Affine Correspondences between Central Cameras for Rapid Relative Pose Estimation

Iván Eichhardt[0000−0003−2294−5905] and Dmitry Chetverikov

MTA SZTAKI, Kende u. 13-17, 1111 Budapest, Hungary
{ivan.eichhardt,dmitry.chetverikov}@sztaki.mta.hu

Abstract. This paper presents a novel algorithm to estimate the relative pose, *i.e.* the 3D rotation and translation of two cameras, from two affine correspondences (ACs) considering any central camera model. The solver is built on new epipolar constraints describing the relationship of an AC and any central views. We also show that the pinhole case is a specialization of the proposed approach. Benefiting from the low number of required correspondences, robust estimators like LO-RANSAC need fewer samples, and thus terminate earlier than using the five-point method. Tests on publicly available datasets containing pinhole, fisheye and catadioptric camera images confirmed that the method often leads to results superior to the state-of-the-art in terms of geometric accuracy.

Keywords: relative pose, affine correspondences, central cameras

1 Introduction

Methods solving geometric computer vision problems using ACs typically use three times fewer correspondences [2] compared to point–based counterparts. This is also the case for this work, since with the proposed epipolar constraints, a total of three linear equations are yielded per correspondence.

A Local Affine Frame (LAF) is a pair $(\mathbf{x}, \mathbf{M})$ of a point $\mathbf{x} \in \mathbb{R}^2$ and a 2D affine transformation $\mathbf{M} \in \mathbb{R}^{2\times 2}$ which describes the local shape and orientation of the region. Scale invariant features are often sufficient for establishing correct Point Correspondences (PCs), which can then be used for solving computer vision problems. However, there are cases when affine invariant feature/region detectors are preferable [26]. LAFs can be obtained using affine invariant feature extractors [12–14, 16, 28]. From a pair of corresponding LAFs, $(\mathbf{x}_1, \mathbf{M}_1)$ and $(\mathbf{x}_2, \mathbf{M}_2)$, an AC $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{A})$ can be constructed, where $(\mathbf{x}_1, \mathbf{x}_2)$ is a PC and $\mathbf{A} = \mathbf{M}_2\mathbf{M}_1^{-1}$ is the affinity. In the planar case for perspective views (*i.e.* the pinhole camera model is valid), $\mathbf{A}$ is the gradient of the underlying homography.

Mainstream methods using PCs to solve computer vision problems completely disregard the information in $\mathbf{M}$. Hartley proposed the normalized eight–point algorithm [8] for determining the epipolar geometry between pinhole views. Nistér [17] developed a minimal, five–point solver for the relative pose problem. Minimal methods, *i.e.* minimal in the number of samples used for estimation,

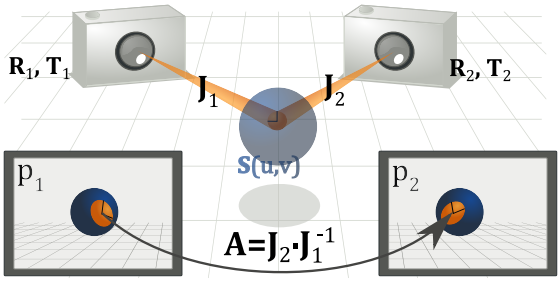

Fig. 1. Illustration of cameras represented by projection functions p_i and poses $\mathbf{R}_i$, $\mathbf{T}_i$, $i = 1, 2$. Parametric surface $S(u, v)$ has gradients $\mathbf{J}_i$ and the affinity $\mathbf{A}$.

are useful when dealing with combinatorially intensive problems, often used with robust methods [7, 10, 25] enabling the removal of numerous outliers.

Methods using ACs to estimate geometry, using the extra information in $\mathbf{A}$, are a new kind of "minimal" solvers typically trisecting the number of minimum samples compared to PC–based counterparts. However, to solve computer vision problems these methods consider only the strictly pinhole case [2, 9, 18, 20, 21] ignoring real–world cameras with a distortion or wide Field–of–View (FoV). These works rely on the fact that $\mathbf{A}$ and $\mathbf{H}$ are related and deduce their results using components of $\mathbf{H}$ and known properties of the epipolar geometry. Based on this relation, Köser and Koch [9] describes a method for camera resection using a single AC. For the epipolar geometry estimation, Riggi *et al.* [22] and Perdoch *et al.* [18] used ACs to generate additional PCs for a PC–based solver. Bentolila *et al.* [2] demonstrated that *three* ACs are sufficient for fundamental matrix estimation. Raposo and Barreto [21] presented a method for relative pose estimation using two ACs, for pinhole views. They demonstrated its applicability in rather controlled conditions. Baráth *et al.* [1] proposed a method to estimates the focal length with the fundamental matrix for pinhole views using two ACs.

In contrast to the above mentioned works, few use ACs for geometric model estimation between non-pinhole views [5, 15, 19]. Molnár and Eichhardt [15] generalized epipolar geometry using ACs. They proposed non-linear equations that constrain the geometry without using the essential matrix. Since the essential matrix is central to the relative pose estimation problem, the current work proposes novel *linear* constraints on its elements, directly applicable to arbitrary central cameras (including wide–FoV or omnidirectional). A new method for the estimation of the relative pose is also presented.

Contribution. We present, to the best of our knowledge, the first algorithm to solve the relative pose problem considering general camera models and using two ACs. New epipolar geometric constraints are introduced for an AC between general central cameras. The pinhole case [21] is a special case of the proposed one. Our approach needs no prior image un-distortion to operate. Using only two ACs for model estimation enables the faster operation of RANSAC and LO-

RANSAC as they take fewer samples compared to the five-point method. The method is validated on publicly available datasets consisting of pinhole, fisheye and catadioptric (360° FoV) camera images. The results presented are often superior to the state-of-the-art in terms of accuracy.

2 Mapping Between General Projective Views

Notations. 2D points are denoted as $\mathbf{x}$. Vectors are written in bold lower–case letters and matrices in bold capitals. Jacobians are denoted by the ∇ operator, *i.e.* $\nabla f(\mathbf{x}) = \left[\partial_1 f \dots \partial_n f\right](\mathbf{x}) \in \mathbb{R}^{m \times n}$, where f is differentiable at $\mathbf{x} \in \mathbb{R}^n$. Hereafter "chain rule" stands for the chain rule for differentiation.

Local Approximation of Projection. Consider a continuously differentiable parametric surface $S(\mathbf{z}) \in \mathbb{R}^3$, $\mathbf{z} \in \mathbb{R}^2$ and some function $p_i : \mathbb{R}^3 \to \mathbb{R}^2$ (basically, the camera model) projecting the 3D points of S onto the image plane:

$$\mathbf{x}_i \doteq p_i\left(\mathbf{R}_i\, S(\mathbf{z}_0) + \mathbf{T}_i\right) \tag{1}$$

for a point $\mathbf{z}_0$ of the parameter space of S, where $\mathbf{R}_i$ and $\mathbf{T}_i$ are the global rotation and translation (the pose) of view i, respectively. Applying the chain rule, the Jacobian of Eq. (1) is

$$\mathbf{J}_i \doteq \nabla_{\mathbf{z}}\left[\mathbf{x}_i\right] = \nabla p_i\left(\mathbf{R}_i\, S(\mathbf{z}_0) + \mathbf{T}_i\right)\, \mathbf{R}_i\, \nabla S(\mathbf{z}_0). \tag{2}$$

$\mathbf{J}_i$ can be interpreted as a local relative affine transformation between infinitesimal environments of the surface S at the point $\mathbf{z}_1$ and its projection at the point $\mathbf{x}_i$. See Fig. 1 for an additional explanation.

The "Affinity". Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be a mapping between the views as follows:

$$f(\mathbf{x}_1) = \mathbf{x}_2. \tag{3}$$

Assume that for all $\mathbf{z} \in \mathrm{dom}(S)$

$$f\left(p_1\left(\mathbf{R}_1\, S(\mathbf{z}) + \mathbf{T}_1\right)\right) = p_2\left(\mathbf{R}_2\, S(\mathbf{z}) + \mathbf{T}_2\right), \tag{4}$$

with respective p_i, $i \in \{1, 2\}$ and poses as denoted before, thus f being compatible with the epipolar geometry of the two views. The Taylor expansion of Eq. (4) around $\mathbf{x}_1$ is $f(\mathbf{y}) \approx \mathbf{x}_2 + \mathbf{A}(\mathbf{y} - \mathbf{x}_1)$, where $\mathbf{A}$ is the Jacobian of f, an *affinity*, *i.e.* a mapping between the infinitesimal environments of $\mathbf{x}_1$ and $\mathbf{x}_2$. The affinity can be expressed using $\mathbf{J}_i$, $i = 1, 2$ and the chain rule:

$$\mathbf{A} = \mathbf{J}_2 \mathbf{J}_1^{-1}. \tag{5}$$

In practice, $\mathbf{J}_i$ are related to LAFs $(\mathbf{x}_i, \mathbf{M}_i)$. The components of the affinity $\mathbf{M}_i$ of a pair of corresponding LAFs are related to Jacobians $\mathbf{J}_i$ at $\mathbf{x}_i$ by a mutual transformation $\mathbf{B}$: $\mathbf{M}_i = \mathbf{J}_i\mathbf{B}$. Thus $\mathbf{A}$ can be expressed using corresponding LAFs: $\mathbf{M}_2\mathbf{M}_1^{-1} = (\mathbf{J}_2\mathbf{B})\left(\mathbf{J}_1\mathbf{B}\right)^{-1} = \mathbf{J}_2\mathbf{B}\mathbf{B}^{-1}\mathbf{J}_1^{-1} = \mathbf{J}_2\mathbf{J}_1^{-1} = \mathbf{A}$.

3 Epipolar Constraints Based on an AC

Now consider $q_i : \mathbb{R}^2 \to \mathbb{R}^3$, $p_i \circ q_i = \mathrm{Id}_{\mathbb{R}^2}$, image-to-camera projection functions. The well–known epipolar constraint can be formulated using the PC ($\mathbf{x}_1$, $\mathbf{x}_2$) as

$$q_2\left(\mathbf{x}_2\right)^{\mathrm{T}} \mathbf{E}\, q_1\left(\mathbf{x}_1\right) = 0, \tag{6}$$

where $\mathbf{E} = \mathbf{R}\left[\mathbf{t}\right]_{\times}$ is the essential matrix. Using the bijection f and substituting $f\left(\mathbf{x}_1\right)$ for $\mathbf{x}_2$, the following equation for $\mathbf{x}_1$ is obtained:

$$q_2\left(f\left(\mathbf{x}_1\right)\right)^{\mathrm{T}} \mathbf{E}\, q_1\left(\mathbf{x}_1\right) = 0. \tag{7}$$

New Epipolar Constraints. Applying the gradient operator $\nabla_{\mathbf{x}_1}$ to both sides and using the chain product rule results in the following two *new epipolar constraints* that now use the AC ($\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{A}$):

$$\mathbf{A}^{\mathrm{T}}\left(\nabla q_2\left(\mathbf{x}_2\right)\right)^{\mathrm{T}} \mathbf{E}\, q_1\left(\mathbf{x}_1\right) + \left(\nabla q_1\left(\mathbf{x}_1\right)\right)^{\mathrm{T}} \mathbf{E}^{\mathrm{T}} q_2\left(\mathbf{x}_2\right) = \mathbf{0}, \tag{8}$$

since $\nabla_{\mathbf{x}_1}\left[q_2\left(f\left(\mathbf{x}_1\right)\right)\right] = \nabla q_2\left(\mathbf{x}_2\right)\mathbf{A}$.

Since $\mathbf{x}_1$ has two components, the gradient provides two extra equations (one for each partial derivative) in addition to the epipolar constraint (6). This means that three constraints are given for each correspondence reducing from 8 to 3 the number of samples required to estimate the elements of $\mathbf{E}$.

4 Relative Pose Using Two ACs

The epipolar constraint (6) can also be written as

$$\widetilde{\mathbf{v}}\widetilde{\mathbf{E}} = 0, \tag{9}$$

where

$$\begin{aligned}
\widetilde{\mathbf{v}} &= \left[w_x\mathbf{v}^{\mathrm{T}},\, w_y\mathbf{v}^{\mathrm{T}},\, w_z\mathbf{v}^{\mathrm{T}}\right], \\
\widetilde{\mathbf{E}} &= \left[e_{11},\, e_{12},\, e_{13},\, e_{21},\, e_{22},\, e_{23},\, e_{31},\, e_{32},\, e_{33}\right]^{\mathrm{T}}.
\end{aligned}$$

The line vector $\widetilde{\mathbf{v}}$ is constructed from the components of $\mathbf{v} = q_1\left(\mathbf{x}_1\right)$ and $\mathbf{w} = q_2\left(\mathbf{x}_2\right) = \left[w_x, w_y, w_z\right]^{\mathrm{T}}$. $\widetilde{\mathbf{E}}$ is a vector containing the elements of $\mathbf{E}$:

$$\mathbf{E} = \begin{bmatrix} e_{11} & e_{12} & e_{13} \\ e_{21} & e_{22} & e_{23} \\ e_{31} & e_{32} & e_{33} \end{bmatrix}.$$

The two rows of Eq. (8) can be formulated in a similar manner as follows:

$$\widetilde{\mathbf{Q}}\widetilde{\mathbf{E}} = \mathbf{0}, \tag{10}$$

where

$$\begin{aligned}\widetilde{\mathbf{Q}} &= \left[w_x\mathbf{V},\, w_y\mathbf{V},\, w_z\mathbf{V}\right] + \mathbf{A}^{\mathrm{T}}\left[\mathbf{W}_1\mathbf{v}^{\mathrm{T}},\, \mathbf{W}_2\mathbf{v}^{\mathrm{T}},\, \mathbf{W}_3\mathbf{v}^{\mathrm{T}}\right], \\ \mathbf{V} &= \left(\nabla q_1\left(\mathbf{x}_1\right)\right)^{\mathrm{T}}, \\ \mathbf{W} &= \left(\nabla q_2\left(\mathbf{x}_2\right)\right)^{\mathrm{T}} = \left[\mathbf{W}_1\ \mathbf{W}_2\ \mathbf{W}_3\right].\end{aligned}$$

Now let us construct a matrix $\widetilde{\mathbf{B}} \in \mathbb{R}^{3\times 9}$ whose first row is $\widetilde{\mathbf{v}}^{\mathrm{T}}$, while the second and the third rows are the two rows of $\widetilde{\mathbf{Q}}$, respectively. The compound system that describes the relation of the essential matrix to an AC is as follows:

$$\widetilde{\mathbf{B}}\widetilde{\mathbf{E}} = \mathbf{0}. \tag{11}$$

The matrices $\widetilde{\mathbf{B}}^{(j)}$, $j = 1, 2, 3$, can be constructed using three different ACs. The null–space of the compound system of these matrices is $\widetilde{\mathbf{E}}$, which provides the elements of the essential matrix $\mathbf{E}$ up to a scale.

With more correspondences, an over–determined system can also be constructed. Its solution is the singular vector with the smallest singular value.

4.1 "2AC" Solver – Essential Matrix From Two Correspondences

The essential matrix $\mathbf{E}$ has 5 degrees of freedom since one of its singular values is zero with its two non–zero ones being equal, which leads to the following cubic constraints [6, 17] on $\mathbf{E}$:

$$2\mathbf{E}\mathbf{E}^{\mathrm{T}}\mathbf{E} - \mathrm{tr}\left(\mathbf{E}\mathbf{E}^{\mathrm{T}}\right)\mathbf{E} = \mathbf{0}. \tag{12}$$

Also, since one of the singular values of $\mathbf{E}$ is zero

$$\det\left(\mathbf{E}\right) = 0. \tag{13}$$

The five–point solvers for essential matrix estimation [11, 17] use the null–space of a 5×9 matrix or the four singular vectors of an over–determined system, corresponding to the least four singular values, take their linear combination with coefficients $x, y, z, 1$ and substitute it into equations (12) and (13) which give a polynomial system. The solutions of the polynomial system can then be back–substituted into the linear combination. The essential matrix can be decomposed into rotation and translation, after handling ambiguities [8, 11, 17].

Similarly to the five–point algorithm, one can construct a solution using only two ACs, hence the name of the solver is "2AC". The proposed solver *approximates* the four–dimensional nullspace using SVD. That is, (11) yields 3 equations per correspondence, the resulting 6×9 coefficient matrix would have a 3-dimensional nullspace. Instead, the four right singular vectors are used, corresponding to the least four singular values of the SVD decomposition.

4.2 Special Case: Pinhole Cameras

State of the art methods using LAFs to estimate epipolar geometry [2, 3, 21] rely on perspective views (*i.e.* pinhole camera). Our approach handles any central projection cameras (*e.g.* wide–FoV or panoramic ones) in a stereo configuration, allowing for a wider range of applications.

The pinhole camera case is a special case of the proposed one. From the relation of the homography and the affinity, Raposo and Barreto [21] derived a matrix equation yielding three equations for the epipolar constraint using an AC. Note that in this paper *no existence of a homography was assumed* between the views, f can be any, more general, or higher–order mapping. This work and their formulation shows that: (i) the first row in their work is the well–known epipolar constraint for a point correspondence, Eq. (9); (ii) and the second and third rows are equivalent to Eq. (10).

Let $\mathbf{v} = \begin{bmatrix} x_1 & x_2 & 1 \end{bmatrix}^{\mathrm{T}}$ and $\mathbf{w} = \begin{bmatrix} y_1 & y_2 & 1 \end{bmatrix}^{\mathrm{T}}$, thus, $\nabla q_i(\mathbf{x}_i)$ is also modified:

$$\nabla q_i(\mathbf{x}_i) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{14}$$

Substituting $\mathbf{v}$, $\mathbf{w}$ and $\nabla q_i(\mathbf{x}_i)$ into equations (9) and (10) yields (15) and (16), respectively. Together, they form the 22nd equation of [21].

$$\begin{bmatrix} x_1y_1 & x_1y_2 & x_1 & x_2y_1 & x_2y_2 & x_2 & y_1 & y_2 & 1 \end{bmatrix} \widetilde{\mathbf{E}} = 0, \tag{15}$$

$$\begin{bmatrix} a_1x_1{+}y_1 & a_3x_1{+}y_2 & 1 & a_1x_2 & a_3x_2 & 0 & a_1 & a_3 & 0 \\ a_2x_1 & a_4x_1 & 0 & a_2x_2{+}y_1 & a_4x_2{+}y_2 & 1 & a_2 & a_4 & 0 \end{bmatrix} \widetilde{\mathbf{E}} = \mathbf{0}. \tag{16}$$

5 Handling Noise in ACs

A PC can be considered a "0th–order", an AC a "1st–order" information, which is more sensitive to noise. This section discusses how to cope with noisy ACs.

Extracting LAFs. The VLFeat library [27] is capable of extracting covariant features using different scale–space based detectors and the affine shape adaptation algorithm [12, 14]. It is also capable of extracting dominant gradient directions from shape–adapted local frame of pixels. The number of iterations and the patch size used in these steps are sufficient for obtaining a robust descriptor, but the affine part of the resulting LAF is rather susceptible to noise. By tuning these parameters, one can enhance the applicability of LAFs for AC–based algorithms. Note that in the tests the default settings of VLFeat were used.

Photometric Refinement. After establishing correspondences, the affine part ($\mathbf{A}$) of ACs can be further refined [21] minimizing the photometric discrepancy between the LAFs. The drawback of this approach is an extra time demand over feature extraction, although it can be massively parallelized. Note that in the

tests, photometric refinement was primarily applied in the semi–synthetic and, partially, in the real–world evaluation. The rest of the real–world tests show that using Locally Optimized RANSAC (Sec. 5.1) has additional benefits, *e.g.* it is significantly less time–consuming, but still provides high accuracy.

5.1 Locally Optimized RANSAC

Sampling noisy ACs *without photometric refinement*, compared to PCs, might yield fewer robust hypotheses during traditional RANSAC iterations. However, there are two benefits of using ACs: *(a)* these hypotheses are still close to the true model; *(b)* combinatorially, sampling two elements is much better compared to samples of five elements: $\binom{N}{2} \ll \binom{N}{5}$. These benefits have the potential to boost LO-RANSAC [4, 10] approach, enabling rapid runtime, with significantly fewer RANSAC-iterations and local optimization steps [10].

Hybrid LO-RANSAC. In this paper, a modified version of LO$^+$ [10] was applied as follows: First, (i) sample *minimal two-sets* of correspondences and use the proposed solver "2AC" for generating hypotheses; then (ii) apply the *local optimization step* to refine the support set of the most recently selected maximal hypothesis. See real–world tests (Sec. 6.4) for details of the performance of this LO-RANSAC approach.

6 Experimental results

Since the essential matrix estimation for the pinhole case [21] is a special case of the proposed one, the evaluation will mainly focus on more general, central–projection models, such as (i) cameras with fisheye lens; (ii) catadioptric cameras; (iii) and other cameras with radial and tangential distortion.

Robustified versions of the five–point algorithm "5PT" [17] and variants of the proposed approach are compared using two and five correspondences denoted as "2AC" and "5AC". To obtain their robustified versions, MSAC [25] was applied. The minimum and maximum number of iterations were set to 10 and 2048, respectively, and failure probability of the estimator was set to 10^{-5}. The angular error metric $\sin^{-1}\left(\frac{q_2(\mathbf{x}_2)^{\mathrm{T}}\mathbf{E}q_1(\mathbf{x}_1)}{\|\mathbf{E}q_1(\mathbf{x}_1)\|}\right)$ was used with the MSAC whose threshold was normally set to 0.15° unless otherwise stated.

6.1 Synthetic Tests

In the synthetic tests 2AC and 5PT are compared. The synthetic scene consisted of 5 oriented points uniformly sampled from the range $\left[-1,\,1\right]^3$, with surface normals sampled on the unit sphere, viewed by two pinhole cameras having radial distortion. The distance of the camera centers from the origin varied from 2 to 3 units, the distance between the cameras from 0.1 to 1.0 units. The optical axes intersected in a point uniformly sampled from $\left[-1,\,1\right]^3$.

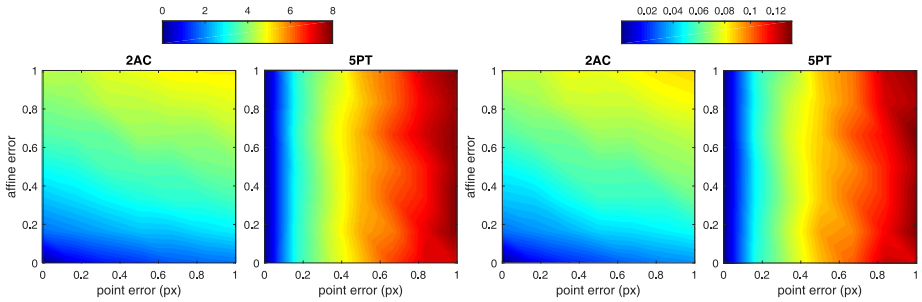

Fig. 2. Plots of sensitivity to noise in points (*axis x*) and affine (*axis y*) components.

To obtain PCs, oriented points were projected to the camera images. The affine parameters of ACs were calculated using the surface normals based on Eq. (2). Two uncorrelated sources of Gaussian noise, σ_p and σ_a, were added to the points in $\mathbb{R}^2$ and the affine parameters in $\mathbb{R}^{2\times 2}$, respectively. For each level of noise σ_p and σ_a, the test was repeated 100 times building the synthetic scene, using 2AC and 5PT, and averaging rotation and translation errors. The results are shown on Fig. 2. For low levels of σ_a, 2AC always outperforms 5PT. However, stronger noise in affine parameters deteriorates the results of 2AC. 5PT is of course not affected by σ_a. Note that noise added to 2D positions is a realistic model of real–world conditions, while noise added to the affinity is a less realistic one.

6.2 Stability Tests

In this section the numerical stability of the proposed solver is compared to existing work and their behavior on different levels of synthetic noise is also investigated. It is important to note that these stability tests are all performed using a *pinhole camera with no distortion* since the solver of [21] is only designed for the pinhole camera model. The setup for these tests are similar to the one described in the previous section. The stability test shows the distribution of the matrix error $\min\left(\|\mathbf{E}-\mathbf{E}_{gt}\|, \|\mathbf{E}+\mathbf{E}_{gt}\|\right)$ from 30000 samples. All results can be seen in Fig. 3.

With no noise added, the pinhole camera based solver [21] shows a slightly better stability than the 5-point solver of Nistér [17] and 2AC. The proposed solver here performs worse since 2AC uses an approximate nullspace acquired using SVD. All six linear equations are used that can be formed from two Affine Correspondences, to estimate a four–dimensional nullspace instead of their true, three–dimensional nullspace.

As the level of synthetic noise added to point coordinates increases, the 5-point solver becomes the worst among the three. 2AC and the pinhole-based method [21] show similar stability to the previous test. However, the solvers begin

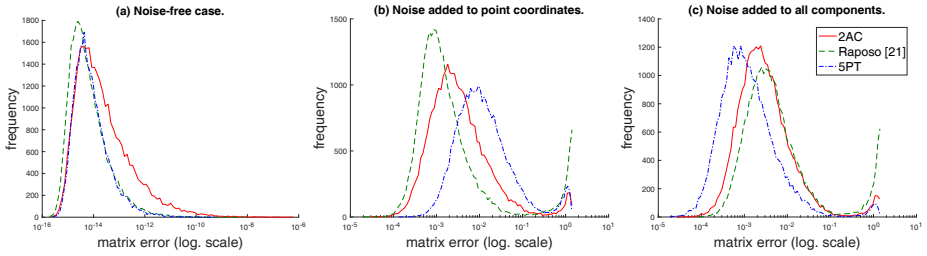

Fig. 3. Histograms of stability tests with (*left*) no noise; (*center*) noise in 2D coordinates; and (*right*) noise in 2D coordinates and affinities. Horizontal axes are log. scales of error exponents and vertical axes are their frequency.

to produce larger errors to the right of the "10^0" on the horizontal logarithmic scale of the diagrams. These estimations failed. The largest number of failed cases are produced by the method in [21].

Adding synthetic noise to both point coordinates and affinity components results in the third diagram of Fig. 3. In this case solvers based on AC–s perform worse compared to the 5-point solver. The second best is 2AC and [21] is third.

6.3 Semi–Synthetic Tests

The semi–synthetic tests were based on the Multi–FoV dataset [29], with ray–traced views of scenes through perspective, fisheye and catadioptric [23] cameras. The dataset provides two scenes with the cameras traversing them, obtaining ground–truth poses, color images and depth maps. For the tests, the ground truth 3D points were sampled from the depth maps of the scene "vfr".

Similarly to the synthetic tests, PCs are established using the known spatial points. The affine transformation part ($\mathbf{A}$) of each AC was initially set to the identity matrix, then refined using gradient–descent based photometric refinement on local areas of the color images, similarly as in the work of Raposo and Barreto [21]. The refinement used the symmetric cost function of summed squared differences between local patches of the color images. The patches were 20×20 pixel size windows centered on the points of a PC. The matrix $\mathbf{A}$ was refined maximizing photometric similarity. Outliers were uniformly sampled from the image space prior to photometric refinement.

In these tests, the robustified versions of 5PT and 2AC were compared. Using both methods, essential matrices and the corresponding sets of inliers were estimated. Essential matrices were then decomposed to relative rotation and translation, to be further refined using bundle adjustment (BA) on the inlier set. For different numbers of input inliers and outliers and levels of 2D noise, the performance of the methods were evaluated based on (i) mean and root mean square (RMS) errors of relative rotation and translation; (ii) runtime and number of iterations; and (iii) precision and recall.

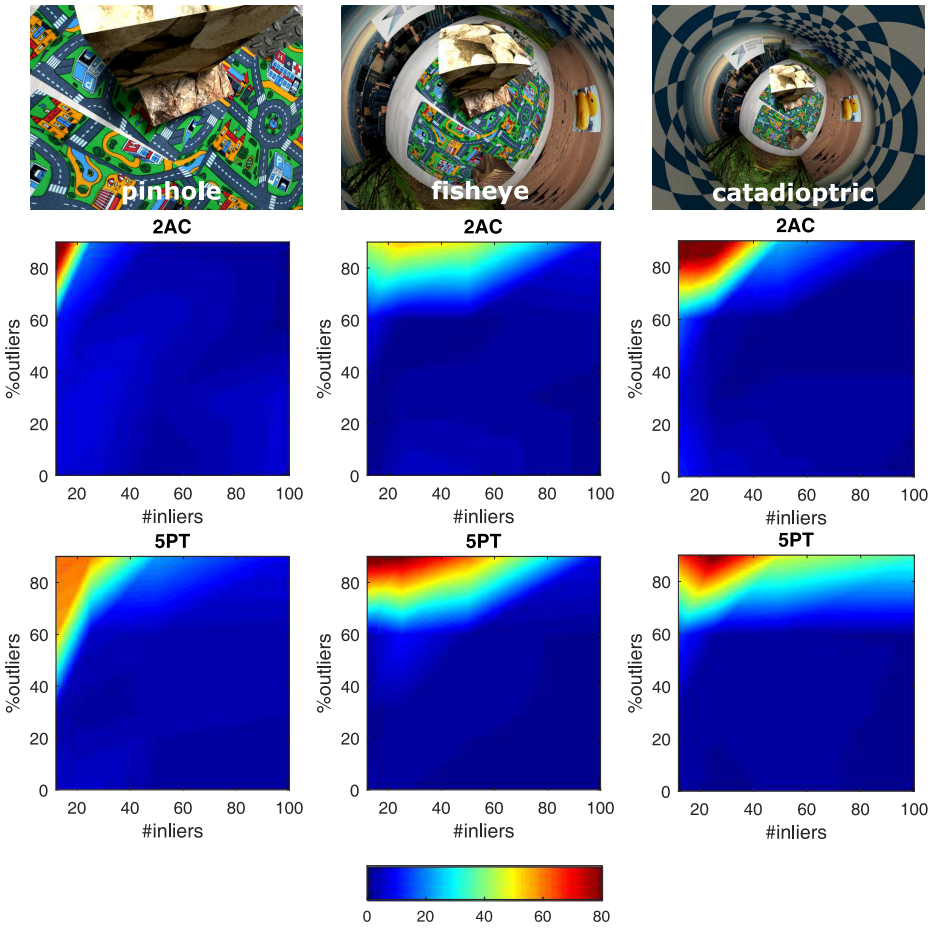

Fig. 4. Rotation errors versus number of inliers and percentage of outliers for 2AC (*2nd row*) and 5PT (*3rd row*). Frames (*1st row*) from the "vfr" scene [29]. Left to right: perspective, fisheye, Catadioptric views. Errors are measured after bundle adjustment.

Fig. 4 shows rotation errors for the pinhole, fisheye and catadioptric cameras, comparing the effect of different levels of inliers and outliers in the sample set. The plots indicate that 5PT is the most sensitive to decreasing number of inliers and increasing number of outliers.

The effect of noise on 2D coordinates was also analyzed while adding more outliers. The fisheye model was used on a dataset of 100 matches. The results are presented in Fig. 5. The plots show the average precision, recall, number of iterations, and runtime. 2AC has the highest precision and the smallest runtime and number of iterations, but its recall decreases with increasing noise more rapidly compared to 5PT. However, higher precision is usually of greater importance since BA for higher precision, *i.e.* higher rates of inliers result in better pose estimation.

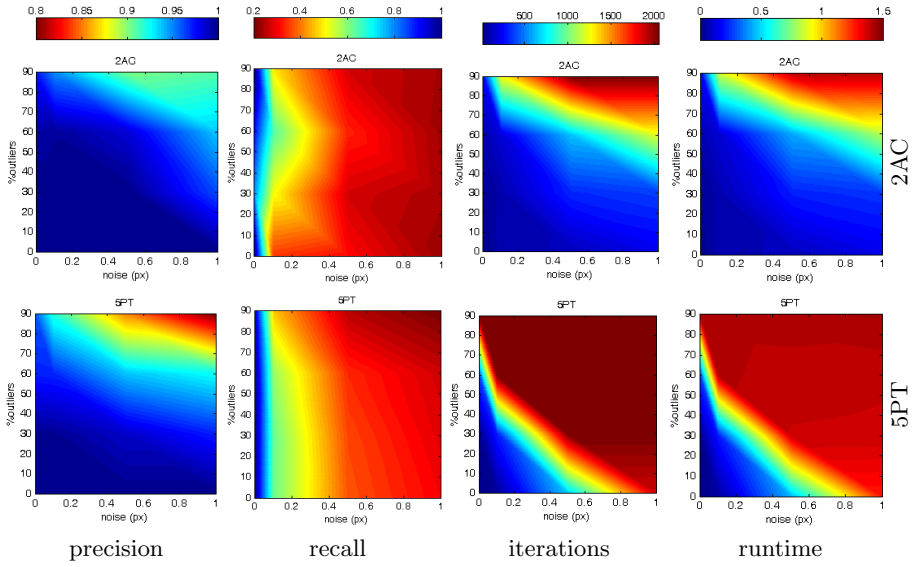

Fig. 5. Results for various levels of noise and outliers: 2AC (*top*) and 5PT (*bottom*).

6.4 Real–world Tests

There are two parts of this real–world evaluation: *(A)* extracted correspondences are further enhanced using simple photometric refinement (see Sec. 5), and in *(B)* the features are used without refinement, but locally optimized RANSAC (see Sec. 5.1) is used to provide high–quality results.

(A) With Photometric Refinement. In this section the proposed approach is compared to the five–point algorithm using image pairs from the Strecha Dense

MVS dataset [24]. The input features were extracted using an affine–invariant version of the Difference of Gaussians (DoG) extractor [27] and photometrically refined as in the semi–synthetic tests. As before, the estimated relative pose was refined by performing BA on the inlier set obtained by the robust estimator. Each test was repeated 100 times with the input features and matches unaltered. Table 1 shows the evaluation of the methods 2AC, 3AC, 5AC and 5PT on the Strecha Dense MVS dataset [24]. The table contains four columns for each scene and for each method: rotation RMSE in degrees, translation RMSE normalized to the ground truth, timing in seconds and number of RANSAC iterations. Regarding rotation and translation errors, 5AC performs best while 2AC and 3AC perform worse than 5PT. As for the runtime and number of RANSAC iterations, the estimators using two or three affine correspondences are the best, and for two scenes out of three, 5AC has lower runtime and number of iterations compared to 5PT. The Dense MVS dataset [24] contains scenes with rather

Table 1. Evaluation of relative motion estimation on the Dense MVS dataset [24]. Top rows: three scenes of, brackets containing images pairs and numbers of correspondences extracted. Columns: solvers, (ρ) rotation and (τ) translation errors, (t) timing in seconds and (n) number of iterations for each scene and for each method. Errors are measured after bundle adjustment. Best results are highlighted.

	`castle` (0001–0002)				`fountain` (0004–0006)				`herzjesus` (0005–0006)			
#	7153				7530				1992			
	ρ	τ	t	n	ρ	τ	t	n	ρ	τ	t	n
`2AC`	0.073°	0.0038	**0.0143**	**10**	0.038°	0.0020	**0.0166**	**10**	0.029°	0.0045	0.0180	**15**
`3AC`	0.056°	0.0031	0.0145	**10**	0.035°	0.0019	0.0195	**10**	0.000°	0.0020	**0.0169**	17
`5AC`	**0.043°**	**0.0025**	0.0244	15	**0.025°**	**0.0015**	0.0194	**10**	**0.051°**	**0.0009**	0.0266	23
`5PT`	0.052°	0.0032	0.0256	15	0.027°	0.0016	0.0202	**10**	0.080°	0.0015	0.0213	21

diverse geometry and texture. The extracted affine correspondences can be less reliable compared to the ones in the semi–synthetic tests. In the real–world tests, 3AC outperforms 2AC. We believe that adding more correspondences to the otherwise minimal solver of 2AC, *e.g.* using 5AC, will increase the reliability of estimations resulting in a better inlier set facilitating the BA of relative pose.

(B) Using Locally Optimized RANSAC. These tests, in contrast to above, were performed without using photometric refinement. RANSAC ("RSC") and a modified version of "LO$^+$" [10] performed robust estimation, using the five-point solver "5PT" and the proposed one "2AC". See Sec. 5.1 for more details on

the proposed locally optimized approach. These different pairings are denoted as follows: 2AC-RSC, 2AC-LO$^+$, 5PT-RSC, 5PT-LO$^+$, undistort+[21]. Images of the test database are shown in Fig. 7. The images were taken by a Point Gray Blackfly camera with YV2.8x2.8SA-2 wide-FoV lens attached.

Feature extraction was performed on the raw images, *without un-distortion*. No photometric refinement was performed on ACs, they were directly fed to the robust methods with the solvers "2AC" and "5PT". "undistort+[21]" denotes [21], applied to undistorted ACs (see Eq. (13) in supp. material).

Fig. 6 shows the evaluation results on the first two images of Fig. 7. It is clear that 2AC-LO$^+$ outperforms all other variants in terms of speed (3 to 8 times better runtime), number of iterations (orders of magnitude fewer) and local optimization steps. The inlier sets returned by 2AC and 5PT using LO$^+$ are nigh-identical, but larger than the RANSAC-only variants. The overall performance of [21] is the worst. The supplementary contains further comparative evaluation using several real-world cases and feature extractors.

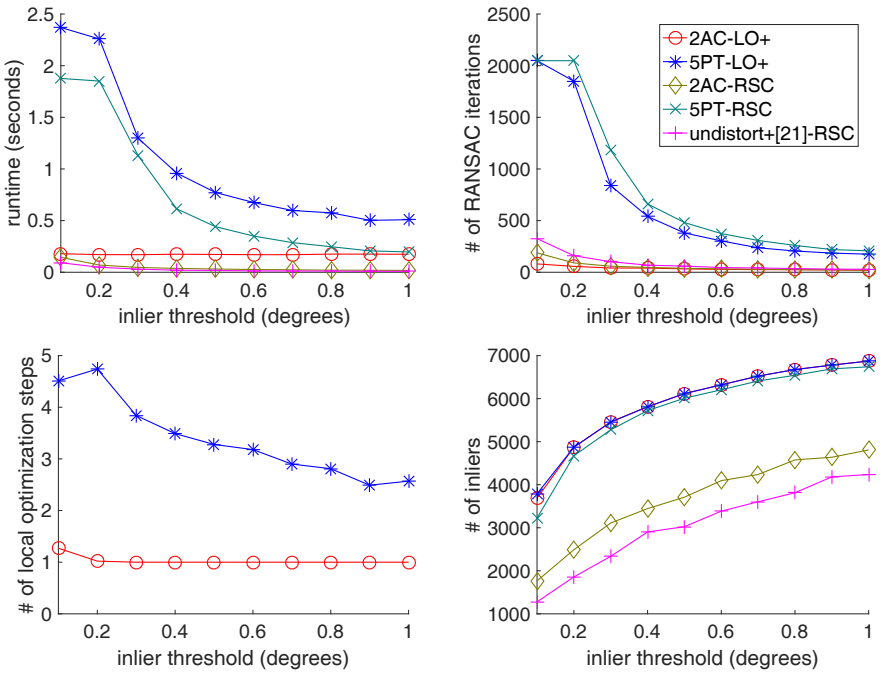

Fig. 6. Real-world ("Sarok" dataset) evaluation of RANSAC "RSC" and LO-RANSAC "LO$^+$" robust estimation using the proposed two-point "2AC" and the five-point "5PT" solvers. "undistort+[21]" denotes [21] using RANSAC applied to undistorted ACs (see d^{-1} in supplementary material). Diagrams compare (*top*) runtime and number of iterations; and (*bottom*) number of LO-steps and number of inliers.

Fig. 7. Image of a scene "Sarok" used for real-world evaluation taken by a Point Gray Blackfly camera with YV2.8x2.8SA-2 wide-FoV lens attached.

7 Conclusion

In the paper a new method (2AC) was presented for relative pose estimation based on novel epipolar constraints using Affine Correspondences. The minimum number of correspondences needed for pose estimation is reduced to two. The method is applicable to arbitrary central–projection models including cameras [23] with wide fields of view (*e.g.* over 180° or omnidirectional). The pinhole–camera based approach [21] was shown to be a specialization of the proposed one. Stability tests showed that if the "affinity" is noisy, the pinhole based method [21] is outperformed. Additionally, 2AC needs no prior image un-distortion. Tests indicate that the five–point algorithm [17] is inferior in runtime and in the number of iterations when using MSAC [25] and LO$^+$ [10]. The quality of estimated pose is also worse after bundle adjustment. The proposed LO-RANSAC approach uses raw ACs to provide state-of-the-art quality in less time.

Based on the new epipolar constraints, other AC-based solvers can be constructed, *e.g.* to estimate additional camera parameters. With more constraints given per correspondence, fewer samples are needed for model estimation, thus robust estimation combined with such a solver will terminate earlier. The supplementary material contains additional evaluation and other material, *e.g.* Jacobians of projection functions.

References

1. Barath, D., Toth, T., Hajder, L.: A Minimal Solution for Two-View Focal-Length Estimation Using Two Affine Correspondences. In: Conf. on Computer Vision and Pattern Recognition (July 2017)
2. Bentolila, J., Francos, J.M.: Conic epipolar constraints from affine correspondences. Computer Vision and Image Understanding **122**, 105–114 (2014)
3. Bentolila, J., Francos, J.M.: Homography and Fundamental Matrix Estimation from Region Matches Using an Affine Error Metric. Journal of Mathematical Imaging and Vision **49**, 481–491 (2014)
4. Chum, O., Matas, J., Kittler, J.: Locally optimized RANSAC. In: Michaelis, B., Krell, G. (eds.) Pattern Recognition. pp. 236–243. Springer, Berlin, Heidelberg (2003)

5. Eichhardt, I., Hajder, L.: Computer vision meets geometric modeling: Multi-view reconstruction of surface points and normals using affine correspondences. In: International Conf. on Computer Vision Workshops. pp. 2427–2435 (Oct 2017)
6. Faugeras, O.: Three-dimensional computer vision: a geometric viewpoint. M. I. T. Press (1993)
7. Fischler, M.A., Bolles, R.C.: 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)
8. Hartley, R.I.: In defense of the eight-point algorithm. IEEE Trans. Pattern Analysis and Machine Intelligence **19**(6), 580–593 (1997)
9. Köser, K., Koch, R.: Differential spatial resection-pose estimation using a single local image feature. In: Proc. European Conf. on Computer Vision. pp. 312–325. Springer (2008)
10. Lebeda, K., Matas, J., Chum, O.: Fixing the locally optimized RANSAC–full experimental evaluation. In: Proc. British Machine Vision Conf. pp. 1–11. Citeseer (2012)
11. Li, H., Hartley, R.: Five-point motion estimation made easy. In: Proc. International Conf. on Pattern Recognition. vol. 1, pp. 630–633. IEEE (2006)
12. Lindeberg, T., Gårding, J.: Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. Image and Vision Computing **15**(6), 415–434 (1997)
13. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing **22**(10), 761–767 (2004)
14. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Proc. European Conf. on Computer Vision. pp. 128–142. Springer (2002)
15. Molnár, J., Eichhardt, I.: A differential geometry approach to camera-independent image correspondence. Computer Vision and Image Understanding (2018)
16. Morel, J.M., Yu, G.: ASIFT: A new framework for fully affine invariant image comparison. SIAM Journal on Imaging Sciences **2**(2), 438–469 (2009)
17. Nistér, D.: An efficient solution to the five-point relative pose problem. IEEE Trans. Pattern Analysis and Machine Intelligence **26**(6), 756–770 (2004)
18. Perdoch, M., Matas, J., Chum, O.: Epipolar geometry from two correspondences. In: Proc. International Conf. on Pattern Recognition. vol. 4, pp. 215–219. IEEE (2006)
19. Pritts, J., Kukelova, Z., Larsson, V., Chum, O.: Radially-distorted conjugate translations. In: Conf. on Computer Vision and Pattern Recognition (June 2018)
20. Raposo, C., Barreto, J.P.: πMatch: Monocular vSLAM and Piecewise Planar Reconstruction Using Fast Plane Correspondences. In: Proc. European Conf. on Computer Vision. pp. 380–395. Springer (2016)
21. Raposo, C., Barreto, J.P.: Theory and Practice of Structure-from-Motion using Affine Correspondences. In: Conf. on Computer Vision and Pattern Recognition. pp. 5470–5478 (2016)
22. Riggi, F., Toews, M., Arbel, T.: Fundamental matrix estimation via TIP-transfer of invariant parameters. In: Proc. International Conf. on Pattern Recognition. vol. 2, pp. 21–24. IEEE (2006)
23. Scaramuzza, D., Martinelli, A., Siegwart, R.: A flexible technique for accurate omnidirectional camera calibration and structure from motion. In: Proc. IEEE Conf. on Computer Vision Systems. pp. 45–45. IEEE (2006)

24. Strecha, C., Von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: Conf. on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
25. Torr, P., Zisserman, A.: Robust computation and parametrization of multiple view relations. In: Conf. on Computer Vision and Pattern Recognition. pp. 727–732. IEEE (1998)
26. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: a survey. Foundations and trends® in computer graphics and vision **3**(3), 177–280 (2008)
27. Vedaldi, A., Fulkerson, B.: VLFeat - an open and portable library of computer vision algorithms. In: Proc. ACM Conf. on Multimedia (2010)
28. Xu, Y., Monasse, P., Géraud, T., Najman, L.: Tree-based morse regions: A topological approach to local feature detection. IEEE Trans. Image Processing **23**(12), 5612–5625 (2014)
29. Zhang, Z., Rebecq, H., Forster, C., Scaramuzza, D.: Benefit of large field-of-view cameras for visual odometry. In: Proc. IEEE Conf. on Robotics and Automation. pp. 801–808. IEEE (2016)