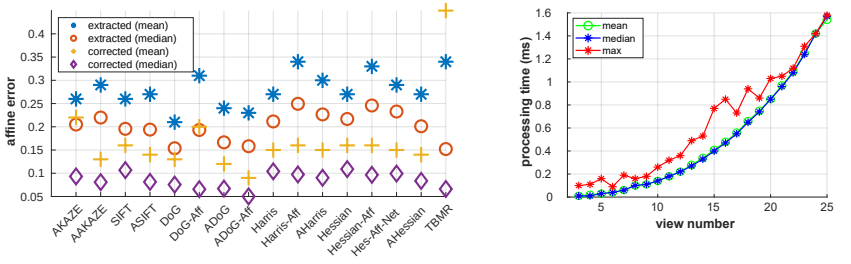
Application: homography estimation using affine correspondences (ACs). We used the Strecha dataset and, solely for validation purposes, the manually annotated homographies, similarly as in the previous section. Affine correspondences were estimated by the AAKAZE method since it leads to the most accurate LAFs (see Table 1). As homography estimator, we chose the HAF method from [2] which estimates the homography from a single affine correspondence and the fundamental matrix.

To test the proposed method, we iterated through every possible image pair in each sequence. For each pair, the following procedure was applied to every AC:

1. The AC is assigned to the closest, in terms of re-projection error, homography $\mathbf{H}^*$ from the manual annotation. If the error is bigger than 3.0 px, the AC is rejected.

2. Homography $\mathbf{H}$ is estimated from the AC and fundamental matrix by the HAF method.

3. Given the ground truth inliers $\mathcal{I}^*$ of $\mathbf{H}^*$ from the manual annotation, the proportion of them being inlier of $\mathbf{H}$ as well (*i.e.*, $|\mathcal{I}|/|\mathcal{I}^*|$, where $\mathcal{I} \subseteq \mathcal{I}^*$ and $\forall \mathbf{p} \in \mathcal{I}$ is inlier of $\mathbf{H}$) is calculated. The threshold is set to 3.0 px.



(a) (b)

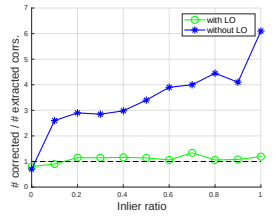
Figure 3: (a) *Comparison of feature detectors* (horizontal axis). The mean and median errors (vertical; on all scenes of the Strecha dataset) of the extracted and corrected LAFs are shown. More detailed evaluation is shown in Table 1 for the most accurate detectors. (b) *The processing time* in milliseconds (mean, median and max) of the proposed method is plotted as the function of the view number. The values are calculated from all of the real-world experiments using our C++ implementation.

4. To measure how a state-of-the-art robust estimator benefits from the proposed method, we applied the local optimisation step of USAC [30] to **H**.

In Fig. 4(a), the average improvement of the corrected LAFs is plotted as the function of the inlier ratio (horizontal axis) with and without local optimisation. In the left of the plot, lower values are better. In the right side, higher values are preferred. We explain the figure through an example. The value of the blue curve at 0.4 inlier ratio is approx. 3. This means that there are three times more ACs amongst the corrected ones than in the extracted correspondence set which led to 0.4 inlier ratio. Accordingly, there are more than 6 times more correspondences leading to ≈ 1 inlier ratio. Also, the ratio of ACs leading to 0 inliers is decreased significantly. Therefore, *there are fewer inaccurate and more accurate ACs* among the corrected correspondences. Originally, $99,331$ extracted ACs led to ≈ 0 inliers and $29,848$ of them were upgraded by the proposed method to have higher inlier ratio. This improvement is slightly less significant, although consistent, when the local optimisation is applied. *Note*: the two curves should not be compared to each other since they show how the proposed algorithm improves homography estimation if LO is or is not applied.

In conclusion, *homography estimation benefits from the corrected ACs* significantly. The **H**s estimated from the corrected ACs are more capable of finding the sought inliers. This holds even if a state-of-the-art robust procedure is used (*i.e.*, LO) after the initial LSQ fitting. **Application: surface normal estimation** using affine correspondences. We applied the multi-view least-squares optimal method from [8] to estimate surface normals from the extracted and corrected LAFs. The used sequences from the Strecha dataset are `fountain-p11`, `herzjesus-p8` and `herzjesus-p25` since those are the only ones with publicly available ground truth 3D point cloud. We estimated the ground truth surface normals from the point clouds. The error is calculated as the angular error (in degrees) between the reconstructed surface normal and the ground truth one.

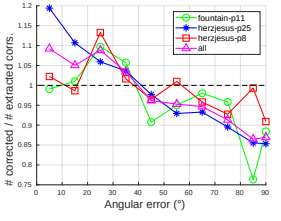
Fig. 4(d) shows the improvement (vertical axis), by using the proposed method as a pre-processing step, plotted as the function of the angular error (horizontal). The same property is shown as for homographies in Fig. 4(a). In Fig. 4(d), higher values on the left side (*i.e.*, increased number of accurate normals) and lower values on the right side (*i.e.*, decreased number of inaccurate normals) indicate the improvement caused by applying proposed algo-



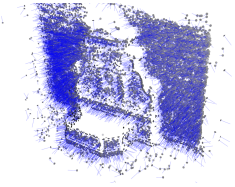
(a)

(b)

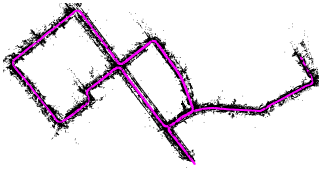
(c)



(d)



(e)



(f)

Figure 4: *Accuracy of AC-wise homography and surface normal estimation.* **(a)** Homographies. The avg. improvement (vertical axis) of the corrected LAFs compared to the extracted ones are plotted as the function of the inlier ratio (horizontal) with and without local optimisation. We explain the figure via examples. The blue curve at 0.4 inlier ratio is ≈ 3. Thus, three times more ACs are leading to 0.4 inlier ratio among the corrected ones than in the extracted set. Accordingly, there are > 6 times more ACs leading to ≈ 1 inlier ratio. **(b–c)** Left image of an example image pair from the `castle-p30` sequence. **H** is estimated from the (b) extracted and (c) corrected ACs centred on the green point. The inliers of **H** (blue points) and the rest of the points from the same plane (red) are drawn. The corrected AC led to $\approx$12 times more inliers than the extracted one. **(d)** shows the same for normal estimation as (a) for **H** estimation. *E.g.*, in the `herzjesus-p8` scene (blue curve) there are 1.2 times more (vertical axis) corrected ACs leading to 5° angular error (horizontal) than in the extracted set. **(c)** An example scene with reconstructed normals (blue lines) and points. **(f)** Example scene from the KITTI dataset [13].

rithm. For example, in the `herzjesus-p25` scene (blue curve), there are 1.2 times more (vertical axis) corrected ACs leading to 5° angular error (horizontal) than in the extracted set. Also, if all scenes are considered (purple curve), there are significantly fewer corrected LAFs (the curve is under 1.0) leading to $> 40°$ angular error than in the extracted LAF set. In conclusion, the proposed method *improves surface normal estimation* via improving its input significantly. In Fig. 4(e), an example scene with reconstructed normals (blue lines) and points are shown. *Note*: to the best of our knowledge, surface normals cannot be recovered without knowing the relative pose (*i.e.*, essential matrix) between the cameras. Consequently, since the essential matrices are already given, the proposed method can be applied straightforwardly.

Application: relative motion estimation of a camera rig using affine correspondences. In this case, the relative poses, *i.e.* the essential matrices, among the cameras in the rig have usually been pre-calculated by, *e.g.*, using chessboards. When estimating the motion of the rig, the affine correspondences found across multiple views *can be straightforwardly corrected by the proposed technique* using the a priori known essential matrices.

robust method	LAF type	iters.	t (ms)	inliers	mean ρ	med. ρ	mean τ	med. τ
MSAC	Extracted	28	5.1	59.0%	0.66	0.18	8.11	2.62
	Corrected	**26**	**4.5**	**60.4%**	**0.61**	**0.17**	**7.31**	**2.35**
LO$^+$-MSAC	Extracted	23	6.5	74.8%	0.45	**0.09**	5.00	1.30
	Corrected	**22**	**5.7**	**75.2%**	**0.38**	**0.09**	**4.18**	**1.28**

Table 2: *Relative motion estimation of a camera rig* (from the KITTI dataset [13]) using the extracted and corrected LAFs. MSAC [34] and LO$^+$-MSAC [16] were used as robust estimators and 2AC [11] as a minimal solver. The reported properties (averaged over 2020 frames) are: number of iterations (3rd column), runtime (in ms; 4th), proportion of inliers (in %; 5th), rotation (ρ; 6–7th) and translation (τ; 8–9th) errors in degrees.

We used trajectory "00" from the KITTI dataset [13]. Multi-view ACs were established in the frames each consisting of a stereo view pair. Each two consecutive stereo pairs were used together simulating a rig of four cameras, and the LAFs were corrected using this rig. The relative motion was then estimated between the consecutive four-tuples of images (*i.e.* a frame of the rig) using MSAC [34] and LO$^+$-MSAC [16] robust methods. The 2AC solver [11] was used as a minimal solver estimating the essential matrix from two affine correspondences. The error of the estimated poses was calculated using the high-quality ground truth trajectory provided in the KITTI dataset. In total, 2020 four-tuples of images, *i.e.* a frame of the rig, were used in the experiments.

Table 2 reports the accuracy of the robust estimation applied to the extracted and corrected LAFs. Due to the improved LAFs, the robust estimation did fewer iterations (3rd column) and, thus, it sped up (4th). Also, the proportion of found inliers is higher (5th), and the *estimated pose is more accurate* if the corrected LAFs were used (6–8th). In Fig. 4(f), the ground truth camera trajectory is shown.

5 Conclusions

A closed-form solution is proposed, optimal in the least-squares sense, for correcting the parameters of multi-view affine correspondences represented as a set of LAFs. The technique requires the epipolar geometry to be pre-estimated between each pair of views and makes the extracted LAFs consistent with the camera movement. It is validated both in synthetic experiments and on publicly available real-world datasets that the method almost always improves the input LAFs. As a by-product, a number of affine-covariant detectors are compared. On the used datasets, AKAZE with the view synthesizer of [23] leads to the most accurate LAFs. Also, it is shown that it makes the affine frames built on the output of partially affine-covariant detectors, *e.g.* SIFT, significantly more accurate. As potential applications, it is shown that the proposed correction improves homography, surface normal and relative motion estimation via improving the input of these methods. When affine frames are used, we see no reason for not applying the proposed technique.

Acknowledgements

Ivan Eichhardt and Daniel Barath were supported by the Hungarian Scientific Research Fund (No. NKFIH OTKA KH-126513 and K-120499). Also, Daniel Barath was supported by OP VVV project CZ.02.1.01/0.0/0.0/16019/000076 Research Center for Informatics.

References

[1] P. Alcantarilla, J. Nuevo, and A. Bartoli. Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. In *Proc. British Machine Vision Conf.*, pages 13.1–13.11, Bristol, 2013. British Machine Vision Association. ISBN 978-1-901725-49-0. 00423.

[2] D. Barath and L. Hajder. A theory of point-wise homography estimation. *Pattern Recognition Letters*, 94:7 – 14, 2017. ISSN 0167-8655.

[3] D. Barath and J. Matas. Multi-class model fitting by energy minimization and mode-seeking. In *Proc. European Conf. on Computer Vision*, pages 221–236, 2018.

[4] D. Barath, J. Molnar, and L. Hajder. Optimal Surface Normal from Affine Transformation. In *Proc. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Appl.*, 2015.

[5] D. Barath, L. Hajder, and J. Matas. Accurate closed-form estimation of local affine transformations consistent with the epipolar geometry. In *Proc. British Machine Vision Conf.*, 2016.

[6] D. Barath, J. Molnar, and L. Hajder. Novel methods for estimating surface normals from affine transformations. In *Proc. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Appl.* Springer International Publishing, 2016.

[7] D. Barath, T. Toth, and L. Hajder. A minimal solution for two-view focal-length estimation using two affine correspondences. In *Conf. on Computer Vision and Pattern Recognition*, 2017.

[8] D. Barath, I. Eichhardt, and L. Hajder. Optimal multi-view surface normal estimation using affine correspondences. *IEEE Trans. Image Processing*, 2019. ISSN 1057-7149.

[9] A. Baumberg. Reliable feature matching across widely separated views. In *Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 774–781, Hilton Head Island, SC, USA, 2000. IEEE Comput. Soc. ISBN 978-0-7695-0662-3.

[10] J. Bentolila and J. M. Francos. Conic epipolar constraints from affine correspondences. *Computer Vision and Image Understanding*, 2014.

[11] I. Eichhardt and D. Chetverikov. Affine correspondences between central cameras for rapid relative pose estimation. In *Proc. European Conf. on Computer Vision*, pages 488–503, 2018.

[12] I. Eichhardt and L. Hajder. Computer Vision Meets Geometric Modeling: Multi-view Reconstruction of Surface Points and Normals using Affine Correspondences. In *International Conf. on Computer Vision Workshops*, pages 2427–2435, 2017.

[13] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 32(11):1231–1237, 2013.

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

[15] Kevin Köser. *Geometric estimation with local affine frames and free-form surfaces*. PhD thesis, Kiel University, 2009.

[16] K. Lebeda, J. Matas, and O. Chum. Fixing the locally optimized RANSAC. In *Proc. British Machine Vision Conf.* Citeseer, 2012.

[17] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[18] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. British Machine Vision Conf.*, 2002.

[19] 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(1-2):43–72, 2005.

[20] D. Mishkin, J. Matas, and M. Perdoch. MODS: Fast and robust method for two-view matching. *Computer Vision and Image Understanding*, 2015.

[21] D. Mishkin, F. Radenovic, and J. Matas. Repeatability is not enough: Learning affine regions via discriminability. In *Proc. European Conf. on Computer Vision*, pages 284–300, 2018.

[22] J. Molnár and D. Chetverikov. Quadratic transformation for planar mapping of implicit surfaces. *Journal of Mathematical Imaging and Vision*, 2014.

[23] J-M. Morel and G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2009.

[24] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *Proc. International Conf. on Computer Vision*, pages 3248–3255, 2013.

[25] P. Moulon, P. Monasse, R. Perrot, and R. Marlet. OpenMVG: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016.

[26] S. Obdrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. In *Proc. British Machine Vision Conf.*, volume 1, page 3, 2002.

[27] M. Perdoch, J. Matas, and O. Chum. Epipolar geometry from two correspondences. In *Proc. International Conf. on Pattern Recognition*, volume 4, pages 215–219. IEEE, 2006.

[28] J. Pritts, Z. Kukelova, V. Larsson, and O. Chum. Radially-distorted conjugate translations. *Conf. on Computer Vision and Pattern Recognition*, 2018.

[29] J. Pritts, Z. Kukelova, V. Larsson, and O. Chum. Radially-distorted conjugate translations. In *Conf. on Computer Vision and Pattern Recognition*, pages 1993–2001, 2018.

[30] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and Jan-Michael Frahm. USAC: a universal framework for random sample consensus. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35 (8):2022–2038, 2013.

[31] C. Raposo and J. P. Barreto. Theory and practice of structure-from-motion using affine correspondences. In *Conf. on Computer Vision and Pattern Recognition*, pages 5470–5478, 2016.

[32] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Conf. on Computer Vision and Pattern Recognition*. IEEE, 2008.

[33] A. Vedaldi and B. Fulkerson. VLFeat - an open and portable library of computer vision algorithms. In *Proc. ACM Conf. on Multimedia*, 2010.

[34] H. Wang, D. Mirota, and G. D. Hager. A generalized kernel consensus-based robust estimator. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(1):178–184, 2010.

[35] Y. Xu, P. Monasse, T. Géraud, and L. Najman. Tree-based morse regions: A topological approach to local feature detection. *IEEE Trans. Image Processing*, 23(12):5612–5625, 2014.