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

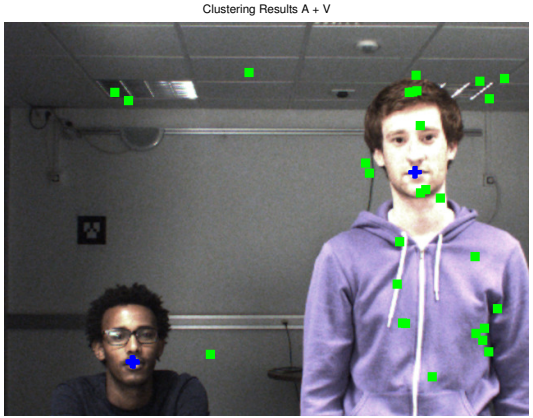
Audio-Event Localization

- An audio event that is recorded with four microphones can be localized in the image plane.
- The cross-correlation function consider a short time interval to estimate the TDOA values.
- Over time, there are many such **audio features** available in the image plane.
- Because of background noise and reverberations, the localization is corrupted by errors.

Visual Feature Localization

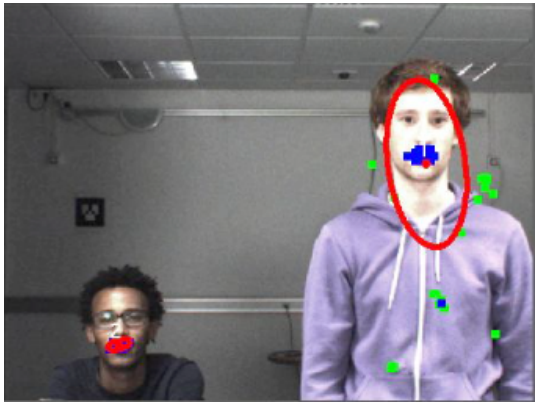
- Face detection and localization
- Face landmarks: lips, eyes, etc.
- In particular, lip detection and localization is quite reliable.

Auditory and Visual Features Side-by-Side



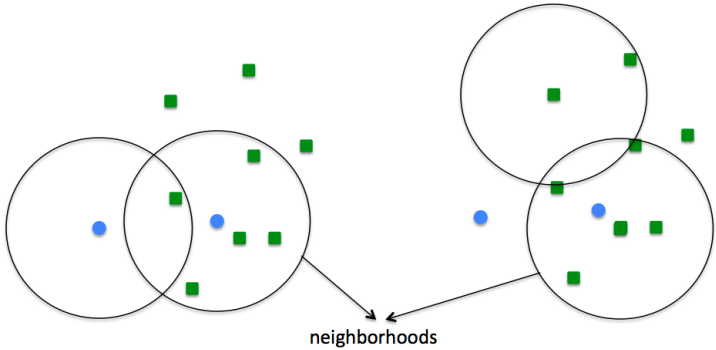
- Visual features (lips): **blue**.
- Audio features (audio events): **green**.

Audio-Visual Clustering



- First cluster contains only visual features (silent person).
- Second cluster contains both visual and audio features (speaking person).

Audio-Visual Weights



- The weight of a visual feature i:

$$w_i = \sum_{j \in \mathrm{N}(i)} \exp^{-d^2(\boldsymbol{X}_i, \boldsymbol{X}_j)}$$

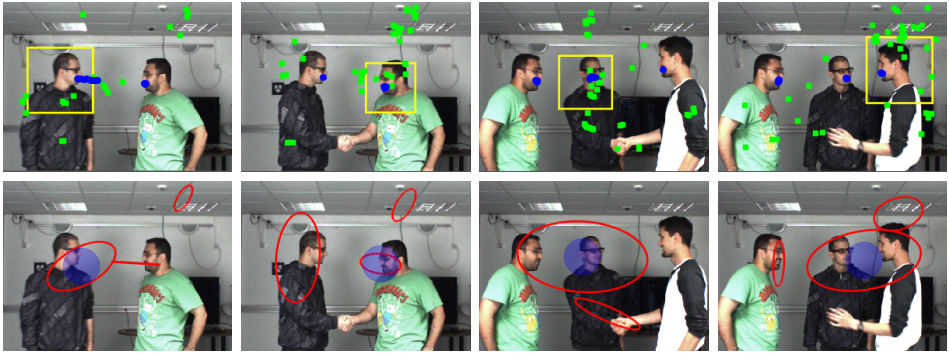
Weighted-Data Gaussian Mixture Model

- Each feature $\boldsymbol{x}_i$ (audio or visual) has a weight w_i and this can be plugged into a GMM:

$$P(\boldsymbol{x}_i|w_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \frac{1}{w_i}\boldsymbol{\Sigma}_k\right)$$

- A weighted-data expectation-maximization algorithm can be used to find **audio-visual clusters**.

Example



Video

Play cocktail-party.mp4

Session Summary

- Audio and visual features in the image plane
- Weighting the features
- Weighted-data Gaussian mixture
- Audio-visual clustering

Week Summary

- Auditory analysis, visual analysis, audio-visual analysis.
- audio-visual feature association.
- Cameras and camera-microphone arrangements.
- Audio-visual alignments.

Week Summary (Continued)

- Visually-guided audition.
- Audio-visual event localization.
- Audio-visual clustering.
- Example of solving a complex audio-visual task.