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

Combining visual and audio data

- The camera-microphone calibration method that we studied allows reconstruct a visual object and to predict its associated TDOA.
- We consider a setup composed of **two cameras** and **two microphones**.
- The microphone positions are known in the camera coordinate system (calibration process).
- This allows to associate a visual object to an audio source.

Associating a TDOA with a Visual Object

- A point $\boldsymbol{S}$ that belongs to a visible object can be reconstructed from its images in the left and right cameras.
- If there is a sound source located at $\boldsymbol{S}$, then one can predict the TDOA associated with a microphone pair:

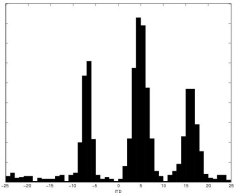
$$\tau(\boldsymbol{S}) = \frac{\|\boldsymbol{S} - \boldsymbol{M}_1\| - \|\boldsymbol{S} - \boldsymbol{M}_2\|}{\nu}$$

- The TDOA of a sound-source can be estimated with:

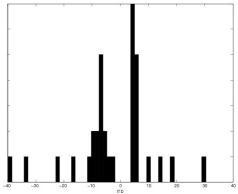
$$\hat{\tau} = \underset{t'}{\operatorname{argmax}}\, \mathrm{NCC}\Big(x_i(t)x_j(t-t')\Big)$$

- If $\hat{\tau} \approx \tau(\boldsymbol{S})$, then an audio-visual object may be at $\boldsymbol{S}$.

Mapping People Onto Sounds

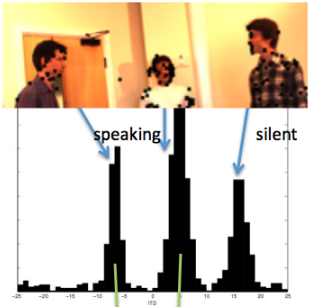


Histogram of visual data



Histogram of estimated TDOAs

Audio-Visual Association

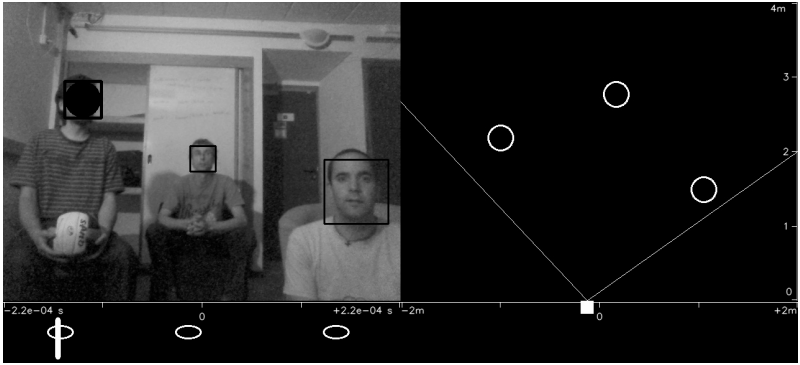


visual data

visual-data histogram

TDOA histogram

Faces That Speak



Video

play humanoids_video.avi

Session Summary

- Two cameras and two microphones
- Estimate the 3D position of a visual object from the two cameras
- Map this 3D position onto the TDOA axis
- Combine the mapped visual object with the estimated TDOA
- associate a person with each TDOA