Chapter 11

# THE GEOMETRY OF MULTIPLE VIEWS

Despite the wealth of information contained in a photograph, the depth of a scene point along the corresponding projection ray is not directly accessible in a single image. With at least two pictures, on the other hand, depth can be measured through triangulation. This is of course one of the reasons why most animals have at least two eyes and/or move their head when looking for friend or foe, as well as the motivation for equipping autonomous robots with stereo or motion analysis systems. Before building such a program, we must understand how several views of the same scene constrain its three-dimensional structure as well as the corresponding camera configurations. This is the goal of this chapter.

In particular, we will elucidate the geometric and algebraic constraints that hold among two, three, or more views of the same scene. In the familiar setting of binocular stereo vision, we will show that the first image of any point must lie in the plane formed by its second image and the optical centers of the two cameras. This *epipolar constraint* can be represented algebraically by a $3 \times 3$ matrix called the *essential matrix* when the intrinsic parameters of the cameras are known, and the *fundamental matrix* otherwise. Three pictures of the same line will introduce a different constraint, namely that the intersection of the planes formed by their preimages be degenerate. Algebraically, this geometric relationship can be represented by a $3 \times 3 \times 3$ *trifocal tensor*. More images will introduce additional constraints, for example four projections of the same point will satisfy certain quadrilinear relations whose coefficients are captured by the *quadrifocal tensor*, etc. Remarkably, the equations satisfied by multiple pictures of the same scene feature can be set up without any knowledge of the cameras and the scene they observe, and a number of methods for estimating their parameters directly from image data will be presented in this chapter.

Computer vision is not the only scientific field concerned with the geometry of multiple views: the goal of photogrammetry, already mentioned in Chapter 5, is precisely to recover quantitative geometric information from multiple pictures. Applications of the epipolar and trifocal constraints to the classical photogrammetry problem of *transfer* (i.e., the prediction of the position of a point in an image given its position in a number of reference pictures) will be briefly discussed in this chapter, along with some examples. Many more applications in the domains of stereo and motion analysis will be presented in latter chapters.

## 11.1 Two Views

### 11.1.1 Epipolar Geometry

Consider the images $p$ and $p'$ of a point $P$ observed by two cameras with optical centers $O$ and $O'$. These five points all belong to the *epipolar plane* defined by the two intersecting rays $OP$ and $O'P$ (Figure 11.1). In particular, the point $p'$ lies on the line $l'$ where this plane and the retina $\Pi'$ of the second camera intersect. The line $l'$ is the *epipolar line* associated with the point $p$, and it passes through the point $e'$ where the *baseline* joining the optical centers $O$ and $O'$ intersects $\Pi'$. Likewise, the point $p$ lies on the epipolar line $l$ associated with the point $p'$, and this line passes through the intersection $e$ of the baseline with the plane $\Pi$.

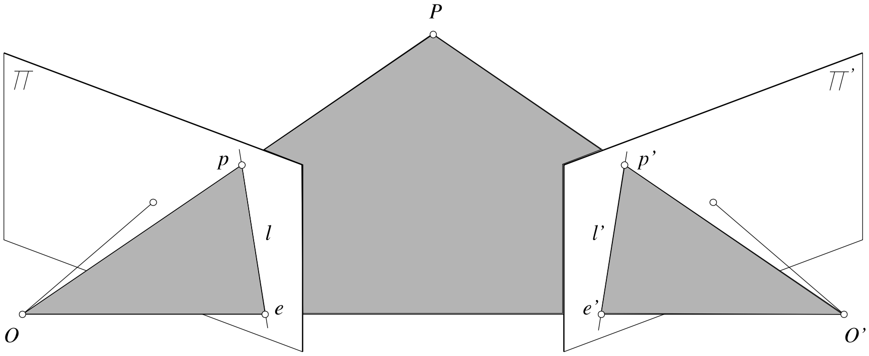

**Figure 11.1.** Epipolar geometry: the point $P$, the optical centers $O$ and $O'$ of the two cameras, and the two images $p$ and $p'$ of $P$ all lie in the same plane.

The points $e$ and $e'$ are called the *epipoles* of the two cameras. The epipole $e'$ is the (virtual) image of the optical center $O$ of the first camera in the image observed by the second camera, and vice versa. As noted before, if $p$ and $p'$ are images of the same point, then $p'$ must lie on the epipolar line associated with $p$. This *epipolar constraint* plays a fundamental role in stereo vision and motion analysis.

Let us assume for example that we know the intrinsic and extrinsic parameters of the two cameras of a stereo rig. We will see in Chapter 12 that the most difficult part of stereo data analysis is establishing correspondences between the two images, i.e., deciding which points in the right picture match the points in the left one. The epipolar constraint greatly limits the search for these correspondences: indeed, since we assume that the rig is calibrated, the coordinates of the point $p$ completely determine the ray joining $O$ and $p$, and thus the associated epipolar plane $OO'p$ and epipolar line. The search for matches can be restricted to this line instead of the whole image (Figure 11.2). In two-frame motion analysis on the other hand, each camera may be internally calibrated, but the rigid transformation separating the two camera coordinate systems is unknown. In this case, the epipolar geometry obviously constrains the set of possible motions. The next sections explore several variants of this situation.

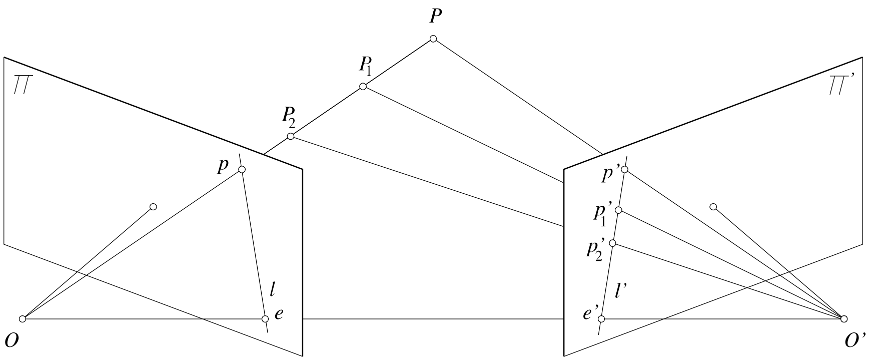

**Figure 11.2.** Epipolar constraint: given a calibrated stereo rig, the set of possible matches for the point $p$ is constrained to lie on the associated epipolar line $l'$.

### 11.1.2 The Calibrated Case

Here we assume that the intrinsic parameters of each camera are known, so $\boldsymbol{p} = \hat{\boldsymbol{p}}$. Clearly, the epipolar constraint implies that the three vectors $\overrightarrow{Op}$, $\overrightarrow{O'p'}$, and $\overrightarrow{OO'}$ are coplanar. Equivalently, one of them must lie in the plane spanned by the other two, or

$$\overrightarrow{Op} \cdot [\overrightarrow{OO'} \times \overrightarrow{O'p'}] = 0.$$

We can rewrite this coordinate-independent equation in the coordinate frame associated to the first camera as

$$\boldsymbol{p} \cdot [\boldsymbol{t} \times (\mathcal{R}\boldsymbol{p}')], \tag{11.1.1}$$

where $\boldsymbol{p} = (u, v, 1)^T$ and $\boldsymbol{p}' = (u', v', 1)^T$ denote the homogenous image coordinate vectors of $p$ and $p'$, $\boldsymbol{t}$ is the coordinate vector of the translation $\overrightarrow{OO'}$ separating the two coordinate systems, and $\mathcal{R}$ is the rotation matrix such that a free vector with coordinates $\boldsymbol{w}'$ in the second coordinate system has coordinates $\mathcal{R}\boldsymbol{w}'$ in the first one (in this case the two projection matrices are given in the coordinate system attached to the first camera by $(\,\text{Id} \quad \mathbf{0}\,)$ and $(\,\mathcal{R}^T, -\mathcal{R}^T\boldsymbol{t}\,)$).

Equation (11.1.1) can finally be rewritten as

$$\boldsymbol{p}^T \mathcal{E} \boldsymbol{p}' = 0, \tag{11.1.2}$$

where $\mathcal{E} = [\boldsymbol{t}_\times]\mathcal{R}$, and $[\boldsymbol{a}_\times]$ denotes the skew-symmetric matrix such that $[\boldsymbol{a}_\times]\boldsymbol{x} = \boldsymbol{a}\times\boldsymbol{x}$ is the cross-product of the vectors $\boldsymbol{a}$ and $\boldsymbol{x}$. The matrix $\mathcal{E}$ is called the *essential matrix*, and it was first introduced by Longuet-Higgins [**?**]. Its nine coefficients are only defined up to scale, and they can be parameterized by the three degrees of freedom of the rotation matrix $\mathcal{R}$ and the two degrees of freedom defining the direction of the translation vector $\boldsymbol{t}$.

Note that $\mathcal{E}\boldsymbol{p}'$ can be interpreted as the coordinate vector representing the epipolar line associated with the point $p'$ in the first image: indeed, an image line $l$ can be defined by its equation $au + bv + c = 0$, where $(u, v)$ denote the coordinates of a point on the line, $(a, b)$ is the unit normal to the line, and $-c$ is the (signed) distance between the origin and $l$. Alternatively, we can define the line equation in terms of the homogeneous coordinate vector $\boldsymbol{p} = (u, v, 1)^T$ of a point on the line and the vector $\boldsymbol{l} = (a, b, c)^T$ by $\boldsymbol{l}\cdot\boldsymbol{p} = 0$, in which case the constraint $a^2 + b^2 = 1$ is relaxed since the equation holds independently of any scale change applied to $\boldsymbol{l}$. In this context, (11.1.2) expresses the fact that the point $p$ lies on the epipolar line associated with the vector $\mathcal{E}\boldsymbol{p}'$. By symmetry, it is also clear that $\mathcal{E}^T\boldsymbol{p}$ is the coordinate vector representing the epipolar line associated with $p$ in the second image.

It is obvious that essential matrices are singular since $\boldsymbol{t}$ is parallel to the coordinate vector $\boldsymbol{e}$ of the left epipole, so that $\mathcal{E}^T\boldsymbol{e} = -\mathcal{R}^T[\boldsymbol{t}_\times]\boldsymbol{e} = 0$. Likewise, it is easy to show that $\boldsymbol{e}'$ is a zero eigenvector of $\mathcal{E}$. As shown by Huang and Faugeras [**?**], essential matrices are in fact characterized by the fact that they are singular with two equal non-zero singular values (see exercises).

### 11.1.3 Small Motions

Let us now turn our attention to *infinitesimal* displacements. We consider a moving camera with translational velocity $\boldsymbol{v}$ and rotational velocity $\boldsymbol{\omega}$ and rewrite (11.1.2) for two frames separated by a small time interval $\delta t$. Let us denote by $\dot{\boldsymbol{p}} = (\dot{u}, \dot{v}, 0)^T$ the velocity of the point $p$, or *motion field*. Using the exponential representation of rotations,[1] we have (to first order):

$$\begin{cases} \boldsymbol{t} = \delta t\,\boldsymbol{v}, \\ \mathcal{R} = \mathrm{Id} + \delta t\,[\boldsymbol{\omega}_\times], \\ \boldsymbol{p}' = \boldsymbol{p} + \delta t\,\dot{\boldsymbol{p}}. \end{cases}$$

Substituting in (11.1.2) yields

$$\boldsymbol{p}^T[\boldsymbol{v}_\times](\mathrm{Id} + \delta t\,[\boldsymbol{\omega}_\times])(\boldsymbol{p} + \delta t\,\dot{\boldsymbol{p}}) = 0,$$

and neglecting all terms of order two or greater in $\delta t$ yields:

$$\boldsymbol{p}^T([\boldsymbol{v}_\times][\boldsymbol{\omega}_\times])\boldsymbol{p} - (\boldsymbol{p}\times\dot{\boldsymbol{p}})\cdot\boldsymbol{v} = 0. \qquad (11.1.3)$$

[1] The matrix associated with the rotation whose axis is the unit vector $\boldsymbol{a}$ and whose angle is $\theta$ can be shown to be equal to $e^{\theta[\boldsymbol{a}_\times]} \overset{\text{def}}{=} \sum_{i=0}^{+\infty} \frac{1}{i!}(\theta[\boldsymbol{a}_\times])^i$.

Equation (11.1.3) is simply the instantaneous form of the Longuet-Higgins relation (11.1.2) which captures the epipolar geometry in the discrete case. Note that in the case of pure translation we have $\boldsymbol{\omega} = 0$, thus $(\boldsymbol{p} \times \dot{\boldsymbol{p}}) \cdot \boldsymbol{v} = 0$. In other words, the three vectors $\boldsymbol{p} = \overrightarrow{op}$, $\dot{\boldsymbol{p}}$ and $\boldsymbol{v}$ must be coplanar. If $e$ denotes the infinitesimal epipole, or *focus of expansion*, i.e., the point where the line passing through the optical center and parallel to the velocity vector $\boldsymbol{v}$ pierces the image plane, we obtain the well known result that the motion field points toward the focus of expansion under pure translational motion (Figure 11.3).

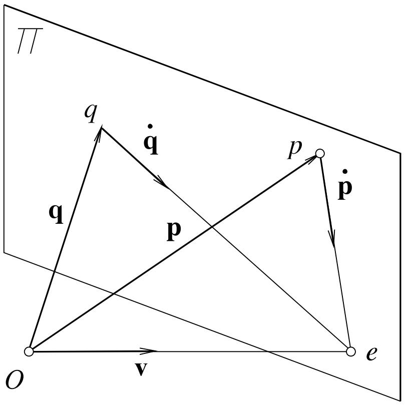

**Figure 11.3.** Focus of expansion: under pure translation, the motion field at every point in the image points toward the focus of expansion.

### 11.1.4 The Uncalibrated Case

The Longuet-Higgins relation holds for *internally calibrated* cameras, whose intrinsic parameters are known so that image positions can be expressed in normalized coordinates. When these parameters are unknown (*uncalibrated* cameras), we can write $\boldsymbol{p} = \mathcal{K}\hat{\boldsymbol{p}}$ and $\boldsymbol{p}' = \mathcal{K}'\hat{\boldsymbol{p}}'$, where $\mathcal{K}$ and $\mathcal{K}'$ are $3 \times 3$ calibration matrices, and $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{p}}'$ are normalized image coordinate vectors. The Longuet-Higgins relation holds for these vectors, and we obtain

$$\boldsymbol{p}^T \mathcal{F} \boldsymbol{p}' = 0, \tag{11.1.4}$$

where the matrix $\mathcal{F} = \mathcal{K}^{-T}\mathcal{E}\mathcal{K}'^{-1}$, called the *fundamental matrix*, is not, in general, an essential matrix.[2] It has again rank two, and the eigenvector of $\mathcal{F}$ (resp. $\mathcal{F}^T$) corresponding to its zero eigenvalue is as before the position $\boldsymbol{e}'$ (resp. $\boldsymbol{e}$) of the epipole. Note that $\mathcal{F}\boldsymbol{p}'$ (resp. $\mathcal{F}^T\boldsymbol{p}$) represents the epipolar line corresponding to the point $\boldsymbol{p}'$ (resp. $\boldsymbol{p}$) in the first (resp. second) image.

[2] Small motions can also be handled in the uncalibrated setting. In particular, Viéville and Faugeras [?] have derived an equation similar to (11.1.3) that characterizes the motion field of a camera with varying intrinsic parameters.

The rank-two constraint means that the fundamental matrix only admits seven independent parameters. Several choices of parameterization are possible, but the most natural one is in terms of the coordinate vectors $\boldsymbol{e} = (\alpha, \beta)^T$ and $\boldsymbol{e}' = (\alpha', \beta')^T$ of the two epipoles, and of the so-called *epipolar transformation* that maps one set of epipolar lines onto the other one. We will examine the properties of the epipolar transformation in Chapter 14 in the context of structure from motion. For the time being, let us just note (without proof) that this transformation is parameterized by four numbers $a, b, c, d$, and that the fundamental matrix can be written as

$$\mathcal{F} = \begin{pmatrix} b & a & -a\beta - b\alpha \\ -d & -c & c\beta + d\alpha \\ d\beta' - b\alpha' & c\beta' - a\alpha' & -c\beta\beta' - d\beta'\alpha + a\beta\alpha' + b\alpha\alpha' \end{pmatrix}. \tag{11.1.5}$$

### 11.1.5 Weak Calibration

As mentioned earlier, the essential matrix is defined up to scale by five independent parameters. It is therefore possible (at least in principle), to calculate it by writing (11.1.2) for five point correspondences. Likewise, the fundamental matrix is defined by seven independent coefficients (the parameters $a, b, c, d$ in (11.1.5) are only defined up to scale) and can in principle be estimated from seven point correspondences. Methods for estimating the essential and fundamental matrices from a minimal number of parameters indeed exist (see [**?**] and Section 11.4), but they are far too involved to be described here. This section addresses the simpler problem of estimating the epipolar geometry from a redundant set of point correspondences between two images taken by cameras with unknown intrinsic parameters, a process known as *weak calibration*.

Note that the epipolar constraint (11.1.4) is a linear equation in the nine coefficients of the fundamental matrix $\mathcal{F}$:

$$(u, v, 1) \begin{pmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \end{pmatrix} \begin{pmatrix} u' \\ v' \\ 1 \end{pmatrix} = 0 \Leftrightarrow (uu', uv', u, vu', vv', v, u', v', 1) \begin{pmatrix} F_{11} \\ F_{12} \\ F_{13} \\ F_{21} \\ F_{22} \\ F_{23} \\ F_{31} \\ F_{32} \\ F_{33} \end{pmatrix} = 0. \tag{11.1.6}$$

Since (11.1.6) is homogeneous in the coefficients of $\mathcal{F}$, we can for example set $F_{33} = 1$ and use eight point correspondences $p_i \leftrightarrow p'_i$ $(i = 1, .., 8)$ to set up an $8 \times 8$

system of non-homogeneous linear equations:

$$\begin{pmatrix} u_1u'_1 & u_1v'_1 & u_1 & v_1u'_1 & v_1v'_1 & v_1 & u'_1 & v'_1 \\ u_2u'_2 & u_2v'_2 & u_2 & v_2u'_2 & v_2v'_2 & v_2 & u'_2 & v'_2 \\ u_3u'_3 & u_3v'_3 & u_3 & v_3u'_3 & v_3v'_3 & v_3 & u'_3 & v'_3 \\ u_4u'_4 & u_4v'_4 & u_4 & v_4u'_4 & v_4v'_4 & v_4 & u'_4 & v'_4 \\ u_5u'_5 & u_5v'_5 & u_5 & v_5u'_5 & v_5v'_5 & v_5 & u'_5 & v'_5 \\ u_6u'_6 & u_6v'_6 & u_6 & v_6u'_6 & v_6v'_6 & v_6 & u'_6 & v'_6 \\ u_7u'_7 & u_7v'_7 & u_7 & v_7u'_7 & v_7v'_7 & v_7 & u'_7 & v'_7 \\ u_8u'_8 & u_8v'_8 & u_8 & v_8u'_8 & v_8v'_8 & v_8 & u'_8 & v'_8 \end{pmatrix} \begin{pmatrix} F_{11} \\ F_{12} \\ F_{13} \\ F_{21} \\ F_{22} \\ F_{23} \\ F_{31} \\ F_{32} \end{pmatrix} = - \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

which is sufficient for estimating the fundamental matrix. This is the *eight-point* algorithm proposed by Longuet-Higgins [**?**] in the case of calibrated cameras. It will fail when the associated $8 \times 8$ matrix is singular. As shown in [**?**] and the exercises, this will only happen, however, when the eight points and the two optical centers lie on a quadric surface. Fortunately, this is quite unlikely since a quadric surface is completely determined by nine points, which means that there is in general no quadric that passes through ten arbitrary points.

When $n > 8$ correspondences are available, $\mathcal{F}$ can be estimated using linear least squares by minimizing

$$\sum_{i=1}^{n} (\boldsymbol{p}_i^T \mathcal{F} \boldsymbol{p}'_i)^2 \tag{11.1.7}$$

with respect to the coefficients of $\mathcal{F}$ under the constraint that the vector formed by these coefficients has unit norm.

Note that both the eight-point algorithm and its least-squares version ignore the rank-two property of fundamental matrices.[3] To enforce this constraint, Luong *et al.* [**?**; **?**] have proposed to use the matrix $\mathcal{F}$ output by the eight-point algorithm as the basis for a two-step estimation process: first, use linear least squares to compute the position of the epipoles $\boldsymbol{e}$ and $\boldsymbol{e}'$ that minimize $|\mathcal{F}^T\boldsymbol{e}|^2$ and $|\mathcal{F}\boldsymbol{e}'|^2$; second, substitute the coordinates of these points in (11.1.5): this yields a linear parameterization of the fundamental matrix by the coefficients of the epipolar transformation, which can now be estimated by minimizing (11.1.7) via linear least squares.

The least-squares version of the eight-point algorithm minimizes the mean-squared *algebraic distance* associated with the epipolar constraint, i.e., the mean-squared value of $e(\boldsymbol{p}, \boldsymbol{p}') = \boldsymbol{p}^T \mathcal{F} \boldsymbol{p}'$ calculated over all point correspondences. This error function admits a geometric interpretation: in particular, we have

$$e(\boldsymbol{p}, \boldsymbol{p}') = \lambda \mathrm{d}(\boldsymbol{p}, \mathcal{F}\boldsymbol{p}') = \lambda' \mathrm{d}(\boldsymbol{p}', \mathcal{F}^T \boldsymbol{p}),$$

where $d(\boldsymbol{p}, \boldsymbol{l})$ denotes the (signed) Euclidean distance between the point $\boldsymbol{p}$ and the line $\boldsymbol{l}$, and $\mathcal{F}\boldsymbol{p}$ and $\mathcal{F}^T\boldsymbol{p}'$ are the epipolar lines associated with $\boldsymbol{p}$ and $\boldsymbol{p}'$. The scale factors $\lambda$ and $\lambda'$ are simply the norms of the vectors formed by the first two

[3] The original algorithm proposed by Longuet-Higgins ignores the fact that essential matrices have rank two and two equal singular values as well.

components of $\mathcal{F}\boldsymbol{p}'$ and $\mathcal{F}^T\boldsymbol{p}$, and their dependence on the pair of data points observed may bias the estimation process.

It is of course possible to get rid of the scale factors and directly minimize the mean-squared distance between the image points and the corresponding epipolar lines, i.e.,

$$\sum_{i=1}^{n}[\mathrm{d}^2(\boldsymbol{p}_i, \mathcal{F}\boldsymbol{p}'_i) + \mathrm{d}^2(\boldsymbol{p}'_i, \mathcal{F}^T\boldsymbol{p}_i)].$$

This is a non-linear problem, regardless of the parameterization chosen for the fundamental matrix, but the minimization can be initialized with the result of the eight-point algorithm. This method was first proposed by Luong *et al.* [**?**], and it has been shown to provide results vastly superior to those obtained using the eight-point method.

Recently, Hartley [**?**] has proposed a normalized eight-point algorithm and has also reported excellent results. His approach is based on the observation that the poor performance of the plain eight-point method is due, for the most part, to poor numerical conditioning. Thus Hartley has proposed to translate and scale the data so it is centered at the origin and the average distance to the origin is $\sqrt{2}$ pixel. This dramatically improves the conditioning of the linear least-squares estimation process. Accordingly, his method is divided into four steps: first, transform the image coordinates using appropriate translation and scaling operators $\mathcal{T} : \boldsymbol{p}_i \to \tilde{\boldsymbol{p}}_i$ and $\mathcal{T}' : \boldsymbol{p}'_i \to \tilde{\boldsymbol{p}}'_i$. Second, use linear least squares to compute the matrix $\tilde{\mathcal{F}}$ minimizing

$$\sum_{i=1}^{n}(\tilde{\boldsymbol{p}}_i^T \tilde{\mathcal{F}} \tilde{\boldsymbol{p}}'_i)^2.$$

Third, enforce the rank-two constraint; this can be done using the two-step method of Luong *et al.* described earlier, but Hartley uses instead a technique, suggested by Tsai and Huang [**?**] in the calibrated case, which constructs the *singular value decomposition* $\tilde{\mathcal{F}} = \mathcal{U}\mathcal{S}\mathcal{V}^T$ of $\tilde{\mathcal{F}}$. Here, $\mathcal{S} = \mathrm{diag}(r, s, t)$ is a diagonal $3 \times 3$ matrix with entries $r \geq s \geq t$, and $\mathcal{U}, \mathcal{V}$ are orthogonal $3 \times 3$ matrices.[4] The rank-two matrix $\bar{\mathcal{F}}$ minimizing the Frobenius norm of $\tilde{\mathcal{F}} - \bar{\mathcal{F}}$ is simply $\bar{\mathcal{F}} = \mathcal{U}\mathrm{diag}(r, s, 0)\mathcal{V}^T$ [**?**]. Fourth, set $\mathcal{F} = \mathcal{T}^T\bar{\mathcal{F}}\mathcal{T}'$. This is the final estimate of the fundamental matrix.

Figure 11.4 shows weak calibration experiments using as input data a set of 37 point correspondences between two images of a toy house. The data points are shown in the figure as small discs, and the recovered epipolar lines are shown as short line segments. The top of the figure shows the output of the least-squares version of the plain eight-point algorithm, and the bottom part of the figure shows the results obtained using Hartley's variant of this method. As expected, the results are much better in the second case, and in fact extremely close to those obtained using the distance minimization criterion of Luong *et al.* [**?**; **?**].

[4] Singular value decomposition will be discussed in detail in Chapter 13.