Chapter 5
Reconstruction from Two Views: Linear Algorithms

Multiple View Geometry
Summer 2016

Prof. Daniel Cremers
Chair for Computer Vision and Pattern Recognition
Department of Computer Science
Technische Universität München

Overview

1. **The Reconstruction Problem**
2. **The Epipolar Constraint**
3. **Eight-Point Algorithm**
4. **Structure Reconstruction**
5. **Four-Point Algorithm**
6. **The Uncalibrated Case**

Problem Formulation

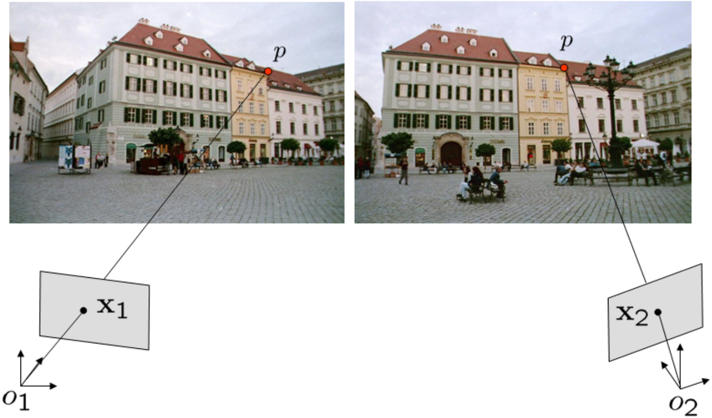
In the last sections, we discussed how to identify point correspondences between two consecutive frames. In this section, we will tackle the next problem, namely that of reconstructing the 3D geometry of cameras and points.

To this end, we will make the following assumptions:

- We assume that we are given a set of corresponding points in two frames taken with the same camera from different vantage points.
- We assume that the scene is static, i.e. none of the observed 3D points moved during the camera motion.
- We also assume that the intrinsic camera (calibration) parameters are known.

We will first estimate the camera motion from the set of corresponding points. Once we know the relative location and orientation of the cameras, we can reconstruct the 3D location of all corresponding points by triangulation.

Problem Formulation



Goal: Estimate camera motion and 3D scene structure from two views.

The Reconstruction Problem

In general $3D$ reconstruction is a challenging problem. If we are given two views with 100 feature points in each of them, then we have 200 point coordinates in $2D$. The goal is to estimate

- 6 parameters modeling the camera motion R, T and
- 100×3 coordinates for the $3D$ points X_j.

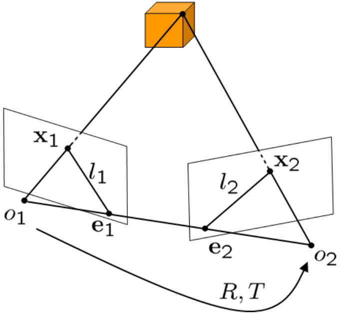
This could be done by minimizing the projection error:

$$E(R, T, X_1, \ldots, X_{100}) = \sum_j \|x_1^j - \pi(X_j)\|^2 + \|x_2^j - \pi(R, T, X_j)\|^2$$

This amounts to a difficult optimization problem called bundle adjustment.

Before we look into this problem, we will first study an elegant solution to entirely get rid of the $3D$ point coordinates. It leads to the well-known 8-point algorithm.

Epipolar Geometry: Some Notation



The projections of a point X onto the two images are denoted by x_1 and x_2. The optical centers of each camera are denoted by o_1 and o_2. The intersections of the line (o_1, o_2) with each image plane are called the epipoles e_1 and e_2. The intersections between the epipolar plane (o_1, o_2, X) and the image planes are called epipolar lines l_1 and l_2. There is one epipolar plane for each $3D$ point X.

The Epipolar Constraint

We know that x_1 (in homogeneous coordinates) is the projection of a $3D$ point X. Given known camera parameters ($K = 1$) and no rotation or translation of the first camera, we merely have a projection with unknown depth λ_1. From the first to the second frame we additionally have a camera rotation R and translation T followed by a projection. This gives the equations:

$$\lambda_1 x_1 = X, \qquad \lambda_2 x_2 = RX + T.$$

Inserting the first equation into the second, we get:

$$\lambda_2 x_2 = R(\lambda_1 x_1) + T.$$

Now we remove the translation by multiplying with $\widehat{T}$ ($\widehat{T} v \equiv T \times v$):

$$\lambda_2 \widehat{T} x_2 = \lambda_1 \widehat{T} R x_1$$

And projection onto x_2 gives the epipolar constraint:

$$\boxed{x_2^\top \widehat{T} R \, x_1 = 0}$$

The Epipolar Constraint

The epipolar constraint

$$\boxed{x_2^\top \, \widehat{T} R \, x_1 = 0}$$

provides a relation between the $2D$ point coordinates of a $3D$ point in each of the two images and the camera transformation parameters. The original $3D$ point coordinates have been removed. The matrix

$$E = \widehat{T} R \quad \in \mathbb{R}^{3\times 3}$$

is called the essential matrix. The epipolar constraint is also known as essential constraint or bilinear constraint.

Geometrically, this constraint states that the three vectors $\overrightarrow{o_1 X}$, $\overrightarrow{o_2 o_1}$ and $\overrightarrow{o_2 X}$ form a plane, i.e. the triple product of these vectors (measuring the volume of the parallelepiped) is zero: In coordinates of the second frame Rx_1 gives the direction of the vector $\overrightarrow{o_1 X}$; T gives the direction of $\overrightarrow{o_2 o_1}$, and x_2 is proportional to the vector $\overrightarrow{o_2 X}$ such that

$$volume = x_2^\top (T \times R x_1) = x_2^\top \, \widehat{T} R \, x_1 = 0.$$

Properties of the Essential Matrix E

The space of all essential matrices is called the essential space:

$$\mathcal{E} \equiv \left\{ \widehat{T}R \mid R \in SO(3),\ T \in \mathbb{R}^3 \right\} \quad \subset \mathbb{R}^{3\times 3}$$

Theorem [Huang & Faugeras, 1989] (Characterization of the essential matrix): A nonzero matrix $E \in \mathbb{R}^{3\times 3}$ is an essential matrix if and only if E has a singular value decomposition (SVD) $E = U\Sigma V^\top$ with

$$\Sigma = \text{diag}\{\sigma, \sigma, 0\}$$

for some $\sigma > 0$ and $U, V \in SO(3)$.

Theorem (Pose recovery from the essential matrix): There exist exactly two relative poses (R, T) with $R \in SO(3)$ and $T \in \mathbb{R}^3$ corresponding to an essential matrix $E \in \mathcal{E}$. For $E = U\Sigma V^\top$ we have:

$$(\widehat{T}_1, R_1) = \left(UR_Z(+\tfrac{\pi}{2})\Sigma U^\top,\ UR_Z^\top(+\tfrac{\pi}{2})V^\top \right), \qquad (1)$$

$$(\widehat{T}_2, R_2) = \left(UR_Z(-\tfrac{\pi}{2})\Sigma U^\top,\ UR_Z^\top(-\tfrac{\pi}{2})V^\top \right), \qquad (2)$$

In general, only one of these gives meaningful (positive) depth values.

A Basic Reconstruction Algorithm

We have seen that the $2D$-coordinates of each $3D$ point are coupled to the camera parameters R and T through an epipolar constraint. In the following, we will derive a $3D$ reconstruction algorithm which proceeds as follows:

- Recover the essential matrix E from the epipolar constraints associated with a set of point pairs.
- Extract the relative translation and rotation from the essential matrix E.

In general, the matrix E recovered from a set of epipolar constraints will not be an essential matrix. One can resolve this problem in two ways:

1. Recover some matrix $E \in \mathbb{R}^{3\times 3}$ from the epipolar constraints and then project it onto the essential space.
2. Optimize the epipolar constraints in the essential space.

While the second approach is in principle more accurate it involves a nonlinear constrained optimization. We will pursue the first approach which is simpler and faster.

The Eight-Point Linear Algorithm

First we rewrite the epipolar constraint as a scalar product in the elements of the matrix E and the coordinates of the points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Let

$$E^s = (e_{11}, e_{21}, e_{31}, e_{12}, e_{22}, e_{32}, e_{13}, e_{23}, e_{33})^\top \quad \in \mathbb{R}^9$$

be the vector of elements of E and

$$a \equiv \boldsymbol{x}_1 \otimes \boldsymbol{x}_2$$

the Kronecker product of the vectors $\boldsymbol{x}_i = (x_i, y_i, z_i)$, defined as

$$a = (x_1 x_2,\ x_1 y_2,\ x_1 z_2,\ y_1 x_2,\ y_1 y_2,\ y_1 z_2,\ z_1 x_2,\ z_1 y_2,\ z_1 z_2)^\top \in \mathbb{R}^9.$$

Then the epipolar constraint can be written as:

$$\boxed{\boldsymbol{x}_2^\top E\, \boldsymbol{x}_1 = a^\top E^s = 0.}$$

For n point pairs, we can combine this into the linear system:

$$\boxed{\chi E^s = 0, \qquad \text{with } \chi = (a^1, a^2, \ldots, a^n)^\top.}$$

The Eight-Point Linear Algorithm

According to

$$\chi E^s = 0, \qquad \text{with } \chi = (a^1, a^2, \ldots, a^n)^\top.$$

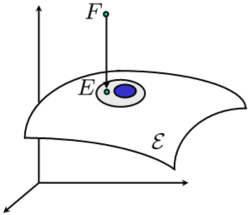
we see that the vector of coefficients of the essential matrix E defines the null space of the matrix χ. In order for the above system to have a unique solution (up to a scaling factor and ruling out the trivial solution $E = 0$), the rank of the matrix χ needs to be exactly 8. Therefore we need at least 8 point pairs.

In certain degenerate cases, the solution for the essential matrix is not unique even if we have 8 or more point pairs. One such example is the case that all points lie on a line or on a plane.

Clearly, we will not be able to recover the sign of E. Since with each E, there are two possible assignments of rotation R and translation T, we therefore end up with four possible solutions for rotation and translation.

Projection onto Essential Space

The numerically estimated coefficients E^s will in general not correspond to an essential matrix. One can resolve this problem by projecting it back to the essential space.



Theorem (Projection onto essential space): Let $F \in \mathbb{R}^{3\times3}$ be an arbitrary matrix with SVD $F = U\,\text{diag}\{\lambda_1, \lambda_2, \lambda_3\}\,V^\top,\ \lambda_1 \geq \lambda_2 \geq \lambda_3$. Then the essential matrix E which minimizes the Frobenius norm $\|F - E\|_f^2$ is given by

$$E = U\,\text{diag}\{\sigma, \sigma, 0\}\,V^\top, \quad \text{with } \sigma = \frac{\lambda_1 + \lambda_2}{2}.$$

Eight Point Algorithm (Longuet-Higgins '81)

Given a set of $n = 8$ or more point pairs $\boldsymbol{x}_1^i, \boldsymbol{x}_2^i$:

- Compute an approximation of the essential matrix. Construct the matrix $\chi = (a^1, a^2, \ldots, a^n)^\top$, where $a^i = \boldsymbol{x}_1^i \otimes \boldsymbol{x}_2^i$. Find the vector $E^s \in \mathbb{R}^9$ which minimizes $\|\chi E^s\|$ as the ninth column of V_χ in the SVD $\chi = U_\chi \Sigma_\chi V_\chi^\top$. Unstack E^s into 3×3-matrix E.

- Project onto essential space. Compute the SVD $E = U \operatorname{diag}\{\sigma_1, \sigma_2, \sigma_3\} V^\top$. Since in the reconstruction, E is only defined up to a scalar, we project E onto the normalized essential space by replacing the singular values $\sigma_1, \sigma_2, \sigma_3$ with $1, 1, 0$.

- Recover the displacement from the essential matrix. The four possible solutions for rotation and translation are:

$$R = U R_Z^\top(\pm \tfrac{\pi}{2}) V^\top, \quad \widehat{T} = U R_Z(\pm \tfrac{\pi}{2}) \Sigma U^\top,$$

with a rotation by $\pm\frac{\pi}{2}$ around z:

$$R_Z^\top(\pm \tfrac{\pi}{2}) = \begin{pmatrix} 0 & \pm 1 & 0 \\ \mp 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Do We Need Eight Points?

The above reasoning showed that we need at least eight points in order for the matrix χ to have rank 8 and therefore guarantee a unique solution for E. Yet, one can take into account the special structure of E. The space of essential matrices is actually a five-dimensional space, i.e. E only has 5 (and not 9) degrees of freedom.

A simple way to take into account the algebraic properties of E is to make use of the fact that $\det E = 0$. If now we have only 7 point pairs, the null space of χ will have (at least) Dimension 2, spanned by two vectors E_1 and E_2. Then we can solve for E by determining α such that:

$$\det E = \det(E_1 + \alpha E_2) = 0.$$

Along similar lines, Kruppa proved in 1913 that one needs only five point pairs to recover (R, T). In the case of degenerate motion (for example planar or circular motion), one can resolve the problem with even fewer point pairs.

Limitations and Further Extensions

Among the four possible solutions for R and T, there is generally only one meaningful one (which assigns positive depth to all points).

The algorithm fails if the translation is exactly 0, since then $E = 0$ and nothing can be recovered. Due to noise this typically does not happen.

In the case of infinitesimal view point change, one can adapt the eight point algorithm to the continuous motion case, where the epipolar constraint is replaced by the continuous epipolar constraint. Rather than recovering (R, T) one recovers the linear and angular velocity of the camera.

In the case of independently moving objects, one can generalize the epipolar constraint. For two motions for example we have:

$$(\boldsymbol{x}_2^\top E_1 \boldsymbol{x}_1)(\boldsymbol{x}_2^\top E_2 \boldsymbol{x}_1) = 0$$

with two essential matrices E_1 and E_2. Given a sufficiently large number of point pairs, one can solve the respective equations for multiple essential matrices using polynomial factorization.

Structure Reconstruction

The linear eight-point algorithm allowed us to estimate the camera transformation parameters R and T from a set of corresponding point pairs. Yet, the essential matrix E and hence the translation T are only defined up to an arbitrary scale $\gamma \in \mathbb{R}^+$, with $\|E\| = \|T\| = 1$. After recovering R and T, we therefore have for point X^j:

$$\lambda_2^j \boldsymbol{x}_2^j = \lambda_1^j R \boldsymbol{x}_1^j + \gamma T, \quad j = 1, \dots, n,$$

with unknown scale parameters λ_i^j. We can eliminate one of these scales by applying $\widehat{\boldsymbol{x}_2^j}$:

$$\lambda_1^j \widehat{\boldsymbol{x}_2^j} R \boldsymbol{x}_1^j + \gamma \widehat{\boldsymbol{x}_2^j} T = 0, \quad j = 1, \dots, n.$$

This corresponds to n linear systems of the form

$$\begin{pmatrix} \widehat{\boldsymbol{x}_2^j} R \boldsymbol{x}_1^j, & \widehat{\boldsymbol{x}_2^j} T \end{pmatrix} \begin{pmatrix} \lambda_1^j \\ \gamma \end{pmatrix} = 0. \quad j = 1, \dots, n.$$

Structure Reconstruction

Combining the parameters $\vec{\lambda} = (\lambda_1^1, \lambda_1^2, \ldots, \lambda_1^n, \gamma)^\top \in \mathbb{R}^{n+1}$, we get the linear equation system

$$\boxed{M\vec{\lambda} = 0}$$

with

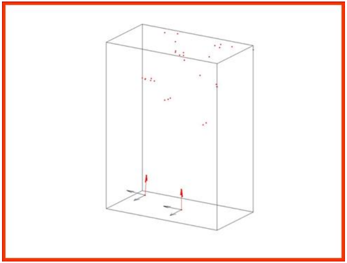
$$M \equiv \begin{pmatrix} \widehat{\boldsymbol{x}_2^1} R \boldsymbol{x}_1^1 & 0 & 0 & 0 & 0 & \widehat{\boldsymbol{x}_2^1} T \\ 0 & \widehat{\boldsymbol{x}_2^2} R \boldsymbol{x}_1^2 & 0 & 0 & 0 & \widehat{\boldsymbol{x}_2^2} T \\ 0 & 0 & \ddots & 0 & 0 & \vdots \\ 0 & 0 & 0 & \widehat{\boldsymbol{x}_2^{n-1}} R \boldsymbol{x}_1^{n-1} & 0 & \widehat{\boldsymbol{x}_2^{n-1}} T \\ 0 & 0 & 0 & 0 & \widehat{\boldsymbol{x}_2^n} R \boldsymbol{x}_1^n & \widehat{\boldsymbol{x}_2^n} T \end{pmatrix}.$$

The linear least squares estimate for $\vec{\lambda}$ is given by the eigenvector corresponding to the smallest eigenvalue of $M^\top M$. It is only defined up to a global scale. It reflects the ambiguity that the camera has moved twice the distance, the scene is twice larger and twice as far away.

Example

Left image

right image



Reconstruction (Author: Jana Košecká)

Degenerate Configurations

The eight-point algorithm only provides unique solutions (up to a scalar factor) if all $3D$ points are in a "general position". This is no longer the case for certain degenerate configurations, for which all points lie on certain $2D$ surfaces which are called critical surfaces.

Typically these critical surfaces are described by a quadratic equation in the three point coordinates, such that they are referred to as quadratic surfaces.

While most critical configurations do not actually arise in practice, a specific degenerate configuration which does arise often is the case that all points lie on a $2D$ plane (such as floors, table, walls,...).

For the structure-from-motion problem in the context of points on a plane, one can exploit additional constraints which leads to the so-called four-point algorithm.

Planar Homographies

Let us assume that all points lie on a plane. If $\boldsymbol{X}_1 \in \mathbb{R}^3$ denotes the point coordinates in the first frame, and these lie on a plane with normal $N \in \mathbb{S}^2$, then we have:

$$N^\top \boldsymbol{X}_1 = d \quad \Leftrightarrow \quad \frac{1}{d} N^\top \boldsymbol{X}_1 = 1.$$

In frame two, we therefore have the coordinates:

$$\boldsymbol{X}_2 = R\boldsymbol{X}_1 + T = R\boldsymbol{X}_1 + T\frac{1}{d}N^\top \boldsymbol{X}_1 = \left(R + \frac{1}{d}TN^\top\right)\boldsymbol{X}_1 \equiv H\boldsymbol{X}_1,$$

where

$$H = R + \frac{1}{d}TN^\top \quad \in \mathbb{R}^{3\times 3}$$

is called a homography matrix. Inserting the $2D$ coordinates, we get:

$$\lambda_2 \boldsymbol{x}_2 = H\lambda_1 \boldsymbol{x}_1 \quad \Leftrightarrow \quad \boxed{\boldsymbol{x}_2 \sim H\boldsymbol{x}_1,}$$

where $\sim$ means equality up to scaling. This expression is called a planar homography. H depends on camera and plane parameters.

From Point Pairs to Homography

For a pair of corresponding $2D$ points we therefore have

$$\lambda_2 \boldsymbol{x}_2 = H \lambda_1 \boldsymbol{x}_1.$$

By multiplying with $\widehat{\boldsymbol{x}_2}$ we can eliminate λ_2 and obtain:

$$\boxed{\widehat{\boldsymbol{x}_2} H \boldsymbol{x}_1 = 0}$$

This equation is called the planar epipolar constraint or planar homography constraint.

Again, we can cast this equation into the form

$$\boldsymbol{a}^\top H^s = 0,$$

where we have stacked the elements of H into a vector

$$H^s = (H_{11}, H_{21}, \ldots, H_{33})^\top \in \mathbb{R}^9,$$

and introduced the matrix

$$\boldsymbol{a} \equiv \boldsymbol{x}_1 \otimes \widehat{\boldsymbol{x}_2} \ \in \ \mathbb{R}^{9\times 3}.$$

The Four Point Algorithm

Let us now assume we have $n \geq 4$ pairs of corresponding $2D$ points $\{\boldsymbol{x}_1^j, \boldsymbol{x}_2^j\},\ j = 1, \ldots, n$ in the two images. Each point pair induces a matrix $\boldsymbol{a}^j$, we integrate these into a larger matrix

$$\chi \equiv (\boldsymbol{a}^1, \ldots, \boldsymbol{a}^n)^\top \ \in \ \mathbb{R}^{3n \times 9},$$

and obtain the system

$$\chi H^s = 0.$$

As in the case of the essential matrix, the homography matrix can be estimated up to a scale factor.

This gives rise to the four point algorithm:

- For the point pairs, compute the matrix χ.
- Compute a solution H^s for the above equation by singular value decomposition of χ.
- Extract the motion parameters from the homography matrix $H = R + \frac{1}{d} T N^\top$.

General Comments

Clearly, the derivation of the four-point algorithm is in close analogy to that of the eight-point algorithm.

Rather than estimating the essential matrix E one estimates the homography matrix H to derive R and T. In the four-point algorithm, the homography matrix is decomposed into R, N and T/d. In other words, one can reconstruct the normal of the plane, but the translation is only obtained in units of the offset d of the plane and the origin.

The $3D$ structure of the points can then be computed in the same manner as before.

Since one uses the strong constraint that all points lie in a plane, the four-point algorithm only requires four correspondences.

There exist numerous relations between the essential matrix $E = \widehat{T}R$ and the corresponding homography matrix $H = R + Tu^\top$ with some $u \in \mathbb{R}^3$, in particular:

$$E = \widehat{T}H, \quad H^\top E + E^\top H = 0.$$

The Case of an Uncalibrated Camera

The reconstruction algorithms introduced above all assume that the camera is calibrated ($K = 1$). The general transformation from a $3D$ point to the image is given by:

$$\lambda \boldsymbol{x}' = K \, \Pi_0 \, g \, \boldsymbol{X} = (KR, KT)\boldsymbol{X},$$

with the intrinsic parameter matrix or calibration matrix:

$$K = \begin{pmatrix} fs_x & fs_\theta & o_x \\ 0 & fs_y & o_y \\ 0 & 0 & 1 \end{pmatrix} \quad \in \mathbb{R}^{3\times 3}.$$

The calibration matrix maps metric coordinates into image (pixel) coordinates, using the focal length f, the optical center o_x, o_y, the pixel size s_x, s_y and a skew factor s_θ. If these parameters are known then one can simply transform the pixel coordinates $\boldsymbol{x}'$ to normalized coordinates $\boldsymbol{x} = K^{-1}\boldsymbol{x}'$ to obtain the representation used in the previous sections. This amounts to centering the coordinates with respect to the optical center etc.

The Fundamental Matrix

If the camera parameters K cannot be estimated in a calibration procedure beforehand, then one has to deal with reconstruction from uncalibrated views.

By transforming all image coordinates $\boldsymbol{x}'$ with the inverse calibration matrix K^{-1} into metric coordinates $\boldsymbol{x}$, we obtain the epipolar constraint for uncalibrated cameras:

$$\boldsymbol{x}_2^\top \widehat{T} R \boldsymbol{x}_1 = 0 \quad \Leftrightarrow \quad \boldsymbol{x}_2'^\top K^{-\top} \widehat{T} R K^{-1} \boldsymbol{x}_1' = 0,$$

which can be written as

$$\boxed{\boldsymbol{x}_2'^\top F \boldsymbol{x}_1' = 0,}$$

with the fundamental matrix defined as:

$$\boxed{F \equiv K^{-\top} \widehat{T} R K^{-1} = K^{-\top} E K^{-1}.}$$

Since the invertible matrix K does not affect the rank of this matrix, we know that F has an SVD $F = U \Sigma V^\top$ with $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, 0)$. In fact, any matrix of rank 2 can be a fundamental matrix.

Prof. Daniel Cremers

Limitations

While it is straight-forward to extend the eight-point algorithm, such that one can extract a fundamental matrix from a set of corresponding image points, it is less straight forward how to proceed from there.

Firstly, one cannot impose a strong constraint on the specific structure of the fundamental matrix (apart from the fact that the last singular value is zero).

Secondly, for a given fundamental matrix F, there does not exist a finite number of decompositions into extrinsic parameters R, T and intrinsic parameters K (even apart from the global scale factor).

As a consequence, one can only determine so-called projective reconstructions, i.e. reconstructions of geometry and camera position which are defined up to a so-called projective transformation.

As a solution, one typically choses a canonical reconstruction from the family of possible reconstructions.