1 A tour into multiple image geometry

This Chapter provides a conceptual overview of the book by introducing in a simple and intuitive way some of its main ideas.

1.1 Multiple image geometry and three-dimensional vision

The purpose of vision is to infer descriptions of the world from images. We will concentrate on a limited but crucial type of description, that of geometry in space, and will investigate how it can be recovered using only geometric constraints and excluding semantic ones. As it will be explained soon, based on geometry alone, it is not possible to infer the 3-D positions of points in a scene from a single image of this scene. As humans, we rely on our semantic knowledge of the world to perform such an inference, but this capability can be easily fooled, as illustrated in Figure 1.1.

The central problem which we wish to address is therefore concerned with multiple images: given two (or more) images of a scene, a partially instantiated camera model, and points in these images which correspond to the same point in the world, construct a description of the 3-D spatial relations between the points in the world. In addition, one would like to complete the instantiation of the camera models and describe the 3-D spatial relations between the cameras which were used to create the images. Indeed, from the difference in position of image points, it is possible to infer spatial relations by taking advantage of the geometric rules which govern the formation of images. The theoretical focus of the book is on the rigorous exposition of these rules, using geometric and algebraic tools. Unlike standard texts on projective geometry, we concentrate on the relation between 3-D space and 2-D space, between points and their projection.

To give the reader an idea of the applications that we have in mind, we give two examples. Throughout this Chapter, we will pause to see which progress we will have made towards being able to process them. In the first example, we are handed ten images of a scene taken by an unknown and possibly zooming camera, three of which are shown in Figure 1.2 (the others are in Figure 10.8). Our goal is to be able to make accurate length measurements in space. From these images, we construct a 3-D geometric model of the buildings and illustrate the correctness of the recovered geometry by showing two rotated views of the reconstruction in Figure 1.3. The model can be used to generate a synthetic view, obtained from a higher viewpoint than the original images, as shown in Figure 1.4 (see also Figure 1.26). Figure 1.5 illustrate that the cameras positions and orientations can also be estimated as part of the reconstruction process.

The second example demonstrates the capacity of the techniques described in this book to deal with continuous streams of images for applications to augmented reality. The sequence (from which images in Figure 1.6 are extracted) is taken with an unknown camera from a helicopter flying over a site where a power plant is to

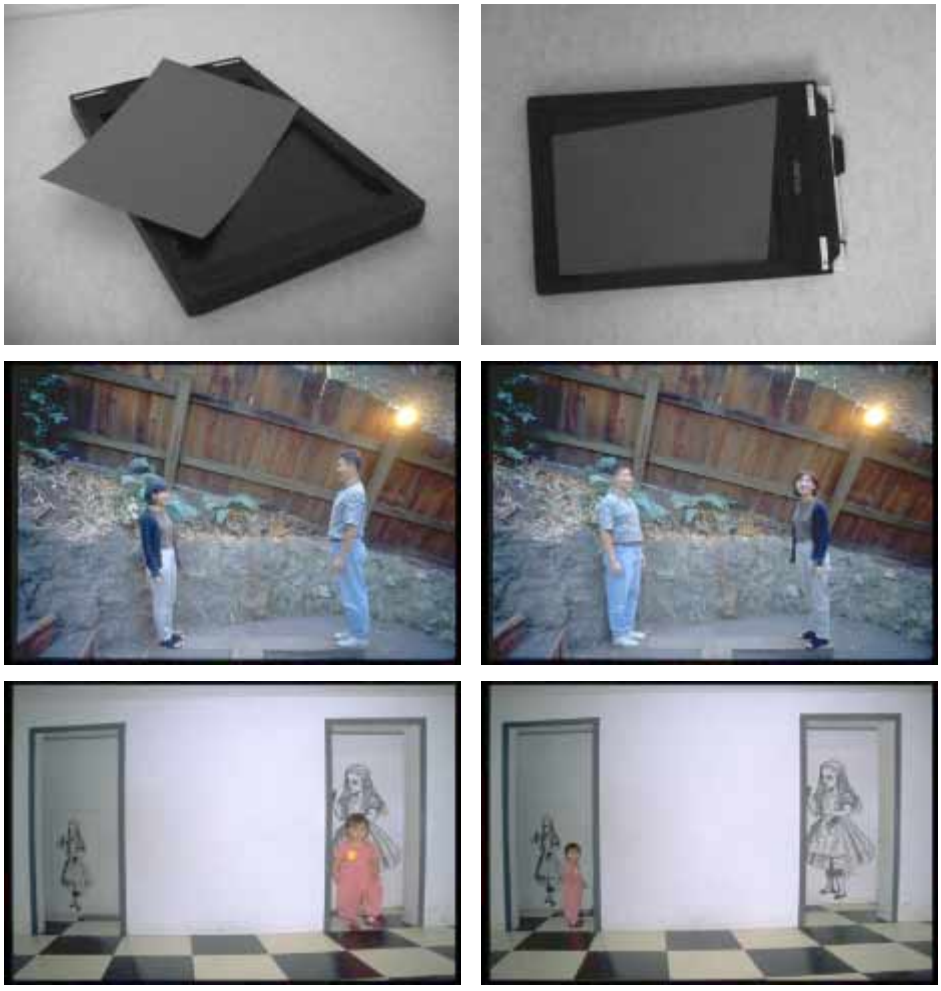

Figure 1.1: A few examples illustrating the difficulty, even for humans, of inferring the geometry of a scene from a single image. Viewers tend to assume that the angles formed by the grey card are right angles (top left). A bird's eye view (top right) reveals that only one of the four angles is a right angle. Relative size judgment can be easily misled by particular spatial configurations which defeat common assumptions (middle and bottom images).

Figure 1.2: A few images of the Arcades square in Valbonne, taken with a digital camera. Courtesy **Sylvain Bougnoux**, INRIA.

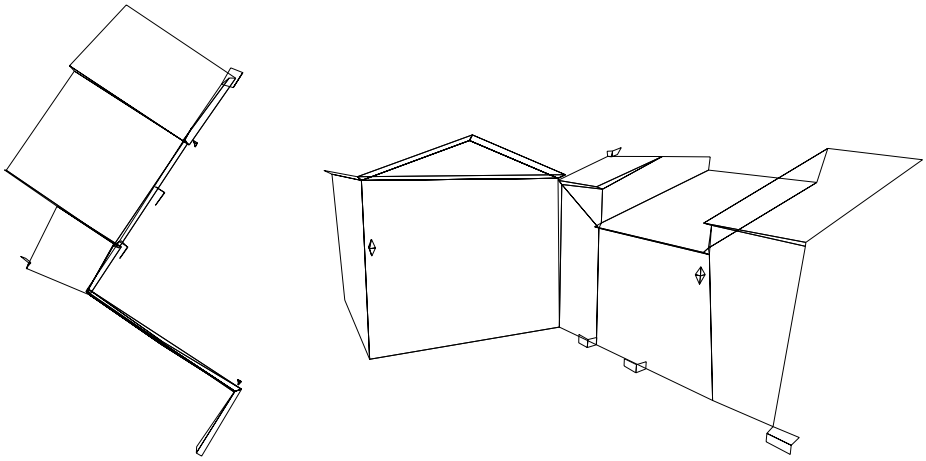

Figure 1.3: Top and front view of the reconstruction of the Arcades square in Valbonne, illustrating the metric correctness of the recovered 3-D model. Courtesy **Sylvain Bougnoux**, INRIA.

be constructed. We wish to add to the scene an artificial object, consisting of a model of the new power plant shown in Figure 1.7 while respecting the geometry of the scene and the movements of the camera. We show in Figure 1.8 three images of the modified sequence. It can be verified that the synthetic objects appear, as they should, to be static with respect to the environment.

1.2 Projective geometry [Chapters 2 and 3]

Euclidean geometry describes our world well: the measurements we make in terms of lengths, angles, parallelism, and orthogonality are meaningful because they are preserved by a change of coordinates which is a *displacement* (rotation and translation), a *Euclidean transformation*. Because of that, Euclidean geometry has also

Figure 1.4: Textured view of the reconstruction, illustrating that synthetic images can be generated from a novel viewpoint. Courtesy **Sylvain Bougnoux**, INRIA.

traditionally been used by vision scientists to describe the geometry of projection.

However, we believe that for the purpose of describing projections, projective geometry is a more adequate framework. As illustrated by Figure 1.2, the railroad tracks are parallel lines in 3-D space, but in the image they are no longer parallel, and appear to converge as they recede towards the horizon, towards a *vanishing point*. Any set of parallel, horizontal lines, whether they lie on the ground or not, appears to meet at a single point on the horizon line. In addition, all the points at infinity in 3-D have the same projection as the observer moves. The rails always seem to disappear at the same point, and as you move in the night, the moon and stars seem to follow you. Since parallelism is not preserved by projection, clearly neither are distances nor angles. Projective geometry is an extension of Euclidean geometry, which describes a larger class of transformations than just rotations and translations, including in particular the perspective projection performed by a camera. It makes it possible to describe naturally the phenomena at infinity that we just noticed. Between projective geometry and Euclidean geometry there are two other geometries, similarity[1] and affine, as illustrated in Table 1.2 (See also Table 1.4).

Let's start with a point of Euclidean[2] coordinates $[u, v]^T$ in the plane. Its projective coordinates are obtained by just adding 1 at the end: $[u, v, 1]^T$. Having now three coordinates, in order to obtain a "one-to-one" correspondence between

[1] The only difference between displacements and similarities is that the latter ones allow for a global scale factor. Since in the context of reconstruction from images, such an ambiguity is always present, we will designate by abuse of language *Euclidean*. transformations the similarity transformations.

[2] Technically speaking, the term "affine" would be more appropriate, but in the context of this section we use by abuse of language the more familiar term "Euclidean". See Chapter 2 for an explanation.

Figure 1.5: Another textured view of the reconstruction, showing also the estimated positions and orientations of some of the cameras: Courtesy **Sylvain Bougnoux**, INRIA.

Euclidean coordinates and projective coordinates, we have the rule that scaling by a nonzero factor is not significant, so that the two triples of coordinates $[u, v, 1]^T$ and $[\lambda u, \lambda v, \lambda]^T$ represent the same point.

More generally, the space of $(n+1)$-tuples of coordinates, with the rule that proportional $(n+1)$-tuples represent the same point, is called the *projective space* of dimension n and denoted $\mathbb{P}^n$. The object space will be considered as $\mathbb{P}^3$ and the image space as $\mathbb{P}^2$, called the *projective plane*. We will see in Section 1.3 that projective coordinates represent naturally the operation performed by a camera. Given coordinates in $\mathbb{R}^n$ we can build projective coordinates by the correspondence

$$[x_1, \ldots, x_n]^T \rightarrow [x_1, \ldots, x_n, 1]^T.$$

To transform a point in the projective coordinates back into Euclidean coordinates, we just divide by the last coordinate and then drop it:

$$[x_1, \ldots, x_n, x_{n+1}]^T \rightarrow [\frac{x_1}{x_{n+1}}, \ldots, \frac{x_n}{x_{n+1}}]^T.$$

We see that the projective space contains more points than the Euclidean space. Points with coordinates $[x_1, \ldots, x_n, x_{n+1}]^T$ with $x_{n+1} \neq 0$ can be viewed as the usual points, whereas the points with coordinates $[x_1, \ldots, x_n, 0]^T$ have no Euclidean equivalent. If we consider them as the limit of $[x_1, \ldots, x_n, \lambda]^T$, when $\lambda \rightarrow 0$ i.e. the limit of $[x_1/\lambda, \ldots, x_n/\lambda, 1]^T$, then we see that they are the limit of a point of $\mathbb{R}^n$ going to infinity in the direction $[x_1, \ldots, x_n]^T$, hence the appellation *point at infinity*. The projective space $\mathbb{P}^n$ can be viewed as the union of the usual space $\mathbb{R}^n$ (points $[x_1, \ldots, x_n, 1]^T$) and the set of points at infinity $[x_1, \ldots, x_n, 0]^T$. The neat thing about this formalism is that points at infinity are not special and are treated just like any other point.

Figure 1.6: Three images of a sequence taken from a helicopter: Courtesy **Luc Robert**, INRIA.

Let's go back to the projective plane. There is one point at infinity for each direction in the plane: $[1,0,0]^T$ is associated with the horizontal direction, $[0,1,0]^T$ is associated with the vertical direction, and so on.

To represent a line in the projective plane, we begin with the standard equation $au + bv + c = 0$. Since it is independent of scaling, we can write it using projective coordinates $\mathbf{m} = [x, y, z]^T$ of the point m:

$$\mathbf{l}^T\mathbf{m} = \mathbf{m}^T\mathbf{l} = ax + by + cz = 0,$$

where the line l is represented by a vector with three coordinates defined up to a scale factor, exactly like a 2-D point: $\mathbf{l} = [a, b, c]^T$ is the projective representation of the line. Since the representation of points is the same as the representation of lines, several results concerning points can be transposed to lines: this is the notion of *duality*. Please note that we use throughout the book the convention that bold

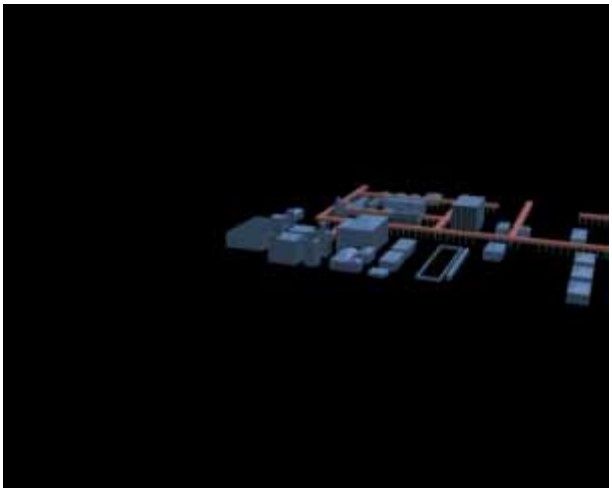

Figure 1.7: Model to insert in the sequence of Figure 1.6. Courtesy **Luc Robert**, INRIA.

type is used to represent the coordinate vector of the geometric object which is in corresponding normal type, such as $\mathbf{m}, m, \mathbf{l}$, and l in this example.

It can be verified with elementary algebra that the line containing the two points m and m' (their *join*) is expressed very simply as the cross-product of their representations:

$$\mathbf{l} \simeq \mathbf{m} \times \mathbf{m}' = \begin{bmatrix} yz' - zy' \\ zx' - xz' \\ xy' - yx' \end{bmatrix}.$$

Note that the three coordinates are just the determinants of the three 2×2 submatrices of $[\mathbf{m}\,\mathbf{m}']$. The points on the line are described by $\mathbf{m}'' = \alpha\mathbf{m} + \beta\mathbf{m}'$. The three points m,m',m'' are aligned if and only if

$$(\mathbf{m} \times \mathbf{m}')^T \mathbf{m}'' = |\, \mathbf{m}, \mathbf{m}', \mathbf{m}'' \,| = 0.$$

By duality, the point at the intersection (their *meet*) of lines l and l' is $\mathbf{m} \simeq \mathbf{l} \times \mathbf{l}'$. The other properties of lines in 2-D space are summarized in Table 1.2.

Therefore, in the projective plane, points and lines have the same representation, and the cross-product describes both meet and join. An important advantage of the representation is that the cross-product is a linear operator, while the description of the meet and join with usual coordinates involves divisions. Being a linear operator, the cross-product can be written as a matrix product $\mathbf{v} \times \mathbf{x} = [\mathbf{v}]_\times \mathbf{x}$ where $[\mathbf{v}]_\times$ is

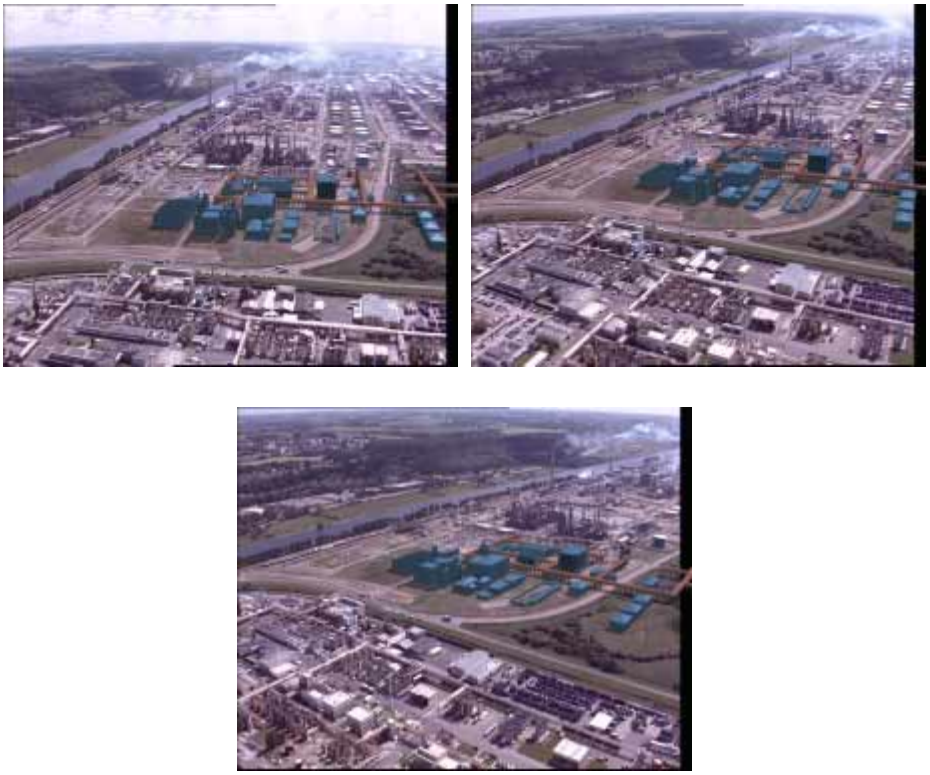

Figure 1.8: Result obtained from Images 1.6 and 1.7. Courtesy **Luc Robert**, INRIA.

the skew-symmetric matrix whose left and right nullspaces are the vector $\mathbf{v}$:

$$[\mathbf{v}]_\times = \begin{bmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{bmatrix}. \tag{1.1}$$

Therefore we can use a simple matrix operator to represent the geometric operation of union of two points to form a line or intersection of two lines to form a point.

If the lines l and l' are parallel, then the previous formula is still valid and gives a point m whose coordinates are found to be proportional to $[b, -a, 0]^T$, or, equivalently, to $[b', -a', 0]^T$. This is a point at infinity which represents the direction of l. We note that all the points at infinity belong in fact to the line of equation $[0, 0, 1]^T$, which is called the *line at infinity* of $\mathbb{P}^2$, and denoted l_∞. The intersection of the line l with the line at infinity l_∞ is, as expected, the point at infinity of l $[b, -a, 0]^T$. This is to be contrasted to Euclidean geometry, where the intersection of

Figure 1.9: Scene with converging lines.

two parallel lines is not defined, and where using the general formula for computing their intersection point leads to a division by zero. In projective geometry, we don't have this problem and therefore we don't need to handle particular cases. All this makes it possible to deal generally with the intersection and union of geometric objects very simply.

If we move to $\mathbb{P}^3$, a number of things are similar, although duality and the representation of lines are more complex [Chapter 3]. A plane is represented by a vector with four coordinates defined up to a scale factor, exactly like a 3-D point: $\mathbf{\Pi} = [\pi_1, \pi_2, \pi_3, \pi_4]^T$ represents the projective equation of the plane $\pi_1 X + \pi_2 Y + \pi_3 Z + \pi_4 = 0$, which means that a point $\mathbf{M} = [X, Y, Z, 1]^T$ belongs to a plane Π if and only if $\mathbf{\Pi}^T \mathbf{M} = 0$. In $\mathbb{P}^3$, the points $[X, Y, Z, 0]^T$ therefore form a plane of equation $[0, 0, 0, 1]^T$, called the *plane at infinity*, and denoted Π_∞. Intuitively, this plane represents directions of the usual planes, since the intersection of Π_∞ with a plane $\mathbf{\Pi} = [\pi_1, \pi_2, \pi_3, \pi_4]^T$ gives $[\pi_1, \pi_2, \pi_3]^T$ which corresponds to the normal of the plane Π, all parallel planes having the same normal. When a point in $\mathbb{R}^3$ tends to the point at infinity M_∞, for example a point on a line receding towards the horizon, its projection tends to the projection of M_∞ which is usually a finite point of $\mathbb{P}_2$ called a *vanishing point*, just as we had seen in Figure 1.2. Now we see another reason why projective geometry will be a useful tool to describe projection: projective transformations mix finite and infinite points, therefore there are less special cases since the points at infinity, which in fact represent directions, are handled just like ordinary points.

	Euclidean	similarity	affine	projective
Transformations				
rotation, translation	×	×	×	×
isotropic scaling		×	×	×
scaling along axes, shear			×	×
perspective projections				×
Invariants				
distance	×			
angles, ratios of distances	×	×		
parallelism, center of mass	×	×	×	
incidence, cross-ratio	×	×	×	×

Table 1.1: An ordering of geometries: particular transformations and properties left invariant by the transformations. Each geometry is a subset of the next. More general transformations mean weaker invariants.

points		lines	
coordinates of m	$\mathbf{m} = [x, y, z]^T$	coordinates of l	$\mathbf{l} = [a, b, c]^T$
incidence $m \in l$	$\mathbf{m}^T\mathbf{l} = 0$	incidence $l \ni m$	$\mathbf{l}^T\mathbf{m} = 0$
line obtained by join	$\mathbf{m} \times \mathbf{m}'$	point obtained by meet	$\mathbf{l} \times \mathbf{l}'$
points in join	$\alpha\mathbf{m} + \beta\mathbf{m}'$	pencil containing meet	$\alpha\mathbf{l} + \beta\mathbf{l}'$
collinearity	$\mid \mathbf{m}, \mathbf{m}', \mathbf{m}'' \mid = 0$	concurrence	$\mid \mathbf{l}, \mathbf{l}', \mathbf{l}'' \mid = 0$
points at infinity	$\mathbf{m}_\infty = [x, y, 0]^T$	line at infinity	$\mathbf{l}_\infty = [0, 0, 1]^T$

Table 1.2: Summary of the properties of points and lines in the projective plane.

1.3 2-D and 3-D [Section 4.1.1]

It is quite easy to describe the geometric aspects of image formation for a single image, which is inferring positions of points in one image from positions in the world. In fact, these laws were already understood by the Italian painters of the Renaissance, who studied geometry in order to reproduce correctly the perspective effects in the images of the world that they where observing. Following them, the transformation from the three-dimensional space to the two-dimensional plane performed by a camera can be described using the *pinhole model* (Figure 1.3):

- a plane $\mathcal{R}$, called the *retinal plane*, or *image plane*,
- a point C which does not belong to $\mathcal{R}$: the *optical center*,

The projection m of a point of the space M is the intersection of the *optical ray* (C, M) with the retinal plane. The *optical axis* is the line going through C and

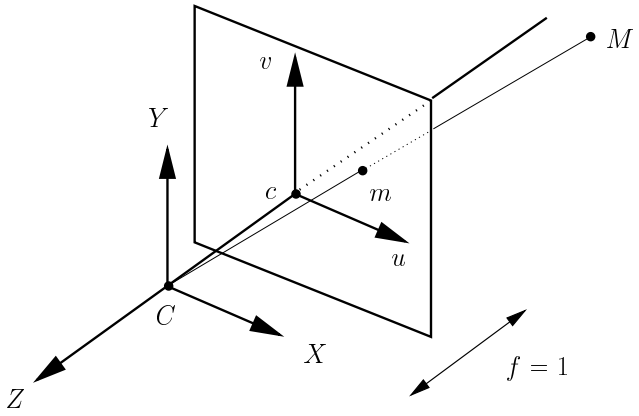

Figure 1.10: The pinhole model expressed in the *camera coordinate system* where the world coordinate system is aligned with the camera and the image coordinate system.

perpendicular to the retinal plane. It pierces that plane at the *principal point* c. If we consider an orthonormal system of coordinates in the retinal plane, centered at c, we can define a three-dimensional orthonormal system of coordinates, called the *camera coordinate system*, centered at the optical center C with two axes of coordinates parallel to the retinal ones and the third one parallel to the optical axis. The *focal length* is the distance between the point C and the plane $\mathcal{R}$. We choose here as unit in the world coordinate system the focal length. Changing this unit corresponds to a simple scaling of the image.

In these two systems of coordinates, the relationship between the coordinates of M, $[X, Y, Z]^T$, and those of its projection m, $[u, v]^T$, is given by Thales theorem:

$$u = \frac{X}{Z} \qquad v = \frac{Y}{Z}. \tag{1.2}$$

Vision is about inferring properties of the world from its images. A central problem of 3-D vision is therefore to invert the projection, which is quite difficult, since one tries to go from a poorer representation (2-D) to a richer representation (3-D). A point m in an image represents a whole incoming light ray, called the *optical ray* of m. By definition, the optical ray contains the optical center, therefore to define its position in 3-D in the camera coordinate system, we just need to specify another point along the ray, say of coordinates $[X, Y, Z]^T$. However, any point of coordinates $[\lambda X, \lambda Y, \lambda Z]^T$ represents the same ray, since both of them are projected to the same 2-D point m. There is an ambiguity along the optical ray, and the consequence of this observation is that using geometry alone we cannot infer the 3-D depth of a point from a single image using geometry alone. This

essential ambiguity is best described by considering $[\lambda X, \lambda Y, \lambda Z]^T$ to be projective coordinates of the optical ray. Because of our choice of unit, $Z = 1$ on the image plane, and therefore the point m of usual coordinates $[u, v]$ has the 3-D coordinates $[u, v, 1]^T$. Projective coordinates of m are $[u, v, 1]^T$, so we see that these projective coordinates represent a point in 3-D on the optical ray of m. This property remains true if another triple of equivalent projective coordinates are used.

Using projective coordinates, the projection equation (1.2) can be written

$$\mathbf{m} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\mathcal{P}_0} \begin{bmatrix} \mathcal{X} \\ \mathcal{Y} \\ \mathcal{Z} \\ \mathcal{T} \end{bmatrix} = \mathcal{P}_0 \mathbf{M}. \tag{1.3}$$

The reward of using projective coordinates is that we have obtained a linear equation instead of a nonlinear one. The usual coordinates are related to projective coordinates by: $u = x/z, v = y/z$ and $X = \mathcal{X}/\mathcal{T}, Y = \mathcal{Y}/\mathcal{T}, Z = \mathcal{Z}/\mathcal{T}$.

Moreover, we can see that the description with projective coordinates is richer than the one with affine coordinates: the points for which $t = 0$ or $\mathcal{T} = 0$ do not have affine correspondents. The points $\mathcal{T} = 0$ are points at infinity (in space), which have been found to be of great utility by such artists-theorists as Piero Della Francesca, Leonardo, and Dürer, when they first formalized perspective projection. As explained in Section 1.2, they are obtained by the intersection of parallel lines and are treated like other points in projective geometry. In particular, they are mapped correctly by the projection matrix producing in general a real vanishing point.

Using projective geometry leads to a simpler, more unified expression of the problem. This make it possible to design more efficient multiple-view algorithms than before. However, the main reward is the ability to deal with a class of problems which couldn't be tackled before, because they depended on *camera calibration*, which we describe next.

The matrix $\mathcal{P}_0$ was particularly simple because of our particular choice of coordinate systems. In general the image coordinate system is defined by the pixels, and the world coordinate system is not aligned with the camera: The general form of the *projection matrix* is

$$\mathcal{P} \simeq \underbrace{\begin{bmatrix} \alpha_u & \gamma & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{A}} \mathcal{P}_0 \underbrace{\begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_3^T & 1 \end{bmatrix}}_{\mathcal{D}} = \mathbf{A}[\mathbf{R}\,\mathbf{t}], \tag{1.4}$$

where

- $\mathbf{A}$ describes the characteristics of the camera or, more precisely, the imaging system. As a 3×3 matrix it represents a change of retinal coordinate system.

Its five entries are called the camera *intrinsic parameters*. α_u and α_v represent the focal length expressed in pixel units in each direction. They describe the total magnification of the imaging system resulting from both optics and image sampling. Their ratio, called the *aspect ratio*, is usually fixed, but is not always equal to 1 due to the digitalization phase. (u_0, v_0) represents the coordinates of the principal point, which usually are not $(0, 0)$ because we count pixels from a corner. The parameter γ, called the skew, is zero except for some very particular imaging situations: non-orthogonal pixels, images of images, and analysis of shadows[3].

- $\mathcal{D}$ describes the location and orientation of the camera with respect to the world coordinate system. It is a 4×4 displacement matrix describing the change of world coordinate system as a rotation $\mathbf{R}$ and a translation $\mathbf{t}$ (the pose of the camera), called the *extrinsic parameters*.

A general projection matrix, being 3×4, depends on eleven parameters (twelve minus a scale factor), which is the number of the intrinsic and extrinsic parameters combined. The decomposition $\mathcal{P} \simeq \mathbf{A}[\mathbf{R}\,\mathbf{t}]$ is unique because of the QR theorem, which states that a non-singular matrix can be factored uniquely as the product of a triangular matrix $\mathbf{A}$ and an orthogonal matrix $\mathbf{R}$.

1.4 Calibrated and uncalibrated capabilities

In the camera coordinate system (the particular coordinate system defined at the beginning of Section 1.3), the projective coordinates of a pixel represent a 3D point on its optical ray, and therefore give us the position of this optical ray in space in the coordinate system of the camera. In general it is not sufficient to measure pixels in order to infer from a pixel m the position of the optical ray in space. The matrix $\mathbf{A}$ is used to transform pixel coordinates into camera coordinates. A camera for which $\mathbf{A}$ is known is said to be *calibrated*. It then acts as a metric measurement device, able to measure the angle between optical rays. Furthermore, if $\mathcal{D}$ is known, then it is possible to relate the camera coordinate system to the world's or other camera's coordinate systems.

The classical (model-based) way to calibrate a camera is by determining its projection matrix using known control points in 3D. Let $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$ represent the three row vectors of $\mathcal{P}$. For each correspondence $m \leftrightarrow M$ from 2-D to 3-D, we

[3] The two latter situations are adequately described by the full projection matrix because the product of two perspective projections, although not always a perspective transformation, is always a projective transformation. Similarly, the product of two perspective projections with a particular change of retinal coordinates (for example an orthogonal one) is not necessarily a perspective transformation with the same particular change of coordinates, but is always a projective transformation.

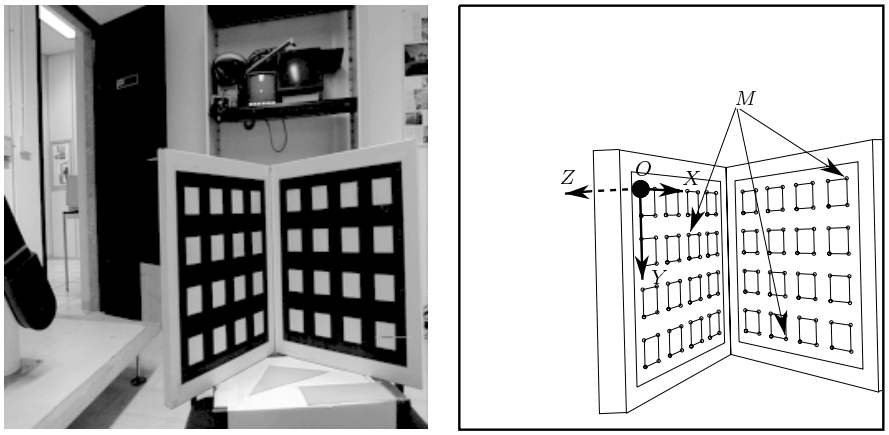

Figure 1.11: The calibration grid used at INRIA and the associated model.

obtain two linear equations in the entries of $\mathcal{P}$:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{m} \simeq \mathcal{P}\mathbf{M} = \begin{bmatrix} \mathbf{U}^T\mathbf{M} \\ \mathbf{V}^T\mathbf{M} \\ \mathbf{W}^T\mathbf{M} \end{bmatrix}; \quad \text{therefore} \quad \begin{cases} u\mathbf{W}^T\mathbf{M} - \mathbf{U}^T\mathbf{M} = 0, \\ v\mathbf{W}^T\mathbf{M} - \mathbf{V}^T\mathbf{M} = 0. \end{cases}$$

The reference points M are measured in some 3-D coordinate frame, and their projections m detected. Usually a special object, like the one shown in Figure 1.11, is engineered so that both operations can be done with a good accuracy [Section 4.6]. Because $\mathcal{P}$ has 11 independent entries, from at least six 2-D to 3-D correspondences in general position it is possible to determine the projection matrix. Once $\mathcal{P}$ is known, it can be decomposed back into $\mathbf{A}$ and $\mathcal{D}$, which are the basis for 3-D measurements from images.

Model-based calibration is not always possible. First, many images such as those available in image libraries or from hand-held video come without calibration data at all or with calibration data which is imprecise (such as the reading of the focal length on the barrel of a zoom lens). Second, even if we have calibrated a system, the calibration data might change either because of involuntary reasons such as mechanical or thermal variations, or because of active controls such as focus or vergence, which add greatly to the flexibility and adaptivity of a vision system. While the first computer vision applications used robotic systems which could be pre-calibrated off-line, the trend today is towards the use of massive and ubiquitous imagery from all kinds of sources.

Because the usual world is Euclidean, a projective framework might seem at first unnecessarily abstract and complicated, but besides allowing us to understand and express the geometry of the problem in a much simpler way, it makes it possible to

consider the world as a redundant superposition of projective, affine, and Euclidean structures, and to deal with these three structures simultaneously. This better understanding of the problem makes it possible:

- To propose linear or analytical approaches, using minimal data if necessary, to problems (such as the bundle adjustment, [Section 10.2]) which have in the past been dealt with only by large-scale minimization, with all the associated global convergence problems. In particular, the previous methods require a precise initialization point which the projective methods can provide.
- To characterize those configurations of points and of cameras which cause degeneracy and instability in the estimation process, and more generally, to improve the stability, robustness and precision of the estimation.

Thanks to the projective approach, we will be able to achieve the same metric results as model-based calibration in many circumstances without the need to use a calibration object or known reference points. For example, we can perform the following photogrammetric tasks.

- Using only constraints such as instances of parallelism and orthogonality in the scene, we obtain a metrically correct reconstruction of a scene (up to a global scale factor) from an arbitrary number of images taken by arbitrarily different cameras [Section 7.3 and 7.4].
- Using only simple constraints about a moving camera (such as zero-skew or constant aspect ratio), we track the 3-D motion and recover the internal parameters of this camera even when they vary over time [Section 11.4 and 11.5].

In a complementary way, we will see that for many applications such as

- the navigation and obstacle avoidance for an autonomous robot or vehicle [Section 7.2.6] and the detection of independent motion, and
- the synthesis of novel images from existing images [Section 7.2.7 and Section 8.1.1],

there is not even a need for metric representations or 3-D reconstruction. Instead, a non-metric description is more general, easier to obtain, and can capture more precisely the properties which are relevant to a given task.

1.5 The plane-to-image homography as a projective transformation [Section 4.1.4]

We begin by introducing an important example of projective transformation, the *homography* between a plane Π in space and the retinal plane.

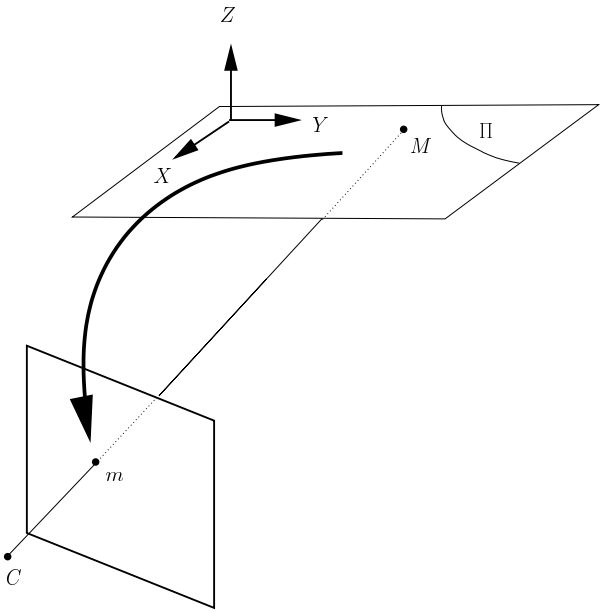

Figure 1.12: The homography between a plane in space and the retinal plane. The world coordinate system is aligned with plane Π.

If we choose the world coordinate system so that the first two axes define the plane, as illustrated in Figure 1.5, the projection of points of Π can be viewed as a transformation between two spaces $\mathbb{P}^2$, since for those points

$$\underbrace{\begin{bmatrix} x \\ y \\ z \end{bmatrix}}_{\mathbf{m}} \simeq \boldsymbol{\mathcal{P}} \begin{bmatrix} \mathcal{X} \\ \mathcal{Y} \\ 0 \\ \mathcal{T} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathcal{P}_{11} & \mathcal{P}_{12} & \mathcal{P}_{14} \\ \mathcal{P}_{21} & \mathcal{P}_{22} & \mathcal{P}_{24} \\ \mathcal{P}_{31} & \mathcal{P}_{32} & \mathcal{P}_{34} \end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix} \mathcal{X} \\ \mathcal{Y} \\ \mathcal{T} \end{bmatrix}}_{\mathbf{p}}.$$

Each point correspondence (m, p) yields two independent proportionality equations:

$$\frac{x}{z} = \frac{h_{11}\mathcal{X} + h_{12}\mathcal{Y} + h_{13}\mathcal{T}}{h_{31}\mathcal{X} + h_{32}\mathcal{Y} + h_{33}\mathcal{T}}, \quad \frac{y}{z} = \frac{h_{21}\mathcal{X} + h_{22}\mathcal{Y} + h_{23}\mathcal{T}}{h_{31}\mathcal{X} + h_{32}\mathcal{Y} + h_{33}\mathcal{T}},$$

which can be linearized in the entries of $\mathbf{H}$:

$$\begin{cases} h_{11}z\mathcal{X} + h_{12}z\mathcal{Y} + h_{13}z\mathcal{T} - h_{31}x\mathcal{X} - h_{32}x\mathcal{Y} - h_{33}x\mathcal{T} = 0, \\ h_{21}z\mathcal{X} + h_{22}z\mathcal{Y} + h_{23}z\mathcal{T} - h_{31}y\mathcal{X} - h_{32}y\mathcal{Y} - h_{33}y\mathcal{T} = 0. \end{cases}$$

$\mathbf{H}$ has eight entries (nine minus a scale factor), therefore from four correspondences $\mathbf{m}, \mathbf{p}$ in general position, $\mathbf{H}$ is determined uniquely by solving a linear system of

equations. Here "general position" means that no three points are collinear, because if that was the case the equations would not be linearly independent. Once $\mathbf{H}$ is computed, we can use it to determine positions of points on Π from a *single* image. This simple example illustrates the power of projective geometry: the mapping between the two planes is done using just linear operations and four reference points, without the need to refer to more complicated representations like rotations, translations and camera parameters.

The transformation $\mathbf{H}$ is called a *homography*, or *projective transformation*[4] of $\mathbb{P}^2$. Generally speaking, a homography is any transformation H of $\mathbb{P}^n$ which is linear in projective coordinates (hence the terminology *linear projective*) and invertible (thus it conserves globally the space, a fact we denote: $H(\mathbb{P}^n) = \mathbb{P}^n$). It can be shown that these properties are equivalent to the fact that collinearity is preserved and subspaces mapped into subspaces of the same dimension. A homography can be described by an $(n+1) \times (n+1)$ non-singular matrix $\mathbf{H}$, such that the image of $\mathbf{x}$ is $\mathbf{x}'$:

$$\mathbf{x}' \simeq \mathbf{H}\mathbf{x}.$$

Like in $\mathbb{P}^2$ we needed four corresponding points in general position to define a homography, in $\mathbb{P}^n$ we need two sets of $n+2$ points such that no $n+1$ of them are linearly dependent to define a homography. Each such set is called a *projective basis*, and it corresponds to the choice of a projective coordinate system.

1.6 Affine description of the projection [Section 4.2]

The projection matrix $\mathcal{P}$ has to be of rank 3, otherwise its image would be a projective line instead of a projective plane. Since it has 4 columns, its nullspace is thus of dimension 1; any vector $\mathbf{C}$ of this nullspace defines a projective point C for which the projection is not defined; this point is the optical center.

Let us now partition the projection matrix $\mathcal{P}$ into the concatenation of a 3×3 sub-matrix $\mathbf{P}$ and a 3×1 vector $\mathbf{p}$. The origin of the world coordinate system, the point $[0, 0, 0, 1]$, is projected onto $\mathbf{p}$.

The optical center is also decomposed by separating its last coordinate from the first three:

$$\mathcal{P} = [\mathbf{P}\ \mathbf{p}], \qquad \tilde{C} = \begin{bmatrix} \mathbf{C} \\ c \end{bmatrix}. \tag{1.5}$$

The equation determining the optical center is $\mathcal{P}\tilde{\mathbf{C}} = \mathbf{0}$. Using the decomposition just introduced, $\mathcal{P}\tilde{\mathbf{C}} = \mathbf{P}\mathbf{C} + \mathbf{p}c$, thus $\mathbf{P}\mathbf{C} = -c\mathbf{p}$. Therefore, if $\det(\mathbf{P}) \neq 0$,

[4] A *perspective* transformation of $\mathbb{P}^2$ is obtained by using a projection matrix in camera coordinates, or in other words, such that its first 3×3 sub-matrix is orthogonal. Unlike projective transformations, perspective transformations do not form a group: the product of two perspective transformations is not necessary a perspective transformation.

then the solution is given by

$$\tilde{\mathbf{C}} \simeq \begin{bmatrix} -\mathbf{P}^{-1}\mathbf{p} \\ 1 \end{bmatrix} \tag{1.6}$$

so the optical center is finite. When $\det(\mathbf{P}) = 0$, it can be verified, using the fact that $\mathcal{P}$ has rank 3, that the optical center lies in the plane at infinity (i.e. $c = 0$).

Any projection matrix arising from a physical system has to satisfy $\det(\mathbf{P}) \neq 0$, since the optical center has to lie in the affine space (we refer to that as *perspective projection*). For simplicity, we will assume that this is the case in this introductory Chapter. The alternative class of models (*parallel projection*) can be considered as approximations to the pinhole model in some particular viewing situations. These include orthographic, weak perspective, and the affine camera [Section 4.4]. They yield a simpler geometry which is affine instead of being projective. The beauty of the projective model is that it handles perspective cameras and parallel cameras equally well. There is no need to distinguish between the two cases. However, it leaves open the possibility to specialize the analysis, which we do now.

At this affine level of description, we can introduce *directions* of optical rays. Since the projection of each point at infinity $[\mathbf{d}^T, 0]^T$ is the vanishing point $\mathbf{v} = \mathbf{P}\mathbf{d}$, $\mathbf{P}$ can be considered as the homography between the plane at infinity Π_∞ and the retinal plane $\mathcal{R}$. Note that parallel lines have the same direction, hence the same point at infinity, thus their projection is a set of lines of $\mathcal{R}$ which contains the vanishing point projection of this point at infinity. The optical ray corresponding to the pixel $\mathbf{m}$ thus has the direction $\mathbf{P}^{-1}\mathbf{m}$. This is illustrated in Figure 1.6.

From the decomposition Equation 1.4, it can be noticed that $\mathbf{P}$ depends only on the orientation of the camera and its intrinsic parameters, not on its position. Therefore we obtain the fact that we had pointed to at the beginning of this Chapter, that the projection of points at infinity is invariant under translation. The dependence of the finite points on the translation is embodied in the vector $\mathbf{p}$, which represents the projection of the origin of the world coordinate system.

Table 1.3 summarizes the descriptions of the projection matrix in the perspective projection case.

1.7 Structure and motion

Let us now add a second image. Two points, m in the first image, and m' in the second image, are said to be *corresponding* if they are the projections of the same 3-D point M in space. Having more than one image opens new possibilities and raises the following questions:

- Given a point m in the first image, where is its corresponding point m' in the second image?
- What is the 3-D geometry of the scene?

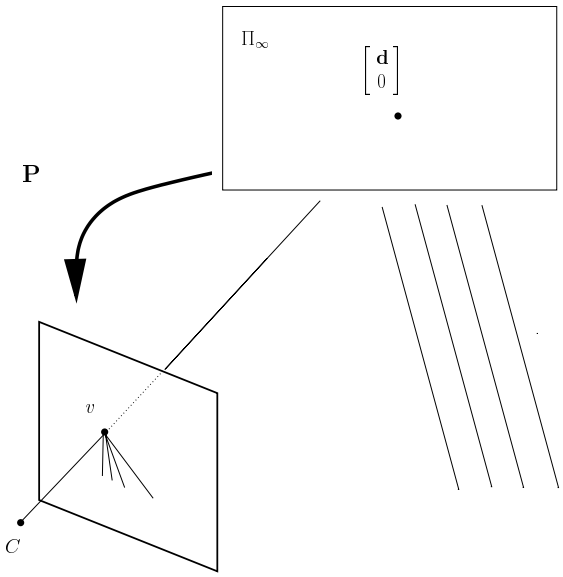

Figure 1.13: The first submatrix **P** of the projection matrix represents a homography between the plane at infinity and the retinal plane. It projects intersections of parallel lines in space to a vanishing point in the image. Its inverse gives the direction of the optical ray.

- What is the relative position of the two cameras?

Note there are several ways of representing the 3-D geometry of a scene. We could recover the depth of a point which is its distance to the image plane, we could recover the 3-D coordinates of a point, or we could recover the relative depths of two points.

We have seen that from a single image, even if we know the parameters of the camera model, we can infer only the position of the optical ray of m, not the position of the 3-D point M. With two images, given the correspondence (m, m'), we can intersect the optical rays of m and m', and so determine M. This is the principle of binocular stereopsis: when two images are taken from different positions, the difference in position of corresponding image points is related to the 3-D position of the object point. To actually infer that 3-D position requires that we can infer the position of the two optical rays in a common coordinate system. We need to know the relative position of the second camera with respect to the first one, which we call its *motion*. Algebraically, if we knew the projection matrices $\mathcal{P}$ and $\mathcal{P}'$, then we could compute M from m and m' by solving the system of four equations (each

level	decomposition	interpretation	
projective	$\mathcal{P}$	$\mathcal{P}$:	projection from object space $\mathbb{P}^3$ to retinal plane $\mathcal{R}$.
affine	$[\mathbf{P}\,\mathbf{p}]$	$\mathbf{P}$:	homography between plane at infinity Π_∞ and $\mathcal{R}$.
		$\mathbf{p}$:	projection of the origin of the world coordinate system.
Euclidean	$\mathbf{A}[\mathbf{R}\,\mathbf{t}]$	$\mathbf{A}$:	change of coordinates in $\mathcal{R}$ (5 intrinsic parameters).
		$(\mathbf{R}, \mathbf{t})$:	camera pose in world coordinates.

Table 1.3: Descriptions of the projection matrix.

proportionality vector equation gives two independent equations):

$$\left\{ \begin{array}{rcl} \mathcal{P}\mathbf{M} & \simeq & \mathbf{m}, \\ \mathcal{P}'\mathbf{M} & \simeq & \mathbf{m}'. \end{array} \right. \tag{1.7}$$

Therefore, in order to be able to determine the 3-D *structure* of the scene, we also need to be able to determine the projection matrices $\mathcal{P}$ and $\mathcal{P}'$ which encode the geometry of the cameras. The two problems of motion determination and structure determination are inter-related, and we will designate them collectively as the *reconstruction problem*.

In the system of equations (1.7), we notice that we have three unknowns (the coordinates of M) and four equations. For a solution to exist, the coordinates of m and m' must satisfy a constraint; in other words, given m, the point m' cannot be an arbitrary point of the second image. In fact, in some particular cases that we are going to examine in Section 1.8 and Section 1.9, it is possible to predict the position of the point m' from the position of the point m.

1.8 The homography between two images of a plane [Section 5.1.1]

We first examine the particular situation when the 3-D points lie on a plane Π. Planes are important entities: in practice because they appear naturally in many scenes and in theory because they are subspaces which have the same dimension as the images. As we have seen in Section 1.5, there is a homography between the retinal plane of the first camera and the plane Π and also a homography between the retinal plane of the second camera and the plane Π; therefore by composition there

is a homography H between the two retinal planes called a *planar homography*, because it is induced by the plane Π and described by a 3×3 matrix $\mathbf{H}$. This homography is illustrated in Figure 1.8. If m and m' are projections of a point M which belongs to Π, then

$$\mathbf{m}' \simeq \mathbf{H}\mathbf{m}.$$

Reversing the roles of the two images transforms $\mathbf{H}$ into its inverse. As in Section 1.5, $\mathbf{H}$ can be determined in general from 4 correspondences. Once $\mathbf{H}$ is known, for any projection m of a point of Π, it is possible to predict the position of its correspondent in the other image. Some care must be taken if Π goes through either of the two optical centers [Section 5.1.1].

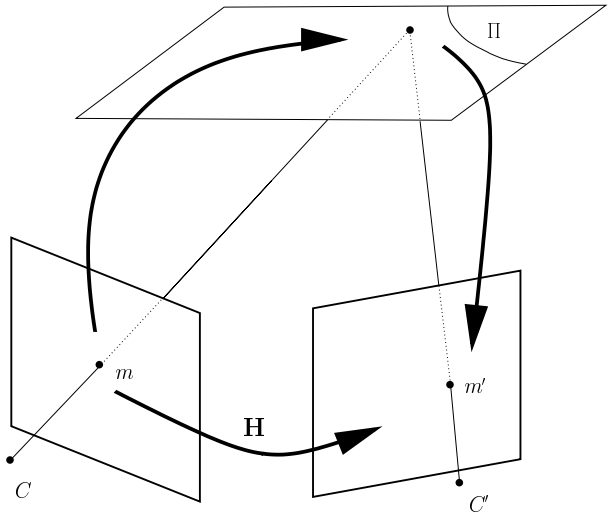

Figure 1.14: The planar homography between two images is obtained by composition of single view homographies.

An important special case occurs when Π is the plane at infinity Π_∞. Then, $\mathbf{H}_\infty$ has a particularly simple expression in terms of the two projection matrices $\mathcal{P}$ and $\mathcal{P}'$, obtained using the decomposition in Section 1.6:

$$\mathbf{H}_\infty \simeq \mathbf{P}'\mathbf{P}^{-1}. \tag{1.8}$$

1.9 Stationary cameras [Section 5.1.2]

A related situation occurs when the two optical centers are identical, i.e. when the two images are taken from the same viewpoint with the camera rotated. Let m and m' be arbitrary corresponding points, i.e. the images of a point in the scene. For

any plane Π, not going through the optical center, they are also the images of the point of intersection of their common optical ray with the plane. They are therefore related by $\mathbf{H}$, which in this case is independent of the plane Π. The homography $\mathbf{H}$ can be used to merge the images and to build image mosaics, as illustrated in Figure 1.15. By applying it to the whole first image, we transform it into a new image which, when overlaid on the second image, forms a larger image representing the scene as it would have appeared from the point of view of the second image but with a larger field of view. If we have several images taken from the same viewpoint, we can iterate the process of choosing one of the images as the reference image and warping all of the other images onto it by applying the corresponding homographies. Note that in this process, we have begun to address the problem of generating new views from existing images: by applying a homography to a single image, we can generate a new view obtained from the same viewpoint but with a rotation of the camera. This process does not handle translation of the camera to a new viewpoint.

Two images taken from the same viewpoint cannot be used to recover the 3-D structure: since the optical ray of two corresponding points is the same, the ambiguity along this ray remains, just like when we have a single image. Therefore, there must be a non-null translational component of the motion for us to be able to recover structure from two images. In the reminder of this section, we assume that the points do not lie on a plane and that the optical centers are distinct.

1.10 The epipolar constraint between corresponding points [Section 5.2.1]

When the points in space and the two cameras are in general position, it is not possible to predict the position of the correspondent m' of a point m, because this position depends on the depth of the 3-D point M along the optical ray. However, geometrically, this position is not arbitrary: M has to lie along the optical ray of m, and therefore m' is necessarily located on the projection of that optical ray in the second camera. This line is called the *epipolar line* of the point m in the second image. See Figure 1.10. If we are able to compute this line, then when looking for the correspondent of m, we need not search the whole second image, but only this line, hence reducing the search space from 2-D to 1-D.

There is another way to view the same construction, by considering it as a way to constrain the cameras rather than the correspondence. Assuming that we know a valid correspondence, $m \leftrightarrow m'$, the relative position of the cameras must be such that optical rays L_m and $L_{m'}$ intersect. Another way to formulate this is to say that the two optical rays and the line between the optical centers (called the *baseline*) are coplanar. The common plane is called the *epipolar plane*.

Algebraically, because the point M depends on three coordinates, and the correspondence $m \leftrightarrow m'$ depends on a total of four parameters, there must be an

Figure 1.15: Two images taken from the same viewpoint, and the composite image obtained by applying a homography to the second image and superimposing it to the first image.

algebraic constraint linking the coordinates of m and m'. We will see next that this constraint is remarkably simple.

1.11 The Fundamental matrix [Section 5.2.1]

The relationship between the point m and its epipolar line l'_m in the second image is projective linear, since the optical ray of m is a linear function of m, and projection is also linear. Therefore, there is a 3×3 matrix which describes this correspondence, called the *Fundamental matrix*, giving the epipolar line of the point m: $\mathbf{l}'_m = \mathbf{F}\mathbf{m}$. If two points m and m' are in correspondence, then the point m' belongs to the

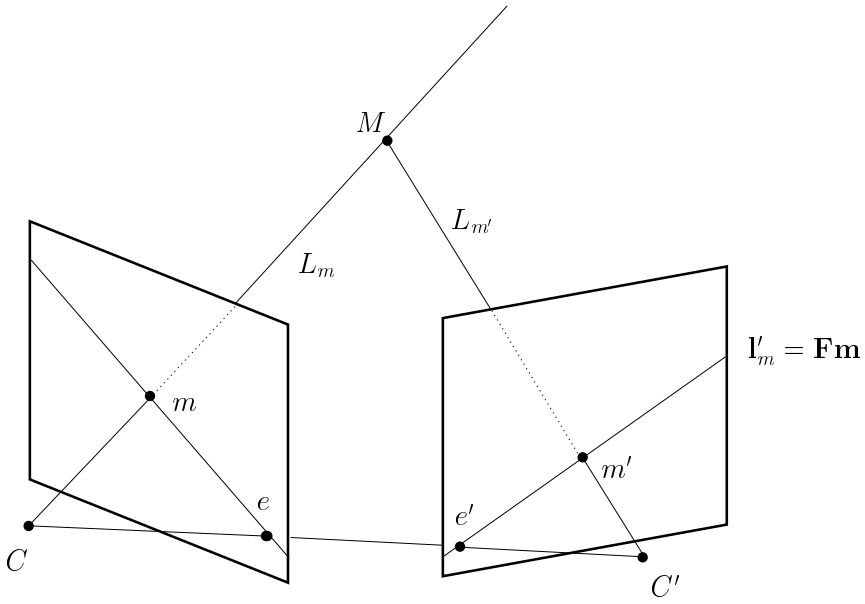

Figure 1.16: Epipolar geometry. Given m, the point m' has to lie on its the epipolar line l'_m. Given a valid correspondence $m \leftrightarrow m'$, the intersection of the optical rays L_m and $L_{m'}$ is not empty, and they are coplanar with the baseline CC'.

epipolar line of m, therefore they satisfy the *epipolar constraint*:

$$\mathbf{m'}^T \mathbf{F} \mathbf{m} = 0, \tag{1.9}$$

which is bilinear in the coordinates of the image points. Reversing the roles of the two images transforms $\mathbf{F}$ into its transpose. Figure 1.11 shows two images and a few epipolar lines.

The Fundamental matrix depends only on the configuration of the cameras (intrinsic parameters, position and orientation) and not on the 3-D points in the scene. In the generic case where we do not assume any spatial relationship between the points in space, the only information available comes from *projective correspondences*, the correspondence of points undergoing linear projection. The epipolar constraint fully describes the correspondence of a pair of generic corresponding points in each image, the ambiguity along the epipolar line being caused by the ambiguity along the optical ray of the projection operation. Since the fundamental matrix depends only on camera geometry, it describes all the epipolar constraints, so it encodes all the information available from projective correspondences. Therefore, no other general constraint is available. This will be confirmed later, with a parameter counting argument, when we will see that it is sufficient to build a 3-D

Figure 1.17: Two images with a few corresponding points and epipolar lines.

reconstruction of points and cameras compatible with projective correspondence.

Since all the optical rays contain the optical center C of the first camera, all the epipolar lines contain the projection of C in the second image (the point where the first camera is seen by the second camera), called the *epipole*. See Figure 1.10. The fact that the epipole in the second image belongs to all the epipolar lines implies $\mathbf{e}'^T\mathbf{F}\mathbf{m} = 0$ for any m, and therefore $\mathbf{e}'^T\mathbf{F} = \mathbf{0}$, or equivalently, $\mathbf{F}^T\mathbf{e}' = 0$. By reversing the role of the two images, $\mathbf{F}\mathbf{e} = 0$. We conclude that $\mathbf{F}$ is a matrix of rank two:

$$\det(\mathbf{F}) = 0.$$

Since it satisfies this algebraic constraint and is only defined up to a scale factor (like all the projective quantities), $\mathbf{F}$ depends on seven parameters.

1.12 Computing the Fundamental matrix [Chapter 6]

Each point correspondence (m, m') yields one equation (1.9), therefore with a sufficient number of point correspondences in general position we can determine $\mathbf{F}$. No knowledge about the cameras or scene structure is necessary. The first step for all the algorithms that we discuss in the book is almost always the computation of the Fundamental matrix, which is of utmost theoretical and practical importance.

Equation 1.9 is linear in the entries of $\mathbf{F}$. It can be rewritten as

$$\mathbf{U}^T\mathbf{f} = 0,$$

where $\mathbf{m} = [u, v, 1]^T$ and $\mathbf{m}' = [u', v', 1]^T$ so

$$\begin{aligned} \mathbf{U} &= [uu', vu', u', uv', vv', v', u, v, 1]^T, \\ \mathbf{f} &= [F_{11}, F_{12}, F_{13}, F_{21}, F_{22}, F_{23}, F_{31}, F_{32}, F_{33}]^T. \end{aligned}$$

Combining the rows $\mathbf{U}$ for each correspondence provides a linear system of the form $\tilde{\mathbf{U}}\mathbf{f} = 0$. Using seven points, it is possible to compute $\mathbf{F}$ using the rank constraint

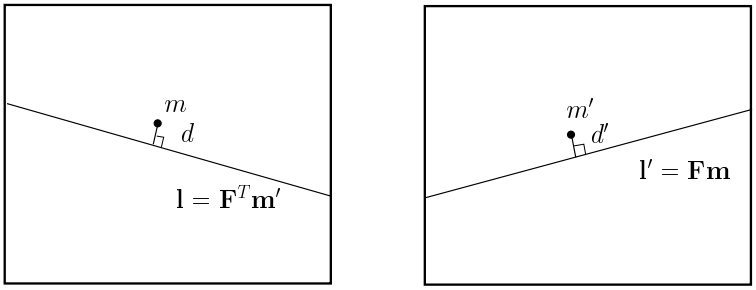

Figure 1.18: Error function used for the computation of the Fundamental matrix: sum of epipolar distances $d^2 + d'^2$.

$\det(\mathbf{F}) = 0$, however because this constraint is cubic there can be three solutions. With eight correspondences in general position, there is a unique solution which is obtained linearly. In practice, we have more than eight correspondences, but they are not exact, so we can seek a least-squares solution:

$$\min_{\mathbf{f}} \|\tilde{\mathbf{U}}\mathbf{f}\| \quad \text{subject to} \quad \|\mathbf{f}\| = 1. \tag{1.10}$$

The constraint $\|\mathbf{f}\| = 1$ is necessary because $\mathbf{F}$ is defined up to a scale factor.

This approach suffers from two difficulties. First, unlike the case of seven points, we notice that the rank constraint is no longer satisfied. Second, the error function in Equation 1.10 was obtained algebraically but has no geometric relevance. However, this approach can give acceptable results if care is taken in renormalizing the pixel coordinates to the interval $[-1, 1]$ to improve the numerical conditioning of matrix $\tilde{\mathbf{U}}$. It has the advantage of simplicity. Practice has shown that the most precise results with noisy data are obtained by using nonlinear minimization techniques which rely on a symmetrized geometric error criterion and enforce the rank constraint by an adequate parameterization. A proven such approach is to minimize the error function illustrated in Figure 1.12:

$$\sum_i \{d(\mathbf{m}'_i, \mathbf{F}\mathbf{m}_i)^2 + d(\mathbf{m}_i, \mathbf{F}^T\mathbf{m}'_i)^2\},$$

where $d(.,.)$ is the perpendicular distance of a point to a line. In practice, it is also important to use robust techniques to reject false correspondences. All these algorithmic refinements are the subject of Chapter 6.

Although eight correspondences are in general sufficient to determine the Fundamental matrix, there are some configuration of 3-D points, called *critical surfaces* for which even with an arbitrarily large number of correspondences, the Fundamental matrix is not uniquely determined. These configurations are important in practice because if the scene is close to such a configuration, the determination of

the Fundamental matrix is quite unstable. Planes are a particular case of critical surfaces, which are quadrics examined in more detail in Section 5.5.

1.13 Planar homographies and the Fundamental matrix [Section 5.2.4]

We have seen in Section 1.8 that for a given plane Π, the correspondence is entirely determined by a planar homography $\mathbf{H}$; in other words, $\mathbf{H}$ can be used to compute corresponding points in the second image from points in the first image. We have just seen that for points in general 3-D position, the Fundamental matrix can be used to constrain correspondences along one direction, that of the epipolar line. There is an important relation between these two matrices, the planar homography H and the Fundamental matrix. As will be seen later, this relation is at the heart of techniques for positioning 3-D points in space from their projections.

Given the two cameras, and therefore the Fundamental matrix, a planar homography is defined by its associated plane. Since a plane depends on three parameters, and a homography on eight parameters, not all 3×3 invertible matrices define a planar homography, so $\mathbf{H}$ must satisfy six constraints, given $\mathbf{F}$. On the other hand, the planar homography constrains the Fundamental matrix because it can be used to generate a point on the epipolar line of any point: if m is a point of the first image, then its optical ray intersects the plane Π at M_Π. $\mathbf{Hm}$ represents the projection of M_Π into the second image. Since by construction the point M_Π belongs to the optical ray of m, the point $\mathbf{Hm}$ belongs to the epipolar line of m. Therefore, given $\mathbf{H}$, it is sufficient to know the epipole e' to determine the Fundamental matrix: the epipolar line l'_m contains $\mathbf{Hm}$ and the epipole e', therefore, $\mathbf{l}'_m = \mathbf{e}' \times \mathbf{Hm}$. Since by definition of $\mathbf{F}$, $\mathbf{l}'_m = \mathbf{Fm}$, we conclude that

$$\mathbf{F} \simeq [\mathbf{e}']_\times \mathbf{H} \tag{1.11}$$

This is illustrated in Figure 1.13. Conversely, it can be shown that any matrix $\mathbf{H}$ which satisfies this constraint is a planar homography generated by some plane.

Applying both sides of Equation (1.11) to the vector $\mathbf{e}$, and using the fact that $\mathbf{F}e = 0$ and $[\mathbf{e}']_\times \mathbf{e}' = 0$ shows that

$$\mathbf{He} \simeq \mathbf{e}';$$

therefore, once the Fundamental matrix is known, the correspondence of three points is sufficient to define a planar homography[5] since the correspondence (e, e') provides the needed fourth point.

The decomposition (1.11) of the Fundamental matrix is not unique, since it is obtained for any planar homography. Considering two planar homographies $\mathbf{H}_1$ and $\mathbf{H}_2$, Equation 1.11 implies that there exist scalars λ_1 and λ_2 such that $[\mathbf{e}']_\times(\lambda_1\mathbf{H}_1 +$

[5] Again, the case where the plane goes through either optical center is special; see Chapter 5.

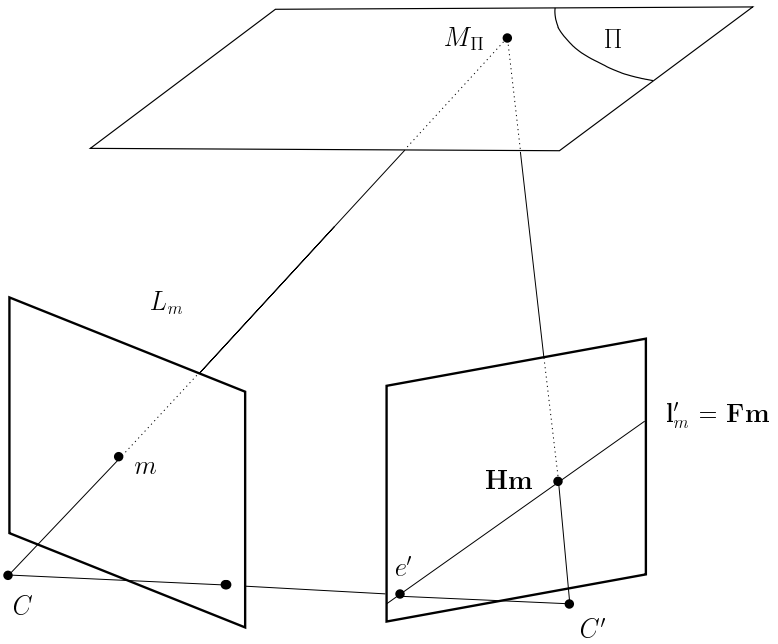

Figure 1.19: Relation between the Fundamental matrix and a planar homography. The points $\mathbf{Hm}$ and e' define the epipolar line of m.

$\lambda_2\mathbf{H}_2) = 0$. It can be shown by inspection that if a matrix $\mathbf{H}$ is such that $[\mathbf{e}']_\times \mathbf{H} = 0$, then there exists a vector $\mathbf{r}$ such that $\mathbf{H} = \mathbf{e}'\mathbf{r}^T$. We therefore conclude that

$$\mathbf{H}_2 \simeq \mathbf{H}_1 + \mathbf{e}'\mathbf{r}^T. \tag{1.12}$$

This equation can be understood geometrically by applying both of its terms to the point m of the first image. Because $\mathbf{r}^T\mathbf{m}$ is a scalar, it says that the point of coordinates $\mathbf{H}_2\mathbf{m}$ in the second image belongs to the line defined by e' and the point of coordinates $\mathbf{H}_1\mathbf{m}$. This is true because, as discussed in the derivation of Equation 1.11, it is the epipolar line of m. In fact, the direction of the vector $\mathbf{r}$ represents the projection in the first image of the intersection of the planes corresponding respectively to $\mathbf{H}_1$ and to $\mathbf{H}_2$. To see that, note that a point belongs to both planes if and only if $\mathbf{H}_2\mathbf{m} \simeq \mathbf{H}_1\mathbf{m}$, and therefore $\mathbf{r}^T\mathbf{m} = 0$. This is illustrated in Figure 1.13. The consequence of this equation is that the family of planar homographies is parameterized by a vector of dimension three, which is expected since the family of planes of $\mathbb{P}^3$ has this dimension. Knowing any of these homographies make it possible to generate all of them.

At this point, one could wonder how to obtain a planar transformation without relying on some prior knowledge of the scene to identify a plane. The trick is that

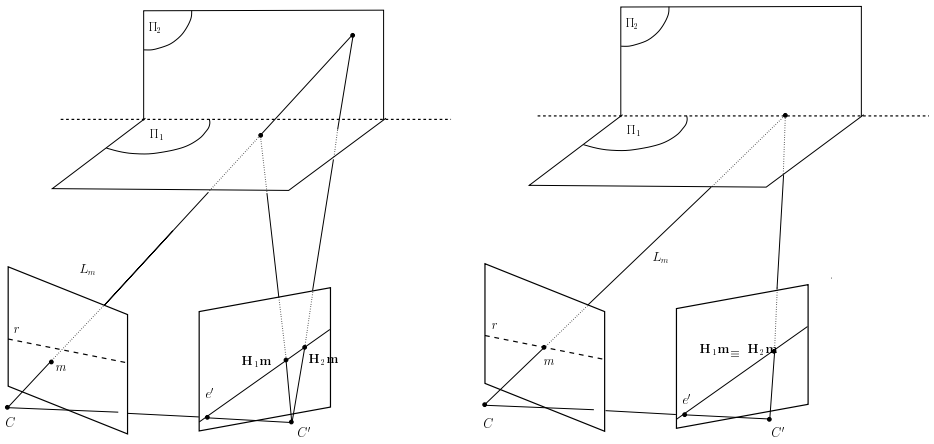

Figure 1.20: Relation between two planar homographies. Left: generic point. The points e', $\mathbf{H}_1\mathbf{m}$, and $\mathbf{H}_2\mathbf{m}$ are aligned. Right: L_m intersect the line $\Pi_1 \cap \Pi_2$. $\mathbf{H}_1\mathbf{m}$, and $\mathbf{H}_2\mathbf{m}$ are identical. m lies on the line r, projection of $\Pi_1 \cap \Pi_2$.

from just $\mathbf{F}$, we can always obtain one by using the *special matrix* defined as

$$\mathbf{S} = [\mathbf{e}']_\times \mathbf{F}. \tag{1.13}$$

It can be verified that $\mathbf{S}$ satisfies $\mathbf{F} \simeq [\mathbf{e}']_\times \mathbf{S}$, using the identity $\mathbf{I} \simeq \mathbf{v}\mathbf{v}^T - [\mathbf{v}]_\times [\mathbf{v}]_\times$; therefore it is equivalent to know $\mathbf{F}$ or $\mathbf{S}$. The small price to pay for this "free" planar transformation is that it is singular, since the two matrices which appear in Equation 1.13 are singular. It can be shown that the plane generating $\mathbf{S}$ is through the optical center C' (hence the singularity) and projects to the line of equation $\mathbf{e}'$ in the second image. Because it is attached to the system of two cameras, it is called the *intrinsic plane* Π_S.

We end this section by giving an expression of the Fundamental matrix as a function of the projection matrices, which can can be used on a calibrated stereo rig to guide the correspondence process by limiting it to epipolar lines. We write the two projection matrices $\mathcal{P} = [\mathbf{P}\ \mathbf{p}]$ and $\mathcal{P}' = [\mathbf{P}'\ \mathbf{p}']$. We place ourselves in the perspective case (as opposed to the parallel case, see Section 1.6) by assuming $\det(\mathbf{P}) \neq 0$. The epipole $\mathbf{e}'$ is the projection $\mathcal{P}'\tilde{\mathbf{C}}$ of the optical center $\tilde{\mathbf{C}}$ given by Equation 1.6; therefore

$$\mathbf{e}' \simeq \mathcal{P}'\tilde{\mathbf{C}} \simeq [\mathbf{P}'\,\mathbf{p}'] \begin{bmatrix} -\mathbf{P}^{-1}\mathbf{p} \\ 1 \end{bmatrix} \simeq \mathbf{p}' - \mathbf{P}'\mathbf{P}^{-1}\mathbf{p}.$$

Now that we know the epipole, to apply Equation 1.11 in order to determine $\mathbf{F}$, we need only a planar homography between the two images. One such homography is the one induced by the plane at infinity, given in Equation 1.8. Using this, we

obtain an expression of the Fundamental matrix as a function of the projection matrices:

$$\mathbf{F} = [\mathbf{p}' - \mathbf{P}'\mathbf{P}^{-1}\mathbf{p}]_\times \mathbf{P}'\mathbf{P}^{-1}$$

1.14 A stratified approach to reconstruction

The reconstruction problem can be stated as that of determining the projection matrices $\mathcal{P}$ and $\mathcal{P}'$, as well as the 3-D points M_i, given a set of N correspondences (m_i, m_i'). The solution is not unique because it depends on the choice of a coordinate system, expressed by the 4×4 matrix $\mathcal{H}$. If $(\mathcal{P}, \mathcal{P}', \mathbf{M}_1, \cdots, \mathbf{M}_N)$ is a solution to the reconstruction problem, then $(\mathcal{P}\mathcal{H}^{-1}, \mathcal{P}'\mathcal{H}^{-1}, \mathcal{H}\mathbf{M}_1, \cdots, \mathcal{H}\mathbf{M}_N)$ is also a solution, since

$$\left\{ \begin{array}{lcll} \mathbf{m} & \simeq & \mathcal{P}\mathbf{M} & = (\mathcal{P}\mathcal{H}^{-1})(\mathcal{H}\mathbf{M}), \\ \mathbf{m}' & \simeq & \mathcal{P}'\mathbf{M} & = (\mathcal{P}'\mathcal{H}^{-1})(\mathcal{H}\mathbf{M}). \end{array} \right. \tag{1.14}$$

In other words, all the pairs of projection matrices of the form $(\mathcal{P}\mathcal{H}, \mathcal{P}'\mathcal{H})$, where $\mathcal{H}$ is an arbitrary projective transformation, are potentially equivalent. However, if we have some constraints about the correspondences, then we can hope to limit the ambiguity $\mathcal{H}$ by enforcing that these constraints have to be satisfied by the pair $(\mathcal{P}\mathcal{H}, \mathcal{P}'\mathcal{H})$.

We will see in Section 1.15 that from uncalibrated images, there are no further restrictions. We can recover only a *projective reconstruction*, which means reconstruction of points up to a general projective transformation $\mathcal{H}$ of space. To obtain an *Euclidean reconstruction* (i.e. up to a Euclidean transformation plus scale), we need to use either some *a priori* information about the world, which makes it possible to determine in succession the plane at infinity (Section 1.17) and the intrinsic parameters of a camera (Section 1.18), or either some *a priori* information about the camera, which makes it possible to perform self-calibration (Section 1.23). The flow chart of the approach to recover Euclidean reconstruction from uncalibrated images is summarized in Figure 1.21.

1.15 Projective reconstruction [Section 7.2]

Point correspondences (m_i, m_i') are the only information that we have in this section. The projective correspondence information can be summarized by the Fundamental matrix $\mathbf{F}$ of the pairs of images, that we compute from the correspondences. Any pair of projection matrices $(\mathcal{P}, \mathcal{P}')$ is a valid solution to the reconstruction problem if and only if its Fundamental matrix is compatible with the point correspondences (m_i, m_i'), or in other words if its Fundamental matrix is $\mathbf{F}$.

It can be shown [Section 7.2] that any pair $(\mathcal{P}, \mathcal{P}')$ has Fundamental matrix $\mathbf{F}$

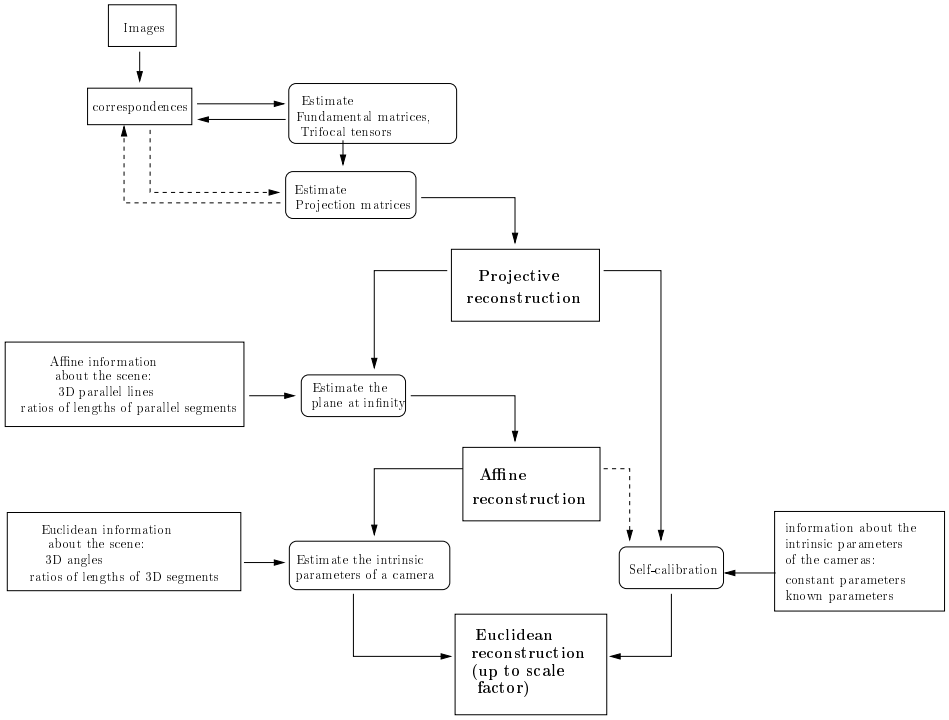

Figure 1.21: Euclidean reconstruction from images can be achieved using information either about the world or about the cameras.

if, and only if it is of the form

$$\left\{ \begin{array}{lcl} \mathcal{P} & \simeq & [\mathbf{I}_3\,\mathbf{0}_3]\mathcal{H}, \\ \mathcal{P}' & \simeq & [\mathbf{H}\,\mu\mathbf{e}']\mathcal{H}, \end{array} \right. \quad \text{with } \mathcal{H} = \left[\begin{array}{c} \mathcal{P} \\ \mathbf{\Pi}^T \end{array} \right], \tag{1.15}$$

where

- $\mathcal{P}$ is an arbitrary projection matrix (11 parameters) $\mathbf{\Pi}$ is the projective equation of an arbitrary plane (3 parameters), μ is an arbitrary constant (1 parameter), which is in fact the common scale of $\mathcal{P}$ and $\mathbf{\Pi}$ in the matrix $\mathcal{H}$. Together, these 15 parameters represent the projective ambiguity in reconstruction: the arbitrary choice of the projective basis in 3-D, or, equivalently, of the matrix $\mathcal{H}$.
- The remaining elements in $\mathcal{P}'$ are: the epipole $\mathbf{e}'$ of $\mathbf{F}$ in the second image and the homography $\mathbf{H}$, compatible with $\mathbf{F}$ and generated by the plane Π. These entities are uniquely determined by $\mathbf{F}$ and Π. This shows that given the Fundamental matrix $\mathbf{F}$, once a projective basis in $\mathbb{P}^3$ is chosen by fixing the 15 previous parameters, $\mathcal{P}'$ is uniquely determined.

So we have partitioned the $22 = 11 \times 2$ parameters of a pair of projection matrices into two types of parameters: the projective correspondence of the pair of cameras embedded in the Fundamental matrix (7 parameters), and a projective transformation (15 parameters), which represents the ambiguity in reconstruction. The Fundamental matrix is invariant to the choice of the projective basis in $\mathbb{P}^3$. From the decomposition in Equation 1.15, it is easy to verify that

$$\begin{array}{c} (\mathcal{P}_1, \mathcal{P}'_1) \text{ and } (\mathcal{P}_2, \mathcal{P}'_2) \text{ have the same Fundamental matrix} \\ \Leftrightarrow \\ \mathcal{P}_1 = \mathcal{P}_2\mathcal{H} \text{ and } \mathcal{P}'_1 = \mathcal{P}'_2\mathcal{H}, \text{ where } \mathcal{H} \text{ is a projective transformation of } \mathbb{P}^3 \end{array}$$

which shows that when the only constraints on the matches come from the Fundamental matrix, all of the reconstructions up to a projective transformations are acceptable. The basic steps of the projective reconstruction are as follows:

- Obtain pairs of correspondences m_i,m'_i.
- Solve for the Fundamental matrix with Equation 1.9
- Compute the special matrix $\mathbf{S} = [\mathbf{e}']_\times \mathbf{F}$ ($\mathbf{e}'$ is given by $\mathbf{F}^T\mathbf{e}' = 0$).
- Compute a particular pair of projection matrices, called the *projective canonical representation*, obtained by choosing the "simplest" pairs among those in Equation 1.15 using $\mathbf{S}$ as a particular instance of $\mathbf{H}$:

$$\left\{ \begin{array}{lcl} \mathcal{P} & \simeq & [\mathbf{I}_3\,\mathbf{0}_3], \\ \mathcal{P}' & \simeq & [\mathbf{S}\,\mu\mathbf{e}']. \end{array} \right. \tag{1.16}$$

This pair is an invariant representation in the sense that its elements do not depend on the choice of projective coordinates in $\mathbb{P}^3$.

- Solve for M_i with Equation 1.7.

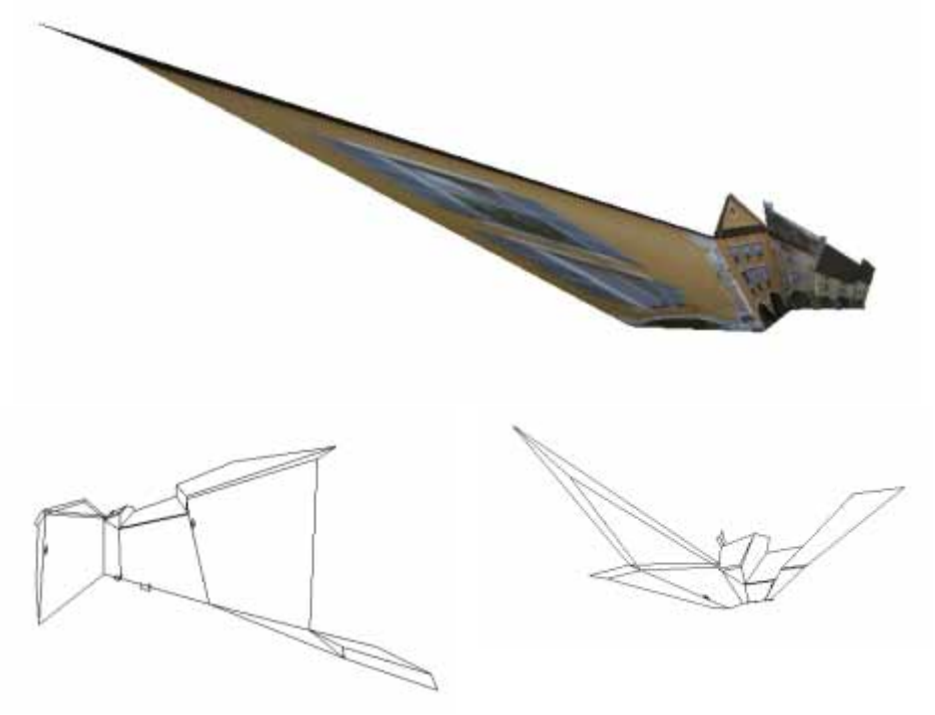

Figure 1.22: Three projective reconstructions, seen from a similar viewpoint. Each reconstruction can be transformed into another by a projective transformation of $\mathbb{P}^3$. Although the deformation can be large, incidence (coplanarity, alignment, and intersection) is preserved.

The previous algorithm will often yield a reconstruction with a very significant projective distortion because a projective transformation does not even preserve depth ordering, let alone relative distances. This means in particular that we cannot make meaningful measurements in such a reconstruction.

1.16 Reconstruction is not always necessary [Section 7.2]

There are applications for which the projective distortion is not a problem because we are not interested in reconstructing the 3-D shape.

A first example is the detection of obstacles for the navigation of a robot or the generation of alarms in a surveillance system, for which we need only qualitative measurements. This can often be achieved by positioning points with respect to a reference plane, assuming that we can identify three correspondences of points lying on this plane so that we can compute its planar homography $\mathbf{H}$. Points above the ground plane or points closer to the robot than a predefined frontal plane can be identified as obstacles. Using the projective canonical form, we can write the reconstructions equations as

$$\mathbf{m} \simeq [\mathbf{I}_3\, \mathbf{0}_3]\mathbf{M}, \qquad \mathbf{m}' \simeq [\mathbf{H}\, \mu\mathbf{e}']\mathbf{M}.$$

The first equation implies $\mathbf{M} = \begin{bmatrix} \mathbf{m} \\ \rho \end{bmatrix}$ where ρ is an unknown related to the position of the point M in a certain projective coordinate system. The substitution into the second equation yields the relation

$$\mathbf{m}' \simeq \mathbf{H}\mathbf{m} + \mu\rho\mathbf{e}', \tag{1.17}$$

where $\kappa = \mu\rho$ is a quantity called *projective parallax*. This quantity can be computed from the correspondence (m, m'), knowing $\mathbf{H}$ and $\mathbf{e}'$, for example by taking the cross-product of both terms with $\mathbf{m}'$. See Figure 1.16. The projective planar parallax turns out [Section 7.2.4] to be proportional to the distance of the point M to the plane Π and inversely proportional to the depth of the point M, which is its distance to the optical center C. This double dependency is illustrated in Figure 1.16. Although projective transformations do not preserve depth ordering, it can be sufficient to know that κ changes sign when M crosses the plane, being zero if M belongs to Π. This observation makes it possible to position a set of points on one side of a reference plane, once again without the need for reconstruction.

A second example is the synthesis of new images from two reference images. Here, the end result is a new image, so the intermediary representation of shape is not important. Once the point M is reconstructed in a projective frame, we could reproject it with any projection matrix $\mathcal{P}''$, generating a new image. In fact, no actual reconstruction is necessary. If $\mathbf{F}_{12}$ is the Fundamental matrix from image 1 to the new image, and $\mathbf{F}_{23}$ is the Fundamental matrix from image 2 to the new image, then given a point m in image 1 and a point m' in image 2, the point in image 3 is obtained in general as the intersection of the epipolar lines $\mathbf{F}_{12}\mathbf{m}$ and $\mathbf{F}_{23}\mathbf{m}'$. This operation is called *transfer*: points m and m' are transferred to the third image using the knowledge of the epipolar geometry. This idea makes it

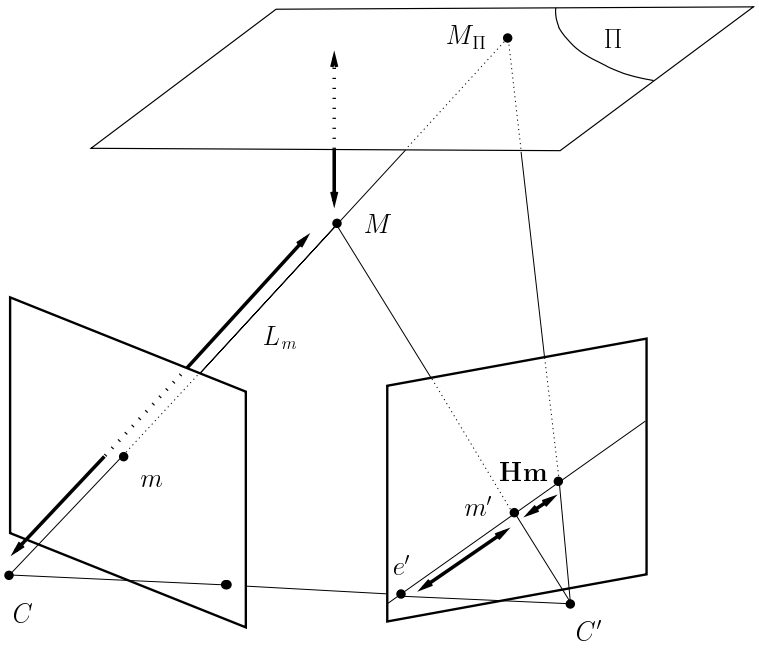

Figure 1.23: Projective parallax. The points **Hm**, e', and m' are aligned. The ratio of their distances is the projective parallax, proportional to the distance of M to Π and inversely proportional to the depth of M.

possible to represent a scene as a collection of images rather than as a 3-D model. We have made some progress towards our goal of generating synthetic views, but to insert a Euclidean model and follow a given camera trajectory we need to recover further information.

1.17 Affine reconstruction [Section 7.3]

Projective reconstruction depends only on the Fundamental matrix, which can be computed from point correspondences. We have just seen that in the general case, from two images we are able to reconstruct points and projection matrices which are obtained from the "true" points by a projective transformation. If, in addition, we have some affine information, we can reduce the ambiguity in reconstruction from a general projective transformation of $\mathbb{P}^3$ to an affine transformation of $\mathbb{P}^3$, which means that we are able to reconstruct points and projection matrices which are obtained from the "true" points by a transformation which is more constrained, and therefore induces less deformations and preserves more properties. However, the affine information has to come from some additional knowledge about the world or

Figure 1.24: Planar parallax. The third image is obtained by warping the second image by the homography of the ground plane, so that points on this plane are mapped to their position in the first image, and superimposing with the first image. The projective parallax is the length of the vector between original and warped points. It is zero for points of the reference plane, increases with height above this plane, and decreases with depth. Remark that the vectors all point towards the epipole in the image, which is near infinity in the direction X.

the system of cameras. Correspondences alone can not provide affine information.

An affine transformation is a particular projective transformation which preserves the plane at infinity Π_∞. It is easy to see that a transformation $\mathcal{A}$ conserves Π_∞ if, and only if the last row of the matrix of $\mathcal{A}$ is of the form $[0, 0, 0, \mu]$, with $\mu \neq 0$. Since this matrix is defined only up to a scale factor, we can take $\mu = 1$, then the transformation $\mathcal{A}$ is fully described by its first 3×3 sub-matrix $\mathbf{A}$ and the 3 first coordinates of the last column vector $\mathbf{b}$:

$$\boldsymbol{\mathcal{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{0}_3^T & 1 \end{bmatrix},$$

which yields the classical description of a transformation of the affine space $\mathbb{R}^3$: $\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{b}$. We have seen that points at infinity represent directions. An affine

transformation therefore preserves parallelism: parallel subspaces are transformed in subspaces which are still parallel. Other properties [Chapter 2] are that depth ordering and the ratios of distances of three aligned points are also preserved. This limits the amount of distortion introduced in the reconstruction.

The affine space is characterized by the plane at infinity Π_∞ in $\mathbb{P}^3$, which has three parameters. Affine information between two images is encoded as the correspondence of projections of points of Π_∞. The correspondence of points of Π_∞, like the correspondence of points of any plane, is described by a planar homography matrix called the *infinity homography* $\mathbf{H}_\infty$ whose expression as a function of the projection matrices was given in Equation 1.8. Once $\mathbf{F}$ is known, the three additional parameters necessary to describe $\mathbf{H}_\infty$ are in the vector $\mathbf{r}_\infty$:

$$\mathbf{H}_\infty \simeq \mathbf{S} + \mathbf{e}'\mathbf{r}_\infty^T.$$

As a particular case of Equation 1.12, in this equation the vector $\mathbf{r}_\infty$ represents the projection in the first image of the intersection of the intrinsic plane Π_S with the plane at infinity, which is the *vanishing line* of Π_S, containing the vanishing points of all sets of parallel lines of Π_S.

Once $\mathbf{F}$ is known, three correspondences of points at infinity are necessary to determine $\mathbf{H}_\infty$. One way to obtain them is to consider three corresponding vanishing points. Since parallel lines in $\mathbb{P}^3$ intersect on the plane at infinity, a vanishing point, which is the intersection of projections of parallel lines, is the projection of one point of Π_∞. Other ways to obtain one constraint on the plane at infinity include using the correspondence of one horizon point which lies at a large distance from the cameras and using knowledge of the ratio of distances for three aligned points.

When affine correspondence is known, we can restrict further the pairs $(\boldsymbol{\mathcal{P}}, \boldsymbol{\mathcal{P}}')$ of possible projection matrices. We require, in addition to the fact that it has $\mathbf{F}$ as its Fundamental matrix, the fact that $\mathbf{H}_\infty$ is its infinity homography. In fact, this requirement is redundant: if $\mathbf{H}_\infty$ is known, then using Equation 1.11, it is sufficient to specify one epipole to define $\mathbf{F}$ and to complete the description of affine correspondence. It can be verified that the pairs of projection matrices which have infinity homography $\mathbf{H}_\infty$ and epipole $\mathbf{e}'$ are of the form

$$\left\{ \begin{array}{rcl} \boldsymbol{\mathcal{P}} & \simeq & [\mathbf{I}_3\, \mathbf{0}_3]\boldsymbol{\mathcal{A}}, \\ \boldsymbol{\mathcal{P}}' & \simeq & [\mathbf{H}_\infty\, \mu\mathbf{e}']\boldsymbol{\mathcal{A}}, \end{array} \right. \quad \text{with } \boldsymbol{\mathcal{A}} = \left[\begin{array}{c} \boldsymbol{\mathcal{P}} \\ \mathbf{0}_3^T\, 1 \end{array} \right]. \tag{1.18}$$

This decomposition is a particular case of Equation 1.16, obtained with $\boldsymbol{\Pi} = [0, 0, 0, 1]^T$. The crucial remark is that the transformation of space is an affine transformation rather than a projective one. This decomposition separates the total 22 parameters into two types of parameters:

- 12 correspond to the affine ambiguity in reconstruction: the arbitrary choice of the affine basis (11 obtained by fixing $\boldsymbol{\mathcal{P}}$, 1 is μ)

- 10 describe the affine correspondence: 8 as the infinity homography $\mathbf{H}_\infty$ and 2 as the epipole $\mathbf{e}'$. That is, given affine correspondence as an infinity homography and an epipole, once an affine basis is chosen by fixing the 12 previous parameters, $\mathcal{P}'$ is uniquely defined.

From the decomposition in Equation 1.18 it is easy to verify that

$$\begin{array}{c}(\mathcal{P}_1, \mathcal{P}'_1) \text{ and } (\mathcal{P}_2, \mathcal{P}'_2) \text{ have the same infinity homography} \\ \text{and Fundamental matrix} \\ \Leftrightarrow \\ \mathcal{P}_1 = \mathcal{P}_2\mathcal{A} \text{ and } \mathcal{P}'_1 = \mathcal{P}'_2\mathcal{A},\ \mathcal{A} \text{ being an affine transformation of } \mathbb{P}^3.\end{array}$$

To summarize, when we have identified the plane at infinity Π_∞, a pair of images with epipole $\mathbf{e}'$ and infinity homography $\mathbf{H}_\infty$ determines a reconstruction up to an affine transformation of $\mathbb{P}^3$ (see Figure 1.25 for an example). The reconstruction can be performed using one particular pair of projection matrices, the *affine canonical representation*, whose elements do not depend on the choice of the affine basis in $\mathbb{P}^3$. We remark that it can be obtained from the projective representation described in Equation 1.16 by multiplication by the matrix $\mathbf{Q}_A^{-1}$:

$$\left\{ \begin{array}{rcccl} \mathcal{P} & = & [\mathbf{I}_3\, \mathbf{0}_3] & = & [\mathbf{I}_3\, \mathbf{0}_3]\mathbf{Q}_A^{-1} \\ \mathcal{P}' & = & [\mathbf{H}_\infty\, \mu\mathbf{e}'] & = & [\mathbf{S}\, \mu\mathbf{e}']\mathbf{Q}_A^{-1} \end{array} \right. \quad \text{with } \mathbf{Q}_A^{-1} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{0}_3 \\ \mathbf{r}_\infty^T & \mu \end{bmatrix}.$$

According to Equation 1.14, to upgrade the projective reconstruction of the points to an affine reconstruction, we need only to apply the transformation $\mathbf{Q}_A$. Because affine transformations include shear and different scalings along axes, the relative distances of points are not preserved. However, the relative distances of *aligned* points are preserved, so we can begin to make quantitative measurements, such as locating the middle of a segment.

1.18 Euclidean reconstruction [Section 7.4]

We now go one step further and reach more familiar ground by examining the case when in addition to the affine correspondence, we have Euclidean information. This information makes it possible to reduce the ambiguity to a *similarity transformation* (displacement plus scale) and to obtain the reconstructions illustrated in Figure 1.26, in which we can measure angles and relative distances. Just like affine transformations are particular projective transformations, similarity transformations are particular affine transformations for which the first 3×3 sub-matrix satisfies $\mathbf{A}\mathbf{A}^T = s\mathbf{I}_3$. It will be seen in Section 1.23 that this algebraic condition corresponds to the invariance to transformation of a geometric object which is a subset of the plane at infinity, the absolute conic Ω, just as the affine transformations are characterized by the invariance of Π_∞. Therefore there is a hierarchy

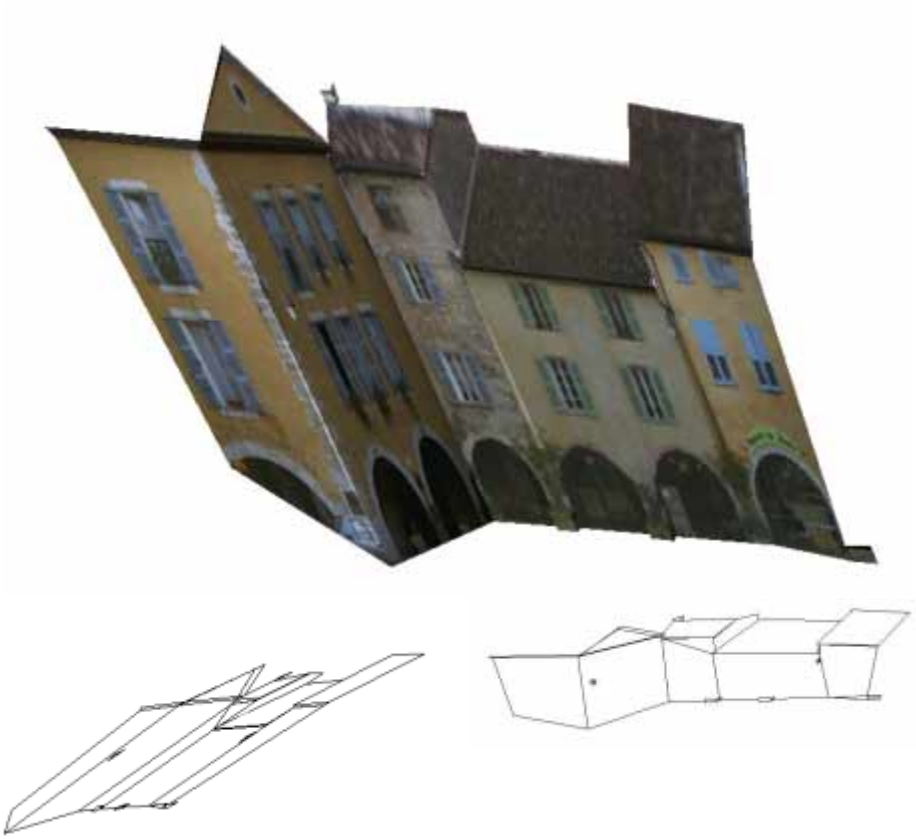

Figure 1.25: Three affine reconstructions, seen from a similar viewpoint. Each reconstruction can be transformed into another by an affine transformation of $\mathbb{P}^3$. There are shear and different scalings along axes, but parallelism is preserved.

of transformations: similarity is a subset of affine, which is a subset of projective. Each time we restrain the transformation, we further constrain the reconstruction, but to do so, we need more information.

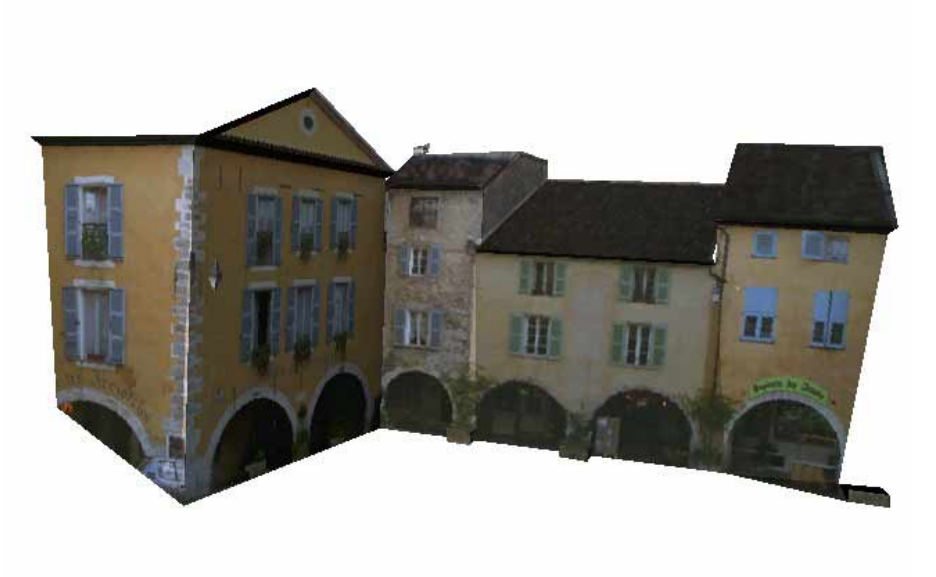

Figure 1.26: A Euclidean reconstruction, seen from a viewpoint similar to the projective and affine reconstructions. Other reconstructions would be transformed into this one by a Euclidean transformation, which is equivalent to a change of viewpoint and a global scaling. Angles and relatives distances are correct (the slight convergence is due to projection).

Euclidean information in an image is encoded as the projection of Ω, or more concretely, as the matrix of intrinsic parameters of the camera which we will call $\mathbf{A}$. In general, this matrix represents five parameters which are known when the camera is calibrated. They can be determined from a combination of knowledge about the camera (for example most real cameras have a zero skew and known aspect ratio, which reduces the number of parameters to three) and of knowledge about the world (such as angles and ratios of distances).

When $\mathbf{A}$ is known, we can restrict further the pairs $(\mathcal{P}, \mathcal{P}')$ of admissible reconstructions. We first note that if we decompose each projection matrix into intrinsic and extrinsic parameters, we have the classical decomposition for any pair of projection matrices:

$$\left\{ \begin{array}{rcccc} \mathcal{P} & \simeq & \mathbf{A}[\mathbf{R}_1\,\mathbf{t}_1] & = & [\mathbf{A}\,\mathbf{0}_3]\boldsymbol{\mathcal{S}} \\ \mathcal{P}' & \simeq & \mathbf{A}'[\mathbf{R}_2\,\mathbf{t}_2] & = & \mathbf{A}'[\mathbf{R}\,\mu\mathbf{t}]\boldsymbol{\mathcal{S}} \end{array} \right. \quad \text{with} \quad \boldsymbol{\mathcal{S}} = \left[\begin{array}{cc} \mathbf{R}_1 & \mathbf{t}_1 \\ \mathbf{0}_3^T & 1/\mu \end{array} \right],$$

where $\mathbf{R} = \mathbf{R}_2\mathbf{R}_1^T$ and $\mathbf{t} = \mathbf{t}_2 - \mathbf{R}_2\mathbf{R}_1^T\mathbf{t}_1$ represents the relative displacement between the two camera coordinate systems. Let's count again: of the total 22 parameters:

- 7 correspond to a similarity transformation representing the arbitrary choice of the Euclidean basis (6 obtained by fixing the coordinate system of the first camera through $\mathbf{R}_1$ and $\mathbf{t}_1$, and 1 being μ which represents the scale),
- 15 describe the intrinsic parameters (5 for each camera) and the relative Euclidean transformation $\mathbf{R}$, $\mathbf{t}$ (position and orientation) of the two cameras.

The direction of the translation is determined, but its norm is not because of the depth-speed ambiguity: one cannot distinguish between a close point moving slowly and a distant point moving proportionally faster.

Computing from the projection matrix, we obtain with easy algebra

$$\mathbf{e}' = \mathbf{A}\mathbf{t} \ , \ \mathbf{H}_\infty = \mathbf{A}'\mathbf{R}\mathbf{A}^{-1}.$$

From that result, we conclude that we can characterize the Euclidean correspondence by either one of the two sets of fifteen parameters:

- the affine correspondence plus intrinsic parameters of one camera: $\mathbf{H}_\infty, \mathbf{e}', \mathbf{A}$
- the intrinsic parameters of both cameras and the displacement between two cameras: $\mathbf{A}$, $\mathbf{A}'$, $\mathbf{R}$, $\mathbf{t}$.

Similarly to the previous situations:

$$\begin{gathered}(\boldsymbol{\mathcal{P}}_1, \boldsymbol{\mathcal{P}}'_1) \text{ and } (\boldsymbol{\mathcal{P}}_2, \boldsymbol{\mathcal{P}}'_2) \text{ have the same Euclidean correspondence} \\ \Leftrightarrow \\ \boldsymbol{\mathcal{P}}_1 = \boldsymbol{\mathcal{P}}_2\boldsymbol{\mathcal{S}} \text{ and } \boldsymbol{\mathcal{P}}'_1 = \boldsymbol{\mathcal{P}}'_2\boldsymbol{\mathcal{S}}, \\ \text{where } \boldsymbol{\mathcal{S}} \text{ is a Euclidean (similarity) transformation of } \mathbb{P}^3.\end{gathered}$$

We can now obtain a *Euclidean canonic representation* as a specialization of affine and projective strata. In this case, this particular pair of projection matrices are obtained just by using as 3-D coordinate system the first camera's coordinate system.

$$\left\{ \begin{array}{ccccc} \boldsymbol{\mathcal{P}} & \simeq & [\mathbf{A}\,\mathbf{0}_3] & = & [\mathbf{I}_3\,\mathbf{0}_3]\mathbf{Q}_A^{-1}\mathbf{Q}_E^{-1} \\ \boldsymbol{\mathcal{P}}' & \simeq & \mathbf{A}'[\mathbf{R}\,\mu\mathbf{t}] & = & [\mathbf{S}\,\mu\mathbf{e}']\mathbf{Q}_A^{-1}\mathbf{Q}_E^{-1} \end{array} \right. \tag{1.19}$$

with

$$\mathbf{Q}_A^{-1} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{0}_3 \\ \mathbf{r}_\infty^T & \mu \end{bmatrix} \quad \text{and} \quad \mathbf{Q}_E^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{0}_3 \\ \mathbf{0}^T & 1 \end{bmatrix}. \tag{1.20}$$

Starting from a projective reconstruction, which requires only point correspondences, we can upgrade to an affine reconstruction when $\mathbf{r}_\infty$ is known (3 degrees of freedom) by applying $\mathbf{Q}_A$ to the points M_i, and to a Euclidean reconstruction

PROJECTIVE	**homography** (incidence, $\mathbb{P}^3$)	
reconstruction ambiguity	$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \boldsymbol{\mathcal{P}} \\ \boldsymbol{\Pi}^T \end{bmatrix}$ $\boldsymbol{\mathcal{H}}$ non-singular	15
invariant description	$\mathbf{S}$: special matrix	7
	$\mathbf{e}'$: epipole	
canonical form	$\begin{cases} \boldsymbol{\mathcal{P}} & \simeq & [\mathbf{I}_3 \, \mathbf{0}_3]\boldsymbol{\mathcal{H}} \\ \boldsymbol{\mathcal{P}}' & \simeq & [\mathbf{S} + \mathbf{e}'\mathbf{r}_\Pi^T \; \mu\mathbf{e}']\boldsymbol{\mathcal{H}} \end{cases}$	22
AFFINE	**affine transformation** (parallelism, Π_∞)	
reconstruction ambiguity	$\boldsymbol{\mathcal{A}} = \begin{bmatrix} \mathbf{P} & \mathbf{p} \\ \mathbf{0}_3^T & 1/\mu \end{bmatrix}$ $\mathbf{P}$ non-singular	12
invariant description	$\mathbf{H}_\infty$: infinity homography	8
	$\mathbf{e}'$: epipole	2
canonical form	$\begin{cases} \boldsymbol{\mathcal{P}} & \simeq & [\mathbf{I}_3 \, \mathbf{0}_3]\boldsymbol{\mathcal{A}} \\ \boldsymbol{\mathcal{P}}' & \simeq & [\mathbf{H}_\infty \; \mu\mathbf{e}']\boldsymbol{\mathcal{A}} \end{cases}$	22
EUCLIDEAN	**similarity** (angles, Ω)	
reconstruction ambiguity	$\boldsymbol{\mathcal{S}} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{t}_1 \\ \mathbf{0}_3^T & 1/\mu \end{bmatrix}$ $\mathbf{R}_1$ orthogonal	7
invariant description	$\mathbf{A}$,$\mathbf{A}'$: intrinsic parameters	5+5
	$\mathbf{R}$: rotation between cameras	3
	$\mathbf{t}$: direction of translation between cameras	2
canonical form	$\begin{cases} \boldsymbol{\mathcal{P}} & \simeq & [\mathbf{A} \, \mathbf{0}_3]\boldsymbol{\mathcal{S}} \\ \boldsymbol{\mathcal{P}}' & \simeq & \mathbf{A}'[\mathbf{R} \; \mu\mathbf{t}]\boldsymbol{\mathcal{S}} \end{cases}$	22

Table 1.4: Canonical forms for the geometries of two images: for each level of description, we have a partition of the 22 parameters describing two projective images into an invariant representation, which represent the correspondence, and the ambiguity in reconstruction. The last column indicates the number of degrees of freedom.

when $\mathbf{A}$ is known (5 DOF) by applying $\mathbf{Q}_E$. $\mathbf{Q}_A$ is a projective transformation which moves the plane at infinity, and $\mathbf{Q}_E$ is an affine transformation which moves the absolute conic in the plane at infinity. Each upgrade reduces the reconstruction ambiguity, first from a general homography (15 DOF) to affine (12 DOF), then from affine to similarity (7 DOF). The representations and ambiguities in reconstruction are summarized in Table 1.4.

We have come a bit closer to our goal of building metric models and generating synthetic augmented images, provided that some information about the scene is available. However, from a practical point of view, using only two images does not afford a lot of robustness towards image noise and imprecision. Moreover, for complex objects, it is necessary to integrate several viewpoints to cover all of the parts which would be occluded from just two views.

1.19 The geometry of three images [Section 8.1]

Although two images make it possible to perform a reconstruction of the scene from point correspondences, adding a third image has two significant geometrical benefits.

First, the point correspondence problem becomes entirely constrained, because, as remarked in Section 1.16, the point in the third image can be transfered from the correspondence in the first two images $m \leftrightarrow m'$, and the fundamental matrices, as

$$\mathbf{m}'' \simeq \mathbf{F}_{12}\mathbf{m} \times \mathbf{F}_{23}\mathbf{m}', \tag{1.21}$$

where $\mathbf{F}_{12}$ (respectively $\mathbf{F}_{23}$) is the Fundamental matrix between view 1 and view 2 (respectively 3). While the epipolar constraint is bilinear in the coordinates of the two image points, this equation is *trilinear* in the coordinates of the three image points.

We note that this method of transfer fails if the two lines represented by $\mathbf{F}_{13}\mathbf{m}$ and $\mathbf{F}_{23}\mathbf{m}'$ are identical. This can happen if the 3-D point M of which m, m', m'' are projections belongs to the plane containing the three optical centers (called *Trifocal plane*) or if the three optical centers are aligned. We are going to see that the three Fundamental matrices $\mathbf{F}_{12}$, $\mathbf{F}_{23}$, and $\mathbf{F}_{13}$ are not independent but are linked by three constraints which express that all the epipoles belong to the Trifocal plane. $\mathbf{F}_{31}\mathbf{e}_{32}$, the epipolar line of $\mathbf{e}_{32}$ in the first image, is the intersection of the trifocal plane with the first retinal plane. The epipoles $\mathbf{e}_{12}$ and $\mathbf{e}_{13}$ also belong to this plane and to the first retinal plane, thus: $\mathbf{e}_{12} \times \mathbf{e}_{13} \simeq \mathbf{F}_{31}\mathbf{e}_{32}$. The two other equations follow by a circular permutation of indices. This is illustrated in Figure 1.19. Therefore the system of Fundamental matrices depends on at most $18 = 7 \times 3 - 3$ parameters. It can be shown [Section 8.1] that this number is exact. The degeneracy of transfer and the non-independence of the Fundamental matrices suggest that they might not be the best way to represent the geometry of three images.

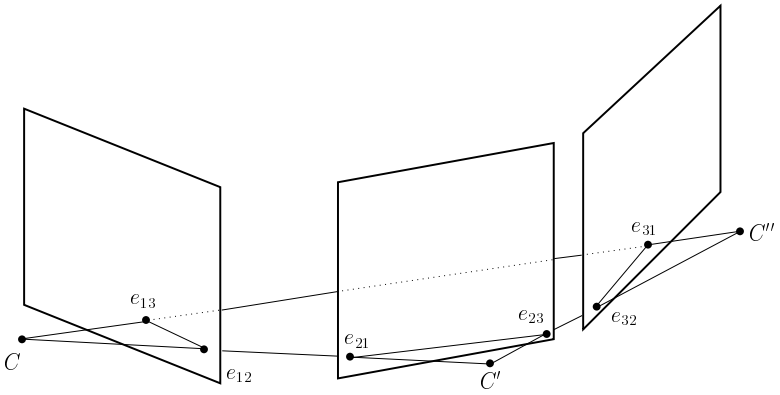

Figure 1.27: The six epipoles of three images lie on the *trifocal plane* defined by the three optical centers.

Second, it is possible to use line correspondences, unlike in the case of two images. As we have discussed in Section 1.10 in general two optical rays L_m and $L_{m'}$ do not intersect. Therefore the correspondence of two points yields one constraint which we described in the image as the epipolar constraint, and is formulated in space as the concurrency of L_m and $L_{m'}$. Knowing the projection l of a line L in space constrains L to lie on a plane Π_l going through the optical center. For any two lines l in the first image and l' in the second image, the planes Π_l and Π'_l will always intersect on a certain line $L_{ll'}$ which projects back to l and l'. Contrast Figure 1.10 and Figure 1.19. Therefore the correspondence of two lines in just two images does not yield any constraint. Let assume we have three images, and the correspondence $l \leftrightarrow l' \leftrightarrow l''$ of a line in each image. We can construct as before a line $L_{l'l''}$ which projects into l' in the second image and l'' in the third image. We can pick any point m on the line l and consider its optical ray L_m. Now that we have two lines, L_m and $L_{l'l''}$, we can obtain a constraint on the cameras called the *Trifocal constraint* by writing that they are concurrent, generalizing the construction we did with points in two images. See Figure 1.19. Note that unlike the case of two images, this construction is not symmetric. One image, the one where we have picked the point, plays a special role, while the two others play symmetric roles.

1.20 The Trifocal tensor [Chapter 8]

Let's put some algebra behind this geometric intuition. In two images, the correspondence of points is described by the 3×3 Fundamental matrix. We will see in this section that in three images, the correspondence of points and lines is described by a $3 \times 3 \times 3$ tensor $\mathcal{T}$, of entries T_i^{jk}, called the *Trifocal tensor*.

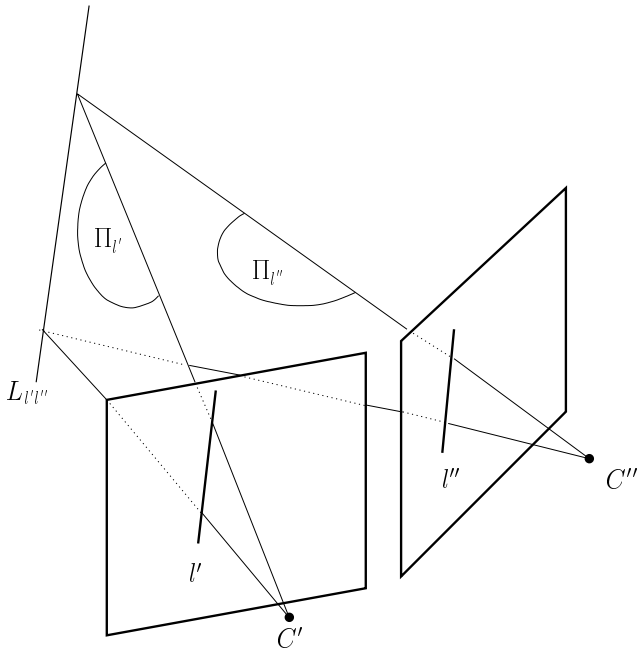

Figure 1.28: A line-line correspondence in two images. For any l' and l'', there is always a valid reconstruction $L_{l'l''}$, obtained by intersecting the planes $\Pi_{l'}$ and $\Pi_{l''}$.

We adopt the usual summation convention for tensors (*Einstein convention*), that any index repeated as subscript and superscript implies a summation over the range value, which in this section will always be 1..3. Any formula involving indices holds for any choice of values of the indices which are not repeated. Because of that, the order in which the terms are listed is unimportant. For instance, the correspondence of points through a homography matrix $\mathbf{H} = (H)_{ij}$ is written as $(m')^i = H^i_j m^j$. Superscripts designate *contravariant* indices (coordinates of points, row index of matrices), subscripts designate *covariant* indices (coordinates of lines, column index of matrices). These transform inversely under changes of basis, so that the contraction (dot product, or sum over all values) of a covariant-contravariant pair is invariant.

It can be shown [Section 8.2] that given the lines l' in the second image and l'' in the third image, the line l, which is the projection in the first image of their reconstruction, is given by

$$l_i = l'_j l''_k \mathcal{T}_i^{jk}. \tag{1.22}$$

The Trifocal tensor lets us predict the position of a line in a third image from its position in two images.

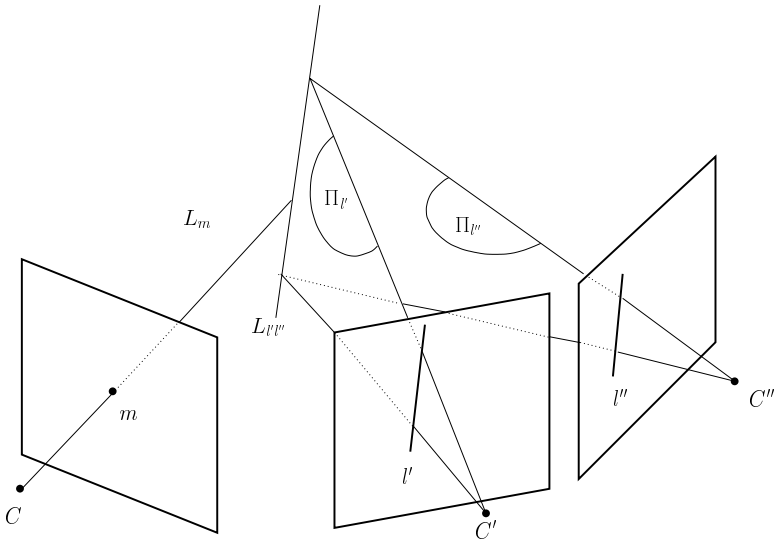

Figure 1.29: A point-line-line correspondence in three images. The optical ray L_m must intersect the line $\Pi_{l'} \cap \Pi_{l''}$.

The basic Trifocal constraint, obtained with a correspondence $m \leftrightarrow l' \leftrightarrow l''$ of a point in the first image and lines in the other images, is obtained by considering a point m on the the line l:

$$m^i l'_j l''_k \mathcal{T}_i^{jk} = 0. \tag{1.23}$$

It expresses the concurrency of the optical ray of m and the 3-D line projected to l' and l''. This constraint is the analogue for three images of the epipolar constraint for two images. While the epipolar constraint was bilinear in image coordinates, the Trifocal constraint is trilinear. It yields a considerably richer and more complex geometry. In the case of two images, the only geometric operation we could do with the Fundamental matrix was to apply it to a point and obtain the epipolar line. There are more possibilities with the Trifocal tensor, which we detail next and summarize in Table 1.5.

If we apply $\mathcal{T}$ only to the line l', i.e. fix it, we obtain a 3×3 matrix $\mathbf{H}'$ which maps a point from the first image to a point in the third image. Equation 1.23 can be read

$$l''_k \overbrace{m^i \underbrace{l'_j \mathcal{T}_i^{jk}}_{H'^k_i}}^{m''^k} = 0.$$

This equation is true for any line l'' containing the projection in the third image of the intersection of the optical ray L_m of m with the plane $\Pi_{l'}$ defined by the line

1	2	3	resulting object
m			$\mathbf{G}$, correlation from image 2 to image 3
	l'		$\mathbf{H}''$, homography from image 1 to image 3
		l''	$\mathbf{H}'$, homography from image 1 to image 2
m	l'		m'', point in image 3
m		l''	m', point in image 2
	l'	l''	l, line in image 1
m	l'	l''	scalar

Table 1.5: Contractions of the Trifocal tensor. This table shows the geometric objects resulting from applying the Trifocal tensor in different ways.

l'. See Figure 1.20. Therefore, the matrix $\mathbf{H}'$ is a planar homography from the first image to the third image, generated by the plane $\Pi_{l'}$. The Trifocal tensor lets us predict the position of a point in a third image from a point in the first image and a line in the second image. A similar result is obtained by exchanging images 2 and 3. In the more frequent case when we have two points $m \leftrightarrow m'$, the projections of M, we can choose a line through one of the points and do the transfer. We notice that this construction always works provided that the point M does not lie on the line joining the optical centers of the two first cameras, whereas the transfer based on the Fundamental matrices described by Equation 1.21 suffered from several other degeneracies.

Last, if we apply $\mathcal{T}$ only to the point m, i.e. fix it, we obtain a 3×3 matrix $\mathbf{G}$ which maps a line l' from the second image to a point m'' in the third image. Such a mapping is called a *correlation*. Equation 1.23 can be read

$$l''_k \overbrace{l'^j \underbrace{m^i \mathcal{T}_i^{jk}}_{G^{jk}}}^{m''^k} = 0.$$

All the points m'' are projections of a point of the optical ray L_m, therefore as l' varies, its mapping by $\mathbf{G}$ describes the epipolar line of m in the third image. Remark that no other points of the third image than this line are reached by $\mathbf{G}$, therefore the range of $\mathbf{G}$ has dimension 1, and we conclude that $\mathbf{G}$ has rank 2. See Figure 1.20. This result indicates that there is a connexion between the Trifocal tensor and the Fundamental matrices. In fact, the three Fundamental matrices can be computed from the Trifocal tensor [Section 8.2.4].

As a particular case of those results, the entries $\mathcal{T}_i^{jk}$ of the Trifocal tensor can themselves be interpreted as homographies or correlations by remarking that those entries are obtained by applying the tensor to the three entities $[1,0,0]^T$, $[0,1,0]^T$, $[0,0,1]^T$. Considering them as lines in the second image yields the three *intrinsic*

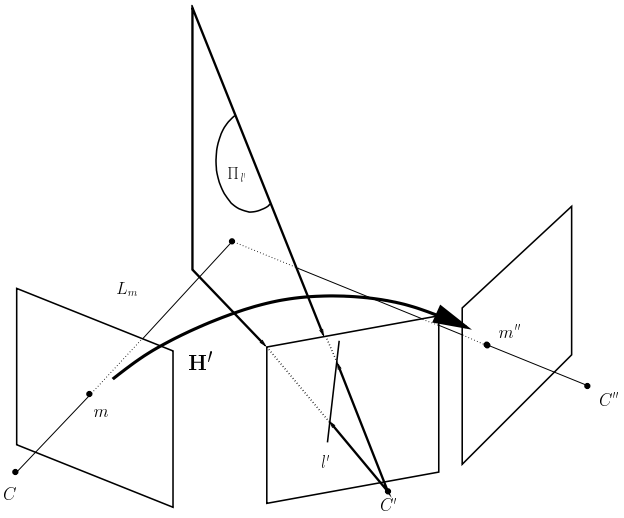

Figure 1.30: The line l' defines the plane Π'_l, which generates an homography $\mathbf{H}'$ between the first image and the third image.

homographies $\mathbf{H}'^k$, of entries $(H'^k)^j_i = \mathcal{T}_i^{jk}$. The tensor can be viewed as a stack of these three homographies. Considering them as points in the first image yields three matrices, the *Trifocal matrices* $\mathbf{G}^i$ of entries $(G^i)^{jk} = \mathcal{T}_i^{jk}$.

As discussed in Section 1.19, the Trifocal tensor depends on 18 independent parameters. $\mathcal{T}$ has $27 = 3 \times 3 \times 3$ entries defined up to a scale factor. Therefore, like the Fundamental matrix had to satisfy one constraint $\det(\mathbf{F}) = 0$, the Trifocal tensor has to satisfy eight algebraic constraints [Section 8.4]. We have seen three of them: the matrices $\mathbf{G}^i$ have zero determinant. Like the Fundamental matrix could be expressed as a function of the epipole and a homography, we have the relation [Section 8.2]

$$\mathcal{T}_i^{jk} = (e')^j (H'')^k_i - (e'')^k (H')^j_i,$$

where e' (respectively e'') is the epipole in the second (respectively third) image with respect to the first image, $\mathbf{H}'$ is a planar homography of a plane Π between the first and the second image, and $\mathbf{H}''$ is the planar homography of the same plane Π between the first and the third images.

1.21 Computing the Trifocal tensor [Chapter 9]

If we have the correspondence of one point m in the first image and two lines l' and l'' respectively in the second and third images, then the basic Trifocal constraint of Equation 1.23 gives one equation which is linear in the entries of $\mathcal{T}$. In practice,

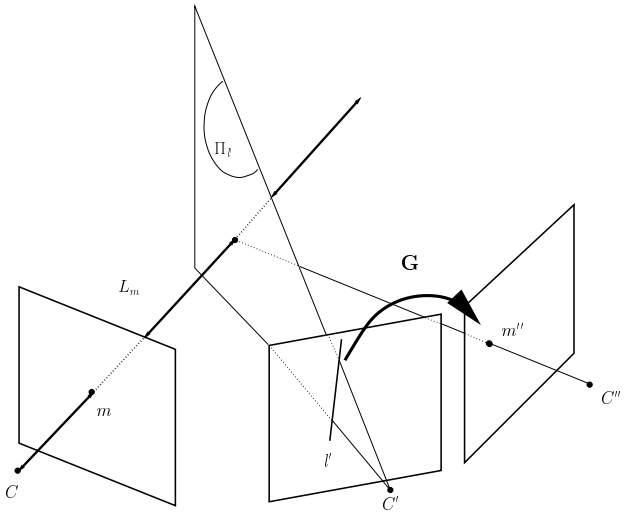

Figure 1.31: The point m defines the optical ray L_m, which generates a correlation G between the second image and the second image.

however, we are more often given the correspondence of three lines or of three points.

The first case, $l \leftrightarrow l' \leftrightarrow l''$ is quite simple: we write that the first line l_i and the line $l'_j l''_k \mathcal{T}_i^{jk}$ transferred from l' and l'' are identical, which can be expressed by the fact that their cross-product is zero. This gives three equations, two of which are independent.

In the second case, we notice that the Trifocal constraint holds for any line l' going through the point m' in the second image, and also for any line l'' going through the point m'' in the third image. Since lines through a point form a projective space of dimension one, there are two independent choices for l', as well as for l'', which yield a total of four independent equations. One possible choice is to consider the horizontal and vertical lines going through the points m' and m'', as illustrated in Figure 1.21

$\mathcal{T}$ has $27 = 3 \times 3 \times 3$ entries defined up to a scale factor. Since each correspondence of lines (respectively points) gives 2 (respectively 4) linear equations in the entries $\mathcal{T}_i^{jk}$, provided that $2\,n_{lines} + 4\,n_{points} \geq 26$ (n_{lines} and n_{points} represent the numbers of lines and points respectively), we can solve linearly for the entries of $\mathcal{T}_i^{jk}$ by constraining them to have sum 1.

Even more than for the Fundamental matrix, data normalization is crucial for the linear method to yield correct results. However, the best methods are obtained by minimizing a symmetrized geometric error function while enforcing algebraic constraints. The problem of correctly parameterizing the Tensor to enforce those constraints is quite tricky, and the symmetrization of the error function is also more

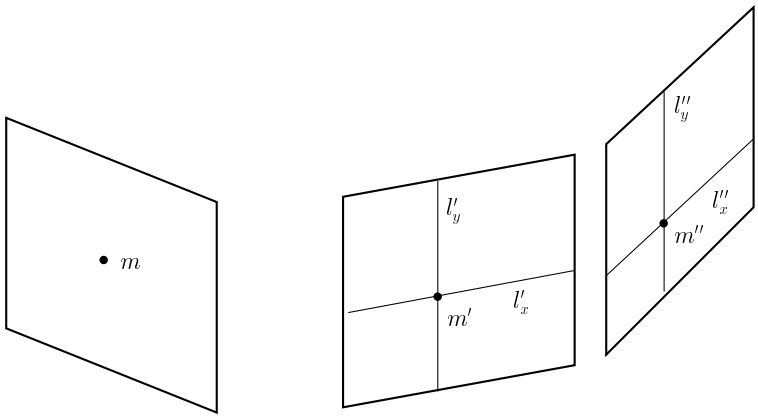

Figure 1.32: A point-point-point correspondence gives four point-line-line correspondences using the cross-hair directions.

difficult to express than for the Fundamental matrix.

1.22 Reconstruction from N images [Chapter 10]

We have seen in Section 1.15 that from two images, with no more information than point correspondences, it is possible to obtain only a projective reconstruction. This situation does not change when more views are added. The reasoning in Section 1.14 still holds, and the new views do not help obtain more constraints. Going from two views to three views, the number of parameters describing the correspondence goes from 7 for the Fundamental matrix to 18 for the Trifocal tensor, while the number of unknown camera parameters grows from $22 = 2 \times 11$ to $33 = 3 \times 11$. The remaining ambiguity is just the same 15 parameters for the projective transformation.

With a fourth view, one might consider quadrilinear constraints as a generalization of the trilinear constraints (Trifocal tensor) and bilinear constraints (Fundamental matrix). However, it turns out [Section 10.2] that the quadrilinear constraints can always be expressed as an algebraic combination of trilinear and bilinear constraints, so they wouldn't help. Moreover, beyond four views there are no further entities.

The projective geometry of N views can be represented using a canonical rep-

resentation which extends the one introduced for two views in Section 1.15:

$$\begin{cases} \mathcal{P}_1 = [\mathbf{I}_3\, \mathbf{0}_3], \\ \mathcal{P}_2 = [\mathbf{P}_2\, \mathbf{p}_2], \\ \mathcal{P}_3 = [\mathbf{P}_3\, \mathbf{p}_3], \\ \vdots \\ \mathcal{P}_N = [\mathbf{P}_N\, \mathbf{p}_N], \end{cases} \tag{1.24}$$

where $\mathbf{P}_2$ is the homography between views 1 and 2 associated with a plane Π, which is determined by the projective basis chosen, and $\mathbf{p}_2$ is the second epipole. As we have just discussed at the beginning of this section, while the first two matrices depend on a total of seven parameters to describe the geometry of the cameras (four additional parameters, the relative scale of $\mathbf{P}_2$ and $\mathbf{p}_2$, and Π are part of the projective ambiguity), all of the remaining matrices $\mathcal{P}_3 \ldots \mathcal{P}_n$ depend on 11 parameters each and therefore do not require a particular parameterization other than by their entries, hence our notation. However, there is a simple geometric interpretation of those entries, illustrated by Figure 1.22. $\mathbf{p}_i$ represents the coordinates of the epipole in the image i with respect to the first image, and $\mathbf{P}_i$ is the homography generated by Π between image one and image i. To see this, let us consider the point $\mathbf{m}$ in the first image, projection of a point M of Π. Because its planar parallax (see Equation 1.17) is zero, we have $\mathbf{M} = \begin{bmatrix} \mathbf{m} \\ 0 \end{bmatrix}$. Therefore, $\mathcal{P}_i\mathbf{M} = \mathbf{P}_i\mathbf{m}$.

The affine case is obtained as the particular case when the plane Π is the plane at infinity Π_∞, and the Euclidean case is obtained by replacing the homographies by rotations and the epipoles by translations, and using camera coordinates in the images.

We now use the more familiar Euclidean case to illustrate that the *local* representations (based on pairs of views) 1-2 and 2-3 are not sufficient to determine the representation 1-3, and therefore the global representation for three views 1-2-3, but that on the other hand, this global representation can be recovered from the three local representations, 1-2,1-3,2-3. Since a similarity has 7 degrees of freedom (see Table 1.4), the representation 1-2-3 has $33 - 7 = 26$ parameters, consisting of 3×5 intrinsic camera parameters, 2×3 rotation parameters, 3×3 translation parameters, minus the global scale μ:

$$\begin{cases} \mathcal{P}_1 = \mathbf{A}_1[\mathbf{I}_3\, \mathbf{0}_3], \\ \mathcal{P}_2 = \mathbf{A}_2[\mathbf{R}_{12}\, \mu\mathbf{t}_{12}], \\ \mathcal{P}_3 = \mathbf{A}_3[\mathbf{R}_{13}\, \mu\mathbf{t}_{13}]. \end{cases} \tag{1.25}$$

This scale factor μ must be the same for $\mathcal{P}_2$ and $\mathcal{P}_3$ because when we have three views, there is only a global scale indetermination, but the ratio of the local scale factors is completely determined, as we shall see soon. Each local (two-view) rep-

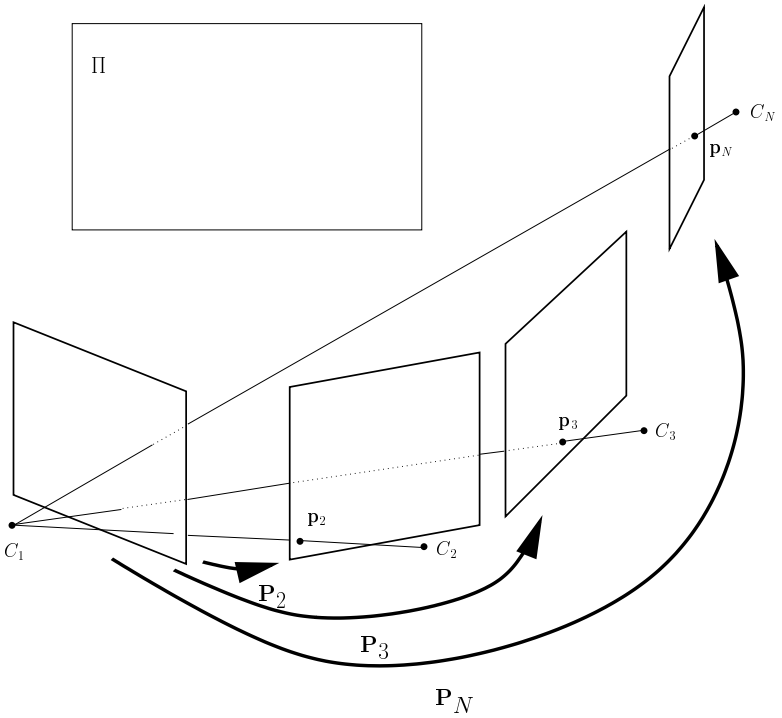

Figure 1.33: In the projective canonical representation for N views, $\boldsymbol{\mathcal{P}}_1 = [\mathbf{I}_3\,\mathbf{0}_3]$, $\boldsymbol{\mathcal{P}}_i = [\mathbf{P}_i\,\mathbf{p}_i]$, $\mathbf{P}_i$ is the homography generated by Π, and $\mathbf{p}_i$ is the epipole in the image.

resentation $i - j$ is written as

$$\begin{cases} \boldsymbol{\mathcal{P}}_{Li} = \mathbf{A}_i[\mathbf{I}_3\,\mathbf{0}_3], \\ \boldsymbol{\mathcal{P}}_{Lj} = \mathbf{A}_j[\mathbf{R}_{ij}\,\mu_{ij}\mathbf{t}_{ij}], \end{cases}$$

where $\mathbf{t}_{ij}$ is the translation from the camera coordinate system of image i to the camera coordinate system of image j, and μ_{ij} is an unknown scale factor which arises from the fact that from correspondences, one cannot determine the absolute translation between two views, but only its direction. Let's try to obtain the direction of $\mathbf{t}_{13}$ from the representations 1-2 and 2-3. The relation between the translations is

$$\mu_{13}\mathbf{t}_{13} = \mu_{12}\mathbf{R}_{23}\mathbf{t}_{12} + \mu_{23}\mathbf{t}_{23}. \tag{1.26}$$

Because the scale factors μ_{12} and μ_{23} are unknown, it is not possible to determine the direction of $\mathbf{t}_{13}$ and therefore to build the representation 1-2-3. This is not surprising because we have here 3×5 intrinsic camera parameters, 2×3 rotation

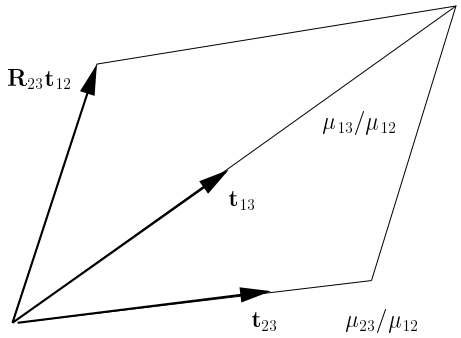

Figure 1.34: From the directions of $\mathbf{R}_{23}\mathbf{t}_{12}$ and $\mathbf{t}_{23}$ alone, it is not possible to recover the direction of $\mathbf{t}_3$. Their relative scale is necessary and sufficient. Reciprocally this relative scale can be recovered if the three directions are known.

parameters, 3×3 translation parameters, minus the two local scales μ_{ij} which totals only 25! However, if we know also the representation 1-3, i.e. $\mathbf{t}_{13}$, then the ratios μ_{23}/μ_{12} and μ_{13}/μ_{12} can be computed by expressing the coplanarity of the three vectors involved, which recovers the translations up to a global scale factor. This is illustrated in Figure 1.22.

The affine case is very similar. In the projective case, the reasoning is just the same, but instead of one unknown scale, four parameters are implied (remember that the system of three fundamental matrices depend on 18 parameters, while two independent fundamental matrices have 14 parameters). It is found [Section 10.2] that from the three Fundamental matrices or from the Trifocal tensor, one can recover a representation of the form (1.24). From that result, one can devise an incremental method which starts from a canonical representation of an $N-1$-tuple of projection matrices and incorporates an Nth view, yielding a representation for N views.

Adding more views in general does not make it possible to upgrade from a projective to an affine or Euclidean representation without specific knowledge (such as for instance the fact that the cameras are identical, as discussed in the next section); however, it improves the robustness of the estimation and makes it possible to deal with more complex scenes. Several algorithms for reconstruction from N views are examined in Chapter 10.

1.23 Self-calibration of a moving camera using the absolute conic [Sections 2.3, 4.3.2]

We have explained how using some knowledge about the world made it possible to upgrade a projective reconstruction to affine and Euclidean. This knowledge has

to be supplied to the system interactively by the user. An alternative, and more automatic, way to supply this information is to use constraints about the camera's intrinsic parameters, such that the fact that they are partially known (for instance for most cameras the skew is zero since the pixel grid is orthogonal, and the aspect ratio is known) or constant across images.

In order to use these constraints for this purpose, we now give a projective encoding of Euclidean structure, using the absolute conic. Let us consider the projective plane $\mathbb{P}^2$. When we move from an affine to a Euclidean representation, we gain the notion of angles and distances, and can define geometric entities such as circles. In projective geometry all the second order loci (ellipses, parabolas, hyperbolas) lose their distinction and are called *conics*, with an equation of the form $\mathbf{m}^T\mathbf{Q}\mathbf{m} = 0$, where $\mathbf{Q}$ is a square matrix. Their projective equivalence is illustrated by the fact that any form can be projected into any other form. Using the duality of points and lines, a conic can be considered not only as a locus of points, but also as a locus of lines, the set of lines which are tangent to the conic. For a given conic of matrix $\mathbf{Q}$, its *dual conic* (the set of its tangents) has matrix the adjoint matrix of $\mathbf{Q}$:

$$\mathbf{Q}^* = \det(\mathbf{Q})\mathbf{Q}^{-1} \tag{1.27}$$

In the projective plane $\mathbb{P}^2$, just like two lines always intersect provided that we add to the affine points the line at infinity, so do two circles, provided that we add to the affine points two complex points in the line at infinity. Indeed, we encourage the reader to verify by simple algebra the surprising fact that all the circles contain the two particular points $I = [1, i, 0]$ and $J = [1, -i, 0]$, called *circular points*, which satisfy $x^2 + y^2 = z = 0$. Similarly, the Euclidean space $\mathbb{P}^3$ is characterized by the *absolute conic* Ω, which is the set of points $[X, Y, Z, T]^T$ satisfying $X^2+Y^2+Z^2 = 0$ and $T = 0$. In other words Ω is the conic in Π_∞ of matrix $\mathbf{I}_3$. Like the plane at infinity was used to characterize affine transformations, the absolute conic can be used to characterize similarity transformations. It can be verified that a projective transformation is a similarity transformation if and only if it leaves the absolute conic invariant.

Because the change of pose corresponds to a Euclidean transformation, which, as a particular case of a similarity transformation, leaves the absolute conic invariant, we conclude that its projection ω which is also a conic with only complex points, does not depend on the pose of the camera. Therefore, its equation in the retinal coordinate system does not depend on the extrinsic parameters and depends only on the intrinsic parameters. Computing in the camera coordinate system, it is easy to see that its matrix is $\mathbf{B} = \mathbf{A}^{-T}\mathbf{A}^{-1}$, whereas its dual conic has matrix, using Equation 1.27,

$$\mathbf{K} = \mathbf{B}^* = \det(\mathbf{B})\mathbf{B}^{-1} \simeq \mathbf{A}\mathbf{A}^T.$$

When the camera is calibrated, in the camera coordinate system, the projection of the absolute conic is just an imaginary circle of radius one. The general uncalibrated case is illustrated in Fig. 1.23. The matrix $\mathbf{K}$ is called the *Kruppa matrix*. It

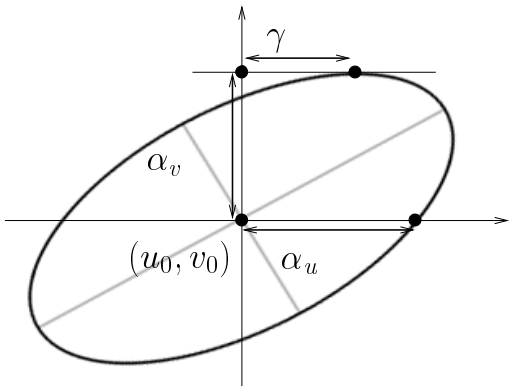

Figure 1.35: The way the five intrinsic parameters of Equation 1.4 affect the image of the absolute conic in the imaginary plane.

is symmetric and defined up to a scale factor, and therefore it depends on five independent parameters which can be recovered uniquely from $\mathbf{K}$.

The principle of self-calibration is to use the absolute conic as a calibration object. It has the advantage of being always available for free, however since it is a rather inaccessible object, all we can do with it is to write constraints across images. The knowledge of the infinity homography or of the Fundamental matrix makes it possible to write equations relating the intrinsic parameters of the two images. When we move a camera, we can obtain several such equations. Combining enough of them with the constraints on intrinsic parameters make it possible to recover the intrinsic parameters. Therefore just by moving a camera, observing its environment, and establishing point correspondences, we are able to calibrate the camera and eventually perform a Euclidean reconstruction of the environment, without ever needing a calibration object.

1.24 From affine to Euclidean [Section 11.1]

We first start with the simpler case when the infinity homography is known. A practically important situation when this occurs is the case of a stationary camera which rotates and zooms while staying at a fixed position. While this situation does not make it possible to recover structure, it is the most favorable to recover the calibration because, as we shall see soon, in this case the constraints take a particularly simple form.

We remark that in $\mathbf{H}_\infty \simeq \mathbf{A}'\mathbf{R}\mathbf{A}^{-1}$, $\mathbf{H}_\infty$ depends on eight parameters, and

the rotation on three parameters, so we should be able to write five constraints on the intrinsic parameters. To express them, we eliminate the motion using the fact that a rotation matrix is orthogonal: $\mathbf{R}\mathbf{R}^T = \mathbf{I}_3$. This step yields a linear relation between the infinity homography and the intrinsic parameters:

$$\mathbf{H}_\infty \mathbf{K} \mathbf{H}_\infty^T \simeq \mathbf{K}', \tag{1.28}$$

where $\mathbf{K}$ and $\mathbf{K}'$ are the two Kruppa matrices. One could think that this is sufficient to solve for the five intrinsic parameters when they are constant ($\mathbf{K} = \mathbf{K}'$), but it turns out that one of the equations in this case is redundant because the infinity homography satisfies the additional constraint $\det(\mathbf{H}_\infty) = 1$. More precisely, it can be verified that there is a two dimensional space of solutions, spanned by the expected solution $\mathbf{K} = \mathbf{A}\mathbf{A}^T$ and the spurious solution $\mathbf{K} = \mathbf{A}\mathbf{U}(\mathbf{A}\mathbf{U})^T$, where $\mathbf{U}$ is the axis of the rotation. Two rotations along different axis are therefore necessary to solve for all of the intrinsic parameters.

We can remark that self-calibration depends only on the rotational component of the motion: it relies on the absolute conic, which being an object at infinity, has an projection not affected by translations but only by rotations. In particular, if there is no rotation, then Equation 1.28 becomes a tautology. In order to recover camera calibration from $\mathbf{H}_\infty$ it is therefore necessary to have a motion with a non-null rotational component. We will see that these conclusions extend to the recovery of camera calibration from $\mathbf{F}$, since it uses equations which are derived from Equation 1.28.

1.25 From projective to Euclidean [Section 11.2]

We now consider the general case when only a projective representation is available.

We note that the representation of Euclidean correspondence consisting of $\mathbf{A}$, $\mathbf{A}'$, $\mathbf{F}$ is redundant since it contains seventeen parameters, while a minimal representation has only fifteen parameters, so there must be two constraints between the Fundamental matrix and the intrinsic parameters. Another way to see it is to remark that the Fundamental matrix can be expressed as

$$\mathbf{F} = \mathbf{A}'^{-T} [\mathbf{t}]_\times \mathbf{R} \mathbf{A}^{-1}.$$

$\mathbf{F}$ depends on seven parameters and the motion on five parameters (the scale of the translation is not determined), so there must be two constraints, obtained again by eliminating the displacement. They can be obtained algebraically by multiplying Equation 1.28 by $[\mathbf{e}']_\times$ left and right,

$$[\mathbf{e}']_\times \mathbf{K}' [\mathbf{e}']_\times \simeq [\mathbf{e}']_\times \mathbf{H}_\infty \mathbf{K} \mathbf{H}_\infty^T [\mathbf{e}']_\times,$$

and using Equation 1.11:

$$\mathbf{F}\mathbf{K}\mathbf{F}^T \simeq [\mathbf{e}']_\times \mathbf{K}' [\mathbf{e}']_\times,$$

which is equivalent to two polynomial equations of degree 2 in the coefficients of $\mathbf{K}$ and $\mathbf{K}'$. These equations are called the *Kruppa equations*. Examples of applications of these equations are:

- Computing the focal lengths of the two cameras from the Fundamental matrix of a pair of images, assuming that the other intrinsic parameters are known.
- Computing all of the intrinsic parameters of a moving camera with constant intrinsic parameters from three images.

The Kruppa equations make it possible to perform self-calibration from the Fundamental matrices by solving only for the intrinsic parameters using polynomial methods. While this is interesting when we have few images, when we have a large number of images, it is advantageous to start from the projective canonical representation, because being global, it is numerically a more stable representation. We recover affine and Euclidean information at the same time by solving for the eight parameters of the projective transformation of Equation 1.20:

$$\mathcal{H} = \mathbf{Q}_A^{-1}\mathbf{Q}_E^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{0}_3 \\ \mathbf{r}_\infty^T\mathbf{A} & \mu \end{bmatrix}.$$

According to Equation 1.19, this transformation maps the set of canonical perspective projection matrices $\mathcal{P}_i$ into canonical Euclidean projection matrices:

$$\mathcal{P}_i\mathcal{H} \simeq \mathbf{A}_i[\mathbf{R}_i\,\mathbf{t}_i], \qquad 2 \leq i \leq n.$$

We use the same trick as we did in the affine case: we multiply the first 3×3 submatrix (remember that only the rotational motion is relevant for self-calibration) by its transpose, eliminating the unknown rotation matrices $\mathbf{R}_i$:

$$\mathcal{P}_i \begin{bmatrix} \mathbf{K} & \mathbf{K}\mathbf{r}_\infty \\ \mathbf{r}_\infty^T\mathbf{K} & \mathbf{r}_\infty^T\mathbf{K}\mathbf{r}_\infty \end{bmatrix} \mathcal{P}_i^T \simeq \mathbf{K}_i, \qquad 2 \leq i \leq n.$$

These equations make it possible to generalize self-calibration to the case of variable intrinsic parameters. Any constraints on the intrinsic parameters $\mathbf{A}_i$ translates into constraints on the Kruppa matrices $\mathbf{K}_i$. For instance, it can be verified with easy algebra that when the pixels are orthogonal, $K_{13}K_{23} - K_{33}K_{12} = 0$. If enough constraints are available, we can solve for $\mathbf{K}$ and $\mathbf{r}_\infty$ by nonlinear minimization. Since there are 8 unknowns, it is crucial to have a good starting point, which can be obtained with the Kruppa equations. This approach was used to self-calibrate the cameras and obtain the results presented at the beginning of the Chapter.

Like there are critical configurations of points which do not allow us to obtain a unique solution for the estimation of the Fundamental matrix, there are critical sequences of motions which do not allow us to obtain a unique solution for the estimation of the camera parameters. We have seen one of them, motions with parallel rotation axis and arbitrary translations. Other configurations are examined in Section 11.6.

1.26 References and further reading

Three dimensional problems involving several images have traditionally been studied under the assumption that the cameras are calibrated, with a few exceptions (the first use of projective geometry to analyze two-view geometry seems to be an unnoticed paper by Thompson (Thompson, 1968)). In the early 90's, Forsyth, Mundy *et al.* (Forsyth et al., 1990) in the context of object recognition, Koenderink and Van Doorn (Koenderink and van Doorn, 1991) in the context of structure from motion, and Barrett *et al.* (Barrett et al., 1992) in the context of image transfer, discovered that useful tasks could be performed using non-metric representations. Following the results of Faugeras (Faugeras, 1992) and Hartley *et al.* (Hartley et al., 1992) on projective reconstruction, an enormous burst of research was launched.

A decade later, there are so many papers that it has not been possible to cover all the important topics. In this book, we concentrate on the geometry of reconstruction and positioning from finite correspondences.

Invariants and object recognition The approach developped in this book relies on the fact that the appearance of an object changes as the viewpoint changes in order to reconstruct the object and the camera motion. On the other hand this variation is one of the fundamental difficulties in recognizing objects from images. Another approach, which one might call the invariance program, seeks to overcome the problem that the appearance of an object depends on viewpoint by using geometric descriptions which are unaffected by the imaging transformation. These invariant measures can be used to index a library of object models for recognition. Many papers representative of this line of research can be found in the book edited by Mundy and Zisserman (Mundy and Zisserman, 1992). The follow-up of this book (Mundy et al., 1994) also has several papers which deal with the reconstruction approach. The two approaches are complementary in the sense that one of them concentrates on relations within the system of cameras while the other concentrates on relations within configurations of 3-D points.

There are numerous invariant descriptions which can be measured from images without any prior knowledge of the position, orientation and calibration of the camera. However, a fundamental limitation of the approach is that no *general* invariants of point sets can be measured from a single image (Burns et al., 1993; Barrett et al., 1991; Clemens and Jacobs, 1991). This means that we need either additional knowledge about the object (such as symmetry (Rothwell et al., 1993), which is the same as having two mirrored images of the same object, or planarity (Forsyth et al., 1990)) or multiple views. A significant part of the survey paper on object recognition using invariance of Zisserman *et al.* (Zisserman et al., 1995b) is a summary of results on the construction of invariants for 3-D objects from a single perspective view. Using continuous invariant descriptors can yield other difficulties. For example, all curves map arbitrarily close to a circle by projective

transformations (Astrom, 1995).

If multiple views are used, the approach that we describe in the book can be applied. An important unifying result of Carlsson and Weinshall (Carlsson and Weinshall, 1998; Carlsson, 1995; Weinshall et al., 1996) is the fundamental duality of the 3-D motion estimation and structure estimation problem. They show that for points and cameras in general position, the problem of computing camera geometry from N points in M views is equivalent to the problem of reconstructing $M + 4$ points in $N - 4$ views.

Infinitesimal displacements In this book, we concentrate on the general case when the displacements are finite. The appropriate data is discrete point correspondences. When the displacements are small, the cameras are closely spaced and some of the image projections are nearly the same. Then some quantities become infinitesimal so that the approximation by a differential analysis is appropriate. Verri and Trucco (Verri and Trucco, 1999) propose a differential technique based on optical flow for estimating the location of the epipole. The method requires a minimum of six points. It is based on a new rewrite of the optical flow equations in terms of a generalization of the time-to-impact, and without decoupling rotation and translation.

The relationship between optical flow and 3-D structure is well understood in the calibrated case (Koenderink and van Doorn, 1975; Longuet-Higgins and Prazdny, 1980; Maybank, 1987). The uncalibrated case was first investigated by Viéville and Faugeras (Viéville and Faugeras, 1996), who used the first-order expansion of the motion equation between two views and analyzed the observability of the infinitesimal quantities. A more ambitious approach was taken by Astrom and Heyden (Astrom and Heyden, 1998). They consider the N-views case and take a Taylor expansion of the projection equations By gathering the projection equations by the powers of Δt using a principle similar to the one described in Section 10.2, they obtain multi-linear constraints which link corresponding points and their derivatives. Recently, Triggs (Triggs, 1999a) proposed a more tractable approach to the N-views case, which bridges the gap between the differential and the discrete formulation, using finite difference expansions.

Brooks *et al.* (Brooks et al., 1997) derived, similarly to Viéville and Faugeras, a differential epipolar equation for uncalibrated optical flow. This equation incorporates two matrices which encode information about the ego-motion and intrinsic parameters of the camera. Given enough points, the composite ratio of some entries of these matrices are determined and, under some conditions, a closed form formula is obtained from these ratios. In (Brooks et al., 1998), a method is presented for the robust determination of the two matrices. The problem of self-calibration from image derivatives was also addressed by Brodsky et al (Brodsky et al., 1998).

Illumination and photometry In this book, we concentrate on geometry, which is only one of the attributes of the 3-D world. Some researchers have begun to combine geometry and photometry. For example, Belhumeur and Kriegman show that from perspective images of a scene where the camera is in fixed viewpoint, but where point light sources vary in each image, one can only reconstruct the surface up to a family of projective transformations from shadows (Kriegman and Belhumeur, 1998), whereas from orthographic images and light sources at infinity, the family of transformation is restricted to affine transformation, and that for Lambertian surfaces one can only reconstruct the surface up to this family of affine transformations (Belhumeur et al., 1997).