COMP417
Introduction to Robotics and Intelligent Systems

Image Formation and Multi-View Geometry

McGill

MRL Mobile Robotics Lab at McGill University

Vision based on cameras

- We have already talked about cameras as a sensor.
- Now let's consider using the data from a camera for robotics applications.

- Obstacle detection and avoidance.
- 3D mapping.
- Detection and recognition tasks.

Motivation

- We have already seen quite successful SLAM methods based on laser sensors. Why bother with vision?
  - Camera technology cheap and ubiquitous
  - Camera is a passive sensor, lower energy
  - Some environments/platforms can't support laser
  - Vision is quite a "rich" source of information

How hard is computer vision?

Marvin Minsky, MIT
Turing award, 1969

"In 1966, Minsky hired a first-year undergraduate (JS) student and assigned him a problem to solve over the summer: connect a television camera to a computer and get the machine to describe what it sees."

Crevier 1993, pg. 88

How hard is computer vision?

Marvin Minsky, MIT
Turing award,1969

Gerald Sussman, MIT

"You'll notice that Sussman never worked in vision again!" – Berthold Horn

What do humans see?

Tricky issue #1

- Vision is a projection of the world into 2D (and image).
  - What are we losing?
  - The extra dimension (3D, or at least 2 1/2D).
  - Relectance.
  - Light source direction.
  - Recall: it's an inverse problem we need to solve.
    - Inverse problems tend to be *ill-posed*.

Dudek,Jenkin
Ch 3

"Solving vision"

- We see an image. What gave rise to it?
  - What 3D surface caused this 2D image?
    - Can we compute a 3D surface that has this thing as it's 2D camera projection?
    - Are such solutions unique? **For a given image, how many different 3D surfaces might explain it?**
- In the algorithmic context, can we develop an algorithm that computes that 3D surface?
- In the machine learning context, can we learn an approach to find the 3D surface that best explains a 2D image?

Depth perception can be ambiguous from just a single image

Ill-posed problems (in the sense of Hadamard)

- Problems where (either):
  - the solution is undefined, <- less typical in our domains
  - the solution is not uniquely defined, <- more typical in our domains
  - the solution is not stable.

  - (can you define each of these formally)?

e.g. Recover $f_0(x)$ from the noise-corrupted version $f'(x)$ as omega tends towards infinity, even if k is small.

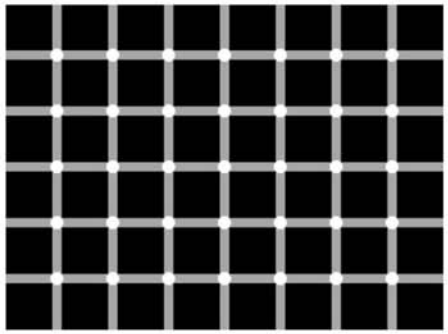
$$f(x) = f_0(x) + k\sin(\omega x). \qquad f'(x) = f_0'(x) + k\omega\cos(\omega x).$$

What to do? Stabilize!

- "Fix" ill-posed problems by stabilizing them.

- Add assumptions, more formally, regularization
  - (most commonly smoothness).
- The idea of regularization is to remove some solutions from the mix, making the solution unique (or more distinct).
  - In machine learning, it is seen as a way of penalizing solutions that are overly complex.

What do humans see?

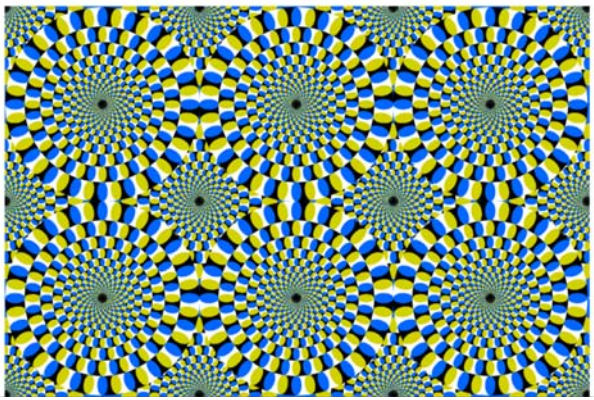


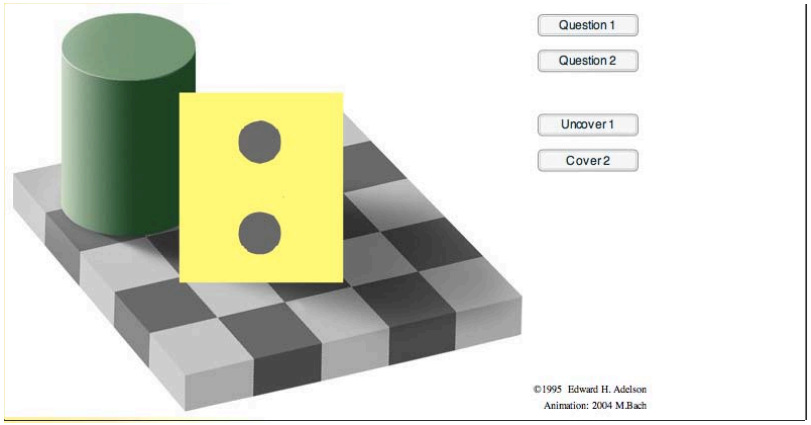
Count the black dots! :o)

Quirks of the human visual system: why?

They are probably there as a way as a side-effect of the assumptions used to address the ill-posed nature of the inverse problem.

Peripheral drift illusion

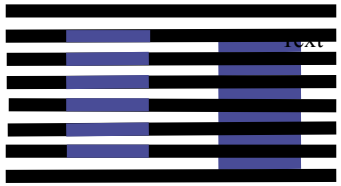




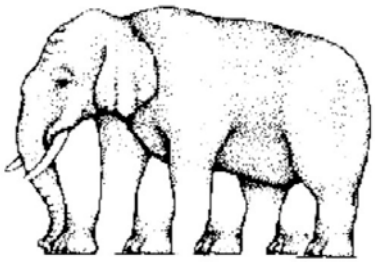
17

Report the colors

"Munker-White Illusion"



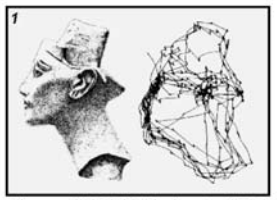
18



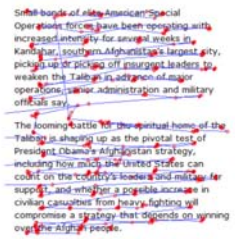
How many legs does this elephant have?

Where do humans fixate?

Top down or bottom up?



"Eye Movements and Vision" by A. L. Yarbus; Plenum Press, New York; 1967



Visual saccades

Camera obscura: dark room

- Known during classical period in China and Greece (e.g., Mo-Ti, China, 470BC to 390BC)

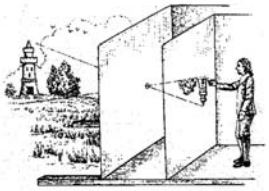
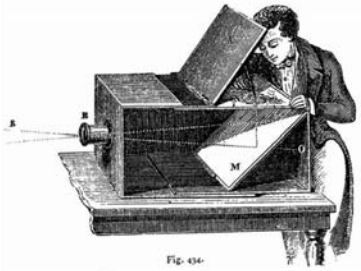


Illustration of Camera Obscura

Freestanding camera obscura at UNC Chapel Hill

Photo by Seth Ilys

James Hays



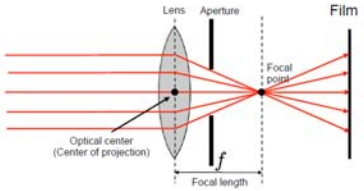
Lens Based Camera Obscura, 1568

Oldest surviving photograph

- Took 8 hours on pewter plate

Joseph Niepce, 1826

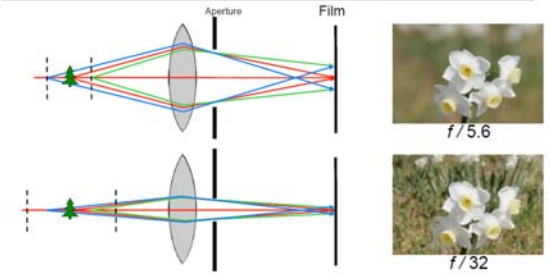
Lenses



A lens focuses parallel rays onto a single focal point

- focal point at a distance *f* beyond the plane of the lens
  - *f* is a function of the shape and index of refraction of the lens
- Aperture of diameter D restricts the range of rays
  - aperture may be on either side of the lens
- Lenses are typically spherical (easier to produce)

Depth of field

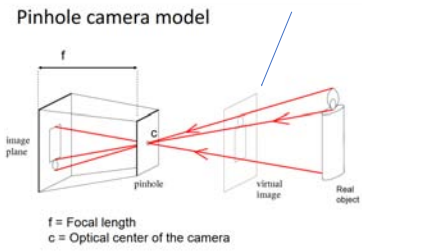


Changing the aperture size affects depth of field

- A smaller aperture increases the range in which the object is approximately in focus

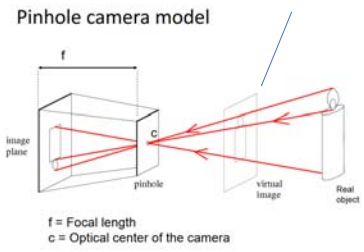
Flower images from Wikipedia http://en.wikipedia.org/wiki/Depth_of_field

To avoid thinking about image inversion

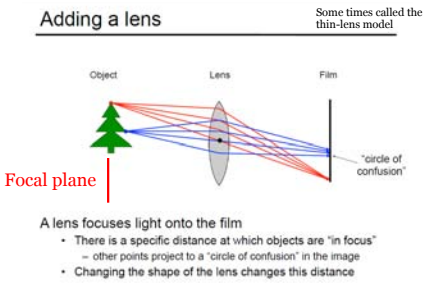


Point aperture → nearly every pixel in the image is in focus

To avoid thinking about image inversion

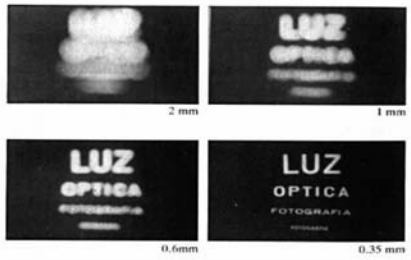


Point aperture → nearly every pixel in the image is in focus → almost infinite depth of field



Aperture of nonzero diameter → only pixels corresponding to objects on the focal plane are in focus → narrow depth of field

Shrinking the aperture

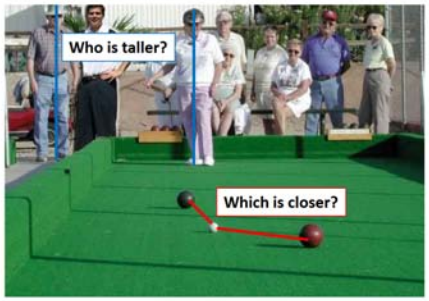


Why not make the aperture as small as possible?

- Less light gets through
- *Diffraction* effects...

Projective Geometry

Length (and so area) is lost.



Length and area are not preserved

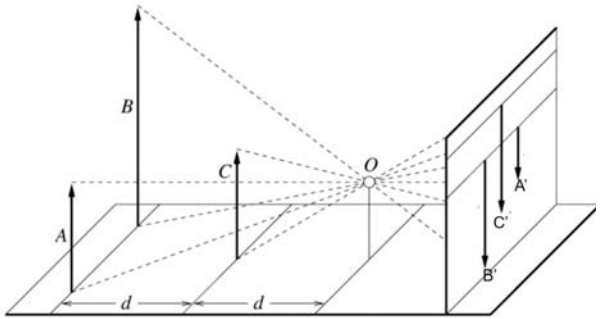
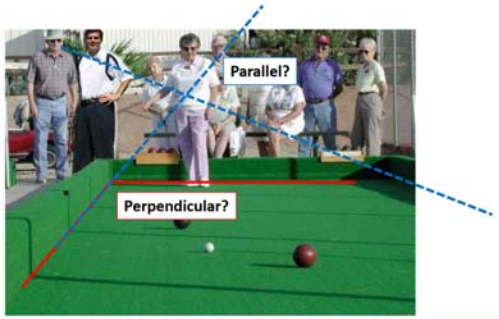


Figure by David Forsyth

Projective Geometry

Angles are lost.



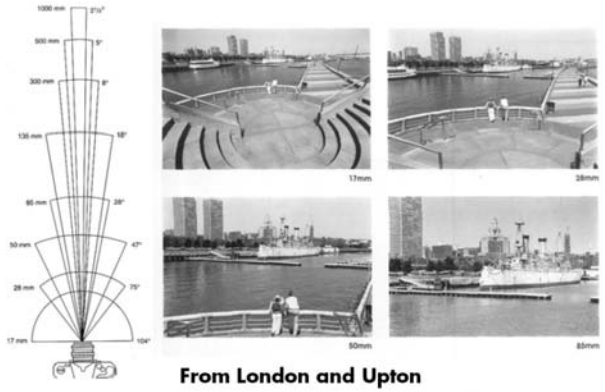
Projective Geometry

What is preserved?

- Straight lines are still straight.

Chromatic aberration
Failure of a lens to focus all colors to the same convergence point.
Due to difference wavelengths having different refractive indeces

Field of View (Zoom, focal length)



From London and Upton

Camera parameters

Focus – Shifts the depth that is in focus.

Focal length – Adjusts the zoom, i.e., wide angle or telephoto lens.

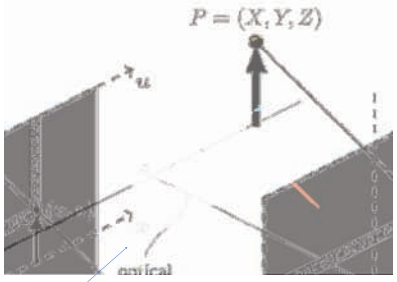
Aperture – Adjusts the depth of field and amount of light let into the sensor.

Exposure time – How long an image is exposed. The longer an image is exposed the more light, but could result in motion blur.

ISO – Adjusts the sensitivity of the "film". Basically a gain function for digital cameras. Increasing ISO also increases noise.

How do we project 3D points to pixels?
What is the measurement model?

From 3D points to pixels: pinhole camera



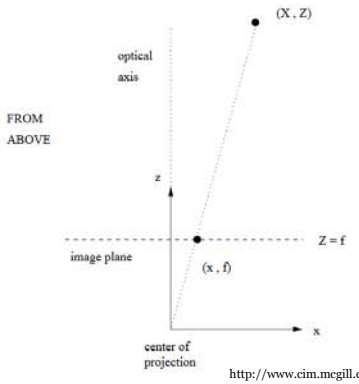
(1) Perspective projection $\begin{bmatrix} x \\ y \end{bmatrix} = \pi(X, Y, Z)$

(2) Conversion from metric to pixel coordinates

$$u = m_x x + c_x$$
$$v = m_y y + c_y$$

m_x, m_y represent number of pixels per mm for the two axes

Perspective projection
$[x,y] = \pi(X,Y,Z)$



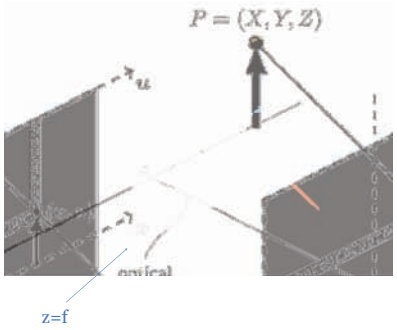
http://www.cim.mcgill.ca/%7Elanger/558.html

By similar triangles: x/f = X/Z

So, x = f * X/Z and similarly y = f * Y/Z

Problem: we just lost depth (Z) information by doing this projection, i.e. depth is now uncertain.

From 3D points to pixels: pinhole camera



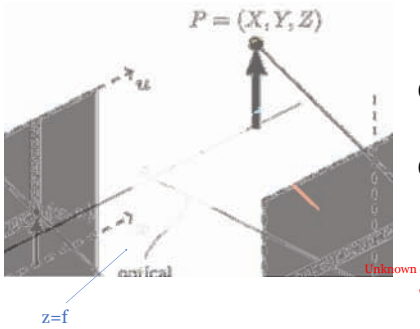
(1) Perspective projection $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} fX/Z \\ fY/Z \end{bmatrix} = \pi(X, Y, Z)$

(2) Conversion from metric to pixel coordinates

$$u = m_x x + c_x$$
$$v = m_y y + c_y$$

$$h_{\text{pinhole}}(X, Y, Z) = \begin{bmatrix} \frac{fm_x X}{Z} + c_x \\ \frac{fm_y Y}{Z} + c_y \end{bmatrix} + \text{noise in pixels}$$

From 3D points to pixels: pinhole camera



(1) Perspective projection $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} fX/Z \\ fY/Z \end{bmatrix} = \pi(X, Y, Z)$

(2) Conversion from metric to pixel coordinates

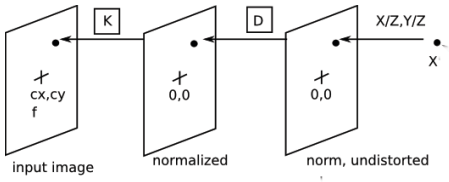
$$u = m_x x + c_x$$
$$v = m_y y + c_y$$

Unknown depth/scale

Camera calibration matrix

Usually presented as $s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} fm_x & 0 & c_x \\ 0 & fm_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$

From 3D points to pixels: thin lens camera



(1) Perspective projection $\begin{bmatrix} x \\ y \end{bmatrix} = \pi(X, Y, Z)$

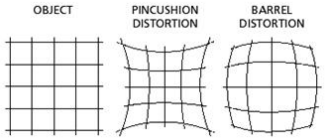
(2) Lens distortion

$$[x^*, y^*] = D(x, y)$$

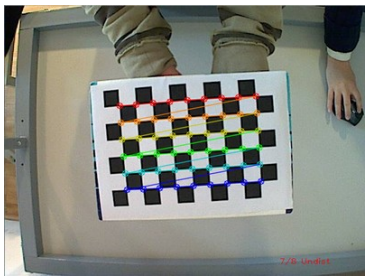
(3) Conversion from metric to pixel coordinates

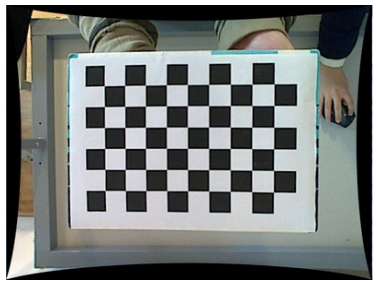
$$u = m_x x^* + c_x$$
$$v = m_y y^* + c_y$$

(2) Lens distortion
[x*, y*] = D(x,y)



(2) Estimating parameters of lens distortion:
[x*, y*] = D(x,y)





$$x^* = x\frac{1+k_1r^2+k_2r^4+k_3r^6}{1+k_4r^2+k_5r^4+k_6r^6} \quad \text{where} \quad r = x^2+y^2$$

$$y^* = y\frac{1+k_1r^2+k_2r^4+k_3r^6}{1+k_4r^2+k_5r^4+k_6r^6} \quad \text{where} \quad r = x^2+y^2$$

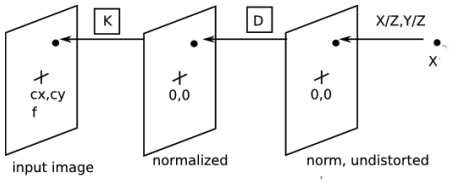
Correcting radial distortion

Barrel distortion

Corrected

from Helmut Dersch

From 3D points to pixels: thin lens camera



(1) Perspective projection $\begin{bmatrix} x \\ y \end{bmatrix} = \pi(X, Y, Z)$

(2) Lens distortion

$$[x^*, y^*] = D(x, y)$$

(3) Conversion from metric to pixel coordinates

$$u = m_x x^* + c_x$$
$$v = m_y y^* + c_y$$

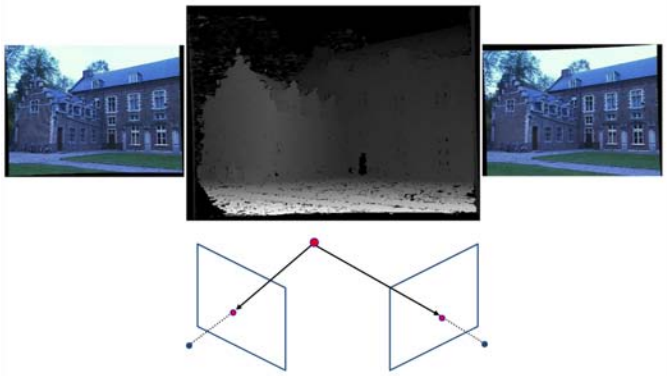
If we have access to camera calibration parameters
we can undo the lens distortion, and treat the measurement
model as in the pinhole camera → single-camera image rectification

What visual or physiological cues help us to perceive 3D shape and depth?

Perspective effects

Image credit: S. Seitz

Stereo



Slides: James Hays and Kristen Grauman

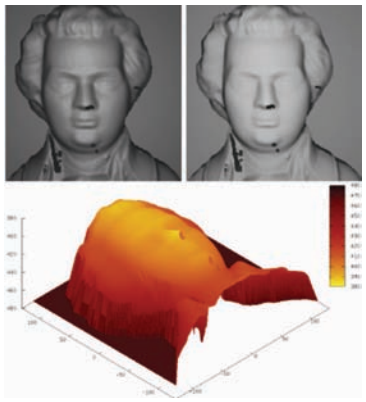
Why multiple views?

Structure and depth can be ambiguous from single views...

Images from Lana Lazebnik

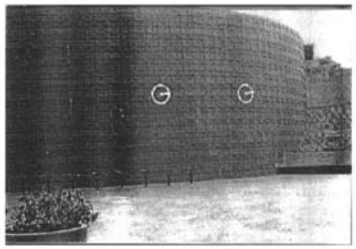
If stereo were critical for depth perception, navigation, recognition, etc., then rabbits would never have evolved.

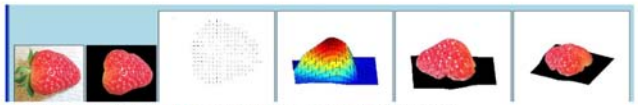
Shape from shading



"Numerical schemes for advanced reflectance models for Shape from Shading", Vogel, Cristiani

Texture



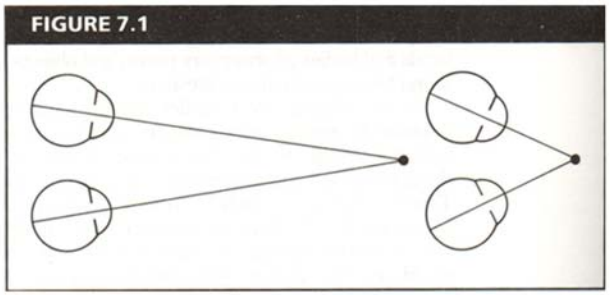


[From A.M. Loh. The recovery of 3-D structure using visual texture patterns. PhD thesis]

Occlusion

Rene Magritt'e famous painting *Le Blanc-Seing* (literal translation: "The Blank Signature") roughly translates as "free hand" or "free rein".

Human stereopsis



From Bruce and Green, Visual Perception, Physiology, Psychology and Ecology

Human eyes **fixate** on point in space – rotate so that corresponding images form in centers of fovea.

Structure from Motion

"SFMedu: A Structure from Motion System for Education", Jianxiong Xiao

Many depth from X methods. We are going to focus on structure from motion and stereo → part of multiple-view geometry