Audio-Visual Tracking by Density Approximation In a Sequential Bayesian Framework

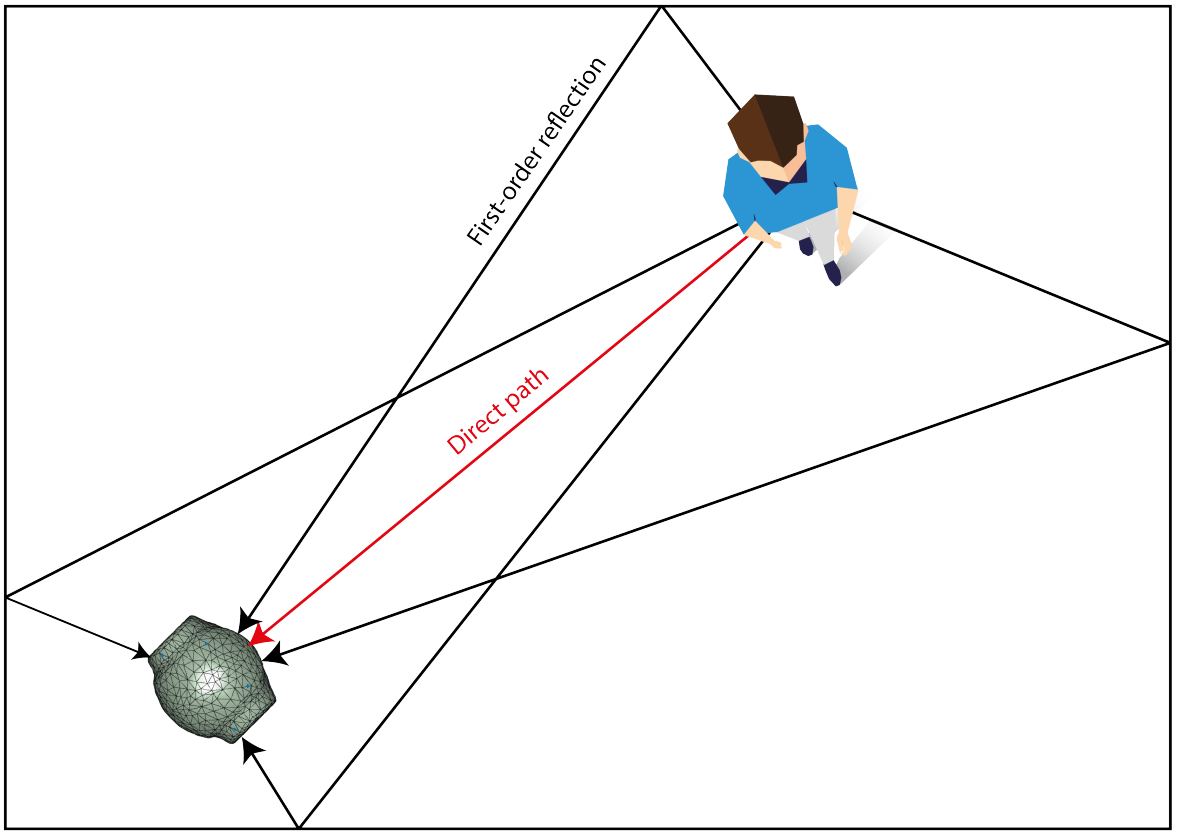
Israel D. Gebru[1], Christine Evers[2], Patrick A. Naylor[2], Radu Horaud [1]

[1]INRIA Grenoble Rhône-Alpes, France
[2]Imperial College London, Department of Electrical and Electronic Engineering, UK

March 3, 2017

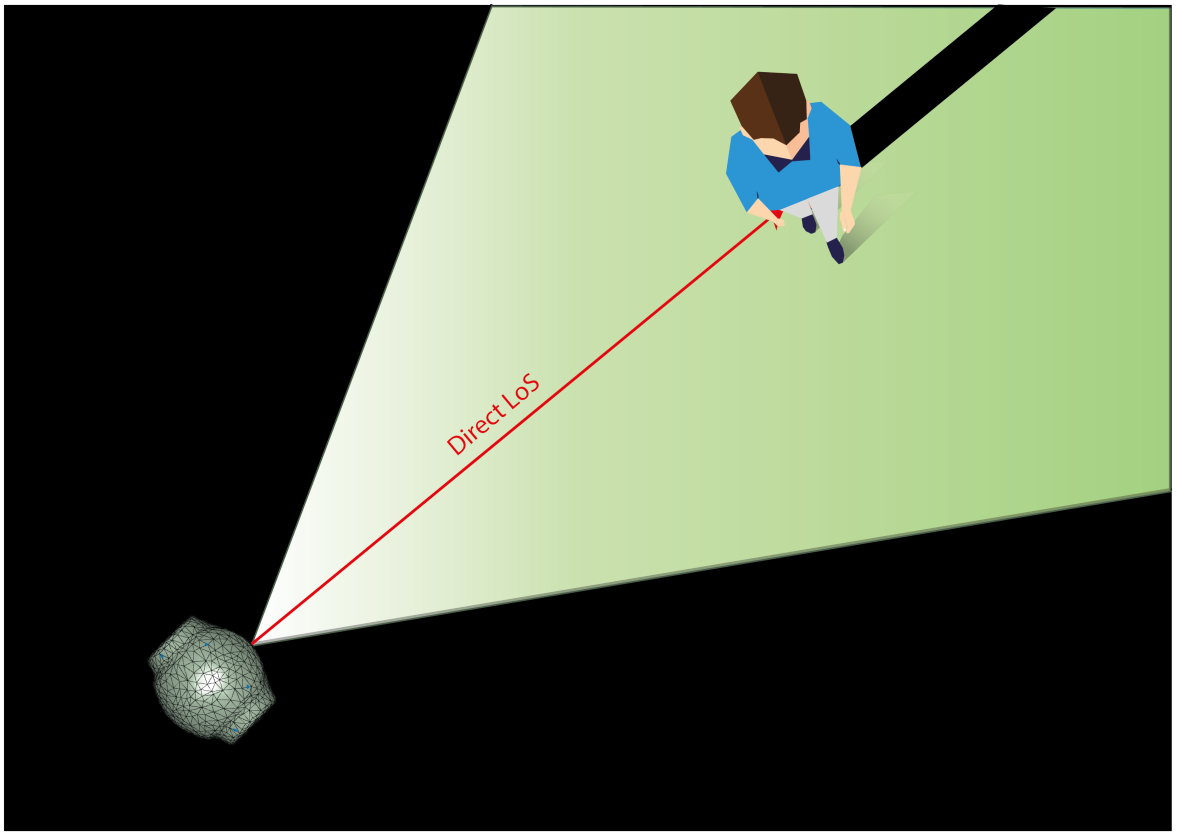
Audio Tracking Challenges

- **Late reverberation & ambient noise:** Localization errors
- **Early reflections:** Spurious detections
- **Speech inactivity:** Missing detections



Visual Tracking Challenges

- **Limited Field of View (FoV):** Blind outside
- **Head / body rotations:** Missing face detections



Objective

Aim: Fuse both modalities for improved source tracking

- Audio used for sources outside camera FoV
- Vision used to disambiguate acoustic features subject to reverb, noise, interference

Challenge:

- How can audio and visual signals be treated harmoniously
- How can any arising non-linearities be overcome

Proposed Model

- Augment sound source localization estimates and face detections:

$$\mathbf{Z}_t = \{\mathbf{Z}_t^a, \mathbf{Z}_t^v\}$$

- Track two-dimensional ($2d$) locations of humans in image plane, $\mathbf{X}$:

$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t}) \propto \mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t)\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1})$$

$$\mathbf{Z}_t^a = h_a(\mathbf{X}_t, \mathbf{U}_t^a) \quad \text{(Audio DoAs)}$$
$$\mathbf{Z}_t^v = h_v(\mathbf{X}_t, \mathbf{U}_t^v) \quad \text{(Visual face detections)}$$

$$\mathbf{X}_t = f(\mathbf{X}_{t-1}, \mathbf{V}_t) \quad \text{(Process model)}$$

Problem Formulation

Models:

- Unknown and potentially non-linear source dynamics
- Non-linear mapping between audio measurements and state space
- Posterior pdf is analytically intractable

Non-linear system: Naturally lends itself to importance sampling, but don't know how to sample

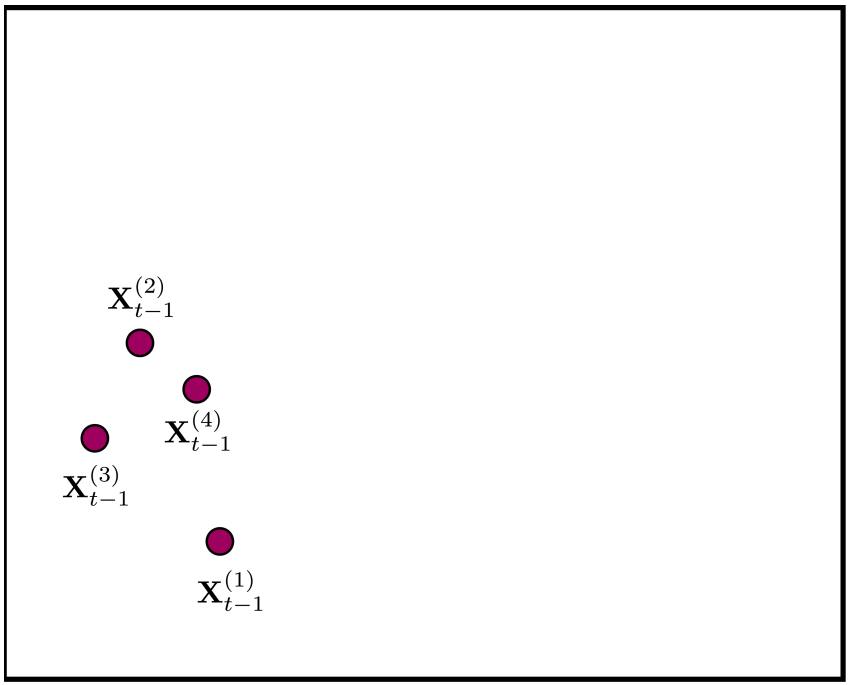
Approach:

1. **Unknown, non-linear transition:** Approximate predicted pdf using the Unscented Transform
2. **Non-linear likelihood:** Approximate by density interpolation.
3. **Mixture reduction:** Retain only peaks in pdf, obtained from Expectation-Maximization

Approximating the Predicted pdf

Predicted pdf: $$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \int \overbrace{\mathrm{P}(\mathbf{X}_t|\mathbf{X}_{t-1})}^{\text{Unknown}} \overbrace{\mathrm{P}(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})}^{\text{GMM}} \mathrm{d}\mathbf{X}_{t-1}$$

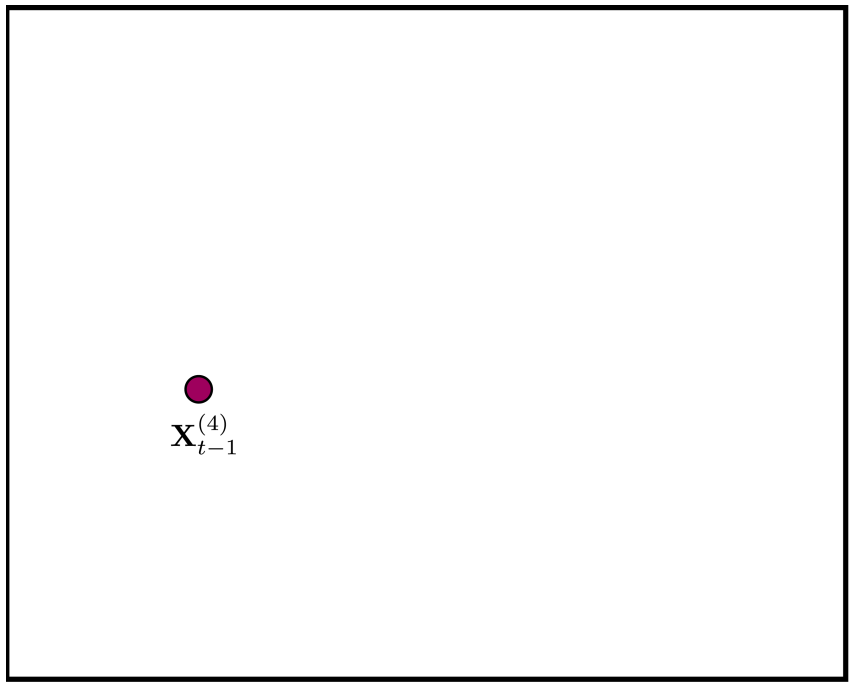
Unscented Transform: Easier to approximate pdf than nonlinear function



Approximating the Predicted pdf

Predicted pdf: $$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \int \overbrace{\mathrm{P}(\mathbf{X}_t|\mathbf{X}_{t-1})}^{\text{Unknown}} \overbrace{\mathrm{P}(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})}^{\text{GMM}} \mathrm{d}\mathbf{X}_{t-1}$$

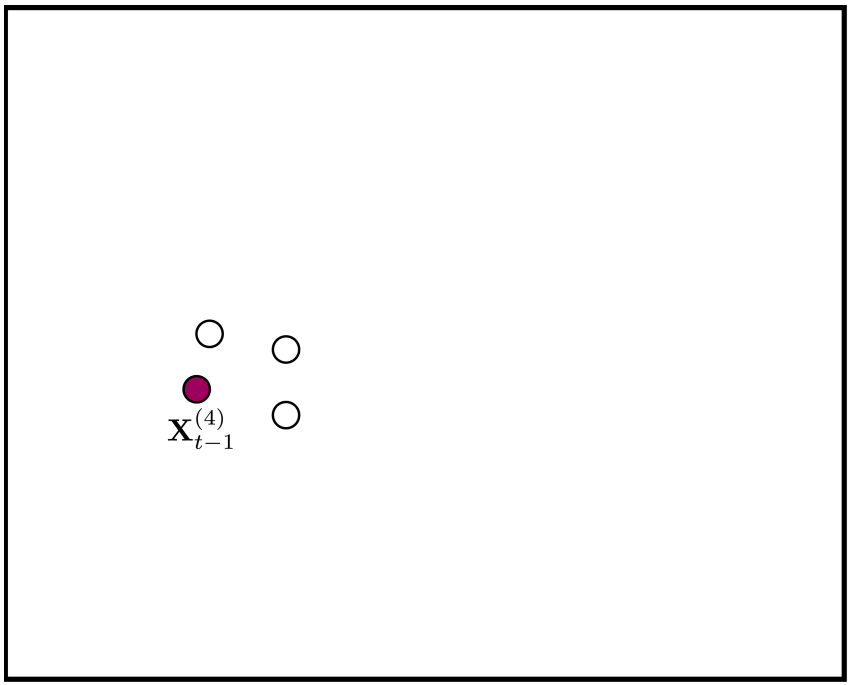
Unscented Transform: Easier to approximate pdf than nonlinear function



Approximating the Predicted pdf

Predicted pdf: $$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \int \overbrace{\mathrm{P}(\mathbf{X}_t|\mathbf{X}_{t-1})}^{\text{Unknown}} \overbrace{\mathrm{P}(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})}^{\text{GMM}} \mathrm{d}\mathbf{X}_{t-1}$$

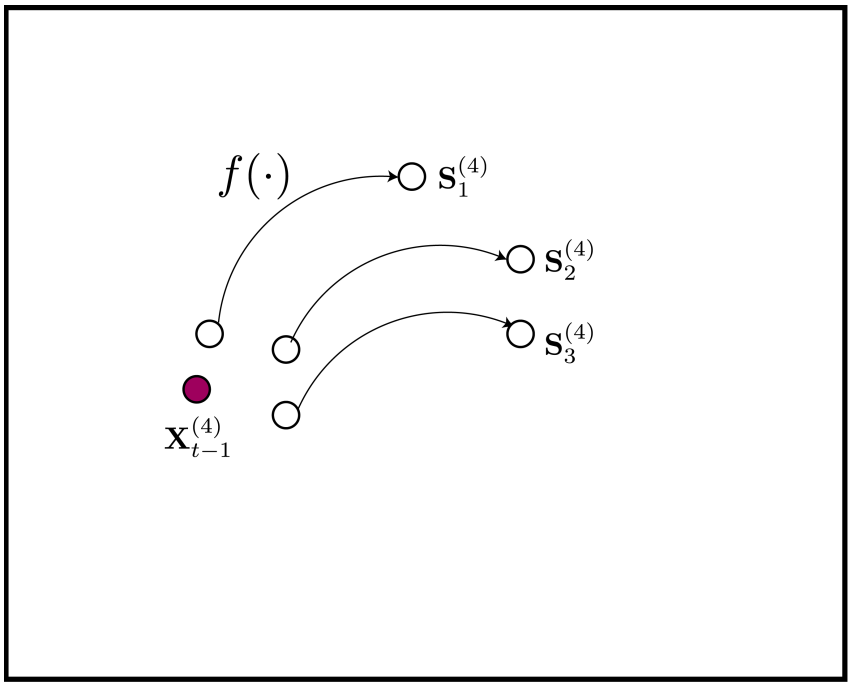
Unscented Transform: Easier to approximate pdf than nonlinear function



Approximating the Predicted pdf

Predicted pdf: $$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \int \overbrace{\mathrm{P}(\mathbf{X}_t|\mathbf{X}_{t-1})}^{\text{Unknown}} \overbrace{\mathrm{P}(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})}^{\text{GMM}} \mathrm{d}\mathbf{X}_{t-1}$$

Unscented Transform: Easier to approximate pdf than nonlinear function



Approximating the Predicted pdf

Predicted pdf: $$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \int \overbrace{\mathrm{P}(\mathbf{X}_t|\mathbf{X}_{t-1})}^{\text{Unknown}} \overbrace{\mathrm{P}(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})}^{\text{GMM}} \mathrm{d}\mathbf{X}_{t-1}$$

Unscented Transform: Easier to approximate pdf than nonlinear function

$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) \approx \sum_{i=1}^{n_{t-1}} \pi_t^i \mathcal{N}(\mathbf{X}_t; \bar{\mathbf{X}}_t^i, \bar{\mathbf{\Sigma}}_t^i)$$

where $\bar{\mathbf{X}}_t^i$, $\bar{\mathbf{\Sigma}}_t^i$ and π_t^i are the deterministic UT samples, covariance matrices, and weights.

Audio-Visual Likelihood

1. Importance sampling from predicted pdf, $\mathbf{X}_i$, $i = 1, \ldots, m$
2. Sample Probability:

$$l_i = \beta_1 \times l_{app}(\mathbf{X}_i) + \beta_2 \times l_a(\mathbf{X}_i) + \beta_3 \times l_v(\mathbf{X}_i)$$

3. Fit Radial Basis Function with Gaussian kernel to predicted GM components

$$\mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t) = \sum_{i=1}^{n_t} w_i \mathcal{N}(\mathbf{X}_t; \mathbf{X}_t^i, \mathbf{P}_t^i).$$

4. **Kernel weights:** Constrained non-negative least square (NNLS) problem:

$$\arg\min_{\mathbf{w}} \|\mathbf{A}\mathbf{w} - \mathbf{b}\|^2 \qquad \text{for } \mathbf{w} \geq 0$$

where $\mathbf{A} \in \mathbb{R}^{m \times m}$: element (i, j) given by kernel probability and $\mathbf{b} \in \mathbb{R}^{m \times 1}$: contains l_i for each row $i = 1, \ldots, m$

Approximating the Posterior pdf

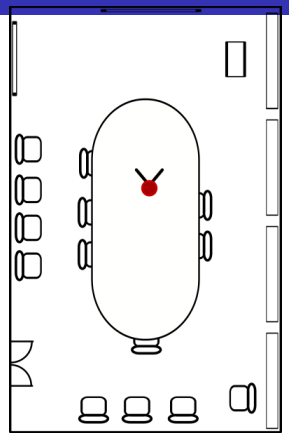
$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t}) \propto \overbrace{\mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t)}^{\text{Gaussian mixture}} \overbrace{\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1})}^{\text{Gaussian mixture}} = \sum_{i=1}^{n_{t-1}} \sum_{j=1}^{n_t} w_t^{ij} \mathcal{N}(\boldsymbol{\mu}_t^{ij}, \boldsymbol{\Sigma}_t^{ij})$$

Mititgate exponential explosion in the number of GM components:

- Cluster using weighted data EM algorithm
- Number of clusters, m_t: Model selection based on Minimum Message Length
- Construct GM from cluster centers:

$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t}) = \sum_{i=1}^{m_t} \pi_t^i \mathcal{N}(\mathbf{X}_t; \mathbf{X}_t^i, \boldsymbol{\Sigma}_t^i)$$

Experimental Validation



- Canon® HD video camera: 1920×1200 pixels resolution and 25 FPS.
- Eigenmike®: 32 channel spherical microphone array.

Scenarios: Two talkers are speaking whilst moving within the room.

1. Scenario **S1**: Consecutively active talkers.
2. Scenario **S2**: Simultaneously active talkers.

Experimental Result: Video

- Matlab code and additional video can be found online:
  `www.team.inria.fr/perception/research/avtracking_by_dabf/`

Experimental Results

Table: Tracking results performance comparison. ↑ denotes higher scores indicate better results, and ↓ denotes lower scores indicate better results.

Sequence	Methods	MOTA (in %)↑	OMAT↓	OSPA↓
S1	GMM AV	83.6	186.6	112.7
	GMM Visual	72.3	245.6	234.0
	VB Visual	45.7	378.6	367.8
S2	GMM AV	89.8	201.4	198.0
	GMM Visual	86.3	200.4	193.6
	VB Visual	44.3	356.7	367.8

MOTA: Multiple Object Tracking Accuracy
OMAT: Optimal MAss Transfer
OSPA: Optimal Sub-Pattern Assignment

Conclusions

- Proposed audio-visual tracking model based on sequential Bayesian Filtering framework
- Fuse detections from audio and visual modalities
- GMM representations are used to approximate relevant density functions
- Results using recordings show improvements compared to visual only results
- Future work will focus on handover for long periods of missing detections in either modality