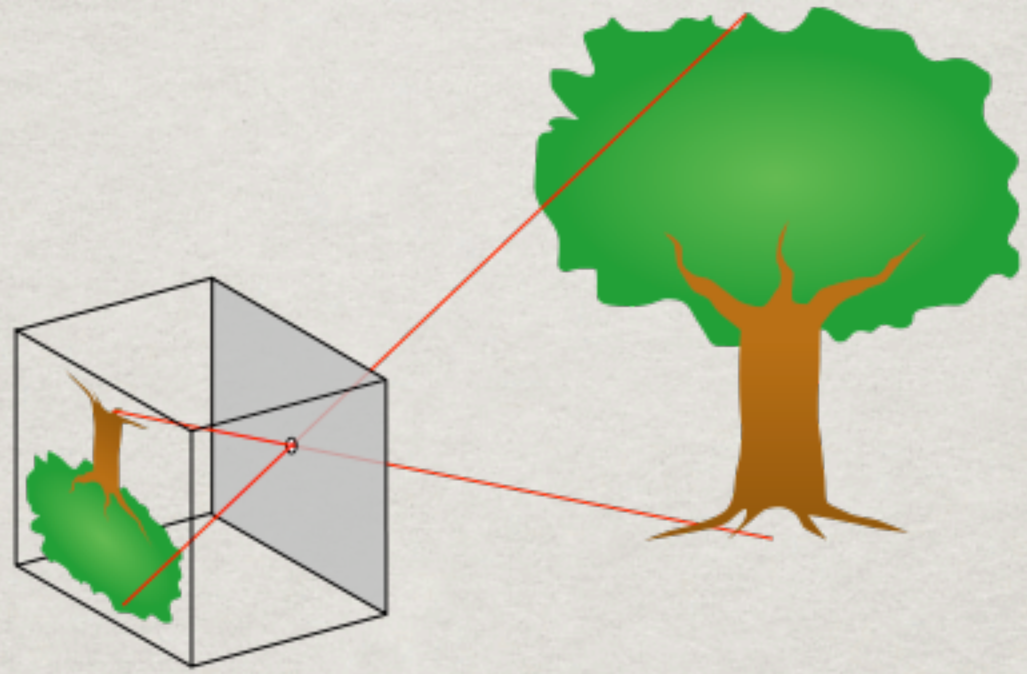
Visual Object Recognition

state-of-the-art techniques and performance evaluation

Lecture 2: Image Formation

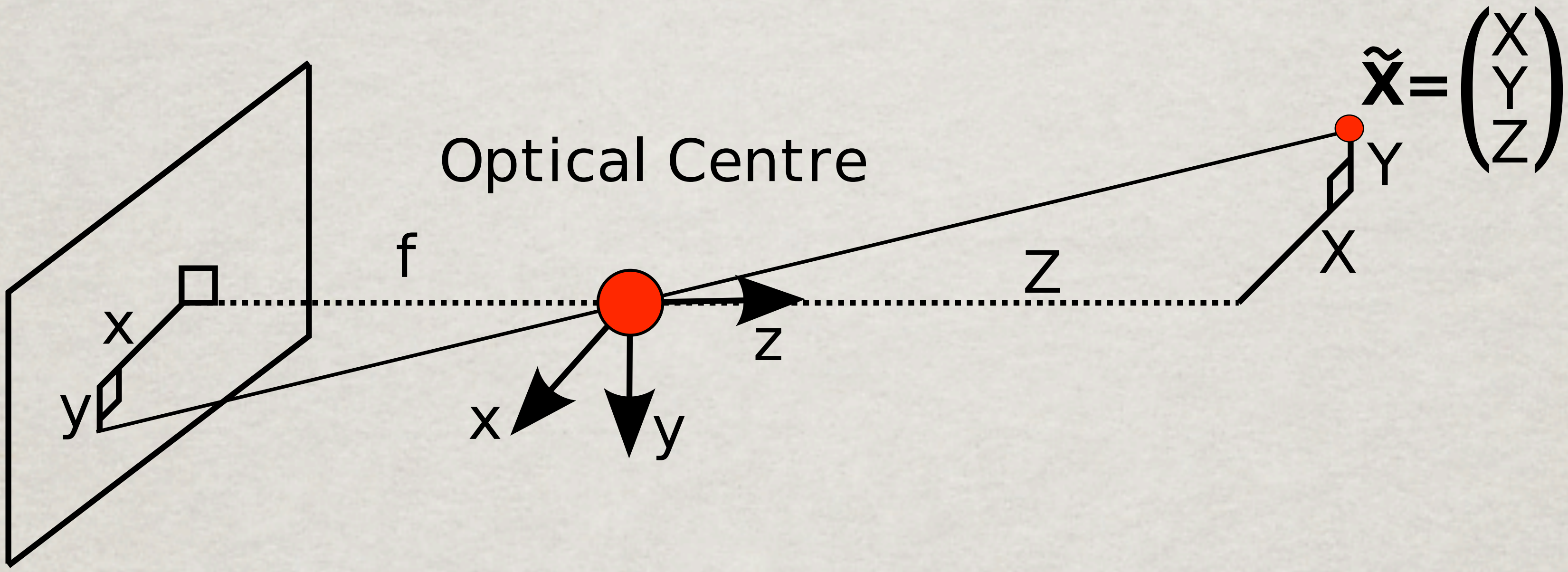
- Pin-hole, and thin lens cameras
- Illumination
- Homographies
- Epipolar Geometry
- Canonical Frames

The Pin-Hole Camera



- A brightly illuminated scene will be projected onto a wall opposite of the pin-hole.
- The image is rotated 180°.

The Pin-Hole Camera



- From similar triangles we get:

$$x = f\frac{X}{Z} \qquad y = f\frac{Y}{Z}$$

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

The Pin-Hole Camera

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

- More generally, we write:

$$\gamma \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & s & c_x \\ 0 & fa & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

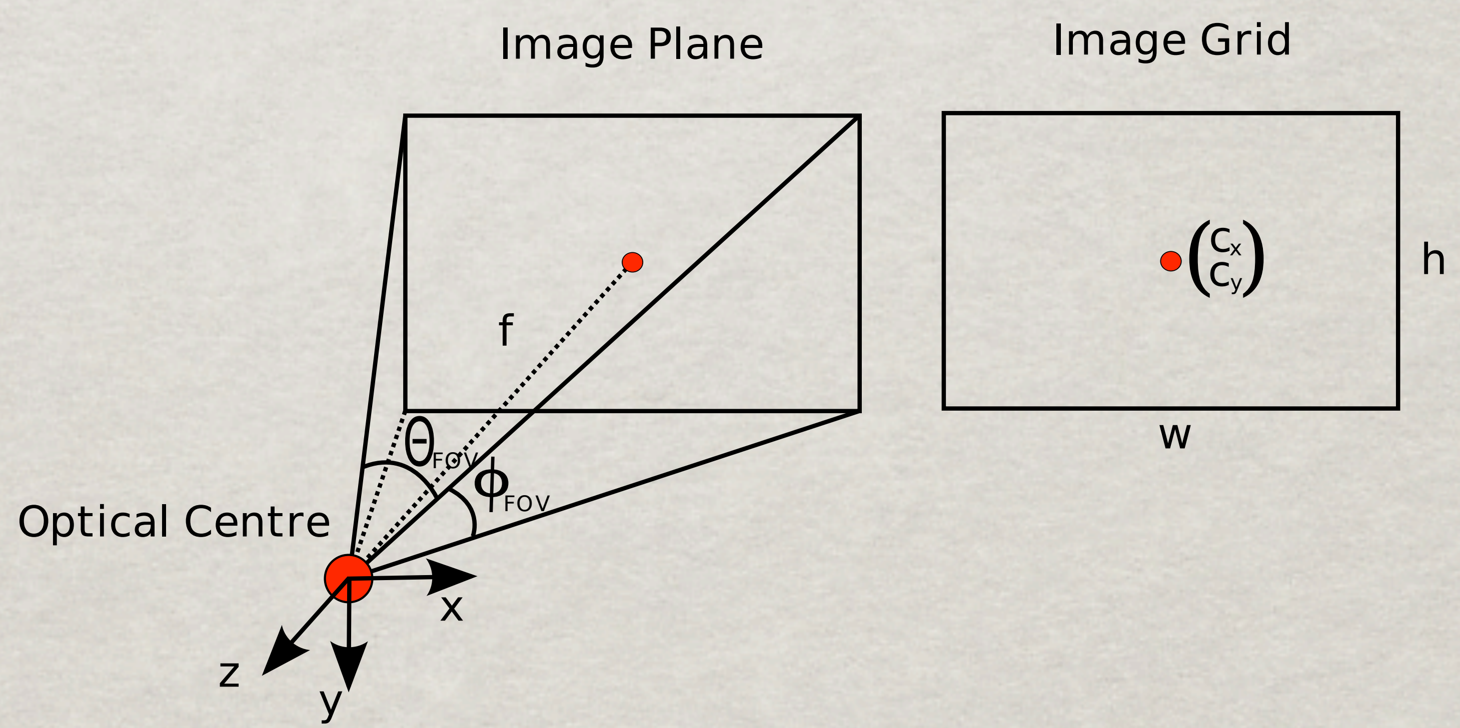
f-focal length, s-skew, a-aspect ratio,
c-projection of optical centre

The Pin-Hole Camera

$$\underbrace{\gamma \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} f & s & c_x \\ 0 & fa & c_y \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{K}} \underbrace{\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}}_{\tilde{\mathbf{X}}}$$

$$\mathbf{x} \sim \mathbf{K}\tilde{\mathbf{X}}$$

- Motivation:



f-focal length, s-skew, a-aspect ratio,
c-projection of optical centre

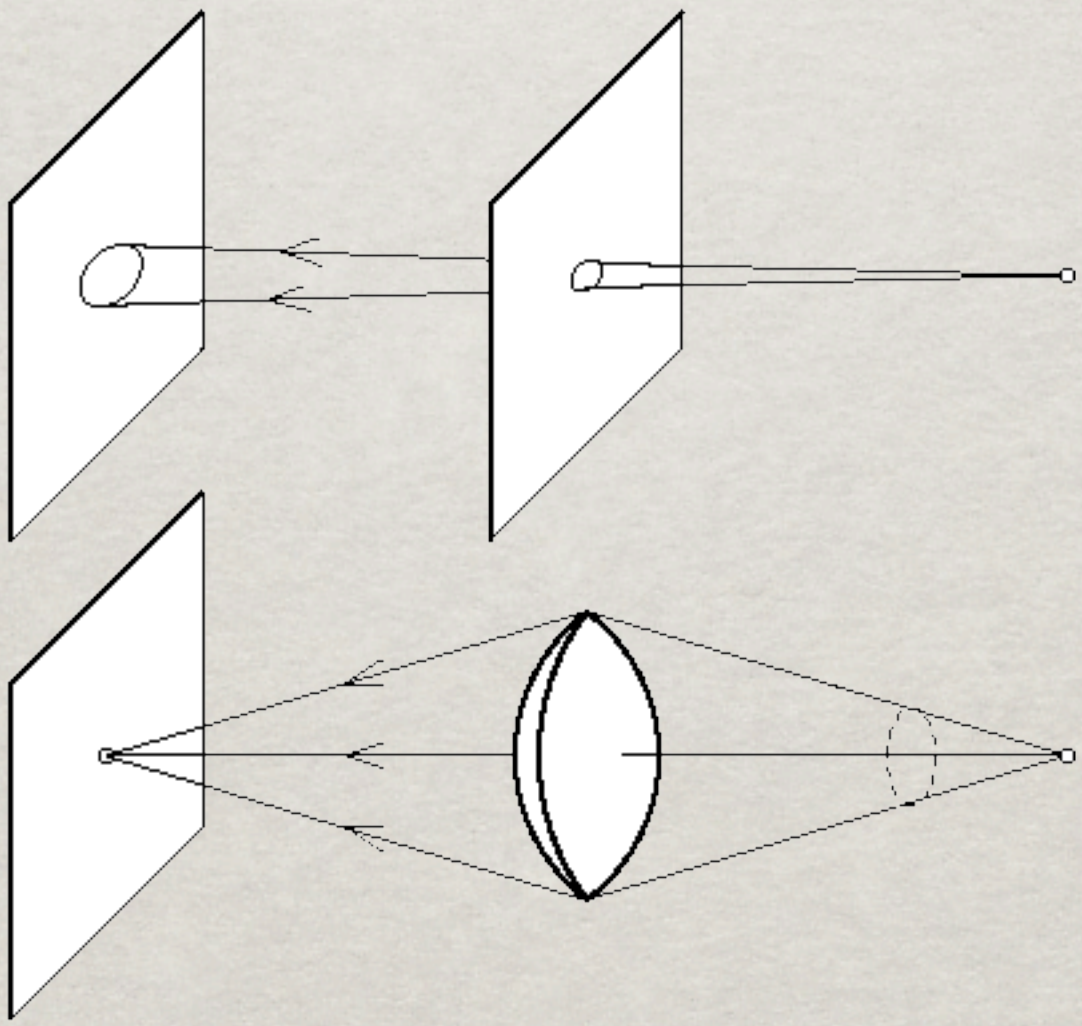
The Pin-Hole Camera

- For a general position of the world coordinate system (WCS) we have:

$$\mathbf{x} \sim \mathbf{K} \underbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix}}_{[\mathbf{R}|\mathbf{t}]} \underbrace{\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}}_{\mathbf{X}}$$

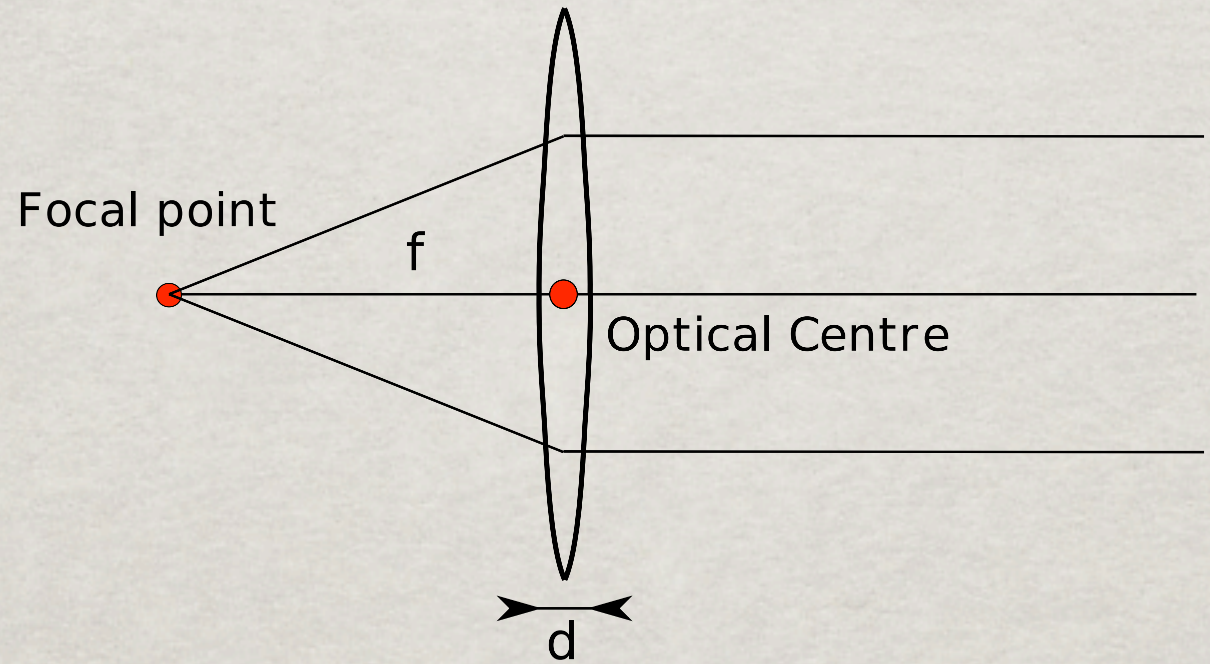
Thin Lens Camera

- But we use lenses, not pin-hole cameras!



Thin Lens Camera

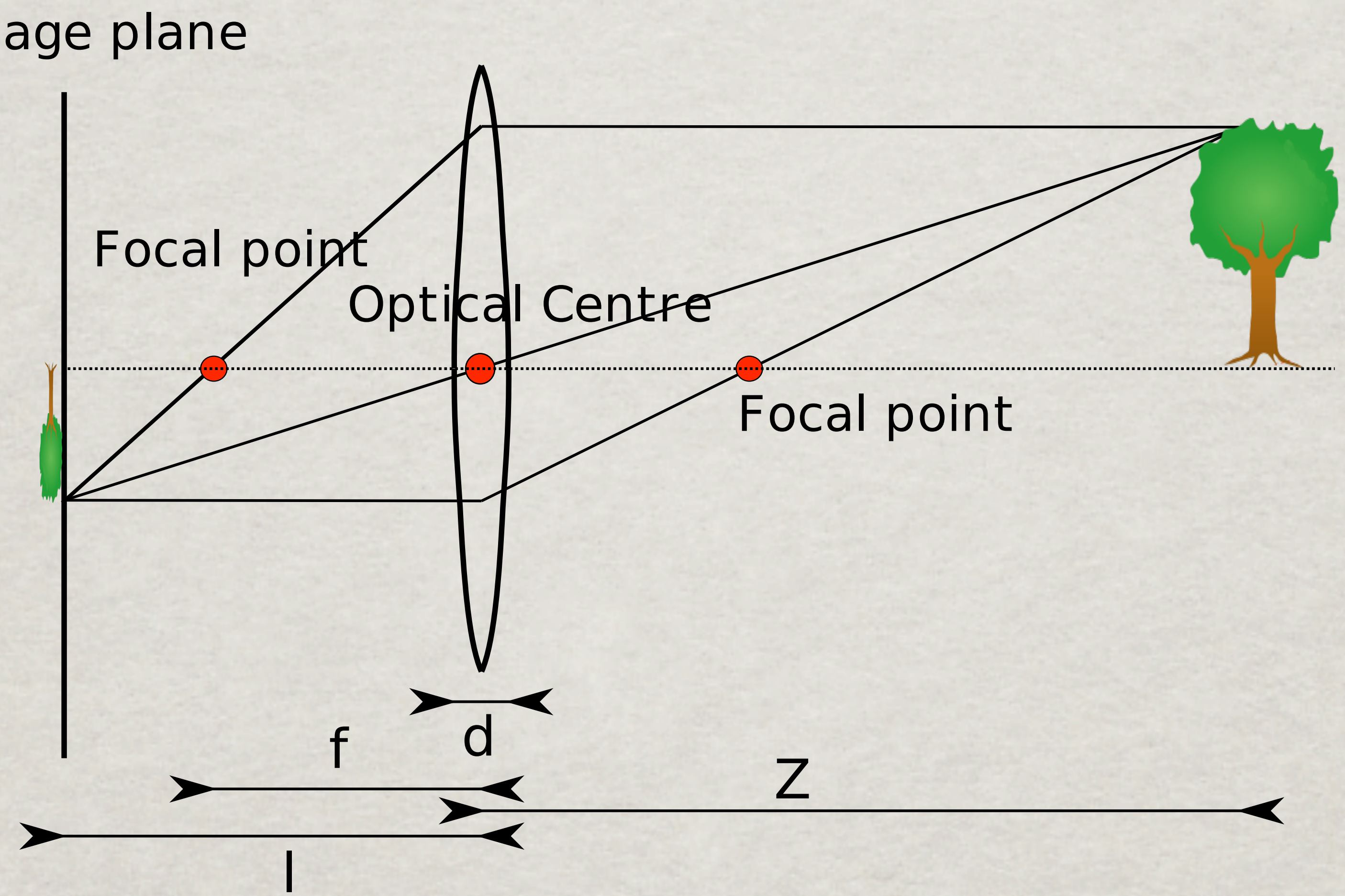
- A thin lens is a (positive) lens with $d << f$



- Parallel rays converge at the focal points

- Rays through the optical centre are not refracted

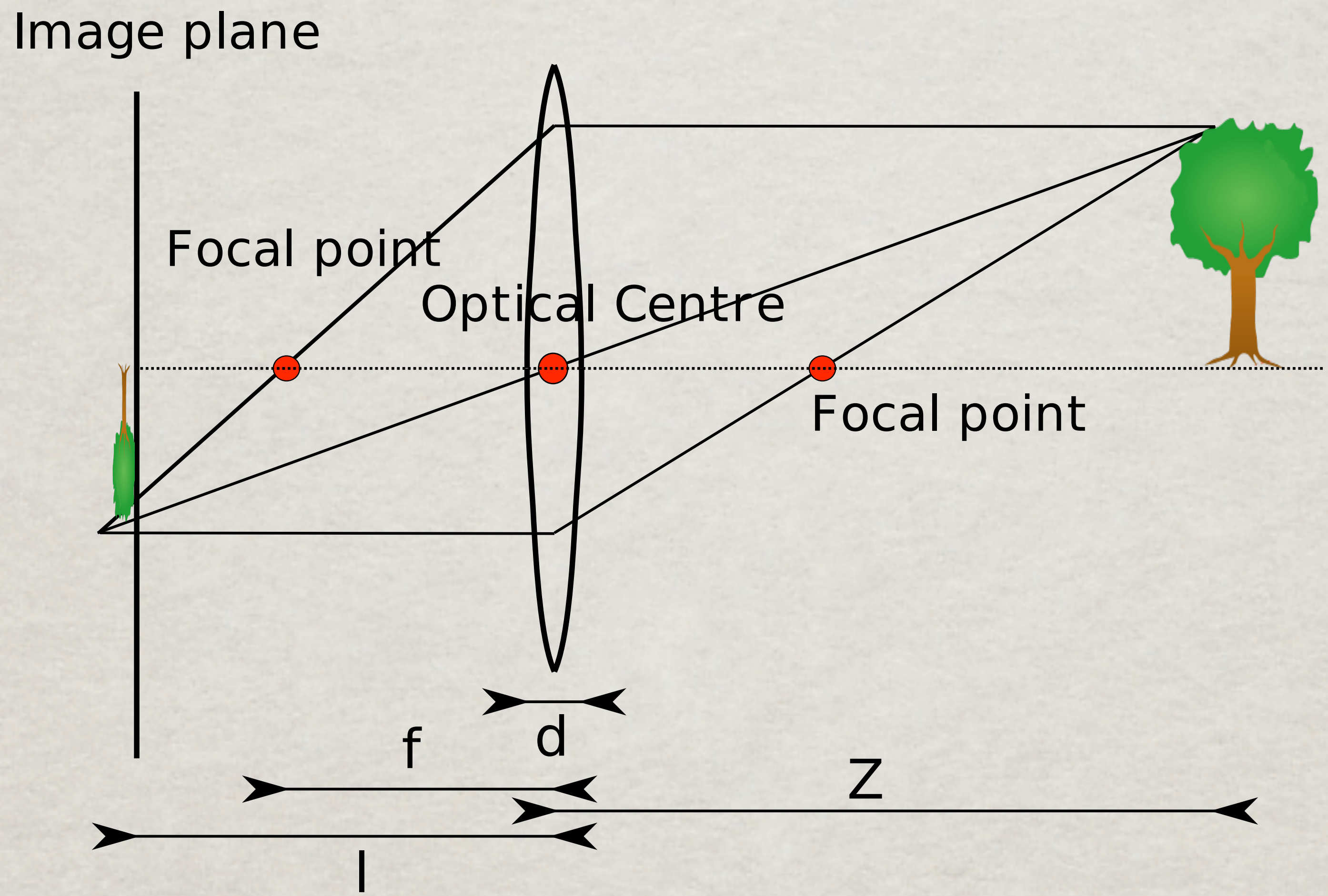
Thin Lens Camera



- Thin lens relation (from similar triangles):

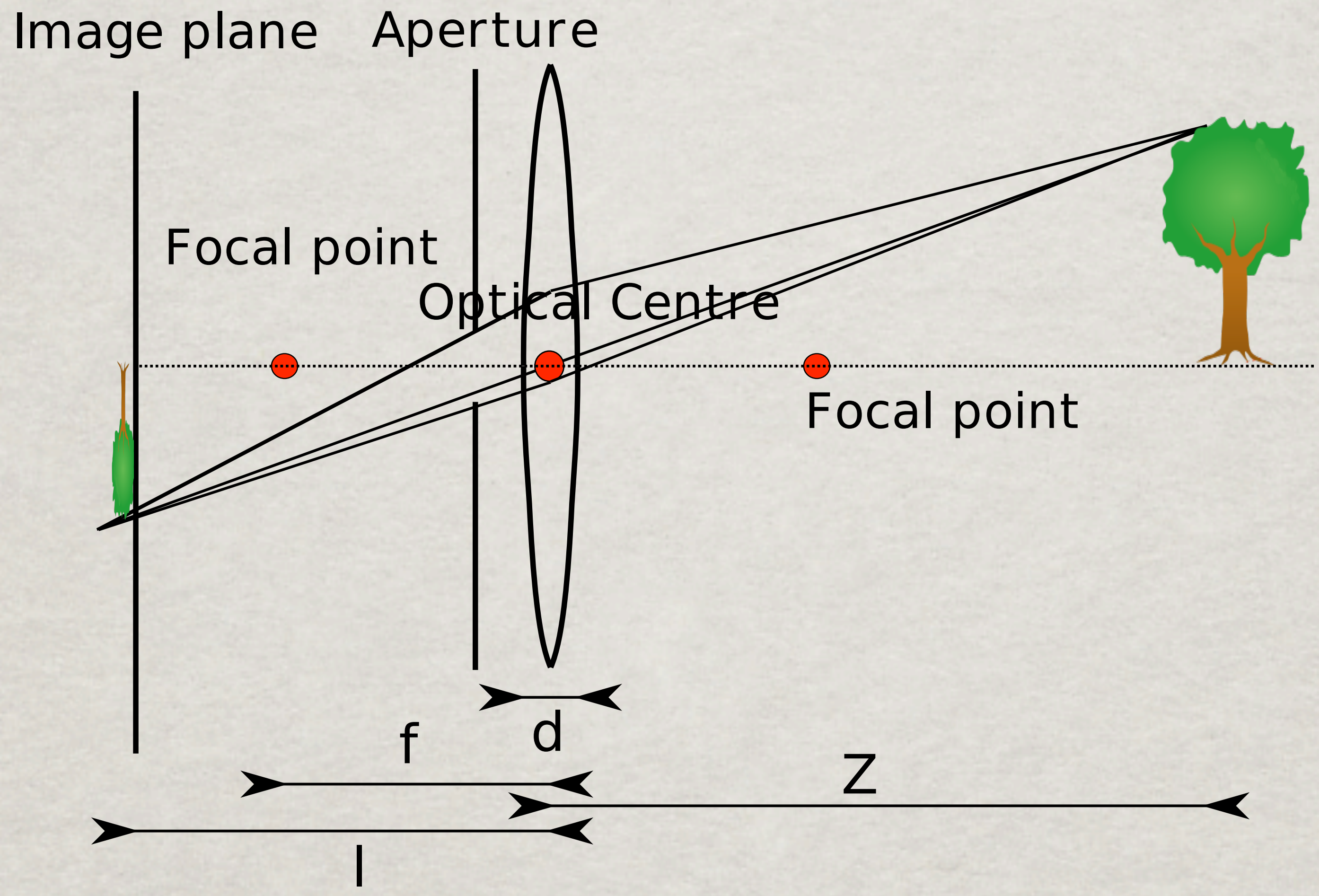
$$\frac{1}{f} = \frac{1}{Z} + \frac{1}{l}$$

Thin Lens Camera



- Focus at one depth only.
- Objects at other depths are blurred.

Thin Lens Camera



- Adding an aperture increases the *depth-of-field*, the range which is sharp in the image.
- A compromise between pinhole and thin lens.

Thin Lens Effects

Correct

Barrel distortion

Pin-cushion distortion

- Radial distortion
- For zoom lenses: Barrel at wide FoV pin-cushion at narrow FoV

Thin Lens Effects

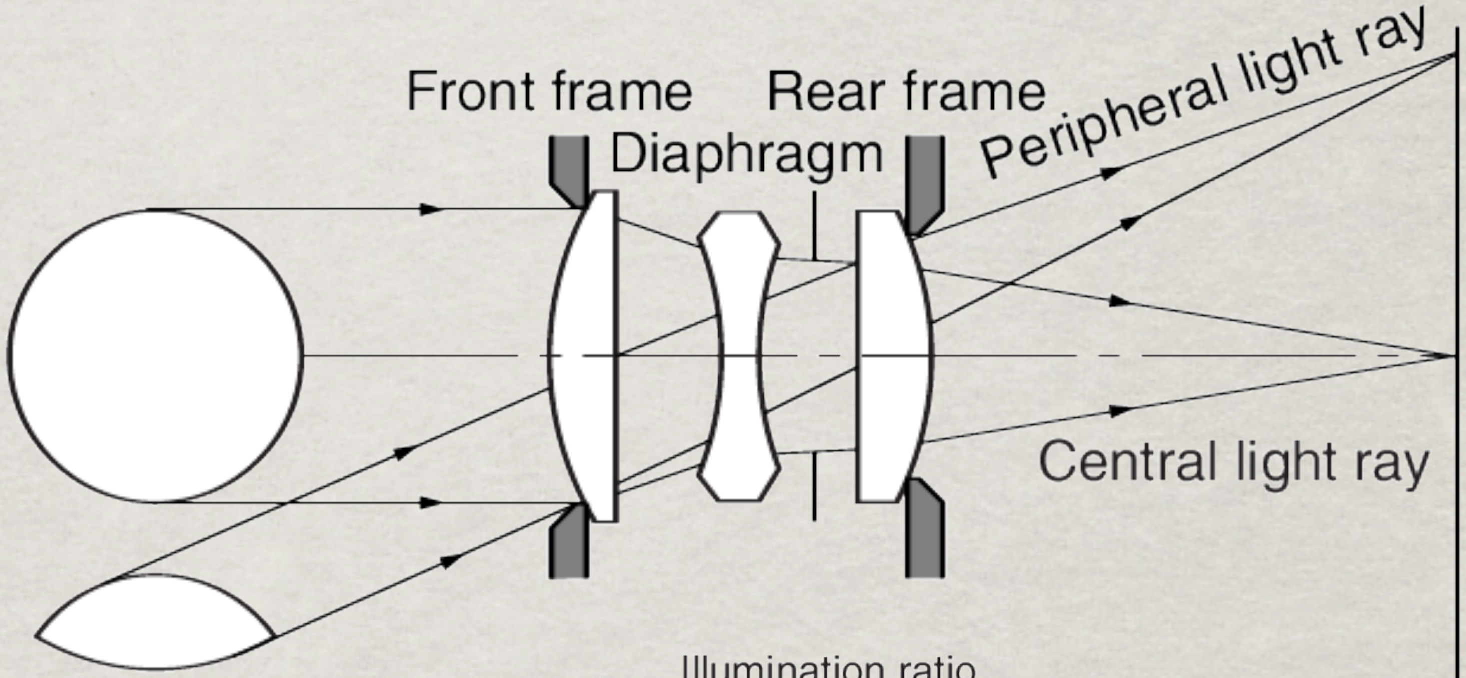
Correct

Darkened periphery

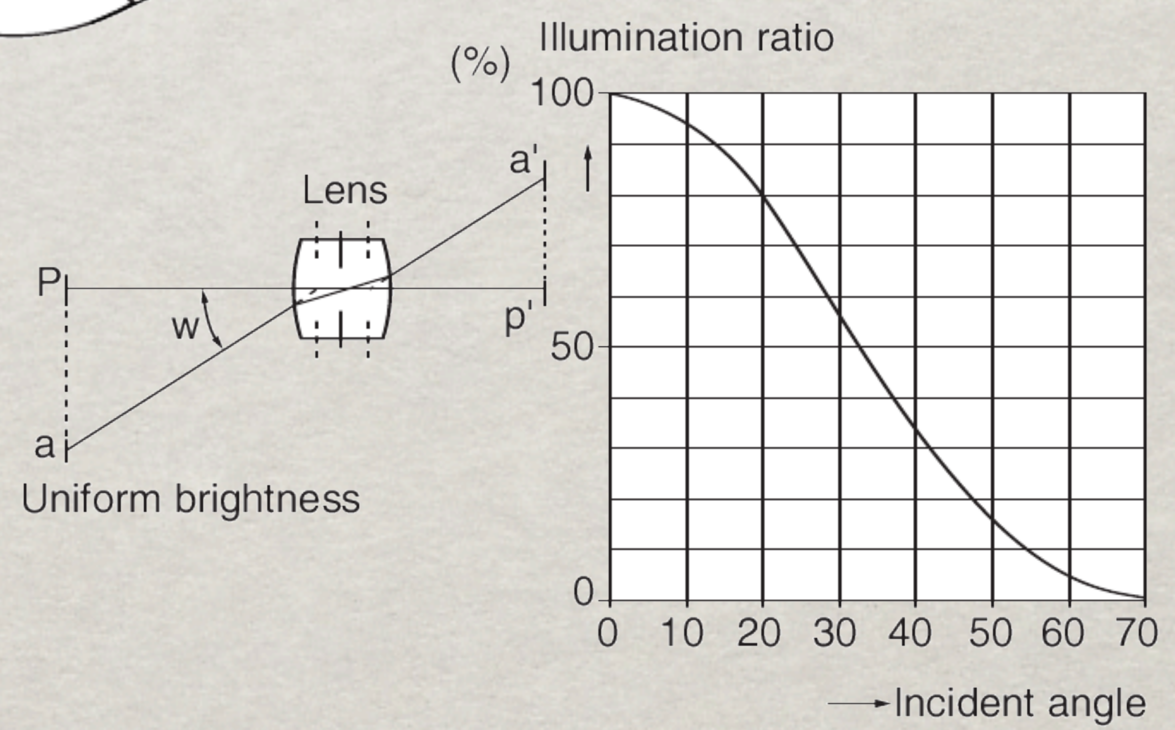
- Vignetting and $\cos^4$-law

Thin Lens Effects

- **Vignetting**



- **cos^4-law** dampening with cos^4(w)



http://software.canon-europe.com/files/documents/EF_Lens_Work_Book_10_EN.pdf

Illumination

- Image intensity is linear in radiance (at least before gamma correction)
- E.g. adding a second, identical light source will double the sensor activation.

$$a(\mathbf{x}) = \int s(\lambda) r(\lambda, \mathbf{x}) e(\lambda) d\lambda$$

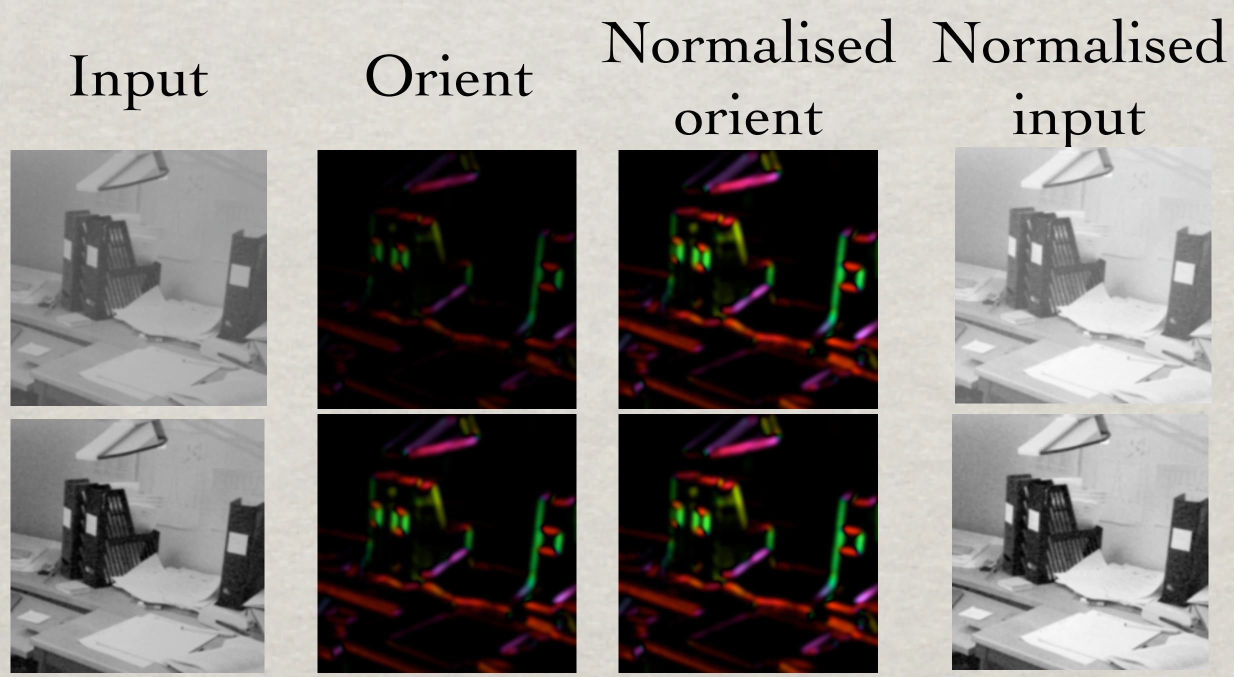
- s-sensor absorption spectrum, r-reflectance spectrum of object, e-emission spectrum of light source (attenuated by the atmosphere)

Illumination

- Mean subtraction, derivatives, and other DC free linear filters remove a *constant offset* in intensity

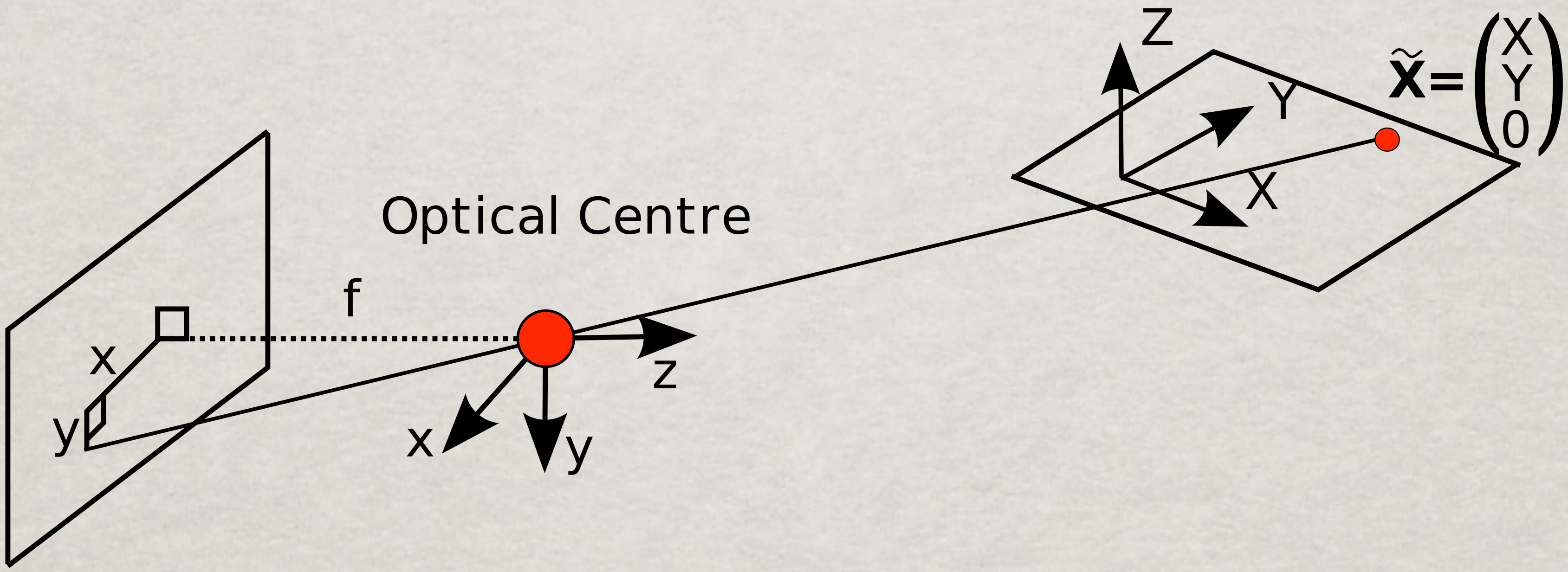
- Normalising a patch by e.g. the l^2-norm removes *scalings* of the intensity.

- Affine invariance by combining both.



Homographies

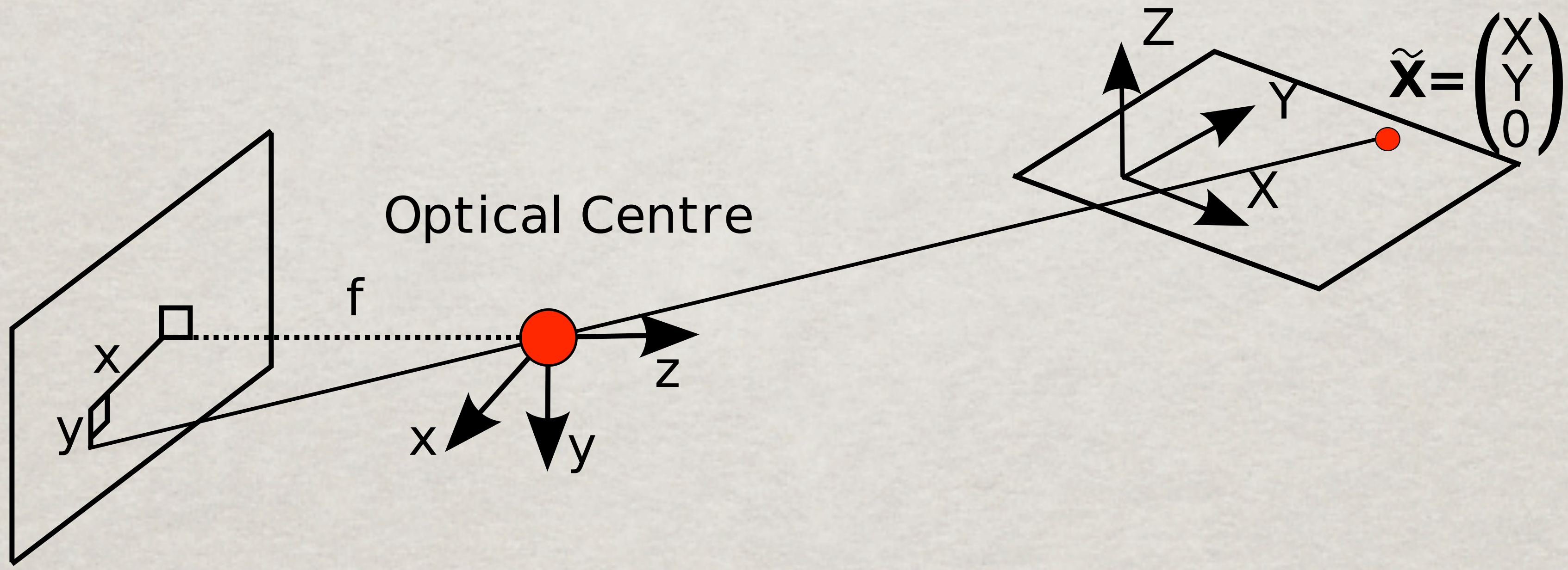
- For a planar object, we can imagine a world coordinate system fixed to the object



$$\gamma \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

Homographies

- For a planar object, we can imagine a world coordinate system fixed to the object



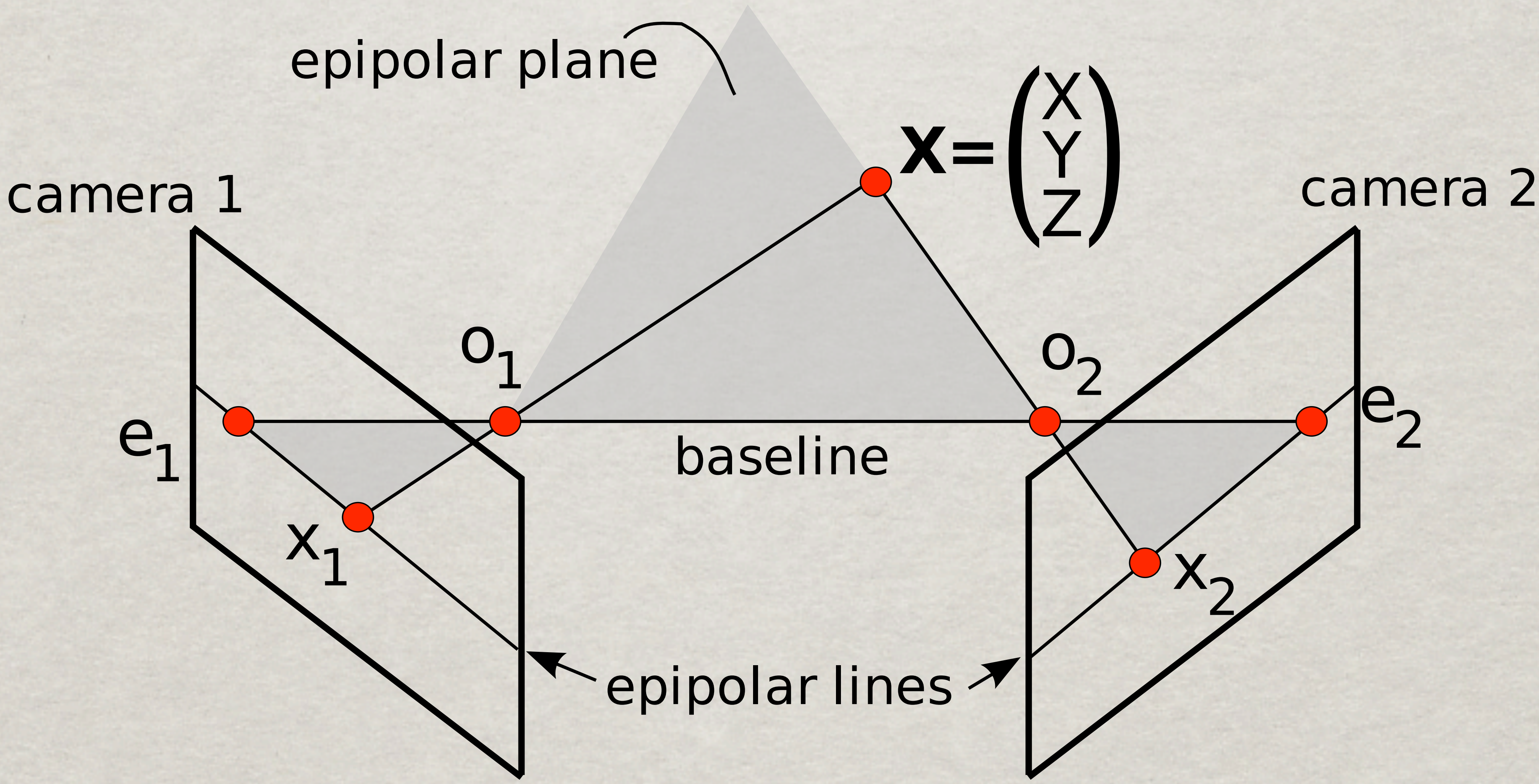
$$\gamma \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} r_{11} & r_{12} & t_1 \\ r_{21} & r_{22} & t_2 \\ r_{31} & r_{32} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}}_{\mathbf{H}} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$$

Homographies

- In general, we can use homographies to describe the transformation between any two planes in 3D.
- Since the matrix **H** is only unique up to scale, it has only 8 degrees of freedom.
- It can be estimated from 4 or more corresponding points on the two planes.
- See e.g. R. Hartley and A. Zisserman, *Multiple View Geometry for Computer Vision*

Epipolar Geometry

- The geometry of two cameras:



- e_1, e_2 are called epipoles. o_1,o_2 are the optical centres.

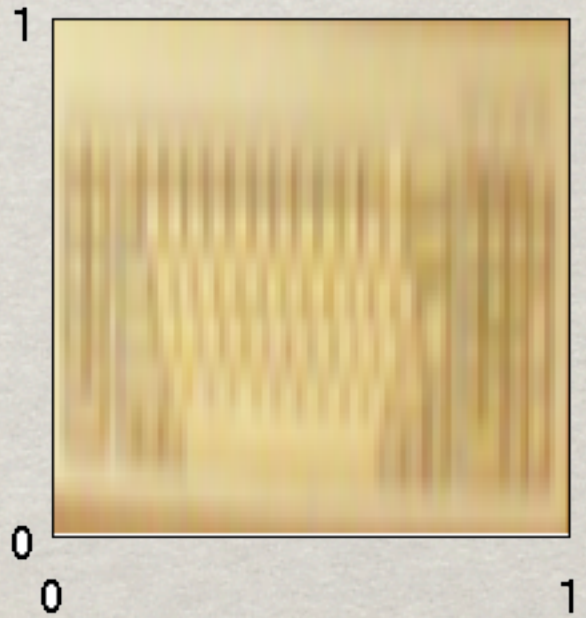
Epipolar Geometry

- So in general, two view geometry only tells us that a corresponding point lies somewhere along a line.
- In practice, we often know more, as objects often have planar, or near planar surfaces. i.e., we are close to the homography case.
- Also: If the views have a **short relative baseline**, we can use even more simple models.

Canonical Frames

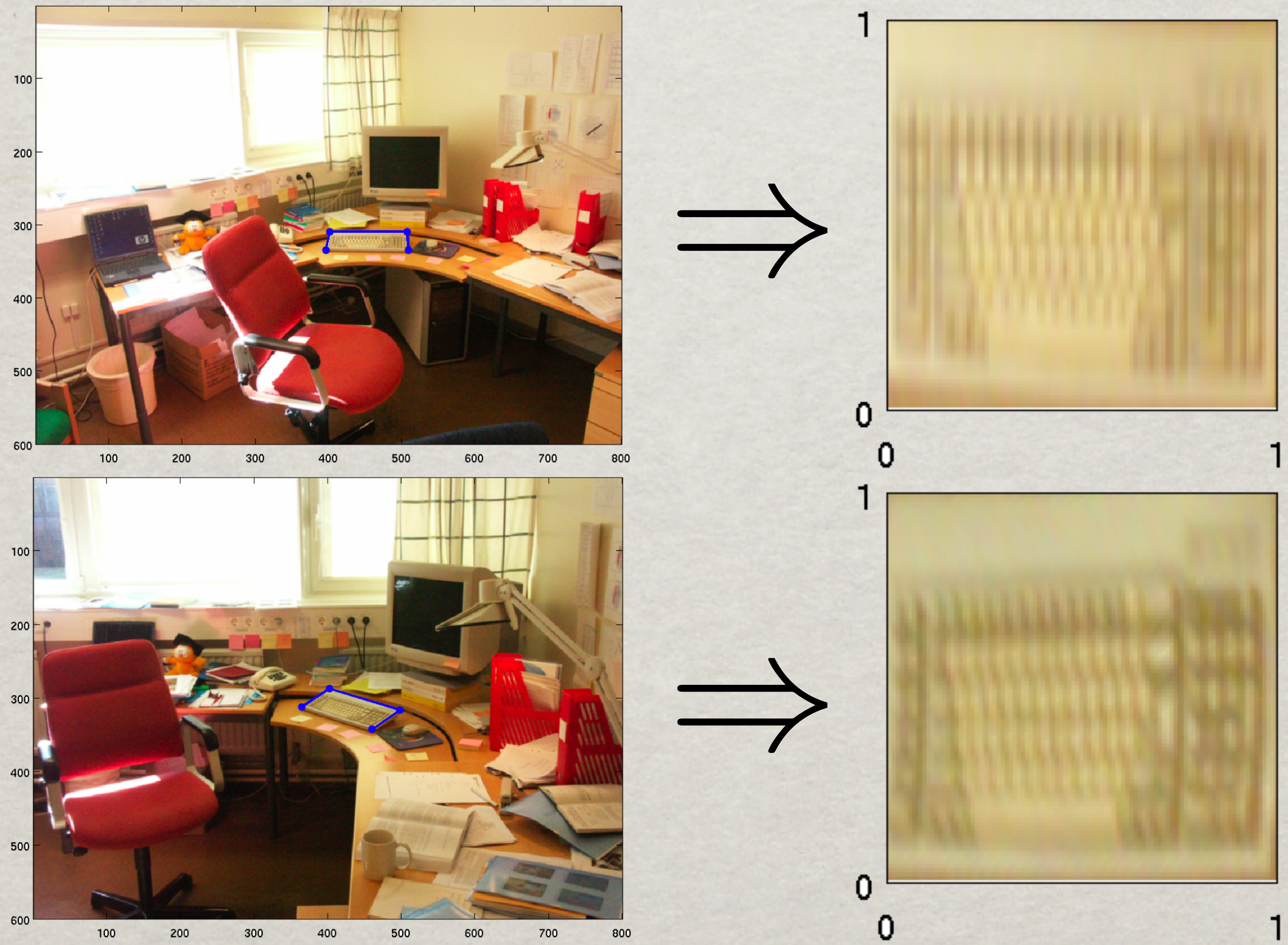
- Aka. covariant frames, and invariant frames.
- Resample patches to canonical frame.
- Points from e.g. Harris detector, or SIFT.

⇒



Canonical Frames

- After resampling matching is much easier.



Canonical Frames

- A hierarchy of transformations:

$$\begin{bmatrix} s & 0 & t_1 \\ 0 & s & t_2 \\ 0 & 0 & 1 \end{bmatrix}$$

- scale+translation

$$\begin{bmatrix} s_1 & s_2 & t_1 \\ -s_2 & s_1 & t_2 \\ 0 & 0 & 1 \end{bmatrix}$$

- similarity (scale+translation+rotation)

$$\begin{bmatrix} a_{11} & a_{12} & t_1 \\ a_{21} & a_{22} & t_2 \\ 0 & 0 & 1 \end{bmatrix}$$

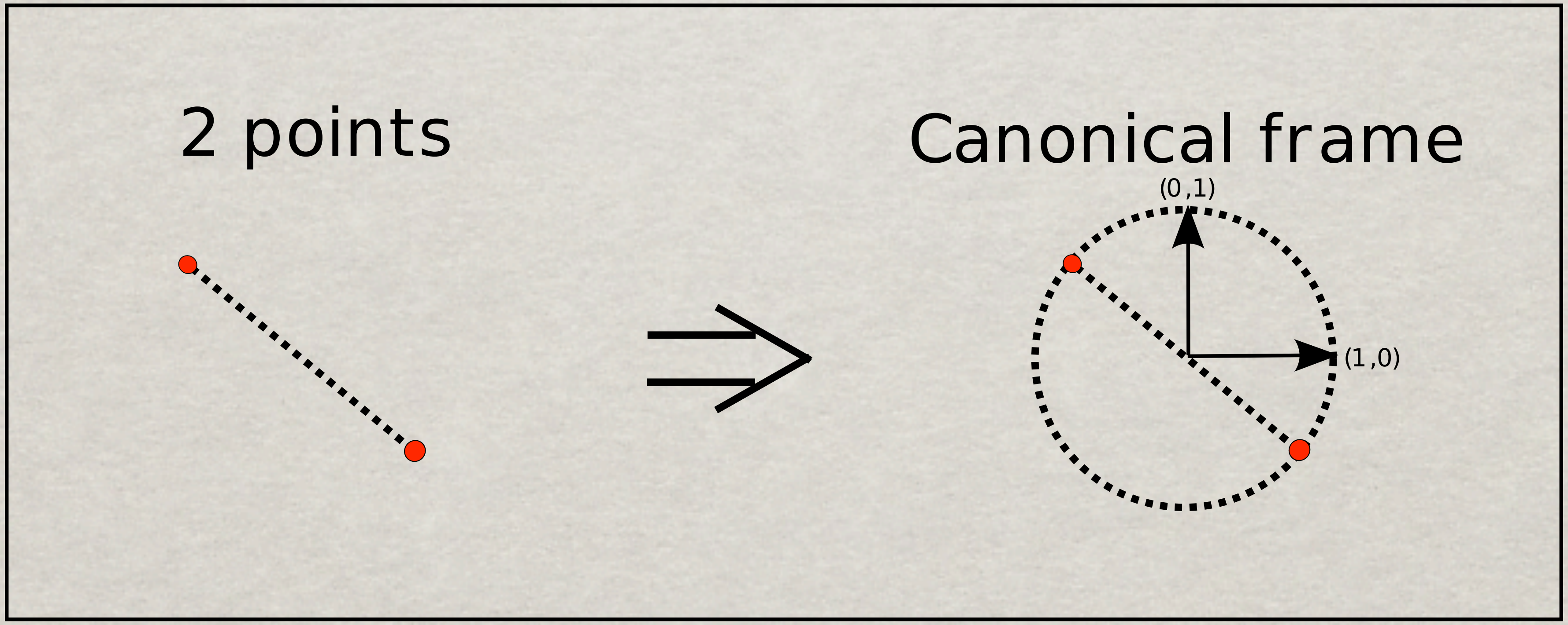
- affine (similarity+skew)

$$\begin{bmatrix} h_{11} & h_{13} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$$

- plane projective (affine+forshortening)

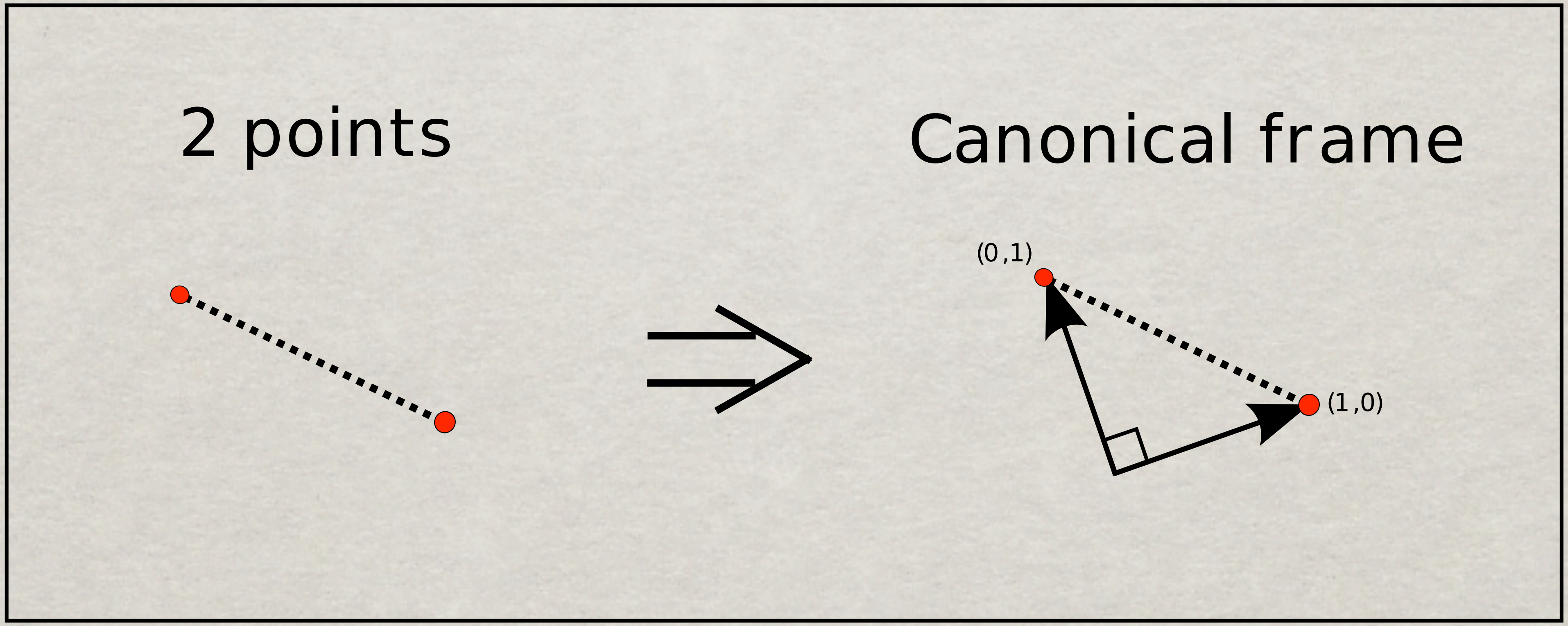
Canonical Frames

- **Scale+translation**: Useful if we know that there is no rotation.
E.g. for a camera mounted in a car, looking at upright pedestrians.



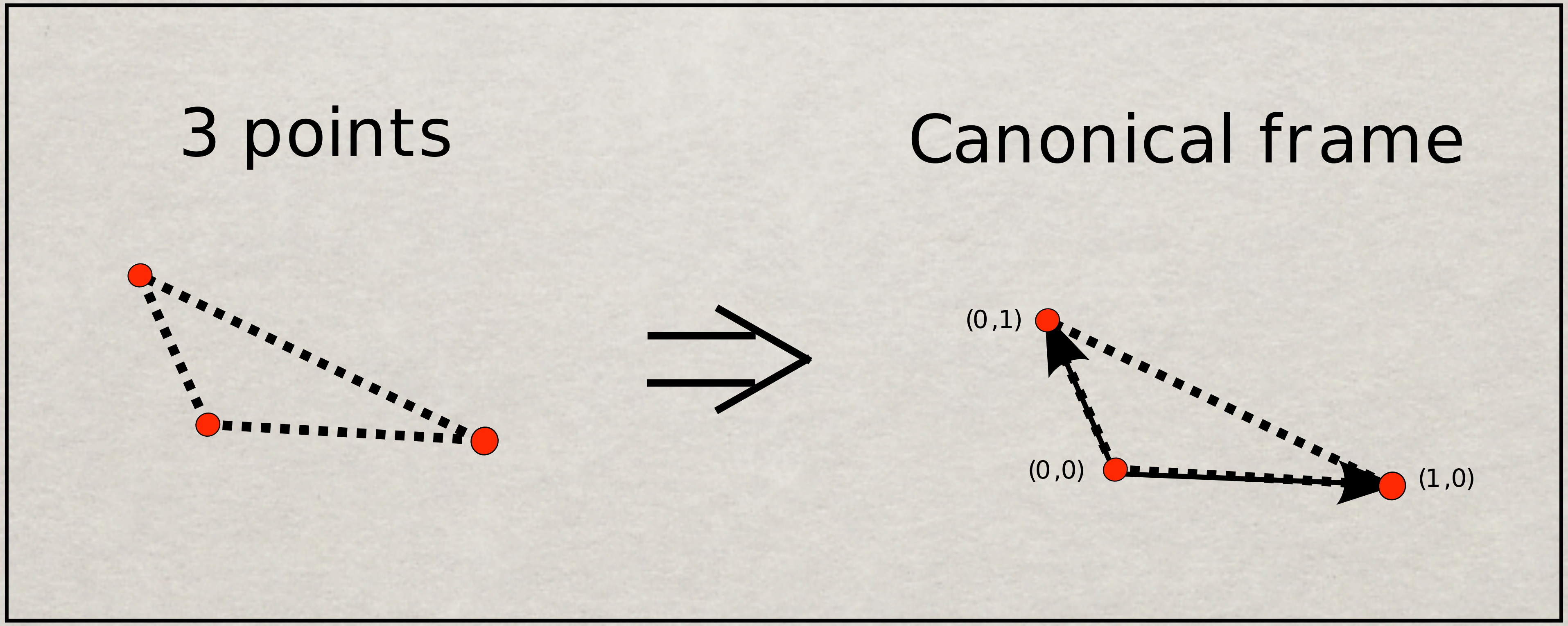
Canonical Frames

- **Similarity**: Full invariance in image plane, none outside image plane.
Useful e.g. for pose estimation.



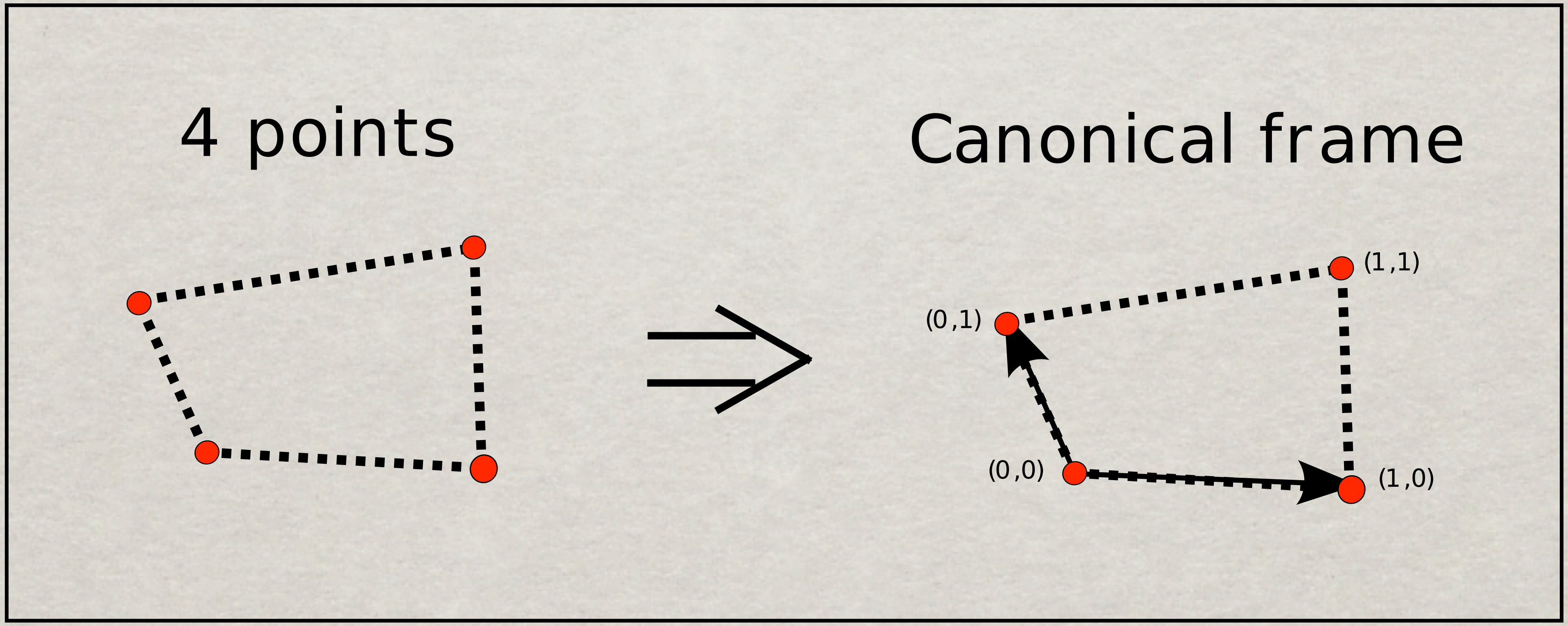
Canonical Frames

- **Affine**: Deals with most common projective distortions. Good if patch size is small relative to distance to patch.



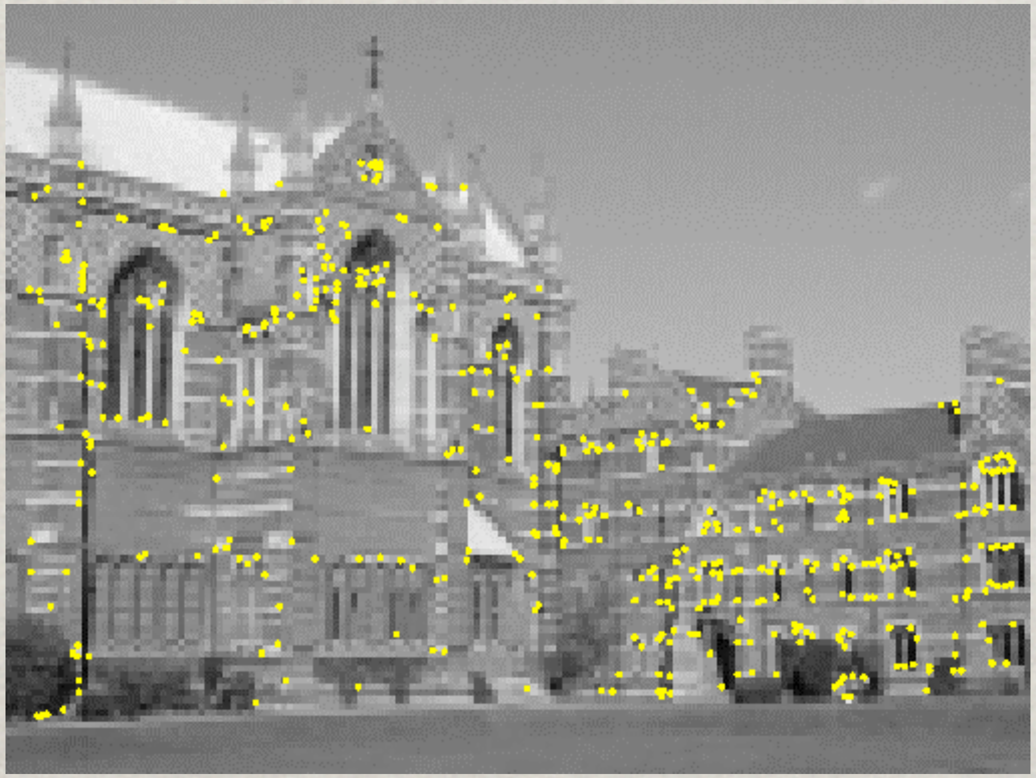
Canonical Frames

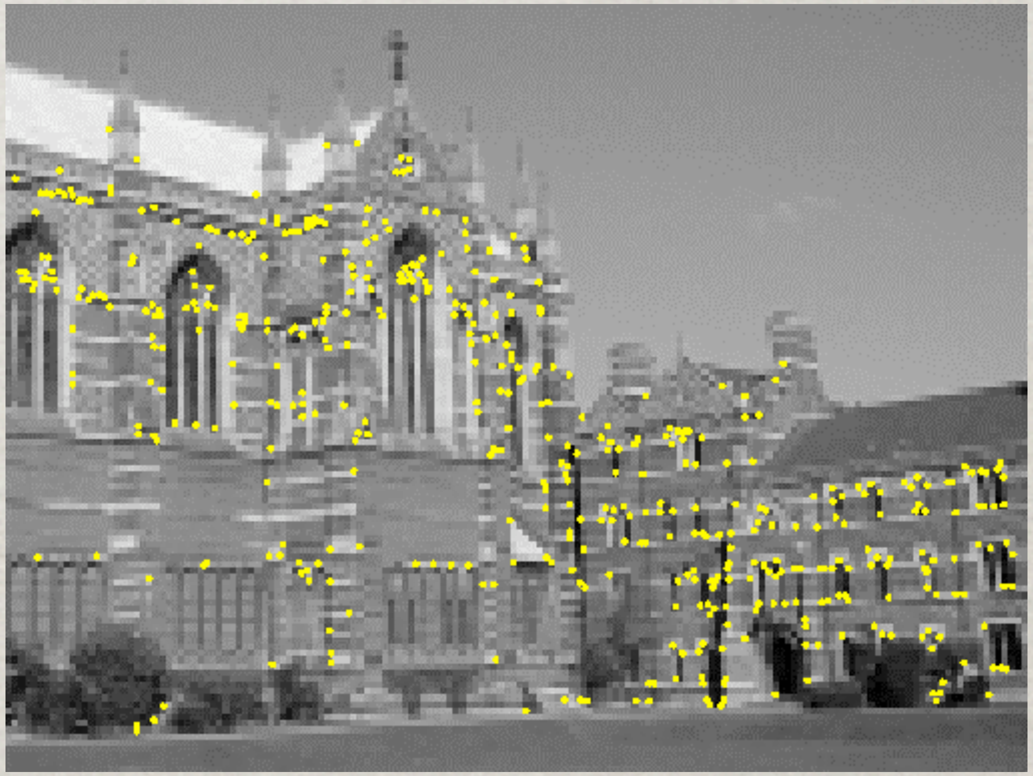
- **Plane projective**: Full modelling of a plane in 3D. Requires more image measurements, but is better for extreme view angles.



Canonical Frames

- **Combinatoral issues**
  - From Harris or SIFT we get images full of keypoints.





Canonical Frames

- **Combinatoral issues**
  - From Harris or SIFT we get images full of keypoints.
  - Using the points, we want to generate frames in both reference and query view and match them.
  - We don't want to miss a combination in one of the images, but we don't want to generate too many combinations either.

Canonical Frames

- Solutions:
  - Use each point as a reference point.
  - Restrict frame construction to k-nearest neighbours in scale space (or image plane).
  - Remove duplicate groupings, and reflections.

Discussion

- Questions/comments on paper and lecture
- E.g. Why k-nearest neighbours in scale-space? Are there other useful canonical frames?...