Computer Vision

CPS 296.1 Supplementary Lecture Notes

Carlo Tomasi – Duke University

Fall 2007

5. Stereo Vision

[Last Modified on October 16, 2007]

Warning. These notes are being written as the course progresses. As errors are discovered, old notes will be corrected and the new version will be posted online.

5 Stereo Vision

5.1 Introduction

The input to stereo vision is a set of two or more images of the same scene taken at the same time from cameras in different positions. The output is a *depth map*, that is, a function $Z_{c1}(\mathbf{x}_1)$ of position in image number 1 that gives the Z_c coordinate of the point visible at $\mathbf{x}_1$ in the reference frame of camera 1.

This Section considers only the two-camera stereo problem. Several papers on multiview stereo exist as well [3, 1, 5, 6].

The positions and orientations of the two cameras in the world are assumed to be known, and can be determined as discussed in the Section on camera calibration. Typically, the world reference frame for stereo is attached to a *rig* to which the cameras are solidly mounted. As a consequence, the rig can be moved, together with the world reference frame, without changing the positions and orientations of the cameras relative to each other. Because of this choice, it is more appropriate to rename the world reference system to be the *stereo rig reference system*.

Even more specifically, the stereo rig reference system is usually made to coincide with the reference system of camera 1. This choice is adopted here as well. Because of this, the depth map can be rewritten more simply as follows:

$$Z_{c1}(\mathbf{x}_1) = Z(\mathbf{x}_1) \ .$$

A point with coordinates $\mathbf{X}$ in the rig reference system projects to point $\mathbf{x}_1$ in image 1 and $\mathbf{x}_2$ in image 2. These two point are said to *correspond* to each other. From the differences in the image positions of corresponding points, a computation called *triangulation* determines the coordinates $\mathbf{X}$ of the point in the world, relative to the rig reference system.

Triangulation is straightforward geometry, while determining pairs of corresponding points is a hard task, the so-called *correspondence problem*. Because the relative position of the two cameras is known, it turns out that the point $\mathbf{x}_2$ corresponding to a given point $\mathbf{x}_1$ is constrained to be on a single, known line in image 2, called the *epipolar line* of point $\mathbf{x}_1$. This relationship holds for images that are immune from distortion, so the first step is to correct the images using knowledge of the distortion coefficients k_2 and k_4, as shown in the Section on camera calibration. After this step, straight lines in the world map to straight lines in the images. The geometry of the epipolar line constraint is explained in Section 5.2.

It is always possible to transform the two images from a stereo pair so that epipolar lines coincide with rows in the image arrays. This transformation is called rectification, and is discussed in Section 5.3. Rectification simplifies the programming of both correspondence and triangulation algorithms, and is therefore performed almost always.

Section 5.4 discusses triangulation for both a general and a rectified camera pair. Section 5.5 tackles the hard correspondence problem.

5.2 Epipolar Geometry

Figure 5.1 shows the main elements of the *epipolar geometry* of a stereo pair.

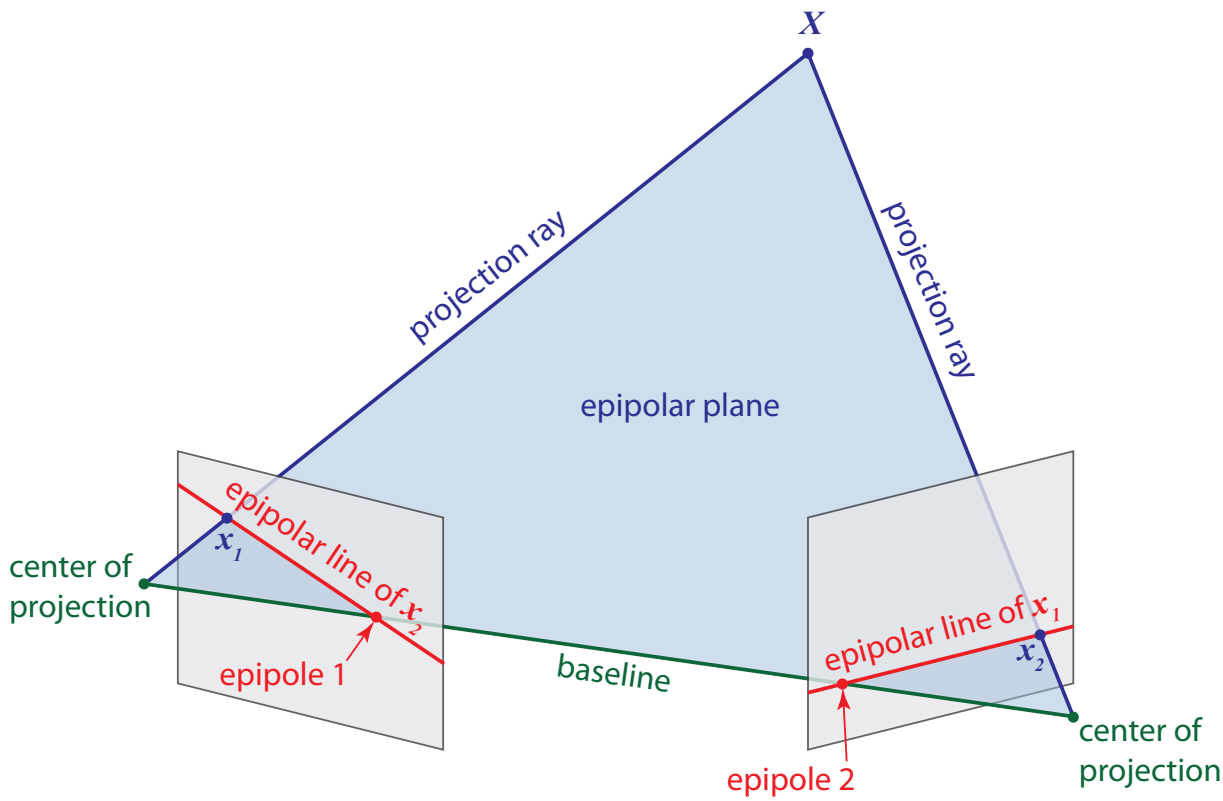

Figure 5.1: Essential elements of the epipolar geometry of a stereo pair.

The world point $\mathbf{X}$ and the centers of projection of the two cameras identify a plane in space, the *epipolar plane* of point $\mathbf{X}$. The Figure shows a triangle of this plane, delimited by the two projection rays and by the *baseline* of the stereo pair, that is, the line segment that connects the two centers of projection.

If the image planes are thought of extending indefinitely, the baseline intersects the two image planes at two points called the *epipoles* of the two images. In particular, if the baseline is parallel to an image plane, then the corresponding epipole is a point at infinity.

The epipoles are fixed points for a given camera pair configuration. With cameras somewhat tilted towards each other, and with a sufficiently wide field of view, the epipoles would be image points. Epipole 1 would be literally the image of the center of projection of camera 2 in image 1, and viceversa. Even if the two cameras do not physically see each other, we maintain this description in an abstract sense: each epipole is the image of one camera in the other image. Note that the epipole takes its number from the image in which it appears, not from the camera of which it is the image.

The epipolar plane intersects the two image planes along the two *epipolar lines* of point $\mathbf{X}$, which by construction pass through the two projection points $\mathbf{x}_1$ and $\mathbf{x}_2$ and through the two epipoles. Thus, epipolar lines come in corresponding pairs, and the correspondence is established

by the single epipolar plane for the given point $\mathbf{X}$.

For a different world point $\mathbf{X}$, the epipolar plane changes, and with it do the image projections of $\mathbf{X}$ and the epipolar lines. However, all epipolar planes contain the baseline. Thus, the set of epipolar planes forms a *pencil* of planes, supported by the line through the baseline.[1]

Suppose now that we are given the two images and point $\mathbf{x}_1$ in image 1. We do not know where the corresponding point $\mathbf{x}_2$ is in the other image, nor where the world point $\mathbf{X}$ is[2]. However, the two centers of projection and point $\mathbf{x}_1$ identify the epipolar plane, and this in turn determines the epipolar line of point $\mathbf{x}_1$ in image 2. The point $\mathbf{x}_2$ must be somewhere on this line, so this construction reduces the search for $\mathbf{x}_2$ from a rectangle (image 2 in its entirety) to a line segment (the part of epipolar line contained in image 2).

This same construction holds for any other point $\mathbf{x}_1$ on the epipolar line in image 1, so the search for stereo correspondences is most efficiently performed for one pair of corresponding epipolar lines at a time.

To find a mathematical relationship between two corresponding epipolar lines, let R, $\mathbf{t}$, be the transformation of coordinates from camera 1 to camera 2. This means that the center of projection of camera 2 is at $\mathbf{t}$ in the system of reference of camera 1, and the rows $\mathbf{i}^T$, $\mathbf{j}^T$, $\mathbf{k}^T$ of the rotation matrix R are the unit vectors of the axes of the camera 2 reference system, expressed in the reference system of camera 1. In other words, a point with coordinates $\mathbf{X}_1$ in the system of reference of camera 1 has the following coordinates in the system of reference of camera 2:

$$\mathbf{X}_2 = R(\mathbf{X}_1 - \mathbf{t}) \ . \tag{1}$$

and conversely

$$\mathbf{X}_1 = R^T\mathbf{X}_2 + \mathbf{t} \ . \tag{2}$$

Then the vector $\mathbf{t}$ is along and has the same length as the baseline.

The projection rays for two arbitrary points in the two images are generically two skew lines in space. The projection rays of two corresponding points, on the the other hand, are coplanar with the baseline. This observation leads to the relationship between epipolar lines as follows.

When expressed in the frame of camera 1, the directions of the projection rays through corresponding image points $\mathbf{x}_{c1}$ and $\mathbf{x}_{c2}$ are along the vectors

$$\mathbf{p}_{c1} \quad \text{and} \quad R^T\mathbf{p}_{c2}$$

where[3]

$$\mathbf{p}_{c1} = \begin{bmatrix} \mathbf{x}_{c1} \\ -1 \end{bmatrix} \quad \text{and} \quad \mathbf{p}_{c2} = \begin{bmatrix} \mathbf{x}_{c2} \\ -1 \end{bmatrix} \ .$$

Coplanarity can be expressed by stating that the triple product of $R^T\mathbf{p}_{c2}$, $\mathbf{t}$, and $\mathbf{p}_{c1}$ is zero:

$$(R^T\mathbf{p}_{c2})^T(\mathbf{t} \times \mathbf{p}_{c1}) = 0$$

[1] We use the term "baseline" for the line *segment*. However, this term is also often used for the entire *line* through the two centers of projection.

[2] Except that $\mathbf{X}$ must be somewhere along the projection ray of $\mathbf{x}_1$.

[3] The vectors $\mathbf{p}_{c1}$ and $\mathbf{p}_{c2}$ can be viewed as the version in homogeneous coordinates of $\mathbf{x}_{c1}$ and $\mathbf{x}_{c2}$.

or

$$\mathbf{p}_{c2}^T R(\mathbf{t} \times \mathbf{p}_{c1}) = 0 \ .$$

The cross product int parentheses is a vector $\mathbf{v}$ orthogonal to the baseline and the first projection ray. The inner product of $R^T\mathbf{p}_{c2}$ and $\mathbf{v}$ being zero expresses the fact that these two vectors are mutually orthogonal, so that $R^T\mathbf{p}_{c2}$ is in the plane of $\mathbf{t}$ and $\mathbf{p}_{c1}$ (literally, $R^T\mathbf{p}_{c2}$ is orthogonal to a vector $\mathbf{v}$ that is orthogonal to both $\mathbf{t}$ and $\mathbf{p}_{c1}$).

As discussed in the Math supplement, the cross product of a vector $\mathbf{t}$ with another vector can be put in matrix form:

$$\mathbf{t} \times \mathbf{p}_{c1} = \mathbf{t}_\times \mathbf{p}_{c1} \quad \text{where} \quad \mathbf{t}_\times = \begin{bmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{bmatrix}$$

so that

$$\mathbf{p}_{c2}^T E \mathbf{p}_{c1} = 0 \quad \text{where} \quad E = R\mathbf{t}_\times \ . \tag{3}$$

The first equation in (3) is called the *epipolar constraint* and the matrix E is called the *essential matrix*.

The key property of the essential matrix is that it lets us find pairs of corresponding epipolar lines in a stereo pair. To this end, consider the geometric meaning of equation (3): This equation expresses the coplanarity between two specific points, $\mathbf{p}_{c1}$ and $\mathbf{p}_{c2}$, on the same epipolar plane. However, the reasoning used to derive that equation does not depend on *which* two points are considered, as long as they are on the same plane. In other words, the equation must hold for *any* pair of points that are on the same epipolar plane. This of course includes any pair of *image* points that are on the same epipolar plane. For instance, if $\mathbf{x}_{c1}$ (and therefore $\mathbf{p}_{c1}$) is fixed in the first image, then any point $\mathbf{x}_{c2}$ for which equation (3) holds ($\mathbf{x}_{c2}$ collects the first two coordinates of $\mathbf{p}_{c2}$, and the third coordinate is 1) must be on the epipolar line of $\mathbf{x}_{c1}$.

More formally, the product

$$\mathbf{l}_2 = E\mathbf{p}_{c1} \tag{4}$$

for fixed $\mathbf{p}_{c1}$ is a column vector, and equation (3) becomes

$$\mathbf{p}_{c2}^T \mathbf{l}_2 = 0$$

(hopefully you now see the reason for the apparently awkward choice of indices in the definition (4)). This is the equation of a line in the second image: if $\mathbf{l}_2 = [l_{21}, l_{22}, l_{23}]^T$, then this equation reads

$$l_{21}x_{c2} + l_{22}y_{c2} - l_{23} = 0 \ .$$

This is the equation of the epipolar line of $\mathbf{x}_{c1}$ in the second image.

Conversely, if $\mathbf{x}_{c2}$ is fixed in the second image, then the product

$$\mathbf{l}_1^T = \begin{bmatrix} l_{11} & l_{12} & l_{13} \end{bmatrix} = \mathbf{p}_{c2}^T E$$

is a row vector, and equation (3) becomes

$$\mathbf{l}_1^T \mathbf{p}_{c1} = 0$$

which is the equation of the epipolar line of $\mathbf{x}_{c2}$ in the first image:

$$l_{11}x_{c1} + l_{12}y_{c1} - l_{13} = 0 \ .$$

In order to *compute* the essential matrix E, it is useful to understand its algebraic properties. To this end, note first that the matrix $\mathbf{t}_\times$ has rank two if $\mathbf{t}$ is nonzero (and obviously rank zero if $\mathbf{t} = \mathbf{0}$). This is because

$$\mathbf{t}_\times \mathbf{t} = \mathbf{t} \times \mathbf{t} = \mathbf{0}$$

so the null space[4] of $\mathbf{t}_\times$ is nontrivial: The vector $\mathbf{t}$ is in the null space of $\mathbf{t}_\times$, so this matrix can only have at most two linearly independent rows. On the other hand, the second and third row of $\mathbf{t}_\times$ are obviously independent if $t_1 \neq 0$; the first and third row are independent if $t_2 \neq 0$; and the first and second row are independent if $t_3 \neq 0$. Thus, if $\mathbf{t}$ is nonzero, at least two rows of $\mathbf{t}_\times$ are independent. In summary, exactly two rows of $\mathbf{t}_\times$ are independent if $\mathbf{t} \neq \mathbf{0}$, so $\mathbf{t}_\times$ has rank two.

Since R is full rank (it has determinant 1), the product $E = R\mathbf{t}_\times$ has rank 2 if $\mathbf{t} \neq \mathbf{0}$. Also, since $\mathbf{t}_\times \mathbf{t} = \mathbf{0}$, we have

$$E\mathbf{t} = \mathbf{0} \ .$$

A similar equation ought to hold even if we switched the order in which the two cameras are listed. From equation (1) with $\mathbf{X}_1 = \mathbf{0}$, we see that the baseline vector as viewed from camera 2 is

$$\mathbf{t}' = R(\mathbf{0} - \mathbf{t}) = -R\mathbf{t}$$

and

$$(\mathbf{t}')^T E = -\mathbf{t}^T R^T E = -\mathbf{t}^T R^T R \mathbf{t}_\times = -\mathbf{t}^T \mathbf{t}_\times = (\mathbf{t}_\times \mathbf{t})^T = \mathbf{0}^T \ .$$

Here we used the fact that $\mathbf{t}_\times$ is skew symmetric:

$$\mathbf{t}_\times^T = -\mathbf{t}_\times \ .$$

Thus,

$$E\mathbf{t} = \mathbf{0} \quad \text{and} \quad E^T R\mathbf{t} = \mathbf{0} \ . \tag{5}$$

In other words, $\mathbf{t}$ spans the null space of E and $R\mathbf{t}$ spans its left null space.[5] The two image points corresponding to the baseline vectors $\mathbf{t}$ and $R\mathbf{t}$ are called the *epipoles*:

$$\mathbf{e}_1 = -\frac{\mathbf{t}}{t_3} \quad \text{and} \quad \mathbf{e}_2 = -\frac{R\mathbf{t}}{\mathbf{k}^T\mathbf{t}} \ .$$

Here, division by the third component makes that component equal to -1, so a point on the image plane is obtained. The following convention[6] is customary when the third component is zero:

$$\mathbf{e}_1 = \frac{1}{\|\mathbf{t}\|} \begin{bmatrix} t_1 \\ t_2 \\ 0 \end{bmatrix} \quad \text{if} \quad t_3 = 0 \quad \text{and} \quad \mathbf{e}_2 = \frac{1}{\|\mathbf{t}\|} \begin{bmatrix} \mathbf{i}^T\mathbf{t} \\ \mathbf{j}^T\mathbf{t} \\ 0 \end{bmatrix} \quad \text{if} \quad \mathbf{k}^T\mathbf{t} = 0 \ .$$

[4]The null space of a matrix is the space of vectors that are orthogonal to its rows.

[5]The left null space of a matrix is the space of vectors that are orthogonal to its columns.

[6]This convention would become natural if camera geometry were restated in homogeneous coordinates. The benefits of doing so, however, do not justify the additional machinery for our purposes.

Of course, equation (5) yields also

$$E\mathbf{e}_1 = \mathbf{0} \quad \text{and} \quad E^T\mathbf{e}_2 = \mathbf{0} \ .$$

Finally, for any vector $\mathbf{v}$ orthogonal to $\mathbf{t}$, the definition of cross product yields

$$\|\mathbf{t}_\times \mathbf{v}\| = \|\mathbf{t}\| \, \|\mathbf{v}\| \ .$$

Since rotation matrices do not change the norm of vectors, we also have

$$\|E\mathbf{v}\| = \|R\mathbf{t}_\times \mathbf{v}\| = \|\mathbf{t}_\times \mathbf{v}\| = \|\mathbf{t}\| \, \|\mathbf{v}\| \ .$$

Together with equations (5), this result shows that the essential matrix E has two nonzero singular values equal to each other (and to $\|\mathbf{t}\|$), and a zero singular value. The left and right singular vectors corresponding to the zero singular value are unit vectors along the epipoles,

$$\mathbf{v}_3 = \frac{\mathbf{e}_1}{\|\mathbf{e}_1\|} \quad \text{and} \quad \mathbf{u}_3 = \frac{\mathbf{e}_2}{\|\mathbf{e}_2\|} \ .$$

Since the other two singular values are equal to each other, the corresponding singular vectors are arbitrary, as long as they are orthogonal to each other and (i) $\mathbf{u}_1$, $\mathbf{u}_2$ are orthogonal to $\mathbf{u}_3$ and (ii) $\mathbf{v}_1$, $\mathbf{v}_2$ are orthogonal to $\mathbf{v}_3$.

The essential matrix is always used in homogeneous equations, so it can be normalized so that its two nonzero singular values are equal to one. This is equivalent to normalizing the length of the baseline translation vector $\mathbf{t}$. Similarly, the norm of the epipoles when the baseline is parallel to the image plane (so that the third component of the epipole is zero) is irrelevant.

The preceding discussion is summarized in Table 1.

The relationship between corresponding points expressed by the epipolar constraint of equation (3) requires camera calibration, because the coordinates $\mathbf{x}_c$ in the camera reference frame are assumed to be known. If lens distortion alone is known, the epipolar constraint can be expressed directly in image coordinates. In that case, the relationship between camera and image coordinates is affine:

$$\mathbf{x}_i = S\mathbf{x}_c + \mathbf{x}_0$$

(see the Section on image formation), and conversely

$$\mathbf{x}_c = S^{-1}(\mathbf{x}_i - \mathbf{x}_0) \ .$$

If we let

$$\mathbf{p}_i = \begin{bmatrix} \mathbf{x}_i \\ -1 \end{bmatrix}$$

we can write

$$\mathbf{p}_c = A\mathbf{p}_i \quad \text{where} \quad A = \begin{bmatrix} S^{-1} & S^{-1}\mathbf{x}_0 \\ \mathbf{0}^T & 1 \end{bmatrix}$$

and the epipolar constraint becomes

$$\mathbf{p}_{i2}^T A^T E A \mathbf{p}_{i1} = 0$$

For a stereo camera pair with nonzero baseline, let

$$\mathbf{X}_2 = R(\mathbf{X}_1 - \mathbf{t})$$

be the coordinate transformation from camera 1 to camera 2, and define

$$\mathbf{b}_1 = \begin{bmatrix} b_{11} & b_{12} & b_{13} \end{bmatrix}^T = \mathbf{t}/\|\mathbf{t}\| \quad \text{and} \quad \mathbf{b}_2 = \begin{bmatrix} b_{21} & b_{22} & b_{23} \end{bmatrix}^T = -R\mathbf{t}/\|\mathbf{t}\|$$

to be respectively (i) the unit vector pointing from the center of projection of camera 1 to that of camera 2, expressed in camera reference system 1, and (ii) the unit vector pointing from the center of projection of camera 2 to that of camera 1, expressed in camera reference system 2. Call these two vectors the first and second *unit epipoles* of the stereo pair, and define the *epipoles* as follows:

$$\mathbf{e}_1 = \begin{cases} -\mathbf{b}_1/b_{13} & \text{if } b_{13} \neq 0 \\ \begin{bmatrix} b_{11} & b_{12} & 0 \end{bmatrix}^T & \text{if } b_{13} = 0 \end{cases} \quad \text{and} \quad \mathbf{e}_2 = \begin{cases} -\mathbf{b}_2/b_{23} & \text{if } b_{23} \neq 0 \\ \begin{bmatrix} b_{21} & b_{22} & 0 \end{bmatrix}^T & \text{if } b_{23} = 0 \end{cases}$$

The *essential matrix* of the stereo pair is the matrix

$$E = R\mathbf{b}_{1\times} = R \begin{bmatrix} 0 & -b_{13} & b_{12} \\ b_{13} & 0 & -b_{11} \\ -b_{12} & b_{11} & 0 \end{bmatrix} .$$

If point $\mathbf{x}_{c1} = (x_{c1}, y_{c1})$ in image 1 corresponds to point $\mathbf{x}_{c2} = (x_{c2}, y_{c2})$ in image 2, then $\mathbf{x}_{c1}$ is on the *epipolar line* of $\mathbf{x}_{c2}$, which has equation

$$l_{11}x_{c1} + l_{12}y_{c1} - l_{13} = 0 \quad \text{where} \quad \mathbf{l}_1^T = \begin{bmatrix} l_{11} & l_{12} & l_{13} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{c2}^T & -1 \end{bmatrix} E .$$

Conversely, $\mathbf{x}_{c2}$ is on the epipolar line of $\mathbf{x}_{c1}$, which has equation

$$l_{21}x_{c2} + l_{22}y_{c2} - l_{23} = 0 \quad \text{where} \quad \mathbf{l}_2 = \begin{bmatrix} l_{21} & l_{22} & l_{23} \end{bmatrix}^T = E \begin{bmatrix} \mathbf{x}_{c1} \\ -1 \end{bmatrix} .$$

The singular value decomposition of E is

$$E = U\Sigma V^T = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 \end{bmatrix} \operatorname{diag}(1,1,0) \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix}^T \tag{6}$$

where

$$\mathbf{u}_3 = \mathbf{b}_2 \quad \text{and} \quad \mathbf{v}_3 = \mathbf{b}_1$$

and $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{v}_1$, $\mathbf{v}_2$ are *any* vectors for which U and V become orthogonal.

Table 1: Definition and properties of the essential matrix.

that is,

$$\mathbf{p}_{i2}^T F \mathbf{p}_{i1} = 0 \quad \text{where} \quad F = A^T E A \ . \tag{7}$$

The matrix F is called the *fundamental matrix* in the literature.

The fundamental matrix allows finding the epipolar lines in image coordinates, rather than in camera coordinates: Given point $\mathbf{p}_{i1}$ in image 1, the corresponding epipolar line in image 2 is

$$\mathbf{p}_{i2}^T \mathbf{m}_2 = 0 \quad \text{where} \quad \mathbf{m}_2 = F \mathbf{p}_{i1}$$

and conversely, given point $\mathbf{p}_{i2}$ in image 2, the corresponding epipolar line in image 1 is

$$\mathbf{m}_1^T \mathbf{p}_{i1} = 0 \quad \text{where} \quad \mathbf{m}_1^T = \mathbf{p}_{i2}^T F \ .$$

The algebraic properties of F are similar to those of E: the fundamental matrix F has rank 2, has a singular value of zero, and the corresponding left and right singular vectors are the epipoles in images 2 and 1 if they are normalized so that the third coordinate is one (with the usual convention when the third coordinate is zero). However, the two nonzero singular values are no longer equal to each other.

The usefulness of the fundamental matrix without the essential matrix is somewhat limited by the need to know the lens distortion parameters: if these are known, then the other intrinsic camera parameters are likely to be known as well. Nonetheless, the fundamental matrix yields reasonably good results with little effort when the lens distortion is small. In addition, the fundamental matrix is useful in stereo matching (see Section 5.3 below), because it allows computing epipolar lines without explicit transformation to camera coordinates. Of course, once the essential matrix E is known, the fundamental matrix F can be computed through the second of equations (7).

One way to compute the essential or the fundamental matrix is through their definitions in (3) or (7). However, either matrix can also be computed without knowing the extrinsic parameters R and $\mathbf{t}$ for the stereo pair and, for the fundamental matrix, also without knowing the intrinsic parameters for the two cameras, except for distortion.

This is now shown for the essential matrix. The epipolar constraint is linear (and homogeneous) in the entries of the essential matrix E. The equation (3) can be spelled out as follows:

$$\mathbf{q}^T \mathbf{e} = 0$$

where

$$\begin{aligned} \mathbf{q} &= \mathbf{p}_{c1} \otimes \mathbf{p}_{c2} = [x_{c1}\mathbf{p}_{c2},\ y_{c1}\mathbf{p}_{c2},\ -\mathbf{p}_{c2}] \\ &= [x_{c1}x_{c2},\ x_{c1}y_{c2},\ -x_{c1},\ y_{c1}x_{c2},\ y_{c1}y_{c2},\ -y_{c1},\ x_{c2},\ y_{c2},\ -1]^T \end{aligned}$$

is the *Kronecker product* of the vectors $\mathbf{p}_{c1}$ and $\mathbf{p}_{c2}$ and

$$\mathbf{e} = [e_{11},\ e_{21},\ e_{31},\ e_{12},\ e_{22},\ e_{32},\ e_{13},\ e_{23},\ e_{33}] = E(:)$$

is the list of the entries in E in column-major order (the last expression is in Matlab-style notation).

Given $N \geq 8$ (usually many more) pairs of corresponding points, perhaps determined by visual inspection, one can then write N such equations in the form of an N by 9 homogeneous matrix equation

$$Q\mathbf{e} = \mathbf{0} \ . \tag{8}$$

This can be solved for $\mathbf{e}$ (see Math supplement) to yield the nine entries of a matrix $\tilde{E}$. Because of noise in the entries of Q, this matrix does not generally satisfy the properties of an essential matrix. These can be enforced by letting

$$E = U\text{diag}(1, 1, 0)V^T \quad \text{where} \quad \tilde{E} = U\Sigma V^T$$

is the singular value decomposition of $\tilde{E}$. The matrix E is the essential matrix that is closest to $\tilde{E}$ in the least-squares sense (more specifically, in matrix 2-norm).

The procedure for the fundamental matrix is similar, except that once $\tilde{F}$ has been found by solving a system analogous to (8) the matrix F is found as

$$F = U\text{diag}(\sigma_1, \sigma_2, 0)V^T \quad \text{where} \quad \tilde{F} = U\text{diag}(\sigma_1, \sigma_2, \sigma_3)V^T$$

is the SVD of $\tilde{F}$.

Since each of these procedures finds the result in two separate optimization steps (for instance, first solve system (8) to find $\tilde{E}$, then find the closest E to $\tilde{E}$), the results are not optimal, and are biased. Various authors[7, 4, 2] have shown ways to improve the results. The method in [4] preconditions the problem by scaling and translating the image coordinates (separately foe each image) so that their centroids are at the origin and their RMS spread is one. This makes the condition number (see Math supplement) of the problem (8) or its equivalent for the fundamental matrix much smaller than without this scaling and centering, and improves results considerably. Because the transformation is affine, it can be undone once either $\tilde{E}$ or $\tilde{F}$ is computed. In other words, if

$$\mathbf{p}_1 = T_1\mathbf{p}_{c1} \quad \text{and} \quad \mathbf{p}_2 = T_2\mathbf{p}_{c2}$$

then $\tilde{E}$ is replaced by $T_2^T\tilde{E}T_1$, and similarly for $\tilde{F}$.

The methods in [7, 2] are more complex, and go ones step further by refining the solution with a numerical procedure that minimizes the geometrical re-projection error, rather than the algebraic error implied by equations (3) and (7). This is analogous to what we did for camera calibration, and yields a consistent (*i.e.*, asymptotically unbiased) estimate.

5.3 Rectification

Given a stereo camera pair, stereo matching, that is, the search for pairs of corresponding points, could loop over the pixels of, say, camera 1. For each pixel $\mathbf{x}_{i1}$, the fundamental matrix would yield the corresponding epipolar line $\mathbf{m}_1$, and a search for a point $\mathbf{x}_{i2}$ corresponding to $\mathbf{x}_{i1}$ could be performed over this line according to one of the methods discussed in Section 5.5 below. This would imply a second loop, nested within the first, over the epipolar line.

However, these loops are inefficient. First, the epipolar line $\mathbf{m}_2$ must be computed. Second, looping over a general line in the image requires sub-pixel interpolation. A more efficient approach is to *rectify* the two images first, that is, transform them to two new images whose epipolar lines coincide with the image rows, and so that corresponding epipolar lines are rows with equal subscripts.

A moment's reflection shows that rectified images would be produced directly by a pair of identical cameras placed side to side, with their optical axes parallel to each other and at the same height, and the image sensors coplanar. This is called the *standard stereo configuration*. It is then tempting to place the cameras physically in this configuration to make rectification unnecessary. However, this setup is very difficult to achieve in practice, because it would require a positioning accuracy of the order of a micrometer and a small fraction of a degree. A more common and more practical approach is to place the cameras only approximately in the standard configuration, to achieve maximum overlap of the fields of view, and then rectify the images. The rectification procedure is described next.

Rectification can be thought of as the process of defining two new reference systems for camera 1 and camera 2 that correspond to the standard stereo configuration, and then transforming the camera-coordinate image points into the two new systems. The only transformations allowed are rotations. This is because translations change the viewpoint, and it is not possible to transform an image from one viewpoint into an image from a different viewpoint without knowing the geometry of the world. In addition, linear transformations other than rotations would deform angles between projection rays, and therefore destroy the integrity of the information contained in the image. Thus, we want to find a rotation R_1 of the reference system of camera 1 and a rotation R_2 of the reference system of camera 2 such that the epipolar constraint (see equation (3)) between the rotated image points in camera coordinates,

$$(R_2\mathbf{p}_{c2})^T E(R\mathbf{p}_{c1}) = 0 \ ,$$

transforms image rows into image rows. In symbols, if we let

$$E' = R_2^T E R_1$$

be the epipolar matrix between the transformed images, we want

$$[x_{c2},\, y_{c2}\ -1]E' \begin{bmatrix} x_{c1} \\ y_{c1} \\ -1 \end{bmatrix} = 0$$

to boil down to the equation

$$y_{c2} = y_{c1} \ .$$

This is the case if

$$E' = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} \ .$$

In summary, we want to find rotation matrices R_1 and R_2 such that

$$R_2^T E R_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} .$$

From equation (6),

$$R_2^T E R_1 = R_2^T U \text{diag}(1,1,0) V^T R_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} .$$

The first step is to flip the sign of the third column of U if the determinant of U is positive, and to flip the sign of the third column of V if the determinant of V is positive. These flips are harmless, because the columns in question are multiplied only by zeros in the matrix diag$(1,1,0)$. These sign flips will make sure that the solutions R_1 and R_2 have determinant 1, rather than -1. The equation above can be solved by letting, say,

$$V^T R_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad R_2^T U = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} .$$

These are permutation and inversion matrices, and are therefore orthogonal, as can be verified directly by multiplying them by their transposes:

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = I \quad \text{and} \quad \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix} = I .$$

Their effect is to switch the first and third column (multiplication from the right) of diag$(1,1,0)$, and to cyclically permute its three rows, while changing the sign of the third resulting row. The determinant of both matrices is -1. Check that this does indeed transform diag$(1,1,0)$ into E' as desired. So

$$R_1 = V \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad R_2 = U \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix} .$$

The matrix R_1 is a rotation matrix, because it is the product of two orthogonal matrices with negative determinant. The same conclusion holds for R_2, the product of two orthogonal matrices with positive determinant.

The transformations

$$\mathbf{p}'_{c1} = R_1 \mathbf{p}_{c1} \quad \text{and} \quad \mathbf{p}'_{c2} = R_2 \mathbf{p}_{c2}$$

rotate the image vectors. The results must be normalized by dividing each vector by the negative of its third component, to project it back onto the image plane.

In practice, and just as we did for the removal of lens distortion, it is best to loop over the target coordinates $\mathbf{p}'_{c1}$ and $\mathbf{p}'_{c2}$ and retrieve the corresponding values at points

$$\mathbf{p}_{c1} = R_1^T \mathbf{p}'_{c1} \quad \text{and} \quad \mathbf{p}_{c2} = R_2^T \mathbf{p}'_{c2}$$

in the source image by bilinear interpolation, with

$$\mathbf{x}_{c1} = -\frac{\mathbf{p}_{c1}(1:2)}{\mathbf{p}_{c1}(3)} \quad \text{and} \quad \mathbf{x}_{c2} = -\frac{\mathbf{p}_{c2}(1:2)}{\mathbf{p}_{c2}(3)} \ .$$

This computation is summarized in Table 2.

To Be Continued

5.4 Triangulation

5.5 Correspondence

References

[1] Nicholas Ayache and Francis Lustman. Fast and reliable passive trinocular stereovision. In *Proceedings of the First International Conference on Computer Vision*, pages 422–427, London, England, June 1987.

[2] O. Faugeras and Q. T. Luong. *The geometry of multiple images*. MIT Press, Cambridge, MA, 2001.

[3] A. Gerhard, H. Platzer, J. Steurer, and R. Lenz. Depth extraction by stereo triples and a fast correspondence estimation algorithm. In *Proc. International Conference on Pattern Recognition*, pages 512–515, 1986.

[4] R. I. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–93, June 1997.

[5] Larry Matthies, Takeo Kanade, and Richard Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3(3):209–236, September 1989.

[6] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4):353–363, April 1993.

[7] G. Xu and Z. Zhang. *Epipolar Geometry in Stereo, Motion and Object Recognition – A Unified Approach*, volume 6 of *Computational Imaging and Vision*. Springer-Verlag, New York, NY, 1996.

Let

$$E = \tilde{U}\text{diag}(1,1,0)\tilde{V}^T$$

be the singular value decomposition of the essential matrix E of a calibrated stereo pair. Let $I_1(\mathbf{x})$, $I_2(\mathbf{x})$ be the images obtained by removing lens distortion from a pair of corresponding images from the stereo pair (see the Section on camera calibration for lens distortion removal). To produce rectified images $I_1'(\mathbf{x}')$ and $I_2'(\mathbf{x}')$, proceed as follows.

- If $\det(\tilde{U}) = \det([\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]) = 1$, then let $U = [\mathbf{u}_1, \mathbf{u}_2, -\mathbf{u}_3]$. Otherwise, let $U = \tilde{U}$.
- If $\det(\tilde{V}) = \det([\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]) = 1$, then let $V = [\mathbf{v}_1, \mathbf{v}_2, -\mathbf{v}_3]$. Otherwise, let $V = \tilde{V}$.
- Define

$$R_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} V \quad \text{and} \quad R_2 = U \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix} .$$

- For images $i = 1, 2$, and for every pixel at position $\mathbf{x}'$ in target image I_i', do the following
  - Let S_{ui} and $\mathbf{x}_{u0i}$ be the intrinsic calibration parameters of camera i (after lens distortion removal). Determine the location $\mathbf{x}$ in the source image I_i corresponding to $\mathbf{x}'$ as follows:

$$\mathbf{x}_c' = S_{ui}^{-1}(\mathbf{x}' - \mathbf{x}_{ui0}) \ , \quad \mathbf{p}_c = R_i^T \begin{bmatrix} \mathbf{x}_c' \\ -1 \end{bmatrix} \ , \quad \mathbf{x}_c = -\frac{\mathbf{p}_c(1:2)}{\mathbf{p}_c(3)} \ , \quad \mathbf{x} = S_{ui}\mathbf{x}_c + \mathbf{x}_{ui0} \ .$$

  - Determine pixel value $I_i(\mathbf{x})$ by bilinear interpolation (see Section on image processing).
  - Copy the resulting pixel value to the target image:

$$I_i'(\mathbf{x}') = I_i(\mathbf{x}) \ .$$

Table 2: Rectification of a stereo pair.