CIS 3990

Mobile and IoT Computing

https://penn-waves-lab.github.io/cis3990-24spring

Lecture 5: Visual Localization and Tracking

Instructor: Mingmin Zhao (mingminz@cis.upenn.edu)

TA: Haowen Lai (hwlai@cis.upenn.edu)

Sensing Modalities for Localization?

Past Lectures:

Radio Signals (EM waves)

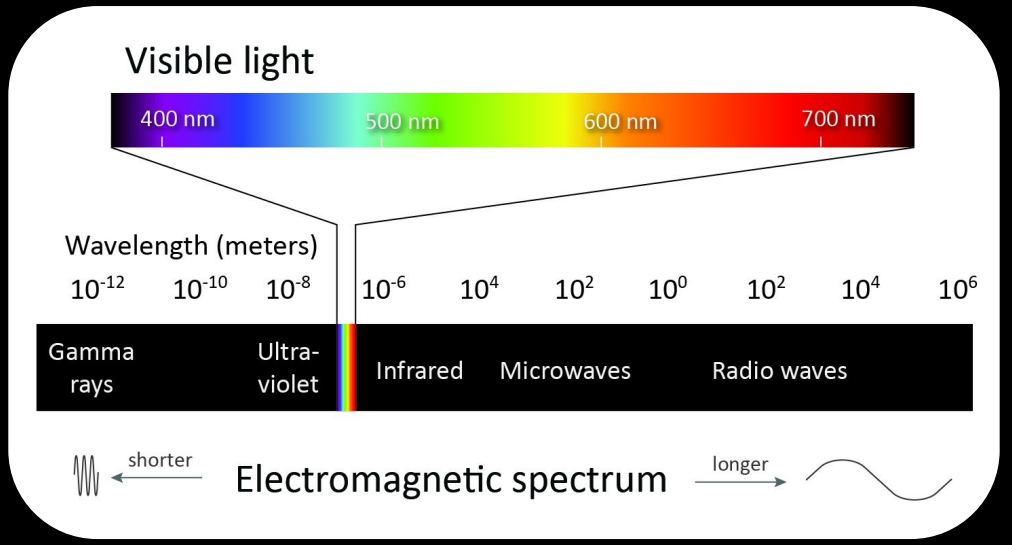
Acoustic Signals (mechanical waves)

Device-based Device-free

This Lecture:

Visual signals: cameras, LiDAR

(Visible light is also EM wave)



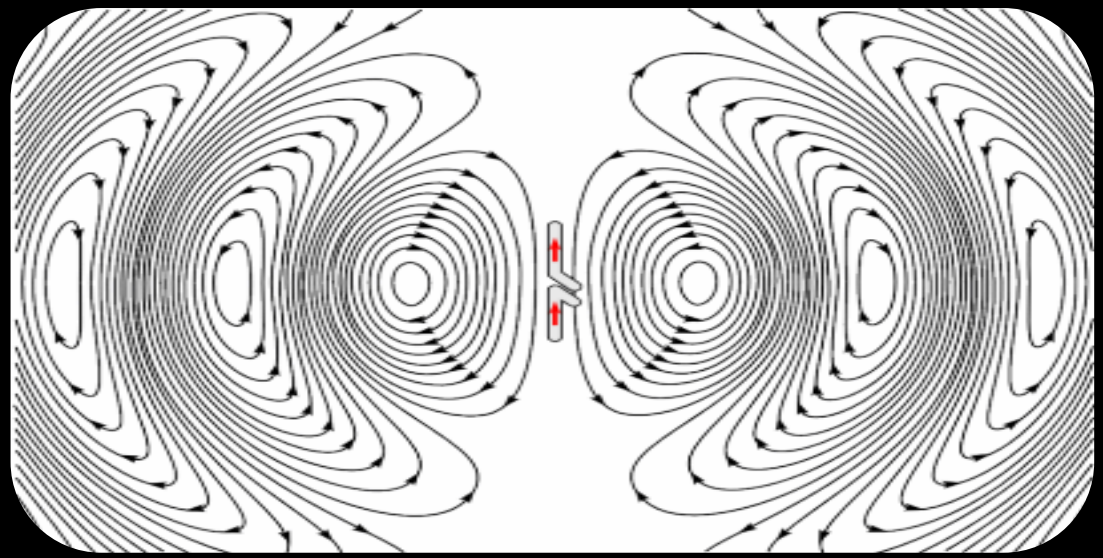
What are their differences?

Sensing Modalities for Localization?

Radio waves

Detection with antennas

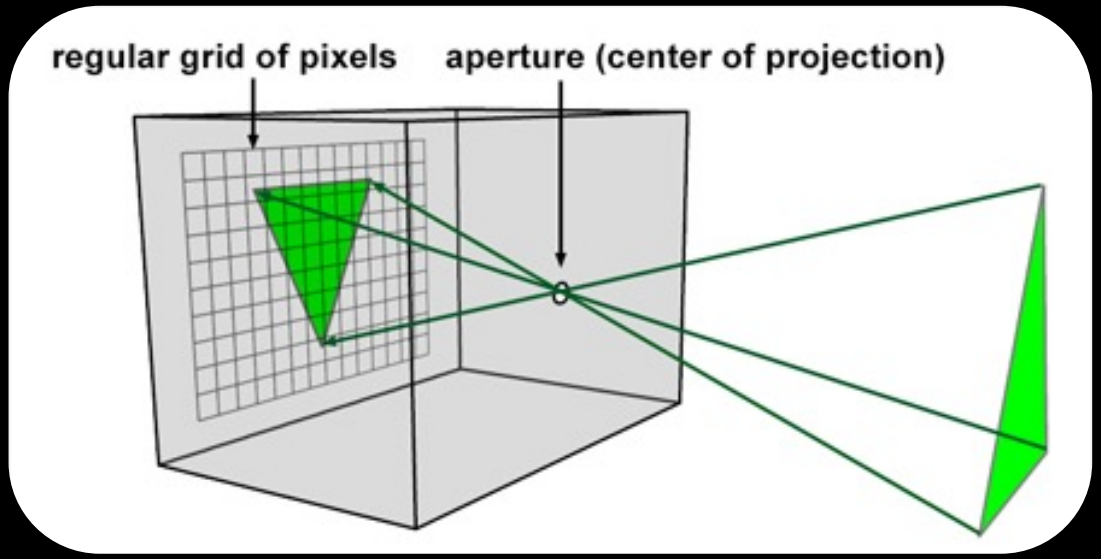
Wide beam & limited angular resolution



Visible light

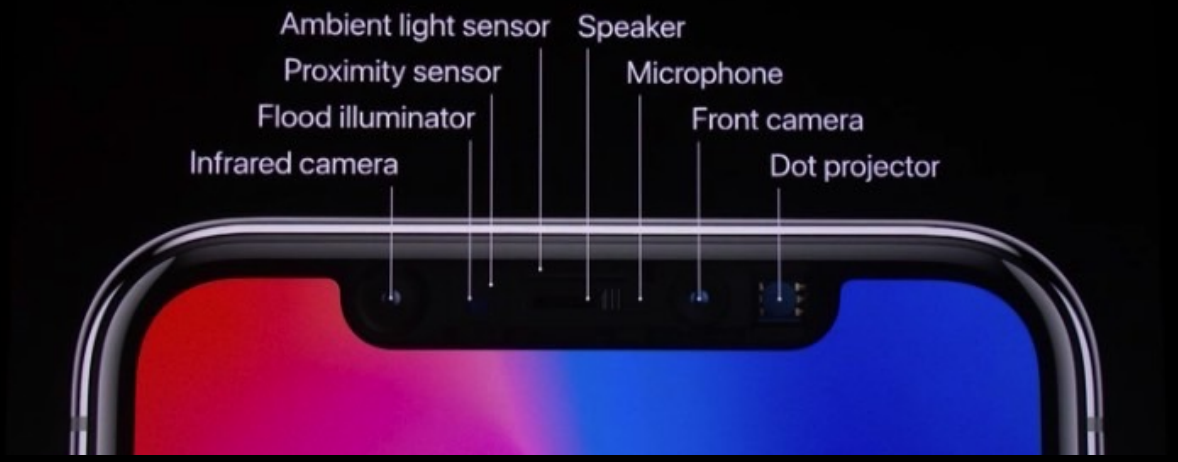
Detection with photo-sensitive sensors

Narrow beam & high angular resolution



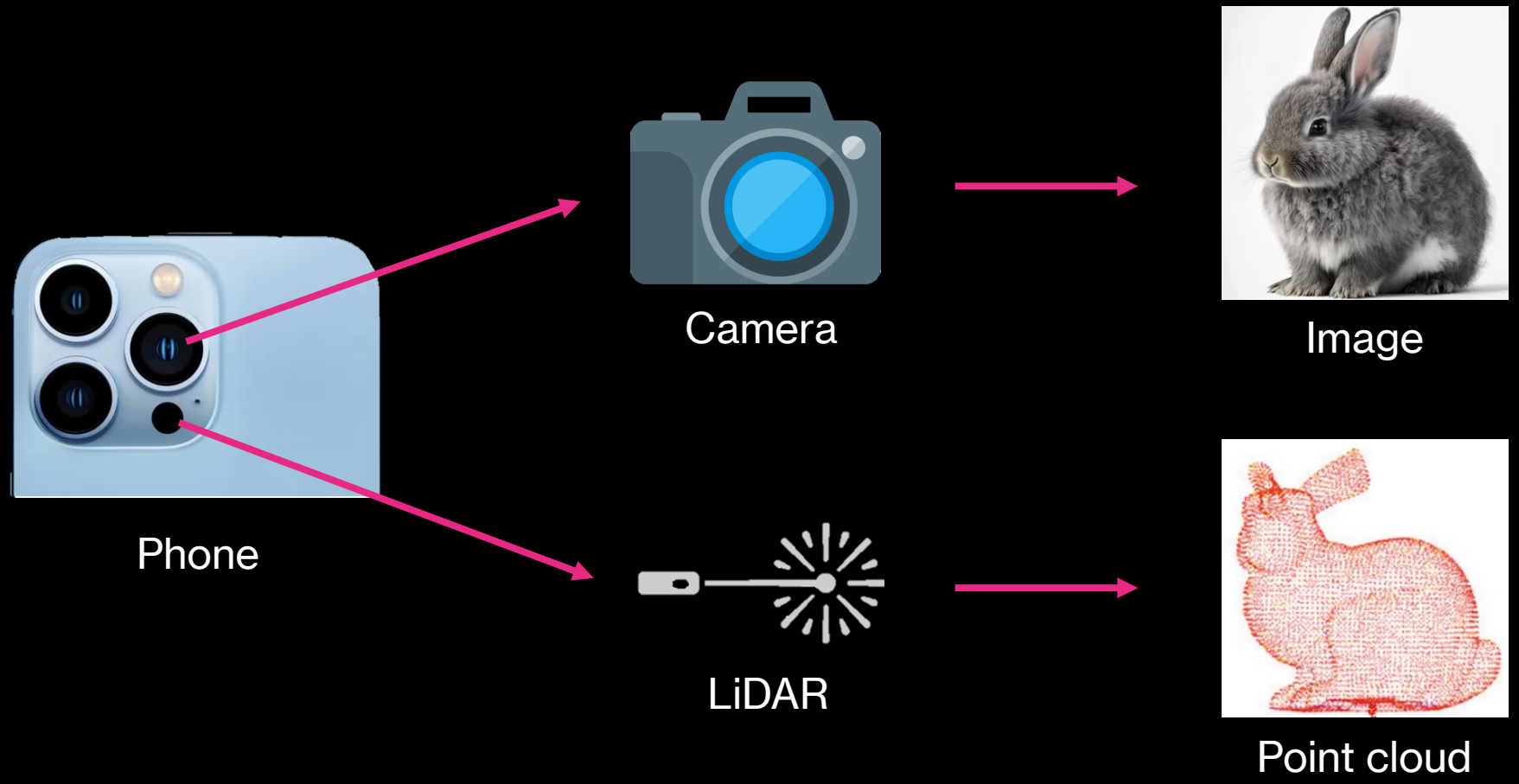
Visual Localization & Tracking

Our smartphones have plenty of visual sensors



Visual Localization & Tracking

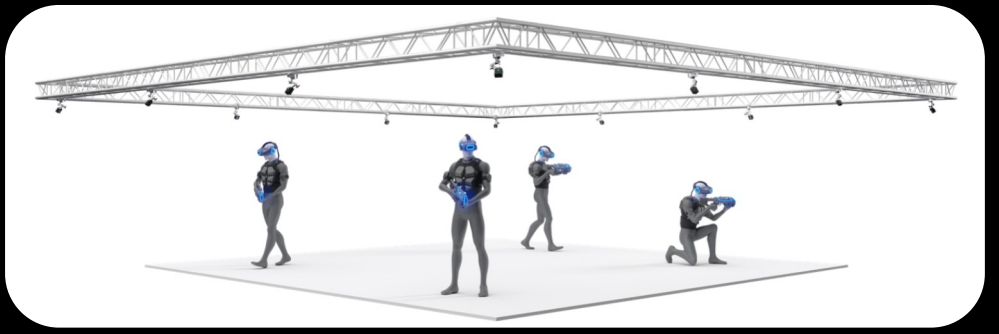
The process of determining location using visual sensors



Visual-based Localization

Share similar principles with wireless/wave-based localization

- Feature matching & pattern recognition
  + c.f. fingerprinting in wireless localization
- Motion tracking by comparing consecutive frames
  + c.f. Doppler shift/speed
- Motion capture system
  + c.f. AoA and triangulation
- Visual tag detection & SLAM
  + c.f. AoA and geometric constraints
- LiDAR & ToF camera
  + c.f. ToF measurement with Radar



Pinhole Camera Model

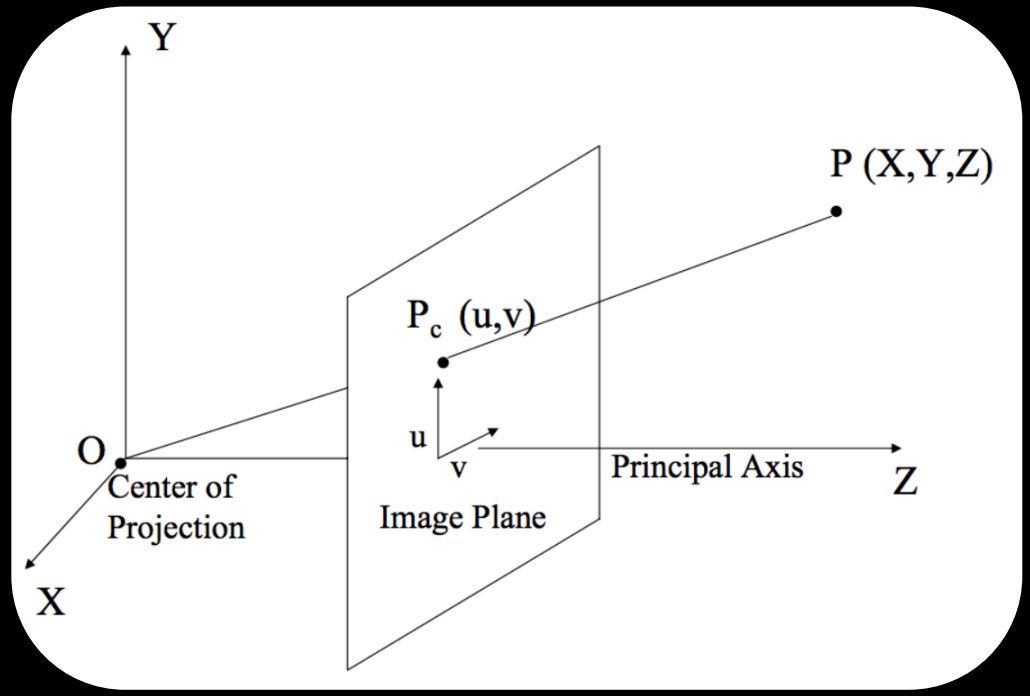
The model describes how a 3D point is projected to a 2D image

For a 3D point $P(X, Y, Z)$ and a pixel $P_c(u, v)$

$$\begin{cases} u = f\alpha\frac{X}{Z} + c_x \\ v = f\beta\frac{Y}{Z} + c_y \end{cases}$$

scaling shifting

- where f is the focal length.
- α and β change the physical unit to pixel
- c_x and c_y shift the point to the desired origin in the image



Pixel ⇔ AoA ⇔ Line in the 3D (light ray)

Pinhole Camera Model

The model describes how a 3D point is projected to a 2D image

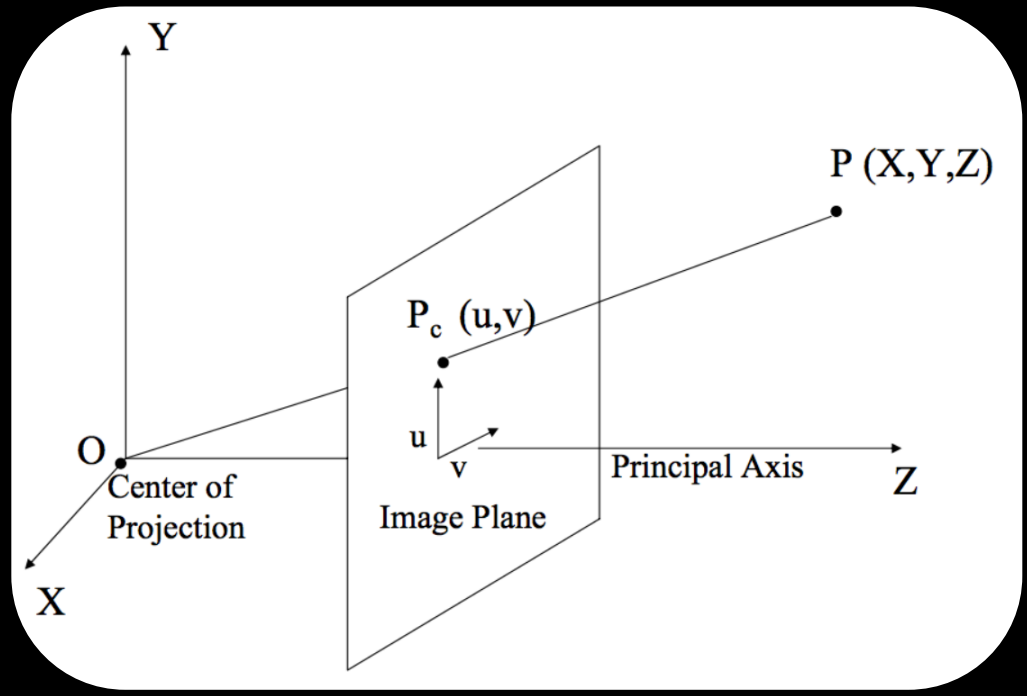
For a 3D point $P(X, Y, Z)$ and a pixel $P_c(u, v)$

$$\begin{cases} u = f\alpha\frac{X}{Z} + c_x \\ v = f\beta\frac{Y}{Z} + c_y \end{cases} \rightarrow \begin{cases} u = f_x\frac{X}{Z} + c_x \\ v = f_y\frac{Y}{Z} + c_y \end{cases}$$

Rewrite P_c as $P_c(u, v, 1)$ (homogeneous coordinates)

$$P_c = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} P = KP$$

Intrinsic matrix



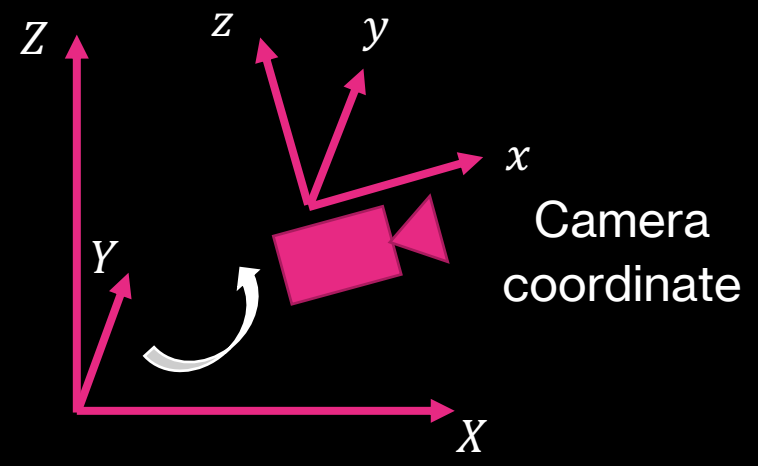
Pinhole Camera Model

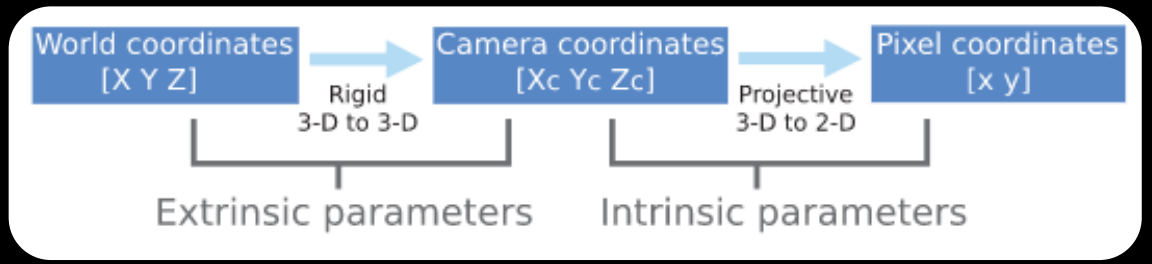
- Usually, the camera coordinate doesn't align with the world coordinate
- There exists a rotation matrix $R \in \mathbb{R}^{3\times3}$ and translation vector $\mathbf{t} \in \mathbb{R}^3$ to align these two coordinates

$$P_c = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} [\mathbf{R}|\mathbf{t}]P = K[\mathbf{R}|\mathbf{t}]P$$

Intrinsic matrix | Extrinsic matrix

Unknown



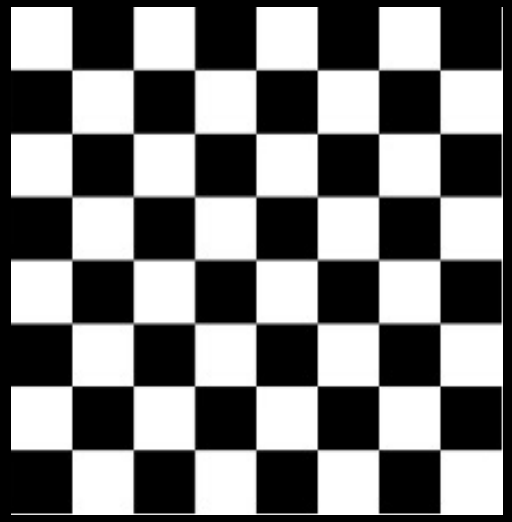


Camera Distortion

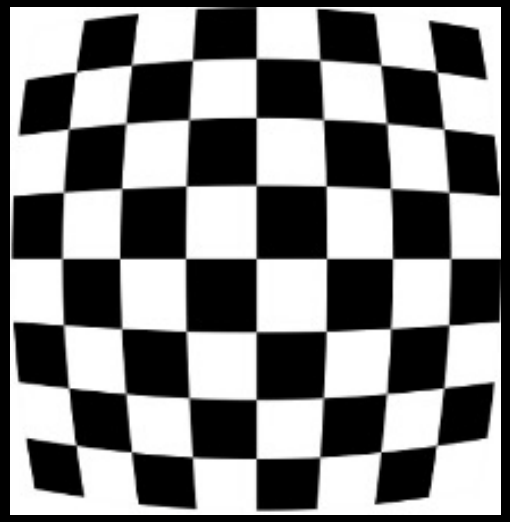
Distortion occurs due to lens and imperfect manufacture

Radial distortion: Light rays bend more near the edges of a lens than they do at its optical center.

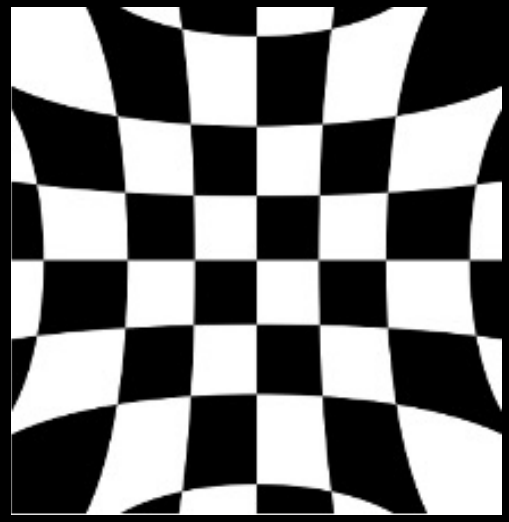
Tangential distortion: occurs when the lens and the image plane are not parallel.



No distortion



Negative radial distortion
(Pincushion distortion)

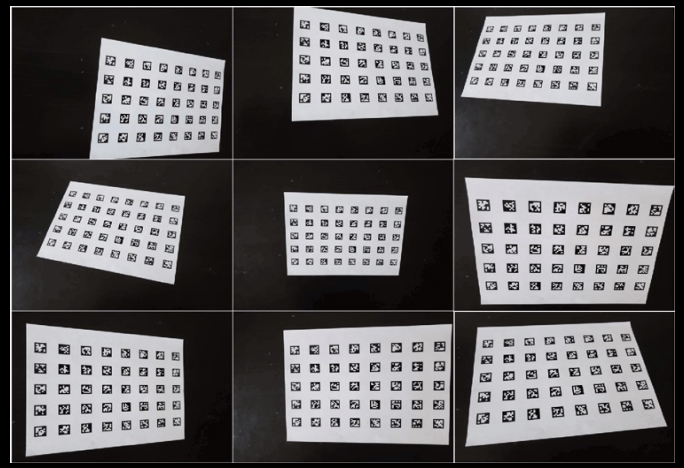


Positive radial distortion
(Barrel distortion)

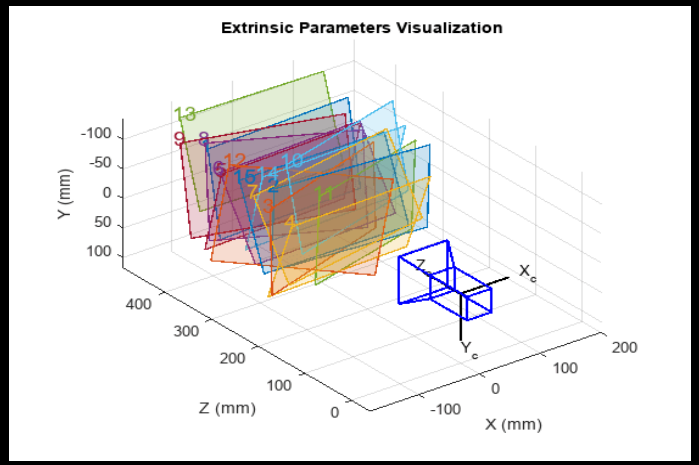
Camera Calibration

Estimate the unknown parameters with multiple observations

- intrinsic matrix K
- extrinsic matrix $[R|\mathbf{t}]$
- distortion parameter $[k_1, k_2, k_3, p_1, p_2]$

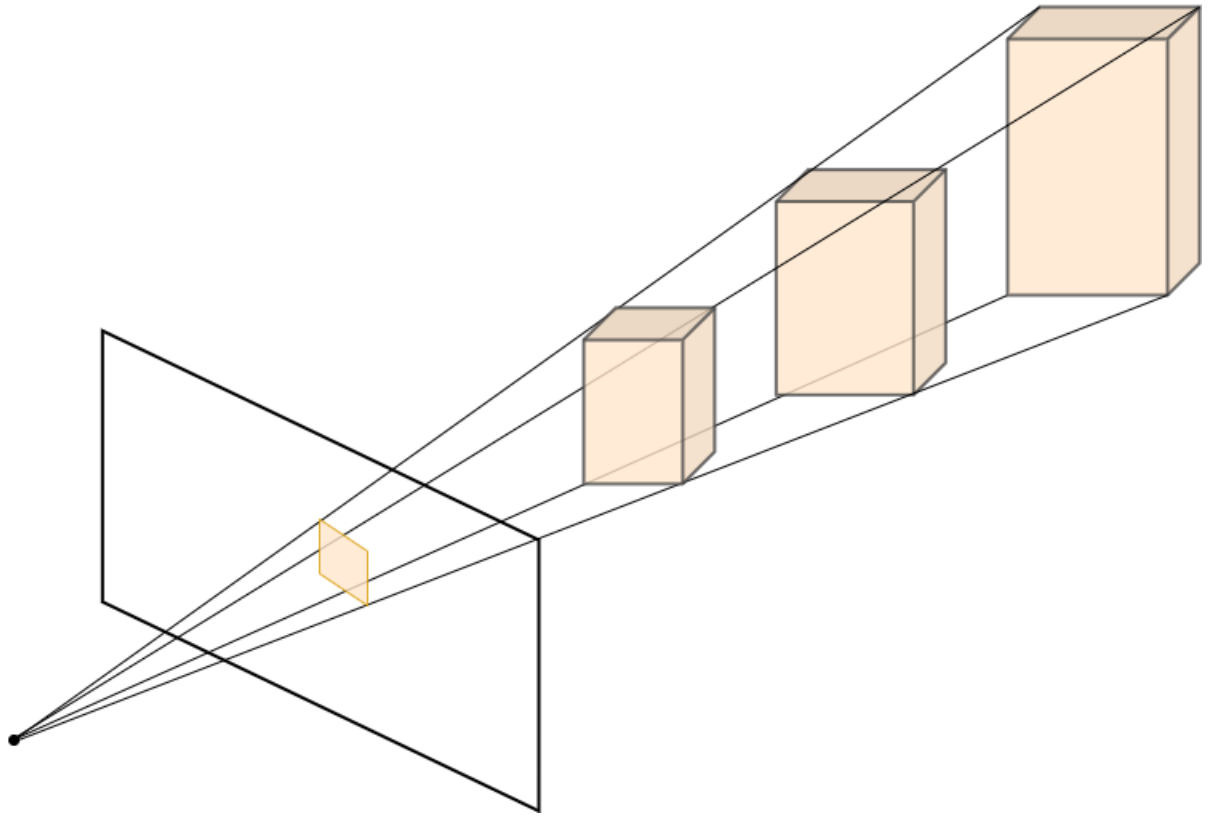


Multiple observations



Extrinsic Visualization

Z. Zhang. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11):1330-1334, 2000.



How to Resolve Depth-Size Ambiguity?

Resolving Depth-Size Ambiguity

Tags with known size

Capturing Depth

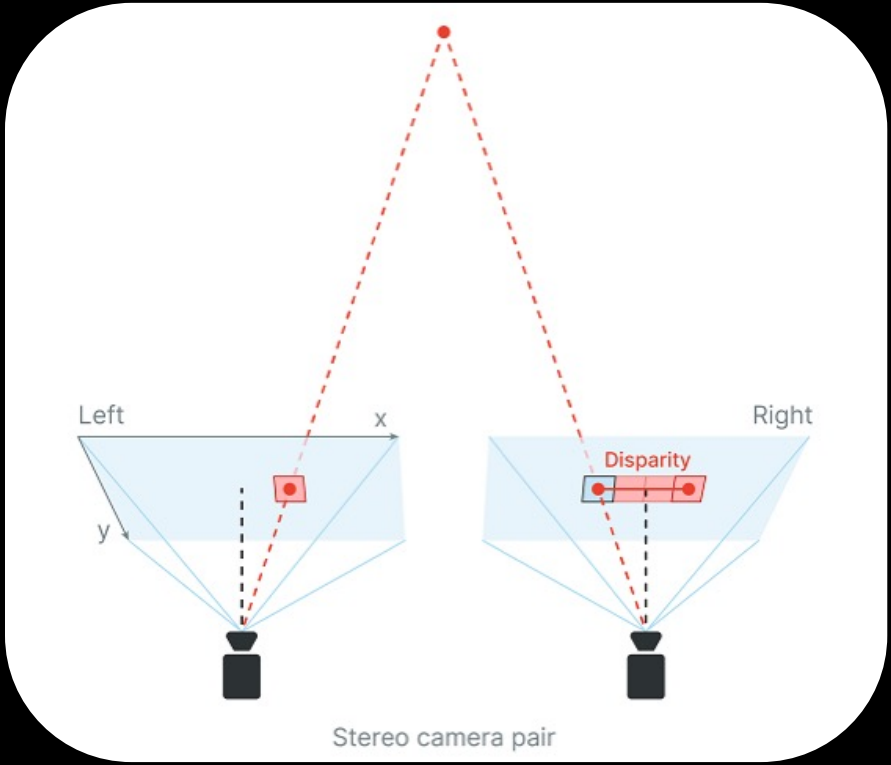
Depth camera

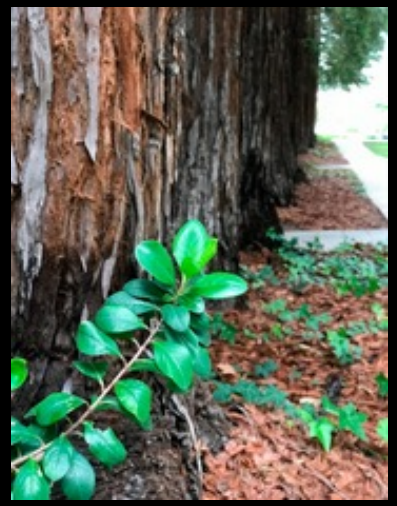
- Measure the depth (i.e., the z value) using **Structured Light** or **ToF**
- RGB-D: RGB image + corresponding depth image

Capturing Depth

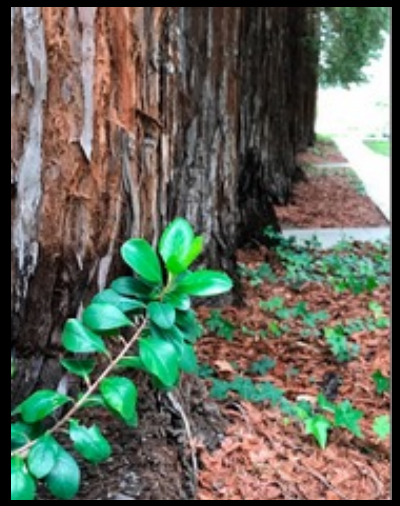
Stereo camera

- Use two cameras at slightly different location (known shift) to capture images
- Calculate depth based on disparity





Left



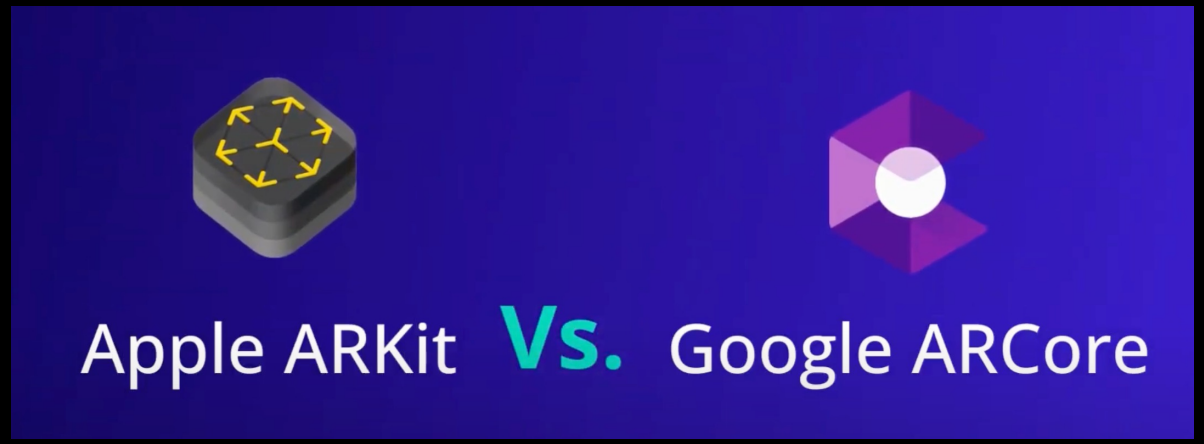
Right

Depth

Visual Odometry

Estimate the motion of the camera/robot based on the images

Visual Odometry in ARKit and ARCore



Motion tracking (visual odometry) serves as a core component for immersive AR/VR experience.

Visual Odometry: Limitations

- **Sufficient illumination** in the environment
- Dominance of **static scene** over moving objects
- **Enough texture** for feature extraction
- Sufficient **scene overlap** between consecutive frames



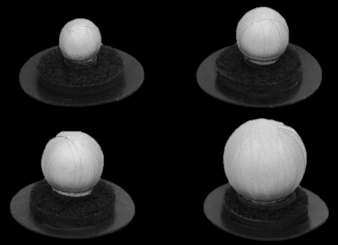

Visual-Inertial Odometry can help mitigate some of them

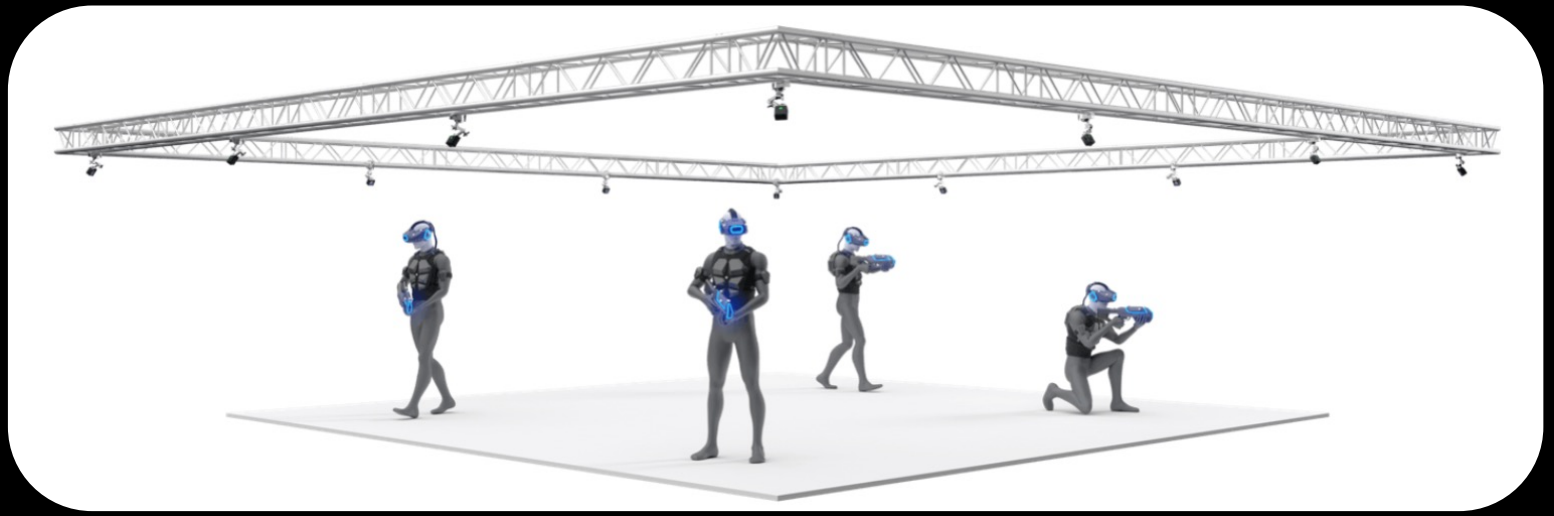
Camera Systems for Motion Capture

Motion Capture Camera System

- OptiTrack, VICON, etc.
- High temporal resolution (120+ FPS) infrared cameras
- Use markers that are highly reflective to infrared

Reflective markers





Multi-Camera System Localization

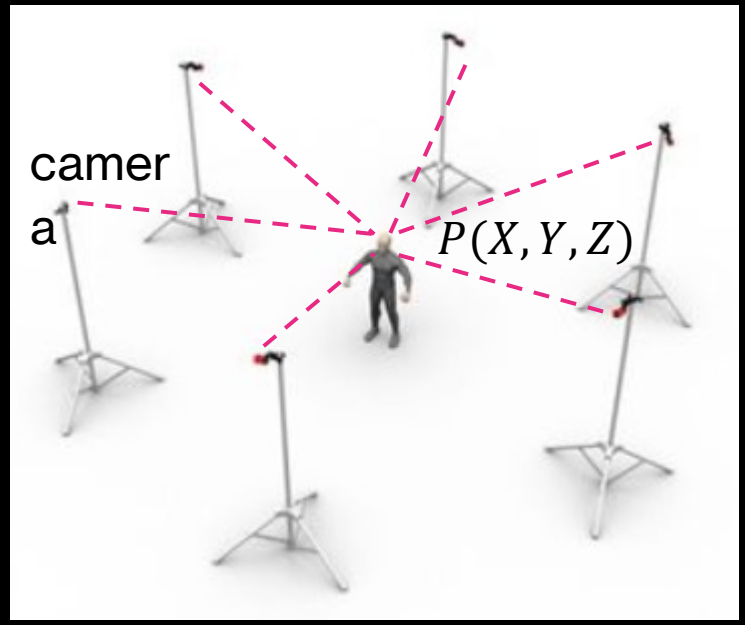
How to localize a point P?

Triangulation by minimize the reprojection error.

After calibration, we know the **intrinsic matrix** K_k and **extrinsic matrix** $[\mathbf{R}_k|\mathbf{t}_k]$ of the k-th camera

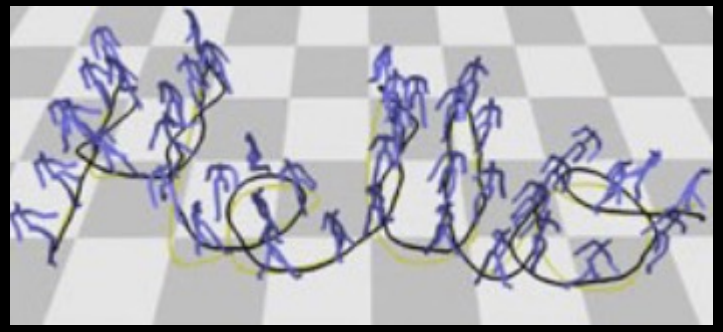
$$P = \arg\min_P \sum_k \|P_k - K_k[\mathbf{R}_k|\mathbf{t}_k]P\|^2$$

where P_k is the point in the k-th image



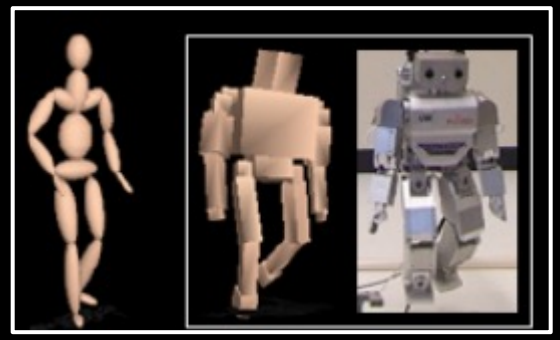
Multiple cameras observe the same point

Camera Systems for Motion Capture



Computer animation

Biomechanics



Robotics

Movies

Video games

Anthropology

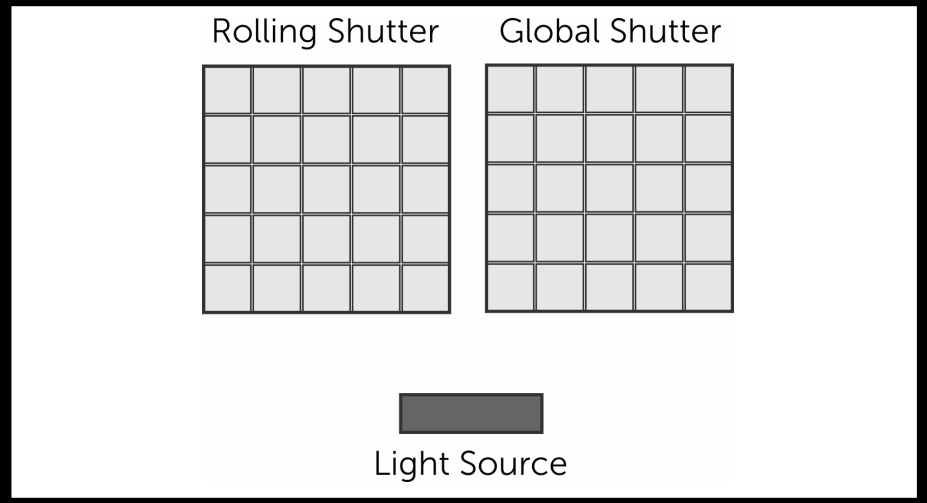
Camera System Calibration

Wave the wand to provide sufficient points for calibration

Camera Shutters

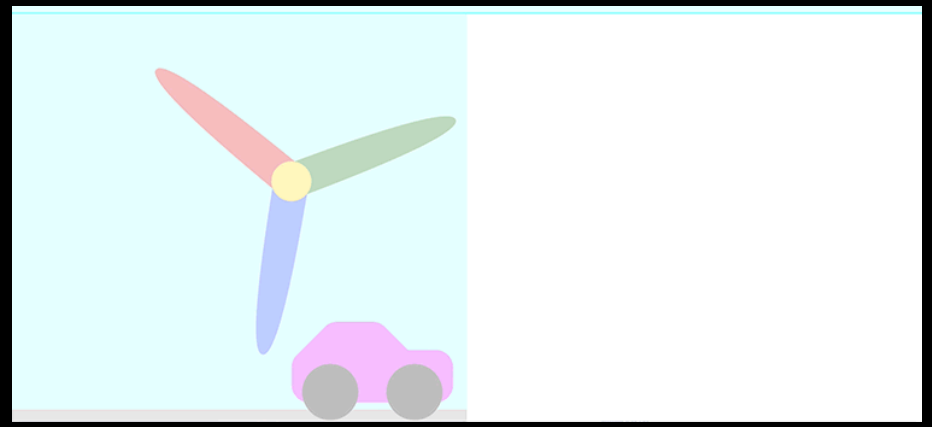
Global shutter cameras:

- All sensor pixels are read out **simultaneously**
- **Pros:** no time difference across the image
- **Cons:** slower frame rate, increased read noise, etc.



Rolling shutter cameras:

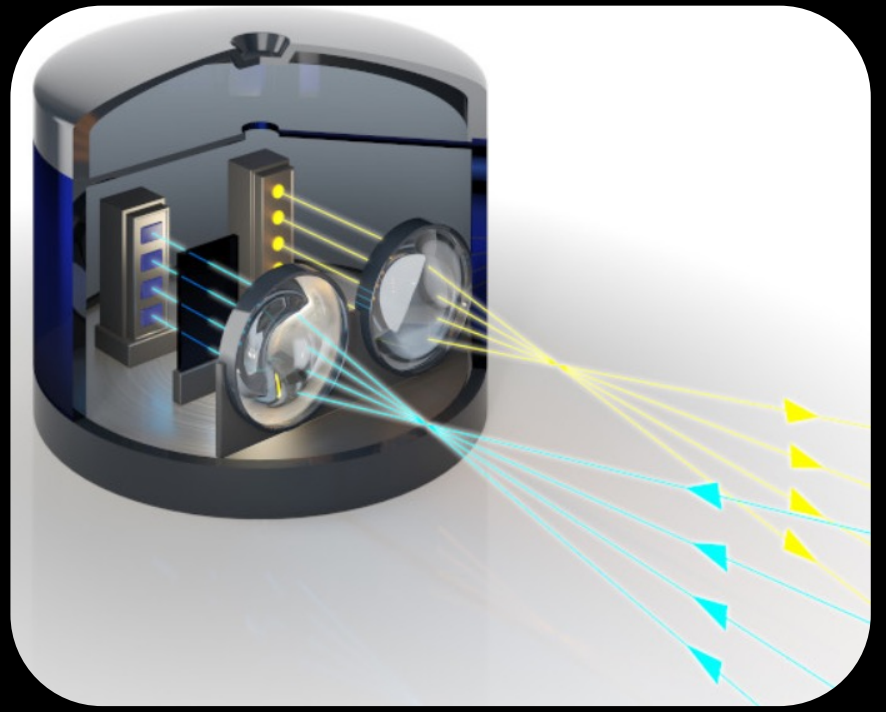
- Read out data row by row when exposed
- The rolling is very fast (~10 μs)
- Pros: capture information at higher frame rate
- Cons: artifacts for fast-moving objects



Global shutter Rolling shutter

LiDAR Basis

LiDAR (Light Detection and Ranging)

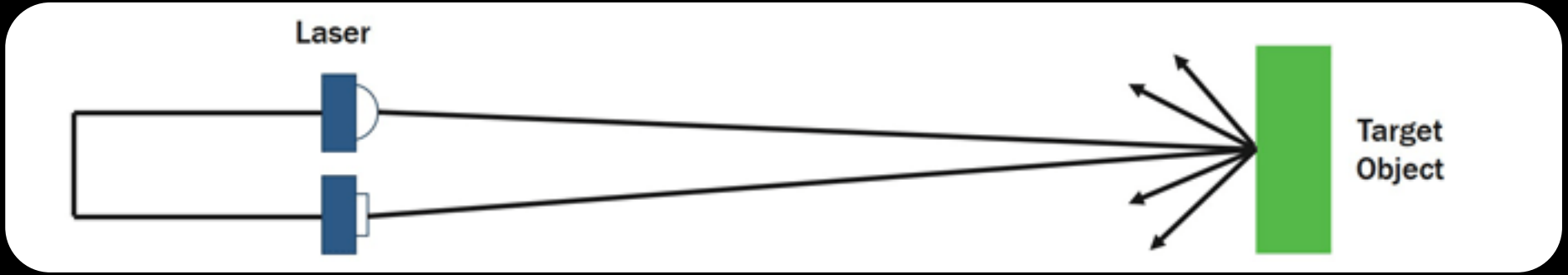


measure the distance for each direction

How does LiDAR detect the range?

LiDAR Basis

How does LiDAR detect the range?



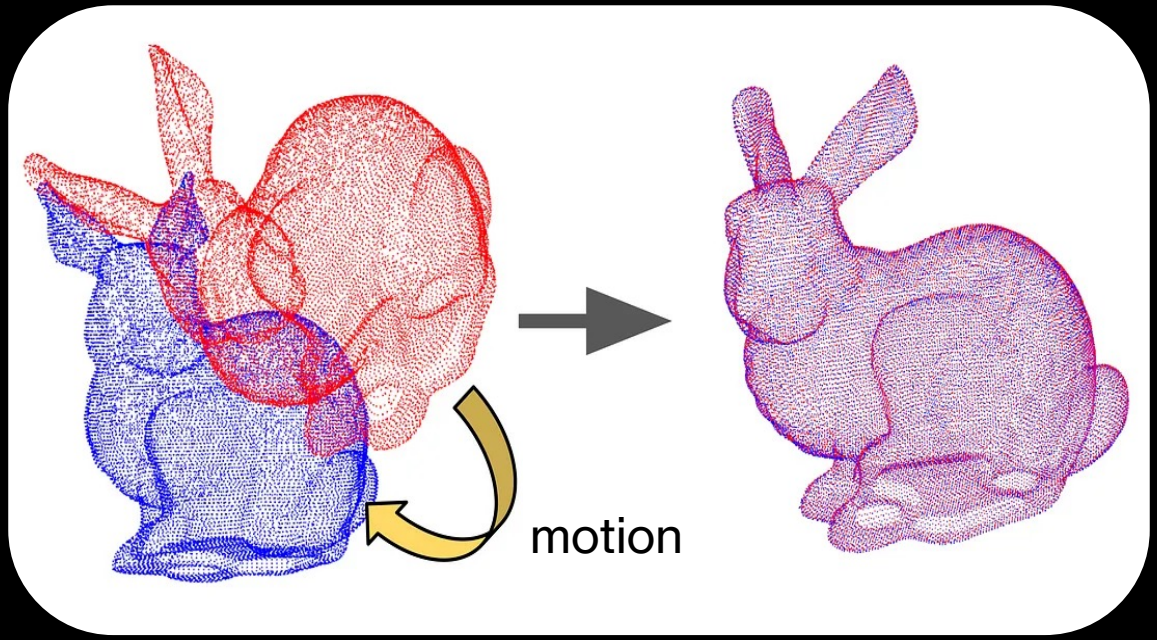
Round-Trip Distance = Reflection time x Speed of light

FMCW or AMCW encoding can be used to measure ToF

LiDAR Odometry: Registration

Estimate sensor motion give consecutive LiDAR measurements

- Point cloud registration
- Find the transformation (motion) that aligns two observed point clouds



LiDAR Odometry: Registration

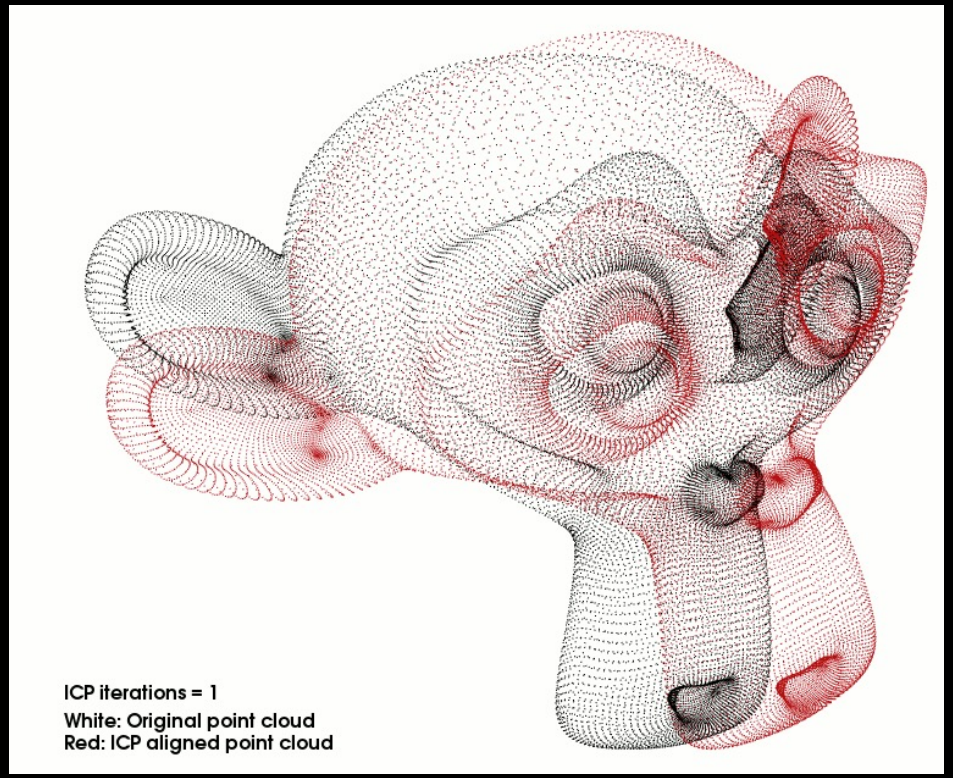
Iterative Closest Point (ICP) algorithm:

Determine correspondence set $\mathcal{K} = \{(\boldsymbol{p}, \boldsymbol{q})\}$ from point cloud $\boldsymbol{P}$ and $\boldsymbol{Q}$

Estimate motion

$$\mathbf{T} = \min_{\mathbf{T}} \sum_{(p,q)\in\mathcal{K}} \|\boldsymbol{p} - \mathbf{T}\boldsymbol{q}\|^2$$

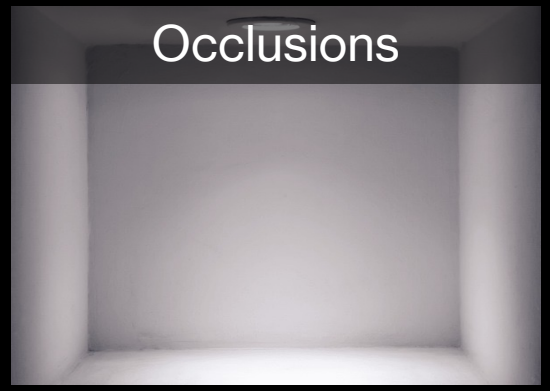
Apply $\mathbf{T}$ to $\boldsymbol{Q}$



Iterative Closest Point (ICP)

Limitations for Visual Sensors

Occlusions



Bad lighting: glare

Bad lighting: dimness

Fog

Sandstorm

Smoke & Fire

RF-based vision: see-through-x, but with limited angular resolution

Multi-modal sensing: come the best of visual, radio, acoustic, and inertial sensors

Objectives of This Module

Learn the fundamentals, applications, and implications of localization, motion tracking, and sensing

1. What are some motivating applications of localization and location-based services?
2. What are the unifying principles of positioning?
3. How do wireless positioning like GPS, Wi-Fi positioning, and Bluetooth ranging work?
4. What is wireless sensing?
5. How do visual positioning and tracking systems work?

Module review due 1 week after the last lecture

Module 1 due: Feb 5th 11:59PM

Next Lecture

- Time: Wed. Feb. 7th
- Module: Sensing
- Topic: Health & Vitals Sensing
- Readings & Questions: VitalRadio (details on the course website)