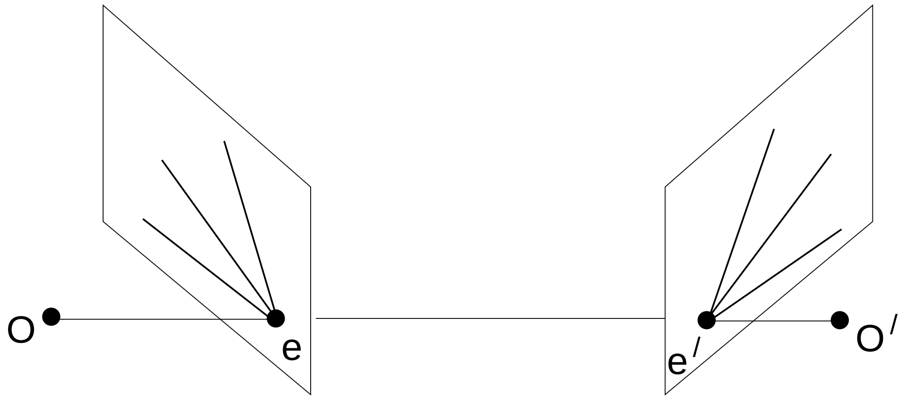
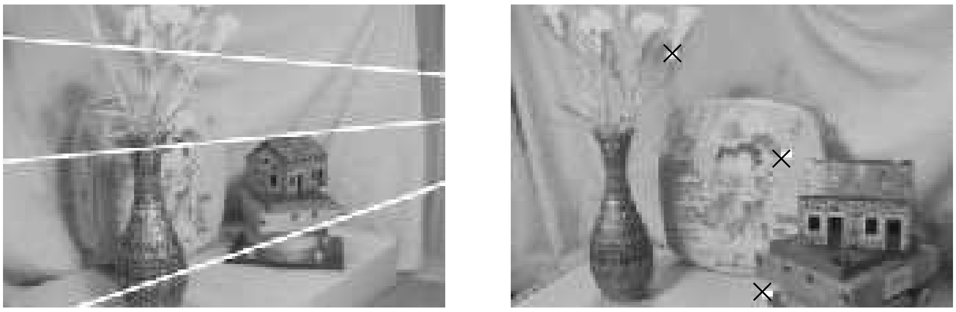
University of Cambridge
Engineering Part IIB

Module 4F12: Computer Vision

Handout 4: Stereo and Structure from Motion

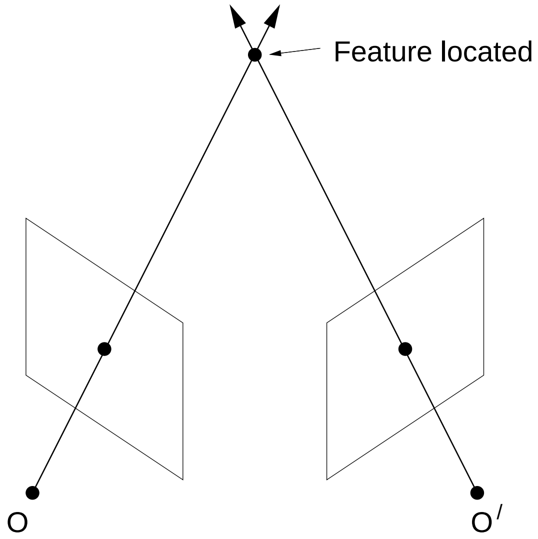




Roberto Cipolla
November 2023

Stereo vision

We have seen how it is impossible in general to recover 3D scene structure from a single image. Even if the camera is calibrated, we can only deduce the *ray* on which each image feature lies.



If we can observe the same feature from two different viewpoints, however, we can solve for the intersection of the rays and recover the 3D location of the feature. This is the essence of triangulation in **stereo vision**.

While this might sound straightforward, there are many subtleties to stereo vision. For instance, to what extent do we need to calibrate the cameras? How do we establish correspondences between features in the two views?

Recovering 3D structure

If the left and right cameras are calibrated with respect to the world coordinate system, then it is not difficult to recover 3D structure.

Recall from handout 3 that each point observed by one camera gives us two equations in three unknowns (X, Y, Z):

$$\begin{aligned} u &= \frac{su}{s} = \frac{p_{11}X + p_{12}Y + p_{13}Z + p_{14}}{p_{31}X + p_{32}Y + p_{33}Z + p_{34}} \\ v &= \frac{sv}{s} = \frac{p_{21}X + p_{22}Y + p_{23}Z + p_{24}}{p_{31}X + p_{32}Y + p_{33}Z + p_{34}} \end{aligned}$$

Observing the same point with the other camera provides two further equations. These can be re-arranged as four linear equations in the three unknowns (X, Y, Z), and geometrically correspond to 4 planes defining 2 rays. The four equations in three unknowns are over-constrained and a solution can be found by least-squares.

To understand what is required for the equations to be consistent, we will first reformulate the equations in terms of 3D vectors. The analysis will also identify a key constraint to help with the correspondence problem.

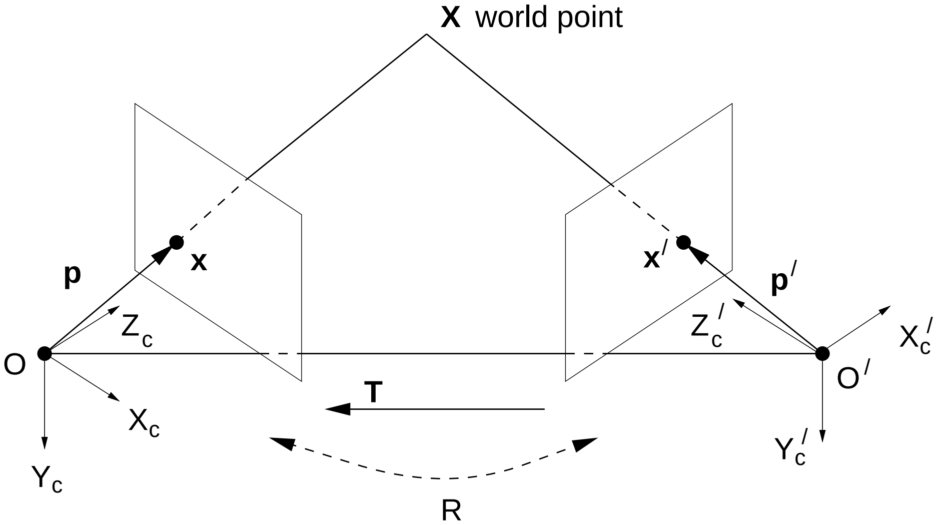
Rays

Suppose we know the *relative* positions of the cameras and their *intrinsic* parameters[1]. Given the CCD parameters, we can translate pixel coordinates (u, v) into image plane coordinates (x, y):

$$u = u_0 + k_u x \ , \quad v = v_0 + k_v y$$

With the focal length, we can translate image plane coordinates into a ray in 3D space. Let's define the ray by the point $\mathbf{p}$ (in camera-centered coordinates) where it pierces the image plane:

$$\mathbf{p} = \begin{bmatrix} x \\ y \\ f \end{bmatrix}$$



[1] We can extract most of this information from the two calibration matrices.

From pixels to rays

We can conveniently express the relationship between an image point with pixel coordinates $\mathbf{w}$ and a ray in 3D $\mathbf{p}$. These are related by the CCD calibration matrix:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} k_u & 0 & u_0 \\ 0 & k_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

We can modify this to derive a relationship between pixel coordinates and rays:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} k_u & 0 & u_0/f \\ 0 & k_v & v_0/f \\ 0 & 0 & 1/f \end{bmatrix} \begin{bmatrix} x \\ y \\ f \end{bmatrix}$$

Recall that we defined the camera calibration matrix K as follows:

$$\mathrm{K} = \begin{bmatrix} fk_u & 0 & u_0 \\ 0 & fk_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

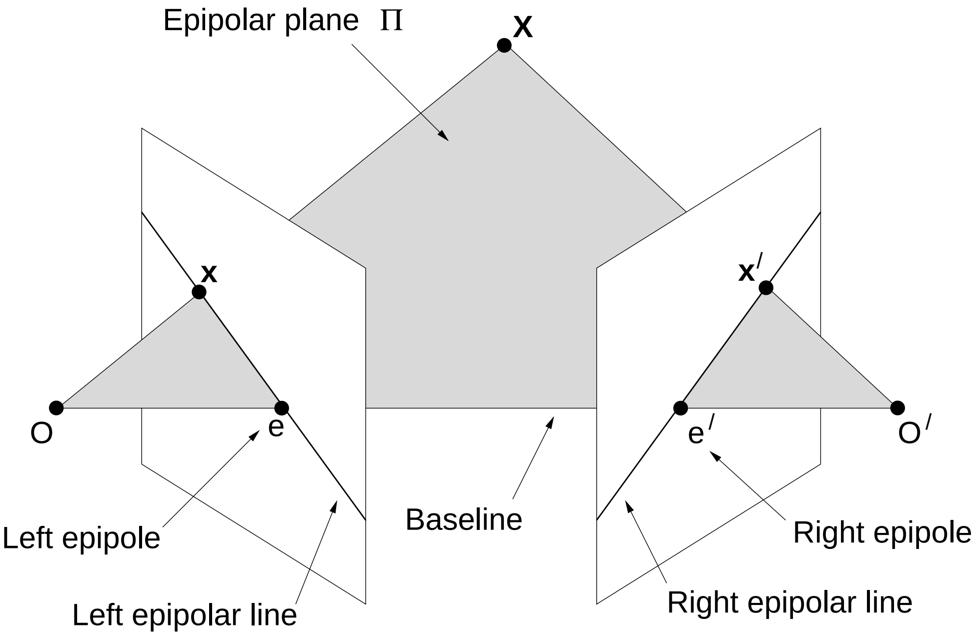
then we can write

$$\tilde{\mathbf{w}} = \mathrm{K}\mathbf{p}$$

Epipolar geometry

An important part of stereo is triangulating 2 rays from a pair of image correspondences. The most important matching constraint which can is used is the **epipolar constraint**, and follows directly from the fact that the rays must intersect in 3D space.

Epipolar constraints facilitate the search for correspondences: they constrain the search to a line in each image. To derive general epipolar constraints, consider the **epipolar geometry** of two cameras:



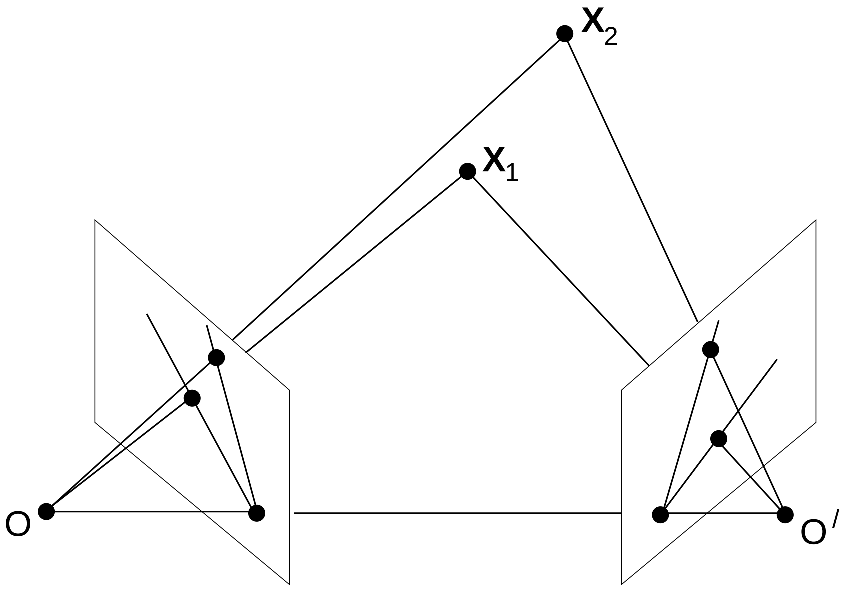
The **epipolar plane** is the plane defined by a 3D point **X** and the optical centres.

Epipolar geometry

The **baseline** is the line joining the optical centres.

An **epipole** is the point of intersection of the baseline with the image plane. There are two epipoles, one for each image.

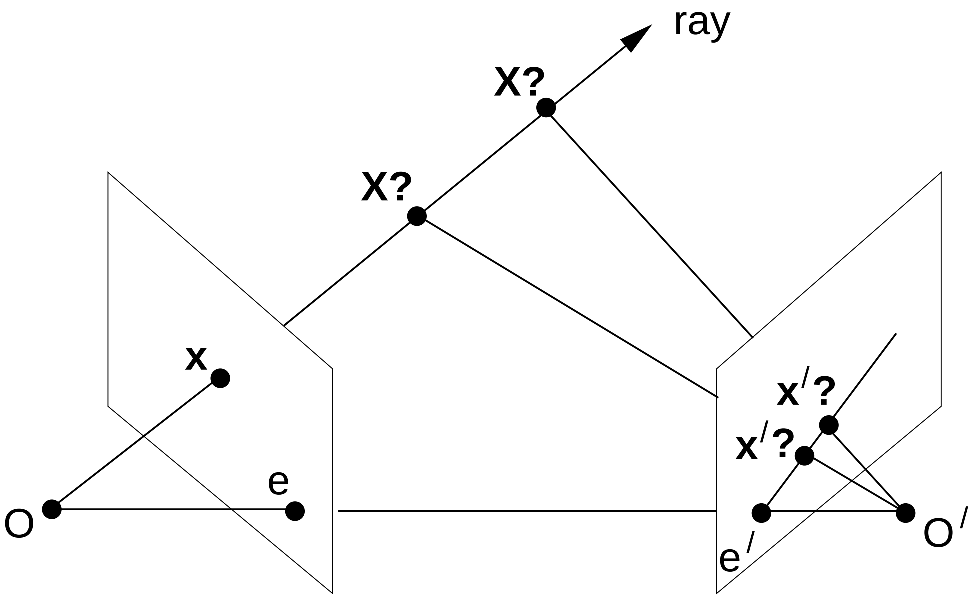
An **epipolar line** is a line of intersection of the epipolar plane with an image plane. It is the image in one camera of the ray from the other camera's optical centre to the point **X**.



For different world points **X**, the epipolar plane rotates about the baseline. All epipolar lines intersect at the epipole.

Epipolar geometry

The epipolar line constrains the search for correspondence from a region to a line. If a point feature $\mathbf{x}$ is observed in one image, then its location $\mathbf{x}'$ in the other image must lie on the epipolar line.



We can derive an expression for the epipolar line. The two camera-centered coordinate systems are related by a rotation R and translation $\mathbf{T}$:

$$\mathbf{X}'_\mathrm{c} = \mathrm{R}\mathbf{X}_\mathrm{c} + \mathbf{T}$$

Taking the vector product with $\mathbf{T}$, we obtain

$$\begin{aligned} \mathbf{T} \times \mathbf{X}'_\mathrm{c} &= \mathbf{T} \times \mathrm{R}\mathbf{X}_\mathrm{c} + \mathbf{T} \times \mathbf{T} \\ \Leftrightarrow \mathbf{T} \times \mathbf{X}'_\mathrm{c} &= \mathbf{T} \times \mathrm{R}\mathbf{X}_\mathrm{c} \end{aligned}$$

The essential matrix

Taking the scalar product with $\mathbf{X}'_{\rm c}$, we obtain

$$\begin{aligned} \mathbf{X}'_{\rm c}.(\mathbf{T} \times \mathbf{X}'_{\rm c}) &= \mathbf{X}'_{\rm c}.(\mathbf{T} \times {\rm R}\mathbf{X}_{\rm c}) \\ \Leftrightarrow \mathbf{X}'_{\rm c}.(\mathbf{T} \times {\rm R}\mathbf{X}_{\rm c}) &= 0 \qquad (1) \end{aligned}$$

Recall that a vector product can be expressed as a matrix multiplication:

$$\begin{aligned} \mathbf{T} \times \mathbf{X}_{\rm c} &= {\rm T}_\times \mathbf{X}_{\rm c} \\ \text{where } {\rm T}_\times &= \begin{bmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{bmatrix} \end{aligned}$$

So equation (1) can be rewritten as

$$\begin{aligned} \mathbf{X}'_{\rm c}.({\rm T}_\times {\rm R}\mathbf{X}_{\rm c}) &= 0 \\ \Leftrightarrow \mathbf{X}'^{T}_{\rm c} {\rm E}\mathbf{X}_{\rm c} &= 0\ , \quad \text{where } {\rm E} = {\rm T}_\times {\rm R} \end{aligned}$$

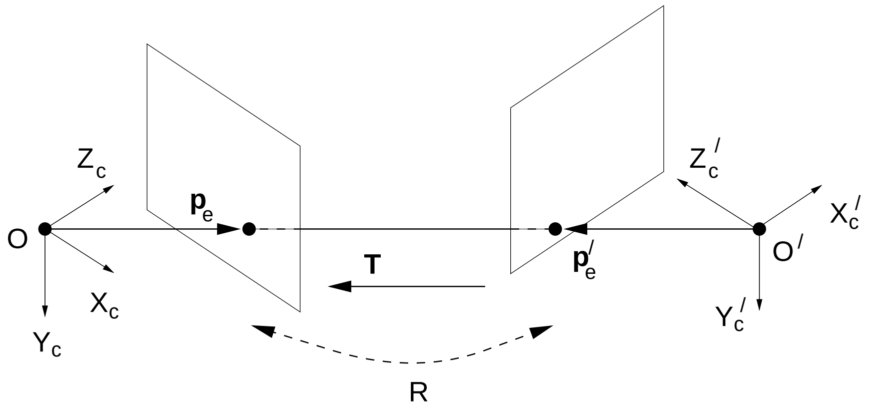
E is a 3 × 3 matrix known as the **essential matrix**. The constraint also holds for rays $\mathbf{p}$, which are parallel to the camera-centered position vectors $\mathbf{X}_{\rm c}$:

$$\mathbf{p}'^{T}{\rm E}\mathbf{p} = 0 \qquad (2)$$

This is the epipolar constraint. If we observe a point $\mathbf{p}$ in one image, then its position $\mathbf{p}'$ in the other image must lie on the line defined by (2).

The essential matrix

The essential matrix can also be used to find the locations of the epipoles.



Referring to the figure, the position of the left camera's epipole is $\mathbf{p}_\mathrm{e}$ in the left camera's coordinate system and $\lambda\mathbf{T}$ in the right camera's coordinate system. Relating the coordinate systems, we obtain

$$\lambda\mathbf{T} = \mathrm{R}\mathbf{p}_\mathrm{e} + \mathbf{T}$$

Taking the vector product with $\mathbf{T}$, we obtain

$$\begin{aligned} \mathbf{0} &= \mathbf{T} \times \mathrm{R}\mathbf{p}_\mathrm{e} \\ \Leftrightarrow \mathrm{E}\mathbf{p}_\mathrm{e} &= \mathbf{0} \end{aligned}$$

So the location of the epipole in the left image lies in the null space of E. It follows that E is non-invertible (det E = 0) and is therefore of maximum rank 2. The result for the other epipole is $\mathrm{E}^T\mathbf{p}_\mathrm{e}' = \mathbf{0}$.

Essential matrix: example

Let's calculate the essential matrix for the basic parallel camera configuration:

$$\mathrm{R} = \mathrm{I}\ , \quad \mathbf{T} = \begin{bmatrix} -d \\ 0 \\ 0 \end{bmatrix}\ , \quad \mathrm{E} = \mathrm{T}_\times \mathrm{R} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & d \\ 0 & -d & 0 \end{bmatrix}$$

The epipolar constraint $\mathbf{p}'^T \mathrm{E}\mathbf{p} = 0$ is therefore

✎

$$\begin{aligned} \begin{bmatrix} x' & y' & f \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & d \\ 0 & -d & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ f \end{bmatrix} &= 0 \\ \Leftrightarrow \begin{bmatrix} x' & y' & f \end{bmatrix} \begin{bmatrix} 0 \\ df \\ -dy \end{bmatrix} &= 0 \\ \Leftrightarrow y &= y' \end{aligned}$$

Hence the image of any point $\mathbf{X}$ must lie on the same horizontal line in each image plane.

For parallel cameras, the epipolar lines are parallel, and the epipole is at infinity. This is what we'd expect: neither camera can "see" the optical centre of the other camera.

From rays to pixels

Up until now we have been assuming calibrated cameras, so we can go from pixel coordinates **w** to rays **p**. But what if we do not know the calibration?

We have seen how pixel coordinates and image plane coordinates are related by the CCD calibration matrix:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} k_u & 0 & u_0 \\ 0 & k_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

We can modify this to derive a relationship between pixel coordinates and rays:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} k_u & 0 & u_0/f \\ 0 & k_v & v_0/f \\ 0 & 0 & 1/f \end{bmatrix} \begin{bmatrix} x \\ y \\ f \end{bmatrix}$$

If we define the matrix K as follows:

$$\mathrm{K} = \begin{bmatrix} fk_u & 0 & u_0 \\ 0 & fk_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

then we can write

$$\tilde{\mathbf{w}} = \mathrm{K}\mathbf{p}$$

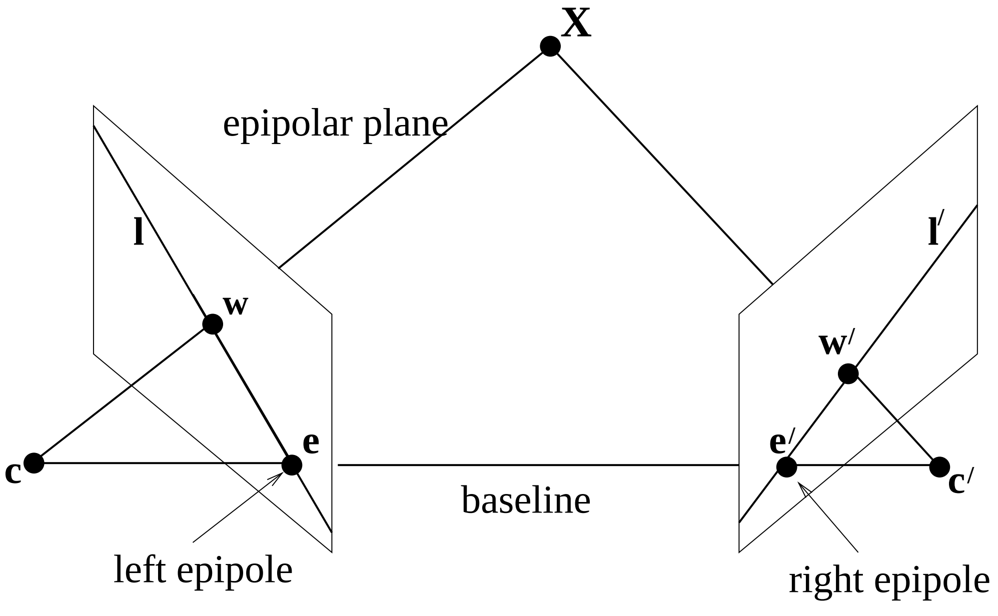
The fundamental matrix

The epipolar constraint becomes

$$\begin{aligned} \mathbf{p}'^T \mathrm{E}\mathbf{p} &= 0 \\ \Leftrightarrow \tilde{\mathbf{w}}'^T \mathrm{K}'^{-T} \mathrm{E} \mathrm{K}^{-1} \tilde{\mathbf{w}} &= 0 \\ \Leftrightarrow \tilde{\mathbf{w}}'^T \mathrm{F} \tilde{\mathbf{w}} &= 0 \ , \quad \text{where } \mathrm{F} = \mathrm{K}'^{-T} \mathrm{E} \mathrm{K}^{-1} \end{aligned}$$

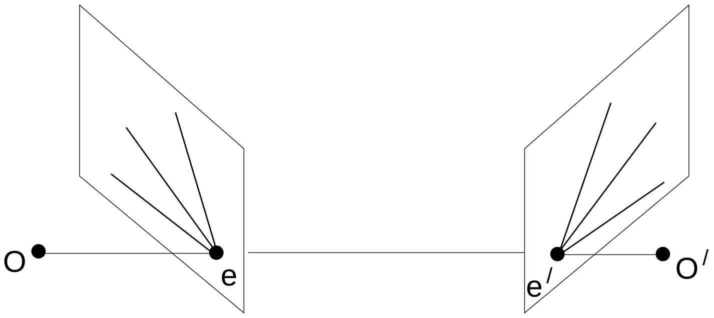
F is the 3×3 **fundamental matrix** and the epipolar constraint can be expressed in terms of image/pixel co-ordinates:

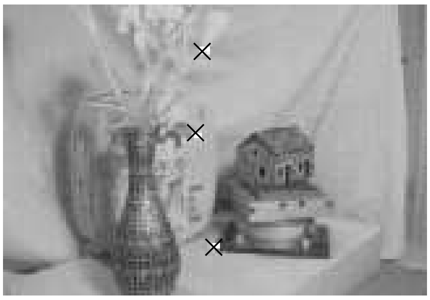
$$\begin{bmatrix} u' & v' & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = 0$$



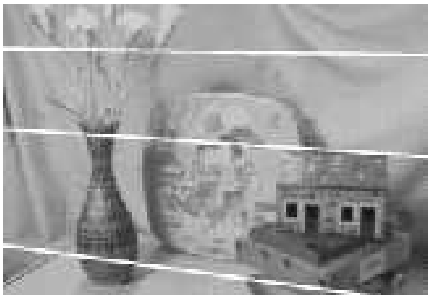
Epipolar geometry examples

Converging cameras

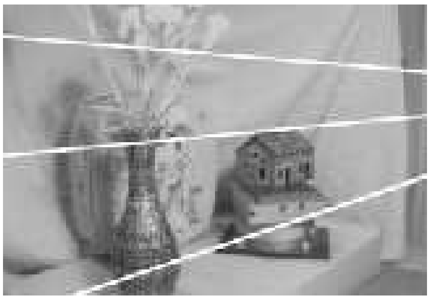




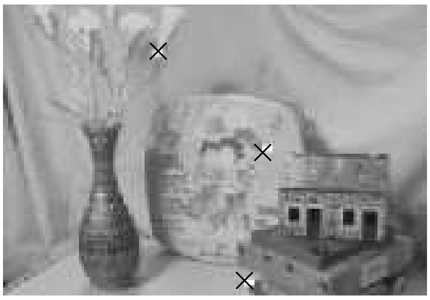
3 corner features
in left image



Epipolar lines
in right image



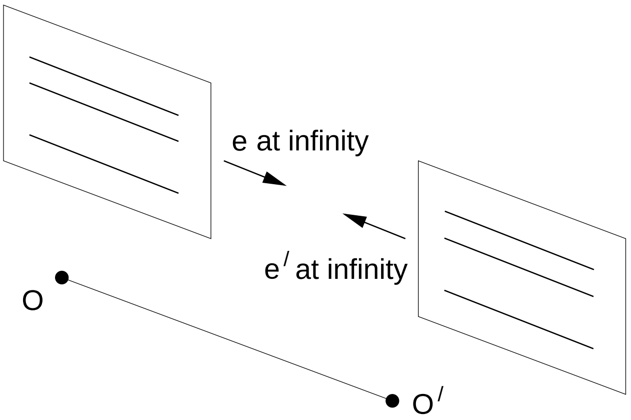
Epipolar lines
in left image

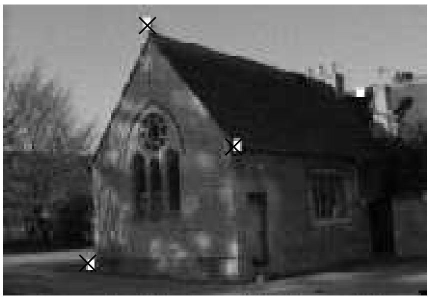


3 corner features
in right image

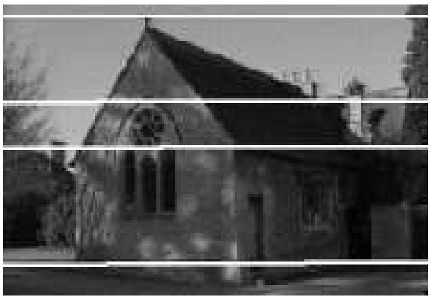
Epipolar geometry examples

Near parallel cameras

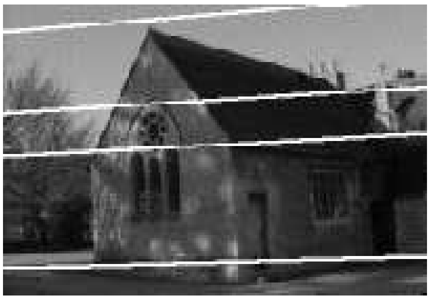




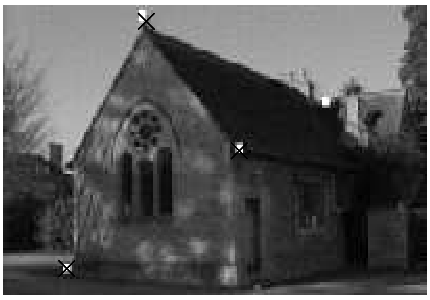
3 corner features
in left image



Epipolar lines
in right image



Epipolar lines
in left image



3 corner features
in right image

Epipolar lines and epipoles

For any given point $\tilde{\mathbf{w}}$ in the left image, using the known F we can derive an epipolar constraint on the corresponding location in the right image, $\tilde{\mathbf{w}}'$.

The **epipolar line** in the right view can be expressed simply in in homogeneous co-ordinates, $\tilde{\mathbf{l}}'$:

$$\begin{aligned} \tilde{\mathbf{w}}'^T \mathrm{F} \tilde{\mathbf{w}} &= 0 \\ \tilde{\mathbf{w}}'^T \tilde{\mathbf{l}}' &= 0 \\ \Leftrightarrow \tilde{\mathbf{l}}' &= \mathrm{F} \tilde{\mathbf{w}} \end{aligned}$$

Similarly for a point in the right image, $\tilde{\mathbf{w}}'$, the epipolar line in the left image is given by $\tilde{\mathbf{l}}$:

$$\tilde{\mathbf{l}} = \mathrm{F}^{\mathrm{T}} \tilde{\mathbf{w}}'$$

The locations of the epipoles $\tilde{\mathbf{w}}_{\mathrm{e}}$ and $\tilde{\mathbf{w}}'_{\mathrm{e}}$ (in pixels) are given by

$$\begin{aligned} \mathrm{E}\mathbf{p}_{\mathrm{e}} &= \mathbf{0} \Leftrightarrow \mathrm{E}\mathrm{K}^{-1}\tilde{\mathbf{w}}_{\mathrm{e}} = \mathbf{0} \\ \Leftrightarrow \mathrm{K}'^{-T}\mathrm{E}\mathrm{K}^{-1}\tilde{\mathbf{w}}_{\mathrm{e}} &= \mathbf{0} \\ \Leftrightarrow \mathrm{F}\tilde{\mathbf{w}}_{\mathrm{e}} &= \mathbf{0} \text{ and likewise } \quad \mathrm{F}^T \tilde{\mathbf{w}}'_{\mathrm{e}} = \mathbf{0} \end{aligned}$$

The fundamental matrix

At first sight, F appears to have 9 degrees of freedom. However, its overall scale does not matter (so we could set f_{33} to 1) and, as with E, it has zero determinant (maximum rank 2). So F has only 7 degrees of freedom.

Computing F from correspondences

Since the cameras are uncalibrated, we do not know E, K or K′ and so we do not know F a-priori. However, we can estimate F from point correspondences.

Each point correspondence $\tilde{\mathbf{w}} \leftrightarrow \tilde{\mathbf{w}}'$ generates one constraint on F:

$$\begin{bmatrix} u' & v' & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = 0$$

n of these constraints can be arranged in the following form:

$$\begin{bmatrix} u'_1u_1 & u'_1v_1 & u'_1 & v'_1u_1 & v'_1v_1 & v'_1 & u_1 & v_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u'_nu_n & u'_nv_n & u'_1 & v'_nu_n & v'_nv_n & v'_n & u_n & v_n & 1 \end{bmatrix} \begin{bmatrix} f_{11} \\ f_{12} \\ f_{13} \\ f_{21} \\ f_{22} \\ f_{23} \\ f_{31} \\ f_{32} \\ f_{33} \end{bmatrix} = \mathbf{0}$$

Computing F from correspondences

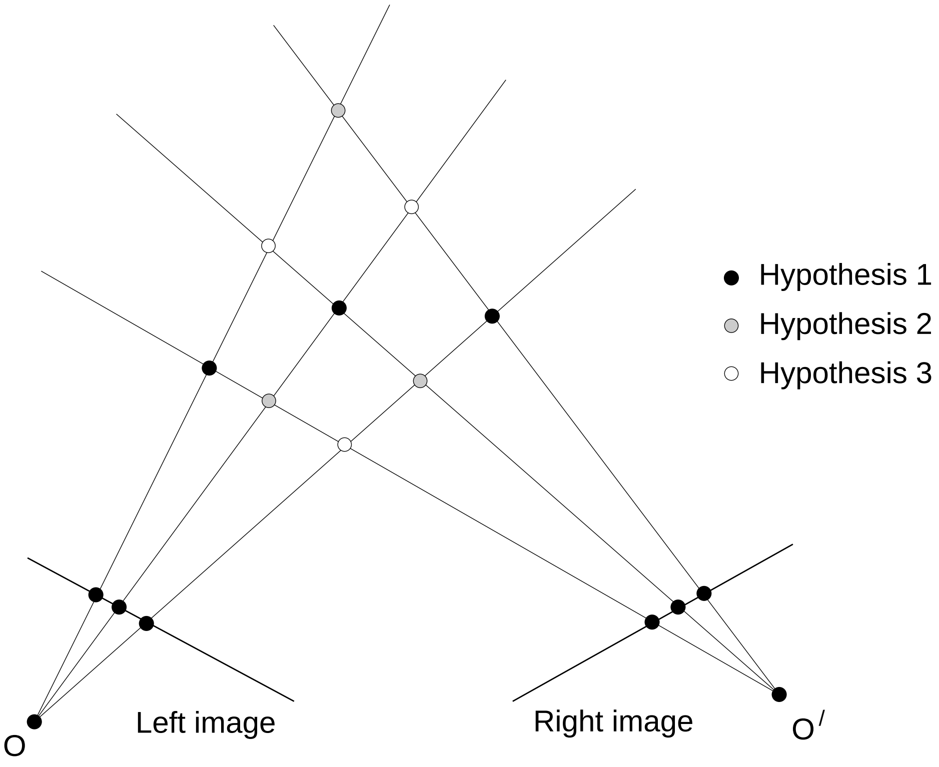
Given 8 or more perfect correspondences (image points in *general* position, no noise), F can be determined uniquely up to scale. In practice, we may have more than 8 correspondences and the image measurements will be noisy. The system of equations is then solved by least squares.

Note that we have not attempted to enforce the constraint that $\det \mathrm{F} = 0$. If the 8 image points are noisy, then we will find that our estimate of F does *not* have zero determinant and the epipolar lines do not meet at a point. Nonlinear techniques exist to estimate F from 7 point correspondences, enforcing the rank 2 constraint.

Given F, we can establish correspondences with *relative* ease. If we know the intrinsic camera parameters K, we can also find the essential matrix, decompose E into $\mathrm{T}_\times$ and R, and recover metric structure by triangulation. Without K we can only recover structure up to a 3D projective transformation, which can later be disambiguated using further constraints.

The correspondence problem

Even with the epipolar constraint, establishing correspondences between points in the left and right image is not trivial. Comparing image patches by correlation is unreliable since the grey levels are viewpoint dependent.



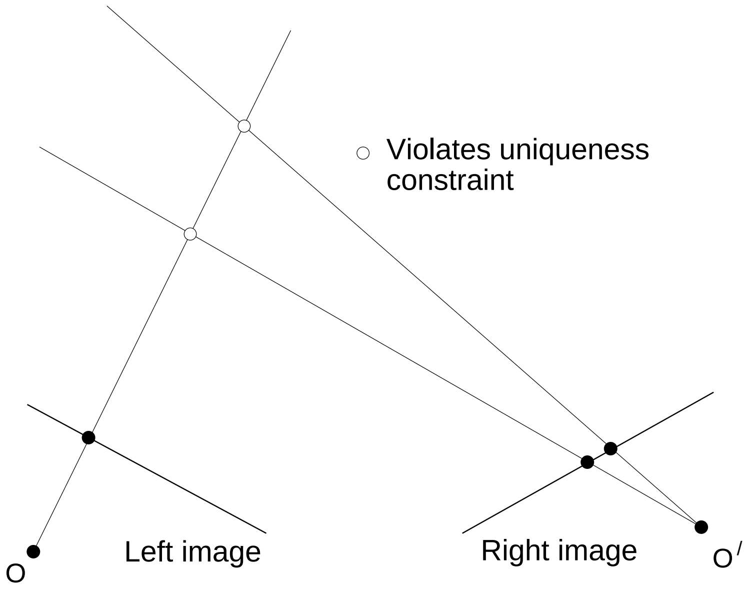
In the illustration, we are trying to match three corners in the left image with three corners in the right image. We have three hypotheses, all of which satisfy the epipolar constraint. How can we discover which hypothesis is correct?

The correspondence problem

The correspondence problem is difficult to solve, but we can make progress by identifying more constraints.

Uniqueness

The most obvious constraint is uniqueness. For opaque objects, each point in the left image has at most one match in the right image.

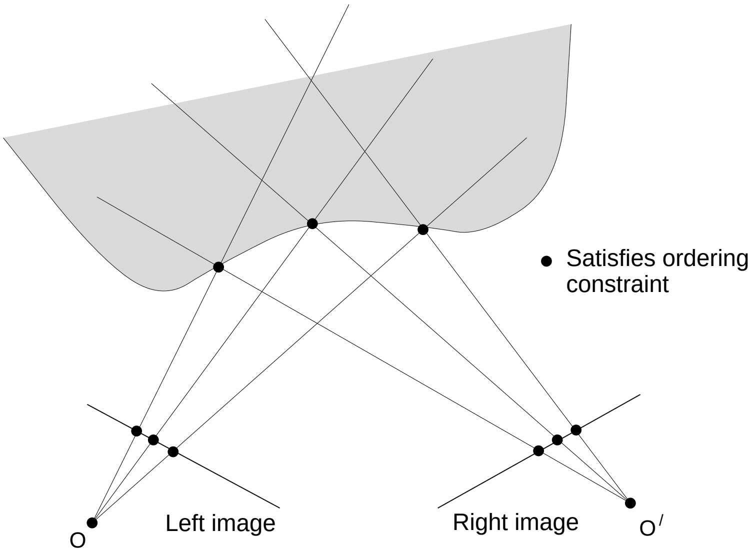


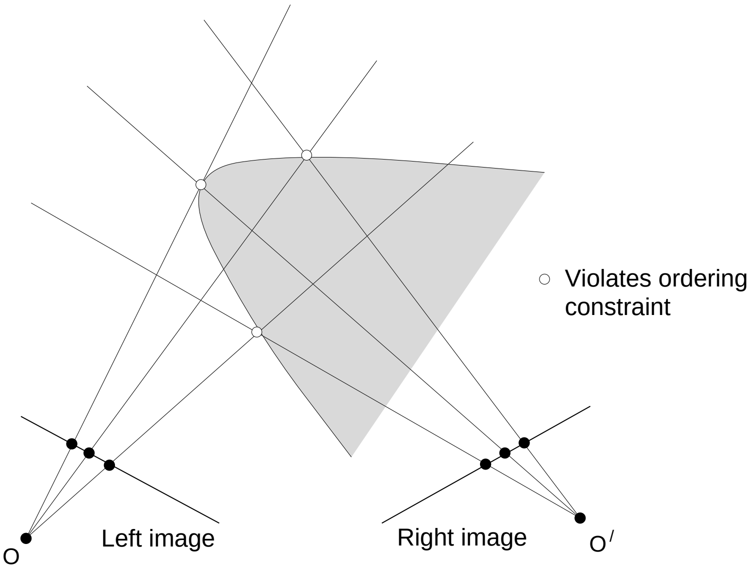
For transparent objects, we cannot rely on the uniqueness constraint. Two features may be visible in the right image but instantaneously aligned in the left image.

The correspondence problem

Ordering

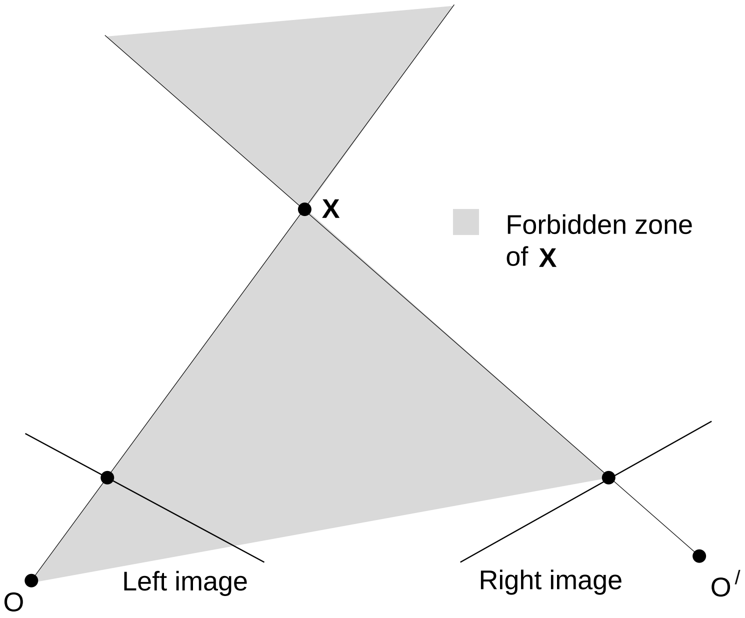
Corresponding points lying on the surface of an opaque object will be ordered identically in left and right images.





The correspondence problem

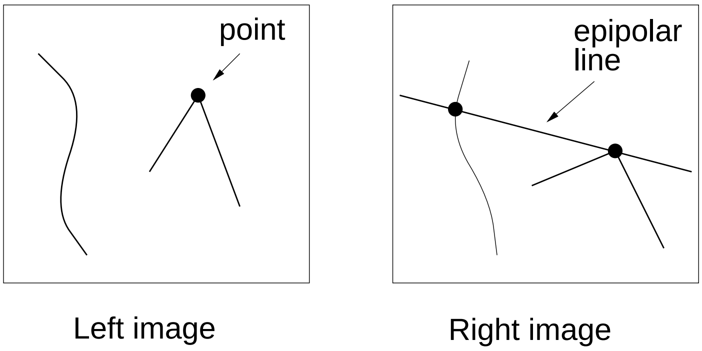
The ordering constraint will not necessarily hold if the points do not lie on the surface of the same opaque object. Given point $\mathbf{X}$ observed in both images, any point lying in $\mathbf{X}$'s "forbidden zone" will violate the ordering constraint.



Figural continuity

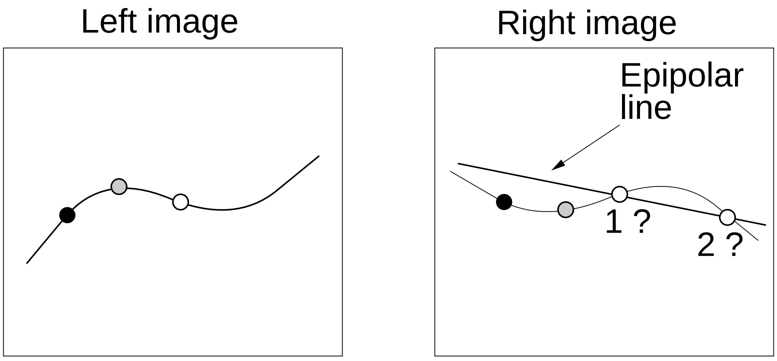
When distinguished points lie on image contours, we can sometimes use figural continuity as a matching constraint. In the following example, the point in the left image must match the point towards the right of the right image.

The correspondence problem



Disparity gradient

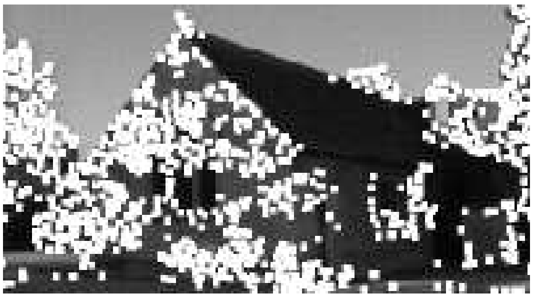
If surfaces are smooth, then disparities (differences in location between points in the left and right images) must be locally smooth. So, away from occluding boundaries, a further constraint comes from imposing a limit on the allowable spatial derivatives of disparity.

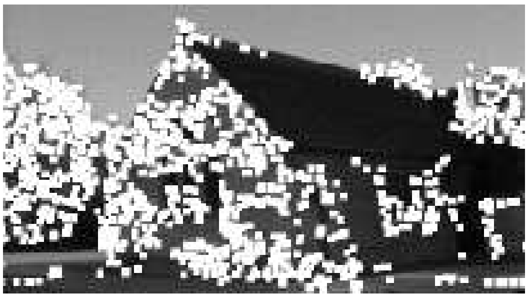


Given matches ● and ◉, point ○ in the left image must match point 1 in the right image. Point 2 would exceed the disparity gradient limit.

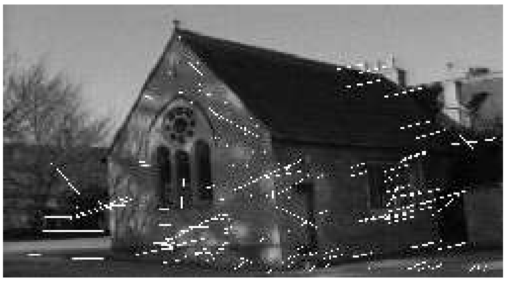
Finding correspondences

Here is the outline of an algorithm for finding correspondences between corners (typically 200–300 per image).





1. Unguided matching.

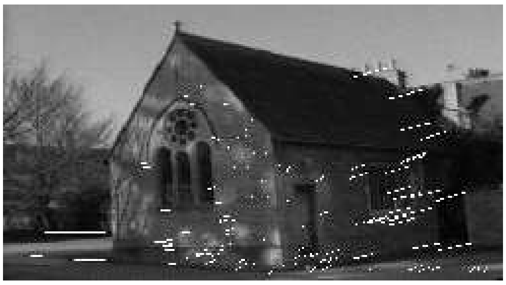


Seed matches

Use normalized cross-correlation to obtain a small number of seed matches.

2. Compute epipolar geometry. Use seed matches and robust regression to compute F.

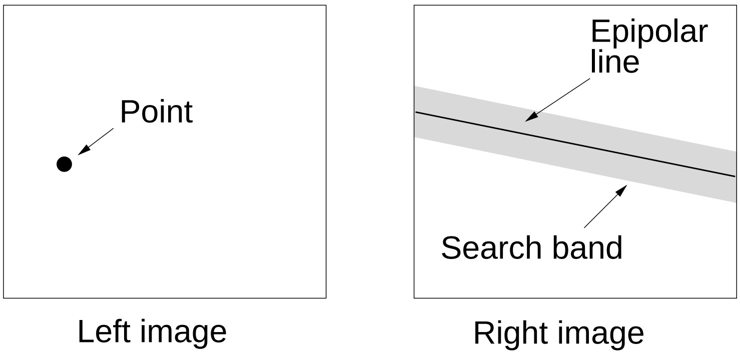
Find an F which is consistent with many of the seed matches, reject the rest as outliers.



Matches consistent with F

Finding correspondences

3. Guided matching. Now that we know F, the search for matches can be restricted to a narrow band around epipolar lines.



Using the epipolar and other constraints (ordering etc.), we obtain a large number of matches.

With intrinsically calibrated cameras, we can now recover structure by triangulation. Practical implementations first obtain the two projection matrices via a singular value decomposition of the essential matrix, then solve for structure using least squares.

Recovering metric structure

The SVD of the essential matrix is given by

$$\mathrm{E} = \mathrm{K}'^T \mathrm{F} \mathrm{K} = \mathrm{T}_\times \mathrm{R} = \mathrm{U} \Lambda \mathrm{V}^T$$

It can be shown that

$$\hat{\mathrm{T}}_\times = \mathrm{U} \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \mathrm{U}^T \text{ and } \mathrm{R} = \mathrm{U} \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathrm{V}^T$$

Then, aligning the left camera and world coordinate systems, we have

$$\mathrm{P} = \mathrm{K}\,[\, \mathrm{I} \mid \mathbf{0} \,] \text{ and } \mathrm{P}' = \mathrm{K}'\,[\, \mathrm{R} \mid \mathbf{T} \,]$$

Given the two projection matrices, we can recover structure (only up to scale, since we do not know $\|\mathbf{T}\|$) using least squares, as described on page 2. Ambiguities in $\mathbf{T}$ and R are resolved by ensuring that visible points lie in front of the two cameras.

We only recover structure at the detected corners: to reconstruct more of the scene, we use dense, intensity-based matching between corners. The recovered 3D points are then triangulated, and the visual appearance of the model improved by mapping texture from the images onto the model.

Factorization of fundamental matrix

Like the essential matrix, the fundamental matrix can be factorized into a skew-symmetric matrix corresponding to translation and a 3×3 non-singular matrix corresponding to rotation.

$$\begin{aligned} \mathbf{F} &= \mathbf{K}'^{-\top}[\mathbf{t}]_\times \mathbf{R}\mathbf{K}^{-1} \\ &= [\mathbf{K}'\mathbf{t}]_\times \mathbf{K}'\mathbf{R}\mathbf{K}^{-1} \\ &= [\mathbf{e}']_\times \mathbf{M}_\infty \end{aligned}$$

where $\mathbf{M}_\infty$ represents a 2D projective transformation induced by the plane at infinity:

$$\mathbf{M}_\infty = \mathbf{K}'\mathbf{R}\mathbf{K}^{-1}$$

The factorization of the fundamental matrix, however, is not unique. Any 2D projective transformation $\mathbf{M}$ of the form

$$\mathbf{M} = [\mathbf{e}']_\times \mathbf{F} + \mathbf{e}'\mathbf{v}^\top$$

will give the same fundamental matrix. This property leads to an ambiguity in the recovered projection matrices, known as the **projective ambiguity**.

Projective reconstruction

The factorization of the fundamental matrix can be used to compute the canonical cameras – the normalized projection matrices. These are real projection matrices, $\mathbf{P}$ and $\mathbf{P}'$, up to an arbitrary 3D projective transformation represented algebraically by a 4×4 matrix $\mathbf{H}$, and known as a projective ambiguity.

$$\begin{aligned} \mathbf{PH} &= [\mathbf{I} \mid \mathbf{0}] \\ \mathbf{P'H} &= [\mathbf{M} \mid \mathbf{e}'] \end{aligned}$$

The projective ambiguity is a 3D projective transformation and can be represented by a non-singular 4×4 matrix, $\mathbf{H}$, of the form

$$\mathbf{H} = \begin{bmatrix} s\mathbf{R_w} & \mathbf{t_w} \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{K}^{-1} & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{v}^\top & 1 \end{bmatrix}.$$

This ambiguity can only be removed with additional information derived from scene constraints or knowledge of the camera parameters, $\mathbf{K}$ and $\mathbf{K}'$. In particular the ambiguity is completely removed by using the 3D position of 5 known scene points to determine the transformation $\mathbf{H}$ or $\mathbf{H}^{-1}$.

Affine stereo

Recall that when depth variations in the scene are small compared with the viewing distance, an affine camera is appropriate:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

The affine camera can be calibrated by observing four points in space.

With two calibrated affine cameras, it is straightforward to triangulate to recover structure:

$$\begin{bmatrix} u \\ v \\ u' \\ v' \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p'_{11} & p'_{12} & p'_{13} & p'_{14} \\ p'_{21} & p'_{22} & p'_{23} & p'_{24} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \qquad (3)$$

Each point observed in left and right images gives us 4 equations in the 3 unknowns (X, Y, Z). These can be solved using least squares.

But what about an epipolar constraint to help with the correspondence problem?

Affine stereo

Assume (without loss of generality), that the left camera is aligned with the world coordinate system: this will simplify the algebra considerably. It is straightforward to show (by inspection of the weak perspective camera matrix) that the left camera matrix reduces to

✎

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} p_{11} & 0 & 0 & p_{14} \\ 0 & p_{22} & 0 & p_{24} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

We can now easily eliminate X and Y from the equations for u' and v' in (3):

$$\begin{aligned} u' &= p'_{11}\frac{(u-p_{14})}{p_{11}} + p'_{12}\frac{(v-p_{24})}{p_{22}} + p'_{13}Z + p'_{14} \\ v' &= p'_{21}\frac{(u-p_{14})}{p_{11}} + p'_{22}\frac{(v-p_{24})}{p_{22}} + p'_{23}Z + p'_{24} \end{aligned}$$

Rewriting these equations, we obtain

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} p'_{11}\frac{(u-p_{14})}{p_{11}} + p'_{12}\frac{(v-p_{24})}{p_{22}} + p'_{14} \\ p'_{21}\frac{(u-p_{14})}{p_{11}} + p'_{22}\frac{(v-p_{24})}{p_{22}} + p'_{24} \end{bmatrix} + Z \begin{bmatrix} p'_{13} \\ p'_{23} \end{bmatrix}$$

Affine stereo

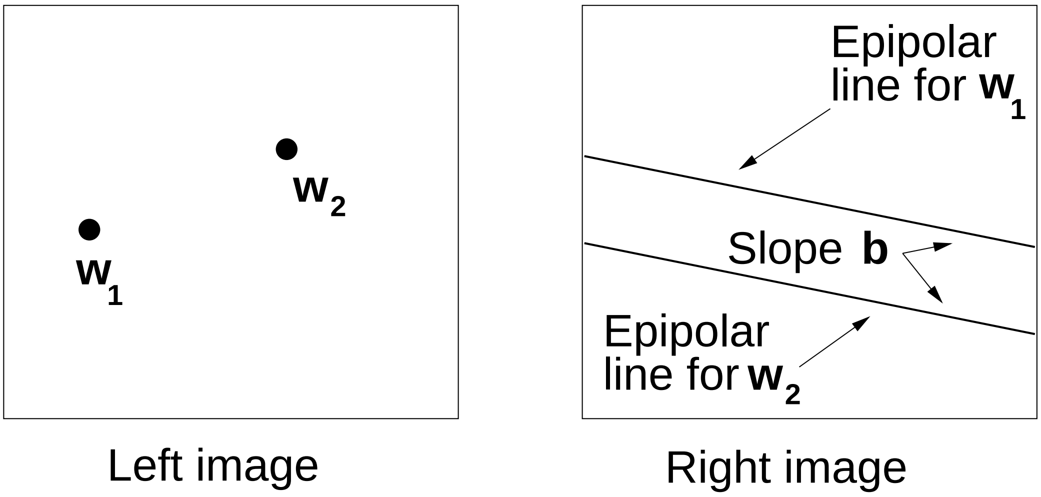
We can rewrite the preceding equation as

$$\mathbf{w}' = \mathbf{a} + Z\mathbf{b} \tag{4}$$

This is one form of the epipolar constraint for affine stereo (see the examples paper for another form).

Given calibrated cameras and a point $\mathbf{w}$ in the left image, we do not know Z but we do know $\mathbf{a}$ and $\mathbf{b}$. Thus, the corresponding point $\mathbf{w}'$ must lie on the epipolar line in the right image described by (4).

Since $\mathbf{b}$ is independent of $\mathbf{w}$, it follows that all epipolar lines are parallel under affine stereo.


$\mathbf{w_2}$
$\mathbf{w_1}$
Left image
Epipolar line for $\mathbf{w_1}$
Slope **b**
Epipolar line for $\mathbf{w_2}$
Right image

Affine Fundamental Matrix

The left (u, v) and right (u', v') pixel positions of a point in space viewed through *weak* perspective cameras also satisfy the epipolar constraints. By eliminating Z we can show that the pixel coordinates are related by

$$[\, u' \;\; v' \;\; 1 \,]\, F_A \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = 0$$

where F_A is the affine fundamental matrix which has maximum rank 2 and can be expressed in the form

$$F_A = \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & b \\ c & d & 1 \end{bmatrix}$$

The epipolar lines under weak perspective are parallel.

Summary

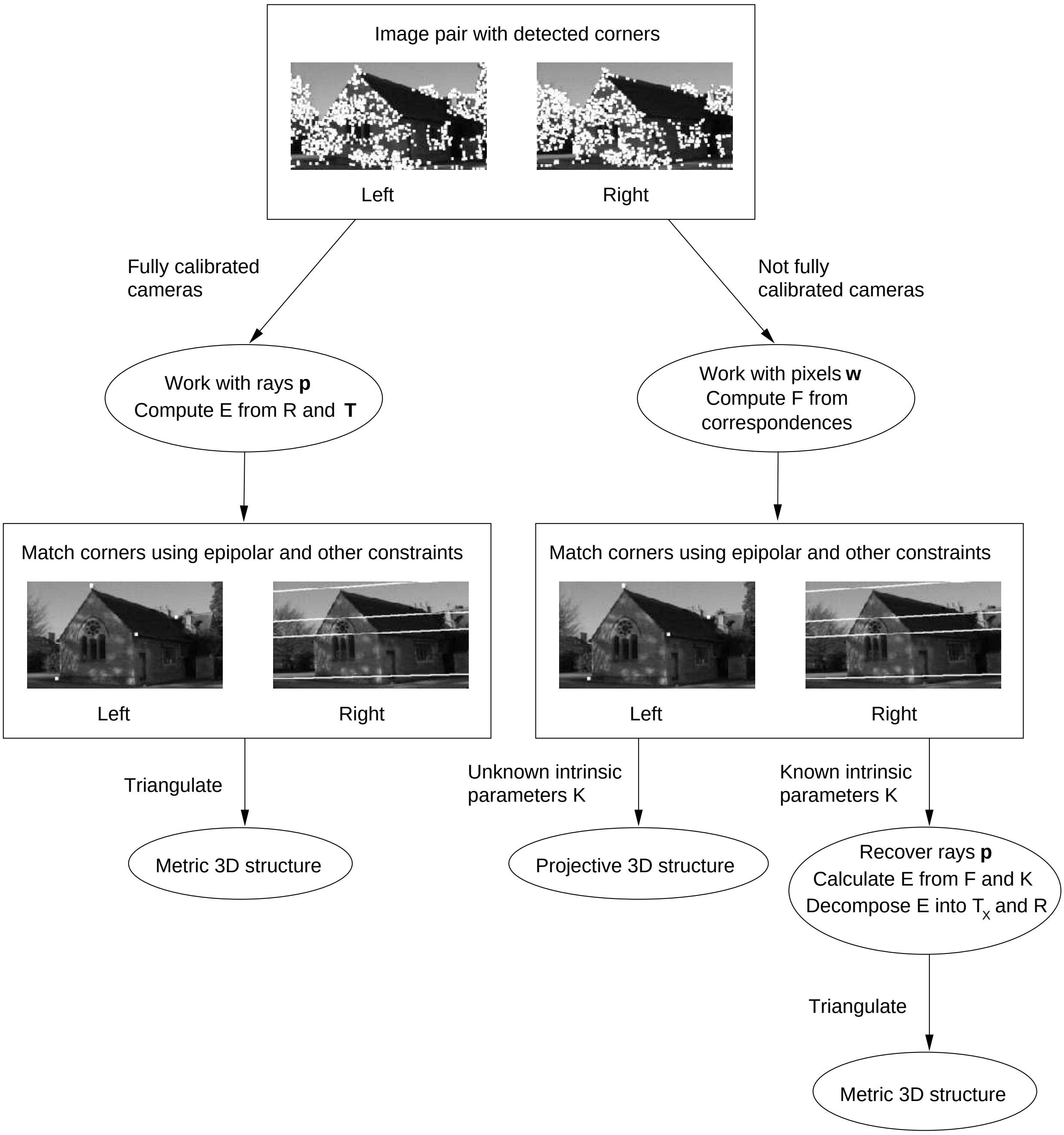


Image pair with detected corners
Left
Right
Fully calibrated cameras
Not fully calibrated cameras
Work with rays **p**
Compute E from R and **T**
Work with pixels **w**
Compute F from correspondences
Match corners using epipolar and other constraints
Left
Right
Match corners using epipolar and other constraints
Left
Right
Triangulate
Unknown intrinsic parameters K
Known intrinsic parameters K
Metric 3D structure
Projective 3D structure
Recover rays **p**
Calculate E from F and K
Decompose E into T_x and R
Triangulate
Metric 3D structure

Bibliography

The figures on pages 13, 14, 24 and 25 were reproduced (with thanks) from Andrew Zisserman's vision course notes. The following publications are included for additional reading.

Epipolar geometry and scene reconstruction

H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

O. D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *Proceedings of the 2nd European Conference on Computer Vision*, 563–578, 1992.

R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. CUP, 2003.

R. Cipolla and P. Giblin *Visual Motion of Curves and Surfaces.* CUP, 1999.

Stereo matching

S. B. Pollard, J. E. W. Mayhew and J. P. Frisby. PMF: a stereo correspondence algorithm using a disparity gradient. *Perception*, 14:449–470, 1985.