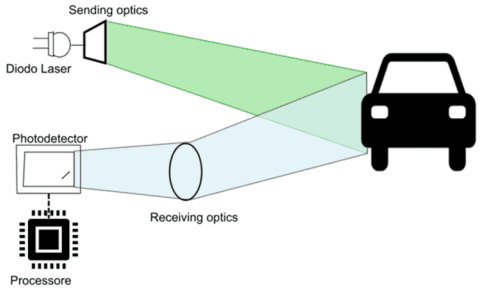
UniGe | MaLGa

Principles of 3D computer vision

Francesca Odone francesca.odone@unige.it

Introduction

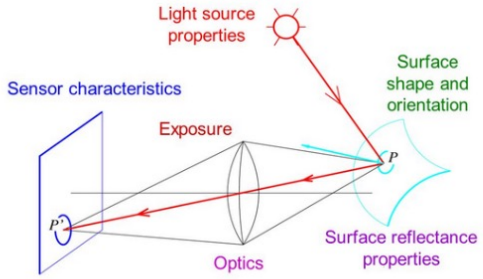
Active 3D Sensors



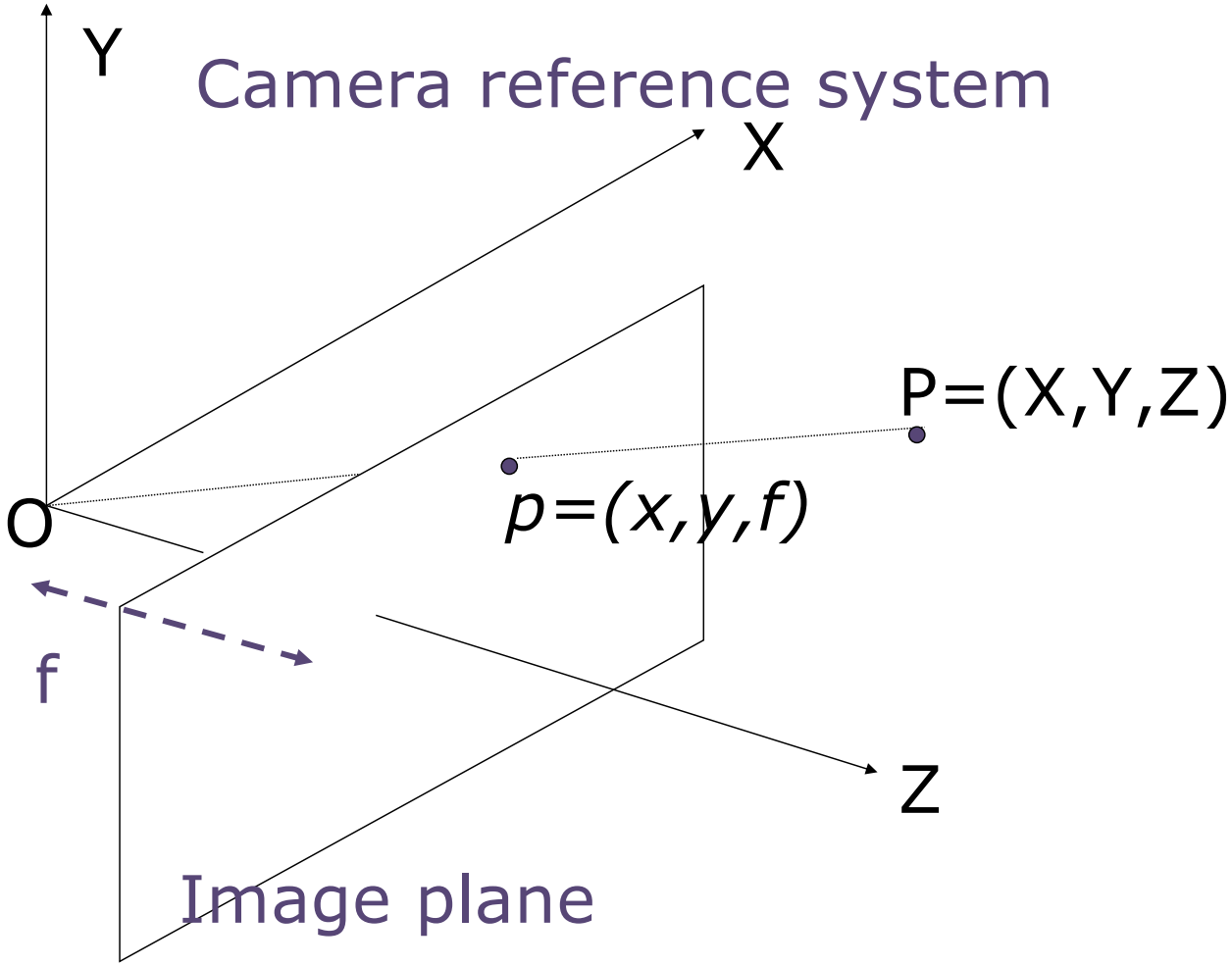
Time-of-Flight active sensors to measure distance

- LIDAR (laser based,short to medium range)
- IR or RGBD (IR based, short range)
- RADAR (RF based, medium to long range)

RGB camera sensors



The geometry of image formation

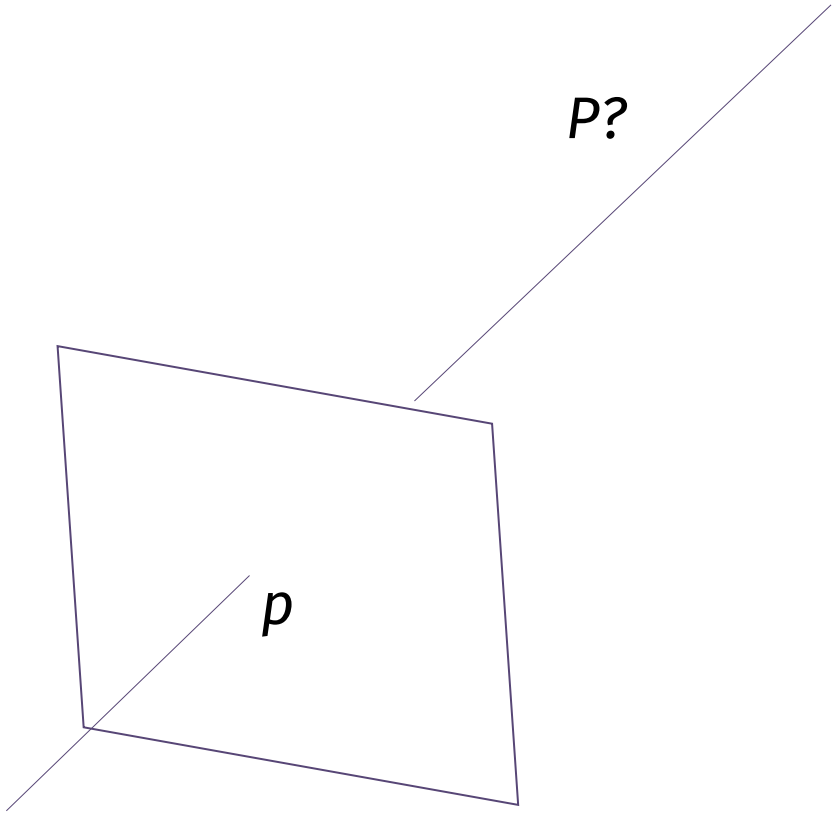


f focal length

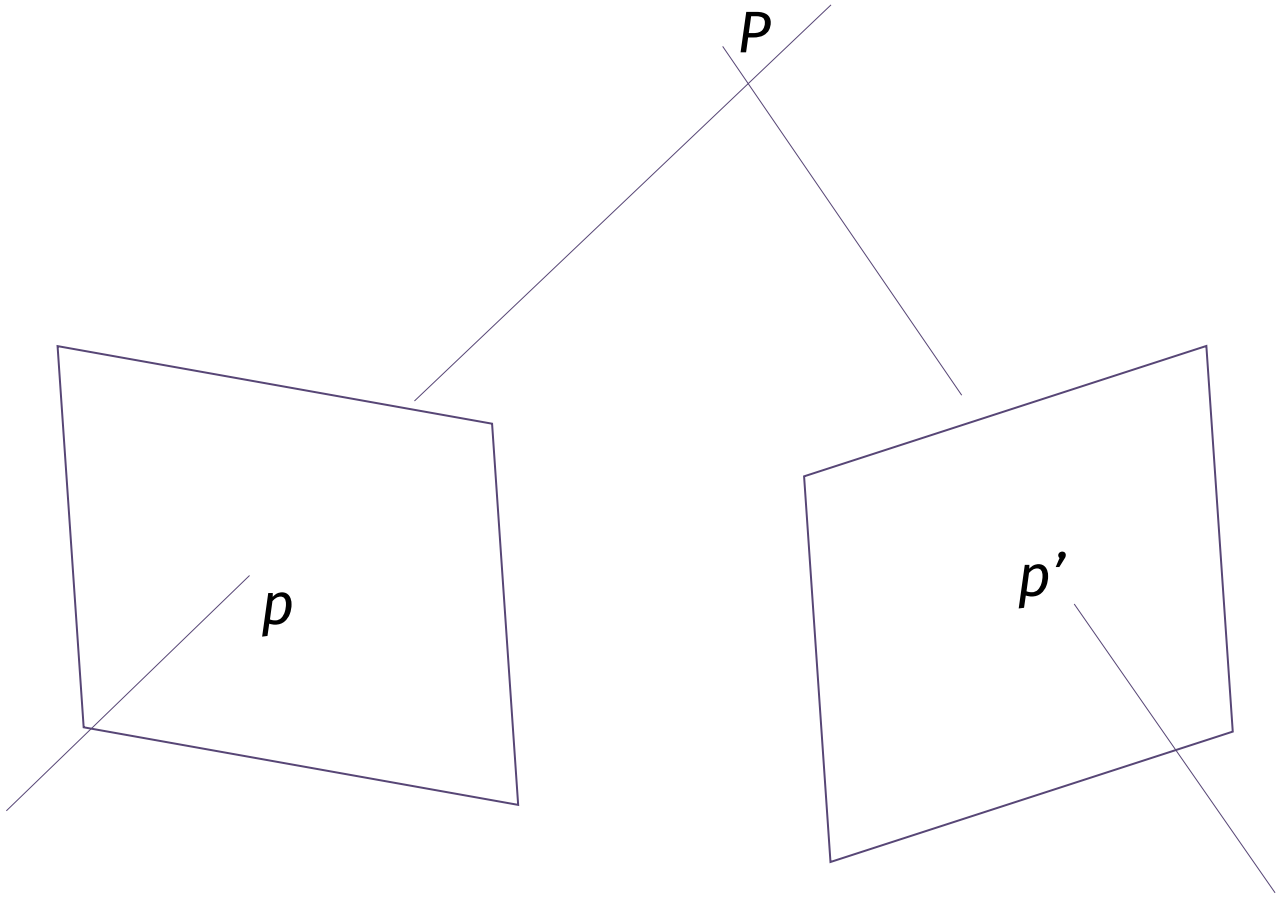
Scale Ambiguity

- 3D Information is lost during the projection to the image plane (or the retina)
- It is not possible (without a proper prior knowledge) to distinguish small objects from far objects

The perspective projection taking place when an image is acquired is a “lossy” non invertible transformation



With a second view…



UniGe | MaLGa

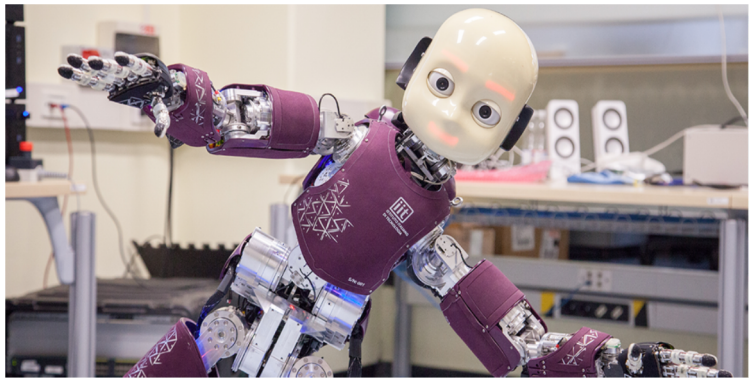
Stereopsis

Introduction

we refer to **stereo vision** as the problem of **inferring 3D information** (structure and distances) from two or more images taken from different viewpoints

we consider an acquisition system with 2 cameras

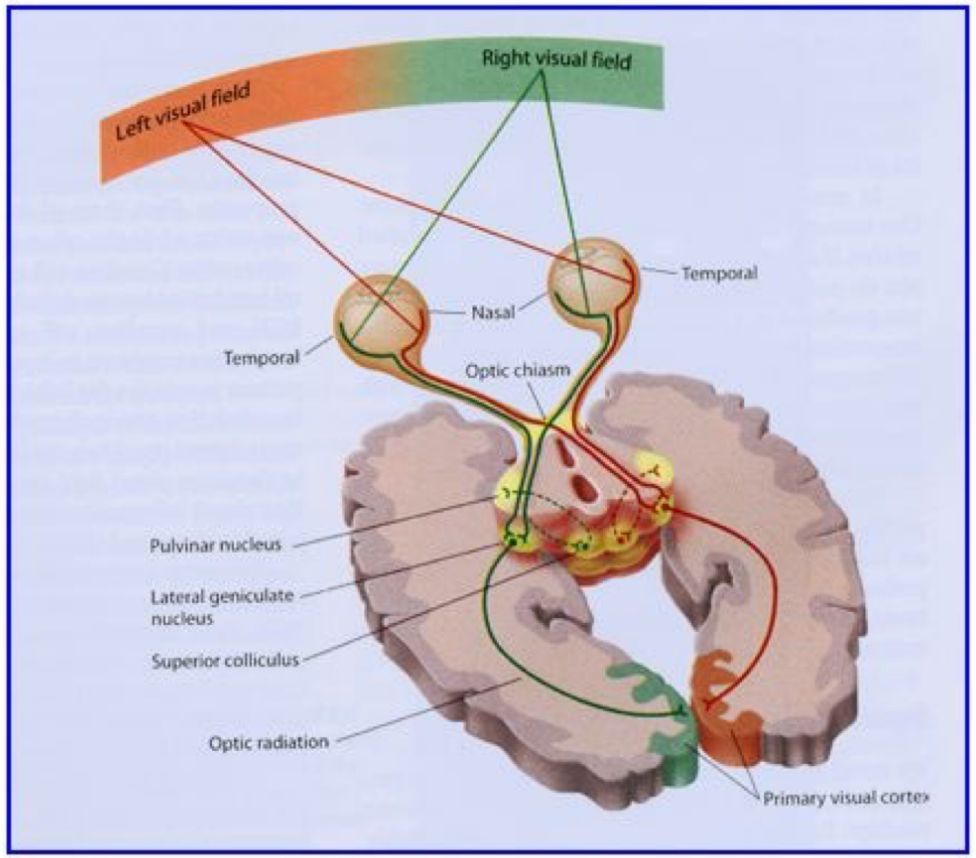
- “explicit” system: a stereo rig



- “implicit”: one moving camera

Stereopsis: perception of depth

Our 3D perception of the world is due to the interpretation that the brain gives of the computed difference in retinal position, called *disparity* between corresponding items



A simple stereo system

Relationship between depth and disparity

We can recover depth information

$$Z = \frac{fT}{x_r - x_l} = \frac{fT}{d}$$

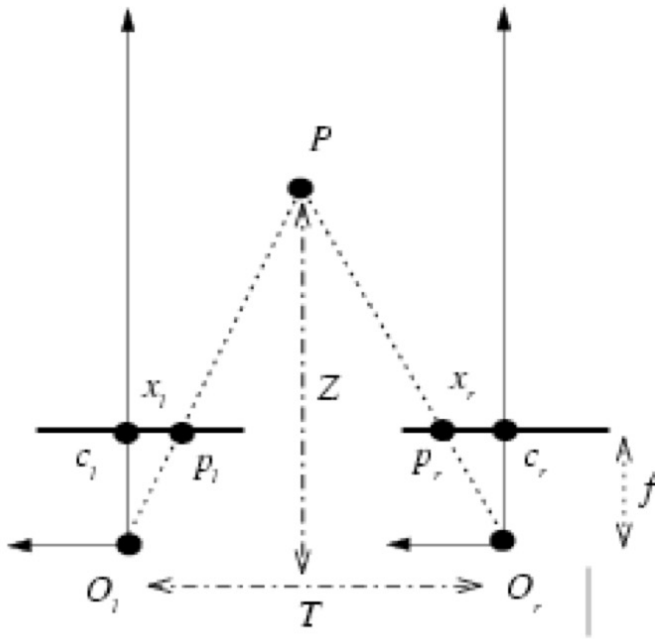
focal length

baseline

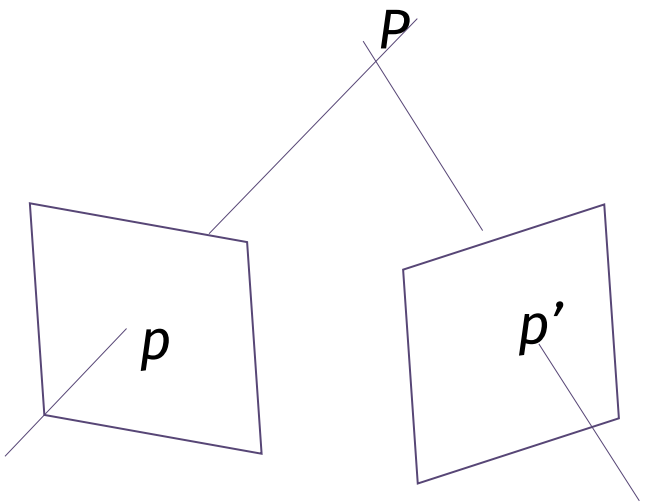
disparity

Depth is inversely proportional to disparity

Special case:
fixation point at infinity



The problems of stereopsis



stereo covers two main problems:

- finding **correspondences** between image pairs (p,p’)
- **reconstructing** the 3D position of a point P given its corresponding projections on the images

step 1: produces 2.5 disparity maps (or depth maps)

step 2: produces a 3D point cloud

UniGe | MaLGa

Disparity

Correspondence problem

the correspondence problem involves two decisions:

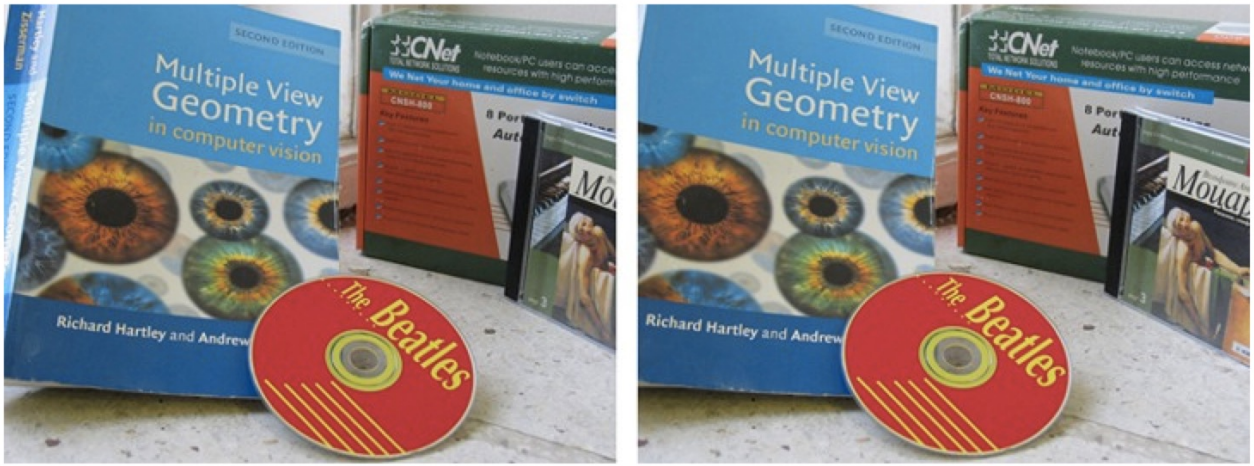
1. which image elements to match
2. which feature description + similarity measure to adopt

(1). can be solved in two ways:

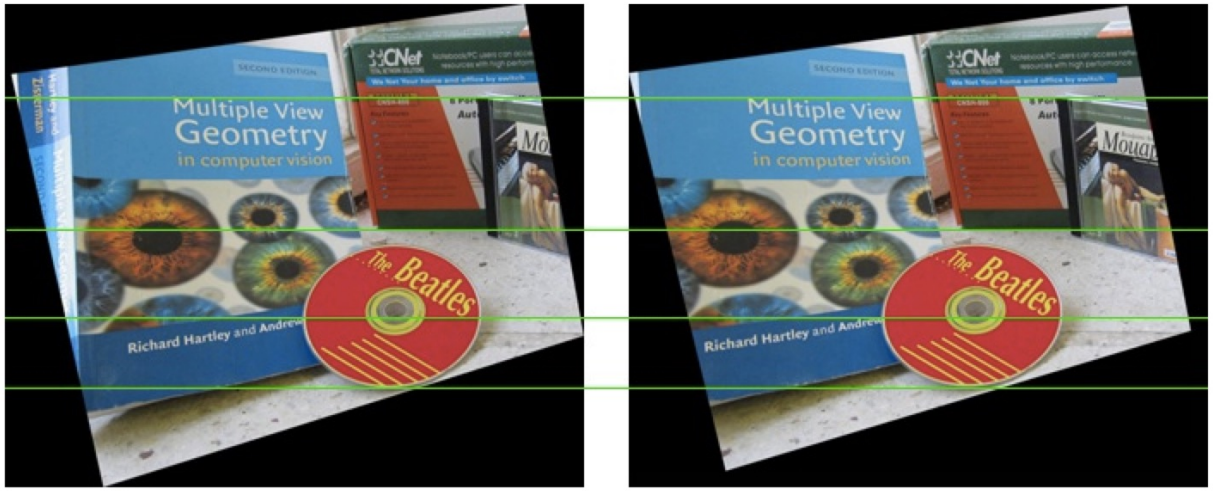
- use all pixels in the image (obtaining *dense correspondences or disparity maps*)
- use subsets of pixels meeting some requirements (obtaining *sparse correspondences)*

Computing disparity maps

Two words on input images



A stereo pair



A rectified stereo pair

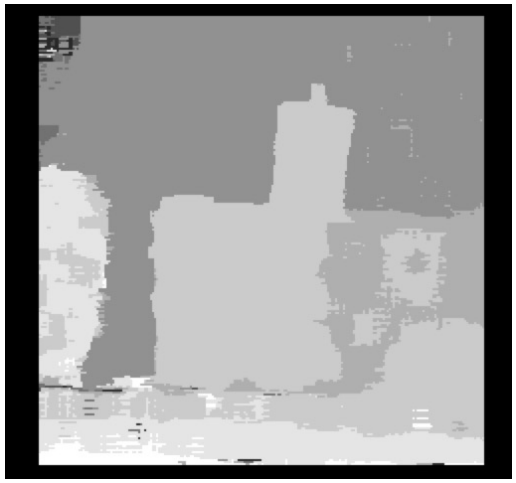
Equiv to fixation point at infinity

Dense correspondences

we consider the so-called *correlation methods* applied to image patches (neighbourhoods of pixels)

we assume we have two *rectified images*, where conjugate points lie on corresponding scanlines of the image (“rows”)

our goal is to obtain a **disparity map** giving the relative displacement for each pixel



assuming a fixation point at infinity disparity is proportional to the inverse of the distance

in a standard color coding bright areas correspond to high disparities (closer objects)

Dense correspondences: algorithm sketch

input:

- a stereo pair of rectified images *Il* and *Ir*
- size of a correlation window W
- a search range [dmin, dmax]

for each pixel *pl* of *(i,j)* coordinates in *Il*

- for each disparity d in the search range
  - estimate the similarity $c(d) = \phi(N1(i,j), N2(i,j+d))$
  - the disparity of the pixel is $\overline{d} = \text{argmax}_{d \in [d_{min}, d_{max}]}\{c(d)\}$

For instance SSD or NCC (see image matching class)

Dense correspondences: left-right consistency

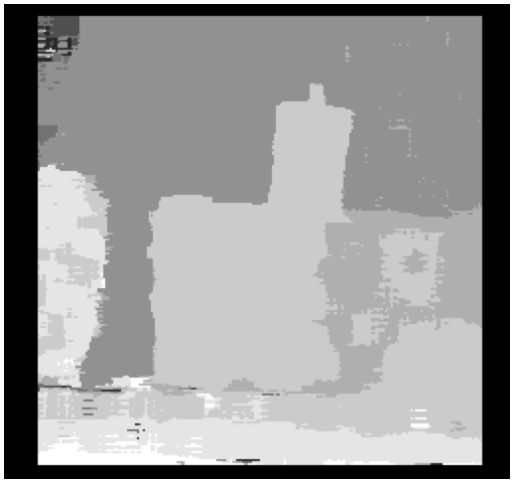
- correspondences are made more difficult by occlusions (points with no counterpart on the other image)

- let us compute

  - *Dlr*: disparity map from *Il* to *Ir*
  - *Drl*: disparity map from *Ir* to *Il*

- then D(i,j)=d iff Dlr(i,j)= -Drl (i,j+d)= d

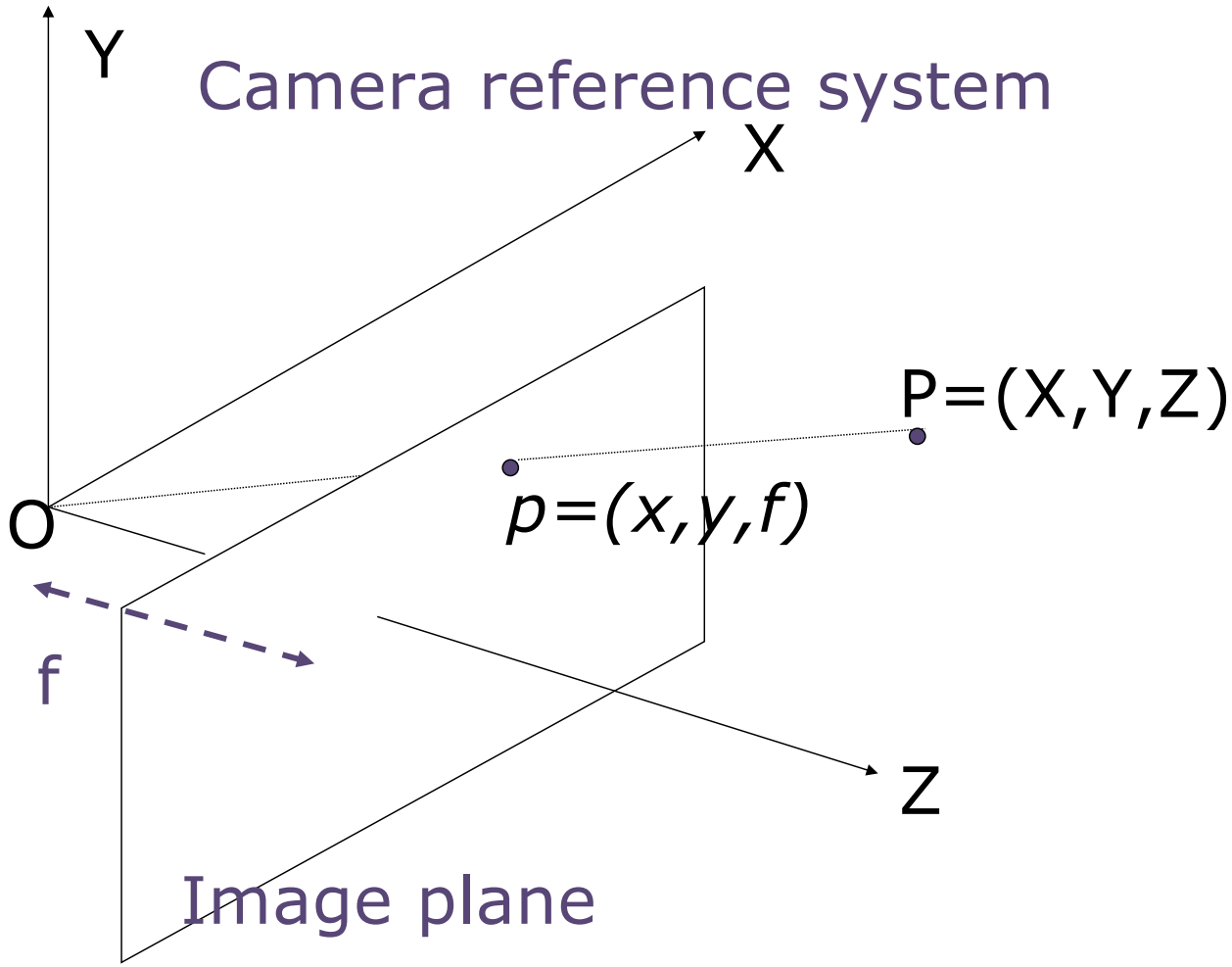
Examples



UniGe | MaLGa

3D reconstruction

Geometry of image formation: perspective or pin-hole model



$$x = f\frac{X}{Z}$$

$$y = f\frac{Y}{Z}$$

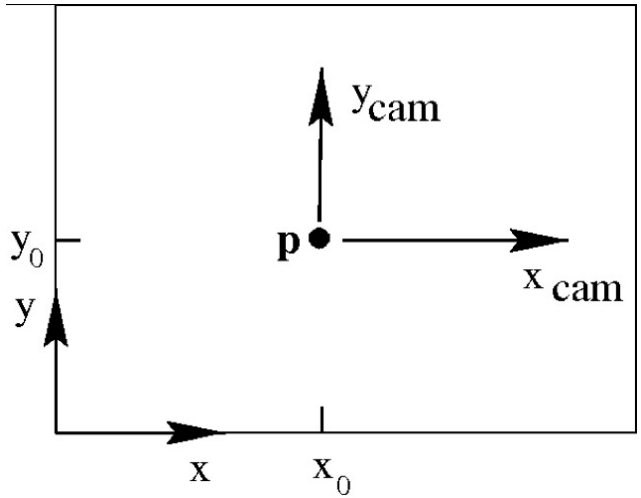
f focal length

Parameters of a stereo system

How to relate points in the world with pixels in the image

Intrinsic/internal parameters

- characterize the mapping of an image point from camera to pixel coordinates in each camera
- They include focal length, the mm-pixel change of coordinates (scaling), the position of the principal point (translation)



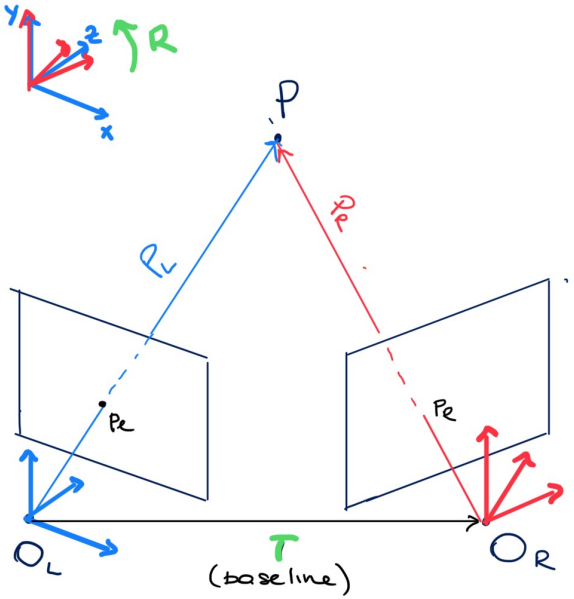
Parameters of a stereo system

How to relate points in the world with pixels in the image

Extrinsic/external parameters

- describe the relative position and orientation of the two cameras (R, **T**)

$$\mathbf{P}_r = R(\mathbf{P}_l - \mathbf{T})$$



System calibration: sketch

Internal parameters can be know a priori or estimated by camera calibration procedures

External parameters can be obtained starting from point correspondences

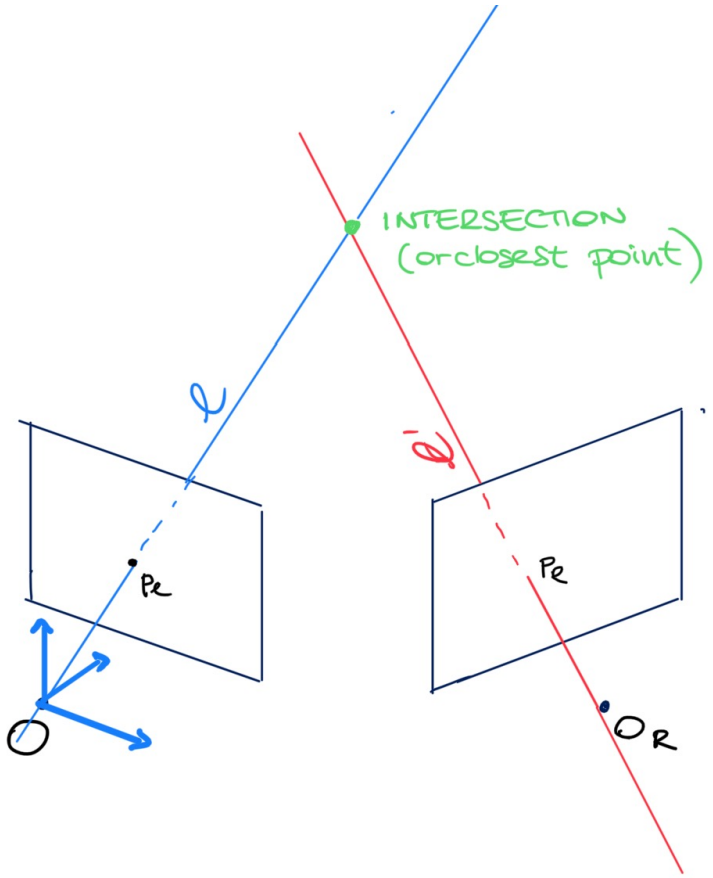
2D to 3D: points triangulation (sketch)

Assuming we know the internal and external parameters

Given a pair of corresponding points in 2D (metric coordinates on a common reference frame) (p_l, p_r)

Estimate the equation of the two projection rays ℓ, ℓ'

Compute the intersection (or the closest point wrt both lines) it will be the 3D reconstruction P



UniGe | MaLGa

quick detour on geometry background

Homogeneous coordinates

Points on a plane

A point on a plane $(x, y) \in \mathbb{R}^2$

Lines

Equation of a line on a plane $ax + by + c = 0$

We may also represent it as $(a, b, c)^\top$

Notice that $(ka, kb, kc)^\top \quad k \neq 0$ it is the same line (they belong to the same **equivalence class**

This equivalence class is called a **homogeneous vector**

Homogeneous coordinates

Projective plane

The projective plane (or 2D projective space) is formed by any vector $(a, b, c)^\top$ representative of an equivalence class:

$$\mathbb{R}^3 - (0, 0, 0)$$

Homogeneous coordinates

A point and a line

A point $(x, y)^\top$ lies on a line $\mathbf{l} = (a, b, c)^\top$ IFF

$$ax + by + c = 0$$

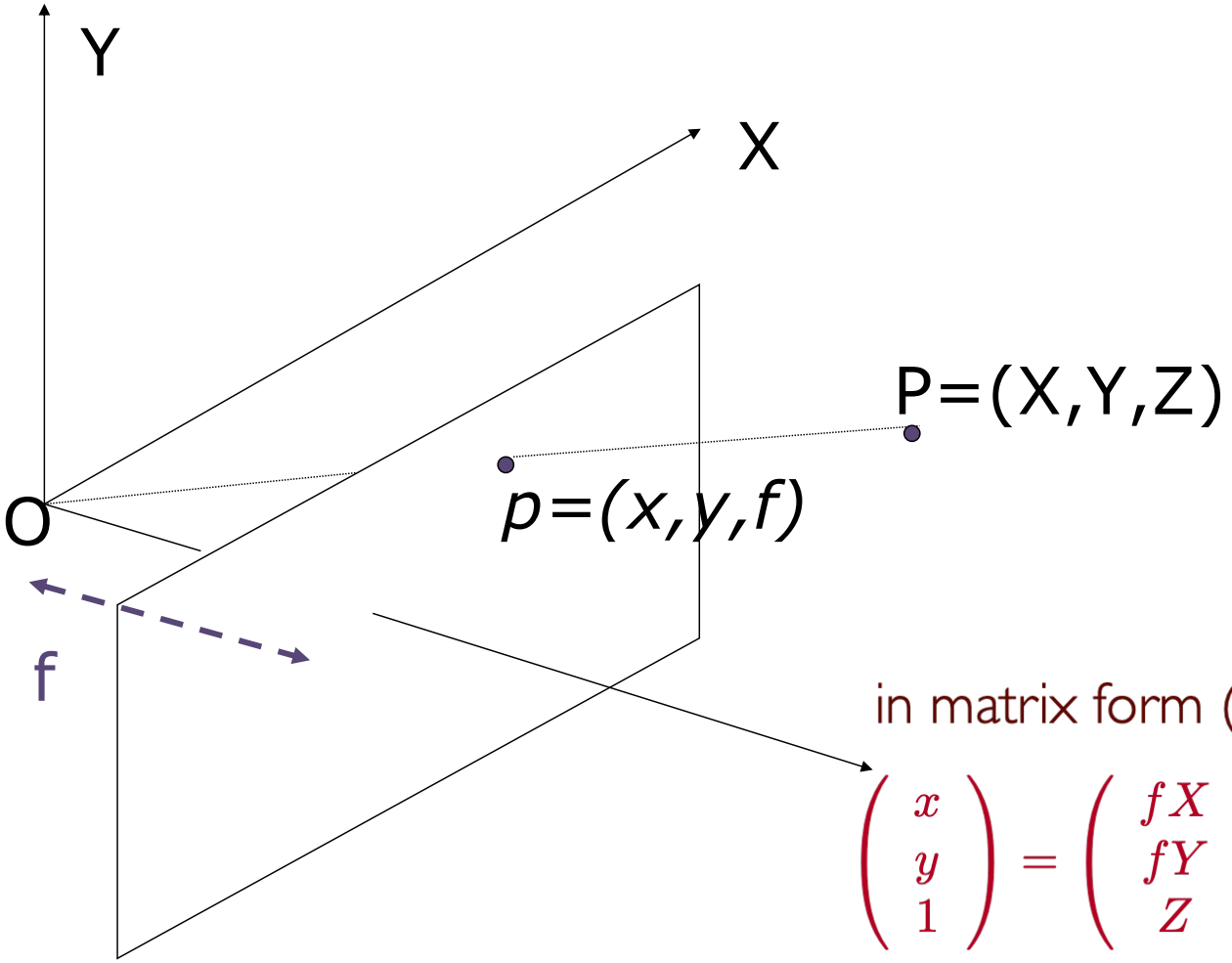
If we define $\mathbf{x} = (x, y, 1)^\top$ then

$$\mathbf{x}^\top \mathbf{l} = 0$$

A point on the Euclidean plane may be represented as a point in the projective plane (as a 3D vector with a final 1).
Notice that here again points define equivalence classes

Perspective model

in homogeneous coordinates becomes a linear transformation



$$x = f\frac{X}{Z}$$

$$y = f\frac{Y}{Z}$$

in matrix form (homogeneous coordinates)

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} fX \\ fY \\ Z \end{pmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$$

UniGe | MaLGa

$$\mathbf{p} = M\mathbf{P} \quad M = \mathrm{diag}(f, f, 1)[I; \mathbf{0}]$$

M is the projection matrix

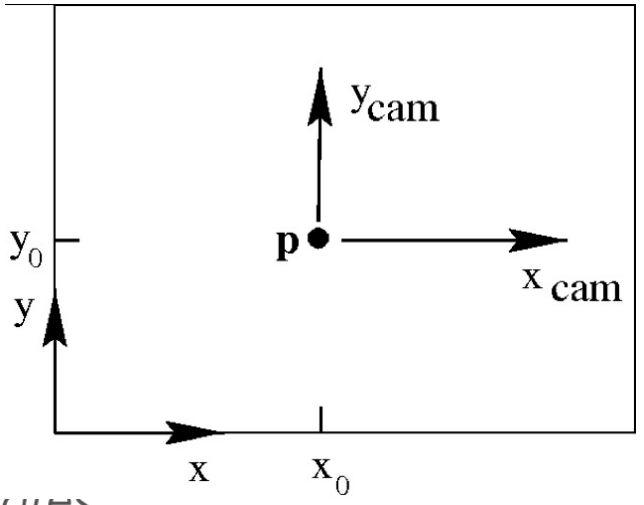
Parameters of a stereo system

How to relate points in the world with pixels in the image

Intrinsic/internal parameters

- characterize the mapping of an image point from camera to pixel coordinates in each camera

- (translation and scaling)



$$K = \begin{bmatrix} \alpha_x & 0 & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}$$

all entries in pixel coordinates

$$\alpha_x = f m_x$$
$$\alpha_y = f m_y$$

number of pixels per unit distance

$$x_0 = m_x p_x$$
$$y_0 = m_y p_y$$

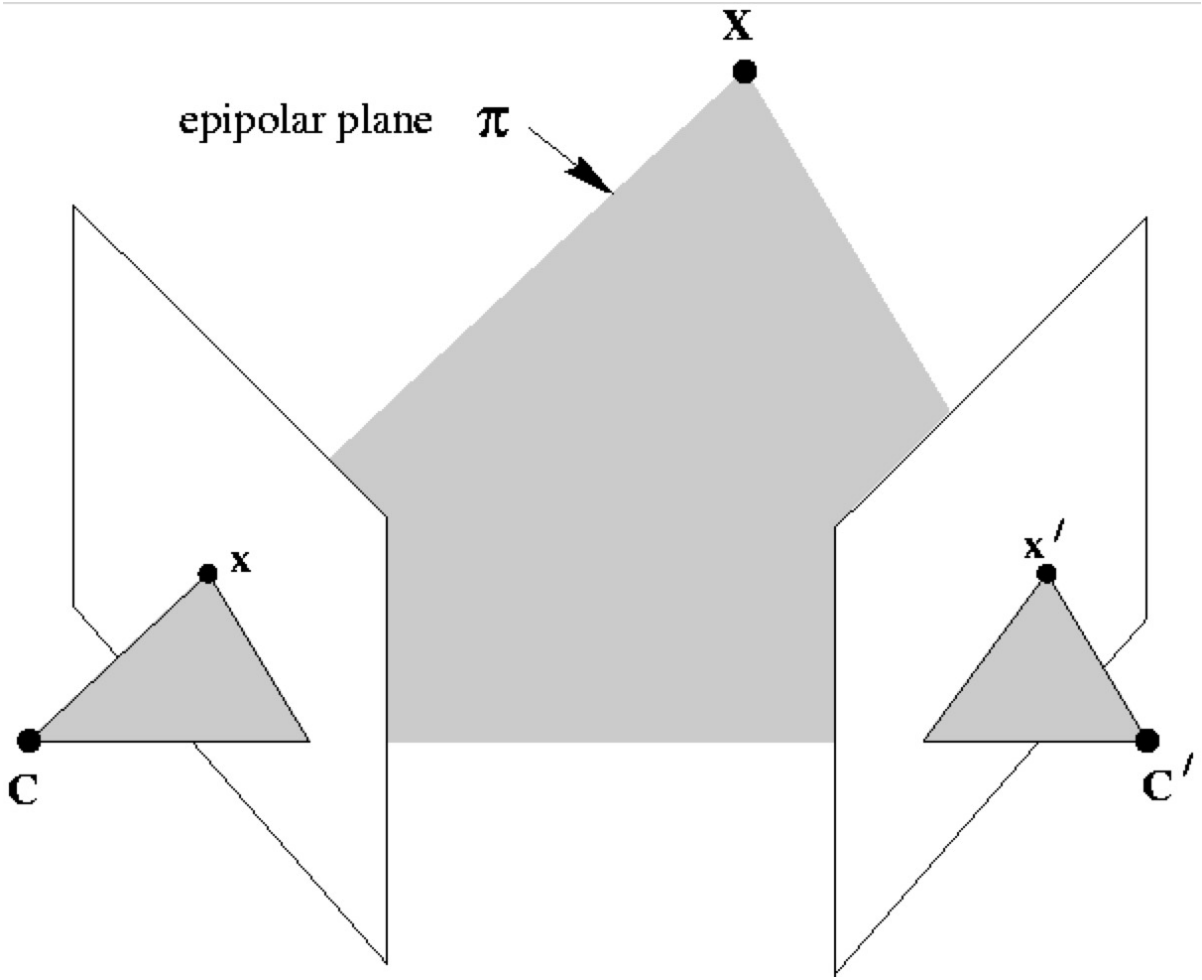
$$\mathbf{p}_{pix} = KM\mathbf{P}$$

UniGe | MaLGa

Epipolar geometry

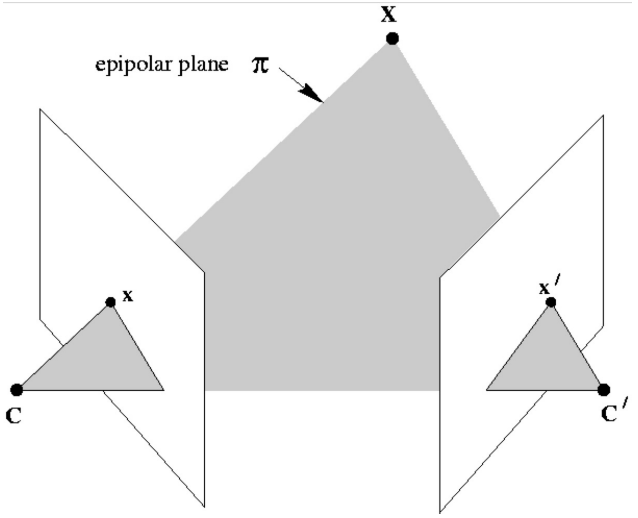
Geometry of a stereo system

Epipolar geometry



Geometry of a stereo system

Epipolar constraints



let us consider ***x*** How is ***x'*** constrained?

x' lies on the line ***l'*** intersection between the epipolar plane and the (right) image plane

l' is the projection on the (right) image plane of the ray passing through **C** and ***x***

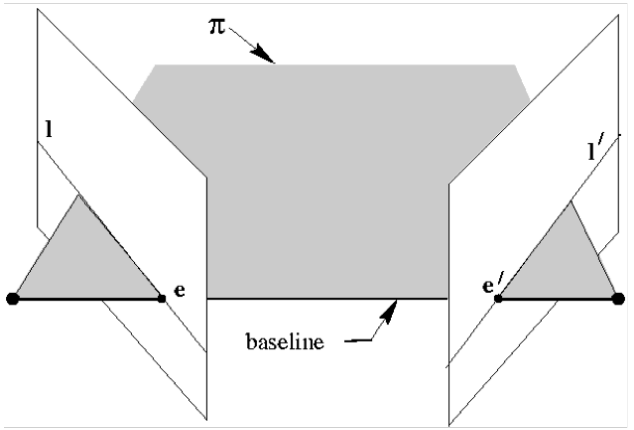
In practice, when we look for the corresponding point ***x'*** we can limit our search to a line, ***l'***

from a 2D to a 1D problem!

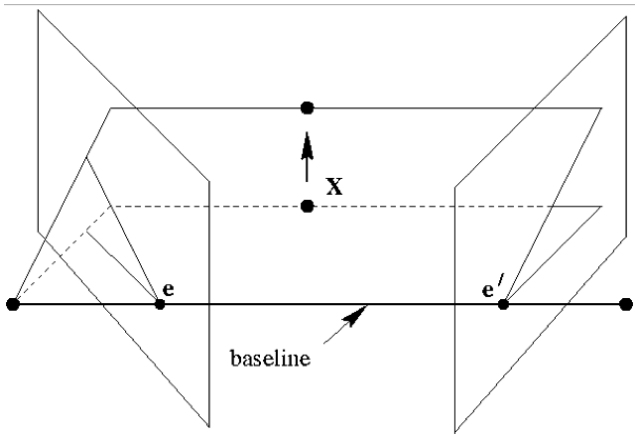
Geometry of a stereo system

Terminology

- ***epipole:*** the intersection of the baseline with the 2 image planes. Also: the projection of one camera centre to the other image plane



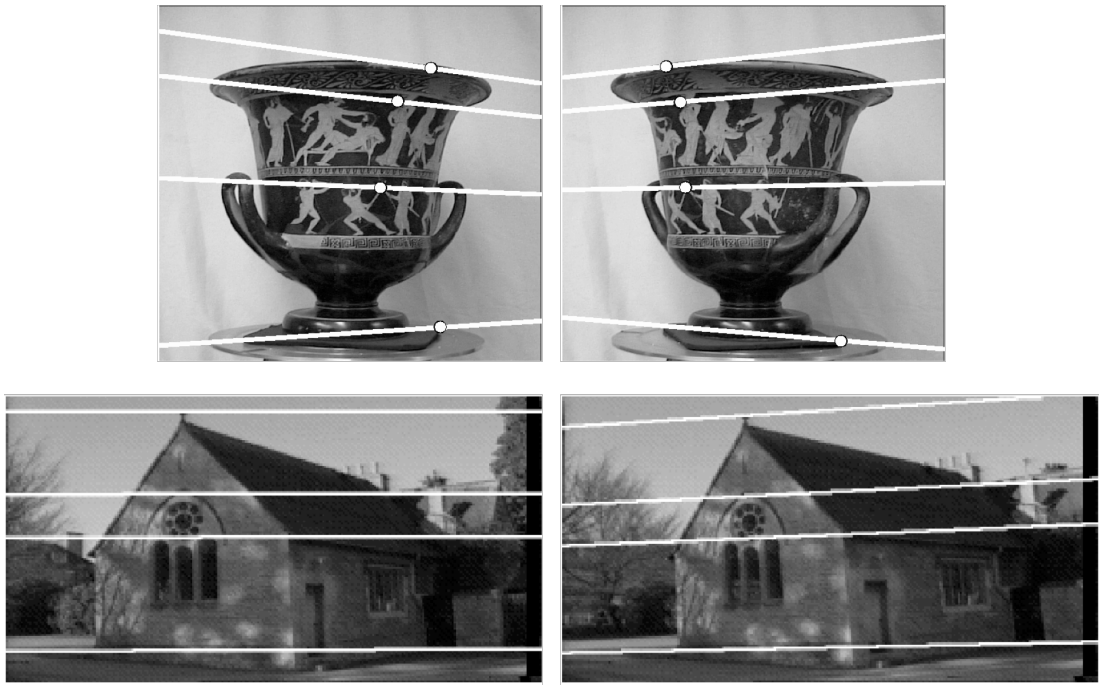
- ***epipolar plane:*** a plane containing the baseline and passing through a 3D point ***X***. It is a *one parameter family (stencil)* of planes



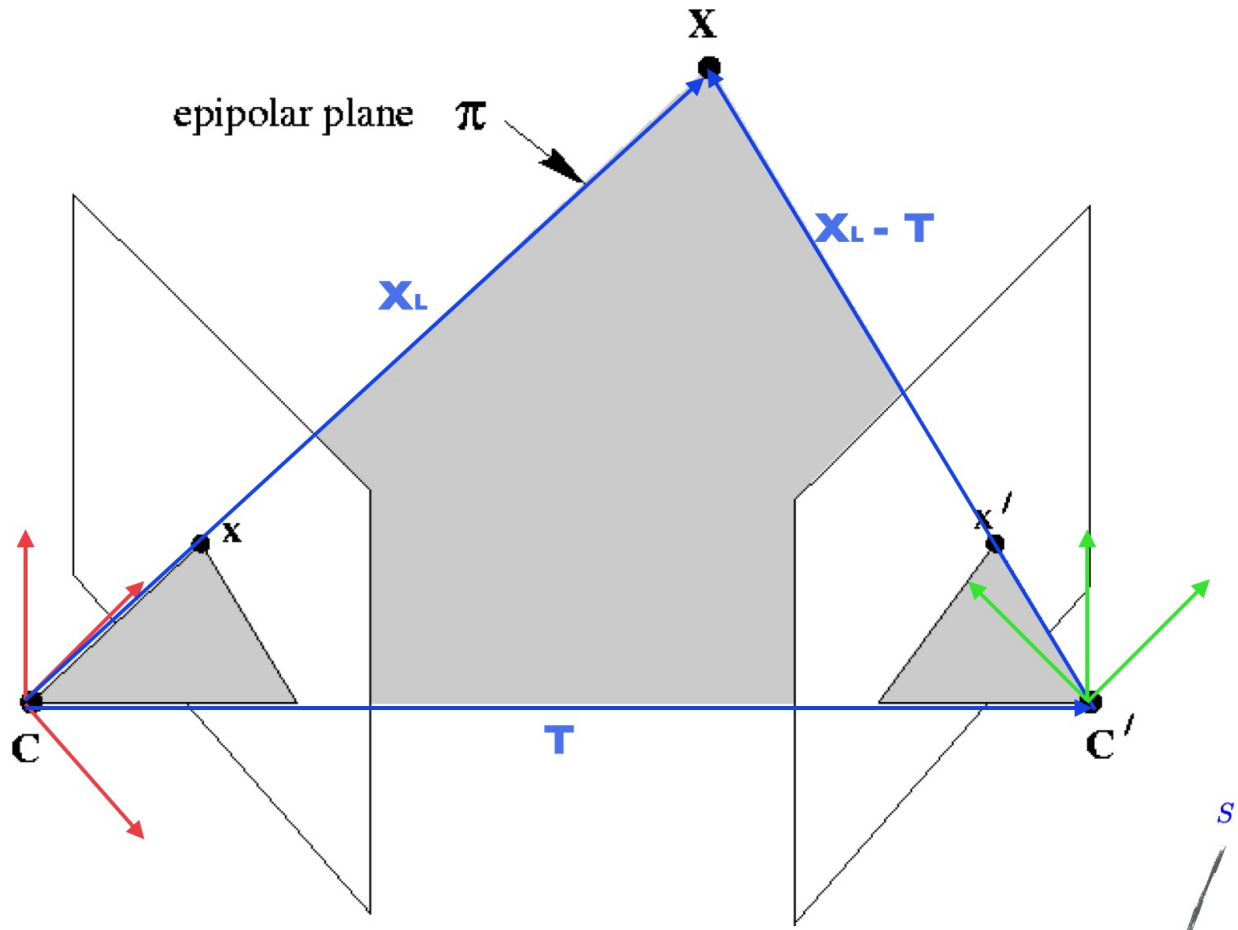
- ***epipolar line:*** the intersection of an epipolar plane with one image plane. All epipolar lines contain the epipole.

Geometry of a stereo system

Epipolar lines



Geometry of a stereo system



$$S = \begin{bmatrix} 0 & -T_3 & T_2 \\ T_3 & 0 & -T_1 \\ -T_2 & T_1 & 0 \end{bmatrix}$$

$\mathbf{X}_l, \mathbf{T}, \mathbf{X}_l - \mathbf{T}$ are coplanar $\Rightarrow$ $(\mathbf{X}_l - \mathbf{T})^\top \mathbf{T} \times \mathbf{X}_l = 0$

$\mathbf{X}_r = R(\mathbf{X}_l - \mathbf{T})$ $\Rightarrow$ $(R^\top \mathbf{X}_r)^\top \mathbf{S} \mathbf{X}_l = 0$

$\mathbf{X}_r^\top R \mathbf{S} \mathbf{X}_l = 0$

Geometry of a stereo system

Essential matrix

•The essential matrix is a 3x3 matrix E=SR $\mathbf{X}_r^\top E \mathbf{X}_l = 0$

•by projecting the point on the two image planes we obtain the equation which is *satisfied by each (x_m x_m') pair of corresponding points in mm coordinates*

$$\mathbf{x_m}'^\top E \mathbf{x_m} = 0$$

- E has rank 2, since S has rank 2 and R is full rank
- E has 5 d.o.f. : 3 from R 3 from **T** but with a scale ambiguity 3+3-1=5

Geometry of a stereo system

Fundamental matrix

we now derive an equivalent equation relating points in **pixel coordinates:**

$$\mathbf{x}_m = K_l^{-1}\mathbf{x}$$

$$\mathbf{x}'_m = K_r^{-1}\mathbf{x}'$$

- Then

$$\mathbf{x}'^{\top} K_r^{-T} E K_l^{-1} \mathbf{x} = 0$$

F

- The fundamental matrix satisfies the equation

$$\mathbf{x}'^{\top} F \mathbf{x} = 0$$

for each pair of (x, x') of corresponding points in pixel coordinates

Geometry of a stereo system

Fundamental matrix

Suppose we have two images acquired by cameras with no coinciding centres.

$$\mathbf{x}'^{\top} F \mathbf{x} = 0$$

Then the fundamental matrix F is the unique 3x3 rank 2 matrix so that, for each corresponding pair (x,x')

Fundamental (and essential) matrix is a map between points and (epipolar) lines

$$\mathbf{l}' = F\mathbf{x}$$

Geometry of a stereo system

8 points algorithm (sketch)

$$\mathbf{x} = (x, y, 1)^\top \quad \mathbf{x}' = (x', y', 1)^\top$$

we may obtain a linear algorithm from 8 points correspondences (notice this is not the minimal problem)

$$\begin{bmatrix} x_1'x_1 & x_1'y_1 & x_1' & y_1'x_1 & y_1'y_1 & y_1' & x_1 & y_1 & 1 \\ x_2'x_2 & x_2'y_2 & x_2' & y_2'x_2 & y_2'y_2 & y_2' & x_2 & y_2 & 1 \\ & . & . & . & . & . & . & . & \\ x_n'x_n & x_n'y_n & x_n' & y_n'x_n & y_n'y_n & y_n' & x_n & y_n & 1 \end{bmatrix} \begin{pmatrix} f_{11} \\ f_{12} \\ . \\ . \\ . \\ f_{33} \end{pmatrix} = \mathbf{0}$$

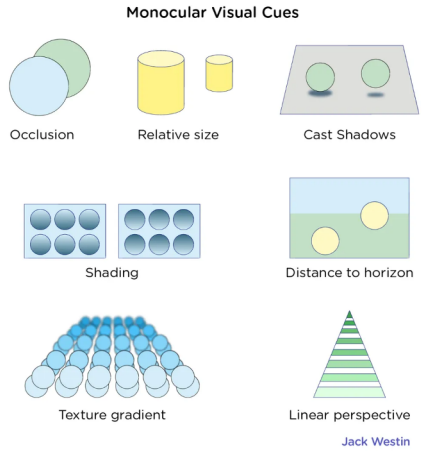
It is a homogeneous system Af=0, overdetermined, which may be solved with SVD

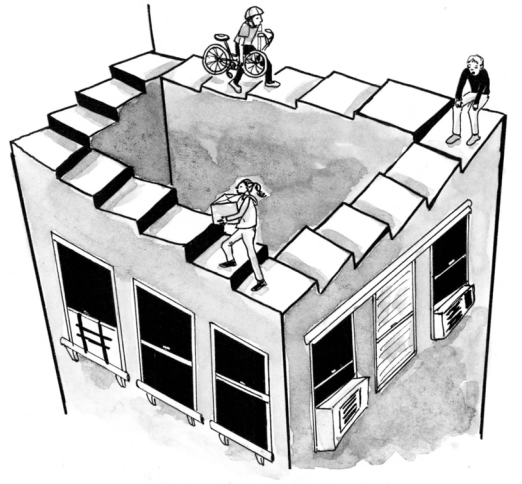
Deep learning and 3D vision

It is not all about geometry

- Stereopsis in humans (given the narrow baseline) works primarily in short range (up to 10 meters)

- Beyond this range, the brain uses other cues
- Humans can still infer depth from single images
- This provides a biological justification to the monocular deep learning approaches that have been proposed in the last years





Monocular Vision and Deep Learning

- Many recent works uses deep

  ConvNets to learn monocular depth

  directly from RGB images
- Not only depth

- We do not have access to very large general purpose datasets for 3D tasks :
  - Expensive to acquire and annotate images
  - Using the limited amount of available annotations results in poor generalization (despite performing well on benchmark data)
  - Research is focusing on unsupervised or self-supervised approaches

UniGe