Visual Object Recognition

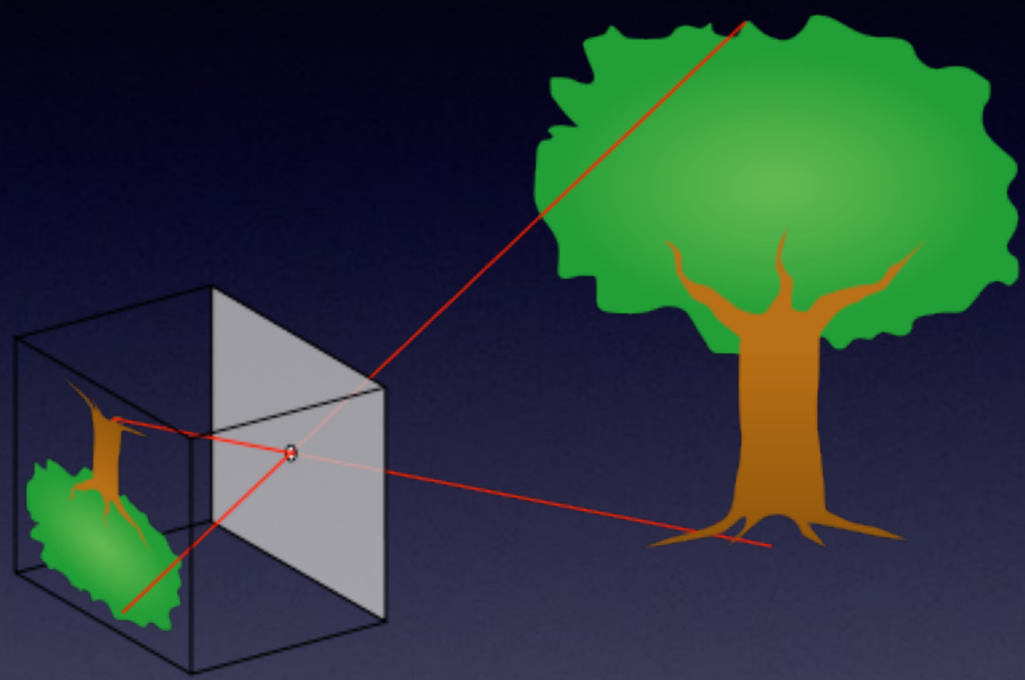
Lecture 2: Image Formation

Per-Erik Forssén, docent
Computer Vision Laboratory
Department of Electrical Engineering
Linköping University

Lecture 2: Image Formation

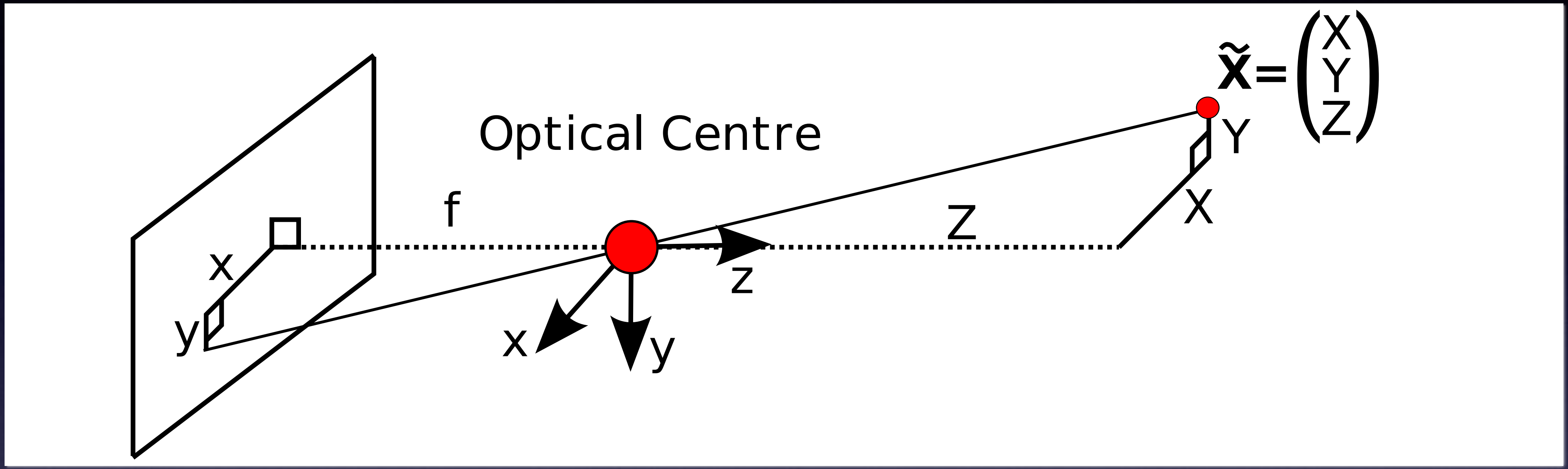
- Pin-hole, and thin lens cameras
  Projective geometry, lens distortion, vignetting, intensity, colour
- Geometric and Photometric Invariance
  Colour constancy, colour spaces, affine illumination model, homographies, epipolar geometry, canonical frames

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:

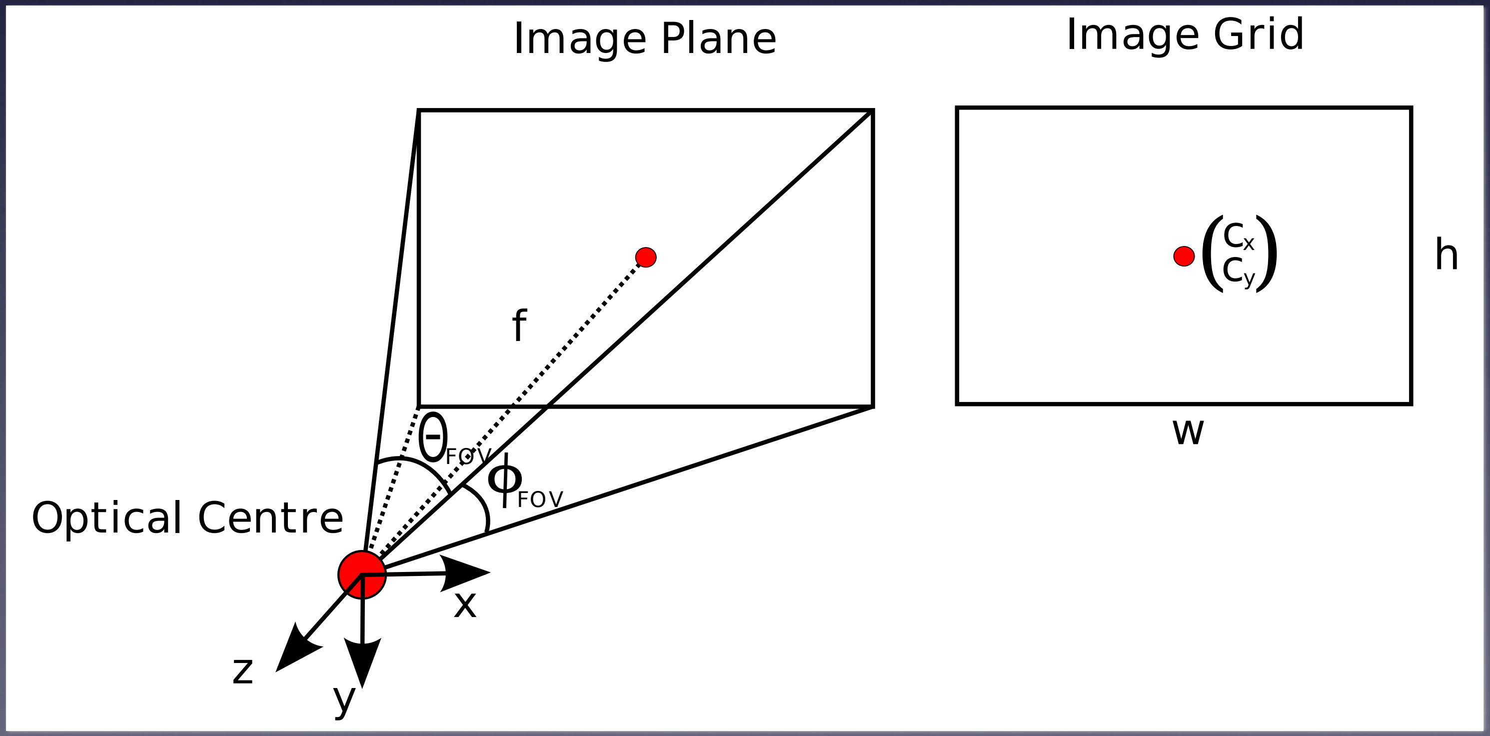
$$\gamma \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f & s & c_x \\ 0 & fa & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

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

The Pin-Hole Camera

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

- Motivation:

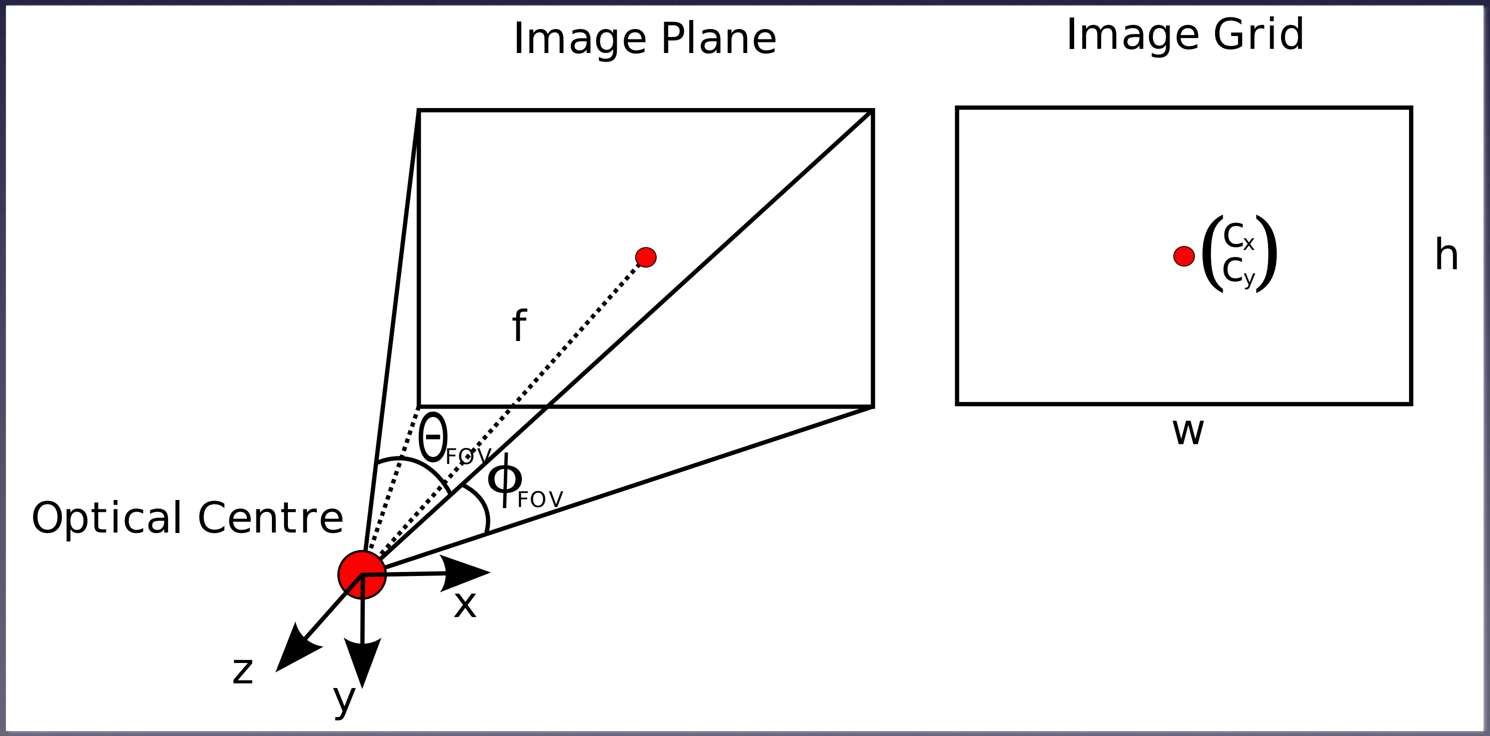


f-focal length, s-skew, a-aspect ratio, **c**-projection of optical centre

The Pin-Hole Camera

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

- Motivation:



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

The Pin-Hole Camera

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

- Also **normalized image coordinates**:

$$\mathbf{u} = \begin{bmatrix} X/Z \\ Y/Z \\ 1 \end{bmatrix}$$

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

$$\mathbf{u} = \mathbf{K}^{-1}\mathbf{x}$$

The Pin-Hole Camera

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

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

The Pin-Hole Camera

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

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

The Pin-Hole Camera

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

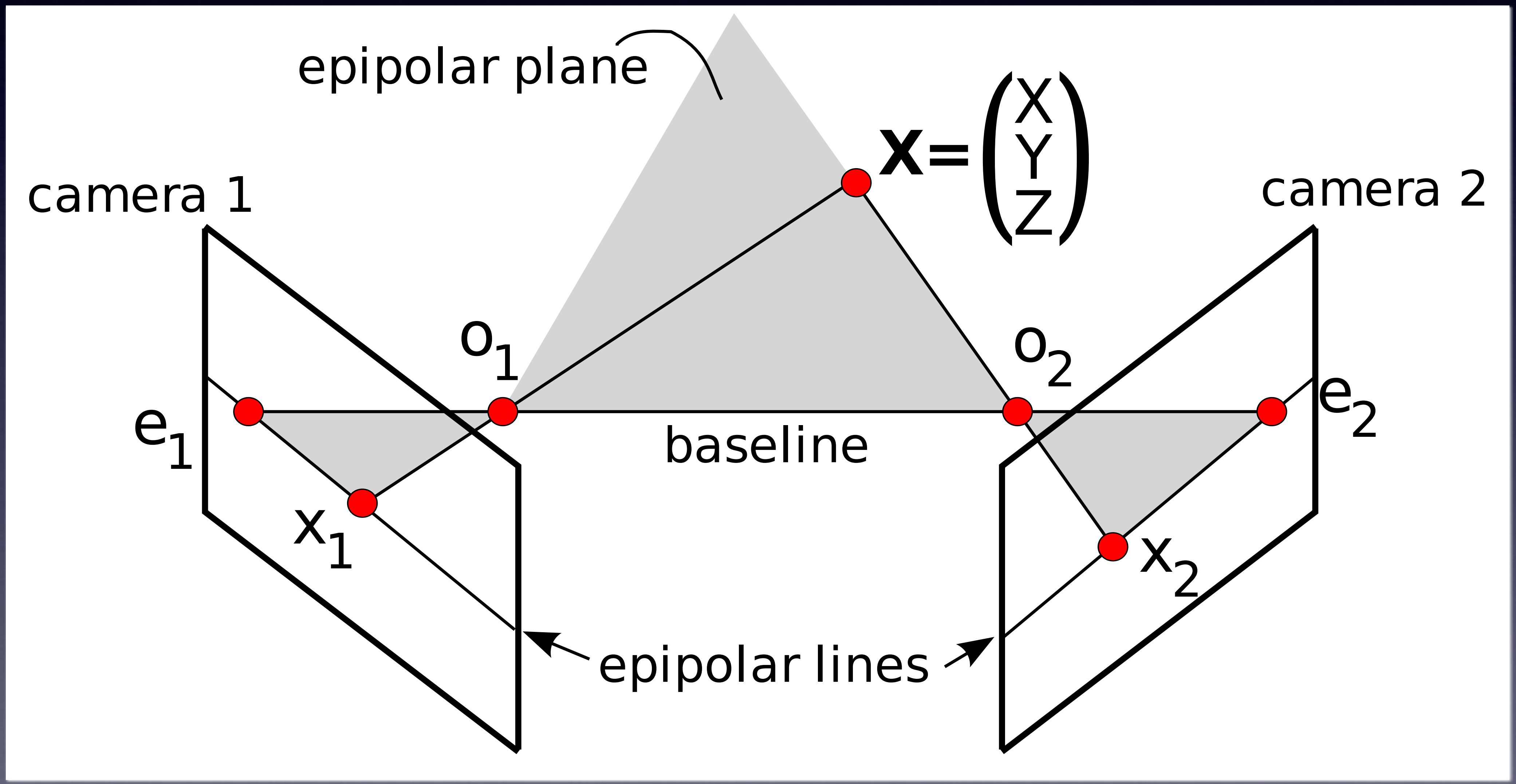
$$\mathbf{u} \sim \underbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix}}_{[\mathbf{R}|\mathbf{t}]} \underbrace{\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}}_{\mathbf{X}} \quad \Leftrightarrow \quad \mathbf{u} \sim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

and thus

$$\mathbf{x} \sim \mathbf{K}[\mathbf{R}|\mathbf{t}]\mathbf{X}$$

Epipolar geometry

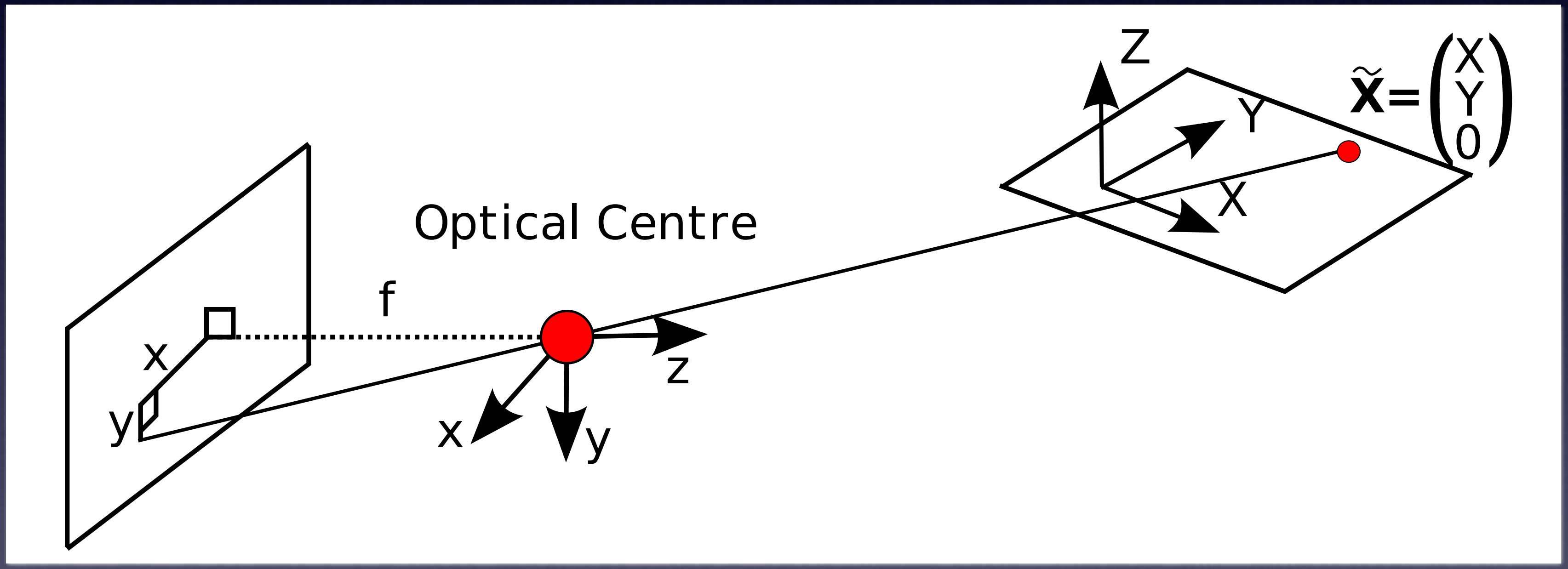
- The **epipolar geometry** of two cameras:



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

Homographies

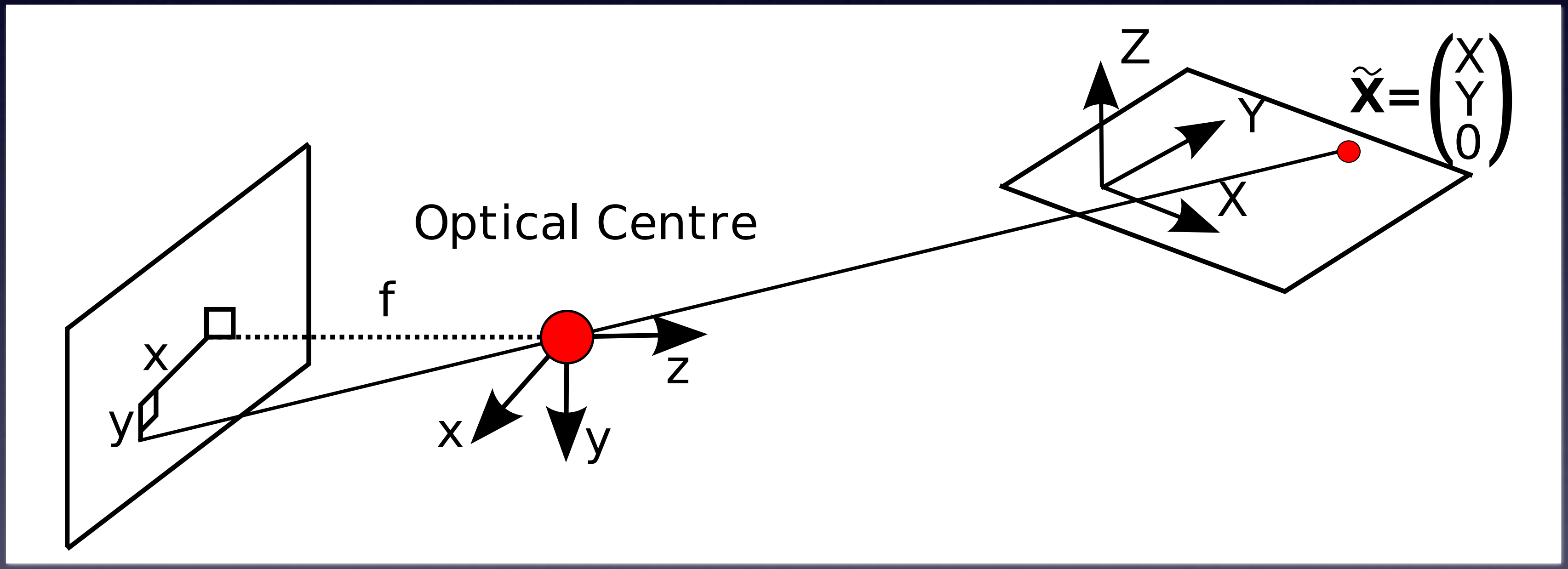
- For a planar object, we can imagine a world coordinate system fixed to the object



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

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

- Projections into two cameras:

$$x_1 \sim \mathbf{H}_1\mathbf{X} \text{ and } x_2 \sim \mathbf{H}_2\mathbf{X}$$

- As the homography is invertible, we can now map from camera 2 to the object and on to camera 1:

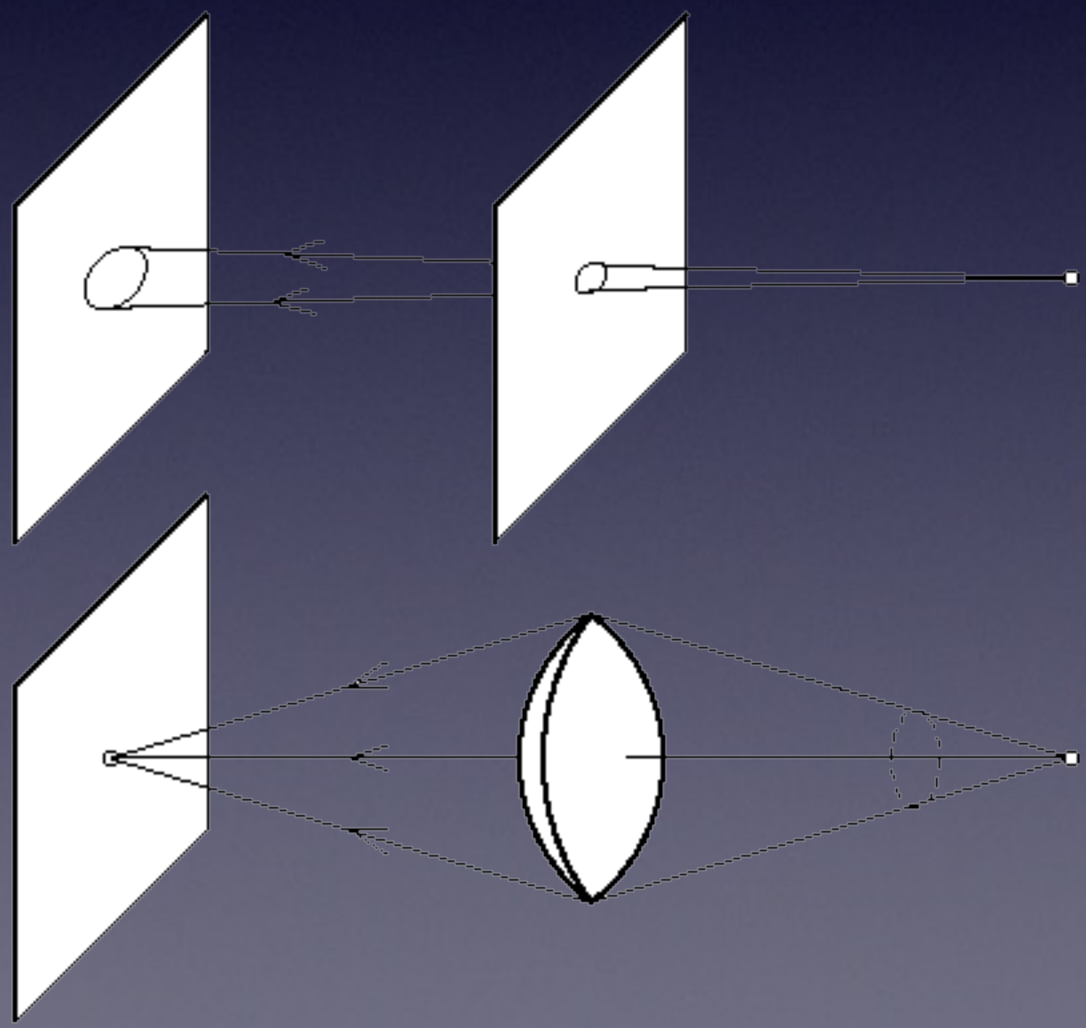
$$\Rightarrow \quad x_1 \sim \mathbf{H}_1\mathbf{H}_2^{-1}x_2$$

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.

Thin Lens Camera

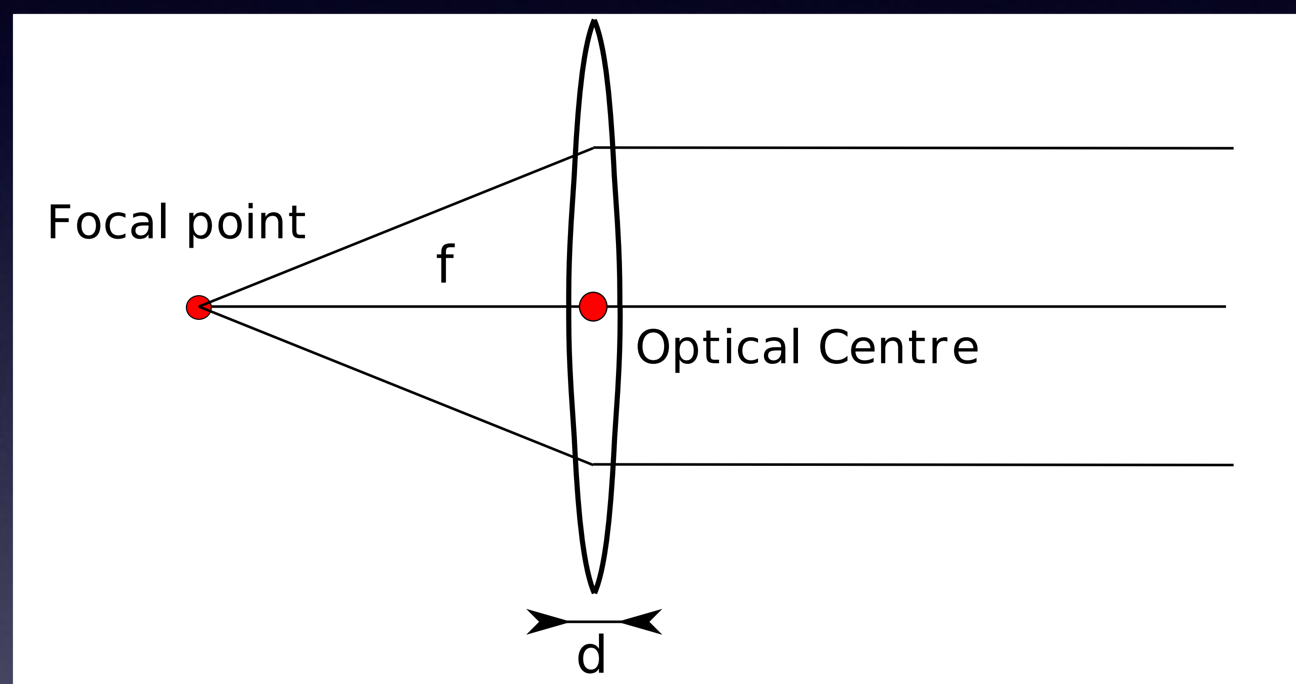
- An actual pinhole lets in too little light, and a bigger hole blurs the picture.



- Real cameras instead use lenses to obtain a sharp image using more light.

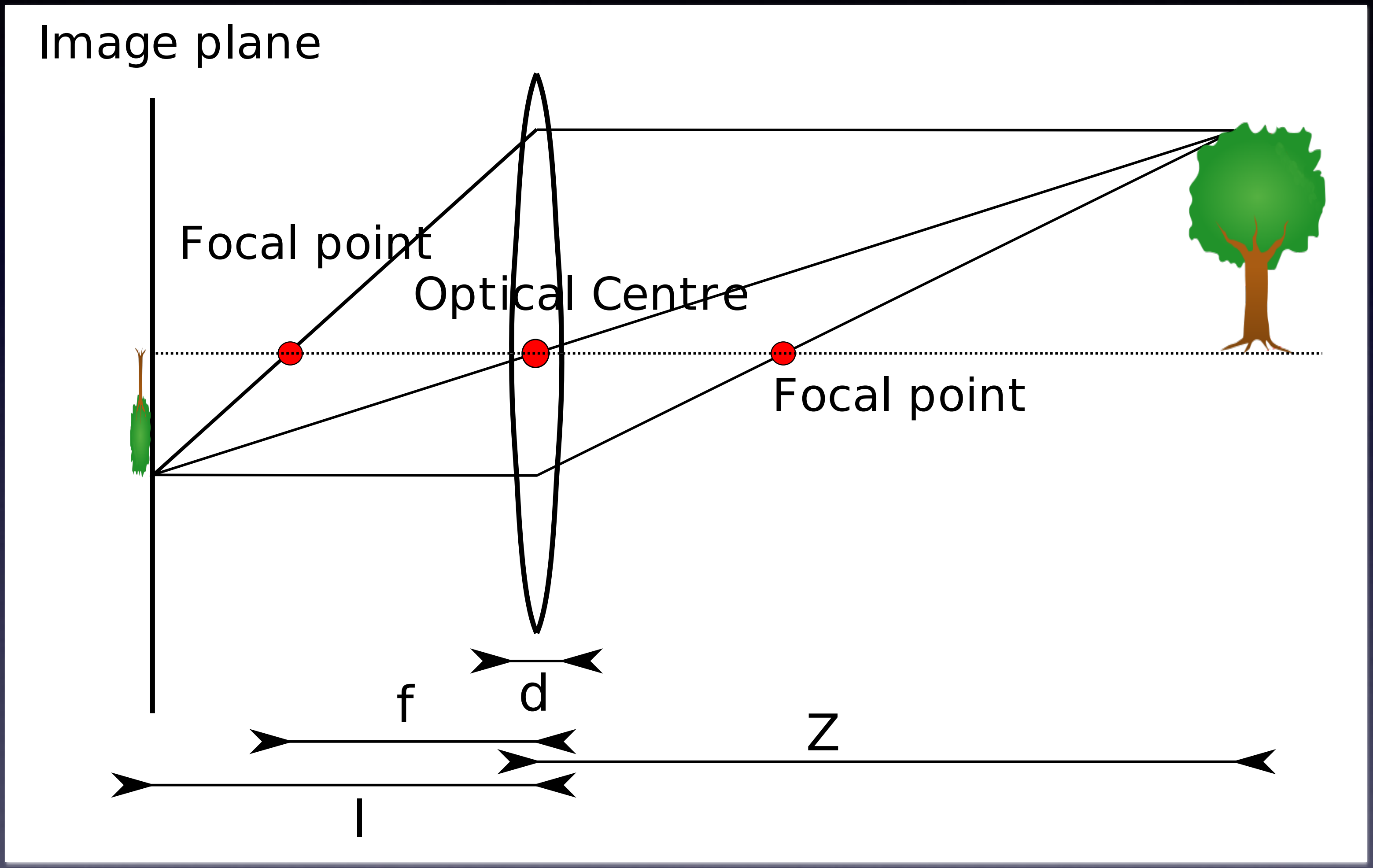
Thin Lens Camera

- 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

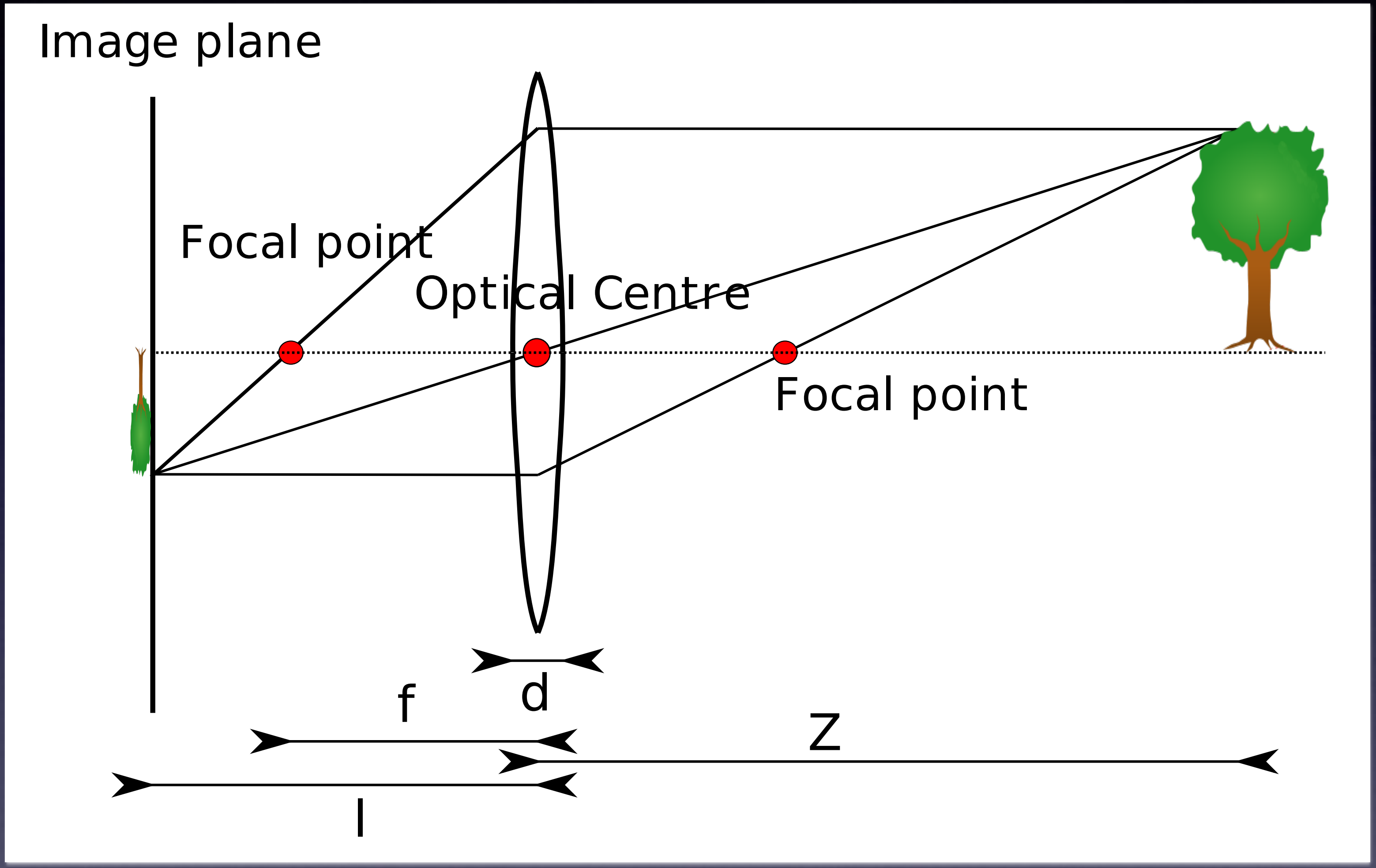
Thin Lens Camera



- Thin lens relation (from similar triangles):

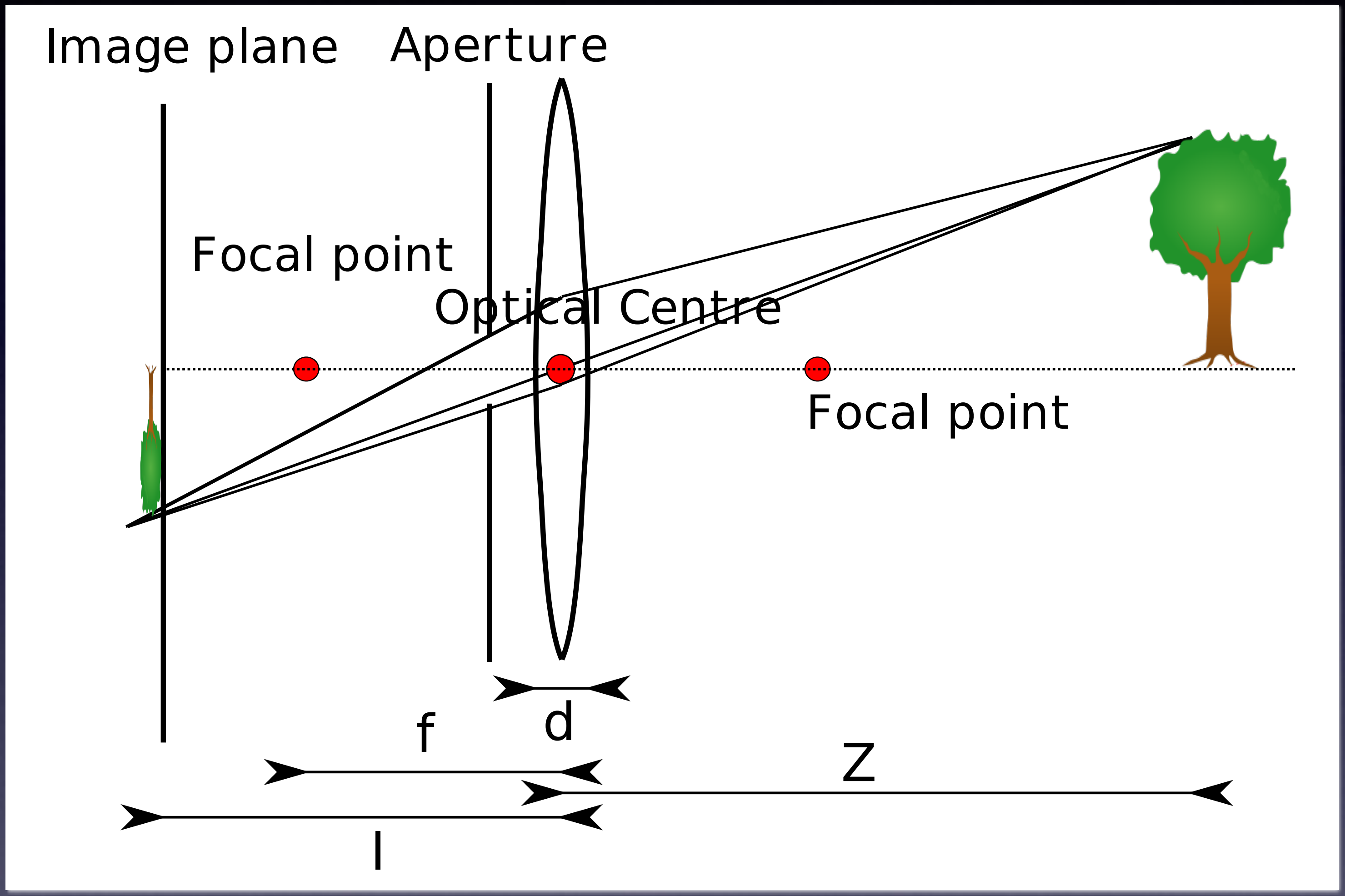
$$\frac{1}{f} = \frac{1}{Z} + \frac{1}{l}$$

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.

Lens distortion

Correct

Barrel distortion

Pin-cushion distortion

- For zoom lenses:

  - Barrel at wide FoV
  - pin-cushion at narrow FoV

Lens distortion

Correct image

$\Rightarrow$

Distorted

- Modelling $\mathbf{x} \sim \mathbf{K} f(\mathbf{u}, \boldsymbol{\Theta}')$
- Used in optimisation such as BA

Lens distortion

Distorted image

$\Rightarrow$

Correct

- Rectification $\mathbf{x}' \sim f^{-1}(\mathbf{K}\mathbf{u}, \boldsymbol{\Theta})$
- Used in dense stereo

Lens Effects

Correct

Darkened periphery

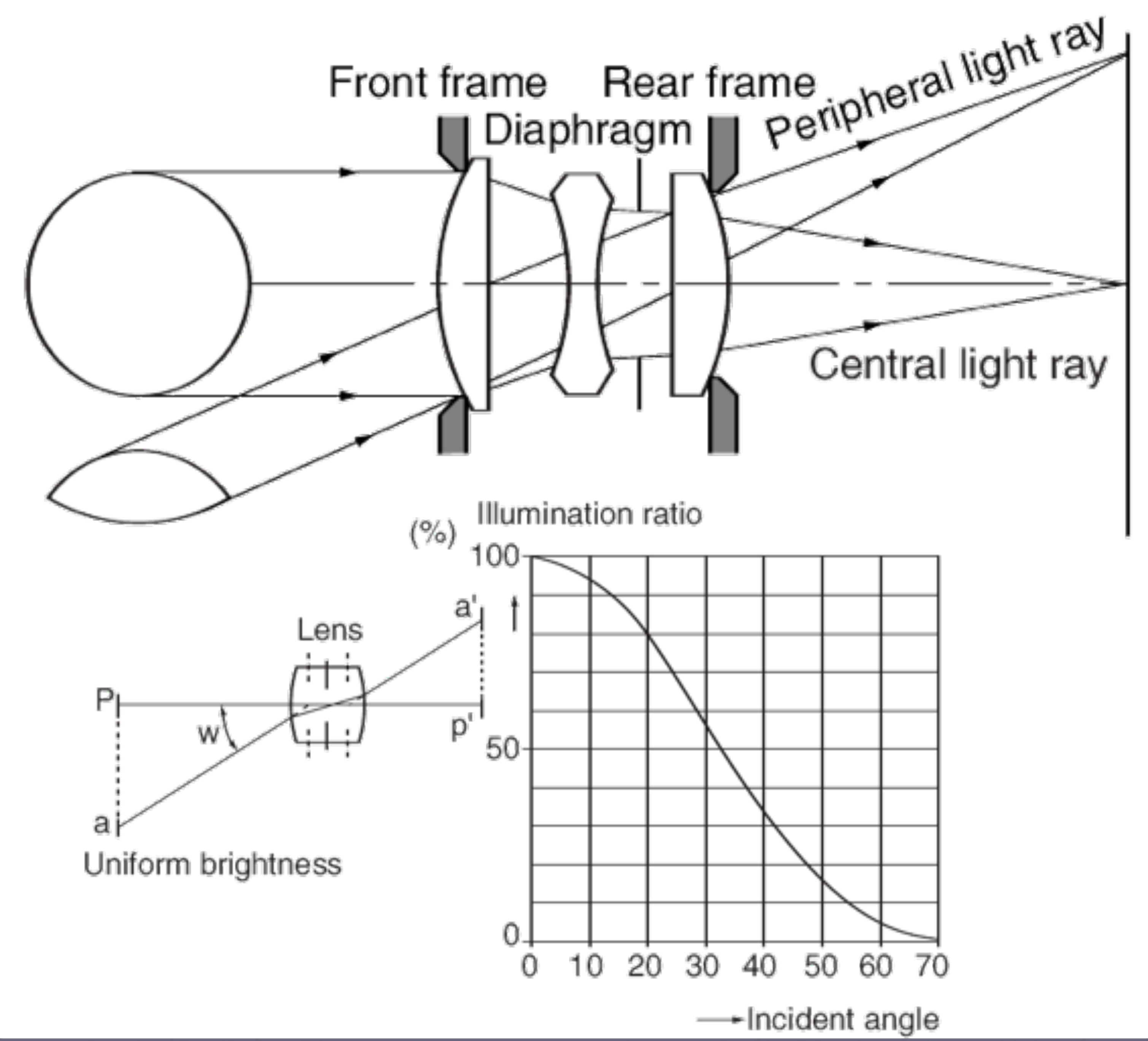
- Vignetting and $\cos^4$-law

  - more severe on wide angle cameras

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

Image intensity

- Sensor activation is linear

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

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

Image intensity

- Sensor activation is linear

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

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

- However, most images are gamma corrected

$$i(\mathbf{x}) = a(\mathbf{x})^\gamma$$

Image intensity

- Sensor activation is linear

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

- HVS handles a 10^{10} dynamic range on $e(\lambda)$

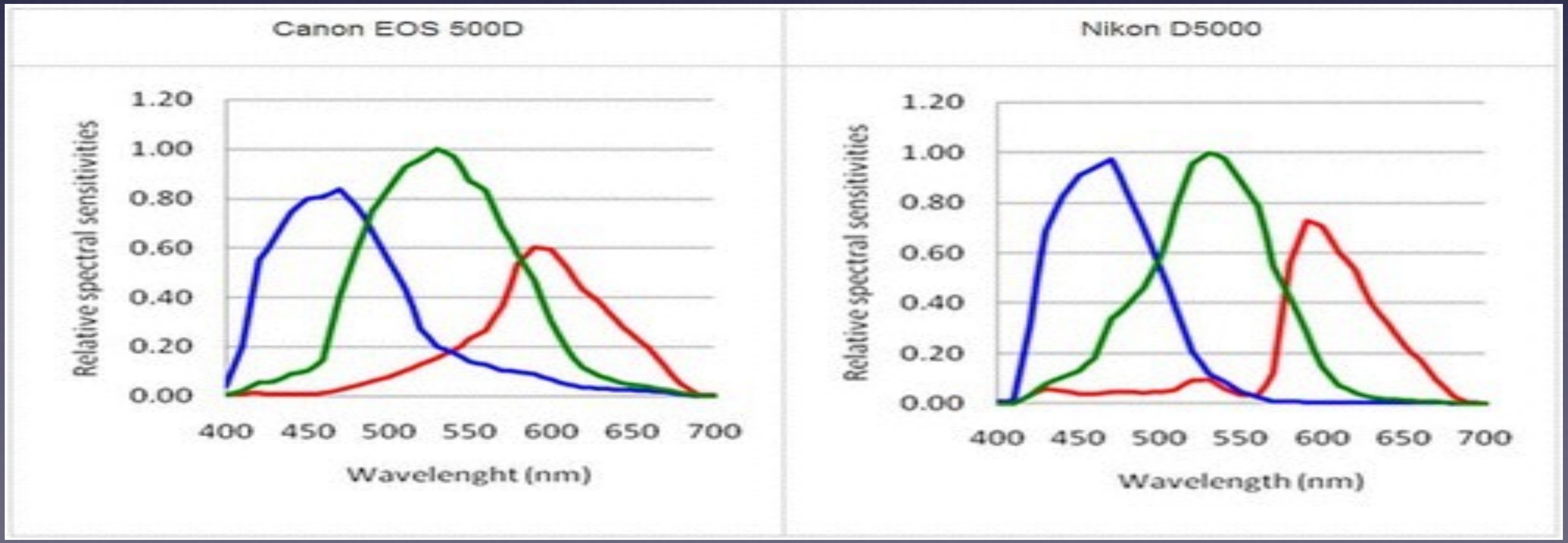
- Exposure time and aperture size also scale the activation.

- Unless we know all aspects of image formation, we cannot trust the absolute intensity value.

Colour

- Colour perception is done using three different activation functions

$$a_k(\mathbf{x}) = \int s_k(\lambda) r(\lambda, \mathbf{x}) e(\lambda) d\lambda,\ k = 1, 2, 3$$



http://publiclab.org/wiki/ndvi-plots-ir-kit

Colour

- Colour perception is done using three different activation functions

$$a_k(\mathbf{x}) = \int s_k(\lambda) r(\lambda, \mathbf{x}) e(\lambda) d\lambda,\ k = 1, 2, 3$$

- Sensor activation is **not** colour. Colour is an object property, i.e. a representation of $r(\lambda)$.

- In order to estimate colour, we need to somehow compensate for the illumination $e(\lambda)$.

Invariance Transformations

varying illumination

varying camera pose

- Two categories of nuisance factors for recognition/matching

Invariance Transformations

varying illumination

varying camera pose

- For matching we need either to know the changes, or an **invariance transformation**
- Ideally, an invariance transformation should keep information intrinsic to the object, but remove all influence from the imaging process

Invariance Transformations

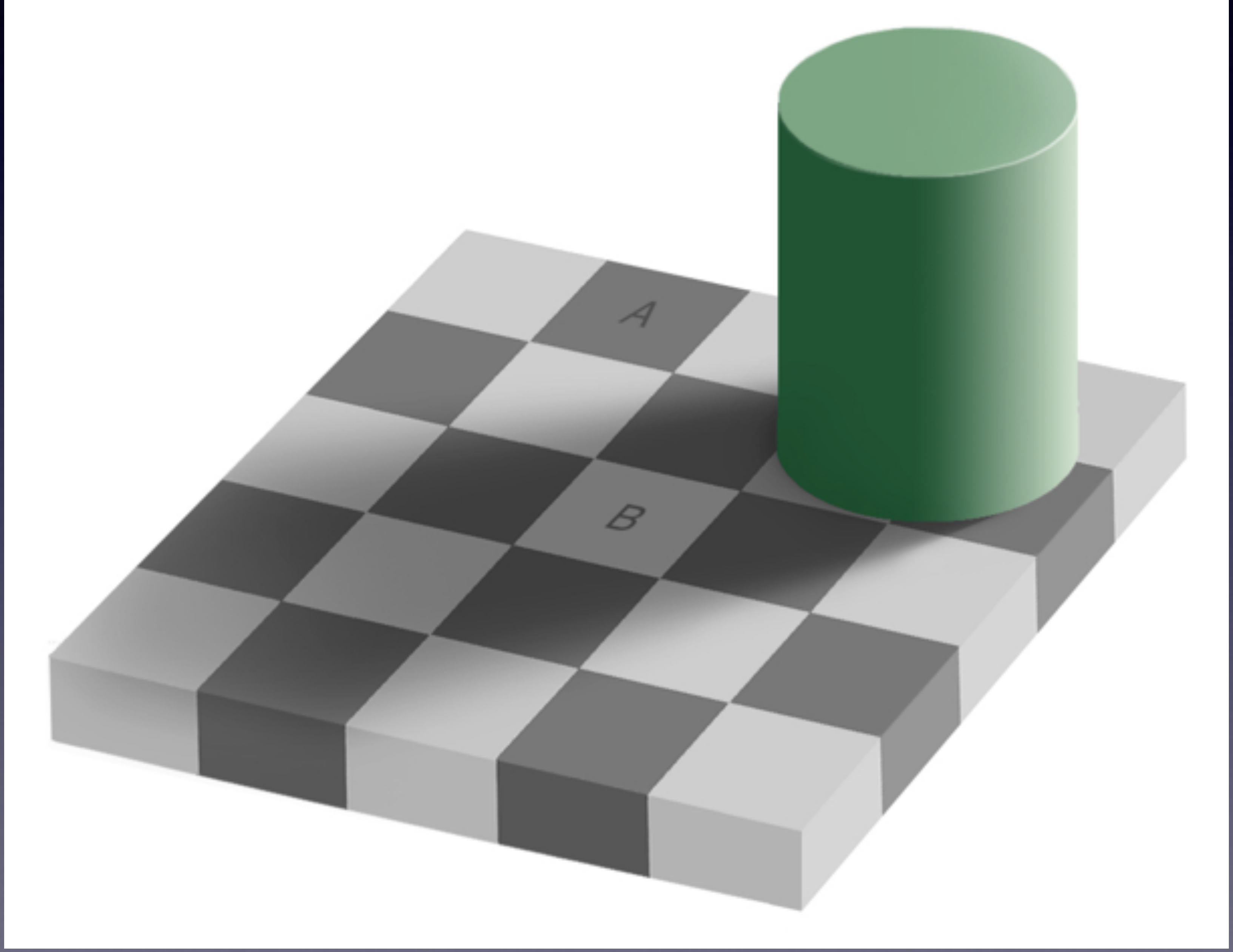
- **Photometric invariance** gives robustness to illumination changes

varying illumination

- **Geometric invariance** gives robustness to view changes

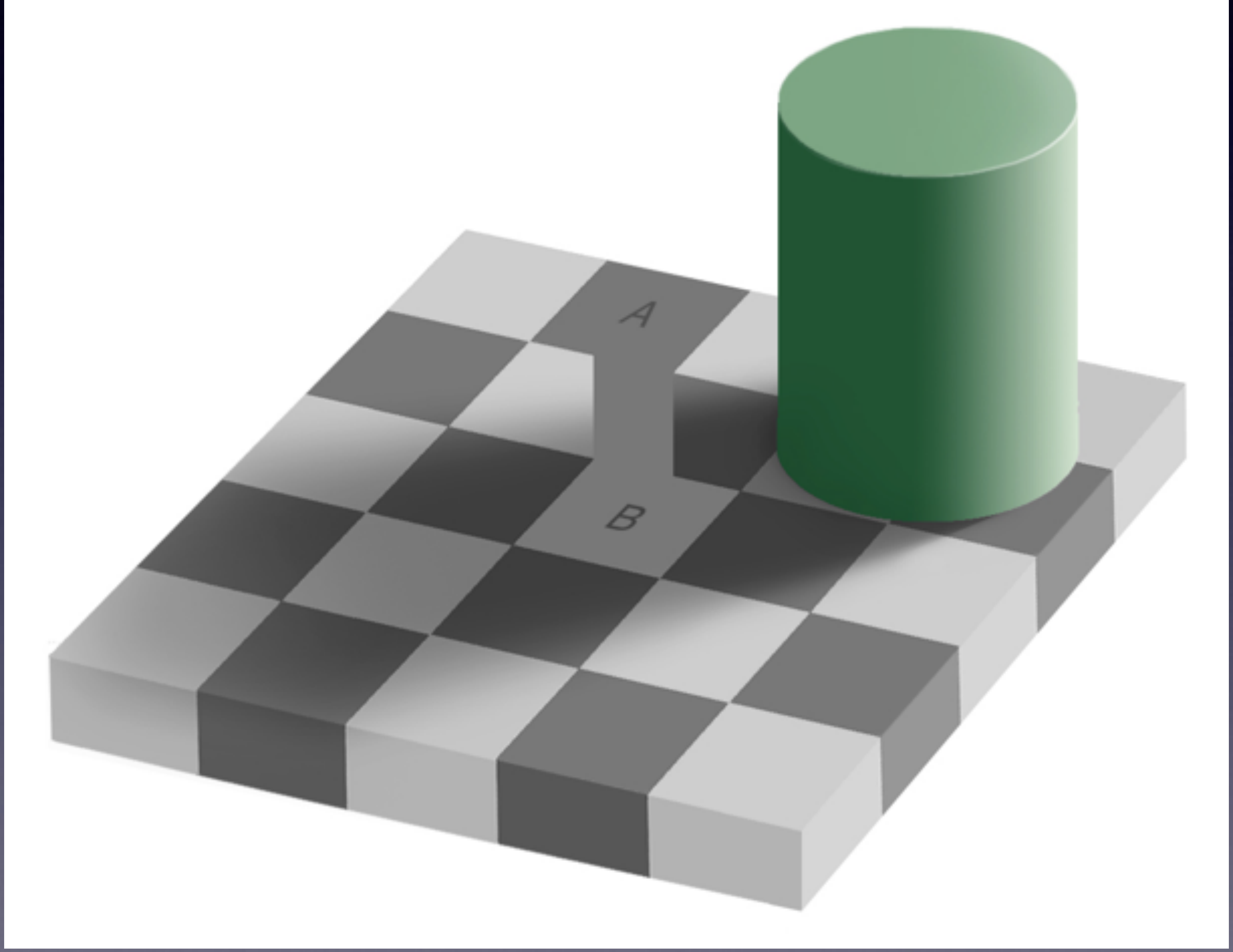
varying camera pose

Colour Constancy



http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html

Colour Constancy



http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html

Colour Constancy

- The colours we perceive are not the activations of cones in the retina.
- **Colour constancy** is an attempt by the HVS to transform the retinal activation into a normalized (white) reference illumination.
- Complex process that takes place at many levels (retina, V2,…) and uses high level information (e.g. known object colour).
- White balancing in cameras is a low-level technical equivalent.

Colour Constancy

- Object based colour transfer
- If you know what you are looking at, you also know something about the illumination

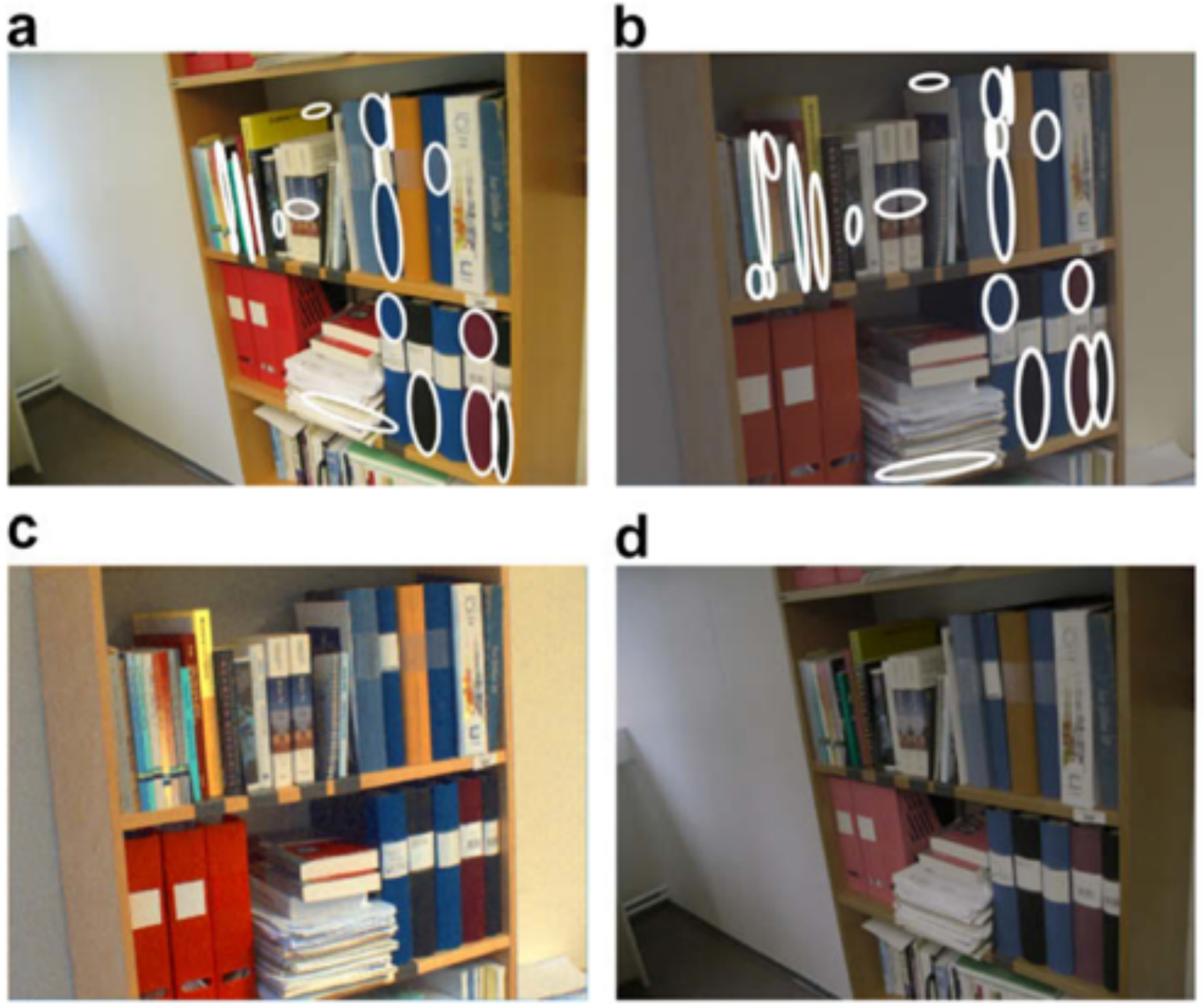


Fig. 10. (a and b) Two images with corresponding detected blobs shown, (c) image in b, after colour correction to match a. (d) image in a, after colour correction to match b.

Forssén & Moe, View Matching with Blob Features, JIVC'09

Photometric invariance

Input

$I(\mathbf{x}) = k_1 I_0(\mathbf{x})$

$J(\mathbf{x}) = k_2 I_0(\mathbf{x})$

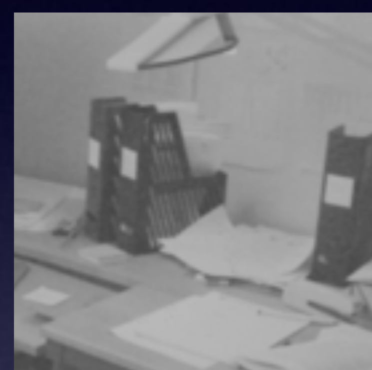
- If illumination changes, direct matching fails

$$\sum_{\mathbf{x}} ||I(\mathbf{x}) - J(\mathbf{x})|| = \text{large}$$

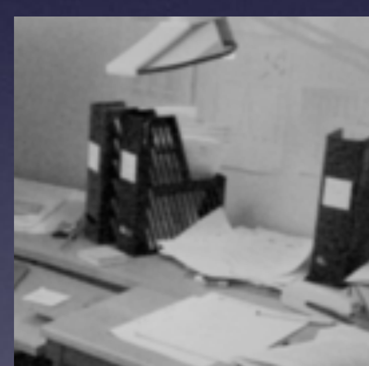
Photometric invariance

Input

$I(\mathbf{x}) = k_1 I_0(\mathbf{x})$



$J(\mathbf{x}) = k_2 I_0(\mathbf{x})$



- If illumination changes, direct matching fails

$$\sum_{\mathbf{x}} ||I(\mathbf{x}) - J(\mathbf{x})|| = \text{large}$$

- We seek a function that is invariant to scalings

$$\sum_{\mathbf{x}} ||f(I(\mathbf{x})) - f(J(\mathbf{x}))|| = \text{small}$$

Photometric invariance

- For cameras with non non-linear radiometric response (and e.g. gamma correction), or if two different cameras are used we may use the **affine model**:

$$I(\mathbf{x}) = k_1 I_0(\mathbf{x}) + k_0$$

- How should we choose f ? we want:

$$\sum_{\mathbf{x}} ||f(I(\mathbf{x})) - f(J(\mathbf{x}))|| = \text{small}$$

Photometric invariance

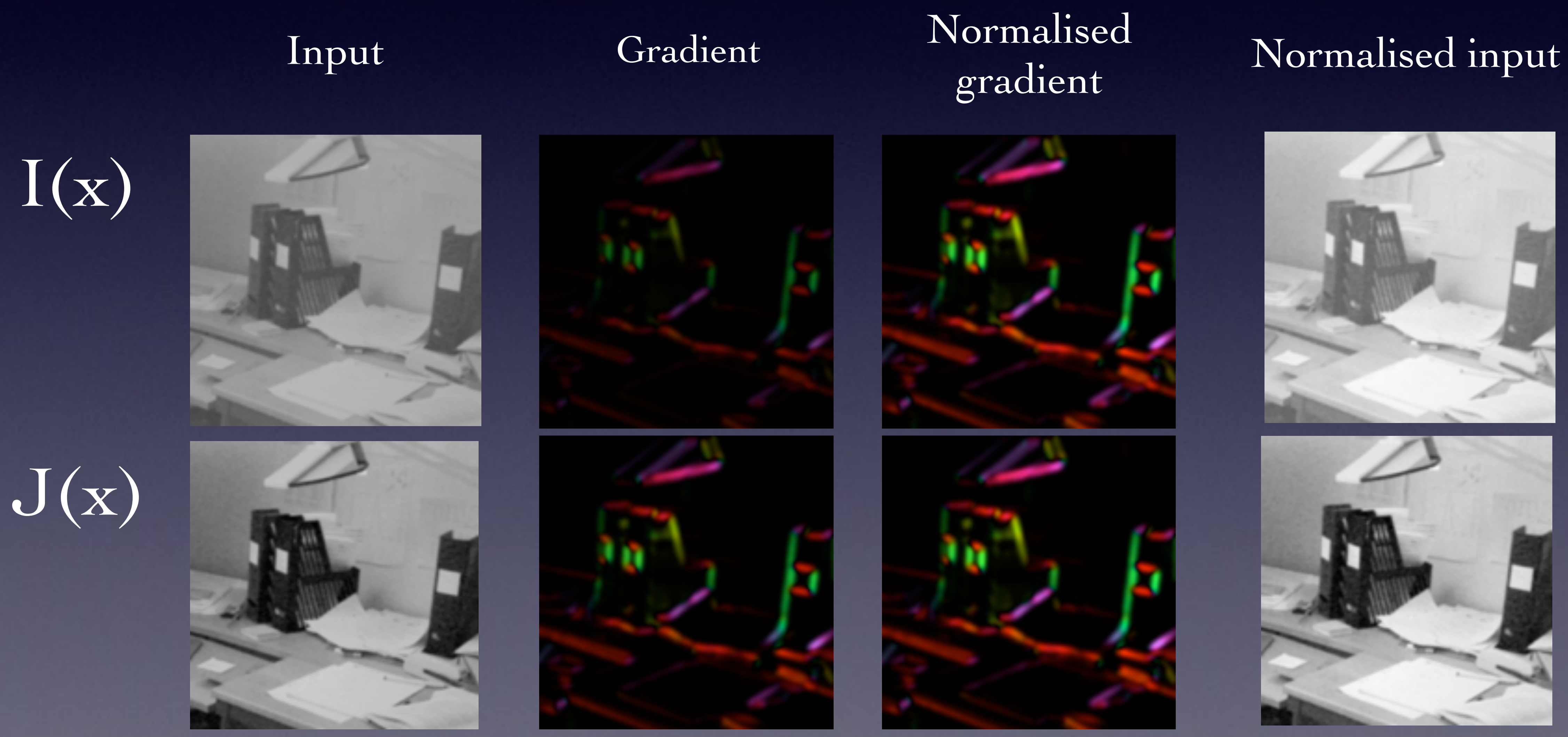
- Mean subtraction, derivatives, and other DC free linear filters remove a constant offset in intensity
- Normalising a patch by e.g. the standard deviation, removes scalings of the intensity.
- Affine invariance by combining both:

$$f(I(\mathbf{x})) = (I(\mathbf{x}) - \mu_I)/\sigma_I$$

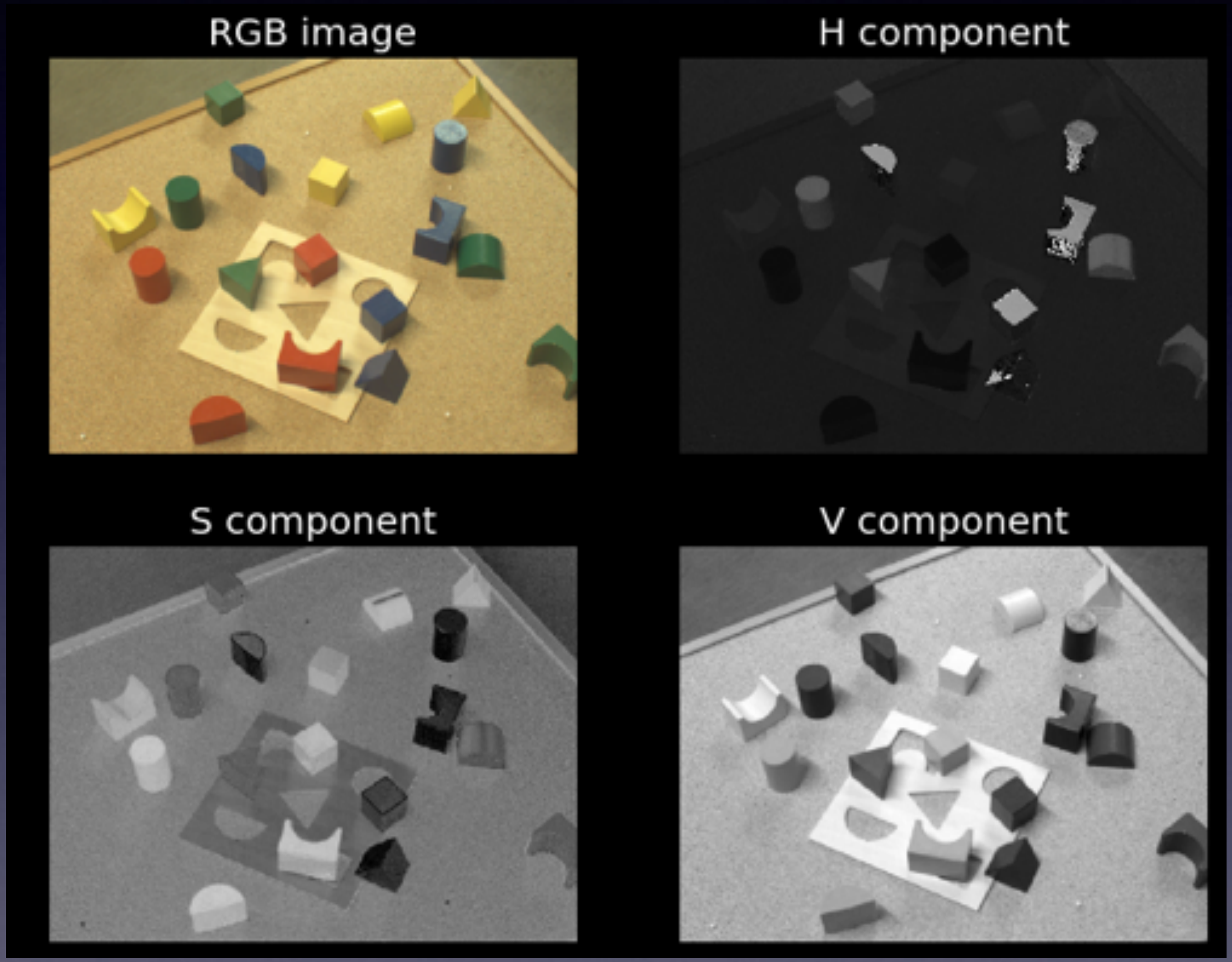
$\mu_I = \text{mean of patch}$ $\sigma_I = \text{std of patch}$

Photometric invariance

- Illustration

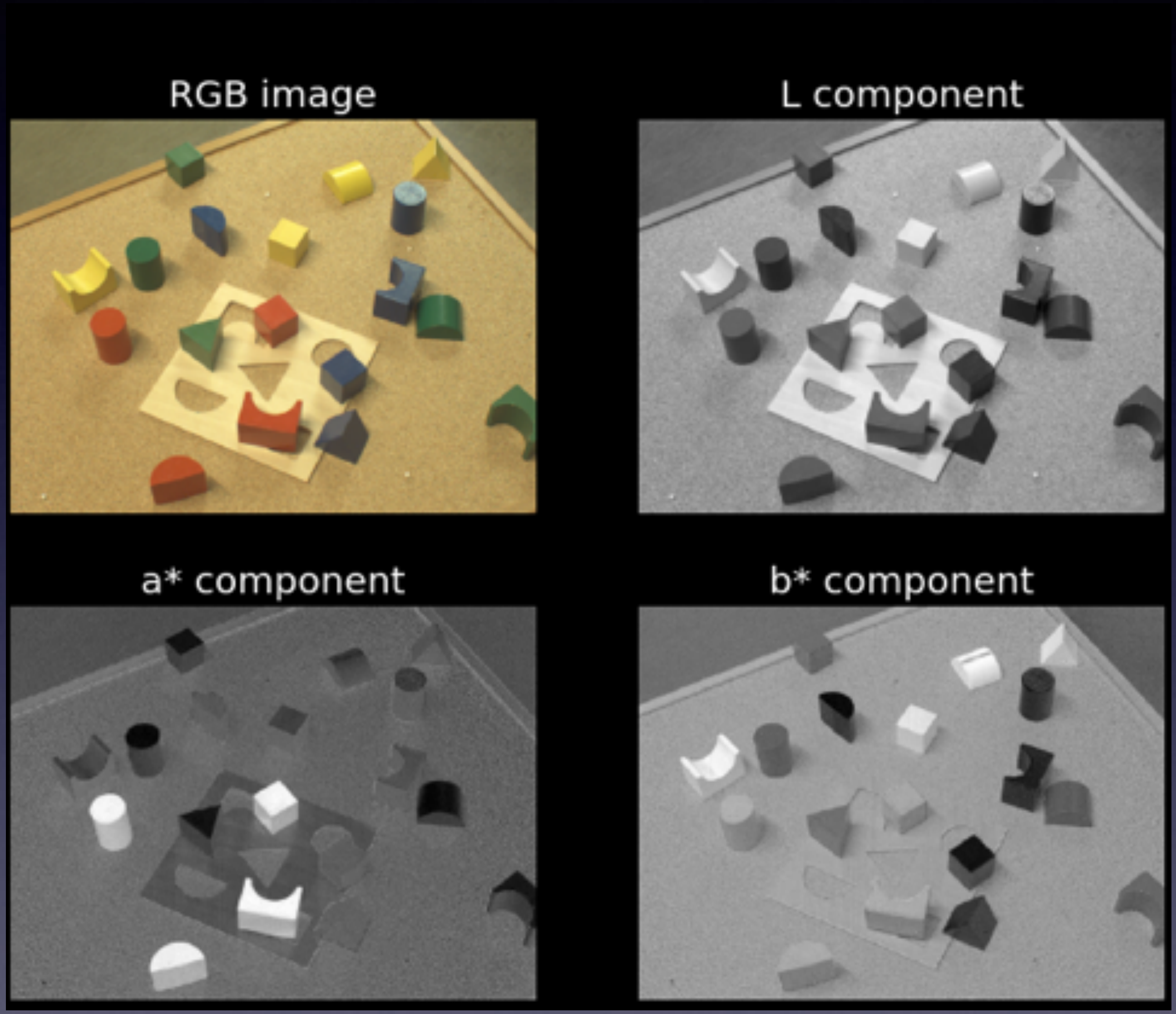


Other colour spaces



- Transform each pixel separately
- Move intensity change into one dimension. E.g. HSV space.
- In matching, V is then downweighted (or discarded)

Other colour spaces



- Other popular choices are the CIE Lab space (left)
- and the Yuv space.
- Colour space changes do not handle changes in the **illumination colour**.

Geometric invariance

- The geometric invariances we use make a **locally planar assumption**. (see also today's paper)
- They can thus be described using **homographies**.

Geometric invariance

- A **homography** is a transformation between points **x** on one plane, and points **y** on another.

$$\gamma \begin{bmatrix} y_1 \\ y_2 \\ 1 \end{bmatrix} = \mathbf{H} \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}$$

- At most 8 degrees-of-freedom (dof) as $\mathbf{H}$ and $k\mathbf{H}, k \in R \setminus 0$ define the same transformation

- See e.g. R. Hartley and A. Zisserman, *Multiple View Geometry for Computer Vision*

Geometric invariance

- A hierarchy of transformations:

  scale+translation (3dof)
  $$\begin{bmatrix} s & 0 & t_1 \\ 0 & s & t_2 \\ 0 & 0 & 1 \end{bmatrix}$$

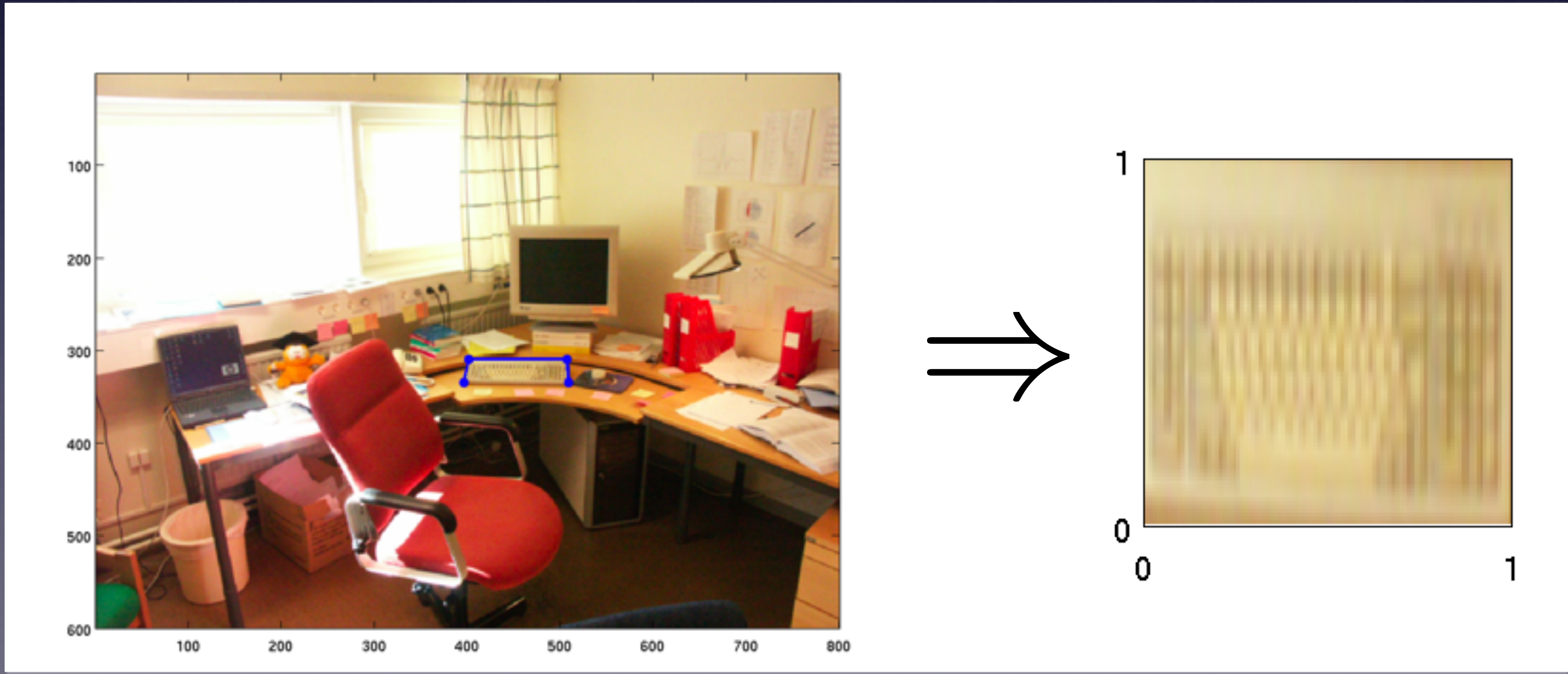
- similarity (4dof)
  (scale+translation+rotation)
  $$\begin{bmatrix} c & s & t_1 \\ -s & c & t_2 \\ 0 & 0 & 1 \end{bmatrix}$$

- affine (6dof)
  (similarity+skew)
  $$\begin{bmatrix} a_{11} & a_{12} & t_1 \\ a_{21} & a_{22} & t_2 \\ 0 & 0 & 1 \end{bmatrix}$$

- plane projective (8dof)
  (affine+forshortening)
  $$\begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$$

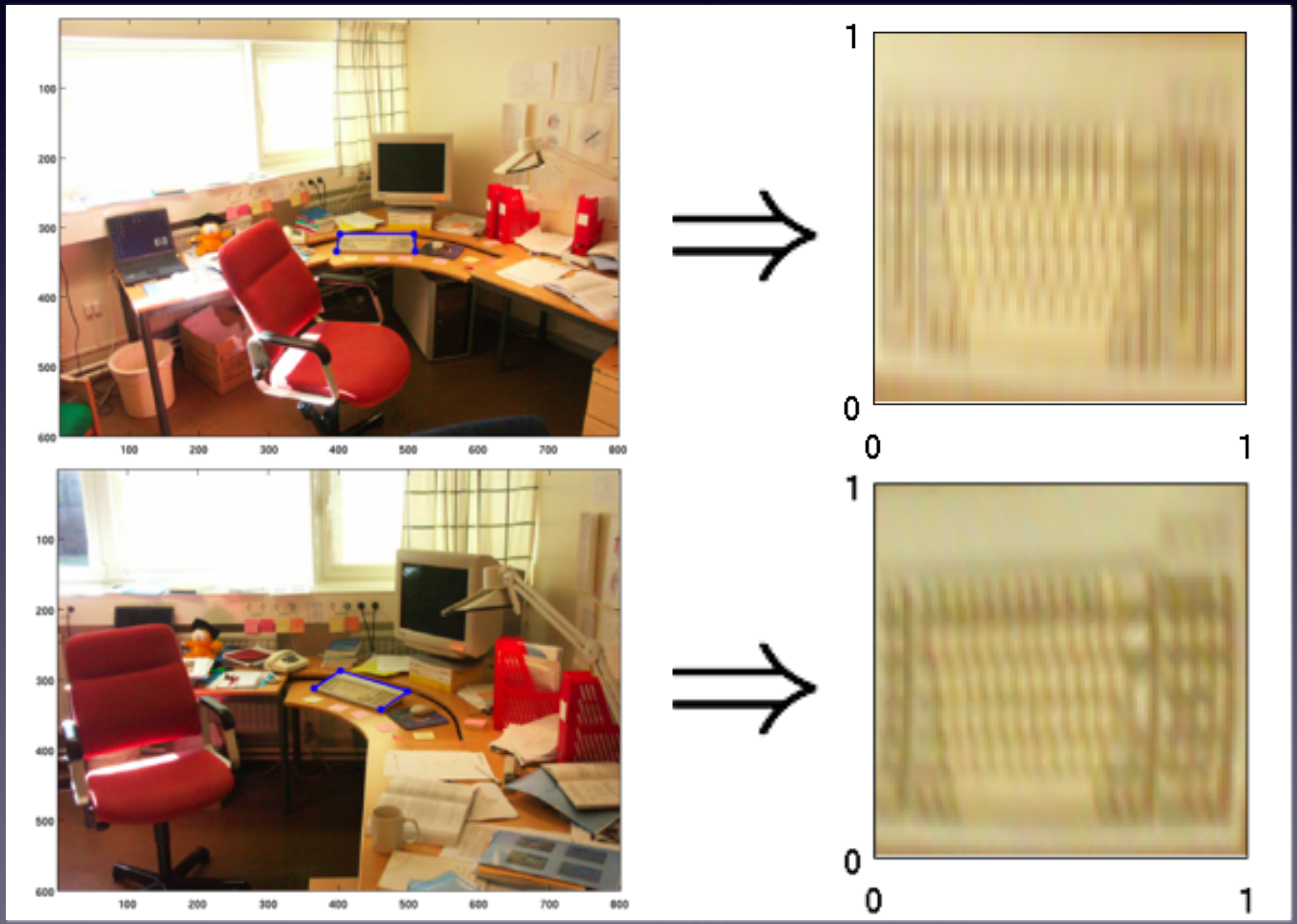
Canonical Frames

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



Canonical Frames

- After resampling matching is much easier!



Canonical Frames

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

Canonical Frames

- **Combinatoral issues**
  - From Harris or DoG 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.

Summary

- **Photometric invariance** is needed to handle changes in illumination.
- Common approaches: **colour constancy**, other **colour spaces**, and **normalisation**
- **Geometric invariance** handles changes in object pose
- A **locally planar** assumption is very useful for geometric invariance

Discussion

- Questions/comments on paper:

  M. Brown, D. Lowe, "Invariant Features from Interest Point Groups", BMVC 2002

- PhD students only, round robin scheduling