COMPUTER VISION
Two-view Geometry

Emanuel Aldea <emanuel.aldea@u-psud.fr>
http://hebergement.u-psud.fr/emi/

Computer Science and Multimedia Master - University of Pavia

Outline

The 3D representation of points

In the 3D space :

$$\underbrace{\mathbf{p} = (X, Y, Z)^T = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}}_{\text{initial point}} \quad \underbrace{\mathbf{p}' = (X', Y', Z')^T = \begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix}}_{\text{same point in different coordinate system}}$$

Euclidean transform $\mathbf{p}' = \mathbf{R}\mathbf{p} + \mathbf{t}$ becomes in homogeneous coordinates :

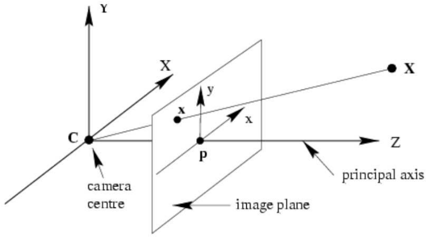
$$\begin{bmatrix} X' \\ Y' \\ Z' \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

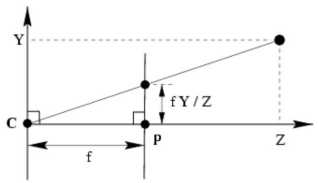
or otherwise $\tilde{\mathbf{p}}' = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix} \tilde{\mathbf{p}}$, avec $\mathbf{R}^T\mathbf{R} = \mathbf{I}$, $\det \mathbf{R} = 1$

- the transform has six degrees of freedom (three elementary rotations, three elementary translations)
- we discard the ˜ for the sake of simplicity, but when it makes sense the variables are homogeneous

Outline

The pinhole camera model





3D $\Rightarrow$ 2D projection

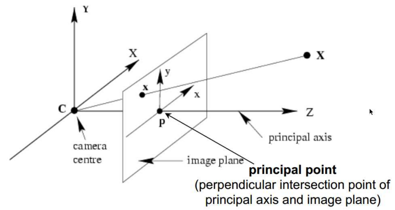
- In the 3D focal plance : $(X, Y, Z)^T \Rightarrow (fX/Z, fY/Z, f)^T$
- In the image 2D plane : $(X, Y, Z)^T \Rightarrow (fX/Z, fY/Z) = (x, y)$

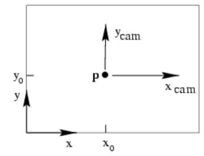
The pinhole camera model

The image plane projection $(fX/Z, fY/Z)$ gives in homogeneous coordinates :

$$\begin{bmatrix} fX \\ fY \\ Z \end{bmatrix} = \begin{bmatrix} f & & \\ & f & \\ & & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & & & 0 \\ & 1 & & 0 \\ & & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \mathrm{diag}(f, f, 1)[\mathbf{I}|\mathbf{0}]\mathbf{X}$$

Problem : usually, the chosen reference in the image plane is not the projection of the optical axis :





This gives in the reference system we use commonly :

$$(X, Y, Z) \Rightarrow (fX/Z + p_x, fY/Z + p_y)$$

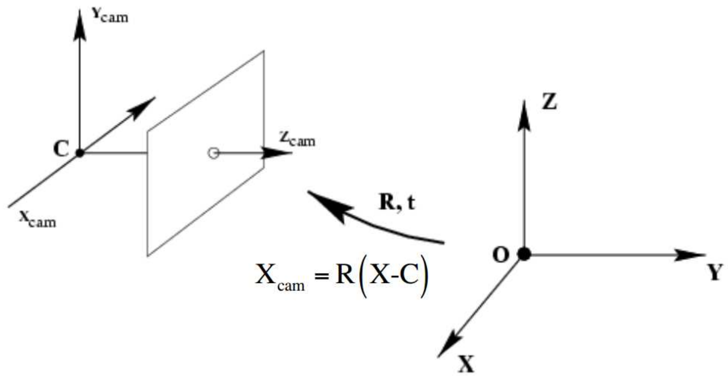
$$\begin{bmatrix} fX \\ fY \\ Z \end{bmatrix} = \begin{bmatrix} f & & p_x \\ & f & p_y \\ & & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & & & 0 \\ & 1 & & 0 \\ & & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \mathrm{diag}(f, f, 1)[\mathbf{I}|\mathbf{0}]\mathbf{X}$$

Outline

- The 3D representation of points
- The pinhole camera model
- **Applying a coordinate transformation**
- Homogeneous representations and algebraic operations
- The fundamental matrix
- The essential matrix
- Rectification

Transformation to an inertial (fixed) frame

Final step of the modelling : we express the 3D variables in a frame which is not attached to the camera and which is fixed (typical setting for mobile robotics) :



By denoting as **C** the center of the camera in "world" coordinates, the transform world to camera is expressed as

$$\mathbf{X}_{cam} = \begin{bmatrix} \mathbf{R} & -\mathbf{RC} \\ \mathbf{0}^T & 1 \end{bmatrix} \mathbf{X}$$

Outline

- The 3D representation of points
- The pinhole camera model
- Applying a coordinate transformation
- Homogeneous representations and algebraic operations
- The fundamental matrix
- The essential matrix
- Rectification

Homogeneous representation of 2D lines and points

- A 2D line is defined by $ax + by + c = 0$ i.e. a parametrization $\mathbf{l} = (a, b, c)$.
- However, $kax + kby + kc = 0$ corresponds to the same line, thus $\mathbf{l} = (ka, kb, kc), \forall k \in \mathbb{R} \setminus \{0\}$
- A 2D point (x, y) lies on a line (a, b, c) if $ax + by + c = 0$.
- This may be expressed as $(x, y, 1)^T \cdot (a, b, c) = (x, y, 1)^T \cdot \mathbf{l} = 0$.
- $\forall k \in \mathbb{R} \setminus \{0\}, (kx, ky, k)^T \cdot \mathbf{l} = 0$ if and only if $(x, y, 1)^T \cdot \mathbf{l} = 0$.
- $\forall k \in \mathbb{R} \setminus \{0\}$, we denote thus (kx, ky, k) as the homogeneous representation of the 2D point (x, y).
- An arbitrary homogeneous $\mathbf{x} = (x_1, x_2, x_3)$ corresponds to the 2D point $(x_1/x_3, x_2/x_3)$.
- Result : the point $\mathbf{x}$ lies on the line $\mathbf{l}$ if and only if $\mathbf{x}^T\mathbf{l} = 0$.
- Result : the intersection of two lines $\mathbf{l}$ and $\mathbf{l}'$ is the point $\mathbf{x} = \mathbf{l} \times \mathbf{l}'$.
- Result : the line through two points $\mathbf{x}$ and $\mathbf{x}'$ is $\mathbf{l} = \mathbf{x} \times \mathbf{x}'$.

Some quick vector operations

$$\mathbf{x} \times \mathbf{y} = \mathbf{x}_\times \cdot \mathbf{y} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} = \begin{pmatrix} x_2y_3 - x_3y_2 \\ x_3y_1 - x_1y_3 \\ x_1y_2 - y_1x_2 \end{pmatrix}$$

$$\mathbf{x}_\times = \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}$$

Mixed product : $\mathbf{x}^T(\mathbf{y} \times \mathbf{z}) = |\mathbf{x} \;\; \mathbf{y} \;\; \mathbf{z}|$ (the volume of the parallelepiped defined by the three vectors)

Singular value decomposition

Theorem (SVD) :
Let $\mathbf{A}$ be an $m \times n$ matrix. $\mathbf{A}$ may be expressed as :

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{i=1}^{\min(m,n)} \sigma_i U_i V_i^T$$

where $\mathbf{\Sigma}$ is a $m \times n$ diagonal matrix with $\sigma_i = \mathbf{\Sigma}_{ii} \geq 0$, and $\mathbf{U}$ $(m \times m)$ and $\mathbf{V}$ $(n \times n)$ are composed of orthornormal columns

- The rank of $\mathbf{A}$ is the number of $\sigma_i > 0$
- An orthonormal basis for the null space of $\mathbf{A}$ is composed of V_i for indices i such that $\sigma_i = 0$
- By convention, the σ_i are aligned in descending order by the decomposition algorithms.

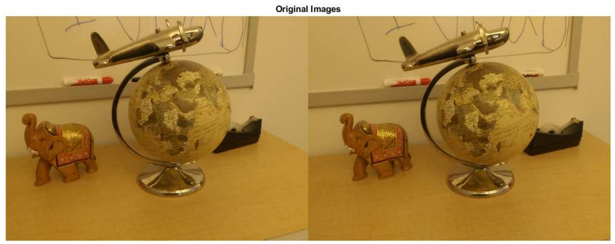
Outline

- The 3D representation of points
- The pinhole camera model
- Applying a coordinate transformation
- Homogeneous representations and algebraic operations
- The fundamental matrix
- The essential matrix
- Rectification

Why is this part "fundamental" ? (cheap joke)

What we can get from two views :

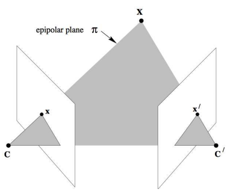
- Sparse 3D reconstruction
- Relative camera pose estimation
- Parametric surface fitting
- Dense 3D reconstruction (more complex work required for this)
- ... but also many multi-view algorithms extend nicely from two-view analysis



The anatomy of two views

Some important observations :

- the pixel projection is along the ray defined by the 3D point and the camera center (i.e. as for $\mathbf{x}$, $\mathbf{X}$ and $\mathbf{C}$)
- conversely, if $\mathbf{x}$ and $\mathbf{x}'$ do correspond to the same 3D point, the two rays intersect
- the two rays define a plane π denoted as *epipolar plane*
- the epipolar plane also contains the ray defined by the camera centers



The anatomy of two views

From the projection in the two views we have :

$$\lambda \mathbf{x} = \mathbf{K}\mathbf{X} \quad \lambda' \mathbf{x}' = \mathbf{K}'(\mathbf{R}\mathbf{X} + \mathbf{t})$$

By eliminating $\mathbf{X}$ we get :

$$\mathbf{X} = \lambda \mathbf{K}^{-1}\mathbf{x} \quad \lambda' \mathbf{x}' = \mathbf{K}'(\lambda \mathbf{R}\mathbf{K}^{-1}\mathbf{x} + \mathbf{t})$$

$$\lambda' \mathbf{K}'^{-1}\mathbf{x}' = \lambda \mathbf{R}\mathbf{K}^{-1}\mathbf{x} + \mathbf{t}$$

We eliminate the sum by applying a cross product with $\mathbf{t}$:

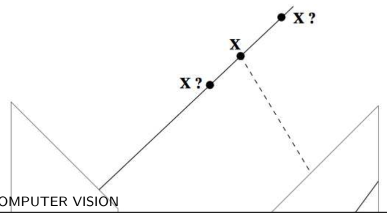
$$\lambda' \mathbf{t}_\times \mathbf{K}'^{-1}\mathbf{x}' = \lambda \mathbf{t}_\times \mathbf{R}\mathbf{K}^{-1}\mathbf{x}$$

We multiply by $\mathbf{K}'^{-1}\mathbf{x}'$ in order to get a null mixed product :

$$0 = \lambda \mathbf{K}'^{-1}\mathbf{x}' \mathbf{t}_\times \mathbf{R}\mathbf{K}^{-1}\mathbf{x}$$

Finally, by transposing $\mathbf{K}'^{-1}\mathbf{x}'$ and ignoring the scalar λ we get :

$$\mathbf{x}'^T \underbrace{\mathbf{K}'^{-T}\mathbf{t}_\times \mathbf{R}\mathbf{K}^{-1}}_{\mathbf{F}} \mathbf{x} = 0$$



The fundamental matrix F

$$\boxed{\mathbf{x}'^T \mathbf{F}\mathbf{x} = 0}$$

- applying the **F** constraint does not require information about the scene 3D structure
- **F** is valid for the whole image
- we may apply the constraint without performing/knowing the camera calibration
- For a given point $\mathbf{x}'$,we denote by $\mathbf{l}'$ its corresponding *epipolar line*. It follows from $\mathbf{x}'^T \mathbf{F}\mathbf{x} = 0$ that
$$\mathbf{l}' = \mathbf{F}\mathbf{x}$$
- Similarly, $\mathbf{l} = \mathbf{F}^T\mathbf{x}'$
- The fundamental matrix constraint translates to a search along the epipolar line ...
- ... but also $\mathbf{F} = \mathbf{K}'^{-T}\mathbf{t}_\times \mathbf{R}\mathbf{K}^{-1}$ encodes, along with the calibration matrices, *the rotation and translation* between views

The fundamental matrix F

Theorem

The condition which is necessary and sufficient for a matrix **F** to be a fundamental matrix is that
$$\det(\mathbf{F}) = 0$$

Multiple ways to notice that **F** is rank deficient :

- it follows from the fact that $\det(\mathbf{t}_\times) = 0$
- it follows from the fact that $\mathbf{F}\mathbf{e} = 0$

Computing F - the 8 point algorithm

Straightforward approach :

- each observation (match) provides a constraint on F as $\mathbf{x_i}'^T \mathbf{F}\mathbf{x_i} = 0$
- if we group the unknowns as the column vector $\mathbf{f} = [f_{11}\ f_{12} \dots f_{33}]$, the constraint may be expressed as $\mathbf{a_i}\mathbf{f} = 0$, with $\mathbf{a_i}$ a row vector
- only 8 parameters are independent, since the scale is not determined
- the search for **f** may be expressed as :
$$\min_{\mathbf{f}} \|\mathbf{A}\mathbf{f}\|, \text{subject to } \|\mathbf{f}\| = 1$$
where $\mathbf{A} = [\mathbf{a_1}\ \mathbf{a_2} \dots \mathbf{a_8}]$
- Solution : **f** is the last column of **V**, where $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ is the SVD of **A**
- Proof :
$\|\mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{f}\| = \|\mathbf{D}\mathbf{V}^T\mathbf{f}\|$, and $\|\mathbf{f}\| = \|\mathbf{V}^T\mathbf{f}\|$. We have to minimize $\|\mathbf{D}\mathbf{V}^T\mathbf{f}\|$ subject to $\|\mathbf{V}^T\mathbf{f}\| = 1$. If $\mathbf{y} = \mathbf{V}^T\mathbf{f}$, then we minimize $\|\mathbf{D}\mathbf{y}\|$ subject to $\|\mathbf{y}\| = 1$. Since **D** is diagonal with values in descending order, it means that $\mathbf{y} = (0, 0 \dots, 1)$, and $\mathbf{f} = \mathbf{V}\mathbf{y}$ is the last column of **V**. (*A5.3, Hartley and Zisserman*)

Considerations - the 8 point algorithm

Straightforward approach :

- major issue : the solution **F** may violate the rank constraint !
- Hack : decompose **F** using SVD, set $\sigma_3 = 0$ and recompose.
- What about searching directly for a rank 2 solution for **F** ?

The 7 point algorithm :

- Use 7 constraints for $\mathbf{A}\mathbf{f} = \mathbf{0}$
- Use SVD on **A** in order to find the vectors $\mathbf{f_1}$ and $\mathbf{f_2}$ that span the null space (the kernel) of **A**
- Find an element in the kernel expressed by the linear combination $\mathbf{f} = \mathbf{f_1} + \alpha\mathbf{f_2}$ which also satisfies $\det(\mathbf{F}) = 0$
- $\det(\mathbf{F_1} + \alpha\mathbf{F_2})$ is a third degree polynomial, so up to three potential solutions may be recovered
- This algorithm is also preferred as fewer observations are needed

Outline

- The 3D representation of points
- The pinhole camera model
- Applying a coordinate transformation
- Homogeneous representations and algebraic operations
- The fundamental matrix
- The essential matrix
- Rectification

Using the camera calibration and the essential matrix

If the calibration matrices $\mathbf{K}$ and $\mathbf{K}'$ are known :

- we may recover the pose information from $\mathbf{F} = \mathbf{K}'^{-T}\mathbf{t}_\times \mathbf{R}\mathbf{K}^{-1}$:

$$\mathbf{E} = \mathbf{t}_\times \mathbf{R} = \mathbf{K}'^{T}\mathbf{F}\mathbf{K}$$

- $\mathbf{E}$ has five degrees of freedom (and not six) because the relative translation $\mathbf{t}$ has a scale ambiguity (just as $\mathbf{F}$).
- Beside $\det(\mathbf{E}) = 0$, there is an additional constraint with respect to $\mathbf{F}$, which results from the structure of $\mathbf{E}$:
  Theorem : The condition which is necessary and sufficient for a matrix $\mathbf{E}$ to be an essential matrix is that two of its singular values be equal, and the third one be 0.
- There are thus at least five points needed for recovering directly $\mathbf{E}$ from an image pair, assuming that the calibration matrices are known, and there is an algorithm which solves this minimal problem(Nistér, David. "An efficient solution to the five-point relative pose problem." IEEE Transactions on Pattern Analysis and Machine Intelligence (2004).)
- Knowing $\mathbf{E}$: interesting for relative pose estimation
- Main disadvantage : $\mathbf{K}$ and $\mathbf{K}'$ are required to get to $\mathbf{E}$

Recovering R and t from E

It has been shown that the decomposition of $\mathbf{E}$ is possible and there are actually four valid solutions (*9.6.2, Hartley and Zisserman*) :

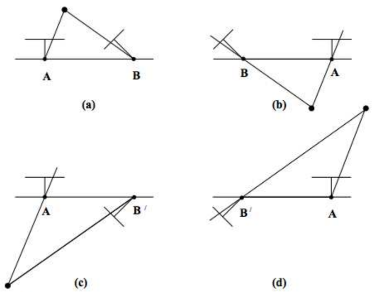


Fig. 9.12. **The four possible solutions for calibrated reconstruction from E.** *Between the left and right sides there is a baseline reversal. Between the top and bottom rows camera B rotates 180° about the baseline. Note, only in (a) is the reconstructed point in front of both cameras.*

- Identify the correct solution : cheirality check (the 3D points have to be in front of the camera) with an additional match from the two views

Outline

- The 3D representation of points
- The pinhole camera model
- Applying a coordinate transformation
- Homogeneous representations and algebraic operations
- The fundamental matrix
- The essential matrix
- Rectification

Rectification

Using $\mathbf{F}$, we restrict the search for the corresponding projection $\mathbf{x}'$ of a point $\mathbf{x}$ to a line (the epipolar line $\mathbf{l}' = \mathbf{F}\mathbf{x}$).

Stereo rectification

- Apply an adjustment to the images in order to get horizontal epipolar lines in both views
- The search for $\mathbf{x}'$ takes place simply along the same corresponding row in the second image : interesting for dense correspondence
- This implies that epipoles are at horizontal infinity : $\mathbf{e} = \mathbf{e}' = [1\ 0\ 0]^T$
- Apply a virtual rotation of cameras (Fusiello, A. ; Trucco, E. ; Verri, A. A compact algorithm for rectification of stereo pairs. Mach. Vision Appl 2000)
- An interpolation is required for creating the new images, but high computation gain overall