A Minimal Solution for Two-view Focal-length Estimation using Two Affine Correspondences

Daniel Barath, Tekla Toth, and Levente Hajder
Machine Perception Research Laboratory
MTA SZTAKI, Budapest, Hungary
{barath.daniel,hajder.levente}@sztaki.mta.hu

Abstract

A minimal solution using two affine correspondences is presented to estimate the common focal length and the fundamental matrix between two semi-calibrated cameras – known intrinsic parameters except a common focal length. To the best of our knowledge, this problem is unsolved. The proposed approach extends point correspondence-based techniques with linear constraints derived from local affine transformations. The obtained multivariate polynomial system is efficiently solved by the hidden-variable technique. Observing the geometry of local affinities, we introduce novel conditions eliminating invalid roots. To select the best one out of the remaining candidates, a root selection technique is proposed outperforming the recent ones especially in case of high-level noise. The proposed 2-point algorithm is validated on both synthetic data and 104 publicly available real image pairs. A Matlab implementation of the proposed solution is included in the paper.

1. Introduction

The recovery of camera parameters and scene structure have been studied for over two decades since several applications, such as 3D vision from multiple views [13], are heavily dependent on the quality of the camera calibration. In particular, two major calibration types can be considered: aiming at the determination of the intrinsic and/or extrinsic parameters. The former ones include focal lengths, principal point, aspect ratio, and non-perspective distortion parameters, while the extrinsic parameters are the relative pose. Assuming two cameras with unknown extrinsic and *a priori* intrinsic parameters except a common focal length is called the *semi-calibrated case* [19]. It leads to the *unknown focal-length problem*: estimation of the relative motion and common focal length, simultaneously. The semi-calibrated case is realistic since (1) the aspect ratio is determined by the shape of the pixels on the sensors, it is usually 1:1; (2) the principal point is close to the center of the image, thus it is a reasonable approximation and (3) the distortion can be omitted if narrow field-of-view lenses are applied. Considering solely the locations of point pairs makes the problem solvable using at least six point pairs [19, 30, 31]. The objective of this paper is to *solve the problem exploiting only two local affine transformations*.

In general, 3D vision approaches [13] including state-of-the-art structure-from-motion pipelines [1, 7, 11, 24] apply a robust estimator, e.g. RANSAC [10], augmented with a minimal method, such as the five [25] or six-point [19] algorithm as an engine. Selecting a method exploiting as few point pairs as possible gains accuracy and drastically reduces the processing time. Benefiting from estimators which use less input data, the understanding of low-textured environment becomes significantly easier [28]. Moreover, minimal methods are advantageous from theoretical point-of-view leading to deeper understanding.

Local affine transformations represent the warp between the infinitely close vicinities of corresponding point pairs [15] and have been investigated for a decade. Their application field includes homography [4] and surface normal [15, 5] estimation; recovery of the epipoles [6]; triangulation of points in 3D [15]; camera pose estimation [16]; structure-from-motion [28]. In practice, local affinities can be accurately retrieved [3, 22] using e.g. affine-covariant feature detectors, such Affine-SIFT [23] and Hessian-Affine [21]. To the best of our knowledge, no paper has dealt with the unknown focal length problem using local affine transformations.

This paper proposes two novel linear constraints describing the relationship between local affinities and epipolar geometry. Forming a multivariate polynomial system and solving it by the *hidden-variable technique* [9], the proposed method is efficient and estimates the focal length and the relative motion using only two affinities. In order to eliminate invalid roots, a novel condition is introduced investigating the geometry of local affinities. To select the best candidate out of the remaining ones, we propose a root

selection technique which is as accurate as the state-of-the-art for small noise and outperforms it for high-level noise.

2. Preliminaries and Notation

Epipolar geometry. Assume two perspective cameras with a common intrinsic camera matrix $\mathbf{K}$ to be known. Fundamental and essential matrices [13] are as follows:

$$\mathbf{F} = \begin{bmatrix} f_1 & f_2 & f_3 \\ f_4 & f_5 & f_6 \\ f_7 & f_8 & f_9 \end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix} e_1 & e_2 & e_3 \\ e_4 & e_5 & e_6 \\ e_7 & e_8 & e_9 \end{bmatrix}.$$

If the cameras are calibrated ($\mathbf{K}$ is known) matrix $\mathbf{F}$ can be transformed to be an essential matrix $\mathbf{E}$ as follows:

$$\mathbf{E} = \mathbf{K}^T \mathbf{F} \mathbf{K}. \tag{1}$$

The epipolar relationship of corresponding point pair $\mathbf{p}_1$ and $\mathbf{p}_2$ are described by $\mathbf{F}$ as

$$\mathbf{p}_2^T \mathbf{F} \mathbf{p}_1 = 0. \tag{2}$$

A valid fundamental matrix must satisfy singularity constraint $\det(\mathbf{F}) = 0$. Considering this cubic constraint and the fact that a fundamental matrix is defined up to an arbitrary scale, its degrees-of-freedom is reduced to seven. Thus seven point pairs are enough for the estimation.

As the essential matrix encapsulates the full camera motion, the orientation and direction of the translation, it has five degrees-of-freedom. The two additional constraints are described by the well-known trace constraint [19] as

$$2\mathbf{E}\mathbf{E}^T\mathbf{E} - \text{tr}(\mathbf{E}\mathbf{E}^T)\mathbf{E} = 0. \tag{3}$$

Even though Eq. 3 yields nine polynomial equations for $\mathbf{E}$, only two of them are algebraically independent.

Semi-calibrated case is assumed in this paper as only the common focal-length f is considered to be unknown. Without loss of generality, the intrinsic camera matrix is $\mathbf{K} = \mathbf{K}^T = \text{diag}(f, f, 1)$, where f is the unknown focal-length. In order to replace $\mathbf{E}$ with $\mathbf{F}$ in Eq. 3 we define matrix $\mathbf{Q}$ as follows:

$$\mathbf{Q} = \text{diag}\left(1, 1, \tau\right), \quad \tau = f^{-2}. \tag{4}$$

Due to the fact that $\mathbf{K}$ is non-singular, and trace($\mathbf{E}\mathbf{E}^T$) identifies a scalar value, Eq. 3 can be simplified by multiplying with $\mathbf{K}^{-T}$ and $\mathbf{K}^{-1}$ from the left and the right sides, respectively. Moreover, trace is invariant under cyclic permutations. As a consequence, Eq. 3 is written as [17, 27]

$$2\mathbf{F}\mathbf{Q}\mathbf{F}^T\mathbf{Q}\mathbf{F} - \text{tr}(\mathbf{F}\mathbf{Q}\mathbf{F}^T\mathbf{Q})\mathbf{F} = 0. \tag{5}$$

This relationship will help us to recover the focal length and the fundamental matrix using two affine correspondences.

An affine correspondence $(\mathbf{p}_1, \mathbf{p}_2, \mathbf{A})$ consists of a corresponding point pair and the related local affinity $\mathbf{A}$ transforming the vicinity of point $\mathbf{p}_1$ to that of $\mathbf{p}_2$. In the rest of the paper, $\mathbf{A}$ is considered as its left 2×2 submatrix

$$\mathbf{A} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$$

since the third column – the translation part – is determined by the point locations.

We use the **hidden variable technique** in the proposed method. It is a resultant technique in algebraic geometry for the elimination of variables from a multivariate polynomial system [9]. Suppose that m polynomial equations in n variables are given. In brief, one can assume an unknown variable as a parameter and rewrite the equation system as $\mathbf{C}(y_1)\mathbf{x} = 0$, where $\mathbf{C}$ is a coefficient matrix depending on the unknown y_1 (hidden variable) and vector $\mathbf{x}$ is the vector of $n-1$ unknowns. If the number of equations equals to that of the unknown monomials in $\mathbf{x}$, i.e. matrix $\mathbf{C}$ is square, the non-trivial solution can be carried out as $\det(\mathbf{C}(y_1)) = 0$. Solving the resultant equation for y_1 and back-substituting it, the whole system is solved.

3. Focal-length using Two Correspondences

This section aims the recovery of the unknown focal length and fundamental matrix using two affine correspondences. First, the connection between the fundamental matrix and local affinity is introduced, then we discuss the estimation technique.

3.1. Exploiting a Local Affine Transformation

Suppose that an affine correspondence $(\mathbf{p_1}, \mathbf{p_2}, \mathbf{A})$ and fundamental matrix $\mathbf{F}$ are known. It is trivial that every affine transformation preserves the direction of the lines going through points $\mathbf{p_1}$ and $\mathbf{p_2}$ on the first and second images. As a consequence, the link between directions $\mathbf{v_1}$ and $\mathbf{v_2}$ of epipolar lines can be described [3] by affine transformation $\mathbf{A}$ as

$$\mathbf{A}\mathbf{v_1} \parallel \mathbf{v_2}. \tag{6}$$

Reformulating Eq. 6 using the well-known fact from Computer Graphics [33] leads to $\mathbf{A}^{-T}\mathbf{R}^{90}\mathbf{v_1} = \beta\mathbf{R}^{90}\mathbf{v_2}$, where matrix $\mathbf{R}^{90}$ is a 2D orthonormal (rotation) matrix rotating with 90 degrees and β is an unknown scale. Vectors $\mathbf{R}^{90}\mathbf{v_1}$ and $\mathbf{R}^{90}\mathbf{v_2}$ are the line normals $\mathbf{n_1}$ and $\mathbf{n_2}$ as

$$\mathbf{A}^{-T}\mathbf{n_1} = \beta\mathbf{n_2}. \tag{7}$$

In Appendix A, it is proven that β is equal to -1 if $\mathbf{n_1}$ and $\mathbf{n_2}$ are calculated from the fundamental matrix using relationships $\mathbf{F}\mathbf{n_1}$ and $\mathbf{F}^T\mathbf{n_2}$ and they are *not normalized*. In brief, it is given as the distance ratio of neighboring epipolar lines on the two images. For the case when the normals are

not normalized – the original scale has not been changed –, β is only a scale inverting the directions.

Normals are expressed from $\mathbf{F}$ as the first two coordinates of the epipolar lines: $\mathbf{n}_1 = (\mathbf{l}_1)_{(1:2)} = (\mathbf{F}^T\mathbf{p}_2)_{(1:2)}$ and $\mathbf{n}_2 = (\mathbf{l}_2)_{(1:2)} = (\mathbf{F}\mathbf{p}_1)_{(1:2)}$ [13], where the lower indices select a subvector. Therefore, Eq. 7 is written as

$$\mathbf{A}^{-T}(\mathbf{F}^T\mathbf{p}_2)_{(1:2)} = -(\mathbf{F}\mathbf{p}_1)_{(1:2)} \tag{8}$$

and forms a system of linear equations consisting of two equations as follows:

$$(u_2 + a_1u_1)f_1 + a_1v_1f_2 + a_1f_3 + (v_2 + a_3u_1)f_4 + a_3v_1f_5 + a_3f_6 + f_7 = 0 \tag{9}$$

$$a_2u_1f_1 + (u_2 + a_2v_1)f_2 + a_2f_3 + a_4u_1f_4 + (v_2 + a_4v_1)f_5 + a_4f_6 + f_8 = 0. \tag{10}$$

Thus each local affine transformation *reduces the degrees-of-freedom by two*.

3.2. Two-point Solver

Suppose that two affine correspondences $(\mathbf{p}_1^1, \mathbf{p}_2^1, \mathbf{A}^1)$ and $(\mathbf{p}_1^2, \mathbf{p}_2^2, \mathbf{A}^2)$ are given. Coefficient matrix

$$\mathbf{C}^i = \begin{bmatrix} u_2 + a_1u_1 & a_1v_1 & a_1 & v_2 + a_3u_1 & a_3v_1 & a_3 & 1 & 0 & 0 \\ a_2u_1 & u_2 + a_2v_1 & a_2 & a_4u_1 & v_2 + a_4v_1 & a_4 & 0 & 1 & 0 \\ u_1u_2 & v_1u_2 & u_2 & u_1v_2 & v_1v_2 & v_2 & u_1 & v_1 & 1 \end{bmatrix}$$

related to the i-th ($i \in \{1, 2\}$) correspondence is formed as the combination of Eqs. 2, 9, 10 and satisfies formula $\mathbf{C}^i\mathbf{x} = 0$, where $\mathbf{x} = \begin{bmatrix} f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7 & f_8 & f_9 \end{bmatrix}^T$ is the vector of unknown elements of the fundamental matrix. We denote the concatenated coefficient matrix of both correspondences as follows:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}^1 \\ \mathbf{C}^2 \end{bmatrix}. \tag{11}$$

It is of size 6×9, therefore, its left null space is three-dimensional. The solution is carried out as

$$\mathbf{x} = \alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c}, \tag{12}$$

where $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are the singular vectors and α, β, γ are unknown non-zero scalar values.

Remember that only the common focal length is unknown from the intrinsic parameters, therefore, we are able to exploit the trace constraint. Eq. 5 yields ten cubic equations for four unknowns α, β, γ and τ, where $\tau = f^{-2}$ encapsulates the unknown focal length. We consider τ as the hidden variable and form coefficient matrix $\mathbf{C}(\tau)$ w.r.t. the other three ones – thus the rows of $\mathbf{C}(\tau)$ are univariate polynomials with variable τ. Even though α, β and γ are defined up to a common scale, we do not fix this scale in order to keep the homogenity of the system. The monomials of this polynomial system are as $\mathbf{y} = [\alpha^3 \;\; \alpha^2\beta \;\; \alpha^2\gamma \;\; \alpha\beta^2 \;\; \alpha\beta\gamma \;\; \alpha\gamma^2 \;\; \beta^3 \;\; \beta^2\gamma \;\; \beta\gamma^2 \;\; \gamma^3]^T$. Table 1 demonstrates the coefficient matrix.

Since the scale of monomial vector $\mathbf{x}$ has not been fixed, the non-trivial solution of equation $\mathbf{C}(\tau)\mathbf{y} = 0$ is when the determinant vanishes as

$$\det(\mathbf{C}(\tau)) = 0. \tag{13}$$

Therefore, the hidden-variable resultant – a polynomial of the hidden variable – is $\det(\mathbf{C}(\tau))$. As the current problem is fairly similar to that of [19], we adopt the proposed algorithm. It is proved that $\det(\mathbf{C}(\tau))$ is actually a 15-th degree polynomial and it obtains the candidate values for τ. Then the solution for α, β, γ and τ is given as $\mathbf{y} = \text{null}(\mathbf{C}(\tau))$. Finally, fundamental matrix $\mathbf{F}$ regarding to each obtained focal length can be directly estimated using Eq. 12.

$\mathbf{C}(\tau)$	1 α^3	2 $\alpha^2\beta$	3 $\alpha^2\gamma$	4 $\alpha\beta^2$	5 $\alpha\beta\gamma$	6 $\alpha\gamma^2$	7 β^3	8 $\beta^2\gamma$	9 $\beta\gamma^2$	10 γ^3
1	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
.	.	.	.	.	.	.	.	.	.	.
10	c_{91}	c_{92}	c_{93}	c_{94}	c_{95}	c_{96}	c_{97}	c_{98}	c_{99}	c_{100}

Table 1: The coefficient matrix $\mathbf{C}(\tau)$ related to the ten polynomial equations of the trace constraint.

4. Elimination and Selection of Roots

In this section, a novel technique is proposed to omit roots on the basis of the underlying geometry. Then we show a heuristics considering the properties of digital cameras to remove invalid focal lengths. In the end, we introduce a root selection algorithm.

4.1. Elimination of Invalid Focal Lengths

A solution is proposed here based on the underlying geometry to eliminate invalid focal lengths. Suppose that a point pair $(\mathbf{p}_1, \mathbf{p}_2)$, the related local affinity $\mathbf{A}$, the fundamental matrix $\mathbf{F}$, and an obtained focal length f are given. As the semi-calibrated case is assumed, $\mathbf{F}$ and f exactly determines the projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ of both cameras [13]. Denote the 3D coordinates and the surface normal induced by point pair $(\mathbf{p}_1, \mathbf{p}_2)$, local affinity $\mathbf{A}$ and the projection matrices with $\mathbf{q} = [x \;\; y \;\; z]^T$ and $\mathbf{n} = [n_x \;\; n_y \;\; n_z]^T$, respectively. According to our experiences, linear triangulation [13] is a suitable and efficient choice to estimate $\mathbf{q}$. Surface normal $\mathbf{n}$ is estimated exploiting affinity $\mathbf{A}$ by the method proposed in [5].[1]

Without loss of generality, we assume that a point of a 3D surface cannot be observed from behind. As a consequence, the angle between vectors $\mathbf{c}_i - \mathbf{q}$ and $\mathbf{n}$ must be smaller than $90°$ for both cameras, where $\mathbf{c}_i$ is the position of the i-th camera ($i \in \{1, 2\}$). This can be interpreted as follows: each camera selects a half unit-sphere around the

[1] http://web.eee.sztaki.hu/~dbarath/

observed point $\mathbf{q}$. Surface normal $\mathbf{n}$ must lie in the intersection of these half spheres. These half spheres are described by a rectangle in the spherical coordinate system as follows: $\text{rect}_i = \begin{bmatrix}\theta_i - \frac{\pi}{2} & \sigma_i - \frac{\pi}{4} & \pi & \frac{\pi}{2}\end{bmatrix}$, where θ_i, σ_i are the corresponding spherical coordinates and rect_i is of format $\begin{bmatrix}\text{corner}_\theta & \text{corner}_\sigma & \text{width} & \text{height}\end{bmatrix}$. The intersection area induced by the two cameras is as

$$\text{rect}_\cap = \bigcap_{i \in [1,2]} \text{rect}_i.$$

Point $\mathbf{q}$ is observable from both cameras *if and only if* surface normal $\mathbf{n}$, represented by spherical coordinates Θ and Σ, lies in the intersection area: $\begin{bmatrix}\Theta & \Sigma\end{bmatrix} \in \text{rect}_\cap$. A setup, induced by focal length f, not satisfying this criteria is an invalid one and can be omitted. Note that this constraint can be straightforwardly extended to the multi-view case making the intersection area more restrictive.

4.2. Physical Properties of Cameras

We introduce restrictions on the estimated roots considering the physical limits of the cameras. The focal length within camera matrix $\mathbf{K}$ is not equivalent to the focal length of the lenses, since it is the ratio of the optical focal length and the pixel size [13]. Particularly, the latter one is a few micrometers, while the optical focal length are within interval $[1 \dots 500]$ mm. Therefore, coarse lower and upper limits for a realistic camera are 100 and 500.000. Focal lengths out of this interval are automatically discarded. Note that these limits can be easily changed considering cameras with different properties.

4.3. Root Selection

To resolve the ambiguity of multiple roots and to minimize the effect of the noise, the classical way is to exploit multiple measurements eliminating the inconsistent ones. Since Eq. 13 is a high-degree polynomial it is sensitive to noise – small changes in the coordinates and affine elements cause significantly different coefficients.

RANSAC [10] is a successful technique for that problem, e.g. in the five-point relative-orientation one [25]. Recent methods, i.e. Kernel Voting, exploit the property that the roots form a peak around the real solution [20, 19, 18]. Kernel Voting maximizes a kernel density function like a maximum-likelihood-decision-maker. To our experiences, this technique works accurately if the noise in the coordinates does not exceed $1-2$ pixels on average. Over that, the roots may form several strongly supported peaks and it is not guaranteed that the true solution is found.

Thus we formulate the problem as a mode-seeking in a one dimensional domain: the real focal length appears as the most supported mode. Among several mode-seeking techniques [14] the most robust one is the Median-Shift [29] according to extensive experimentation. Median-Shift providing Tukey-medians [32] as modes does not generate new elements in the domain it is applied to. In particular, there is no significant difference in the results of Tukey- [32] and Weiszfeld-medians [34], however, the former one is slightly faster to compute. Finally, in order to overcome the discrete nature of Median-Shift – since it does not add new instances, only operates with the given ones –, we apply a gradient descent from the retrieved mode x_0 maximizing function

$$f(x) = \sum_{i=1}^{n} \frac{\kappa(x_i - x)}{h}, \tag{14}$$

where n is the number of focal lengths, κ is a kernel function – we chose Gaussian-kernel –, x_i is the i-th focal length, and h is a bandwidth same as for the Median-Shift.

5. Experimental Results

For the synthesized tests, we used the MATLAB code shown in Alg. 1. For the real world tests, we used our C++ implementation[2] which is a modification of the solver of Hartley et al. [12].

5.1. Synthesized tests

For synthesized testing, two perspective cameras are generated by their projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$. The first camera is at position $[0\ 0\ 1]^T$ looking towards the origin, and the distance of the second one from the first is 0.15 in a random direction. Five random planes passing over the origin are generated and each is sampled in fifty random locations. The obtained 3D points are projected onto the cameras. Zero-mean Gaussian-noise is added to the point coordinates. The local affine transformations are calculated by derivating the homographies induced by the tangent planes at the noisy point correspondences similarly to [2].

Figure 1 reports the kernel density function with Gaussian-kernel width 10 plotted as the function of the relative error (in percentage). Candidate focal lengths are estimated as follows:

1. Select two affine correspondences.
2. Apply the proposed 2-point method.
3. Repeat from Step 1.

The iteration limit is chosen to 100. The blue horizontal line reports the result of Median-Shift, the green one is that of Kernel Voting. The σ value of the zero-mean Gaussian-noise added to the point locations and affinities is (a) 0.01 pixels, (b) 0.1 pixels, (c) 1.0 pixels, (d) 3.0 pixels, (e) 3.0 pixels and there are 10% outliers, (f) 1.0 pixels with some errors in the aspect ratio: the true one is 1.00 but 0.95 is used. The real focal length is 600.

Confirming the validity of the proposed theory, the peak is over the ground truth focal length: 0% relative error. The

[2] http://web.eee.sztaki.hu/~dbarath/

Table 2: Mean (Avg) and median (Med) relative error (in percentage) and the spread (σ) of the relative errors in the estimated focal lengths on the 104 real image pairs. Corr # denotes the required correspondence number.

Method	Corr #	Avg	Med	σ
Proposed	2	**9.62**	**3.88**	**14.08**
Perdoch et al. [26]	2	44.66	45.89	26.43
Hartley et al. [12]	6	21.79	8.61	27.48

proposed root selection is more robust than the Kernel Voting approach since the blue line is closer to the zero relative error even if the noise is high.

Fig. 2 reports the mean (top) and median (bottom) errors of the estimated fundamental matrices plotted as the function of the noise σ and compared with the results of Hartley et al.[12] and Perdoch et al.[26]. The error is the Frobenious norm of the estimated and ground truth fundamental matrices. 100 runs were performed on each noise level. It can be seen that the accuracy of the estimated fundamental matrices is similar to that of Hartley et al. [12].

5.2. Tests on Real Images[3]

To test the proposed method on real world photos, 104 image pairs were downloaded[4] each containing the ground truth focal length in the EXIF data (see Fig. 4 for examples). Affine correspondences are detected by ASIFT [23] and the same procedure is applied as for the synthesized tests. Fig. 3a reports the histogram of the relative errors (in percentage) in the focal length estimates on all the 104 pairs. It can be seen that in most of the cases the obtained results are accurate, the relative error is close to zero. Fig. 3b shows the first image of an example pair and the point correspondences.

In Table 2, the proposed method is compared with the 6-point algorithm [12] and the one creating point correspondences from two local affinities [26]. The reported relative errors are computed as the ratio of the estimation error and the ground truth focal length as $|f_{est} - f_{gt}|/f_{gt}$. It can be seen that the 2-point technique outperforms the other ones in terms of both mean and median accuracy and spread.

5.3. Time Demand

Augmenting RANSAC or other robust statistics with the proposed method significantly reduces the processing time. Table 3 reports the required iteration number [13] of RANSAC to converge using different minimal methods (columns) as engine. Rows show the ratio of the outliers.

[3]Test data are provided as supplemental material.

[4]http://www2c.airnet.ne.jp/kawa/photo/ste-idxe.htm

Table 3: Required iteration number of RANSAC augmented with minimal methods (columns) with 95% probability on different outlier levels (rows).

	# of required points				
Outl.	**2**	5	6	7	8
50%	**11**	95	191	383	766
80%	**74**	$\sim 10^3$	$\sim 10^4$	$\sim 10^5$	$\sim 10^6$

6. Conclusion

A theory and an efficient method is proposed to estimate the unknown focal-length and the fundamental matrix using only two affine correspondences. The 2-point method is validated on both synthesized and real world data. Compared with the state-of-the-art methods, it obtained the most accurate focal lengths with fundamental matrices having similar quality as the recent algorithms. Combining the minimal solver with a robust statistics, e.g. RANSAC, allows significant reduction in computation. Particularly, its time demand is around a few milliseconds, thus it is much faster than affine-covariant detectors providing the input.

The proposed algorithm can also be applied in reconstruction or multi-view pipelines, e.g. that of Bujnak et al. [8], if at least two images of the same camera with fixed focal length are available.

A. Proof of the Linear Affine Constraints

Lemma 1 (Constraints on the Normals of Epipolar Lines)**.** *Given a local affine transformation* $\mathbf{A}$ *transforming the infinitely close vicinities of the related point pair. The normals of the corresponding epipolar lines are* $\mathbf{n}_1$ *and* $\mathbf{n}_2$*. Matrix* $\mathbf{A}$ *is a valid local affinity if and only if* $\mathbf{A}^{-T}\mathbf{n}_1 = -\mathbf{n}_2$.

Proof. It is trivial that affinity $\mathbf{A}$ transforms the direction of the corresponding epipolar lines to each other as $\mathbf{A}\mathbf{v} \parallel \mathbf{v}'$, where $\mathbf{v}$ and $\mathbf{v}'$ are the directions of the lines on the two images. It is well-known from Computer Graphics [33] that this is equivalent to $\mathbf{A}^{-T}\mathbf{n} = \beta\mathbf{n}'$, where $\mathbf{n} = (\mathbf{F}^T\mathbf{p}')_{1:2}$ and $\mathbf{n}' = (\mathbf{F}\mathbf{p})_{1:2}$ are the normals of the epipolar lines ($\beta \neq 0$). Note that lower index $(1:2)$ denotes the first two elements of a vector. We prove here that

$$\mathbf{A}^{-T}\mathbf{n} = -\mathbf{n}'. \tag{15}$$

(Proof) Given a corresponding point pair $\mathbf{p} = [x, y, 1]^T$ and $\mathbf{p}' = [x', y', 1]^T$. Let $\mathbf{n}_1 = [\mathbf{n}_{1,x} \quad \mathbf{n}_{1,y}]^T$ and $\mathbf{n}'_1 = [\mathbf{n}'_{1,x} \quad \mathbf{n}'_{1,y}]^T$ be the normal directions of epipolar lines $\mathbf{l}_1 = \mathbf{F}^T\mathbf{p}' = [\mathbf{l}_{1,a} \quad \mathbf{l}_{1,b} \quad \mathbf{l}_{1,c}]^T$ and $\mathbf{l}'_1 = \mathbf{F}\mathbf{p} = [\mathbf{l}'_{1,a} \quad \mathbf{l}'_{1,b} \quad \mathbf{l}'_{1,c}]^T$. Then it is trivial that $\mathbf{A}^{-T}\mathbf{n}_1 = \beta\mathbf{n}'_1$ due to $\mathbf{A}\mathbf{v} \parallel \mathbf{v}'$, where β is a scale factor.

First, the task is to determine how affinity $\mathbf{A}$ transforms the length of $\mathbf{n}_1$ if $|\mathbf{n}_1| = |\mathbf{n}'_1| = 1$. Introduce point $\mathbf{q} =$

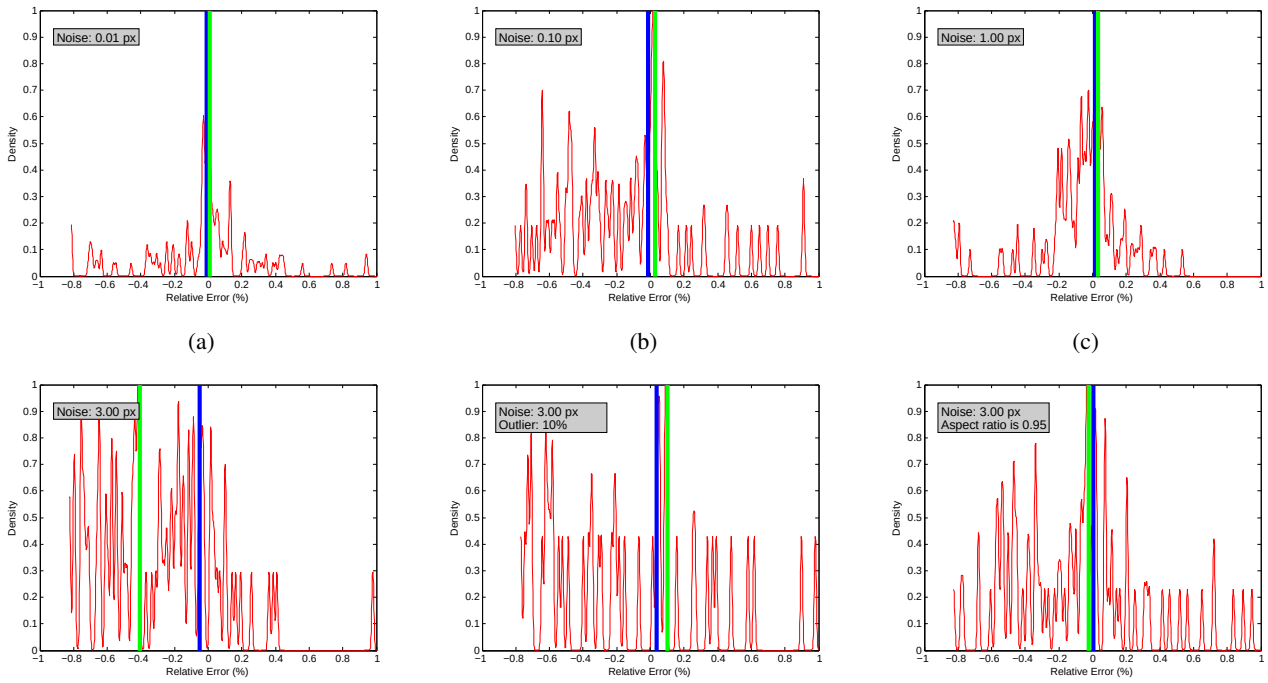


Figure 1: The kernel density function (vertical axis) with Gaussian-kernel width 10 plotted as the function of the relative error (%). Five planes are generated and each is sampled in 20 locations – points are projected onto the cameras and local affinities are calculated. The blue horizontal line is the result of Median-Shift, the green one is that of the Kernel Voting. The σ value of the zero-mean Gaussian-noise added to the point locations and affinities is (a) 0.01 pixels, (b) 0.1 pixels, (c) 1.0 pixels, (d) 3.0 pixels, (e) 3.0 pixels and there are 10% outliers, (f) 1.0 pixels with some errors in the aspect ratio: the true one is 1.00 but 0.95 is used. Ground truth focal length is 600. Best viewed in color.

$\mathbf{p} + \delta\mathbf{n}_1$, where δ is an arbitrary scalar value. This new point determines an epipolar line on the second image as $\mathbf{l}'_2 = \mathbf{F}\mathbf{q} = \mathbf{F}(\mathbf{p} + \delta\mathbf{n}_1) = [\mathbf{l}'_{2,a} \quad \mathbf{l}'_{2,b} \quad \mathbf{l}'_{2,c}]^T$. Scale β is given by distance d' between line $\mathbf{l}'_2$ and point $\mathbf{p}'$ (see Fig. 5a). The calculation of distance d' is written as follows:

$$\begin{aligned} d' &= \frac{|s_{1,a}x' + s_{2,b}y' + s_{3,c}|}{\sqrt{s_{1,a}^2 + s_{2,b}^2}}, \\ s_{i,k} &= \mathbf{l}'_{1,k} + \delta f_{i1}\mathbf{n}_{1,x} + \delta f_{i2}\mathbf{n}_{1,y}, \\ i &\in \{1,2,3\}, k \in \{a,b,c\} \end{aligned} \tag{16}$$

Point $\mathbf{p}'$ lies on $\mathbf{l}'_1$, which can be written as $\mathbf{l}'_{1,a}x' + \mathbf{l}'_{1,b}y' + \mathbf{l}'_{1,c} = 0$. This fact reduces Eq. 16 to

$$d' = \frac{|\hat{s}_1 u' + \hat{s}_2 v^2 + \hat{s}_3|}{\sqrt{s_1^2 + s_2^2}}, \tag{17}$$

where $\hat{s}_i = \delta f_{i1}\mathbf{n}_{1,x} + \delta f_{i2}\mathbf{n}_{1,y}, \quad i \in \{1,2,3\}$. To determine β, the introduced point $\mathbf{q}$ has to be moved infinitely close to $\mathbf{p}$ ($\delta \rightarrow 0$). The square of β is then written as $\beta^2 = \lim_{\delta \rightarrow 0} \frac{\delta^2}{d'^2} = \lim_{\delta \rightarrow 0} \frac{s_1^2 + s_2^2}{|\hat{s}_1 u' + \hat{s}_2 v' + \hat{s}_3|^2}$. After elementary modifications, the formula for scale β is $\beta = \sqrt{\mathbf{l}'_{1,a}\mathbf{l}'_{1,a} + \mathbf{l}'_{1,b}\mathbf{l}'_{1,b}} / \left(|\widetilde{s}_1 x' + \widetilde{s}_2 y' + \widetilde{s}_3|\right)$, where $\widetilde{s}_i = f_{i1}\mathbf{n}_{1,x} + f_{i2}\mathbf{n}_{1,y}, \; i \in \{1,2,3\}$. Therefore, *we can calculate β for unit length normals.*

Consider the case when normals are kept in their original form and not normalized ($|\mathbf{n}_1| \neq |\mathbf{n}'_1| \neq 1$). The normalization indicates the following formula

$$\mathbf{A}^{-T}\frac{\mathbf{n}}{|\mathbf{n}|} = \beta\mathbf{n}'. \tag{18}$$

The epipolar line corresponding to point $\mathbf{p}$ is parameterized as $[\mathbf{l}'_{1,a}, \mathbf{l}'_{1,b}, \mathbf{l}'_{1,c}] = \mathbf{F}[x, y, 1]^T$. Therefore, its normal is as follows: $\mathbf{n}' = \begin{bmatrix}\mathbf{l}'_{1,a} & \mathbf{l}'_{1,b}\end{bmatrix}^T = (\mathbf{F}\begin{bmatrix}x' & y' & 1\end{bmatrix}^T)_{(1:2)}$. Similarly, $\mathbf{n} = (\mathbf{F}^T\begin{bmatrix}x' & y' & 1\end{bmatrix}^T)_{(1:2)}$. The denominator in Eq. 18 for computing β is rewritten as $|\mathbf{n}| = \sqrt{\mathbf{l}_{1,a}^2 + \mathbf{l}_{1,b}^2}$. The numerator is as follows:

$$\begin{aligned} &\widetilde{s}_1 u' + \widetilde{s}_2 v' + \widetilde{s}_3 = \\ &\mathbf{n}_{1,u}(f_{11}u' + f_{21}v' + f_{31}) + \mathbf{n}_{1,v}(f_{12}u' + f_{22}v' + f_{32}) = \\ &\mathbf{n}_{1,u}^2 + \mathbf{n}_{1,v}^2 = |\mathbf{n}_1|^2. \end{aligned}$$

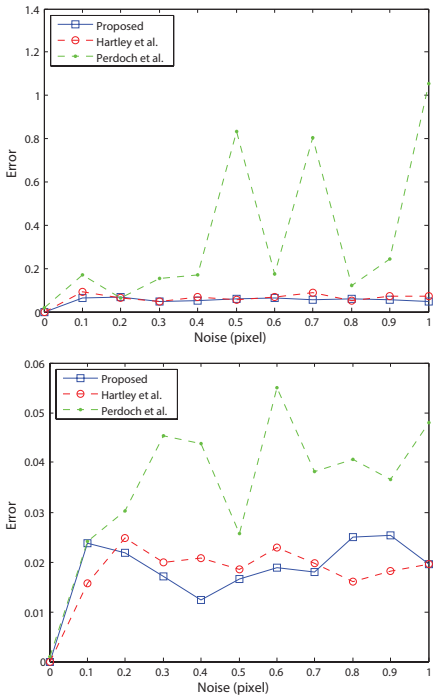
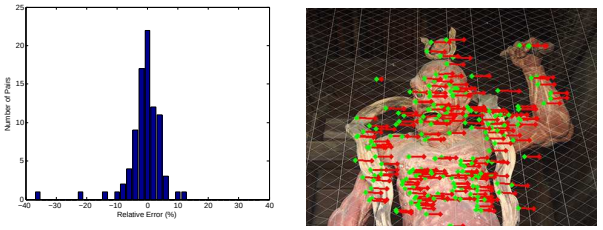


Figure 2: The mean (top) and median (bottom) Frobenious norms of the estimated and the ground truth fundamental matrices plotted as the function of the noise σ. 100 runs on each noise level were performed.



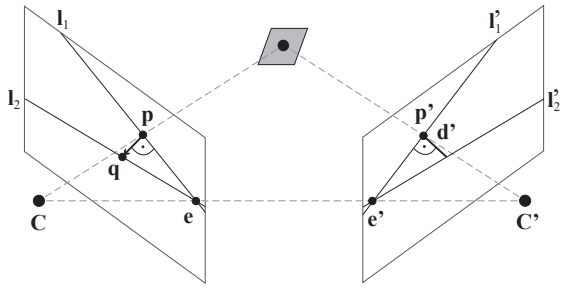
(a) (b)

Figure 3: (a) Histogram of focal length estimation on 104 image pairs. The horizontal axis is the number of the pairs plotted as the function of the relative error (%, vertical axis) in the focal length. (b) The first image of an example pair. Point coordinates on the first image (green dots), on the second one (red dots) and the point movements (red lines).

Thus $\beta = \pm|\mathbf{n}_1|/|\mathbf{n}_1|^2 = \pm 1/|\mathbf{n}_1|$. Therefore, Eq. 18 is modified to $\mathbf{A}^{-T}\mathbf{n} = \pm\mathbf{n}'$.

Since the direction of the epipolar lines on the two images must be the opposite of each other, the positive solution is omitted. The final formula is: $\mathbf{A}^{-T}\mathbf{n} = -\mathbf{n}'$. □

Figure 4: The first images of example pairs. Point coordinates on the first image (green dots), on the second one (red dots) and the point movements (red lines). The ground truth focal lengths, the results of the 6-point [12] and the proposed methods are written in gray rectangle.



(a) The scale between neighboring epipolar lines.

Figure 5: Two projections of a patch. The constraint for scale states that the ratio of $|p - q|$ and d' determines the scale between vectors $\mathbf{A}^{-T}\mathbf{n}$ and $\mathbf{n}'$.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Commun. ACM*, 54(10):105–112, 2011. 1

[2] D. Barath and L. Hajder. Novel ways to estimate homography from local affine transformations. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages

Program 1: The Two-point Algorithm

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% 2-pt focal length algorithm. Use Matlab-7.0(6.5) with SymbolicMath Toolbox.
%% Input: The "Matches" is a 2x8 matrix containing two affine correspondences.
%%        Each row of "Matches": (u1, v1, u2, v2, a1, a2, a3, a4).
%%        Example (the ground truth focal length is 600):
%% Matches = [12.0527 134.0870 -263.1743 679.7212 1.6376 -0.3952 -0.1925 2.2532;
%%         -67.9281 -42.4639 -313.5657 362.3455 1.3758 -0.3845 0.0150 1.4806]
%% Output: focal lengths.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function F = TwoPointFocalLength(Matches)
    syms F f x y z w equ Res Q C
    equ = sym('equ', [1 10]);
    C = sym('C', [10 10]);
    Q = w^(-1) * [1, 0, 0; 0, 1, 0; 0, 0, w];
    M = zeros(size(pts1,1), 9);

    for i = 1 : size(pts1,1)
        u1 = Matches(i,1); v1 = Matches(i,2); u2 = Matches(i,3); v2 = Matches(i,4);
        a1 = Matches(i,5); a2 = Matches(i,6); a3 = Matches(i,7); a4 = Matches(i,8);

        M(3*i + 0,:) = [u1 * u2, v1 * u2, u2, u1 * v2, v1 * v2, v2, u1, v1, 1];
        M(3*i + 1,:) = [u2 + a1 * u1, a1 * v1, a1, v2 + a3 * u1, a3 * v1, a3, 1, 0, 0];
        M(3*i + 2,:) = [a2 * u1, u2 + a2 * v1, a2, a4 * u1, v2 + a4 * v1, a4, 0, 1, 0];
    end;
    [~,~,vm] = svd(M,0);
    N = [vm(:,7), vm(:,8), vm(:,9)];

    f = x*N(:,1) + y*N(:,2) + z*N(:,3);
    F = transpose(reshape(f,3,3)); FT = transpose(F);
    tr = sum(diag(F*Q*FT*Q));

    equ(1) = det(F);
    equ(2:10) = expand(2*F*Q*FT*Q*F-tr*F);
    for i = 1:10
        equ(i) = maple('collect', equ(i), '[x,y,z]', 'distributed');
        for j = 1 : 10
            oper = maple('op', j, equ(i)); C(i,j) = maple('op', 1, oper);
        end
    end

    Res = maple('evalf', det(C)); %%Hidden-variable resultant
    foc = 1.0 ./ sqrt(double([solve(Res)]));
    foc = foc(imag(foc) == 0);
end
```


434–445, 2016. 4

[3] D. Barath, J. Matas, and L. Hajder. Accurate closed-form estimation of local affine transformations consistent with the epipolar geometry. In *British Machine Vision Conference*, 2016. 1, 2

[4] D. Barath, J. Molnar, and L. Hajder. Novel methods for estimating surface normals from affine transformations. In *Computer Vision, Imaging and Computer Graphics Theory and Applications (Selected and Revised Papers)*, pages 316–337. 2015. 1

[5] D. Barath, J. Molnar, and L. Hajder. Optimal Surface Normal from Affine Transformation. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 305–316, 2015. 1, 3

[6] J. Bentolila and J. M. Francos. Conic epipolar constraints from affine correspondences. *Computer Vision and Image Understanding*, 122:105–114, 2014. 1

[7] A. Bódis-Szomorú, H. Riemenschneider, and L. V. Gool. Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In *CVPR*, 2014. 1

[8] M. Bujnak, Z. Kukelova, and T. Pajdla. Robust focal length estimation by voting in multi-view scene reconstruction. *Computer Vision–ACCV 2009*, pages 13–24, 2010. 5

[9] D. A. Cox, J. Little, and D. O'shea. *Using algebraic geometry*. 2006. 1, 2

[10] M. Fischler and R. Bolles. RANdom SAmpling Consensus: a paradigm for model fitting with application to image anal-

ysis and automated cartography. *Commun. Assoc. Comp. Mach.*, 1981. 1, 4

[11] J. Frahm, P. F. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y. Jen, E. Dunn, B. Clipp, and S. Lazebnik. Building rome on a cloudless day. In *11th European Conference on Computer Vision*, pages 368–381, 2010. 1

[12] R. I. Hartley and H. Li. An efficient hidden variable approach to minimal-case camera motion estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(12):2303–2314, 2012. 4, 5, 7

[13] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 1, 2, 3, 4, 5

[14] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, 1999. 4

[15] K. Köser. *Geometric Estimation with Local Affine Frames and Free-form Surfaces*. Shaker, 2009. 1

[16] K. Köser and R. Koch. Differential spatial resection - pose estimation using a single local image feature. In *IEEE Proceedings of the European Conference on Computer Vision*, 2008. 1

[17] Z. Kukelova, M. Bujnak, and T. Pajdla. Polynomial eigenvalue solutions to the 5-pt and 6-pt relative pose problems. In *Proceedings of the British Machine Vision Conference*, 2008. 2

[18] Z. Kukelova, T. Pajdla, and M. Bujnak. *Algebraic methods in computer vision*. PhD thesis, Center for Machine Perception, Czech Technical University, Prague, Czech republic, 2012. 4

[19] H. Li. A simple solution to the six-point two-view focal-length problem. In *IEEE Proceedings of the European Conference on Computer Vision*, 2006. 1, 2, 3, 4

[20] H. Li and R. Hartley. A non-iterative method for correcting lens distortion from nine-point correspondences. In *In Proc. OmniVision05, ICCV-workshop*, 2005. 4

[21] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *IEEE Proceedings of the European Conference on Computer Vision*, 2002. 1

[22] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IEEE Proceedings of the International Journal of Computer Vision*, 2005. 1

[23] J. Morel and G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM J. Imaging Sciences*, 2(2):438–469, 2009. 1, 5

[24] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *International Conference on Computer Vision, ICCV 2013*, pages 3248–3255, 2013. 1

[25] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–777, 2004. 1, 4

[26] M. Perdoch, J. Matas, and O. Chum. Epipolar geometry from two correspondences. In *ICPR*, 2006. 5

[27] Á. Pernek and L. Hajder. Automatic focal length estimation as an eigenvalue problem. *Pattern Recognition Letters*, 34(9):1108–1117, 2013. 2

[28] C. Raposo and J. P. Barreto. Theory and practice of structure-from-motion using affine correspondences. In *IEEE Proceedings on Computer Vision and Pattern Recognition*, 2016. 1

[29] L. Shapira, S. Avidan, and A. Shamir. Mode-detection via median-shift. In *IEEE Proceedings of the International Conference on Computer Vision*, 2009. 4

[30] H. Stewénius, D. Nistér, F. Kahl, and F. Schaffalitzky. A minimal solution for relative pose with unknown focal length. *Image and Vision Computing*, 2008. 1

[31] A. Torii, Z. Kukelova, M. Bujnak, and T. Pajdla. The six point algorithm revisited. In *IEEE Proceedings of the Asian Conference on Computer Vision*, 2010. 1

[32] J. W. Tukey. Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians*, 2:523–531, 1975. 4

[33] K. Turkowski. Transformations of surface normal vectors. In *Technical Report 22, Apple Computer*, 1990. 2, 5

[34] E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 1937. 4