Image Formation

P. Fua
(Taught by M. Salzmann)
IC-CVLab
EPFL

Reminder: Computer Vision

Goal: Inferring the properties of the world from one or more images

- Photographs
- Video Sequences
- Medical images
- Microscopy data

→ **Image Understanding**

Reminder: Challenges

Vision involves dealing with:

- Noisy images
- Many-to-one mapping
- Aperture problem

→ Requires:

- Assumptions about the world
- Statistical and physics-based models
- Training data

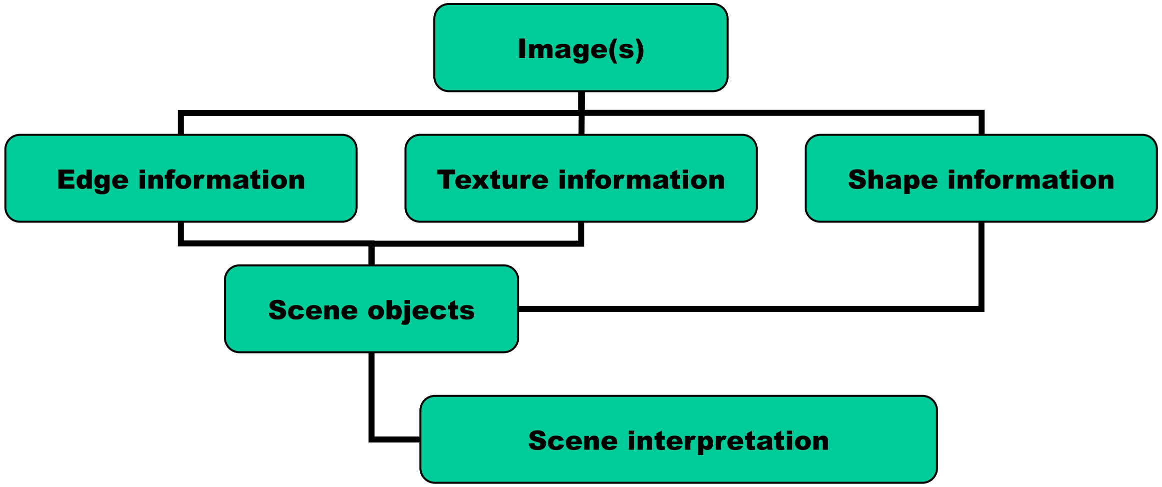
True image understanding seems to require a great deal of thinking. We are not quite there yet.

Reminder: Historical Perspective

- 1960s: Beginnings in artificial intelligence, image processing and pattern recognition.
- 1970s: Foundational work on image formation.
- 1980s: Vision as applied mathematics, geometry, multi-scale analysis, control theory, optimization.
- 1990s: Physics-based models, Probabilistic reasoning.
- 2000s: Machine learning.
- 2010s: Deep Learning.
- 2020s: ?????

--> Improved understanding and successful applications in graphics, mapping, biometrics, and others but still far from human performance.

Reminder: A Teachable Scheme



Decomposition of the vision process into smaller manageable and implementable steps.

--> Paradigm followed in this course

--> May not be the one humans use

Reminder: Human Vision

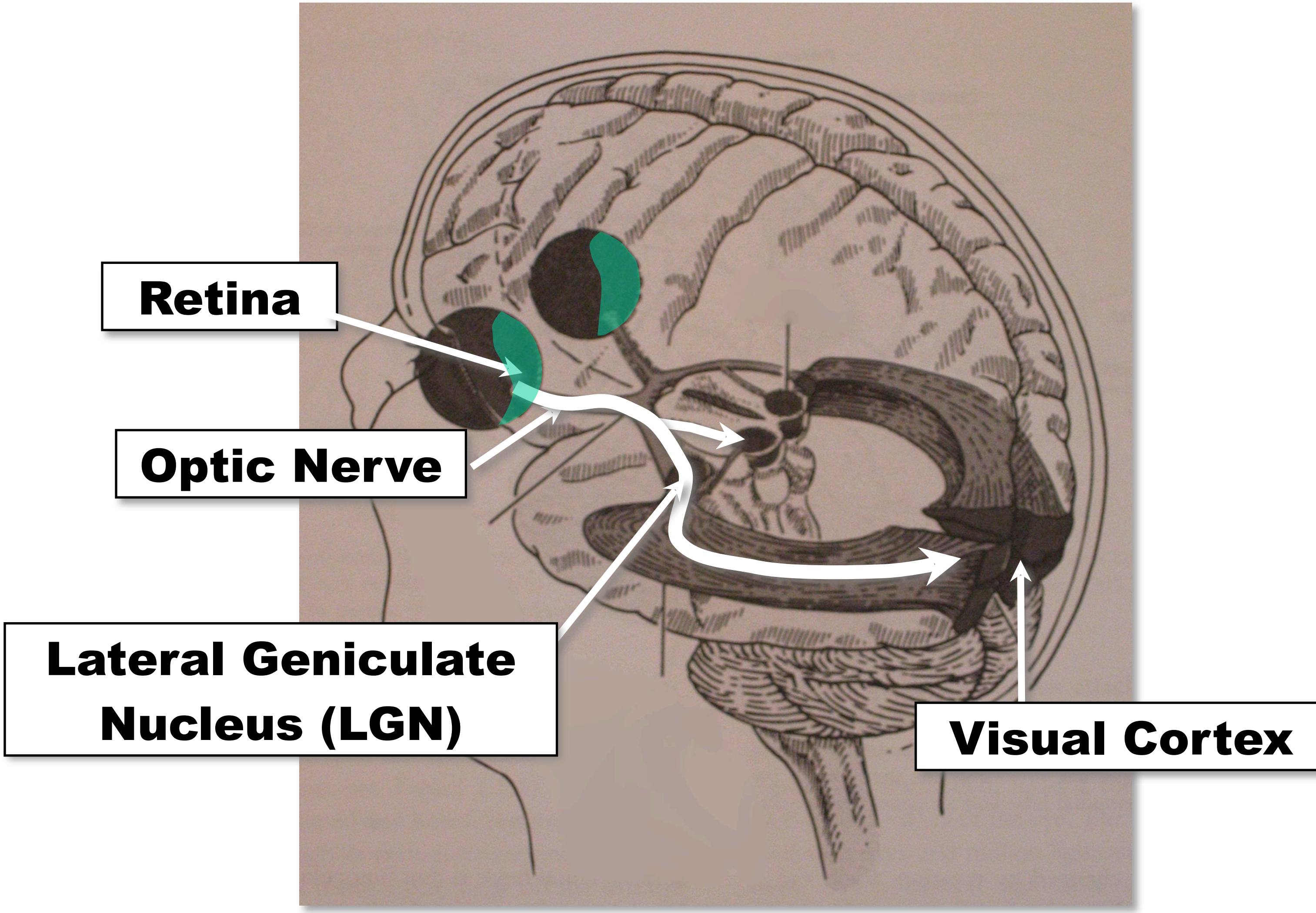
It Works!!

-->Proof of existence.

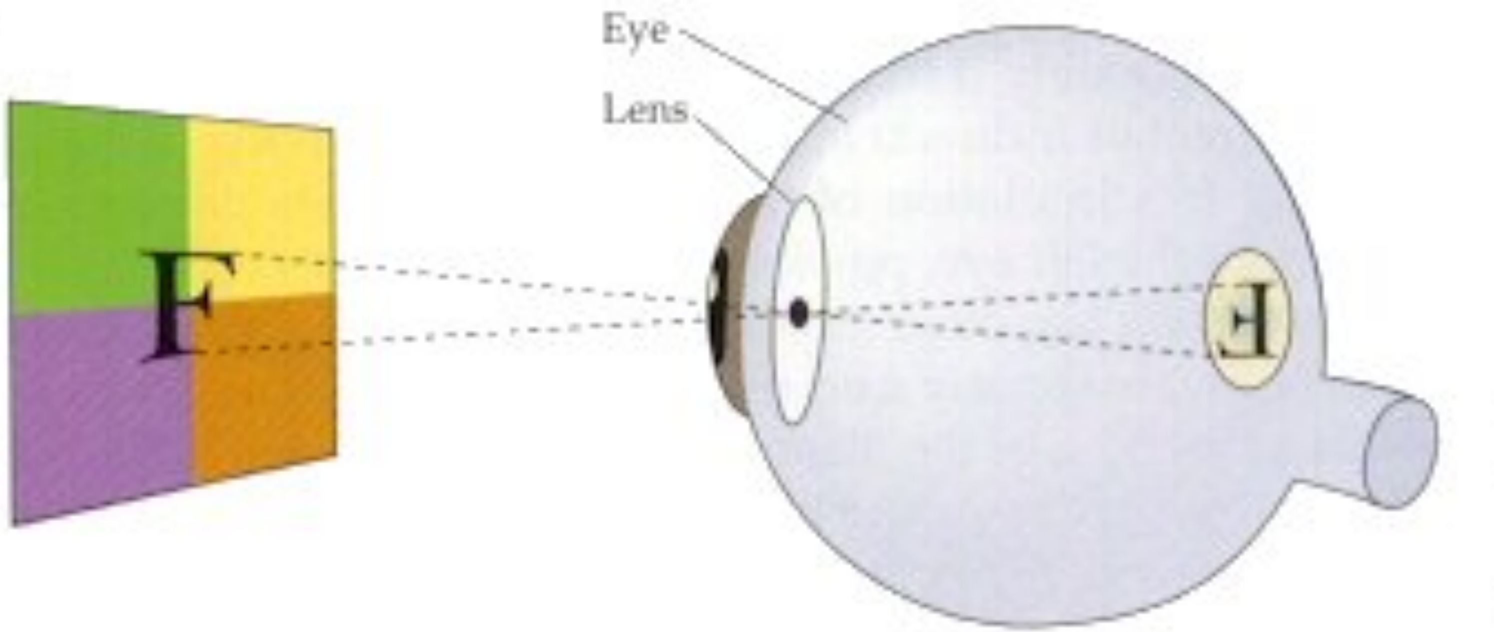
- The image formation process is well understood
- The image understanding one remains mysterious

EPFL

Reminder: Pathways To The Brain

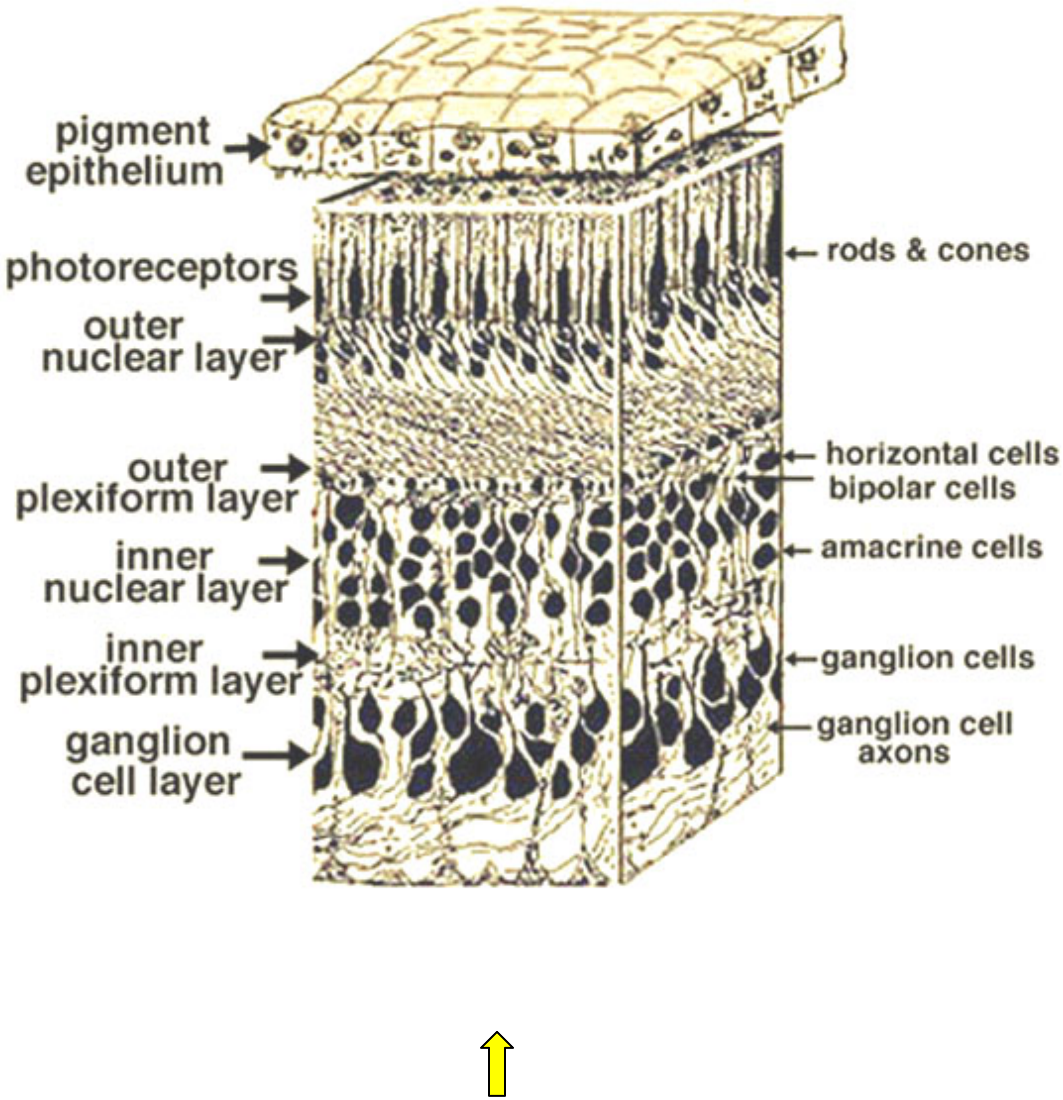


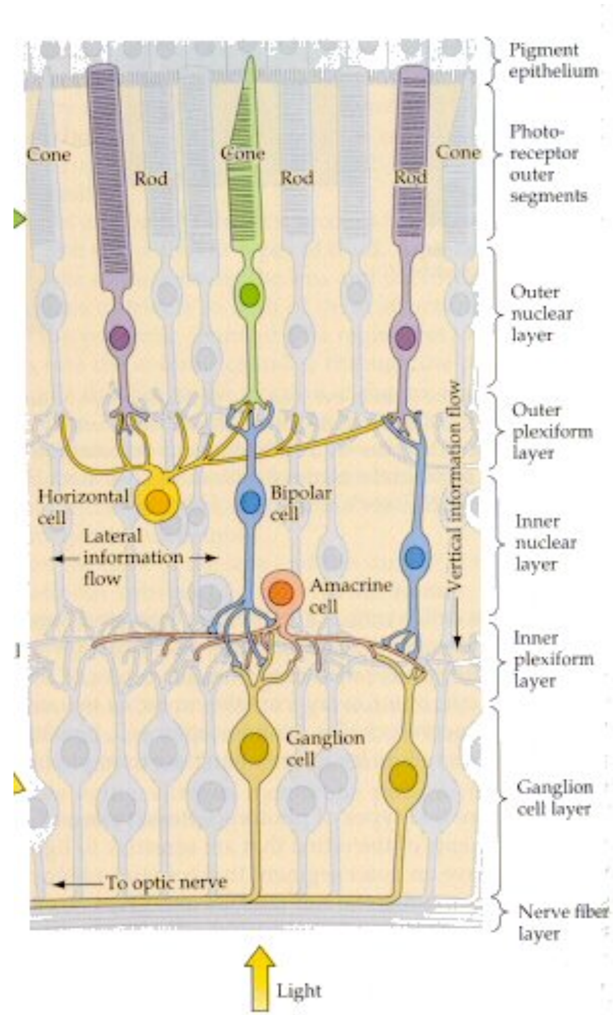
Reminder: Image Formation



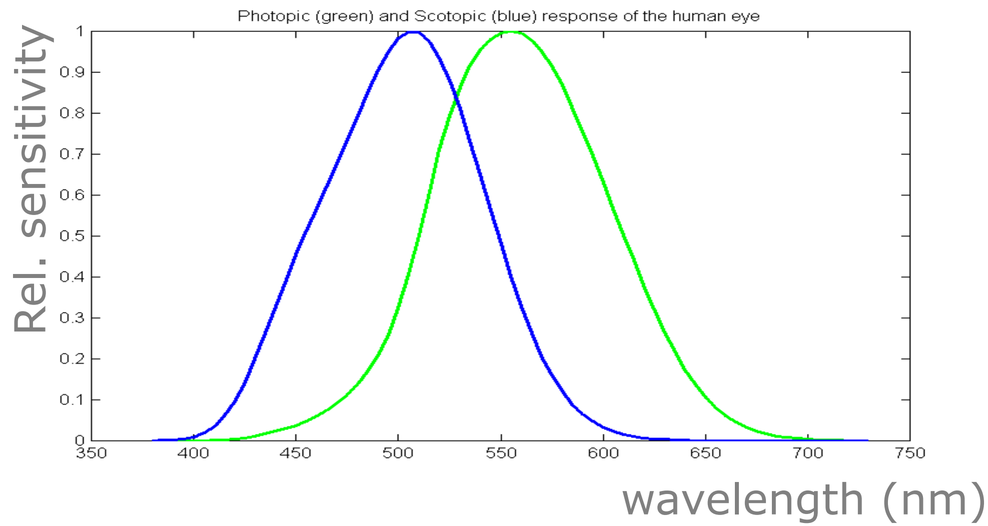
An inverted image forms on the retina.

Reminder: Retina





Reminder: Scotopic vs Photopic



Low luminance (< 1 cd/m²):

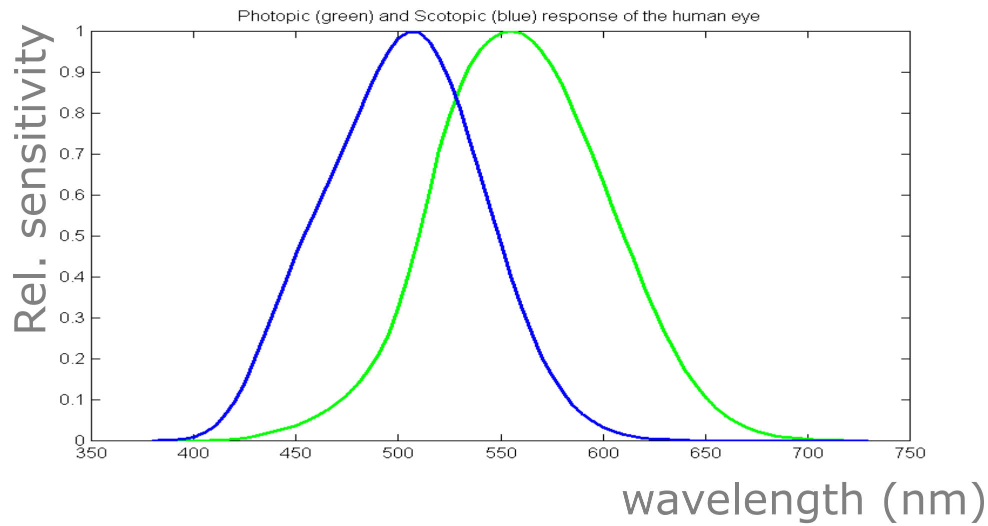
- 120 million rods with peak spectral response around 510 nm.
- Primarily located outside the fovea.

High luminance (> 100 cd/m²):

- 7 million cones per retina.
- Primarily located in the fovea.
- Three types of cones (S, M, L) with peak spectral response at different nm.
- Ratio L:M:S ≅ 40:20:1

EPFL

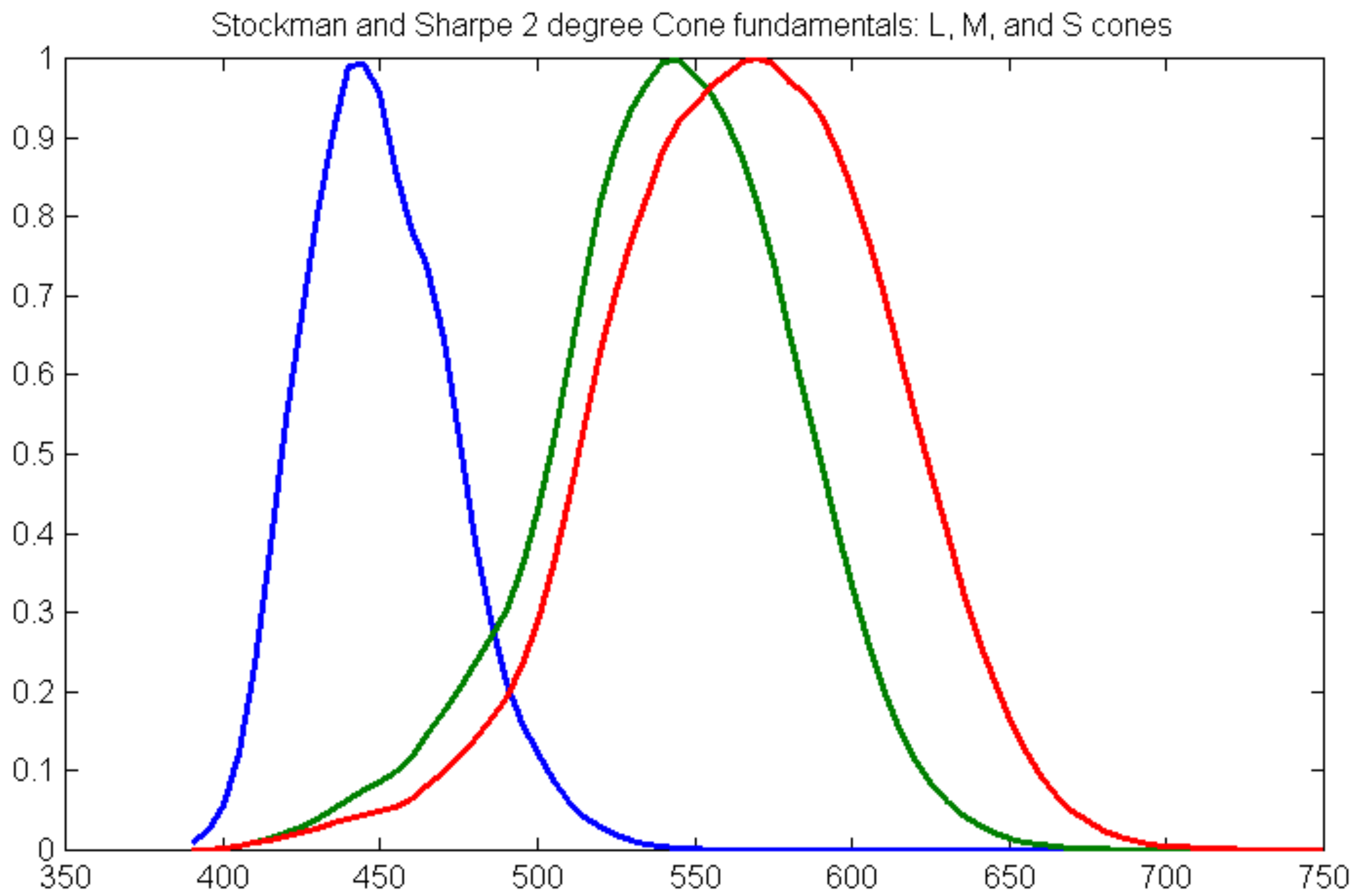
Reminder: Scotopic vs Photopic



Question: How was this data collected?

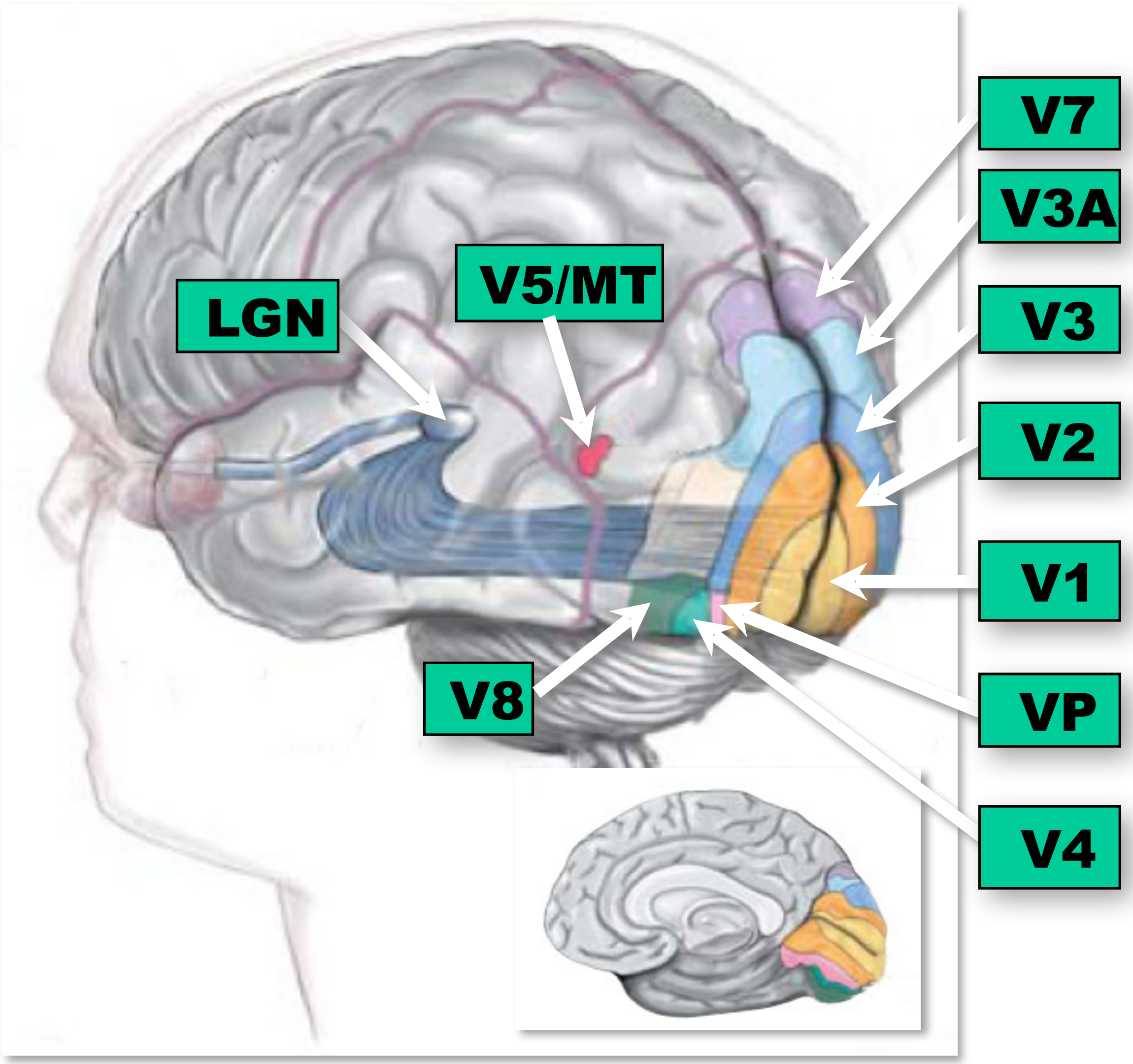
Answer (from S. Süsstrunk):
Psychophysical experiments with color-blind subjects
https://www.sciencedirect.com/science/article/pii/S0042698900000213

Reminder: Sensitivity to Different Wavelengths



Question: What do S, M, L stand for?
Answer: Short, Medium, Long wavelengths

Reminder: Visual Cortex



Logothetis, Scientific American'99

Reminder: Human vs Computer Vision

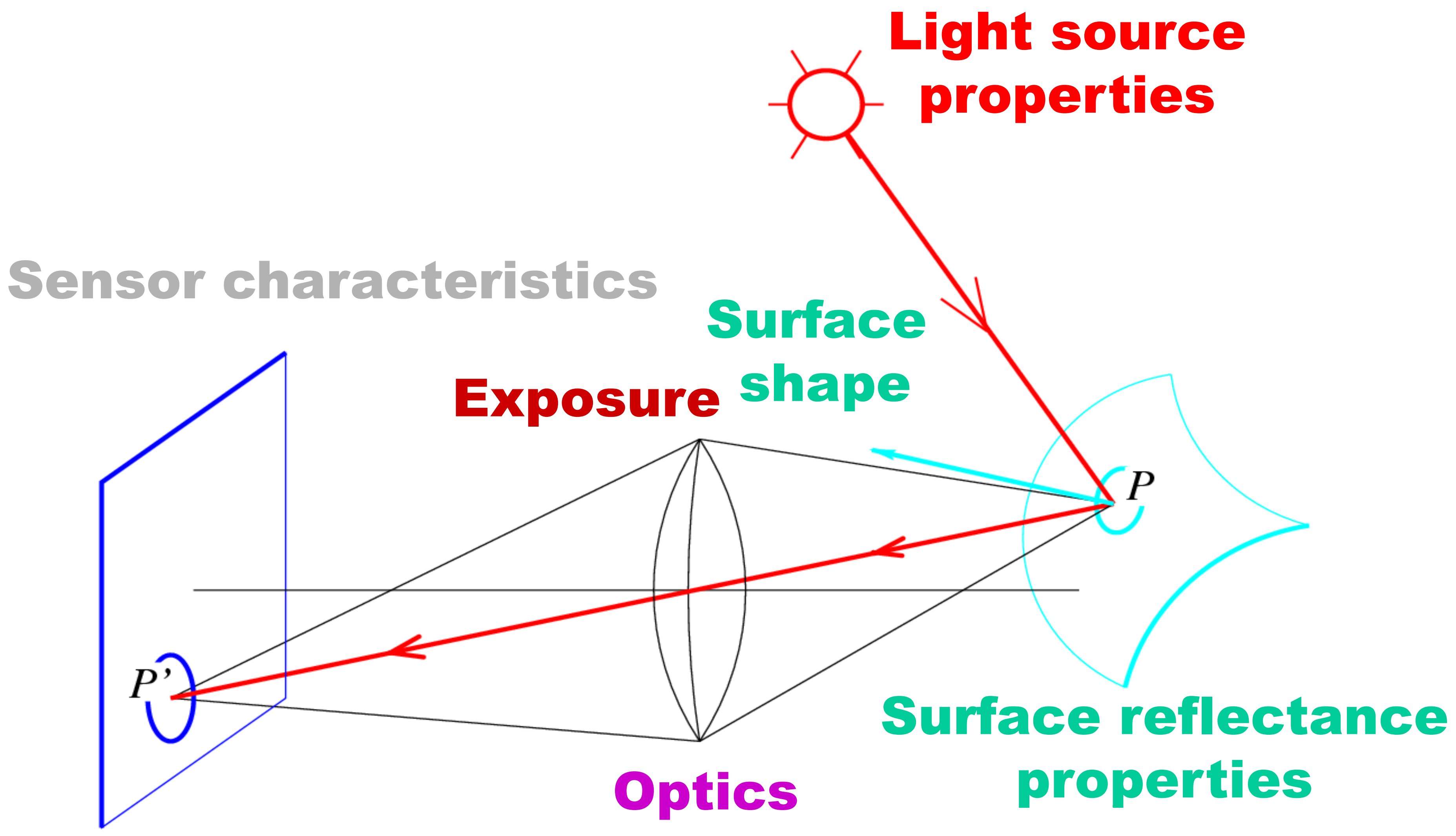
The camera replaces the eye:

- Eye lens -> Camera optics
- Cones and rods -> Sensor array
- Ganglion cells -> Filter banks

The computer replaces the brain:

But how?

Image Formation



Analog Images

An image can be understood as a 2D light intensity function f(x,y) where:

- x and y are spatial coordinates
- f(x, y) is proportional to the brightness or gray value of the image at that point.

→ Cannot be stored as such on a digital computer.

Digital Images

136	134	161	159	163	168	171	173	173	171	166	159	157	155
152	145	136	130	151	149	151	154	158	161	163	163	159	151
145	149	149	145	140	133	145	143	145	145	145	146	148	148
148	143	141	145	145	145	141	136	136	135	135	136	135	133
131	131	129	129	133	136	140	142	142	138	130	128	126	120
115	111	108	106	106	110	120	130	137	142	144	141	129	123
117	109	098	094	094	094	100	110	125	136	141	147	147	145
136	124	116	105	096	096	100	107	116	131	141	147	150	152
152	152	137	124	113	108	105	108	117	129	139	150	157	159
159	157	157	159	135	121	120	120	121	127	136	147	158	163
165	165	163	163	163	166	136	131	135	138	140	145	154	163
166	168	170	168	166	168	170	173	145	143	147	148	152	159
168	173	173	175	173	171	170	173	177	178	151	151	153	156
161	170	176	177	177	179	176	174	174	176	177	179	155	157
161	162	168	176	180	180	180	182	180	175	175	178	180	180

A digitized image is one in which:

- Spatial and grayscale values have been made discrete.
- Intensities measured across a regularly spaced grid in x and y directions are sampled to
  - 8 bits (256 values) per point for grayscale,
  - 3x8 bits per point for color images.

They are stored as two-dimensional arrays of gray-level values. The array elements are called pixels and identified by their x, y coordinates.

Grayscale Images

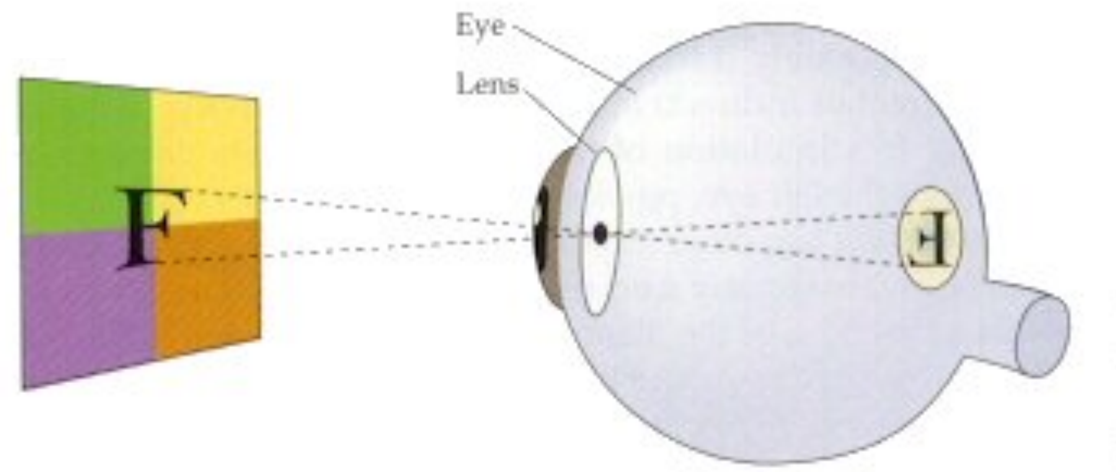
136	134	161	159	163	168	171	173	173	171	166	159	157	155
152	145	136	130	151	149	151	154	158	161	163	163	159	151
145	149	149	145	140	133	145	143	145	145	145	146	148	148
148	143	141	145	145	145	141	136	136	135	135	136	135	133
131	131	129	129	133	136	140	142	142	138	130	128	126	120
115	111	108	106	106	110	120	130	137	142	144	141	129	123
117	109	098	094	094	094	100	110	125	136	141	147	147	145
136	124	116	105	096	096	100	107	116	131	141	147	150	152
152	152	137	124	113	108	105	108	117	129	139	150	157	159
159	157	157	159	135	121	120	120	121	127	136	147	158	163
165	165	163	163	163	166	136	131	135	138	140	145	154	163
166	168	170	168	166	168	170	173	145	143	147	148	152	159
168	173	173	175	173	171	170	173	177	178	151	151	153	156
161	170	176	177	177	179	176	174	174	176	177	179	155	157
161	162	168	176	180	180	180	182	180	175	175	178	180	180

Color Images

A color image is often represented by three 8-bit images, one for red, one for green, and one for blue.

EPFL

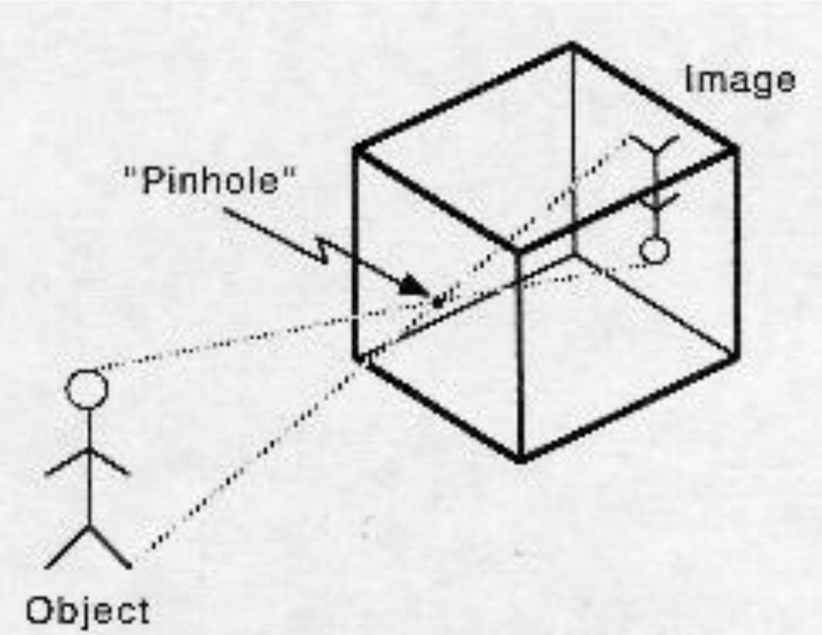
Image Formation



Projection from surfaces to 2-D sensor.

- Where: Geometry
- How bright: Radiometry
- Stored how: Sensing

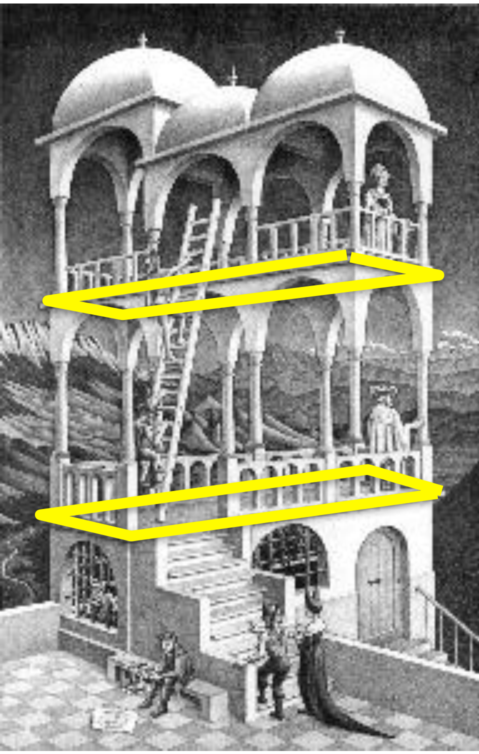
Pinhole Camera Model



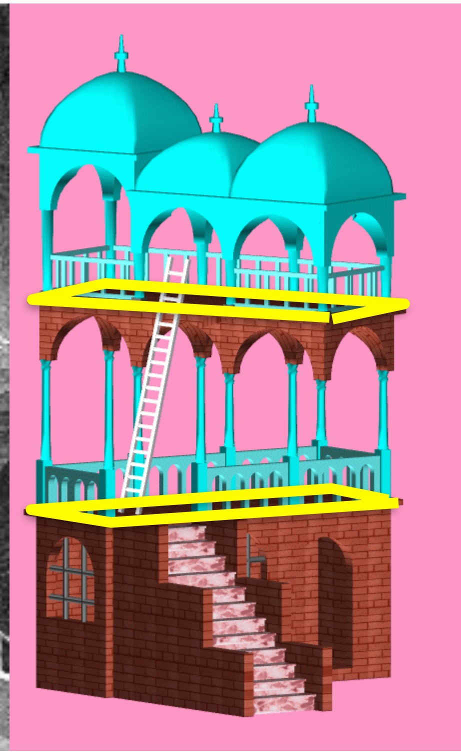
Idealized model of the perspective projection:

- All rays go through a hole and form a pencil of lines.
- The hole acts as a ray selector that allows an inverted image to form.

Escher's Belvedere



M. C. Escher



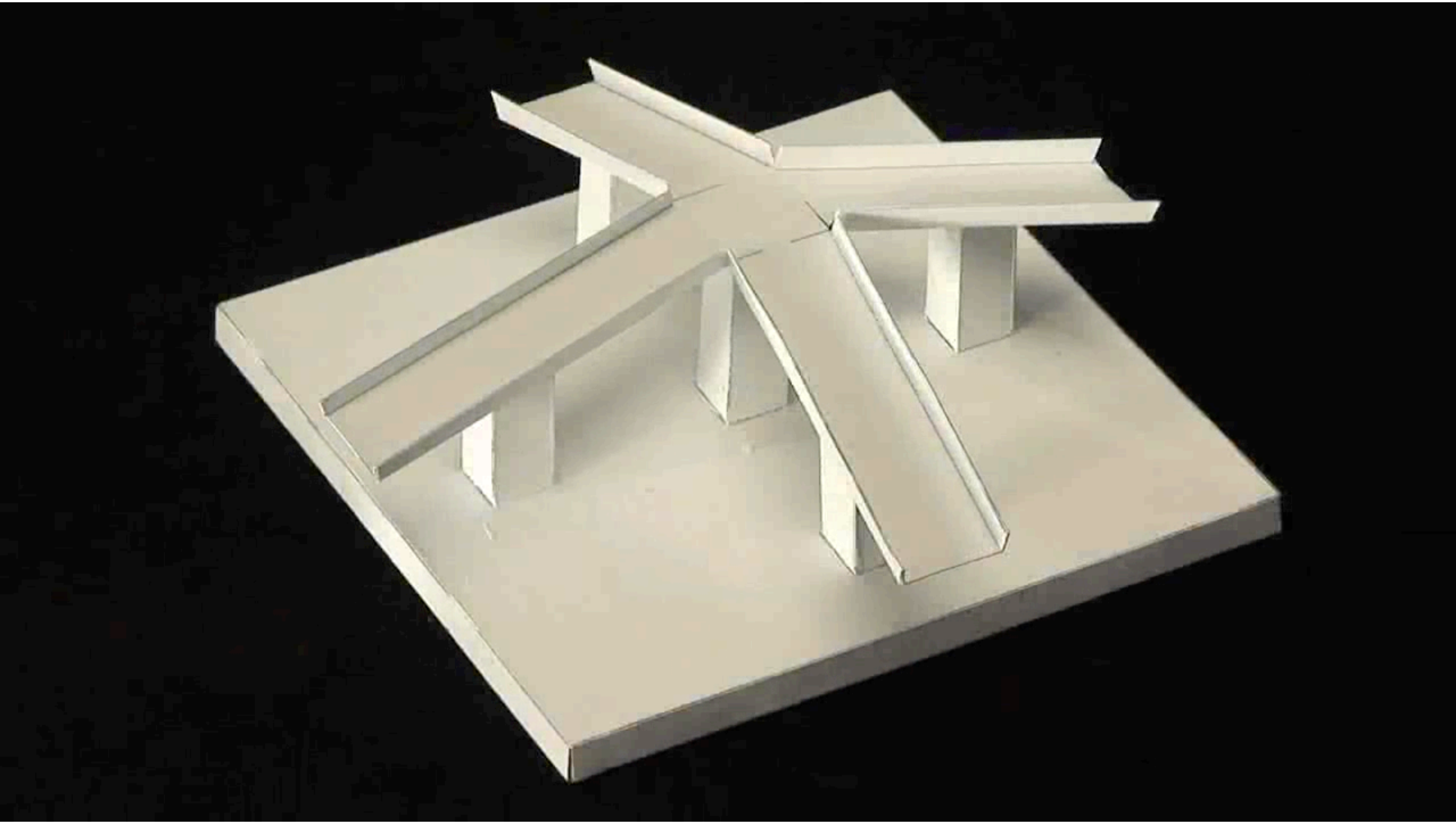
Impossible

Possible

Done

http://www.cs.technion.ac.il/~gershon/

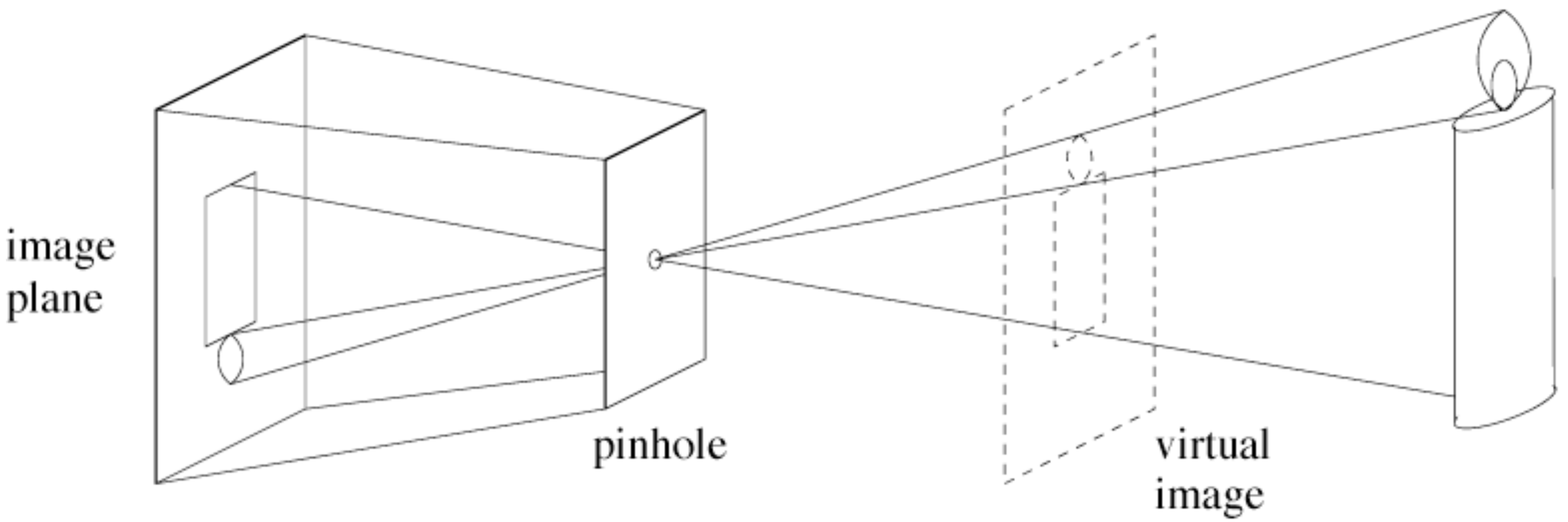
Magnet Like Slopes



Impossible Slopes by Kokichi Sugihara

http://www.isc.meiji.ac.jp/~kokichis/Welcomee.html

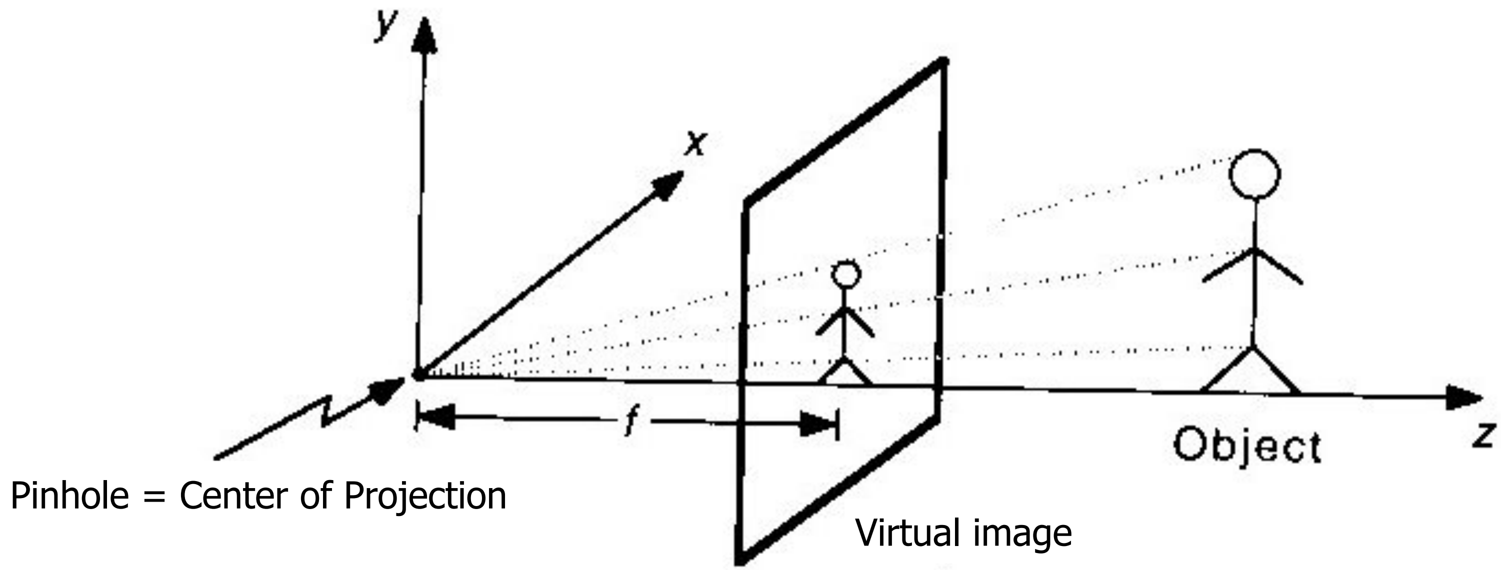
Virtual Image



- The real image forms on the image plane and is inverted.
- Let us consider a virtual plane in front of the camera.
- On this plane, we have a virtual non-inverted image.

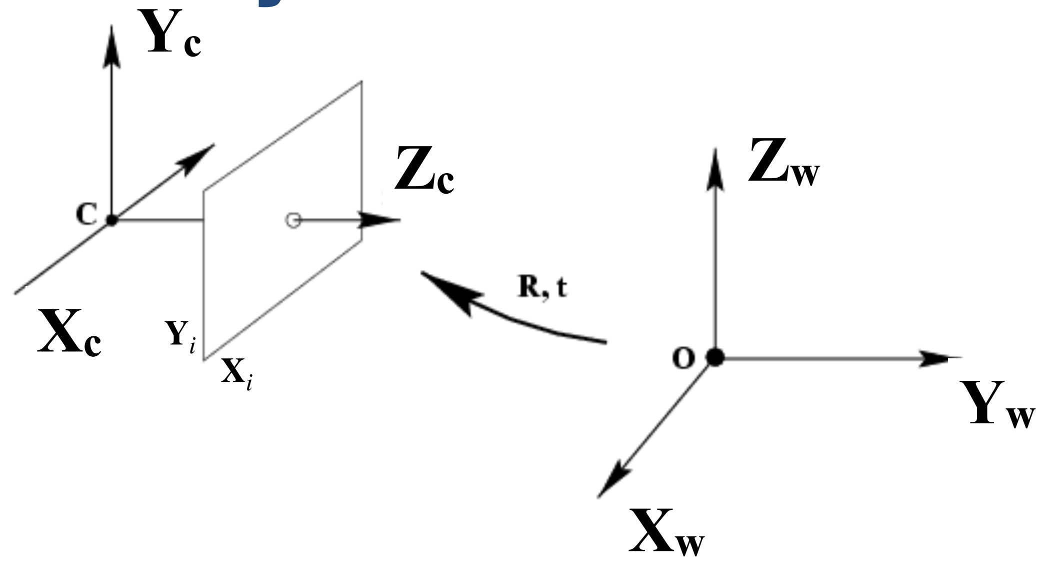
—> It is simpler to reason in terms of that virtual image.

Camera Geometry



From now on, we will use this formalism.

Coordinate Systems



Camera Coordinate System:

$$(X_c, Y_c, Z_c)$$

Image Coordinate System:

$$(X_i, Y_i, Z_i)$$

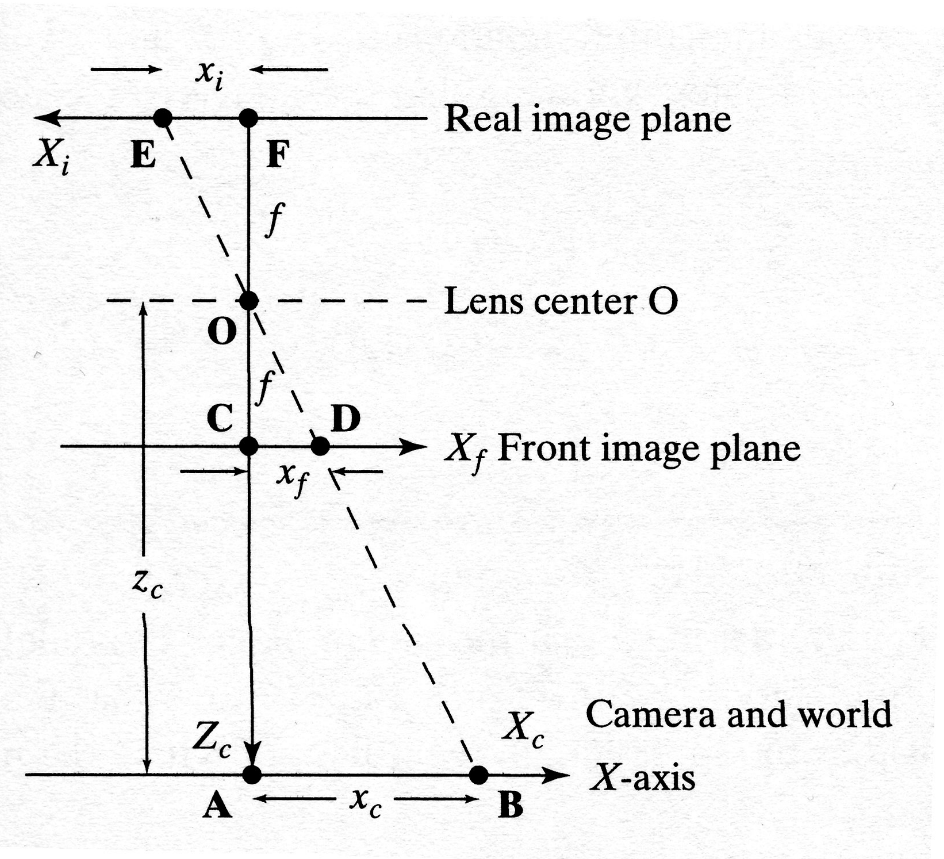
World Coordinate System:

$$(X_w, Y_w, Z_w)$$

Camera Coordinate System

- The center of the projection coincides with the origin of the world.
- The camera axis (optical axis) is aligned with the world's z-axis.
- To avoid image inversion, the image plane is in front of the center of projection.

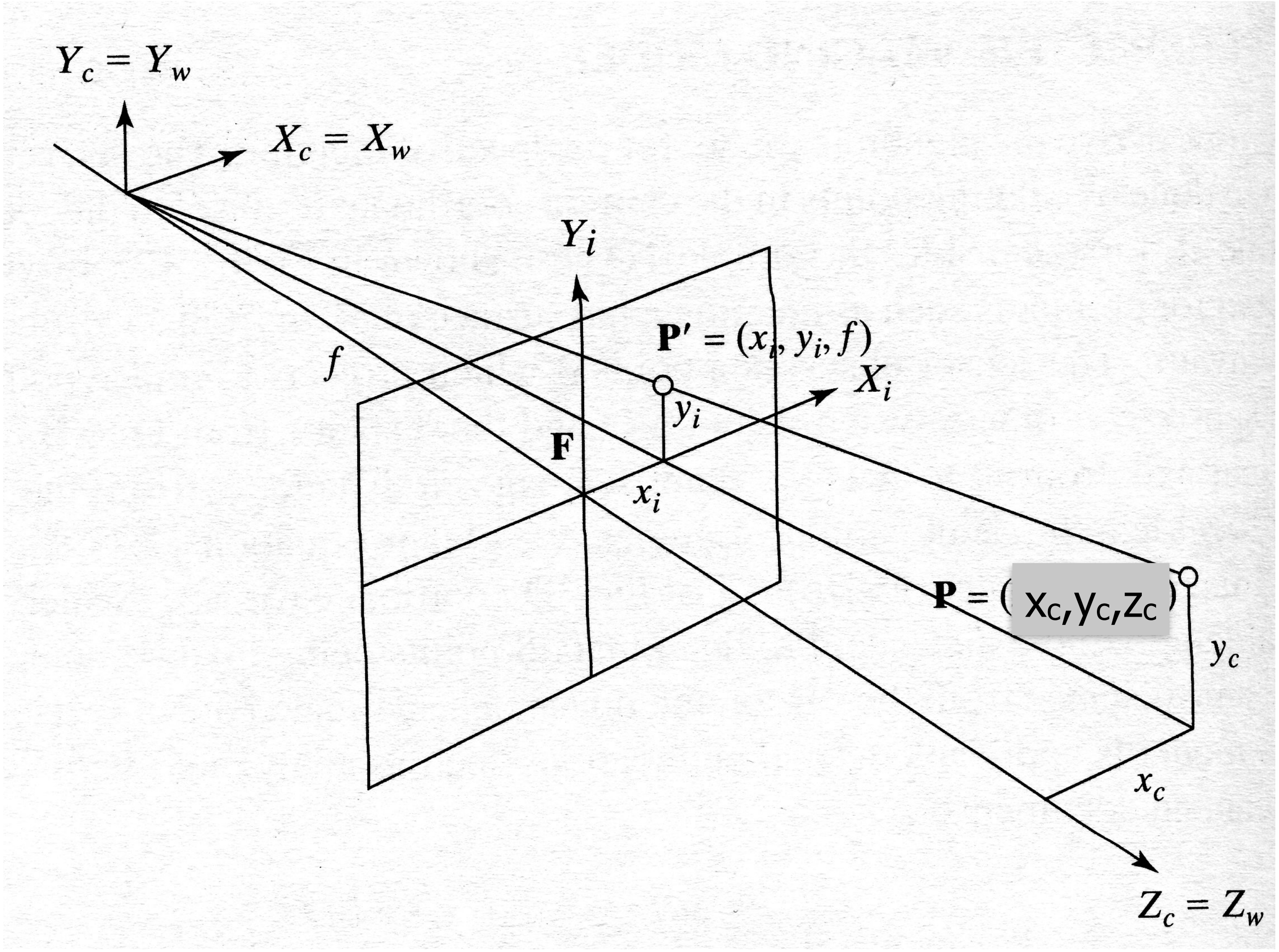
1D Image



$$\frac{x_f}{f} = \frac{x_c}{z_c}$$

$$\Rightarrow x_i = x_f = f\frac{x_c}{z_c}$$

2D Image

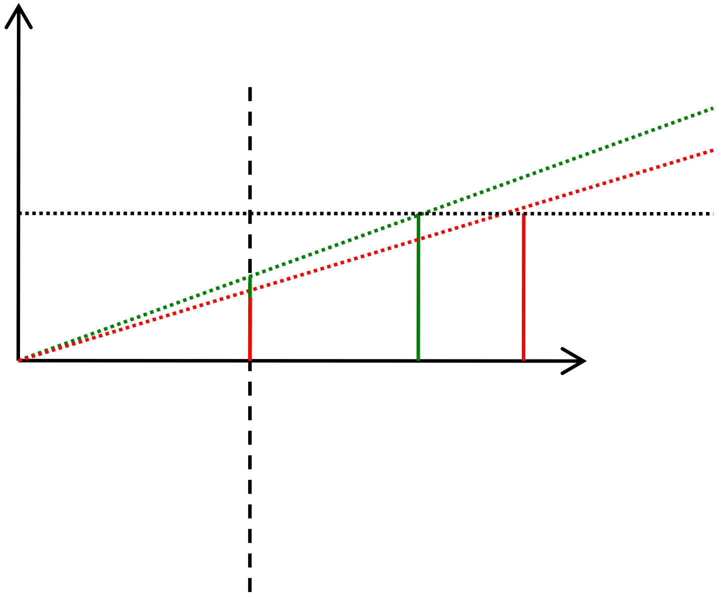


$$x_i = f\frac{x_c}{z_c}$$

$$y_i = f\frac{y_c}{z_c}$$

We dropped the distinction between (x_f, y_f) and (x_i, y_i).

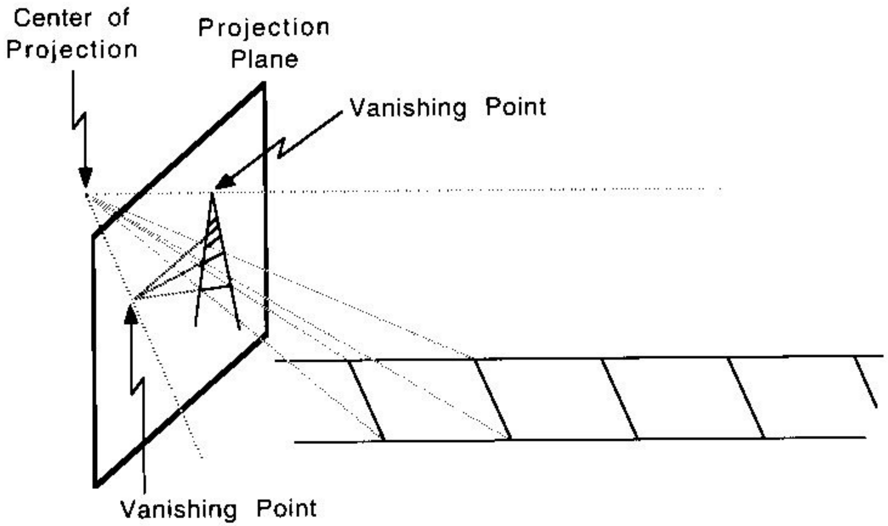
Distant Objects Appear Smaller



The green and red objects are of the same size but the red one is farther and therefore its projection smaller.

Because the car at the back has the same size in projection, we perceive it as being larger.

Projected Parallel Lines Meet



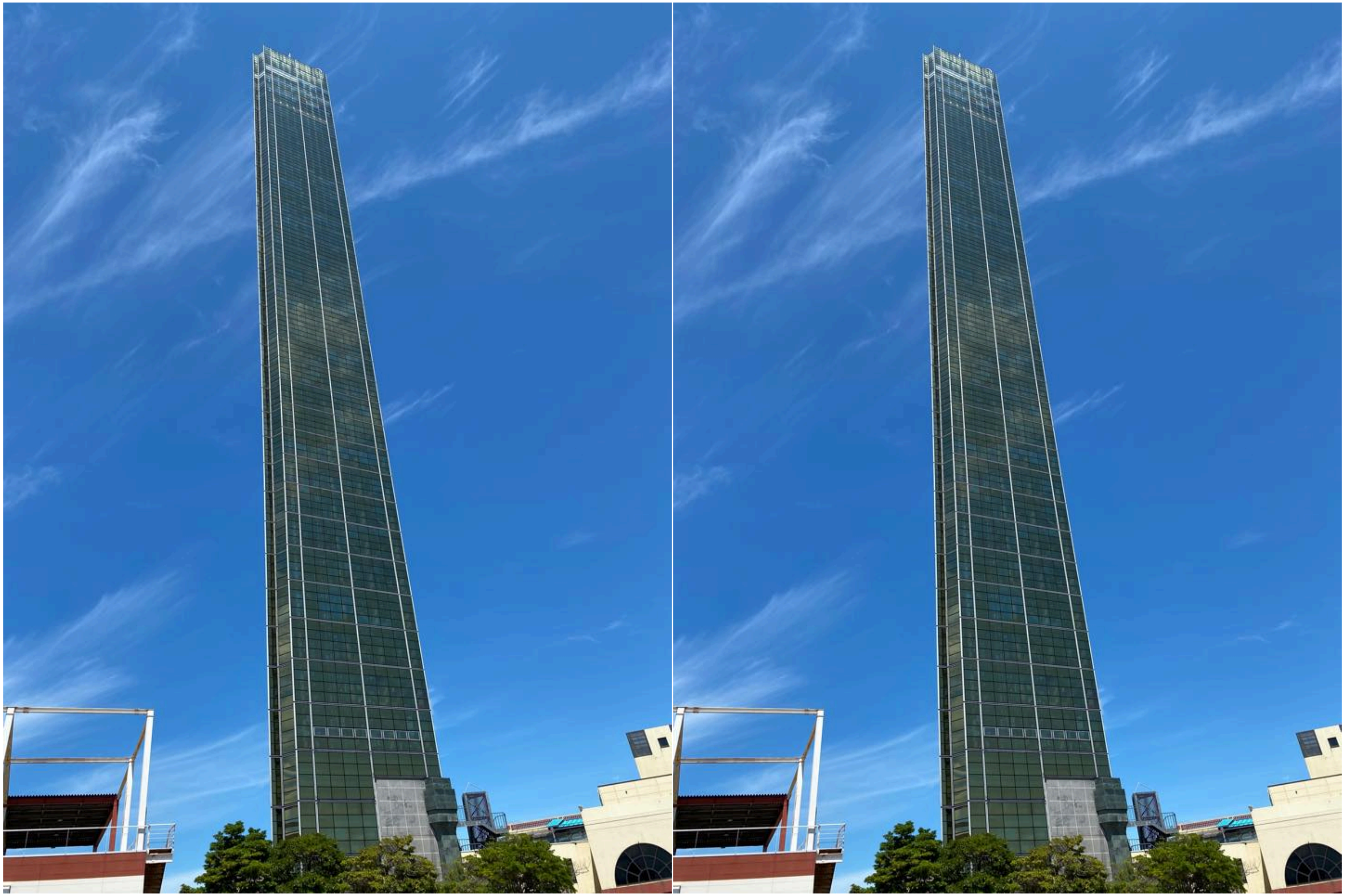
- Their intersection is referred to as the vanishing point.
- There is one per set of parallel 3D lines with the same direction vector.

Vanishing Points

- The projections of parallel lines all meet at one point, called the vanishing point.
- As the focal length increases, the vanishing point moves towards infinity.

EPFL

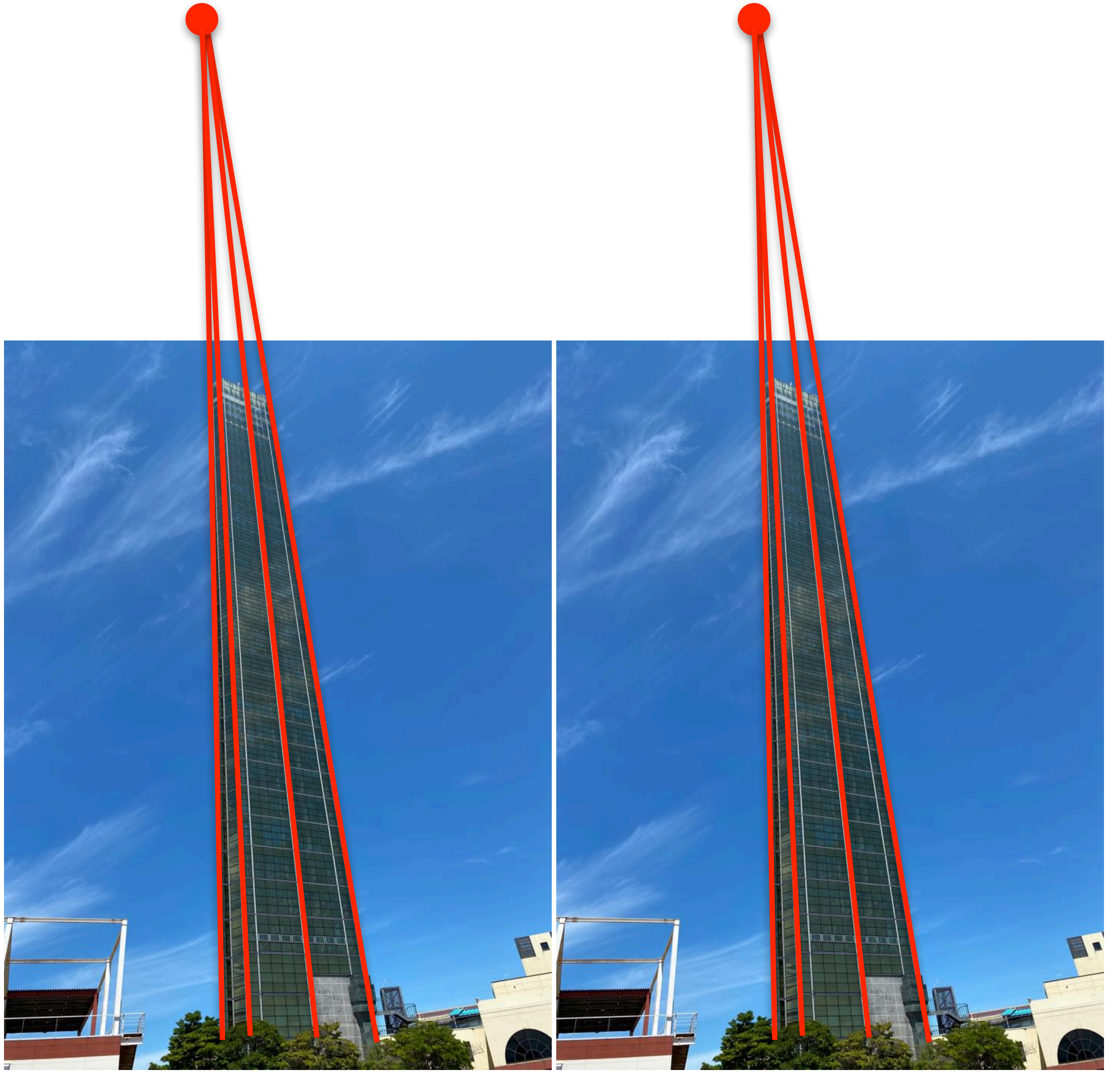
Leaning Towers



The two images are the same!

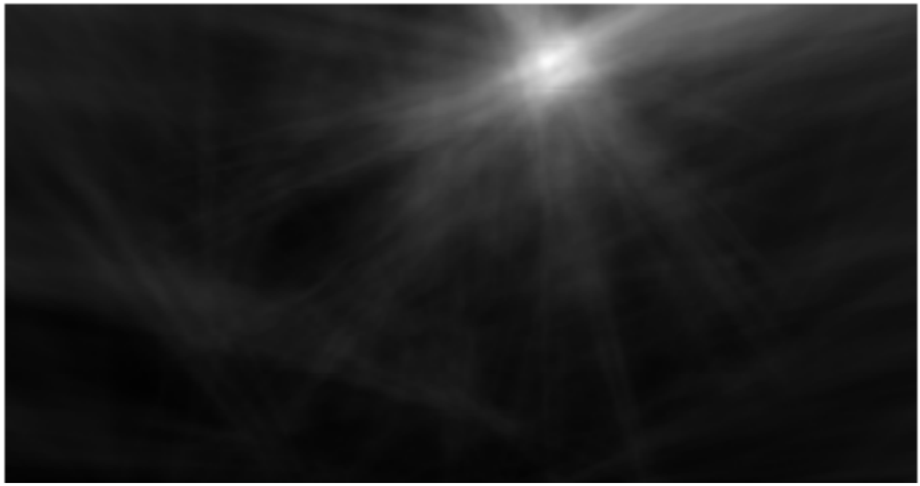
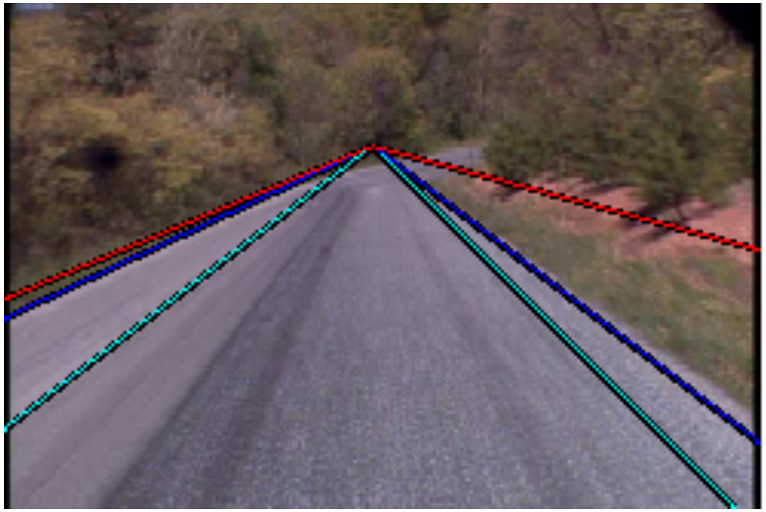
EPFL

Leaning Towers Explained?



There are two different vanishing points. Hence, the two towers are not perceived as being parallel.

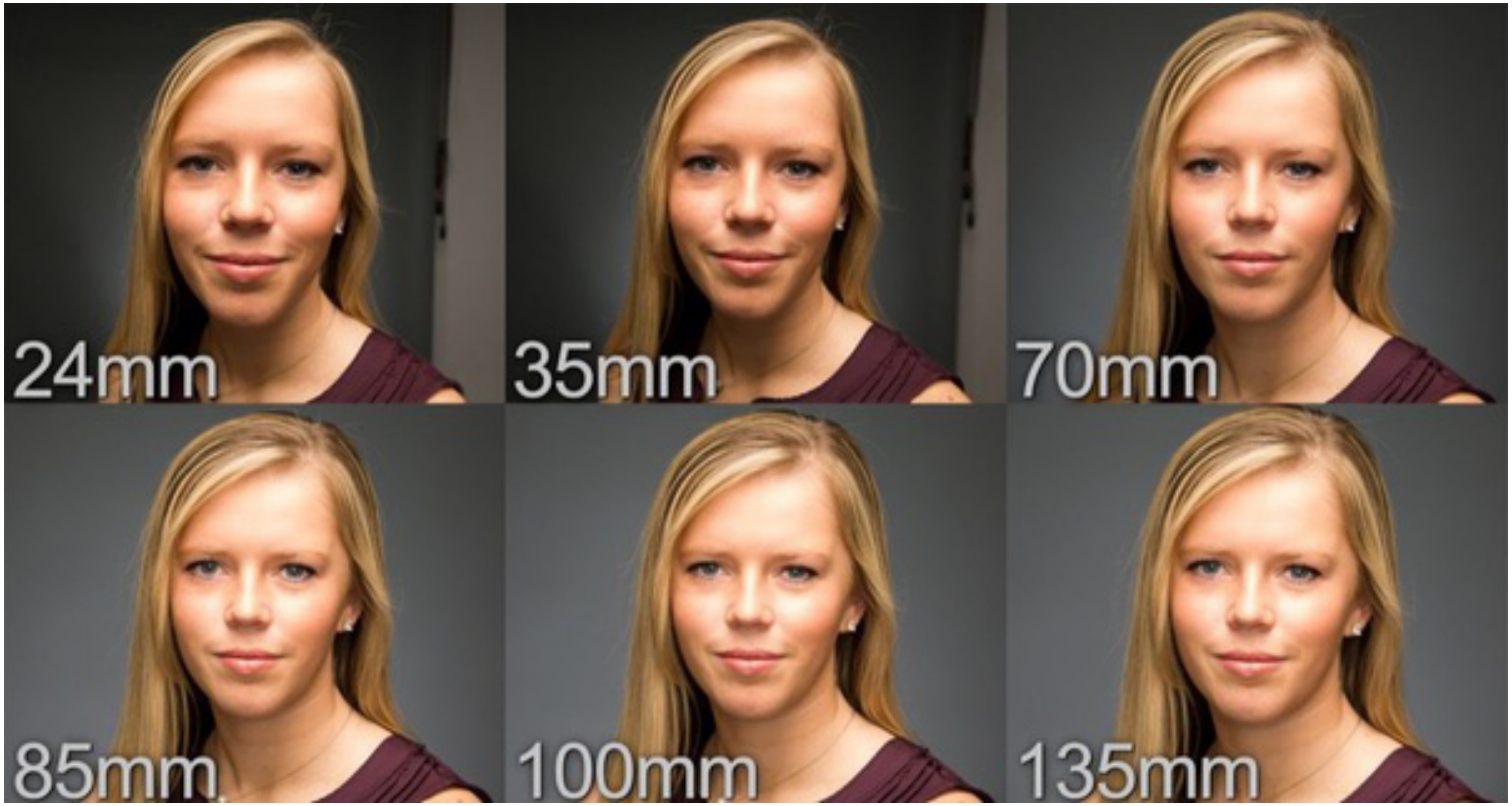
Road Following


The vanishing point can be computed and used to direct an autonomous car.

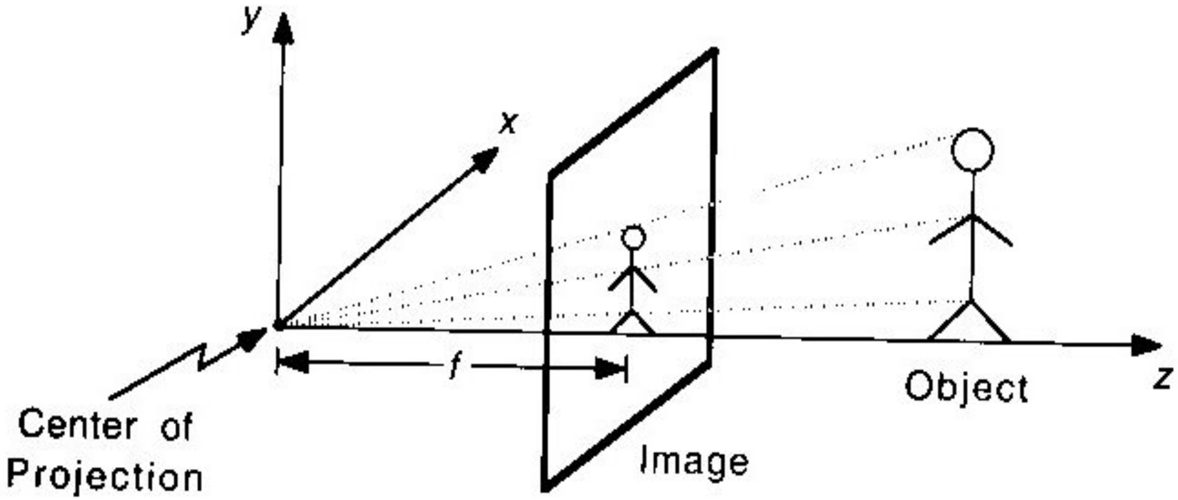
Rasmussen et al. , BMVC'04

Effect of Focal Length on Faces



- Professional portrait: From 85 to 100.
- Typical phone camera: From 24 to 35.

Projection is Non Linear



In pixels

$$u \propto x_i = f\frac{x_c}{z_c}$$

$$v \propto y_i = f\frac{y_c}{z_c}$$

In meters

→Reformulate it as a linear operation using homogeneous coordinates.

Homogeneous Coordinates

- Homogeneous representation of 2D point:

$$\mathbf{x} = (x_1, x_2, x_3) \text{ represents } (x_1/x_3, x_2/x_3)$$

- Homogeneous representation of 3D point:

$$\mathbf{X} = (x_1, x_2, x_3, x_4) \text{ represents } (x_1/x_4, x_2/x_4, x_3/x_4)$$

- Scale invariance:

X and l**X** represent the same point, same for **x** and l**x**.

Simple Projection Matrix

2D point expressed in projective coordinates.

3D point expressed in projective coordinates.

Let us write:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}$$

- $[x, y, z]^T$ represents

$$X_i = \frac{x}{z} = f \frac{X_c}{Z_c}$$

$$Y_i = \frac{y}{z} = f \frac{Y_c}{Z_c}$$

- Therefore $[x, y, z]^T$ is the projection of $[X_c, Y_c, Z_c, 1]^T$.

—> We have expressed the projection of a 3D point as the multiplication of its projective coordinates by a projection matrix.

Intrinsic And Extrinsic Parameters

$$\begin{bmatrix} \mathrm{x} \\ \mathrm{y} \\ \mathrm{z} \end{bmatrix} = \begin{bmatrix} \text{Matrix of} \\ \text{intrinsic parameters} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \text{Matrix of} \\ \text{extrinsic parameters} \end{bmatrix} \begin{bmatrix} \mathrm{X} \\ \mathrm{Y} \\ \mathrm{Z} \\ 1 \end{bmatrix}$$

- Camera may not be at the origin, looking down the z-axis
  - → Extrinsic parameters
- One unit in image coordinates may not be the same as one unit in world coordinates
  - → Intrinsic parameters

Complete Linear Camera Model

2D point expressed in projective coordinates and in pixels.

$$\begin{bmatrix} \mathrm{x} \\ \mathrm{y} \\ \mathrm{z} \end{bmatrix} = \begin{bmatrix} \text{Matrix of} \\ \text{intrinsic parameters} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \text{Matrix of} \\ \text{extrinsic parameters} \end{bmatrix} \begin{bmatrix} \mathrm{X} \\ \mathrm{Y} \\ \mathrm{Z} \\ 1 \end{bmatrix}$$

$$= \mathbf{K} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \mathbf{Rt} \begin{bmatrix} \mathrm{X} \\ \mathrm{Y} \\ \mathrm{Z} \\ 1 \end{bmatrix}$$

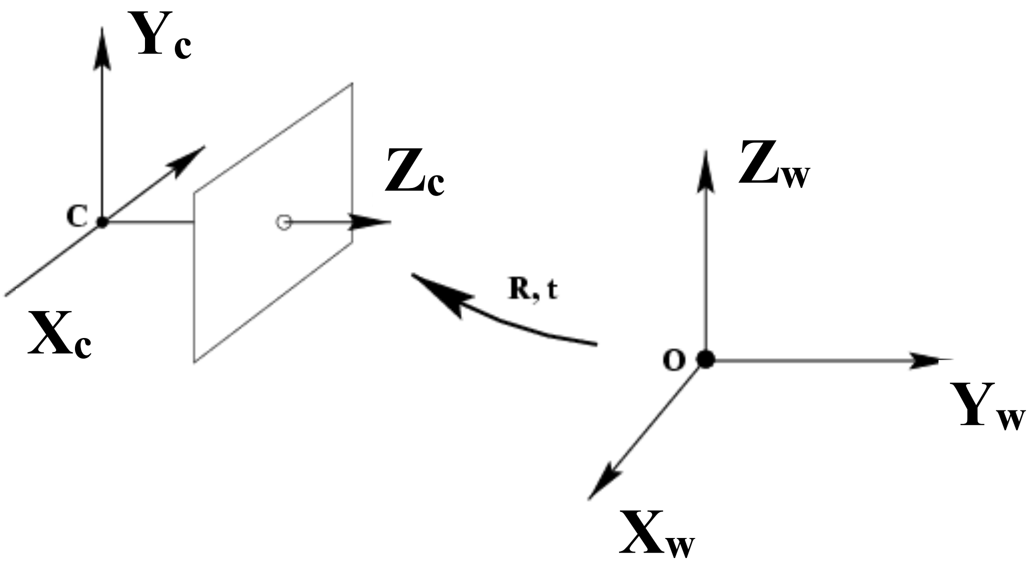
where $\mathbf{K}$ is a 3×3 matrix and $\mathbf{Rt}$ a 4×4 matrix.

3D point expressed in projective coordinates using the world coordinate system.

EPFL

Matrix of Extrinsic Parameters

It converts world coordinates into camera coordinates:



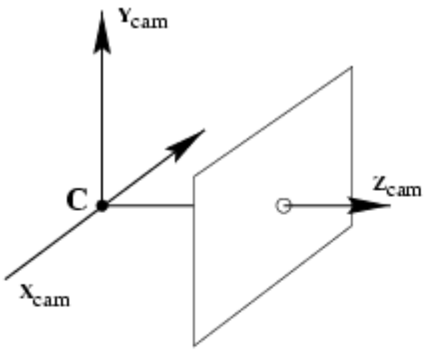
$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = \mathbf{R}\left(\begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} - \tilde{\mathbf{C}}\right) \text{ with } \mathbf{R}^t\mathbf{R} = \mathbf{I}$$

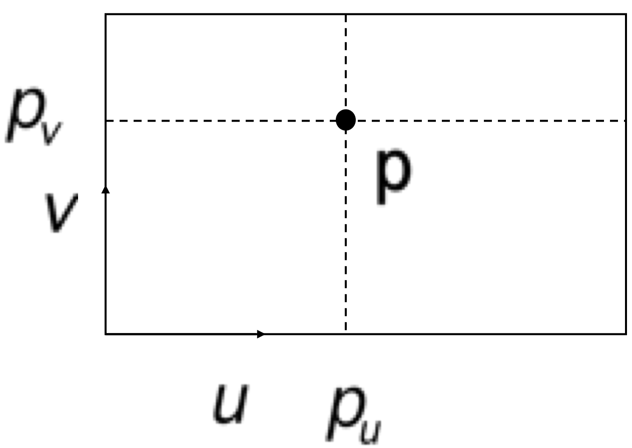
$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \text{ with } \mathbf{T} = -\mathbf{R}\tilde{\mathbf{C}}$$

→Rotations and translations also expressed in terms of matrix multiplications in projective space.

Matrix of Intrinsic Parameters

It converts image coordinates into pixels:





$$
\begin{aligned}
u &= X_i + p_u = fX/Z + p_u \\
v &= Y_i + p_v \ = fY/Z + p_v
\end{aligned}
$$

$$
\mathbf{K} = \begin{bmatrix} f & 0 & p_u \\ 0 & f & p_v \\ 0 & 0 & 1 \end{bmatrix}
$$

Principal point: p

Matrix of Intrinsic Parameters

It converts image coordinates into pixels:

$$
\begin{aligned}
u &= \alpha_u X_i + p_u = \alpha_u X/Z + p_u \\
v &= \alpha_v Y_i + p_v = \alpha_v Y/Z + p_v
\end{aligned}
$$

$$
\mathbf{K} = \begin{bmatrix} \alpha_u & 0 & p_u \\ 0 & \alpha_v & p_v \\ 0 & 0 & 1 \end{bmatrix}
$$

The pixels are not necessarily square, must account for different scaling in x and y.

Matrix of Intrinsic Parameters

It converts image coordinates into pixels:

$$\mathbf{K} = \begin{bmatrix} \alpha_u & \mathrm{s} & p_u \\ 0 & \alpha_v & p_v \\ 0 & 0 & 1 \end{bmatrix}$$

s encodes the non-orthogonality of the u and v directions. It is very close to zero in modern cameras.

Putting it All Together

Projection Matrix

2D projection

3D point

$$\mathbf{x} = \mathbf{P}\mathbf{X}$$

$$\mathbf{P} = \mathbf{K}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\mathbf{Rt}$$

with $\mathbf{K} = \begin{bmatrix} \alpha_u & s & p_u \\ 0 & \alpha_v & p_v \\ 0 & 0 & 1 \end{bmatrix}$, $\mathbf{Rt} = \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{0} & 1 \end{bmatrix}$, and $\mathbf{R}^T\mathbf{R} = \mathbf{I}$.

Intrinsics

Extrinsics

Hartley, Chap 6.

Another Way to Write the Projection Matrix

Projection Matrix

2D projection 3D point

$$\mathbf{x} = \mathbf{P}\mathbf{X}$$

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & 1 \end{bmatrix}$$

- In projective geometry, λ **x** = **x**. Therefore the matrix **P** can always be rescaled so that its last element is one.
- The 3x4 matrix **P** has 11 degrees of freedom.

Hartley, Chap 6.

Camera Calibration

Internal Parameters:

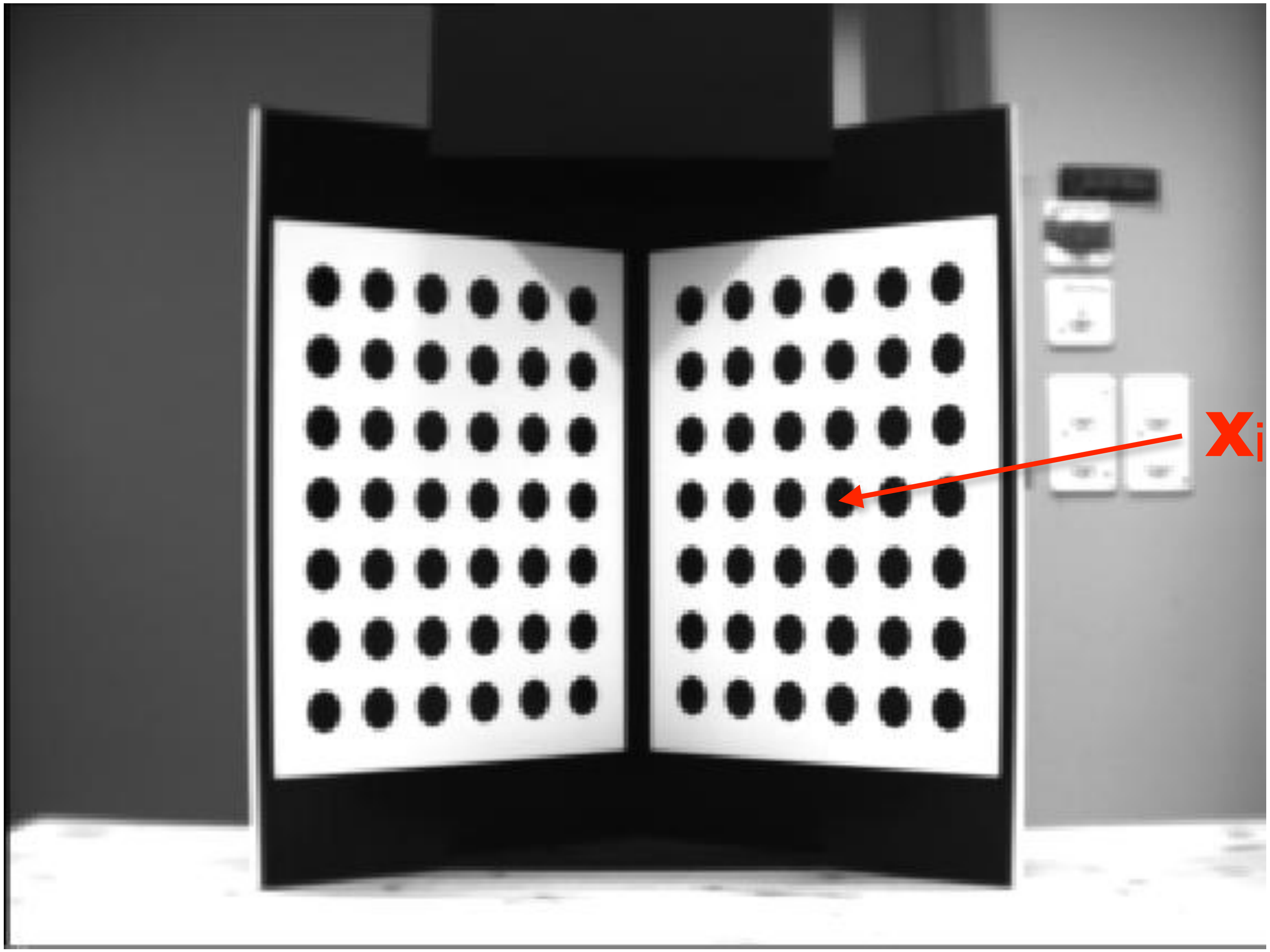
- Horizontal and vertical scaling (2)
- Principal points (2)
- Skew of the axis (1)

External Parameters:

- Rotations (3)
- Translations (3)

→There are 11 free parameters to estimate. This is known as **calibrating** the camera.

Calibration Grid



$$\mathbf{x}_i = \mathbf{P}\mathbf{X}_i$$

One way to calibrate: Take a picture of a calibration grid.

EPFL

Estimating the Camera Parameters

- Number of measurements required:
  - 11 degrees of freedom.
  - 2 constraints per correspondence.

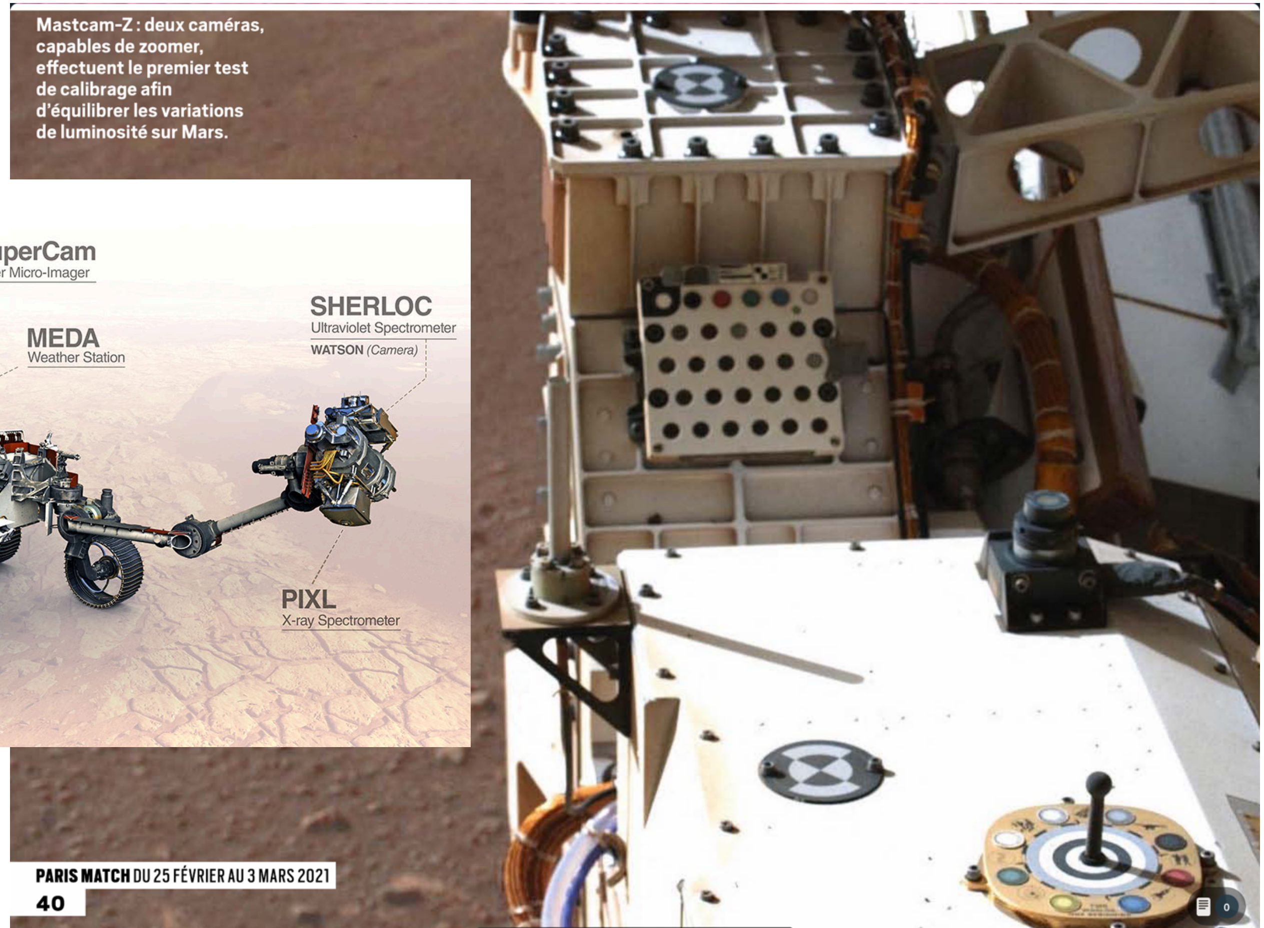
- Direct linear transform:
  - Minimal solution for 6 correspondences
  - Over-constrained solutions by imposing

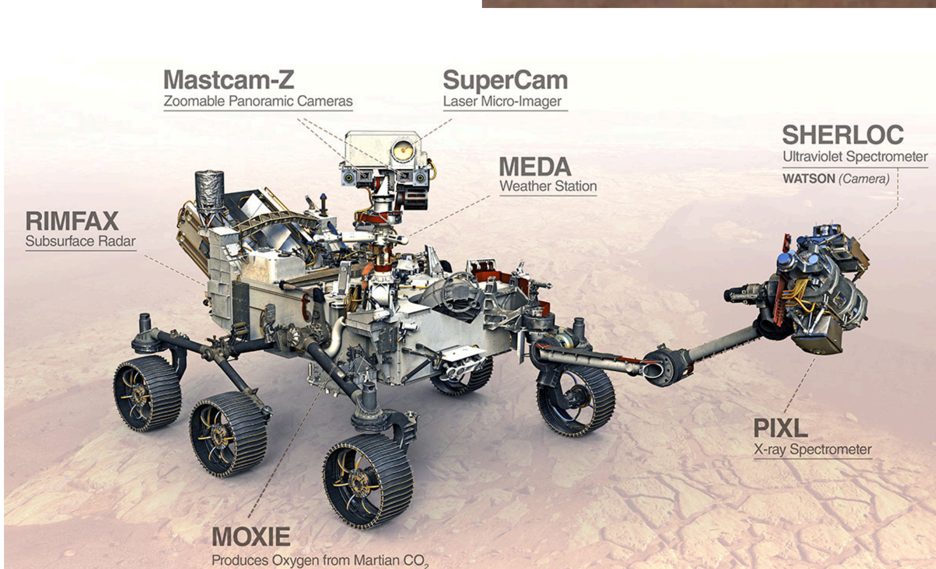
  For all i, $\mathbf{x}_i = \mathbf{P}\mathbf{X}_i$, with $\|P\| = 1$ or $P_{34} = 1$

- Non linear optimization.

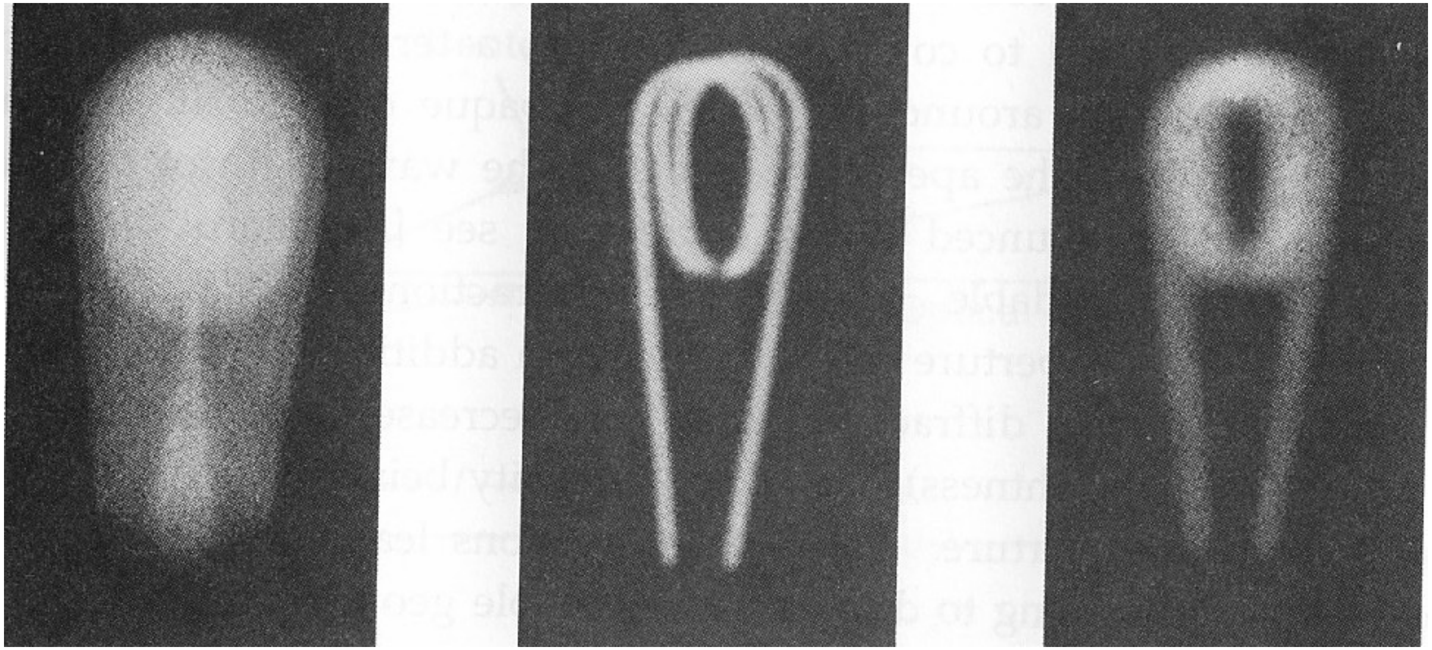
Hartley, Chap 7.

Martian Calibration





Limitations of the Pinhole Model

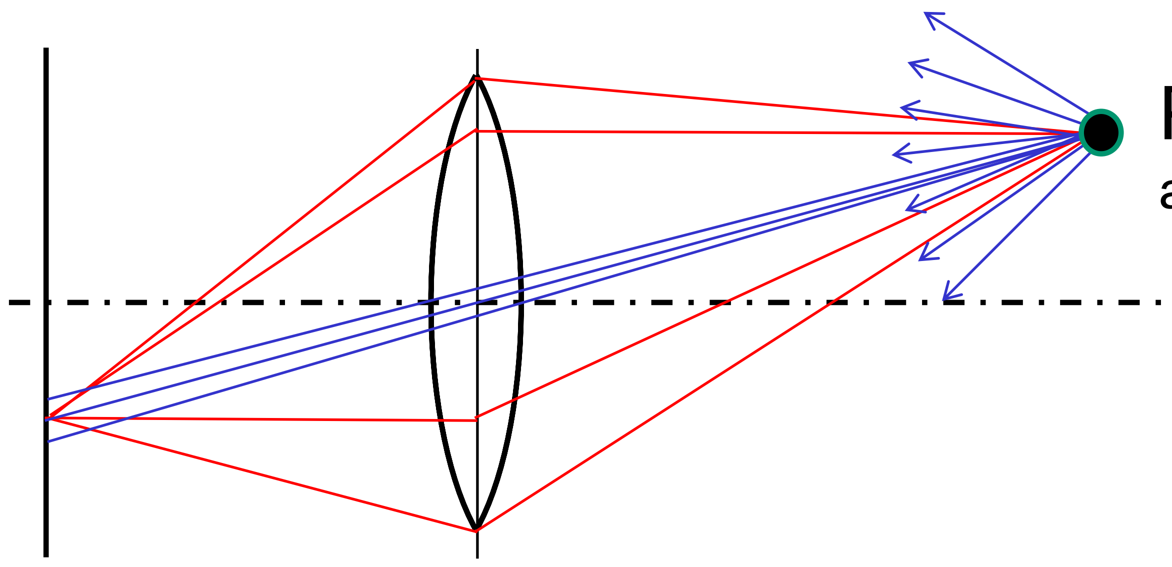


Idealization because the hole cannot be infinitely small

- Image would be infinitely dim
- Diffraction effects

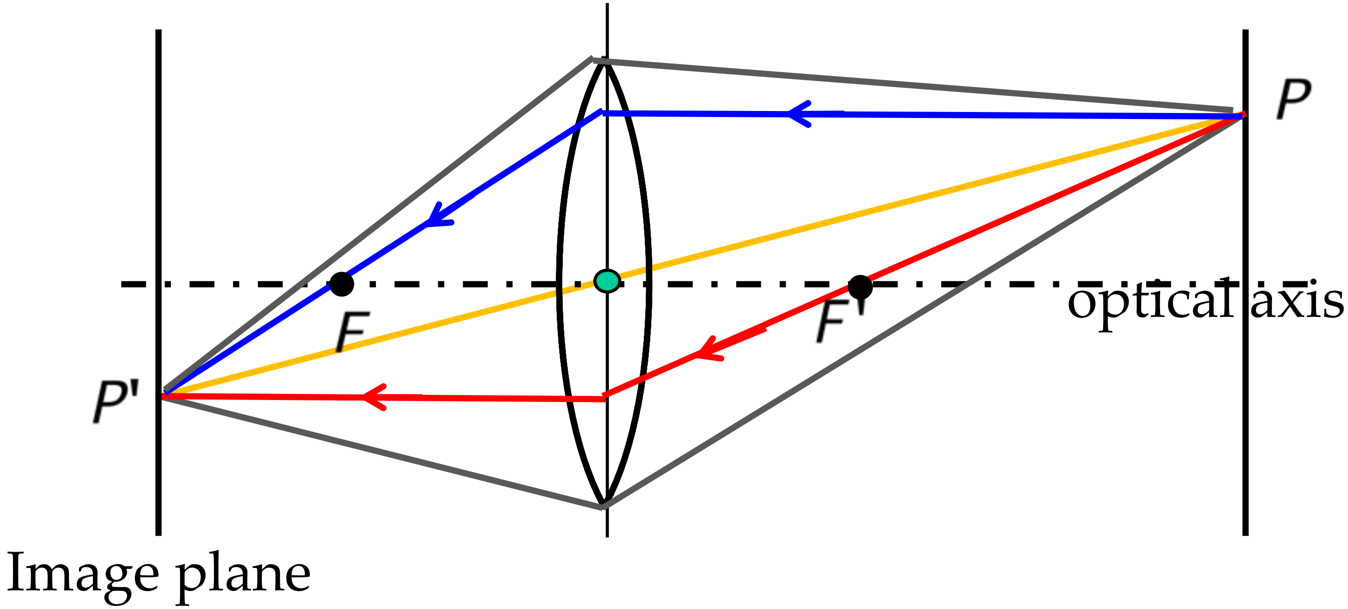
→ Use a lens to overcome this problem.

Imaging With a Lens



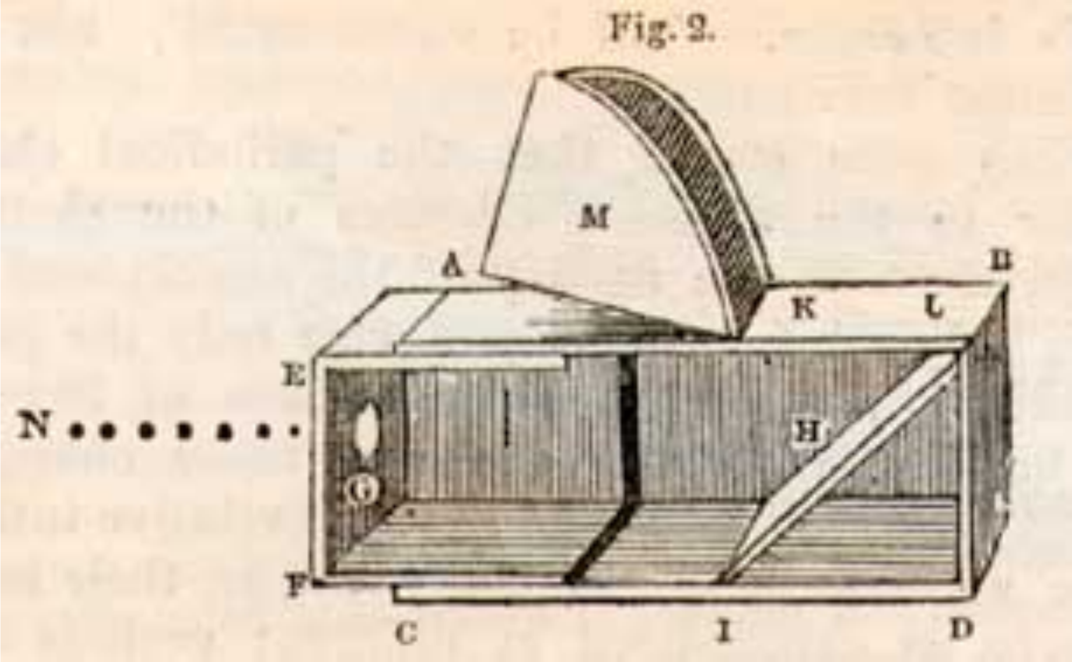
An ideal lens performs the same projection as a pinhole but **gathers much more light**!

Thin Lens Properties



- An incident ray that passes through the center of the lens will in effect continue in the direction it had when it entered the lens.
- Any incident ray traveling parallel to the optical axis, will refract and travel through the focal point on the opposite side of the lens.
- Any incident ray traveling through the focal point on the way to the lens will be refracted and travel parallel to the principal axis.
- All rays emanating from P and entering the lens will converge at P'.

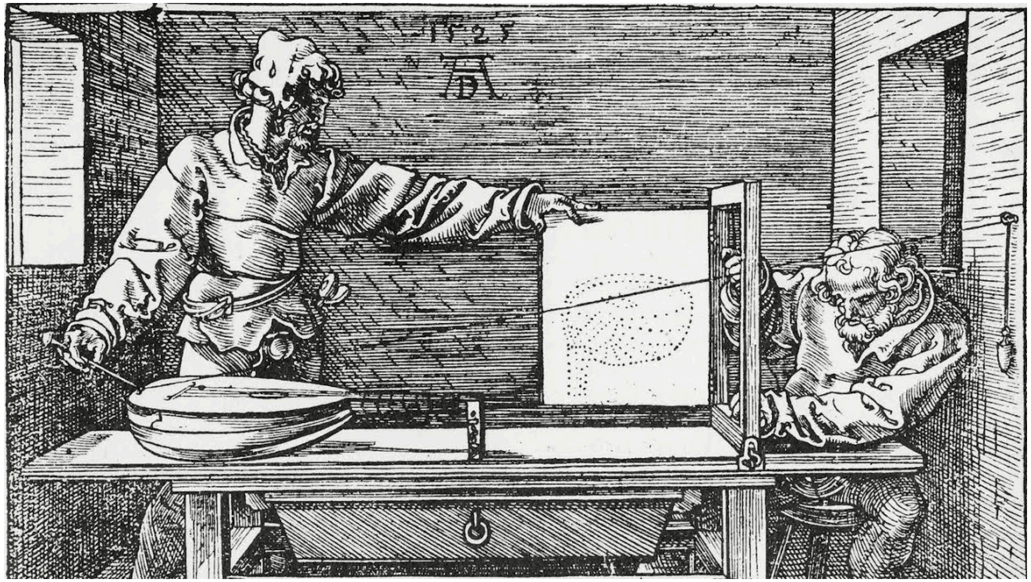
Camera Obscura



- Used by painters since the Renaissance to produce perspective projections.
- Direct ancestors to the first film cameras.

Optional

Durer 1471-1528



- He clearly knew all about the perspective transform!

Optional

Shifting Perspective

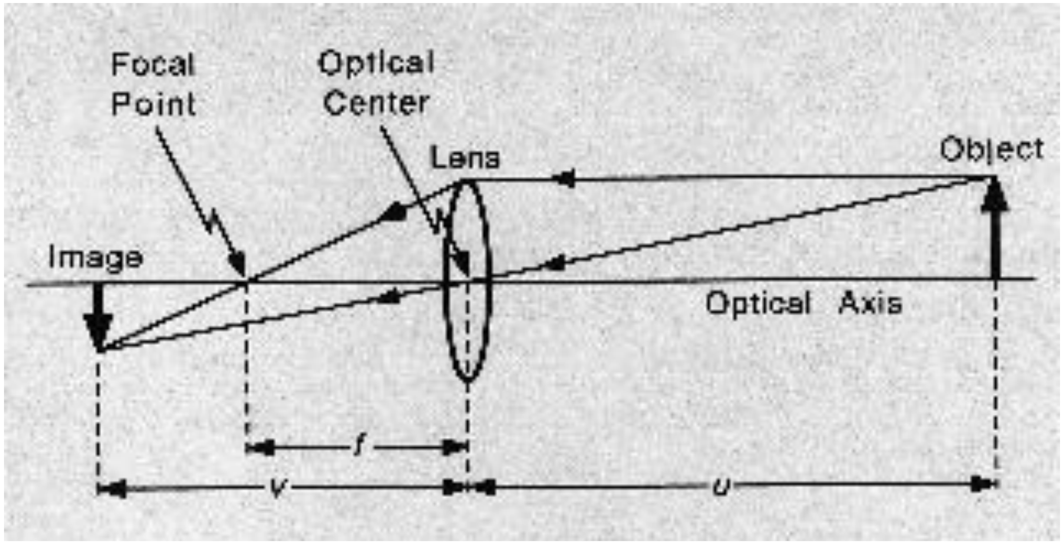
Buddha cutting his hair, 8th c.

China, 8th century:

- The focal point moves from one part of the image to the other.
- The characters are always seen at eye-level as the picture is unrolled.

Optional

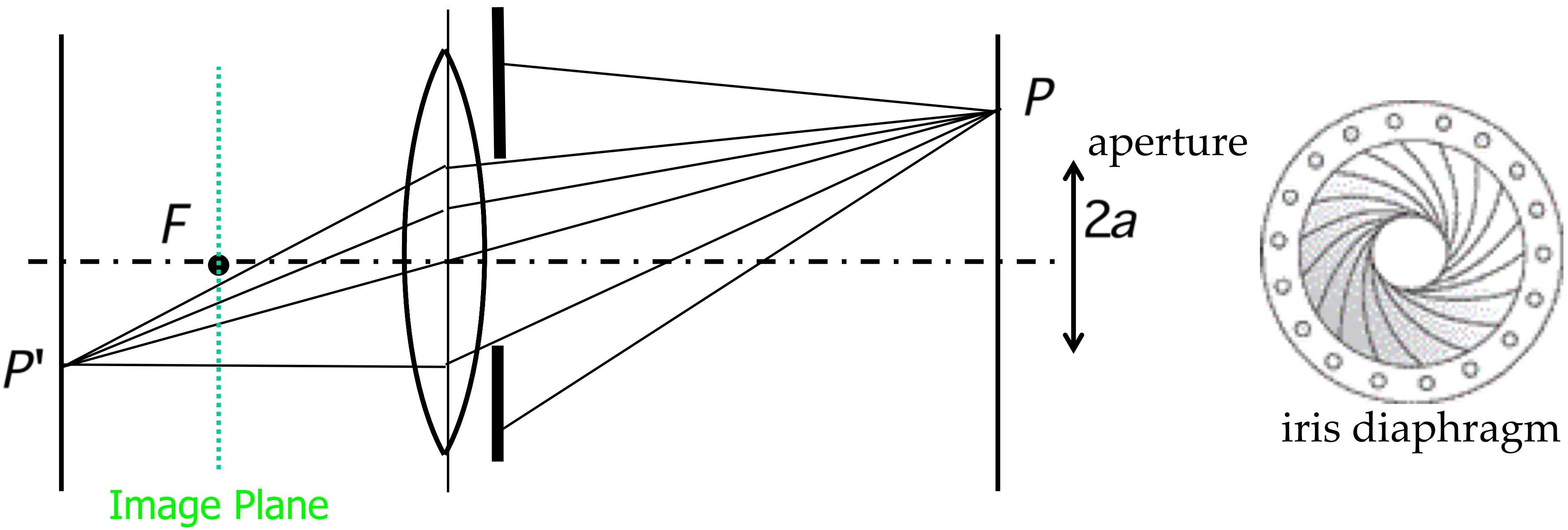
Thin Lens Equation



$$\cancel{\frac{1}{u}} + \frac{1}{v} = \frac{1}{f}$$

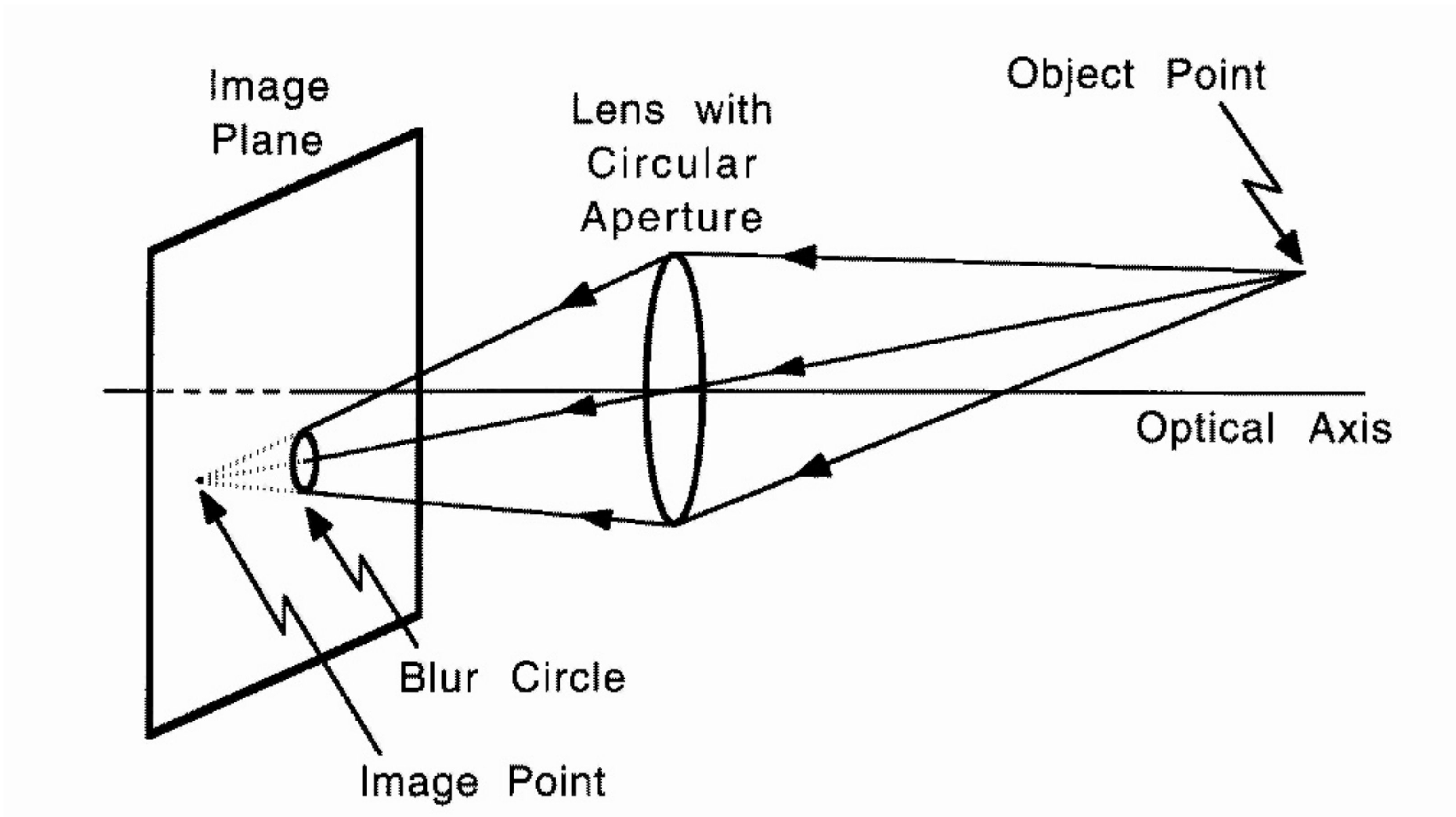
→ Lens with focal distance f equivalent to pinhole camera with similar focal distance but larger aperture.

Aperture



- Diameter d=2a of the lens that is exposed to light.
- The image plane is not located exactly where the rays meet.
- The greater a, the more blur there will be.

Blur Circle



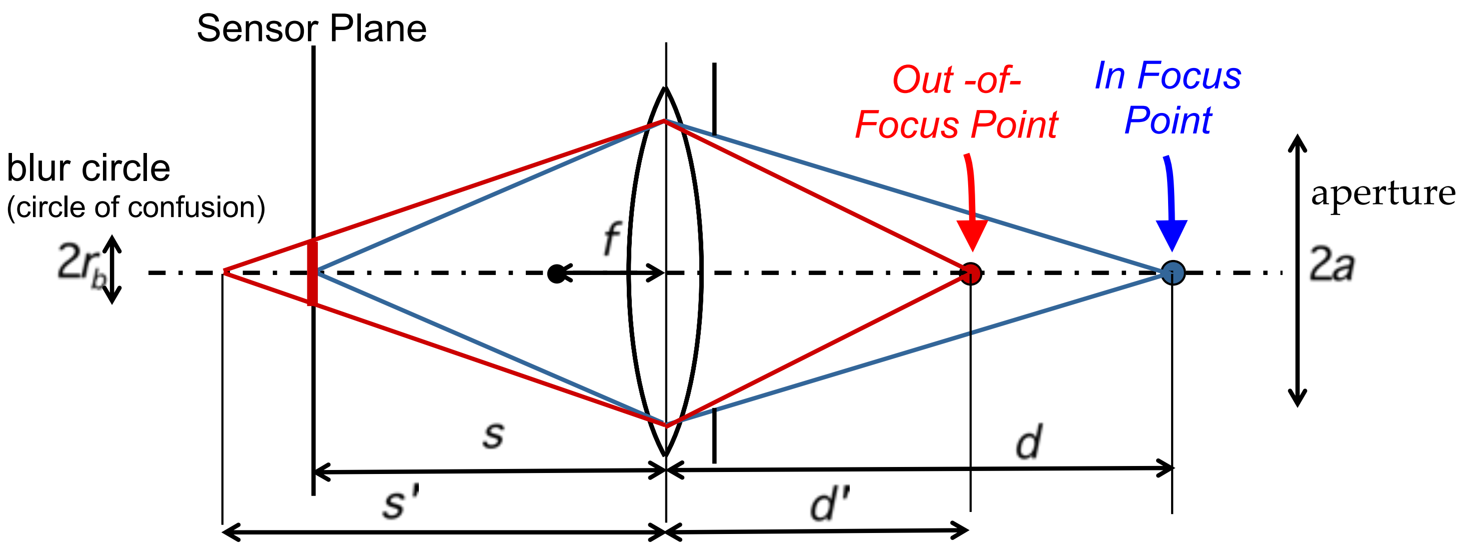
The size of the blur circle is proportional to the aperture.

Depth of Field

- Range of object distances (d-d’) over which the image is sufficiently well focused.
- Range for which blur circle is less than the resolution of the sensor.

Small focal length —> Large depth of field.

Proof



- Simple geometry:

$$r_b = \frac{a}{s'}\,|\,s - s'\,|$$

- Thin lens equation:

$$\frac{1}{d} + \frac{1}{s} = \frac{1}{f} \Rightarrow s = \frac{df}{d-f}$$

$$\frac{1}{d'} + \frac{1}{s'} = \frac{1}{f} \Rightarrow s' = \frac{d'f}{d'-f}$$

$$s' - s = \frac{f}{d-f}\frac{f}{d'-f}(d-d')$$

$$r_b = \frac{af^2}{s'}\left|\frac{(d-d')}{(d-f)(d'-f)}\right|$$

Large f and a —> Large r_b.

Changing the Focal Length

Wide field of view (small f):
Large depth of field.

Narrow field of view (large f):
Small depth of field.

$$r_b \propto \frac{af^2}{s'}$$

Small f —> Small r_b

Changing Aperture

f/11 1/30sec

Small aperture, long exposure:
Large depth of field.

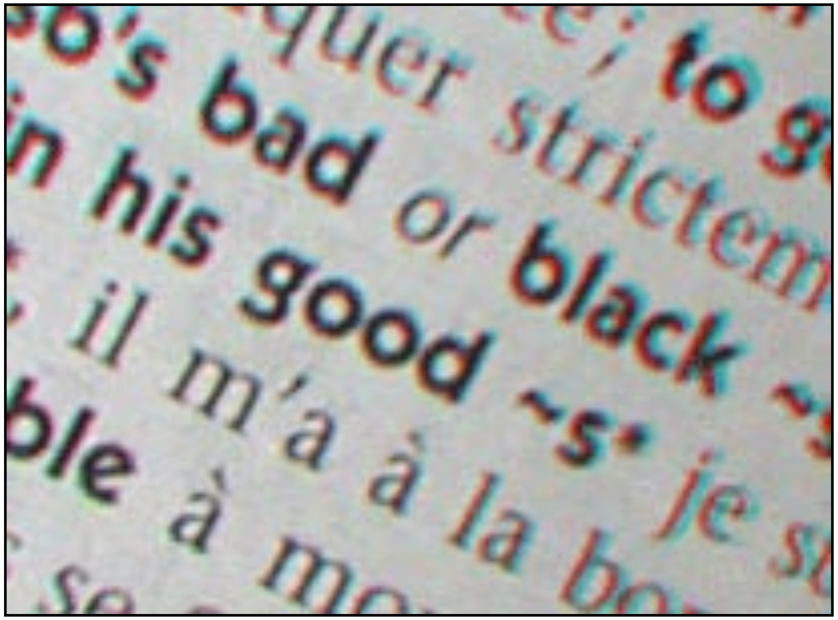
f/2.8 1/500sec

Large aperture, short exposure:
Small depth of field.

$$r_b \propto \frac{af^2}{s'}$$

Small a —> Small r_b

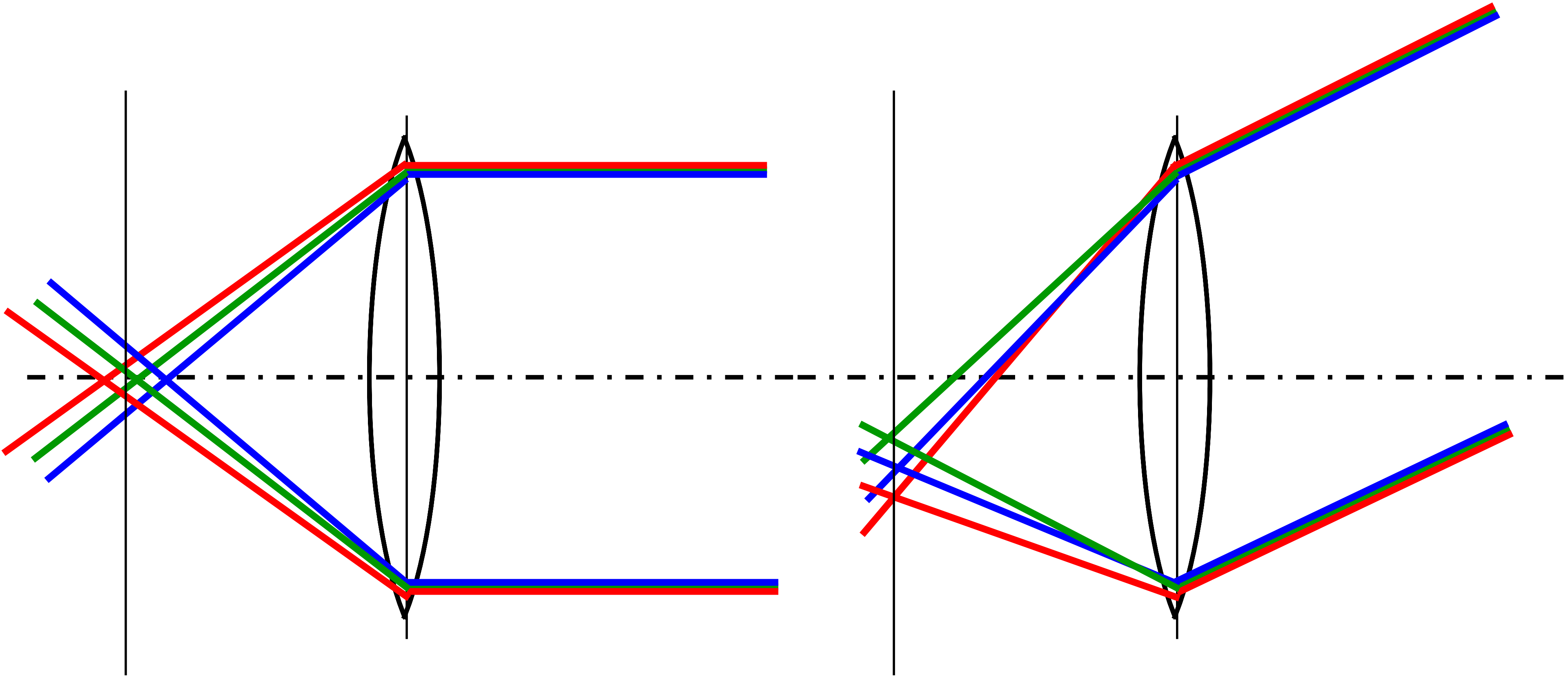
Distortions



The lens is not exactly a "thin lens:"

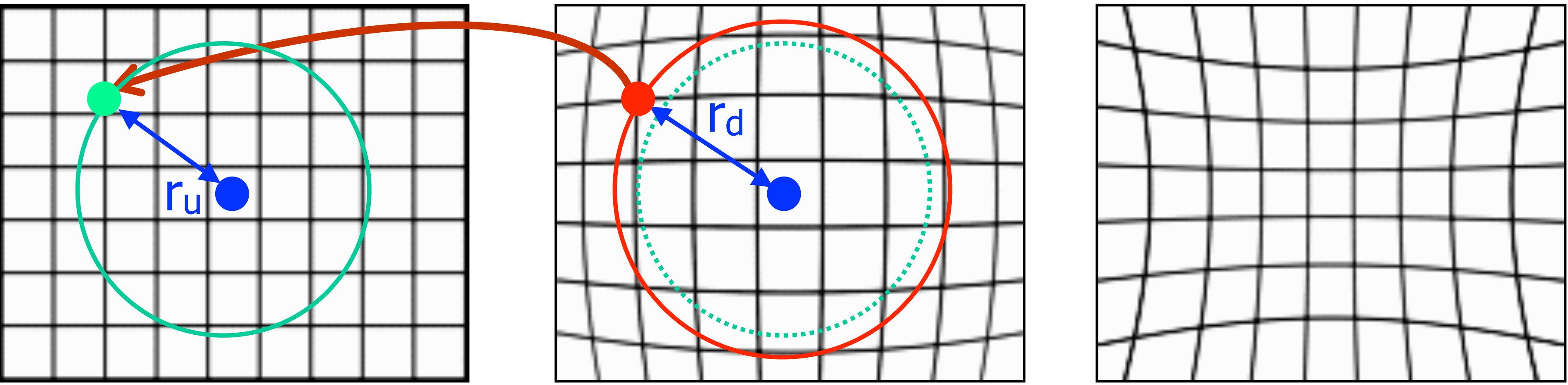
- Different wave lengths are refracted differently,
- Barrel Distortion.

Chromatic Aberration



Different wavelengths are refracted differently.

Radial Lens Distortion



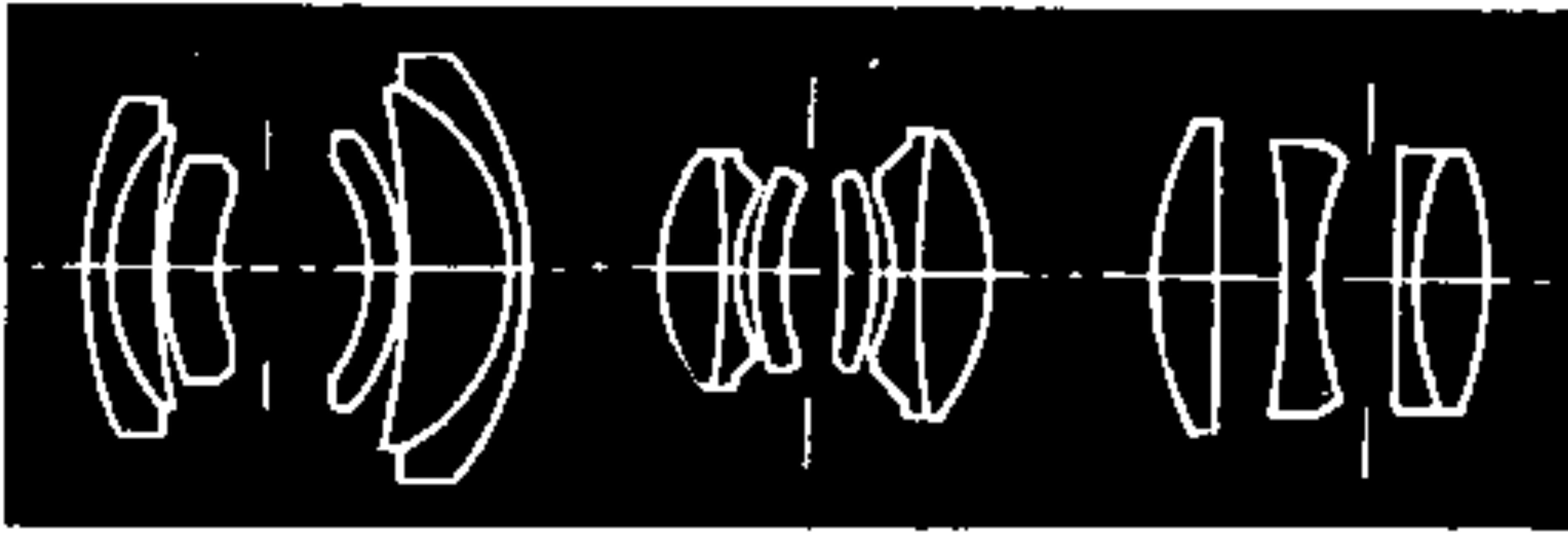
No Distortion Barrel Distortion Pincushion Distortion

The distortion is a function of radial distance to the image center:

$$r_u = r_d(1 + k_1 r^2 + k_2 r^4 + \ldots)$$

- r_d: Observed distance of the projected point to the center.
- r_u: Distance of the point to the center in an image without distortions.

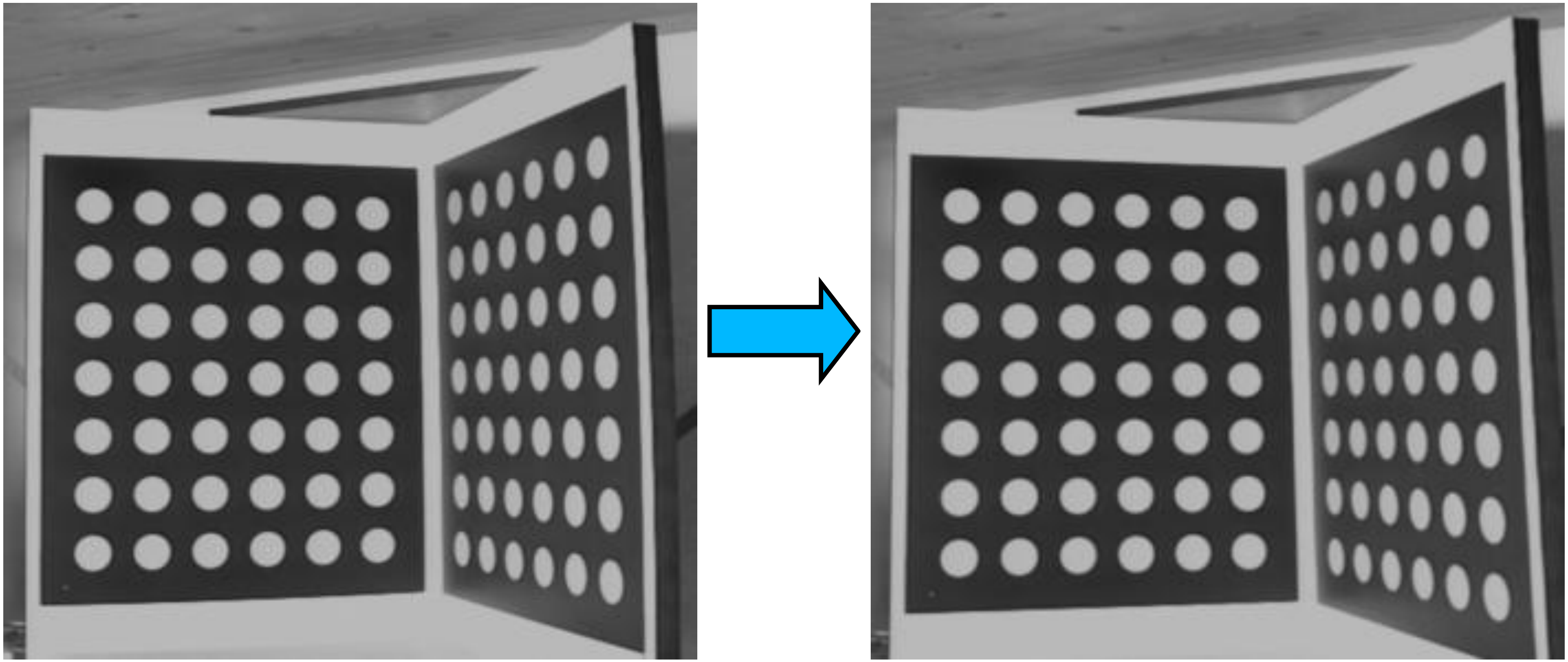
Lens Systems



Aberrations can be minimized by aligning several lenses with well chosen

- Shapes,
- Refraction indices.

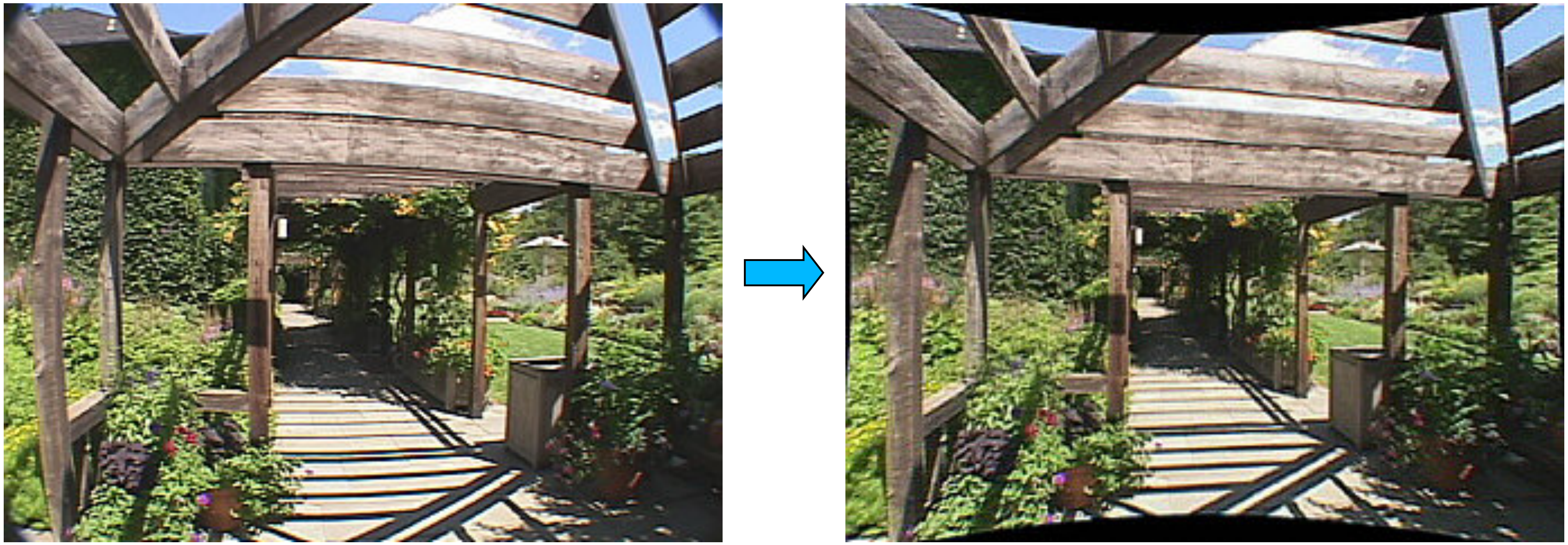
Undistorting



Real image Synthetic image

—> Create the synthetic image a sense without distortion would have produced.

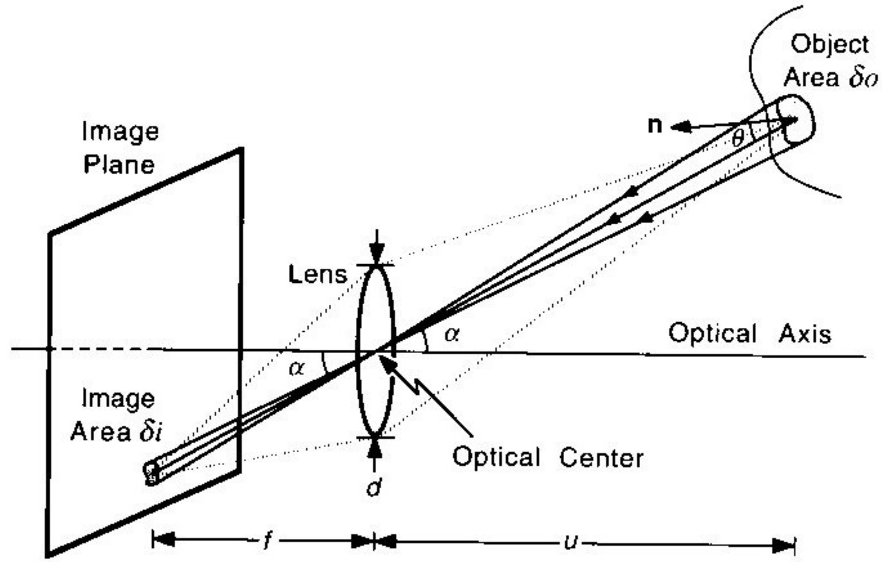
Undistorting



Once the image is undistorted, the camera projection can be formulated as a projective transform.

→ The pinhole camera model applies.

Fundamental Radiometric Equation



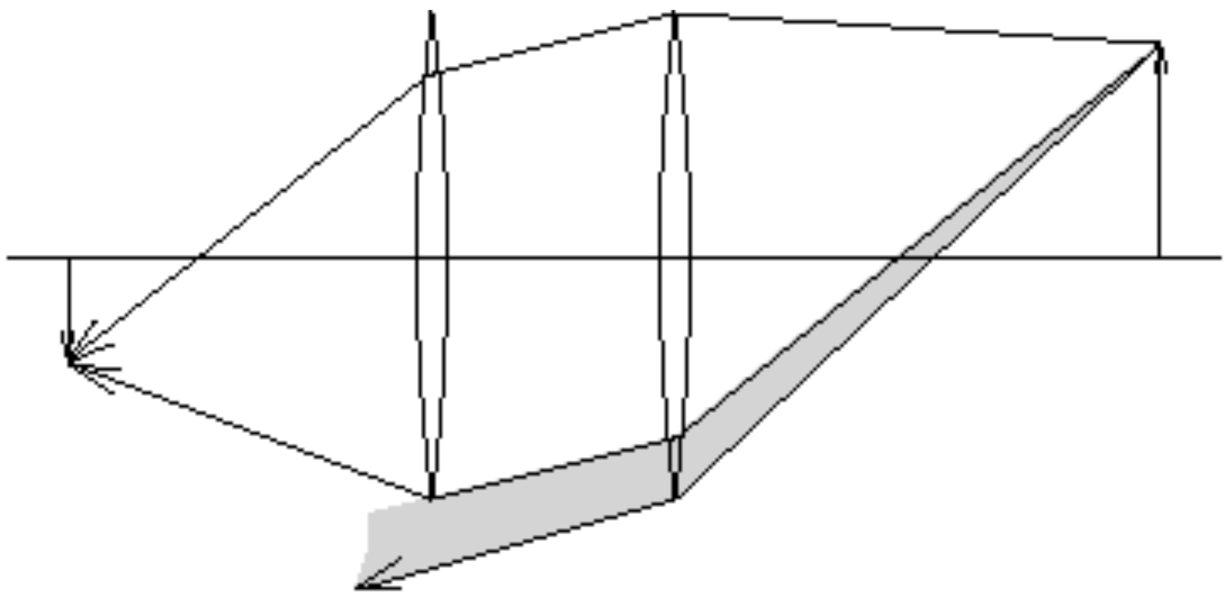
Scene Radiance (Rad) : Amount of light radiation emitted from a surface point (Watt / m2 / Steradian).

Image Irradiance (Irr): Amount of light incident at the image of the surface point (Watt / m2).

$$\text{Irr} = \frac{\pi}{4}\left(\frac{d}{f}\right)^2 \cos^4(\alpha)\,\text{Rad}\ ,$$

$$\Rightarrow \text{Irr} \propto \text{Rad for small values of } \alpha\ .$$

Vignetting



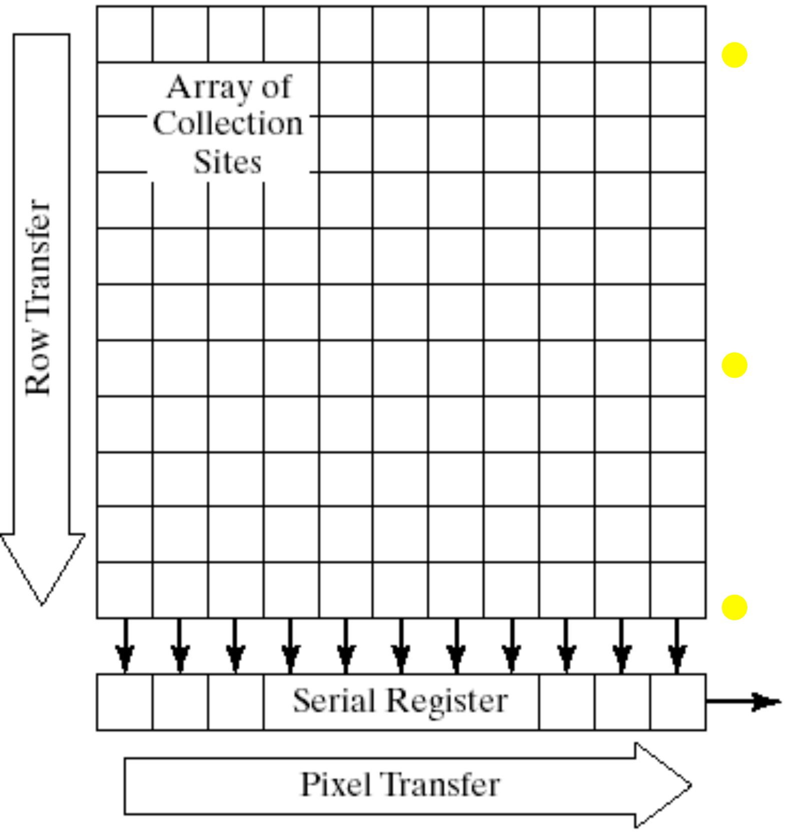
Images can get darker towards their edges because some of the light does not go through all the lenses.

De Vignetting

—> As for geometric undistortion, undo vignetting to create an image that an ideal camera would have produced.

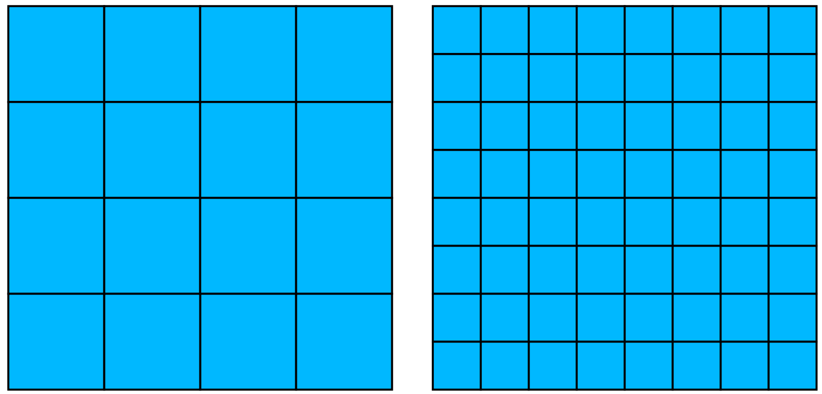
Zheng et al., CVPR'08

Sensor Array



- Photons free up electrons that are then captured by a potential well.
- Charges are transferred row by row wise to a register.
- Pixel values are read from the register.

Sensing



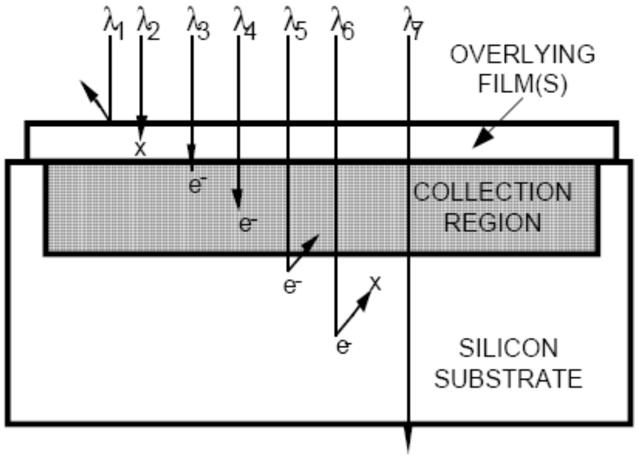
Conversion of the "optical image" into an "electrical image":

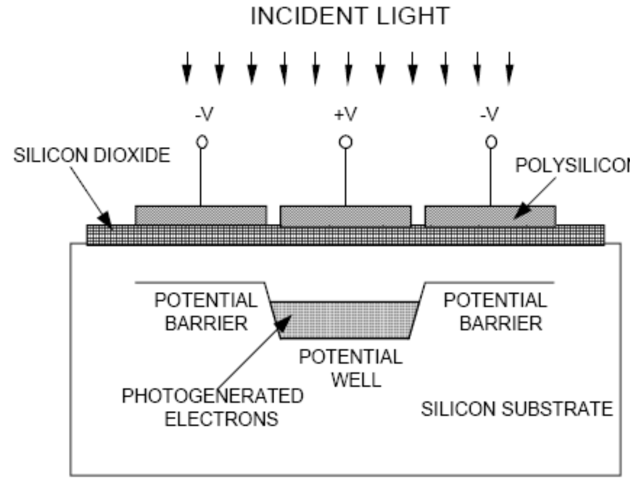
$$
\begin{aligned}
E(x, y) &= \int_{t_0}^{t_1} \int_0^{\Lambda} \mathrm{Irr}(x, y, t, \lambda) s(\lambda) dt d\lambda \\
I(m, n) &= \mathrm{Quantize}\left(\int_{x_0}^{x_1} \int_{y_0}^{y_1} E(x, y) dx dy\right)
\end{aligned}
$$

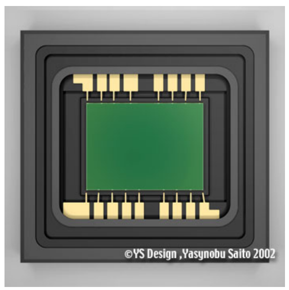
→ Quantization in

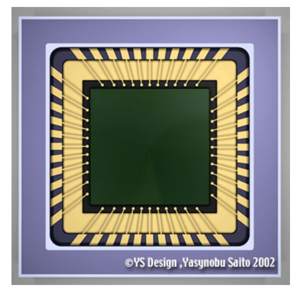
- Time
- Space

Sensors









CCD

CMOS

- Charged Coupling Devices (CCD): Made through a special manufacturing process that allows the conversion from light to signal to take place in the chip without distortion.
- Complimentary Metal Oxide Semiconductor (CMOS): Easier to produce and similar quality. Now used in most cameras except when quantum efficient pixels are needed, e.g. for astronomy.

In Short

- Camera geometry can be modeled in terms of the pinhole camera model, which is linear in projective space.

- Image radiance is roughly proportional to surface radiance and the two can be used interchangeably for our purposes.