Title: Mathematical Foundations of Photogrammetry

Name: Konrad Schindler

Affil./Addr.: Photogrammetry and Remote Sensing, ETH Zürich

schindler@geod.baug.ethz.ch

Mathematical Foundations of Photogrammetry

Introduction

The goal of photogrammetry is to obtain information about the physical environment from images. This chapter is dedicated to the mathematical relations that allow one to extract *geometric 3D measurements* from 2D perspective images. [1] Its aim is to give a brief and gentle overview for students or researchers in neighbouring disciplines. For a more extensive treatment the reader is referred to textbooks such as (Hartley and Zisserman, 2004; McGlone, 2013).

The basic principle of measurement with photographic cameras—and many other optical instrument—is simple: light travels along (approximately) straight rays; these rays are recorded by the camera, thus the sensor measures *directions* in 3D space. The fundamental geometric relation of photogrammetry is thus a simple *collinearity constraint*: a 3D world point, its image in the camera, and the cameras projection center must all lie on a straight line. The following discussion is restricted to the most common type of camera, the so-called perspective camera, which has a single center of projection and captures light on a flat sensor plane. It should however be pointed out

[1] Beyond geometric measurement, photogrammetry also includes the semantic interpretation of images and the derivation of physical object parameters from the observed radiometric intensities. The methodological basis for these tasks is a lot broader and less coherent, and is not treated here.

that the model is valid for all cameras with a single center of projection (appropriately replacing only the mapping from image points to rays in space), and extensions exist to non-central projections along straight rays (e.g. Pajdla, 2002).

In a physical camera the light-sensitive sensor plane is located behind the projection center and the image is captured upside down (the "upside-down configuration"). However, there exists a symmetrical setup with the image plane located in front of the camera, as in a slide projector (the "upright configuration"), for which the resulting image is geometrically identical. For convenience the latter configuration is used here, with the image plane between object and camera.

Preliminaries and Notation

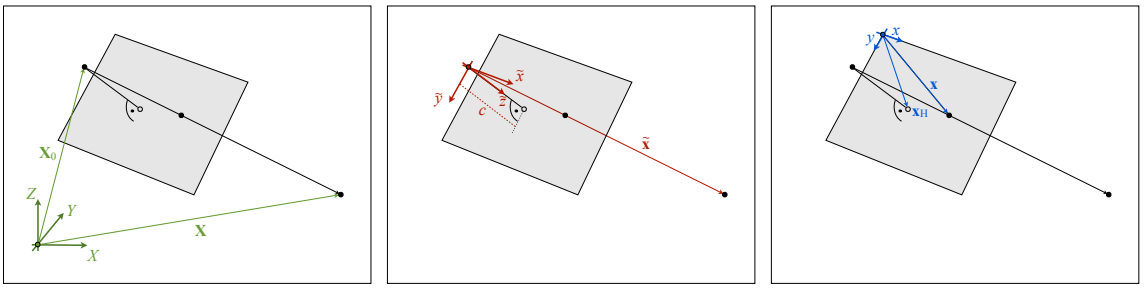
To understand the material in this chapter, the reader should possess basic knowledge of engineering mathematics (linear algebra and calculus) and should be familiar with the basic notions of homogeneous coordinates and projective geometry, found in textbooks such as (Semple and Kneebone, 1952).

In terms of notation, scalars will be denoted in italic font x, vectors in bold font $\mathbf{x}$ and matrices in sanserif font X. The symbol I is reserved for an identity matrix of size 3×3, and $\mathbf{0}$ denotes a 3-vector of zeros.

All coordinate systems are defined as right-handed. Three coordinate systems are required to describe a perspective camera (see Fig. 1):

- the 3D object coordinate system;
- the camera coordinate system, which is another 3D coordinate system, attached to the camera such that its origin lies at the projection center and the sensor plane is parallel to its xy-plane and displaced in positive z-direction;
- the 2D image coordinate system in the sensor plane; its origin lies at the upper left corner of the image, and its x- and y-axis are parallel to those of the camera

coordinate system; for digital cameras it is convenient to also align the x-axis with the rows of the sensor array.



(a) object coordinates (b) camera coordinates (c) image coordinates

Fig. 1. Coordinate systems: $\mathbf{X}$ are the 3D object coordinates of a point; the 3D camera coordinates of the same point are $\tilde{\mathbf{x}}$; the 2D image coordinates of its projection are $\mathbf{x}$.

To distinguish these coordinate systems, uppercase letters $\mathbf{X}$ refer to object coordinates, lowercase letters with tilde $\tilde{\mathbf{x}}$ to camera coordinates, and plain lowercase letters $\mathbf{x}$ to image coordinates.

Transposition of a vector or matrix is written $\mathsf{X}^\top$, and the cross-product between two 3-vectors is denoted either as $\mathbf{x} \times \mathbf{y}$, or using the cross-product matrix $[\mathbf{x}]_\times \mathbf{y}$, where for $\mathbf{x} = [u, v, w]^\top$

$$[\mathbf{x}]_\times = \begin{bmatrix} 0 & -w & v \\ w & 0 & -u \\ -v & u & 0 \end{bmatrix} .$$

The Kronecker product between two vectors or matrices is denoted $\mathsf{X} \otimes \mathsf{Y}$, such that

$$\mathsf{X} \otimes \mathsf{Y} = \begin{bmatrix} x_{11}\mathsf{Y} & x_{12}\mathsf{Y} & \dots & x_{1n}\mathsf{Y} \\ x_{21}\mathsf{Y} & x_{22}\mathsf{Y} & \dots & x_{2n}\mathsf{Y} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}\mathsf{Y} & x_{m2}\mathsf{Y} & \dots & x_{mn}\mathsf{Y} \end{bmatrix} . \tag{1}$$

Some further matrix operators are required: the determinant $\det(\mathsf{X})$, the vectorization $\mathrm{vec}(\mathsf{X}) = \mathbf{x} = [X_{11}, X_{12}, \dots, X_{mn}]^\top$ and the diagonal matrix (here for a 3-dimensional example)

$$\mathrm{diag}(a,b,c) = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix} .$$

Geometric entities are usually assumed to be given in homogeneous coordinates, so a vector $\mathbf{X} = [U, V, W, T]^\top$ refers to a 3D object point with Euclidean coordinates $\frac{1}{T}[U, V, W]^\top$, where $T \neq 0$ is the projective scale.[2] Similarly, a 2D image point is $\mathbf{x} = [u, v, t]^\top$. If Euclidean vectors are needed they are denoted with a superscript e, so for example the Euclidean image point is $\mathbf{x}^e = [x, y]^\top = \frac{1}{t}[u, v]^\top$. Altough stochastic uncertainty modeling is not covered in this chapter, it should be noted that variance propagation is equally possible in homogeneous notation (Förstner, 2010).

Single-view geometry

The collinearity equation

The mapping with an ideal perspective camera can be decomposed into two steps, namely

1. a transformation from object coordinates to camera coordinates, referred to as *exterior orientation*, and
2. a projection from camera coordinates to image coordinates with the help of the cameras *interior orientation*.

The *exterior orientation* is achieved by a translation from the object coordinate origin to the origin of the camera coordinate system (i.e. the projection center), followed by a rotation which aligns the axes of the two coordinate systems. With the Euclidean

[2] Historically, much of the mathematics of photogrammetry was developped in Euclidean notation, and that form is still used in several textbooks. The projective formulation has found widespread use since $\approx$2000.

object coordinates $\mathbf{X}_0^e = [X_0, Y_0, Z_0]^\top$ of the projection center and the 3×3 rotation matrix R this reads as

$$\tilde{\mathbf{x}} = \mathsf{M}\mathbf{X} = \begin{bmatrix} \mathsf{R} & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathsf{I} & -\mathbf{X}_\mathbf{0}^\mathbf{e} \\ \mathbf{0}^\top & 1 \end{bmatrix} \mathbf{X} \,. \tag{2}$$

Let us now have a closer look at the camera coordinate system. In this system, the image plane is perpendicular to the z-axis. The z-axis is also called the *principal ray* and intersects the image plane in the *principal point*, which has the camera coordinates $\tilde{\mathbf{x}}_H = \tilde{t} \cdot [0, 0, c, 1]^\top$ and the image coordinates $\mathbf{x} = t \cdot [x_H, y_H, 1]^\top$. The distance c between the projection center and the image plane is the *focal length* (or *camera constant*). The perspective mapping from camera coordinates to image coordinates then reads

$$\mathbf{x} = [\mathsf{K}|\mathbf{0}]\tilde{\mathbf{x}} = \begin{bmatrix} c & 0 & x_H & 0 \\ 0 & c & y_H & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tilde{\mathbf{x}} \,. \tag{3}$$

This relation holds if the image coordinate system has no shear (orthogonal axes, respectively pixel raster) and isotropic scale (same unit along both axes, respectively square pixels). If a shear s and/or a scale difference m are present, they amount to an affine distortion of the image coordinate system, and the camera matrix becomes

$$\mathsf{K} = \begin{bmatrix} c & cs & x_H \\ 0 & c(1+m) & y_H \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

with five parameters for the *interior orientation*.

By concatenating the two steps from object to image coordinates we get the final projection, i.e. the algebraic formulation of the *collinearity constraint* (Das, 1949)

$$\mathbf{x} = \lambda \mathsf{P}\mathbf{X} \propto \mathsf{P}\mathbf{X} = \mathsf{K}\mathsf{R}[\mathsf{I}| - \mathbf{X}_0^e]\mathbf{X} \,. \tag{5}$$

If an object point $\mathbf{X}$ and its image $\mathbf{x}$ are both given at an arbitrary projective scale, they will only satisfy the relation up to a constant factor. To verify the constraint, i.e. check whether $\mathbf{x}$ is the projection of $\mathbf{X}$, one can use the relation

$$\mathbf{x} \times \mathsf{P}\mathbf{X} = [\mathbf{x}]_\times \mathsf{P}\mathbf{X} = \mathbf{0} \; . \tag{6}$$

Note that due to the projective formulation only two of the three rows of this equation are linearly independent.

Given a projection matrix P it is often necessary to extract the interior and exterior orientation parameters. To that end, observe that

$$\mathsf{P} = [\mathsf{M}|\mathbf{m}] = [\mathsf{KR}| - \mathsf{KR}\mathbf{X}_0^e] \; . \tag{7}$$

The translation part of the exterior orientation immediately follows from $\mathbf{X}_0^e = -\mathsf{M}^{-1}\mathbf{m}$. Moreover, the rotation must by definition be an orthonormal matrix, and the calibration must be an upper triangular matrix. Both properties are preserved by matrix inversion, hence the two matrices can be found by QR-decomposition of $\mathsf{M}^{-1} = \mathsf{R}^\top \mathsf{K}^{-1}$ (or, more efficiently, by RQ-decomposition of M).

Non-linear errors

Equation (5) is valid for ideal perspective cameras. Real physical cameras obey the model only approximately, mainly because the light is collected with the help of lenses rather than entering through an ideal, infinitely small projection center ("pinhole"). The light observed on a real cameras sensor did not travel there from the object point along a perfectly straight path, which leads to errors if one uses only the projective model.

In the image of a real camera we cannot measure the ideal image coordinates $\mathbf{x}^e$, but rather the ones displaced by the non-linear image distortion,

$$\check{\mathbf{x}}^e = \mathbf{x}^e + \Delta\mathbf{x}(\mathbf{x}^e, \mathbf{q}) \ , \tag{8}$$

with $\mathbf{q}$ the parameters of the model that describes the distortion. A simple example would be a radially symmetric lens distortion around the principal point,

$$\Delta\mathbf{x} = \frac{\mathbf{x}^e - \mathbf{x}^e_H}{r}(q_2 r^2 + q_4 r^4) \quad , \quad r = \|\mathbf{x}^e - \mathbf{x}^e_H\| \ . \tag{9}$$

Here, the different physical or empirical distortion models are not further discussed. Instead, the focus is on how to compensate the effect when given a distortion model and its parameters $\mathbf{q}$.

The corrections $\Delta\mathbf{x}$ vary across the image, which means that they depend on the (ideal) image coordinates $\mathbf{x}$. This may be represented by the mapping

$$\check{\mathbf{x}} = \mathsf{H}(\mathbf{x})\mathbf{x} = \begin{bmatrix} 1 & 0 & \Delta x(\mathbf{x}, \mathbf{q}) \\ 0 & 1 & \Delta y(\mathbf{x}, \mathbf{q}) \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x} \ . \tag{10}$$

The overall mapping from object points to observable image points, including non-linear distortions, is now

$$\check{\mathbf{x}} = \check{\mathsf{P}}(\mathbf{x})\mathbf{X} = \mathsf{H}(\mathbf{x})\mathsf{P}\mathbf{X} \ . \tag{11}$$

Note that the non-linear distortions $\Delta\mathbf{x}(\mathbf{x}^e, \mathbf{q})$ are a property of the camera, i.e. they are part of the interior orientation, together with the calibration matrix K.

Equation (11) forms the basis for the correction of non-linear distortions. The computation is split into two steps. Going from object point to image point, one first projects the object point to an ideal image point, $\mathbf{x} = \mathsf{P}\mathbf{X}$, and then applies the distortion, $\check{\mathbf{x}} = \mathsf{H}(\mathbf{x})\mathbf{x}$. Note that for practical purposes the (linear) affine distortion parameters s and m of the image coordinate system are often also included in $\mathbf{H}(\mathbf{x})$ rathern than in K.

For photogrammetric operations the inverse relation is needed, i.e. one measures the coordinates $\check{\mathbf{x}}$ and wants to convert them to ideal ones $\mathbf{x}$, in order to use them as inputs for procedures based on collinearity, such as orientation and triangulation. Often it even makes sense to remove the distortion and synthetically generate perspective (straight line preserving) images as a basis for further processing. To correct the measured coordinates

$$\mathbf{x} = \mathsf{H}^{-1}(\mathbf{x})\check{\mathbf{x}} = \begin{bmatrix} 1 & 0 & -\Delta x(\mathbf{x},\mathbf{q}) \\ 0 & 1 & -\Delta y(\mathbf{x},\mathbf{q}) \\ 0 & 0 & 1 \end{bmatrix} \check{\mathbf{x}} \; . \tag{12}$$

one would need to already know the ideal coordinate one is searching for, so as to evaluate $\mathsf{H}(\mathbf{x})$. One thus resorts to an iterative scheme, starting from $\mathbf{x} \approx \check{\mathbf{x}}$. Usually at most one iteration is required, because the non-linear distortions vary slowly across the image.

Two-view geometry

From a single image of an unknown scene, no 3D measurements can be derived, because the third dimension (the depth along the ray) ist lost during projection. The photogrammetric measurement principle is to acquire multiple images from different viewpoints and measure *corresponding* points, meaning image points which are the projections of the same physical object point. From correspondences on can reconstruct the 3D coordinates via triangulation. The minimal case of two views forms the nucleus for this approach.

The coplanarity constraint

A direct consequence of the collinearity constraint is the *coplanarity* constraint for two cameras: the viewing rays through corresponding image points must be coplanar,

because they intersect in the 3D point. It follows that even if only the *relative orientation* between the two cameras is known one can reconstruct a (projectively distorted) straight line-preserving model of the 3D world, by intersecting corresponding rays; and that if additionally the interior orientations are known (the cameras are calibrated), one can reconstruct an angle-preserving model of the world in the same way. The scale of such a *photogrammetric model* cannot be determined without external reference, because scaling up and down the two ray bundles together does not change the coplanarity.

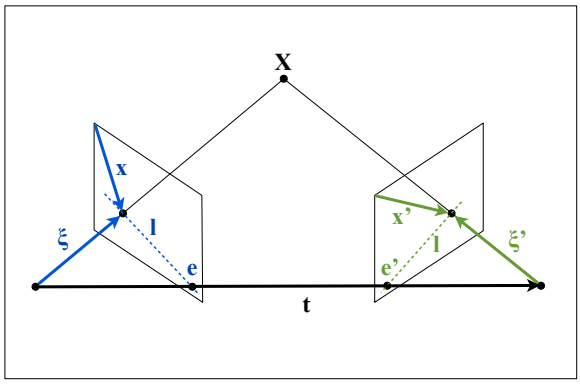


Fig. 2. The coplanarity constraint: the two projection rays $\boldsymbol{\xi}$, $\boldsymbol{\xi}'$ must lie in one plane, which also contains the baseline $\mathbf{t}$ and the object point $\mathbf{X}$. As a consequence, possible correspondences to an image point $\mathbf{x}$ must lie on the epipolar line $\mathbf{l}'$ and vice versa. All epipolar lines $\mathbf{l}$ intersect in the epipole $\mathbf{e}$, the image of the other camera's projection center.

Now let us look at a pair of corresponding image points $\mathbf{x}$ in the first image and $\mathbf{x}'$ in the second image. The coplanarity constraint (or *epipolar constraint*) states that the two corresponding rays in object space must lie in a plane. By construction that plane also contains the *baseline* $\mathbf{t} = \mathbf{X}_0^{e\prime} - \mathbf{X}_0^{e}$ between the projection centers. From (2), (3) the ray direction through $\mathbf{x}$ in object space (in Euclidean corrdinates) is $\boldsymbol{\xi} = \mathsf{R}^\top \mathsf{K}^{-1}\mathbf{x}$, and similarly for the second camera $\boldsymbol{\xi}' = \mathsf{R}'^\top \mathsf{K}'^{-1}\mathbf{x}'$.

Coplanarity between the three vectors implies

$$\boldsymbol{\xi} \cdot (\mathbf{t} \times \boldsymbol{\xi}') = \mathbf{x}^\top \mathsf{K}^{-\top} \mathsf{R} [\mathbf{t}]_\times \mathsf{R}'^\top \mathsf{K}'^{-1} \mathbf{x}' = \mathbf{x}^\top \mathsf{F} \mathbf{x}' = 0 \,. \tag{13}$$

The matrix F is called the *fundamental matrix* and completely describes the relative orientation. It has the following properties:

- $\mathbf{l} = \mathsf{F}\mathbf{x}'$ is the *epipolar line* to $\mathbf{x}'$, i.e. the image of the ray $\boldsymbol{\xi}'$ in the first camera. Corresponding points to $\mathbf{x}'$ must lie on that line, $\mathbf{x}^\top \mathbf{l} = 0$. Conversely, $\mathbf{l}' = \mathsf{F}^\top \mathbf{x}$ is the epipolar line to $\mathbf{x}$.
- The left nullspace of F is the *epipole* $\mathbf{e}$ of the first image, i.e. the image of the second projection center $\mathbf{X}'_0$ in which all epipolar lines intersect, $\mathsf{F}^\top \mathbf{e} = 0$. Conversely, the right nullspace of F is the epipole of the second image.
- F is singular and has rank ≤ 2, because $[\mathbf{t}]_\times$ has rank ≤ 2. Accordingly, F maps points to lines. It thus has seven degrees of freedom (nine entries determined only up to a common scale factor, minus one rank constraint).

The coplanarity constraint is linear in the elements of F and bilinear in the image coordinates, which is the basis for directly estimating the relative orientation.

If the interior orientations of both cameras are known (the cameras have been calibrated), the epipolar constraint can also be written in camera coordinates. The ray from the projection center to $\mathbf{x}$ in the camera coordinate system is given by $\boldsymbol{\eta} = \mathsf{K}^{-1}\mathbf{x}$, and similarly $\boldsymbol{\eta}' = \mathsf{K}'^{-1}\mathbf{x}'$. With these direction vectors the epipolar constraint reads

$$\boldsymbol{\eta}^\top \mathsf{R} [\mathbf{t}]_\times \mathsf{R}'^\top \boldsymbol{\eta}' = \boldsymbol{\eta}^\top \mathsf{E} \boldsymbol{\eta}' = 0 \,. \tag{14}$$

The matrix E is called the *essential matrix* and completely describes the relative orientation between calibrated cameras, i.e. their relative rotation and the direction of the relative translation (the baseline). It has the following properties:

- E has rank 2. Additionally, the two non-zero eigenvalues are equal up to scale. E has five degrees of freedom, corresponding to the relative orientation of an

angle-preserving photogrammetric model (three for the relative rotation, two for the baseline direction).

- The constraint between calibrated rays is still linear in the elements of E and bilinear in the image coordinates.

For completeness it shall be mentioned that beyond coplanarity further constraints, so-called *trifocal constraints* exist between triplets of images: if a corresponding straight line has been observed in three images, then the planes formed by the associated projection rays must all intersect in a single 3D line, see for example (Hartley, 1997; McGlone, 2013). This topic is not further treated here.

Absolute orientation

The transformation from the coordinates of the angle-preserving photogrammetric model $\mathbf{X}^m$ to a given 3D object coordinate system $\mathbf{X}$ is called *absolute orientation*. It corresponds to a similarity transform and thus has seven degrees of freedom (translation, rotation and scaling of the model).

$$\mathbf{X} = \mathsf{SRT}\mathbf{X}^m = \begin{bmatrix} \frac{1}{s}\mathsf{I} & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathsf{R} & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathsf{I} & -\mathbf{T} \\ \mathbf{0}^\top & 1 \end{bmatrix} \mathbf{X}^m \,. \tag{15}$$

Analytical operations

The input to the photogrammetric process are raw images, respectively coordinates measured in those images. This section describes methods to estimate unknown parameters from image coordinates, using the models developed above.

Single image orientation

The complete orientation of a single image has 11 unknowns (5 for the interior orientation and 6 for the exterior orientation). An image point affords two observations,

thus at least six *ground control points* in the object coordinate system and their corresponding image points are required. An algebraic solution, known as the *Direct Linear Transform* or DLT (Abdel-Aziz and Karara, 1971), is obtained directly from equation (6).

$$[\mathbf{x}]_\times \mathsf{P}\mathbf{X} = \left([\mathbf{x}]_\times \otimes \mathbf{X}^\top\right) \mathbf{p} = \mathbf{0}\,, \tag{16}$$

with the vector $\mathbf{p} = \mathrm{vec}(\mathsf{P}) = [P_{11}, P_{12}, \ldots, P_{34}]^\top$. For each control point two of the three equations are linearly independent. Selecting two such equations for each of $N \geq 6$ ground control points and stacking them yields a homogeneous linear system $\mathsf{A}_{2N\times 12}\mathbf{p} = \mathbf{0}$, which is solved with singular value decomposition to obtain the projection matrix P. Note that the DLT fails if all control points are coplanar, and is unstable if they are nearly coplanar. A further critical configuration, albeit of rather theoretical interest, is if all control points lie on a twisted cubic curve (Hartley and Zisserman, 2004).

The direct algebraic solution is not geometrically optimal, but can serve as a starting value for an iterative estimation of the optimal Euclidean camera parameters, see (McGlone, 2013).

A further frequent orientation procedure is to determine the exterior orientation of a single camera with known interior orientation. Three non-collinear control points are needed to determine the six unknowns. The procedure is known as *spatial resection* or P3P problem.

The solution (Grunert, 1841) is sketched in Figure 3. With the known interior orientation the three image points are converted to rays in camera coordinates, $\eta_i = \mathsf{K}^{-1}\mathbf{x}_i, i = 1\ldots 3$. The pairwise angles between these rays are determined via $\cos\alpha = \frac{1}{|\eta_2||\eta_3|}\eta_2^\top\eta_3$ etc. In the object coordinate system the distances between the points are determined, $a = |\mathbf{X}_2^e - \mathbf{X}_3^e|$ etc. The three triangles containing the projection center now give rise to constraints

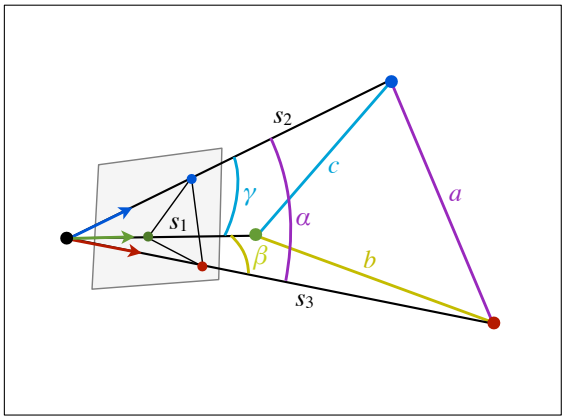


Fig. 3. Spatial resection: the image coordinates together with the interior orientation give rise to three rays in camera coordinates, forming a trilateral pyramid. Applying the cosine law on each of the pyramids faces relates the pairwise angles α, β, γ between the rays and the known distances a, b, c between the object points to the pyramids sides s_1, s_2, s_3. Soving for these three lengths yields the camera coordinates of the three points.

$$\begin{aligned} a^2 &= s_2^2 + s_3^2 - 2s_2 s_3 \cos\alpha \\ b^2 &= s_3^2 + s_1^2 - 2s_3 s_1 \cos\beta \\ c^2 &= s_1^2 + s_2^2 - 2s_1 s_2 \cos\gamma \end{aligned} \tag{17}$$

for the three unknowns s_1, s_2, s_3. Substituting auxiliary variables $u = \frac{s_2}{s_1}$ and $v = \frac{s_3}{s_1}$ yields, after some manipulations, a fourth-order polynomial for v and hence up to four solutions, see (e.g. Haralick et al, 1994). Back-substitution delivers first u and then the three distances s_1, s_2, s_3. Given these distances, the ground control points in camera coordinates follow from $\tilde{\mathbf{x}}_i^e = s_i \cdot \frac{\eta_i}{|\eta_i|}$. The exterior orientation is then found by computing the rotation and translation between the $\tilde{\mathbf{x}}_i$ and the $\mathbf{X}_i$.

There are two critical configurations for spatial resection: one where the projection center is located on (or near) a circular cylinder generated by sweeping the circle through $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ along the triangle's normal vector; and the other when the control points lie on the cubic horopter curve.

Based on the geometric construction (17) several other algebraic schemes exist to solve the equation system, see (e.g. Haralick et al, 1994). For more than three control

points an iterative optimal resection algorithm can be found in the literature (McGlone, 2013).

Relative orientation of two images

A further elementary operation is the relative orientation of images to gain a photogrammetric model. The present chapter deals with the relative orientation of two images. Since relative orientations can be transitively chained, that operation forms the elementary building block for orienting larger image networks (note, in practice it is often preferred to chain image triplets because the associated redundancy affords robustness, however that case is not treated here).

The basis of relative orientation from observed image correspondences is the coplanarity constraint (13). Since the constraint is linear in the unknown elements of the relative orientation, it can be directly reordered and solved. Each corresponding point pair gives rise to an equation

$$\left(\mathbf{x}^\top \otimes \mathbf{x}'^\top\right) \mathbf{f} = 0 \,, \tag{18}$$

with $\mathbf{f} = \mathrm{vec}(\hat{\mathsf{F}}) = [F_{11}, F_{12}, \ldots, F_{33}]^\top$. Stacking $\geq$8 such equations yields a regular, respectively overdetermined, homogeneous equations system for $\mathbf{f}$.

The direct solution ignores the rank-deficiency of the fundamental matrix, instead using at least 8 points to determine 7 unknowns. Due to measurement noise the resulting matrix $\hat{\mathbf{F}}$ will not be a fundamental matrix. To correct this, one can find the nearest (according to the Frobenius norm) rank-2 matrix by decomposing $\hat{\mathsf{F}}$ with SVD and nullifying the smallest singular value,

$$\hat{\mathsf{F}} = \mathsf{U} \cdot \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3) \cdot \mathsf{V}^\top \quad , \quad \mathsf{F} = \mathsf{U} \cdot \mathrm{diag}(\lambda_1, \lambda_2, 0) \cdot \mathsf{V}^\top \tag{19}$$

This so-called “8-point algorithm” (Longuet-Higgins, 1981) can be used in equivalent form to estimate the essential matrix between two calibrated cameras. Reordering (14)

to

$$\left(\boldsymbol{\eta}^\top \otimes \boldsymbol{\eta}'^\top\right) \mathbf{e} = 0 \tag{20}$$

yields the entries of $\mathbf{e} = \mathrm{vec}(\hat{\mathsf{E}})$, and the solution is corrected to the nearest essential matrix by enforcing the constraints on the singular values,

$$\hat{\mathsf{E}} = \mathsf{U} \cdot \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3) \cdot \mathsf{V}^\top \quad , \quad \mathsf{E} = \mathsf{U} \cdot \mathrm{diag}(1, 1, 0) \cdot \mathsf{V}^\top \tag{21}$$

Estimating the 7 unknowns of F or the 5 unknowns of E from 8 points is obviously not a minimal solution, and thus not ideal—especially since a main application of the direct solution is robust estimation in RANSAC-type sampling algorithms. A minimal solution for F, called the "7-point algorithm" can be obtained in the following way (von Sanden, 1908; Hartley, 1994): only 7 equations (18) are stacked into $\mathsf{A}_{7\times 9}\mathbf{f} = 0$. Solving this expression with SVD yields a two-dimensional null-space

$$\mathbf{f}(\delta) = \delta \mathbf{v}_8 + \mathbf{v}_9 \ , \tag{22}$$

with arbitrary δ. To find a fundamental matrix (i.e. a rank-deficient matrix) in that null-space one introduces the nine elements of $\mathbf{f}(\delta)$ into the determinant constraint $\det(\mathsf{F}) = 0$ and analytically expands the determinant with Sarrus' rule. This results in a cubic equation for δ, and consequently in either one or three solutions for F.

Following the same idea a "5-point algorithm" exists for the calibrated case (Nistér, 2004): stacking (20) for five correspondences leads to a 4-dimensional null-space

$$\mathbf{e}(\delta, \epsilon, \zeta) = \delta \mathbf{v}_6 + \epsilon \mathbf{v}_7 + \zeta \mathbf{v}_8 + \mathbf{v}_9 \tag{23}$$

Again this can be substituted back into the determinant constraint. Furthermore, it can be shown that the additional constraints on the fundamental matrix can be written

$$\mathsf{E}\mathsf{E}^\top\mathsf{E} - \frac{1}{2}\mathrm{trace}(\mathsf{E}\mathsf{E}^\top)\mathsf{E} = 0 \tag{24}$$

in which one can also substitute (23). Through further—rather cumbersome—algebraic variable substitutions one arrives at a 10^{th}-order polynomial in ζ, which is solved numerically. For the (up to 10) real roots one then recovers δ and ϵ, and thus E, through back-substitution.

The fundamental matrix is ambiguous if all points are coplanar in object space. The corresponding equations become singular in that case, and unstable near it. On the contrary, the essential matrix does not suffer from that problem, and in fact can be estimated from only 4 correspondences if they are known to be coplanar (Wunderlich, 1982).

Naturally, once initial values for the relative orientation parameters are available an iterative solution exists to find the geometrically optimal solution for an arbitrary number of correspondences, see (McGlone, 2013).

Having determined E it is in many cases necessary to extract explicit relative orientation parameters (rotation and translation direction) for the image pair. Given the singular value decomposition (21) and the two auxiliary matrices

$$\mathsf{W} = \begin{bmatrix} 0 & \pm 1 & 0 \\ \mp 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad , \quad \mathsf{Z} = \begin{bmatrix} 0 & \pm 1 & 0 \\ \mp 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{25}$$

the orientation elements are given by

$$[\mathbf{t}]_\times = \mathsf{U}\mathsf{Z}\mathsf{U}^\top \quad , \quad \mathsf{R} = \mathsf{U}\mathsf{W}\mathsf{V}^\top \tag{26}$$

which can be easily verified by checking $\mathsf{E} = [\mathbf{t}]_\times \mathsf{R}$. The sign ambiguities in W and Z give rise to four combinations, corresponding to all combinations of the "upright" and "upside-down" camera configurations for the two images. The correct one is found by checking in which one a 3D object point is located in front of both cameras.

Reconstruction of 3D points

For photogrammetric 3D reconstruction the camera orientations are in fact only an unavoidable by-product, whereas the actual goal is to reconstruct 3D points (note however that the opposite is true for image-based navigation). The basic operation of reconstruction is triangulation of 3D object points from cameras with known orientations P_i. A direct algebraic solution is found from the collinearity constraint in the form (6). Each image point gives rise to

$$([\mathbf{x}_i]_\times \mathsf{P}_i)\,\mathbf{X} = \mathbf{0} \tag{27}$$

of which two rows are linearly independent. Stacking the equations leads to an equation system for the object point $\mathbf{X}$. Solving with SVD yields a unique solution for two cameras $\mathsf{P}_1, \mathsf{P}_2$, respectively a (projective) least-squares solution for more than two cameras.

A geometrically optimal solution for two views exists, which involves numerically solving a polynomial of degree 6 (Hartley and Sturm, 1997). Iterative solutions for $\geq$3 views also exist. In general the algebraic solution (27) is a good approximation, if one employs proper numerical conditioning.

Orientation of multi-image networks

Most applications require a network of >2 images to cover the object of interest.[3] By combinations of the elementary operations described above, the relative orientation of all images in a common coordinate system can be found: either one can chain two-view relative orientations together while estimating the relative scale from a few object points, or one can generate a photogrammetric model from two views and then interatively add additional views to it by alternating single-image orientation with tri-

[3] In aerial photogrammetry the network is often called an "image block", since the images are usually recorded on a regular raster.

angulation of new object points. Absolute orientation is accomplished (either at the end or at an intermediate stage) by estimating the 3D similarity transform that aligns the photogrammetric model with known ground control points in the object coordinate system.

Obviously such an iterative procedure will lead to error build-up. In most applications the image network is thus polished with a global least-squares optimization of all unknown parameters. The specialization of least-squares adjustment to photogrammetric ray bundles, using the collinearity constraint as functional model, is called *bundle adjustment* (Brown, 1958; Triggs et al, 1999; McGlone, 2013). Adjustment proceeds in the usual way: the constraints $\mathbf{y} = f(\mathbf{x})$ between observations $\mathbf{y}$ and unknowns $\mathbf{x}$ are linearized at the approximate solution $\mathbf{x}_0$, leading to an over-determined equation system $\delta\mathbf{y} = \mathsf{J}\,\delta\mathbf{x}$. The equations are solved in a least-squares sense,

$$\begin{aligned} \mathsf{N}\,\delta\mathbf{x} &= \mathbf{n} \\ \mathsf{N} = \mathsf{J}^\top\mathsf{S}_{yy}^{-1}\mathsf{J} \quad &, \quad \mathbf{n} = \mathsf{J}^\top\mathsf{S}_{yy}^{-1}\delta\mathbf{y}\ , \end{aligned} \tag{28}$$

with S_{yy} the covariance matrix of the observations. The approximate solution is then updated, $\mathbf{x}_1 = \mathbf{x}_0 + \delta\mathbf{x}$, and the procedure iterated until convergence.

In order to yield geometrically optimal solutions the collinearity constraint is first transformed to Euclidean space by removing the projective scale. Denoting cameras by index j, object points by index i, and the rows of the projection matrix by $\mathsf{P}^{(1)}$, $\mathsf{P}^{(2)}$, $\mathsf{P}^{(3)}$, we get

$$x_{ij}^e = \frac{\mathsf{P}_j^{(1)}\mathbf{X}_i}{\mathsf{P}_j^{(3)}\mathbf{X}_i} \quad , \quad y_{ij}^e = \frac{\mathsf{P}_j^{(2)}\mathbf{X}_i}{\mathsf{P}_j^{(3)}\mathbf{X}_i} \tag{29}$$

These equations must then be linearized for all observed image points w.r.t. the orientation parameters contained in the $\mathbf{P}_j$ as well as the 3D object point coordinates $\mathbf{X}_i$. Moreover, equations for the ground control points, as well as additional measurements such as for example GPS/IMU observations for the projection centers, are added.

For maximum accuracy it is also common to regard interior orientation parameters (including non-linear distortions) as observations of a specified accuracy rather than as constants, and to estimate their values during bundle adjustment. This so-called *self calibration* can take different forms, e.g. for crowd-sourced amateur images it is usually required to estimate the focal length and radial distortion of each individual image, whereas for professional aerial imagery it is common to use a single set of orientation parameters for all images, but include more complex non-linear distortion coefficients. For details about GPS/IMU integration, self calibration etc. see (McGlone, 2013).

The normal equations for photogrammetric networks are often extremely large (up to $>10^6$ unknowns). However, they are also highly structured and very sparse (<1% non-zero coefficients), which can be exploited to efficiently solve them. The most common procedure is to eleminate the largest portion of the unknowns, namely the 3D object point coordinates, with the help of the Schur complement. Let index x denote object point coordinates, and index q all other unknown parameters, then the normal equations can be written

$$\begin{bmatrix} \mathsf{N}_{xx} & \mathsf{N}_{xq} \\ \mathsf{N}_{xq}^\top & \mathsf{N}_{qq} \end{bmatrix} \begin{bmatrix} \delta\mathbf{x}_x \\ \delta\mathbf{x}_q \end{bmatrix} = \begin{bmatrix} \mathbf{n}_x \\ \mathbf{n}_q \end{bmatrix} \quad . \tag{30}$$

Inverting N_{xx} is cheap because it is block-diagonal with individual (3×3)-blocks for each object point. Using that fact the normal equations can efficiently be reduced to a much smaller system

$$\begin{gathered} \bar{\mathsf{N}}\ \delta\mathbf{x}_q = \bar{\mathbf{n}}\ , \quad \text{with} \\ \bar{\mathsf{N}} = \mathsf{N}_{qq} - \mathsf{N}_{xq}^\top \mathsf{N}_{xx}^{-1} \mathsf{N}_{xq} \quad , \quad \bar{\mathbf{n}} = \mathbf{n}_q - \mathsf{N}_{xq}^\top \mathsf{N}_{xx}^{-1} \mathbf{n}_x\ . \end{gathered} \tag{31}$$

The standard way to solve the reduced normal equations is to adaptively dampen the equation system with the Levenberg-Marquart method (Levenberg, 1944; Nocedal and Wright, 2006) for better convergence, i.e. the equation system is modified to

$$(\bar{\mathsf{N}} - \lambda \mathsf{I}_q)\delta\mathbf{x}_q = \bar{\mathbf{n}} \;, \tag{32}$$

with I_q the identity matrix of appropriate size, and the damping factor λ depending on the success of the previous iteration. The system (32) is then reduced to triangular form with variants of Cholesky factorization. Using recursive partitioning and equation solvers which exploit sparsity, it is possible to perform bundle adjustment for photogrammetric networks with >10'000 cameras.

Due to automatic tie-point measurement as well as the sheer size of modern photogrammetric campaigns, blunders—mainly incorrect tie point matches—are unavoidable in practice. Therefore bundle adjustment routinely employs robust methods such as iterative reweighted least squares (IRLS) (e.g. Huber, 1981) to defuse, and subsequently eliminate, gross outliers.

Conclusion

A brief summary has been given of the elementary geometry underlying photogrammetric modeling, as well as the mathematical operations for image orientation and image-based 3D reconstruction. The theory of photogrammetry started to emerge in the 19$^{\text{th}}$ century, most of it was developped in the 20$^{\text{th}}$. The geometric relations that govern the photographic imaging process, and their inversion for 3D measurement purposes, are nowadays well understood, the theory is mature and has been compiled—in much more detail than here—in several excellent textbooks (e.g. Hartley and Zisserman, 2004; Luhmann et al, 2006; McGlone, 2013). Still, some important findings, such as the direct solution for relative orientation of calibrated cameras, are surprisingly recent.

References

Abdel-Aziz YI, Karara HM (1971) Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. In: Proceedings of the Symposium on Close-Range Photogrammetry, American Society of Photogrammetry

Brown DC (1958) A solution to the general problem of multiple station analytical stereotriangulation. Tech. Rep. RCA-MTP Data Reduction Technical Report No. 43, Patrick Airforce Base

Das GB (1949) A mathematical approach to problems in photogrammetry. Empire Survey Review 10(73):131–137

Förstner W (2010) Minimal representations for uncertainty and estimation in projective spaces. In: Proceedings of the Asian Conference on Computer Vision, Springer Lecture Notes in Computer Science, vol 6493

Grunert JA (1841) Das Pothenot'sche Problem in erweiterter Gestalt; nebst Bemerkungen über seine Anwendung in der Geodäsie. Grunert Archiv der Mathematik und Physik 1

Haralick RM, Lee CN, Ottenberg K, Nölle M (1994) Review and analysis of solutions of the three point perspective pose estimation problem. International Journal of Computer Vision 13(3):331–356

Hartley R, Zisserman A (2004) Multiple view geometry in computer vision, 2nd edn. Cambridge University Press

Hartley RI (1994) Projective reconstruction and invariants from multiple images. IEEE Transactions on Pattern Analysis and Machine Intelligence 16(10):1036–1041

Hartley RI (1997) Lines and points in three views and the trifocal tensor. International Journal of Computer Vision 22(2):125–140

Hartley RI, Sturm P (1997) Triangulation. Computer Vision and Image Understanding 68(2):146–157

Huber PJ (1981) Robust Statistics. Wiley

Levenberg K (1944) A method for the solution of certain non-linear problems in least squares. Quarterly of Applied Mathematics 2:164168

Longuet-Higgins HC (1981) A computer algorithm for reconstructing a scene from two projections. Nature 293:133–135

Luhmann T, Robson S, Kyle S, Harley I (2006) Close range photogrammetry. Principles, Methods and Applications. Whittles Publishing

McGlone JC (ed) (2013) Manual of Photogrammetry, sixth edn. Americal Society for Photogrammetry and Remote Sensing

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

Nocedal J, Wright SJ (2006) Numerical Optimization, 2nd edn. Springer

Pajdla T (2002) Stereo with oblique cameras. International Journal of Computer Vision 47(1-3):161–170

Semple JG, Kneebone GT (1952) Algebraic projective geometry. Clarendon Press

Triggs B, McLauchlan PF, Hartley RI, Fitzgibbon AW (1999) Bundle adjustment – a modern synthesis. In: Vision Algorithms: Theory and Practice, Springer Lecture Notes in Compuer Science, vol 1883

von Sanden H (1908) Die Bestimmung der Kernpunkte in der Photogrammetrie. PhD thesis, Universität Göttingen

Wunderlich W (1982) Rechnerische Rekonstruktion eines ebenen Objektes aus zwei Photographien. Mitteilungen des Geodätisches Instituts der TU Graz 40:265–377