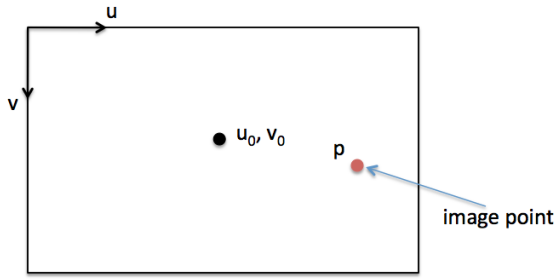
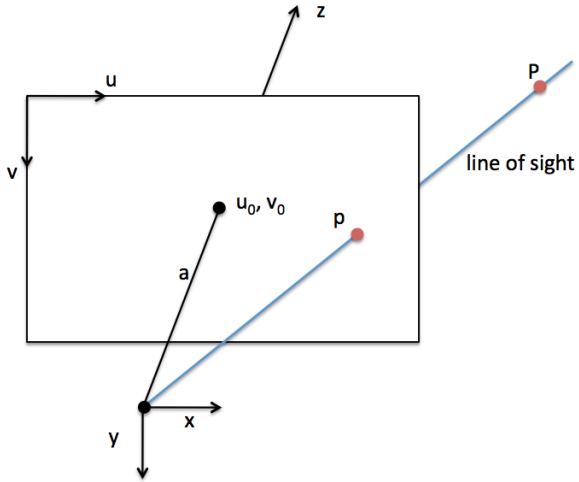
5. Fusion of Audio and Vision

1. Audio-visual processing challenges
2. Representation of visual information
3. **The geometry of vision**
4. Audio-visual feature association
5. Audio-visual alignment
6. Visually-guided audio localization
7. Audio-visual event localization
8. Audio-visual clustering
9. Conclusions

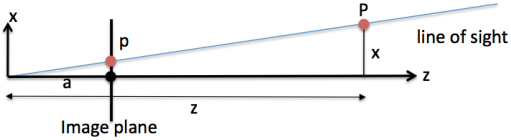
The Image Model



The Camera Model



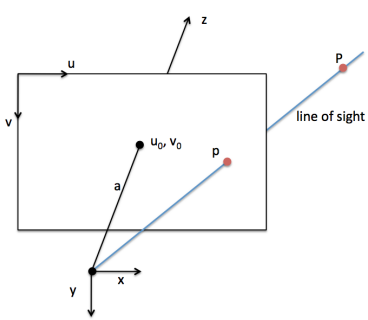
Perspective Camera Model



- Camera parameters: focal length a and image center $(u_0,\ v_0)$
- Coordinates of p in the image plane: $\boldsymbol{p} = (u - u_0,\ v - v_0)^\top$
- Coordinates of P in the camera frame: $\boldsymbol{P} = (x,\ y,\ z)^\top$
- From similar triangles we obtain:

$$\frac{u - u_0}{a} = \frac{x}{z} \quad \text{and} \quad \frac{v - v_0}{a} = \frac{y}{z}$$

Line of Sight

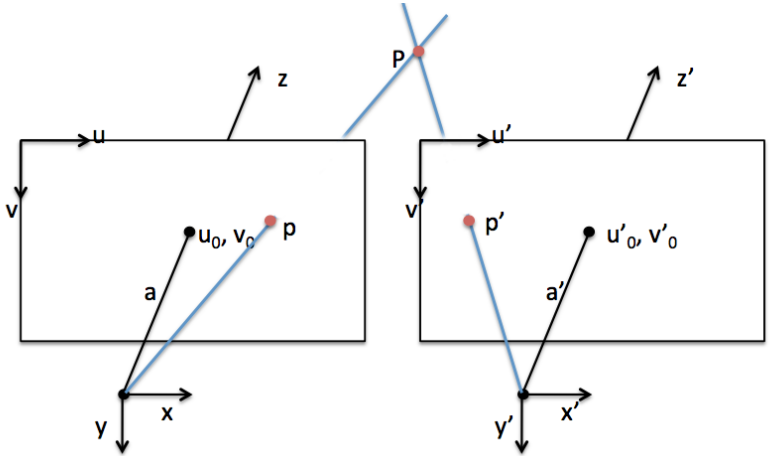


- The line of sight $\mathcal{L}$ through image point p is defined by **camera parameters**, $(a,\ u_0,\ v_0)$:

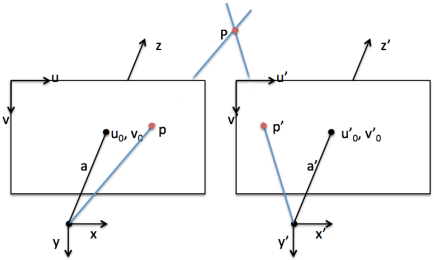
$$P \in \mathcal{L} := \begin{cases} ax - (u - u_0)z = 0 \\ ay - (v - v_0)z = 0 \end{cases}$$

- $(a,\ u_0,\ v_0)$ estimated using calibration

Binocular Camera Pair



Binocular Reconstruction Principle



- Point P is at the intersection of two lines of sight, through p (left) and p' (right)
- Point P in the left camera frame: $P = (x,\ y,\ z)^\top$
- Point P in the right camera frame: $P = (x',\ y',\ z')^\top$
- The two lines of sight must be represented in the same frame (left or right)

Binocular Calibration

- Let $\mathbf{R}$ and $\boldsymbol{t}$ be the rotation matrix (3×3) and translation vector between left and right:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{R} \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} + \boldsymbol{t}$$

- $\mathbf{R}$ and $\boldsymbol{t}$ are the parameters of the binocular system, they are estimated by calibration.
- $\mathbf{R} = [r_{ij}]_{i,j=1}^{i,j=3}$, $\boldsymbol{t} = (t_1,\ t_2,\ t_3)^\top$.

Point Reconstruction

- Point P is the intersection of two lines of sight represented in the same coordinate frame (left camera):

$$P := \begin{cases} ax - (u - u_0)z = 0 \\ ay - (v - v_0)z = 0 \\ a'(r_{11}x + r_{12}y + r_{13}z + t_1) - (u' - u'_0)(r_{31}x + r_{32}y + r_{33}z + t_3) \\ a'(r_{21}x + r_{22}y + r_{23}z + t_2) - (u' - u'_0)(r_{31}x + r_{32}y + r_{33}z + t_3) \end{cases}$$

- This enables reconstruction of P from the left/right pair of image points p, p'.

Session Summary

- Image model
- Camera model
- Binocular model
- Camera calibration
- Binocular reconstruction