Structure-from-Motion, the Bootstrap process, and Self-Driving Cars

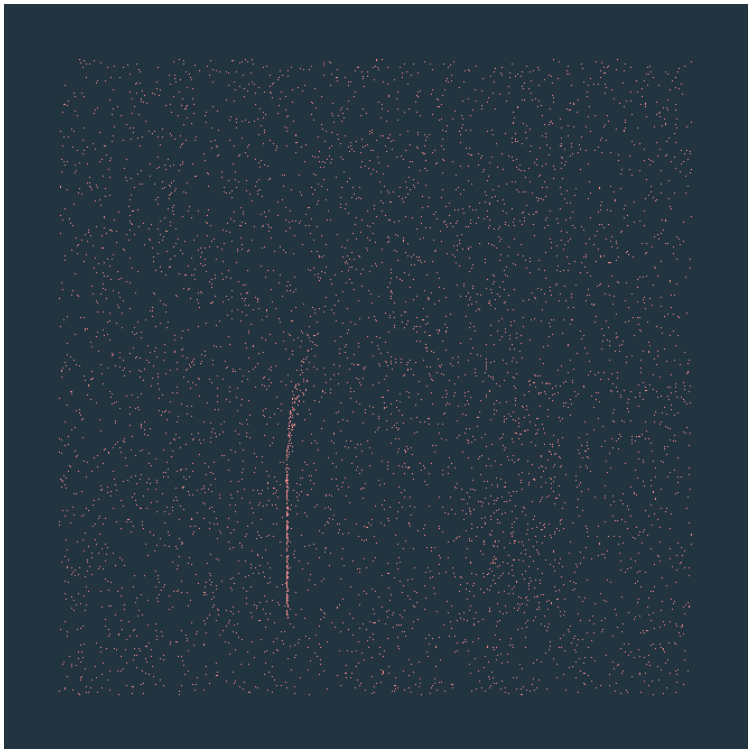
Michael Wang

What you should get out of this session

- What is the relationship between computer vision and biological vision?
- How does invariant over transformation manifest in 3D shape perception?

Two lectures ago

- Optic flow and structure-from-motion (SFM)



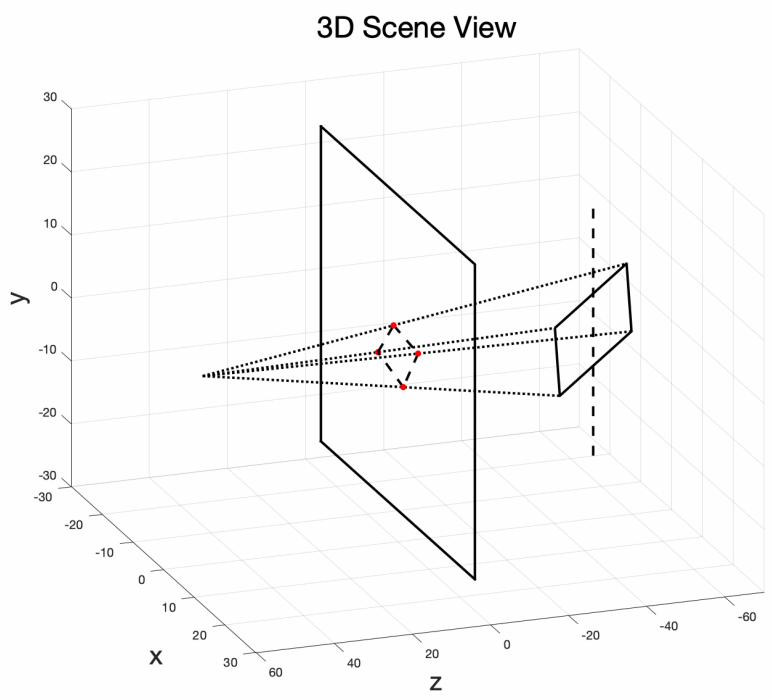
Structure-from-motion

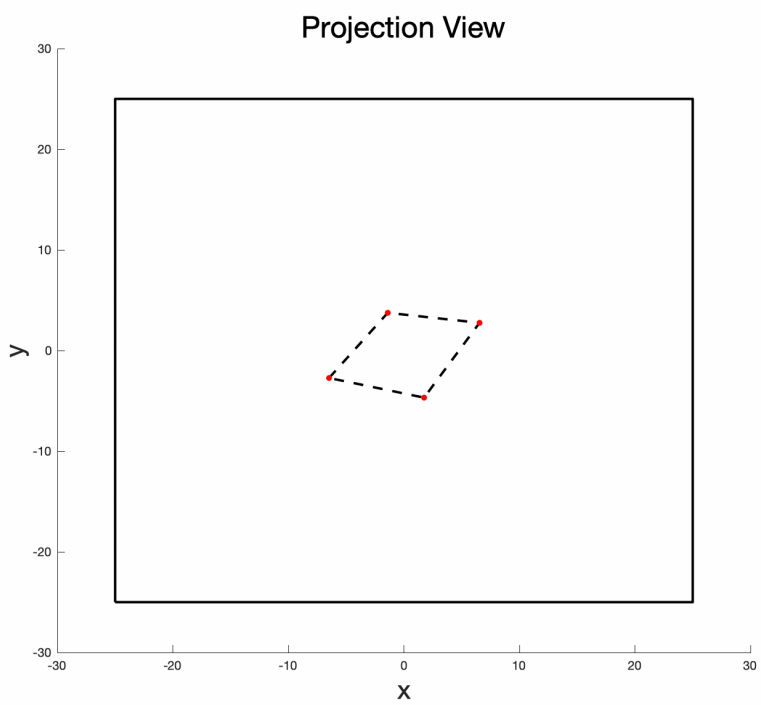
- Use a series of 2D images to reconstruct a 3D scene or object.
- Example algorithm: Lind (1996)
- You can also find other implementations
  - Koenderink, J. J., & Van Doorn, A. J. (1991). Affine structure from motion. *JOSA A*, *8*(2), 377-385.
  - Shapiro, L. S., Zisserman, A., & Brady, M. (1995). 3D motion recovery via affine epipolar geometry. *International Journal of Computer Vision*, *16*(2), 147-182.
- All papers are uploaded to Google Drive folder.

Different Types of Projection

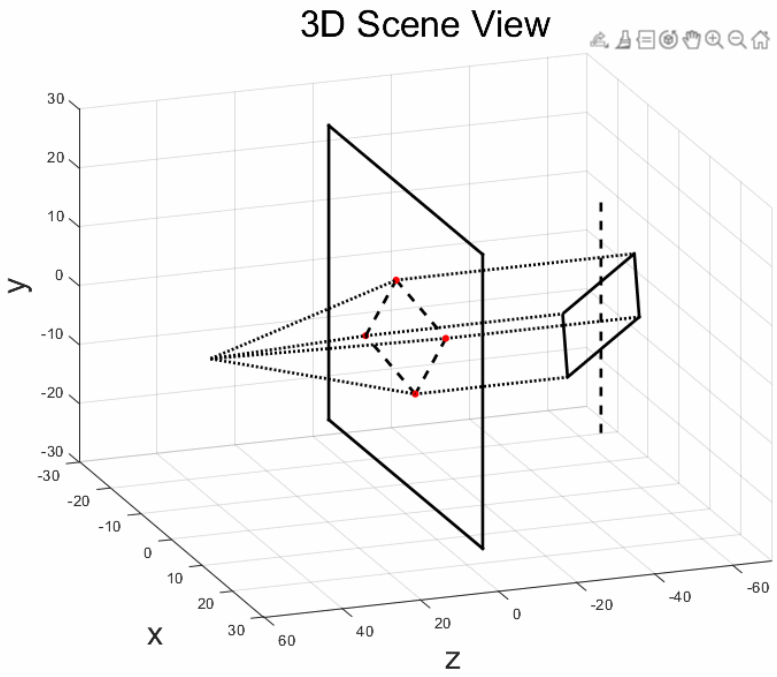
- Perspective Projection
- Orthographical Projection
- Scaled Orthographical Projection

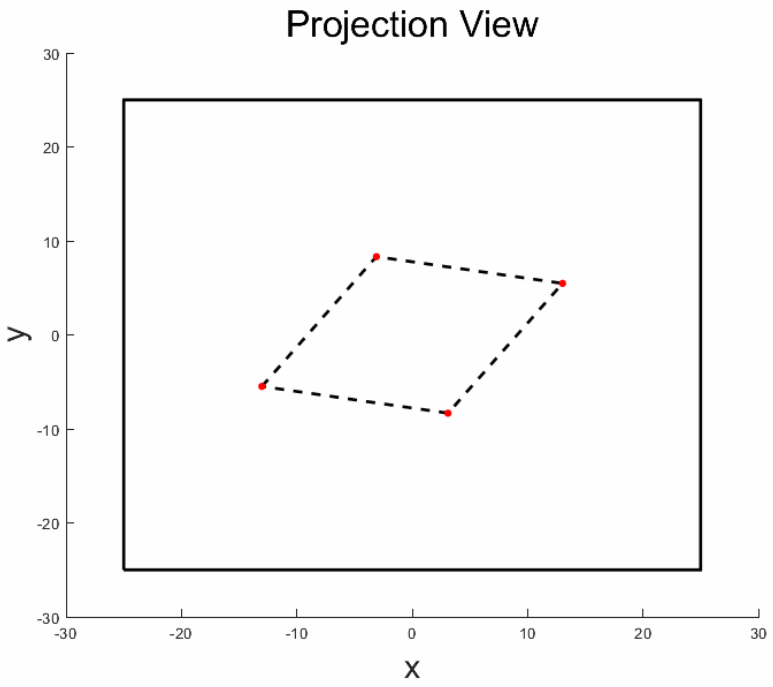
Perspective Projection



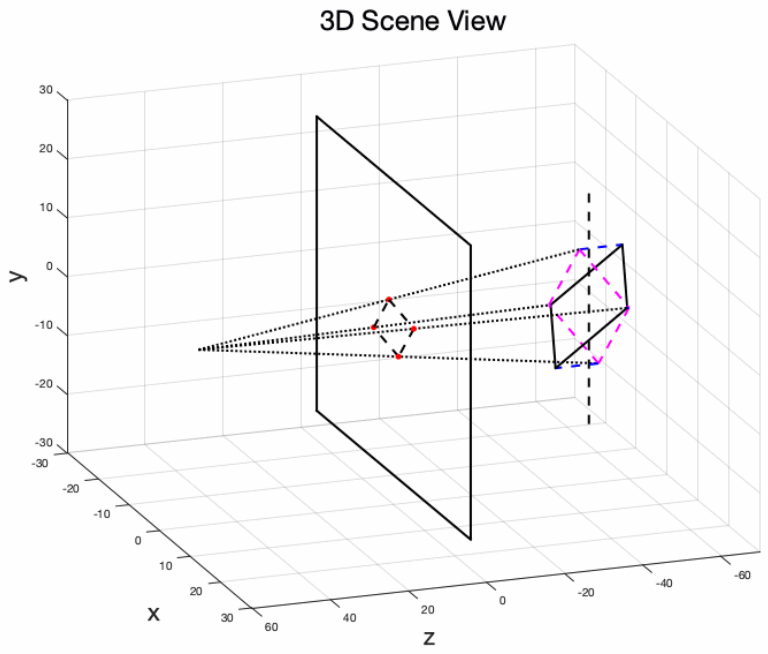


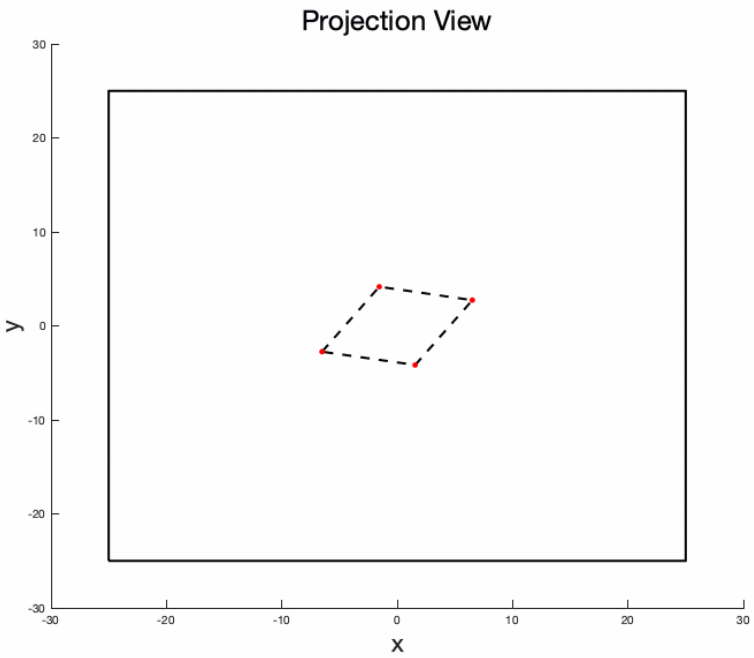
Orthographical Projection





Scaled Orthographical Projection





Motion Decomposition

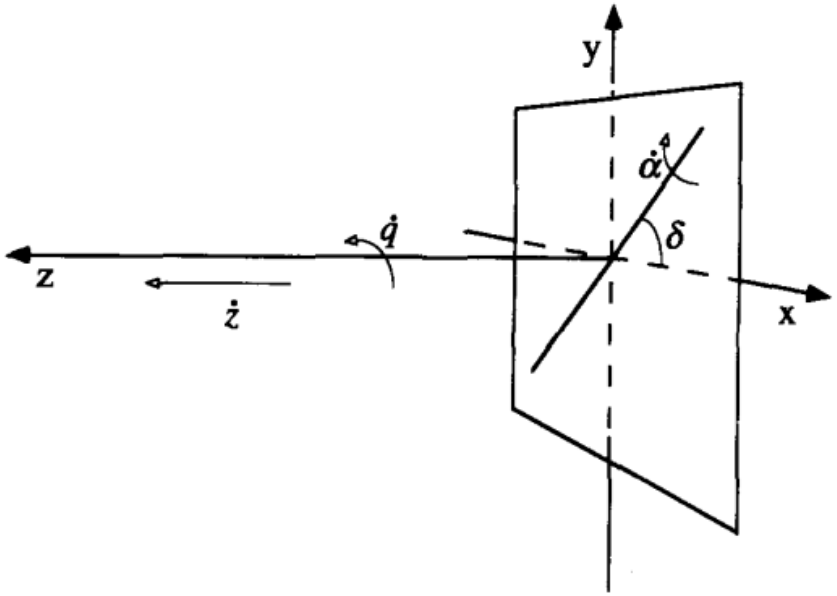


Figure 2. The parameters describing the relative motion.

1. Rotation around the z-axis

$$\dot{x} = -y\dot{q}$$
$$\dot{y} = x\dot{q}$$

2. Translation (of the projection point) along the z-axis

$$\dot{x} = x\dot{z} + x\dot{z}Z$$
$$\dot{y} = y\dot{z} + y\dot{z}Z$$

Note: Z is actually $\frac{Z}{1-Z}$, which is distance scaled by the distance to the unit projection surface.

3. Rotation of the world point around an axis in the xy plane with an angle of δ. Let $\delta = 90°$

$$\dot{x} = x^2\dot{\alpha} - \dot{\alpha}Z$$
$$\dot{y} = xy\dot{\alpha}$$

4. Any motion

$$\dot{x} = x^2\dot{\alpha} - \dot{\alpha}Z + x\dot{z} + x\dot{z}Z - y\dot{q}$$
$$\dot{y} = xy\dot{\alpha} + y\dot{z} + y\dot{z}Z + x\dot{q}$$

Lind, 1996

Small visual angles and weak perspective

- With small visual angles (≤ 5°; weak perspective/scaled orthographical projection; x and y are much smaller than 0), a few terms can be approximated to be 0:

$$x^2\dot{\alpha} \approx 0 \; and \; xy\dot{\alpha} \approx 0$$

Therefore:

$$\dot{x} \approx -\dot{\alpha}Z + x\dot{z} + x\dot{z}Z - y\dot{q}$$
$$\dot{y} \approx y\dot{z} + y\dot{z}Z + x\dot{q}$$

- Additionally, given weak perspective, we can also approximate the effect of the translational velocity along the z-axis to be expressed by only $y\dot{z}$ and $x\dot{z}$:

$$\dot{x} \approx -\dot{\alpha}Z + x\dot{z} - y\dot{q}$$
$$\dot{y} \approx y\dot{z} + x\dot{q}$$

Depth reconstruction

- We have these equations, but now what?

$$\dot{x} \approx -\dot{\alpha}Z + x\dot{z} - y\dot{q}$$
$$\dot{y} \approx y\dot{z} + x\dot{q}$$

- With 2D images, we can directly obtain $x, y, \dot{x}, \dot{y}$ (given that we solve the correspondence problem). Therefore, we can use regression to solve for $\dot{z}_{est}$ and $\dot{q}_{est}$.
- However, this leaves us with on undetermined variable $\dot{\alpha}$, i.e. the reconstructed depth is always scaled by an unknown scaling factor $\dot{\alpha}$:

$$\dot{\alpha}Z = -\dot{x} + x\dot{z}_{est} - y\dot{q}_{est}$$

Relief Depth Compression

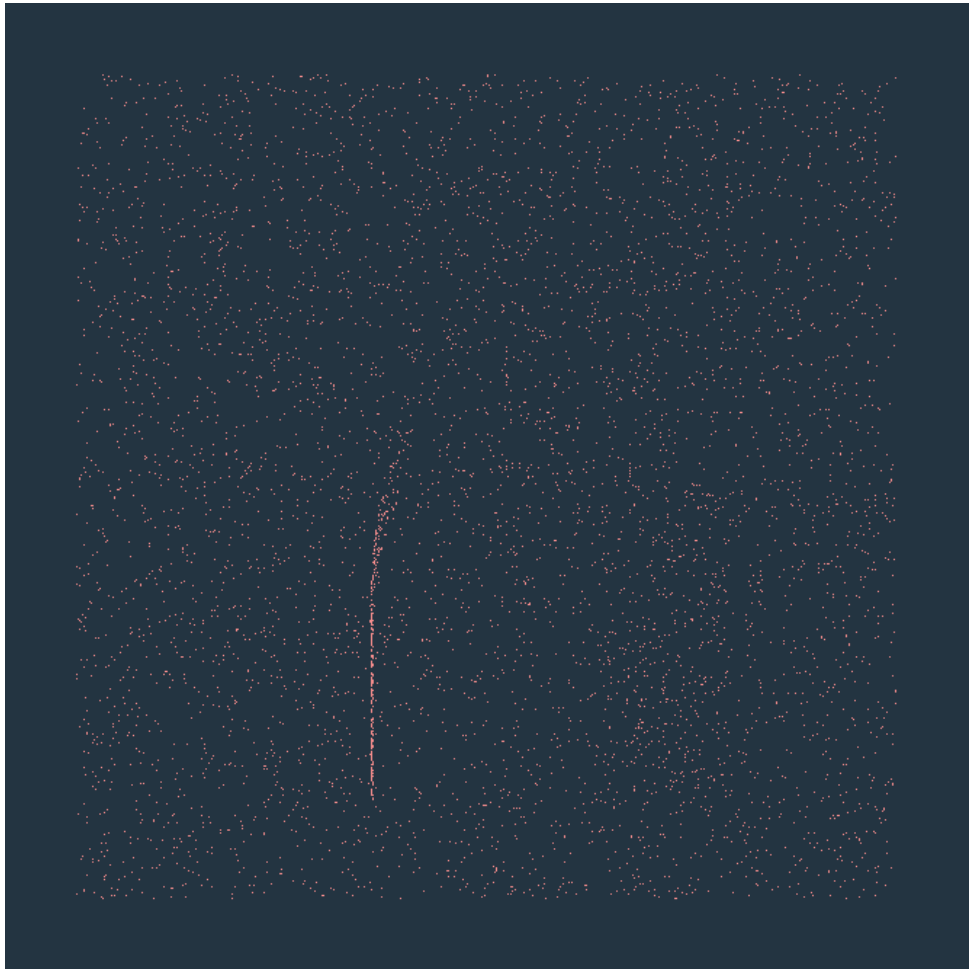
Toronto, Jaume Plensa


Sheraton

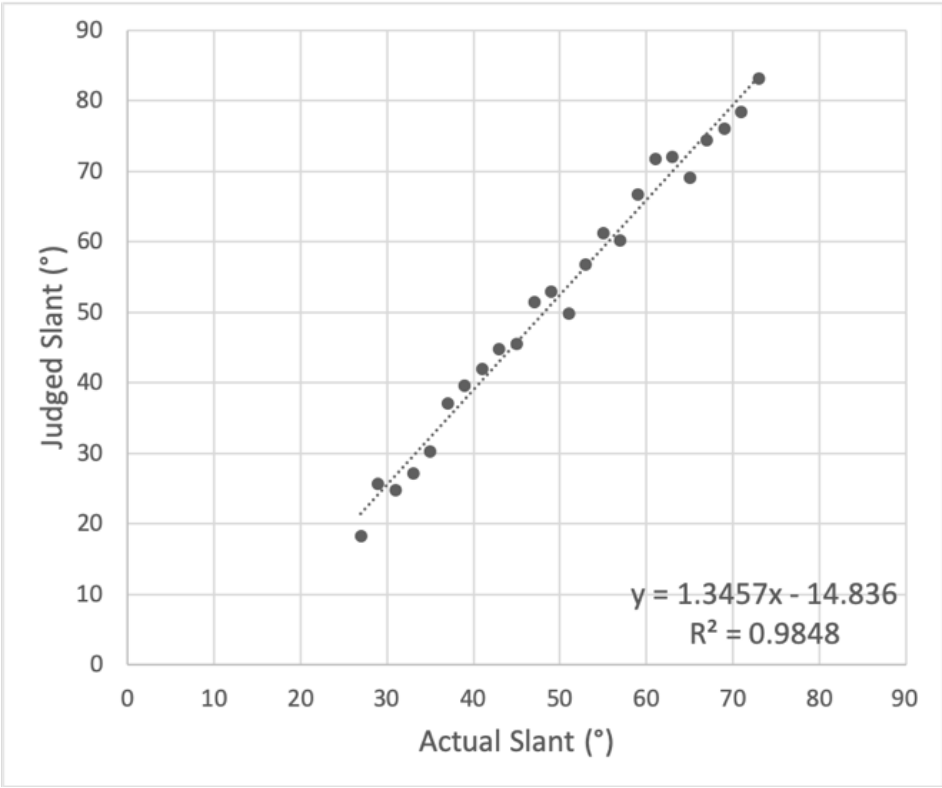

THE KEG

Slant perception

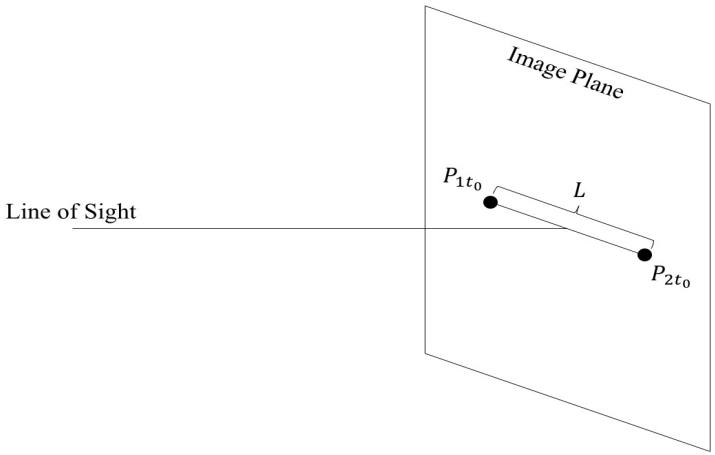
- What is the slant angle of the top surface?
- Try with your palm.



Slant perception



The Bootstrap Process



$$L_0 = \sqrt{(x_{A0} - x_{B0})^2 + (y_{A0} - y_{B0})^2}$$

$$L_t = \sqrt{(x_{At} - x_{Bt})^2 + (y_{At} - y_{At})^2 + (Z_{At} - Z_{Bt})^2}$$

$$Z = \frac{1}{\dot{\alpha}} z$$

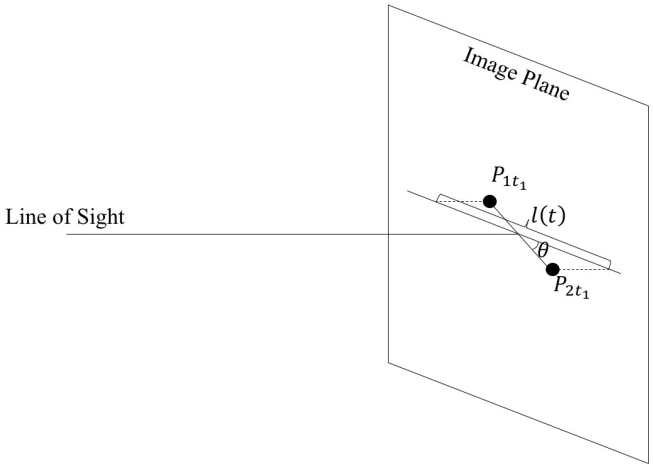
$$L_t = \sqrt{(x_{At} - x_{Bt})^2 + (y_{At} - y_{At})^2 + \frac{1}{\dot{\alpha}^2}(z_{At} - z_{Bt})^2}$$

$$L_0 = L_t$$

$$\dot{\alpha} = \frac{|z_{At} - z_{Bt}|}{\sqrt{L_0^2 - l_t^2}}$$

Wang, Lind, & Bingham (2018, 2019, 2020)

The Bootstrap Process



$$L_0 = \sqrt{(x_{A0} - x_{B0})^2 + (y_{A0} - y_{B0})^2}$$

$$L_t = \sqrt{(x_{At} - x_{Bt})^2 + (y_{At} - y_{At})^2 + (Z_{At} - Z_{Bt})^2}$$

$$Z = \frac{1}{\dot{\alpha}} z$$

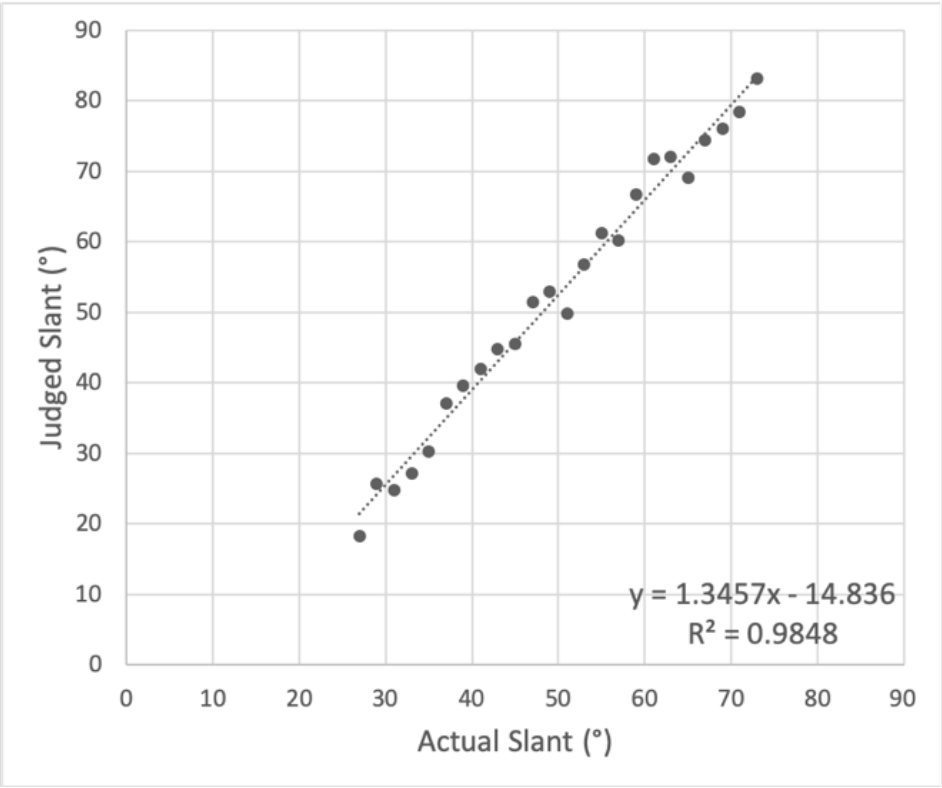
$$L_t = \sqrt{(x_{At} - x_{Bt})^2 + (y_{At} - y_{At})^2 + \frac{1}{\dot{\alpha}^2}(z_{At} - z_{Bt})^2}$$

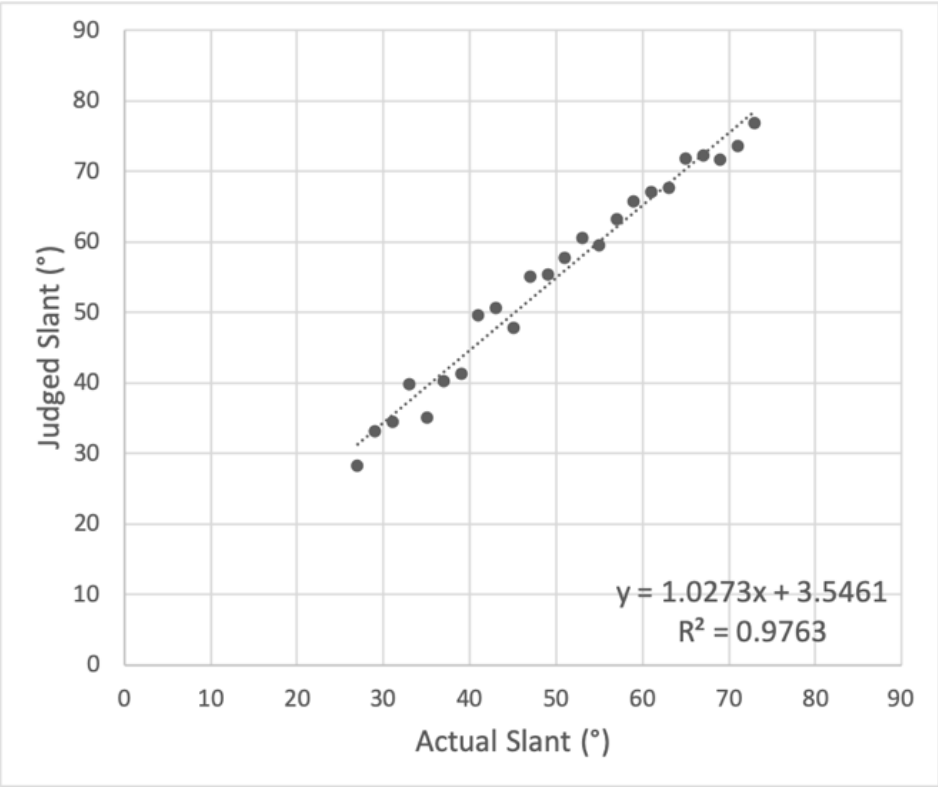
$$L_0 = L_t$$

$$\dot{\alpha} = \frac{|z_{At} - z_{Bt}|}{\sqrt{L_0^2 - l_t^2}}$$

Wang, Lind, & Bingham (2018, 2019, 2020)

Slant perception

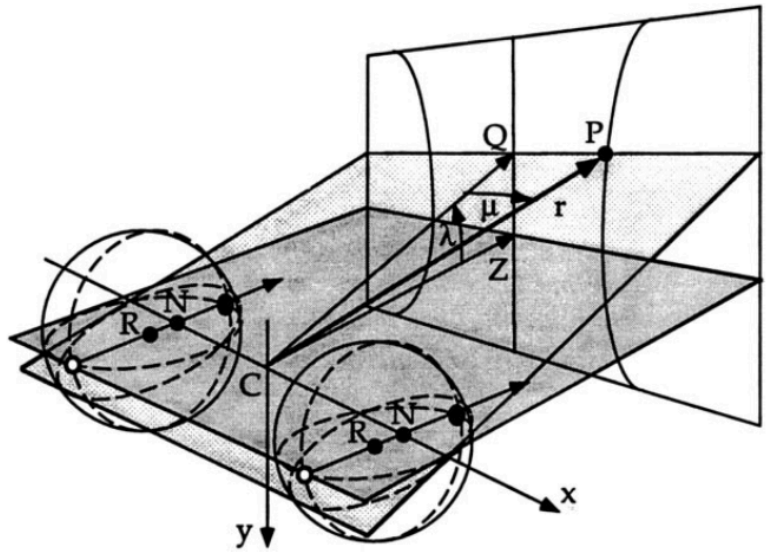




Recall from a previous lecture…

- Stereopsis (binocular disparity)
- If IPD is known, we can uniquely determine distance based on disparity.

$$\tan \delta_{\mu} = \frac{IPD \cos \mu}{r^2 - \frac{IPD^2}{4}}$$



Erkelens & van Ee (1998)

Why does it matter?

- Tesla and it's vision-based self-driving
  - "You are not shooting lasers out of your eyes."
- Zhou, Brown, Snavely, & Lowe (*CVPR* 2017)
  - Unsupervised learning
  - Uses sequences of images as input and aims to explain them by predicting likely camera motion and the scene structure.
    - An estimate of ego-motion (6 DoF transformation matrices)
    - The underlying scene structure (per-pixel depth map under a reference view).
  - "A geometric view synthesis system only performs consistently well when its intermediate predictions of the scene geometry and the camera poses correspond to the physical ground-truth."

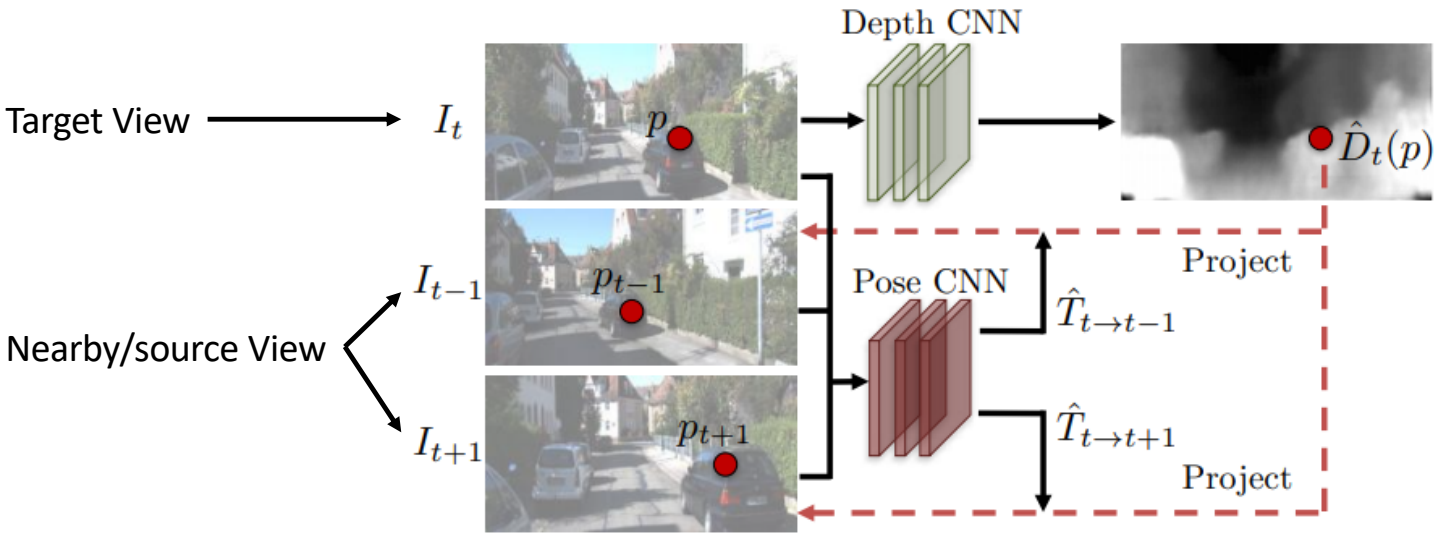


Figure 2. Overview of the supervision pipeline based on view synthesis. The depth network takes only the target view as input, and outputs a per-pixel depth map $\hat{D}_t$. The pose network takes both the target view (I_t) and the nearby/source views (e.g., I_{t-1} and I_{t+1}) as input, and outputs the relative camera poses ($\hat{T}_{t\to t-1}, \hat{T}_{t\to t+1}$). The outputs of both networks are then used to inverse warp the source views (see Sec. 3.2) to reconstruct the target view, and the photometric reconstruction loss is used for training the CNNs. By utilizing view synthesis as supervision, we are able to train the entire framework in an unsupervised manner from videos.

Zhou et al., *CVPR* 2017.

What you should get out of this session

- What is the relationship between computer vision and biological vision?
  - One borrows from the other.
  - SFM comes from computer vision (Ullman, 1979) but has been tested in human visual perception. SFM's principles have been supplying behavioral scientists a lot of new hypotheses.
- How does invariant over transformation manifest in 3D shape perception?
  - The bootstrap process
  - You start with an inaccurate perceived 3D structure, *assuming rigidity*, you see how this perceived structure changes through time and obtain a perceived structure that is more consistent through time.

Dr. Andrew Clement

- The effects of typicality on attention and awareness for object categories.
- Maxfield, Stalder, & Zelinsky (2014)
  - Typicality rating task
  - Search Task
  - Comparison with compute vision algorithms
    - Whether it is reasonable to use computer vision algorithms (scale-invariant feature transform and linear-kernel support vector machine) to predict behavioral typicality judgments.
- What to look for?
  - What is visual attention?
  - Why does it matter?
  - How does it fit into what we have learned so far?

See you next time!