RMIT UNIVERSITY

THE UNIVERSITY OF WESTERN AUSTRALIA

THE UNIVERSITY OF ADELAIDE AUSTRALIA

# Bayesian Integration of Audio and Visual Information for Multi-Target Tracking Using a CB-MeMBer Filter

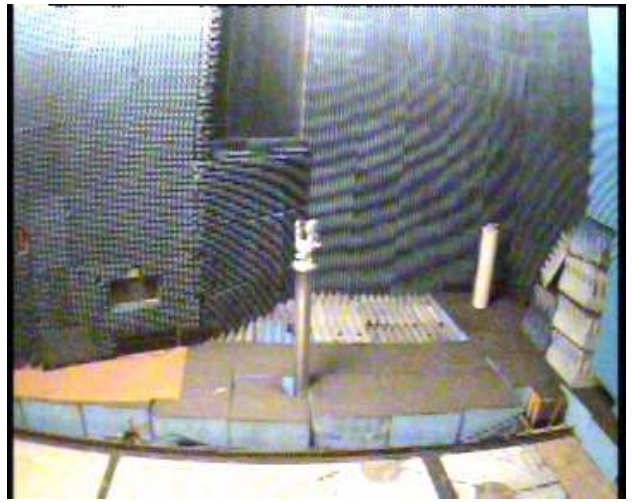
Reza Hoseinnezhad
Ba-Ngu Vo
Ba-Tuong Vo
David Suter

# Contents

- Background
- Random Finite Set Approach to Multi-Object Estimation
- Cardinality Balanced MeMBer Filter
- Audio Visual Tracking
- Simulation Results
- Conclusions

# Background

- Problem: Audio-visual multi-target tracking
- Applications:
  - **monitoring people behaviour**
  - **traffic monitoring**
  - **smart rooms**



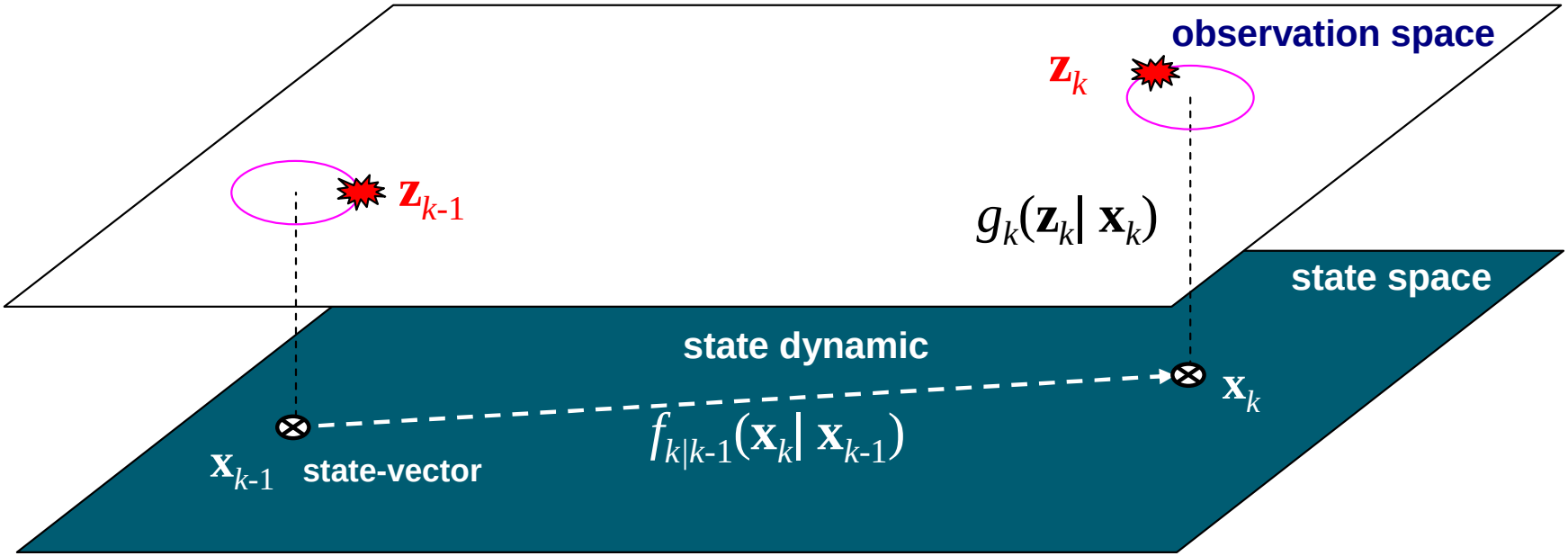
- Main challenges:
  - **Occasionally silent targets**
  - **Occasionally invisible targets (out of visual field of view)**
  - **Clutter measurements**

# Background (Continued)

- **Our contribution:**

  a principled approach to combine audio and video data in a Bayesian framework

# The Bayes (nonlinear) Filter




**Bayes filter**

$$\cdots \longrightarrow p_{k-1}(\bullet\,|\mathbf{z}_{1:k-1}) \xrightarrow{\text{prediction}} p_{k|k-1}(\bullet\,|\,\mathbf{z}_{1:k-1}) \xrightarrow{\text{data-update}} p_k(\bullet\,|\,\mathbf{z}_{1:k}) \longrightarrow \cdots$$

**Kalman filter**

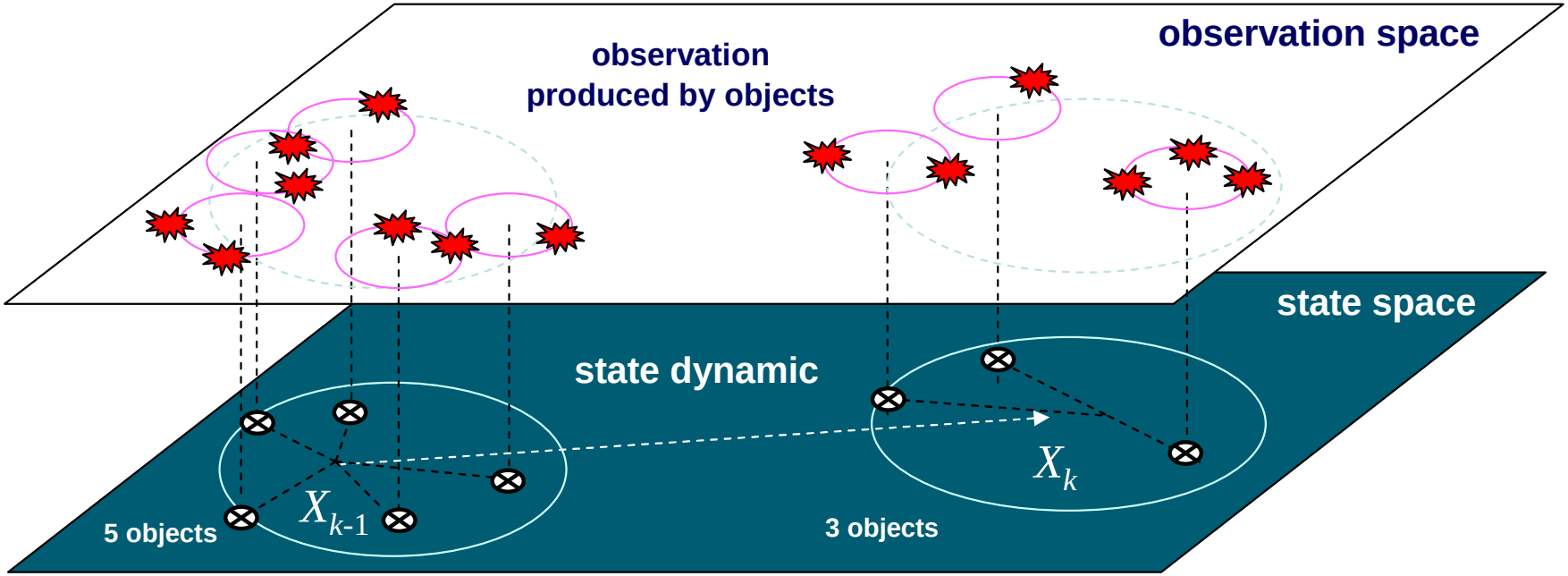
$$\cdots \longrightarrow \mathcal{N}(\bullet;\mathbf{m}_{k-1}, P_{k-1}) \longrightarrow \mathcal{N}(\bullet;\mathbf{m}_{k|k-1}, P_{k|k-1}) \longrightarrow \mathcal{N}(\bullet;\mathbf{m}_k, P_k) \longrightarrow \cdots$$

**Particle filter**

$$\cdots \longrightarrow \{w_{k-1}^{(i)}, \mathbf{x}_{k-1}^{(i)}\}_{i=1}^{N} \longrightarrow \{w_{k|k-1}^{(i)}, \mathbf{x}_{k|k-1}^{(i)}\}_{i=1}^{N} \longrightarrow \{w_k^{(i)}, \mathbf{x}_k^{(i)}\}_{i=1}^{N} \longrightarrow \cdots$$

# Multi-Object Filtering

- Objective: **Jointly estimate** the **number** & **states** of objects
- Challenges:
  - Random number of objects and measurements
  - Detection uncertainty, clutter, association uncertainty

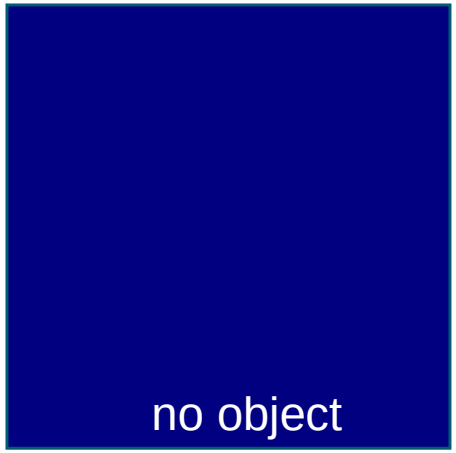


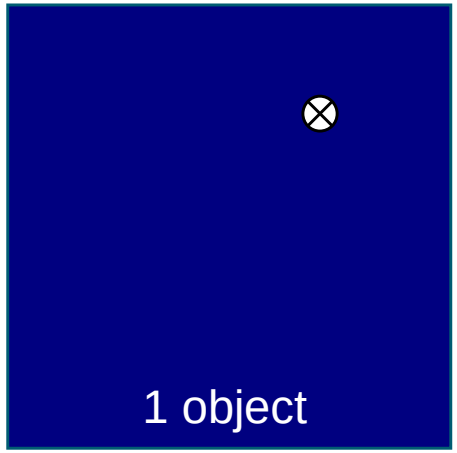

**Multi-object state space model**

# Multi-Bernoulli RFS

**Bernoulli RFS** with existence probability $r$ and distribution $p$,

```
sample u ~ uniform[0,1]
if u < r,
   sample x ~ p,
end;
```




**Multi-Bernoulli RFS**: union of $M$ independent Bernoulli RFSs

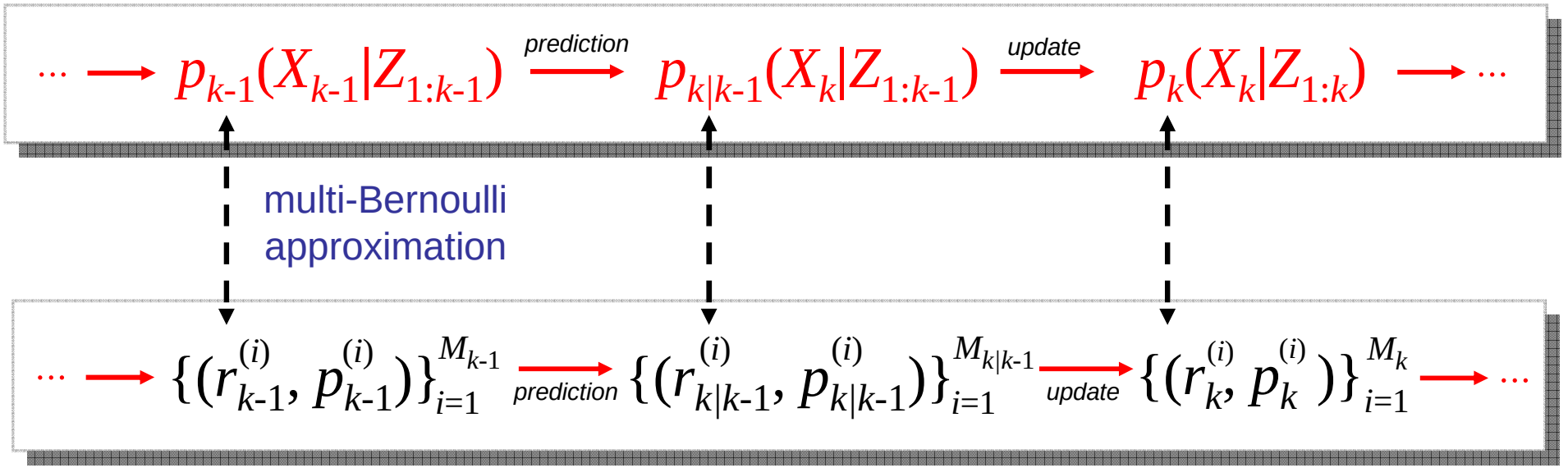
Completely characterised by the set of parameter pairs

$$\{(r^{(i)}, p^{(i)})\}_{i=1}^{M}$$

existence probability of object $i$ pdf of state of object $i$

# Multi-Bernoulli Filter

**Multi-object Bayes filter**




**Cardinality-Balanced MeMBer filter** [VVC 07, 09]

- Approximate predicted/posterior RFSs by Multi-Bernoulli RFSs
- Valid for low clutter rate & high probability of detection
- More useful than PHD filters in particle implementations.

# Multi-Bernoulli Filter

$$\cdots \rightarrow \{(r_{k-1}^{(i)}, p_{k-1}^{(i)})\}_{i=1}^{M_{k-1}} \xrightarrow{\text{prediction}} \{(r_{k|k-1}^{(i)}, p_{k|k-1}^{(i)})\}_{i=1}^{M_{k|k-1}} \xrightarrow{\text{update}} \{(r_{k}^{(i)}, p_{k}^{(i)})\}_{i=1}^{M_{k}} \rightarrow \cdots$$

$$\{(r_{P,k|k-1}^{(i)}, p_{P,k|k-1}^{(i)})\}_{i=1}^{M_{k-1}} \cup \{(r_{\Gamma,k}^{(i)}, p_{\Gamma,k}^{(i)})\}_{i=1}^{M_{\Gamma,k}}$$

$$r_{k-1}^{(i)} \langle p_{k-1}^{(i)}, p_{S,k} \rangle$$

term for object births

$$\frac{\langle f_{k|k-1}(\bullet|\bullet), p_{k-1}^{(i)} p_{S,k} \rangle}{\langle p_{k-1}^{(i)}, p_{S,k} \rangle}$$

CB-MeMBer filter
[VVC 09]

# Multi-Bernoulli Filter

$$\cdots \rightarrow \{(r_{k-1}^{(i)}, p_{k-1}^{(i)})\}_{i=1}^{M_{k-1}} \xrightarrow{\text{prediction}} \{(r_{k|k-1}^{(i)}, p_{k|k-1}^{(i)})\}_{i=1}^{M_{k|k-1}} \xrightarrow{\text{update}} \{(r_{k}^{(i)}, p_{k}^{(i)})\}_{i=1}^{M_{k}} \rightarrow \cdots$$

$$\{(r_{L,k}^{(i)}, p_{L,k}^{(i)})\}_{i=1}^{M_{k|k-1}} \cup \{(r_{U,k}(z), p_{U,k}(z))\}_{z \in Z_k}$$

$$\frac{r_{k|k-1}^{(i)}(1 - \langle p_{k|k-1}^{(i)}, p_{D,k} \rangle)}{1 - r_{k|k-1}^{(i)} \langle p_{k|k-1}^{(i)}, p_{D,k} \rangle}$$

$$\frac{p_{k|k-1}^{(i)}(1 - p_{D,k})}{1 - \langle p_{k|k-1}^{(i)}, p_{D,k} \rangle}$$

$$\frac{p_{D,k} g_k(z|\bullet) \sum_{i=1}^{M_{k|k-1}} \frac{r_{k|k-1}^{(i)} p_{k|k-1}^{(i)}}{1 - r_{k|k-1}^{(i)}}}{\sum_{i=1}^{M_{k|k-1}} \frac{r_{k|k-1}^{(i)} \langle p_{k|k-1}^{(i)}, p_{D,k} g_k(z|\bullet) \rangle}{1 - r_{k|k-1}^{(i)}}}$$

$$\frac{\sum_{i=1}^{M_{k|k-1}} \frac{r_{k|k-1}^{(i)}(1 - r_{k|k-1}^{(i)}) \langle p_{k|k-1}^{(i)}, p_{D,k} g_k(z|\bullet) \rangle}{(1 - r_{k|k-1}^{(i)} \langle p_{k|k-1}^{(i)}, p_{D,k} \rangle)^2}}{\kappa(z) + \sum_{i=1}^{M_{k|k-1}} \frac{r_{k|k-1}^{(i)} \langle p_{k|k-1}^{(i)}, p_{D,k} g_k(z|\bullet) \rangle}{1 - r_{k|k-1}^{(i)} \langle p_{k|k-1}^{(i)}, p_{D,k} \rangle}}$$

CB-MeMBer filter
[VVC 09]

# Audio Visual Tracking

- Our implementation:

$$\mathbf{x} = [x_{\text{im.}}\ y_{\text{im.}}\ \dot{x}_{\text{im.}}\ \dot{y}_{\text{im.}}\ \ w_{\text{im.}}\ h_{\text{im.}}]^\top$$

- Video measurements:
  - kernel-based background subtraction + morphological image operations → a set of rectangular blobs in each frame
  - The result: $Z_v = \{z_{v_i}\}$

$$z_v = [x_{\text{im.}}\ y_{\text{im.}}\ w_{\text{im.}}\ h_{\text{im.}}]^\top$$

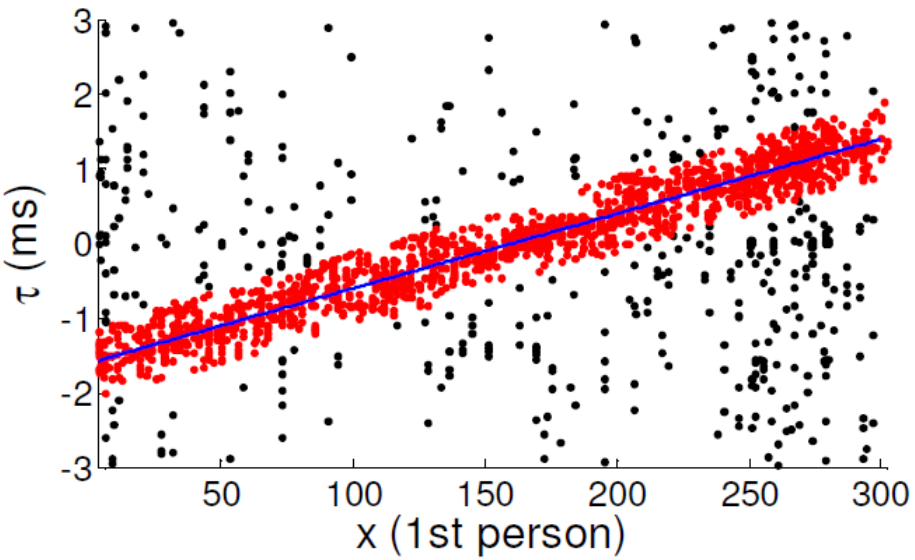
  - Likelihood: $g_v(z_v|x) = \mathcal{N}(z_v; C_v x, \sigma_v^2)$

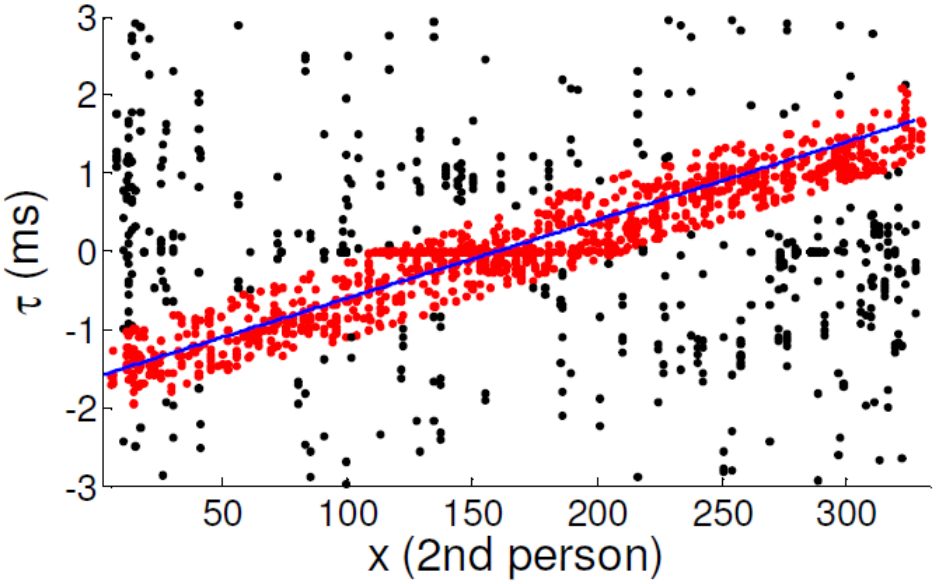
$$\mathcal{N}(z; \mu, \sigma^2) \triangleq \exp(-(z-\mu)^2/(2\sigma^2))/(\sqrt{2\pi}\sigma)$$

$$C_v = \text{diag}(1, 1, 0, 0, 1, 1)$$

- Audio measurements:
  - Audio signals from the microphones on two sides of the camera
  - Signal Processing:
    - time difference of arrival (TDOA)
    - cross-correlation between the signals using the Generalised Cross Correlation function - Phase Transform (GCC-PHAT)
    - reverberation effects → several peaks in the GCC-PHAT curve plotted versus time difference
  - at most five largest peaks of the GCC-PHAT values are picked and considered as TDOA measurements in each frame.

- Likelihood function for audio measurements:
  - Relatively large distance of targets from the microphones compared to the distance between the two microphones
  - Approximately linear relationship between the $x_{im}$ and corresponding TDOA

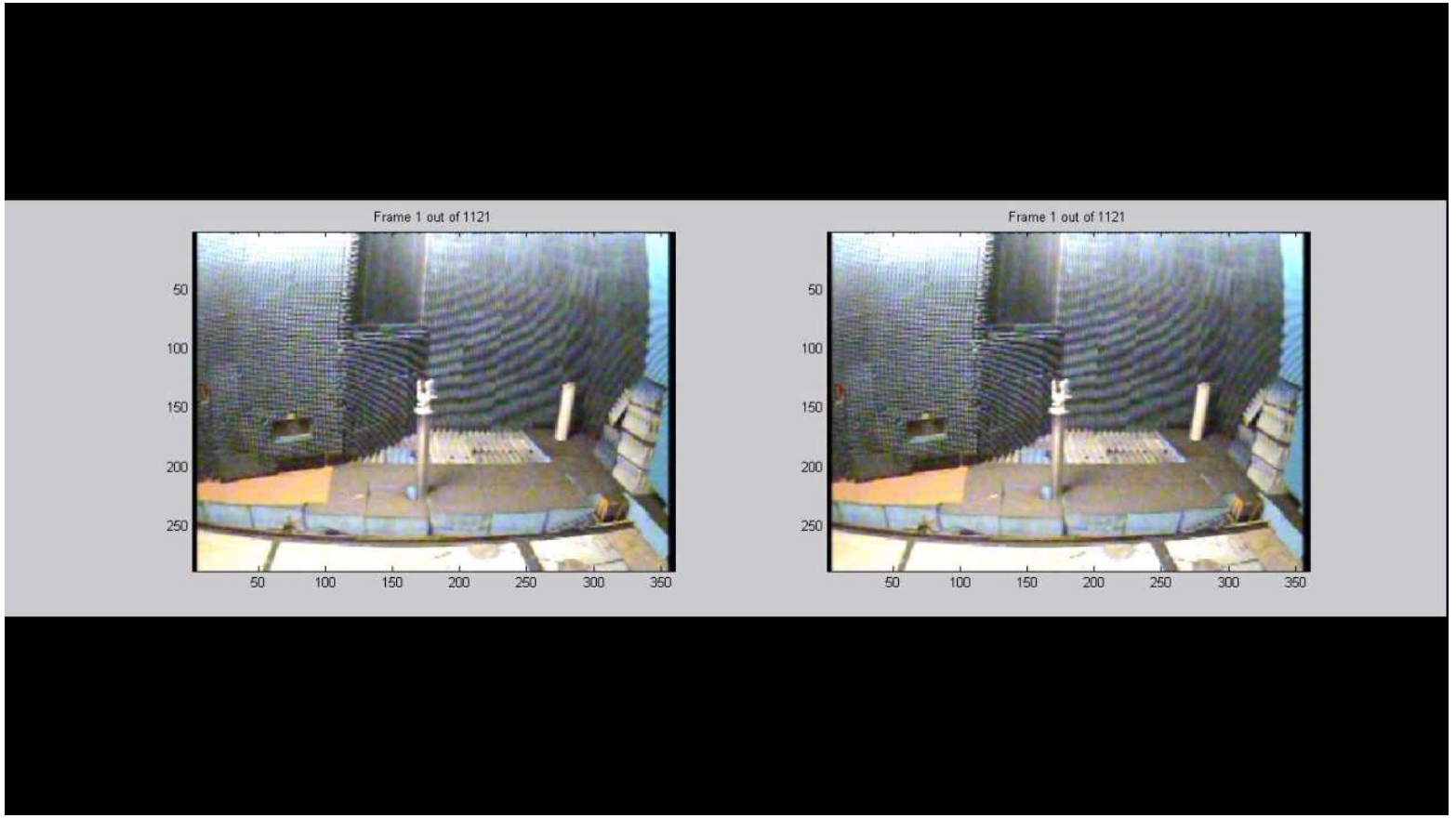


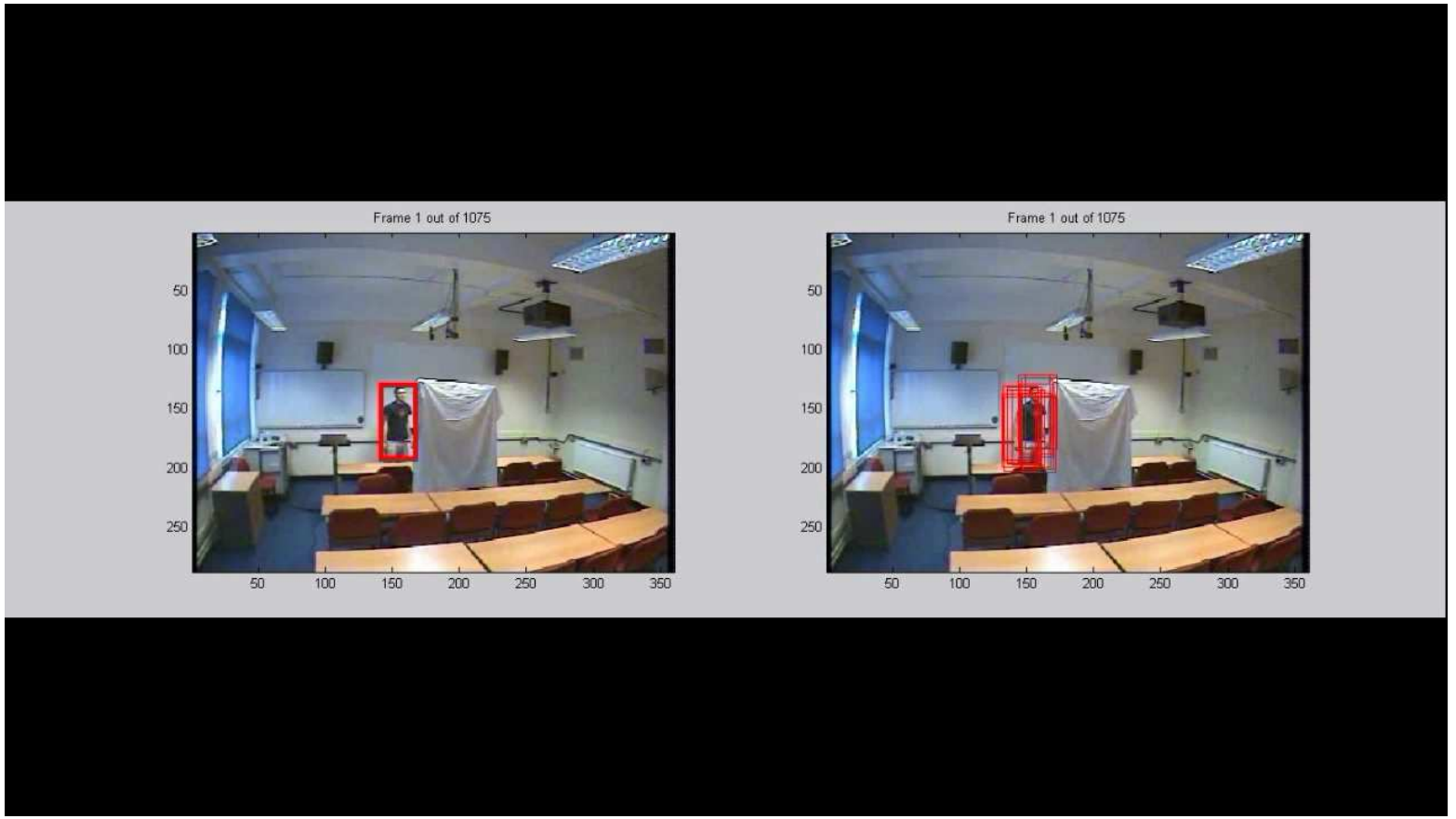
# Sensor Fusion

- Perform CB-MeMBer update twice:
  - **first using the visual measurements**
  - **then audio measurements.**
- Detection probability for each sensor is determined based on our definition of “active speaker”.
  - Example: if an active speaker is considered to be a person who is expected to be visible to the camera in no less than 95% of the time and to be speaking in at least 40% of the time, then we set $p_{Dv} = 0.95$ and $p_{Da} = 0.40$.

# Simulation Results




## SPEVI Database Sequence No. 1




# SPEVI Database Sequence No. 2


# SPEVI Database Sequence No. 3

# Quantitative Results

- In 98.5% of all the frames, the existing targets are all detected, correctly labelled and tracked.
- Labels are never switched after or during occlusions.
- An invisible target is successfully tracked using the audio cues.

| | Without Audio | | | With Audio | | |
|---|---|---|---|---|---|---|
| | FNR | FAR | LSR | FNR | FAR | LSR |
| Seq. 1 | 9% | 2% | 4% | 3% | 0% | 0% |
| Seq. 2 | 32% | 3% | n/a | 5% | 0% | n/a |
| Seq. 3 | 11% | 2% | 3% | 2% | 0% | 0% |

# Conclusions

- A new method for audio-visual tracking of multiple targets was proposed.
- The method is formulated in a random finite set framework based on multi-Bernoulli approximations, and implemented using sequential Monte Carlo techniques.
- Audio and visual cues are integrated by multiple updates.
- The random finite set formulation allows a natural and principled way to model the intermittent nature of sensory data (mainly audio).

- Simulation results show that the proposed method almost perfectly tracks multiple interacting targets, not only when they are silent, but also in times when they are invisible to the camera.


Thank You