# Speaker Localization and Tracking using Multi-modal Signals


Xinyuan Qian


04/09/2019


Centre for Intelligent Sensing
Queen Mary University of London

# Introduction

- Objective
  - **Multiple Objects Tracking (MOT)** in **3D** using a small-size co-located **audio-visual** sensing platform
- Motivations
  - Real-world applications *e.g. surveillance, driver assistance*
  - Complementary advantages of multi-modalities
    - Deal with the rapid changing environment
    - An improved tracking performance

Emotech Olly

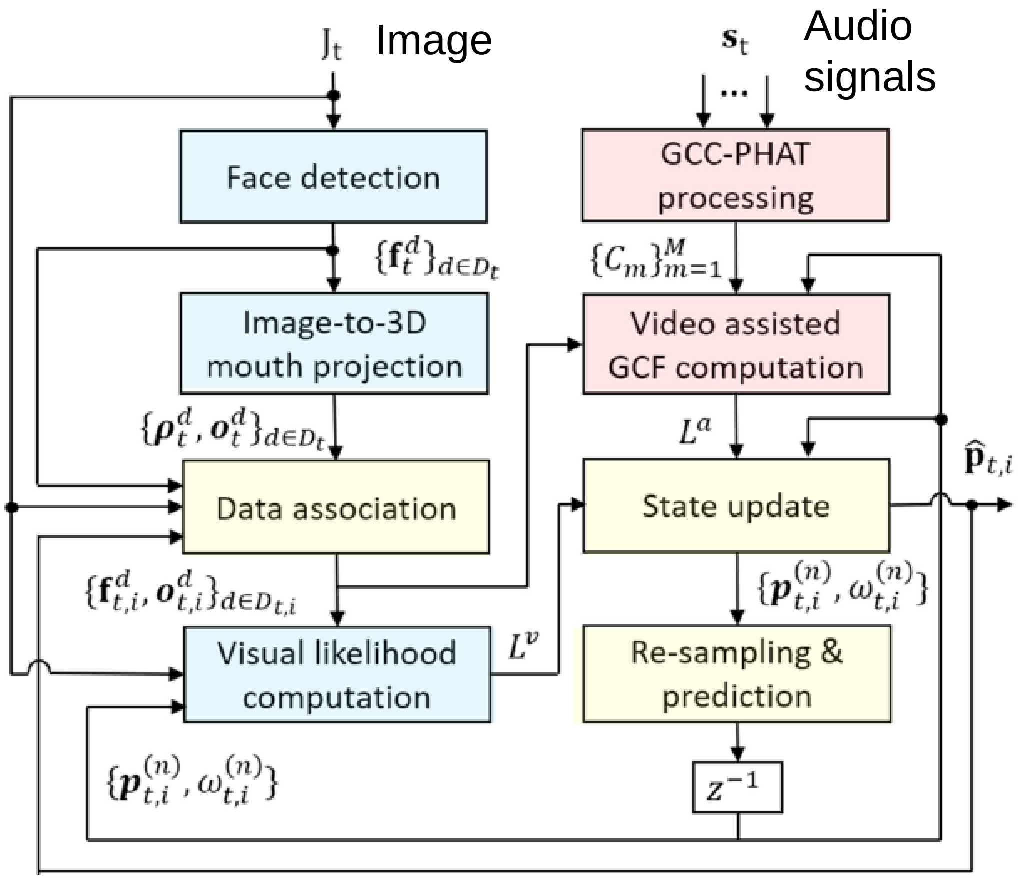
# Challenges

- Traditional problems *e.g. reverberation, background noise, occlusion and body orientation*
- Depth estimate in neither audio nor video *i.e. co-located setup*
- Multi-modality fusion *e.g. what, when and how to fuse*

# State-of-the-Art (SoA) summary

- Increasing popularity for audio-visual MOT in 3D
- SoA Tracking approaches
  - Kalman filter, particle filter framework etc.
  - Time-delay, steered response power etc.
  - Colour, detection, motion etc.
- Limitations
  - MOT in 3D with distributed sensor networks
  - MOT on image with small-size sensing platform
  - Lack of public datasets

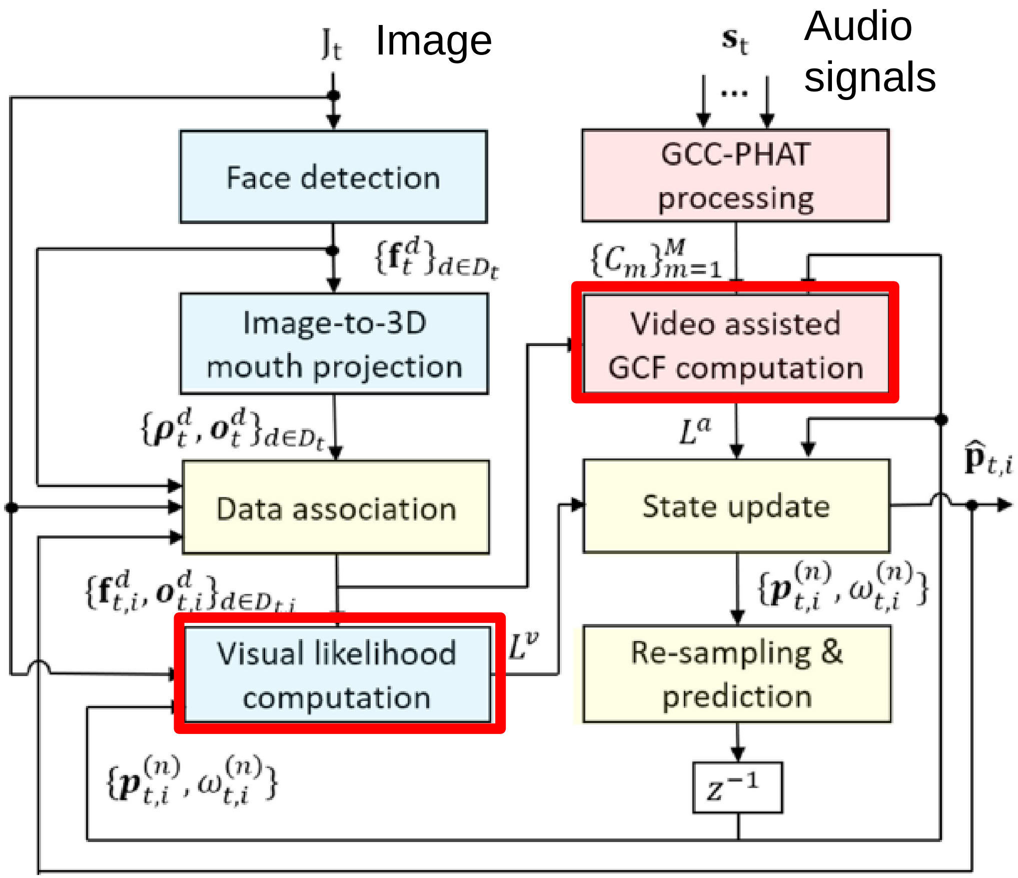
# Proposed method




General block diagram

$\boldsymbol{f}_t^d$: $d^{th}$ face detection, $d \in D_t$
$\boldsymbol{o}_t^d$: 3D mouth location estimate
$\boldsymbol{C}_m$: GCC-PHAT at $m^{th}$ mic pair
$\boldsymbol{P}_t$: particle set
$\boldsymbol{L}^a$: audio likelihood
$\boldsymbol{L}^v$: video likelihood
$\boldsymbol{g}$: Global Coherence Field (GCF)
$\hat{\boldsymbol{p}}_{t,i}$: 3D estimate of target $i$

# Proposed method




General block diagram

$\boldsymbol{f}_t^d$: $d^{th}$ face detection, $d \in D_t$

$\boldsymbol{o}_t^d$: 3D mouth location estimate

$\boldsymbol{C}_m$: GCC-PHAT at $m^{th}$ mic pair

$\boldsymbol{P}_t$: particle set

$\boldsymbol{L}^a$: audio likelihood

$\boldsymbol{L}^v$: video likelihood

$\boldsymbol{g}$: Global Coherence Field (GCF)

$\hat{\boldsymbol{p}}_{t,i}$: 3D estimate of target $i$

**Novelties**

# Video likelihood

- Discriminative *or* Generative likelihood

$$L^v_{\text{det}}(J_t \mid \mathbf{p}) = \sum_{d\in D_{t,i}} \exp\left[-\left(\tilde{\mathbf{o}}^d_{t,i} - \tilde{\mathbf{p}}\right)\Sigma_v^{-1}\left(\tilde{\mathbf{o}}^d_{t,i} - \tilde{\mathbf{p}}\right)^\intercal\right]$$

Face detection set

3D mouth estimate in camera's spherical coordinates

Diagonal covariance matrix

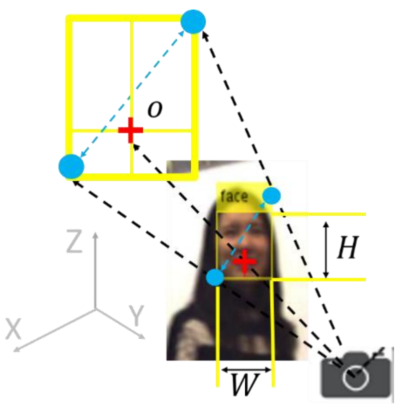
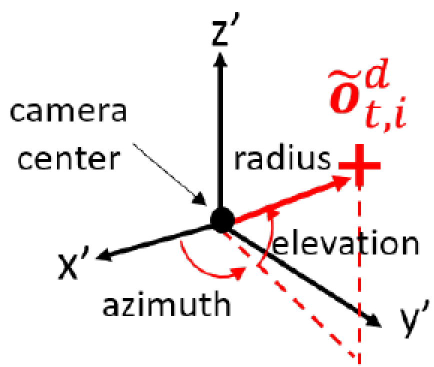
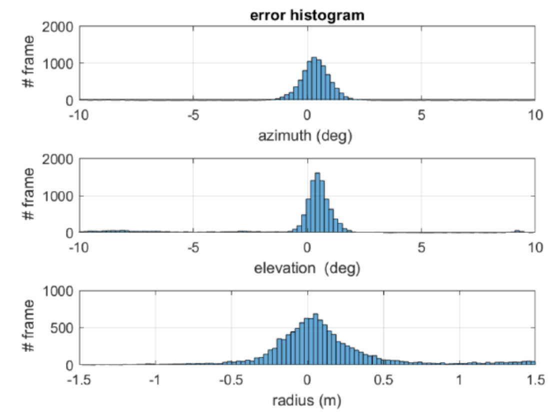



Image-3D projection




Camera's spherical coordinates



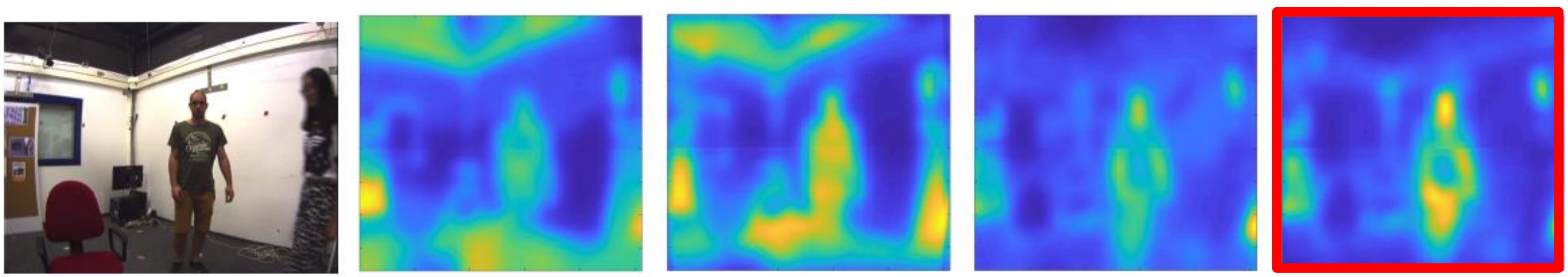

Error in individual coordinates

# Video likelihood

- Discriminative *or* Generative likelihood

$$L^{v}_{\mathrm{HSV}}(J_t \mid \mathrm{p}) = \sum_{b=1}^{B} \sqrt{r^b_{\mathrm{v}} r^b_{\mathrm{f}}} \left[ 8\pi |\Sigma^b_{\mathrm{v}} \Sigma^b_{\mathrm{f}}|^{\frac{1}{4}} \mathcal{N}(\mu^b_{\mathrm{v}} | \mu^b_{\mathrm{f}}, 2(\Sigma^b_{\mathrm{v}} + \Sigma^b_{\mathrm{f}})) \right]$$

Similarity measure between face reference image $f$ and particle's 3D-image projection $v$



(a) test image (b) histogram (RGB) (c) spatiogram (RGB) (d) histogram (HSV) (e) spatiogram (HSV)

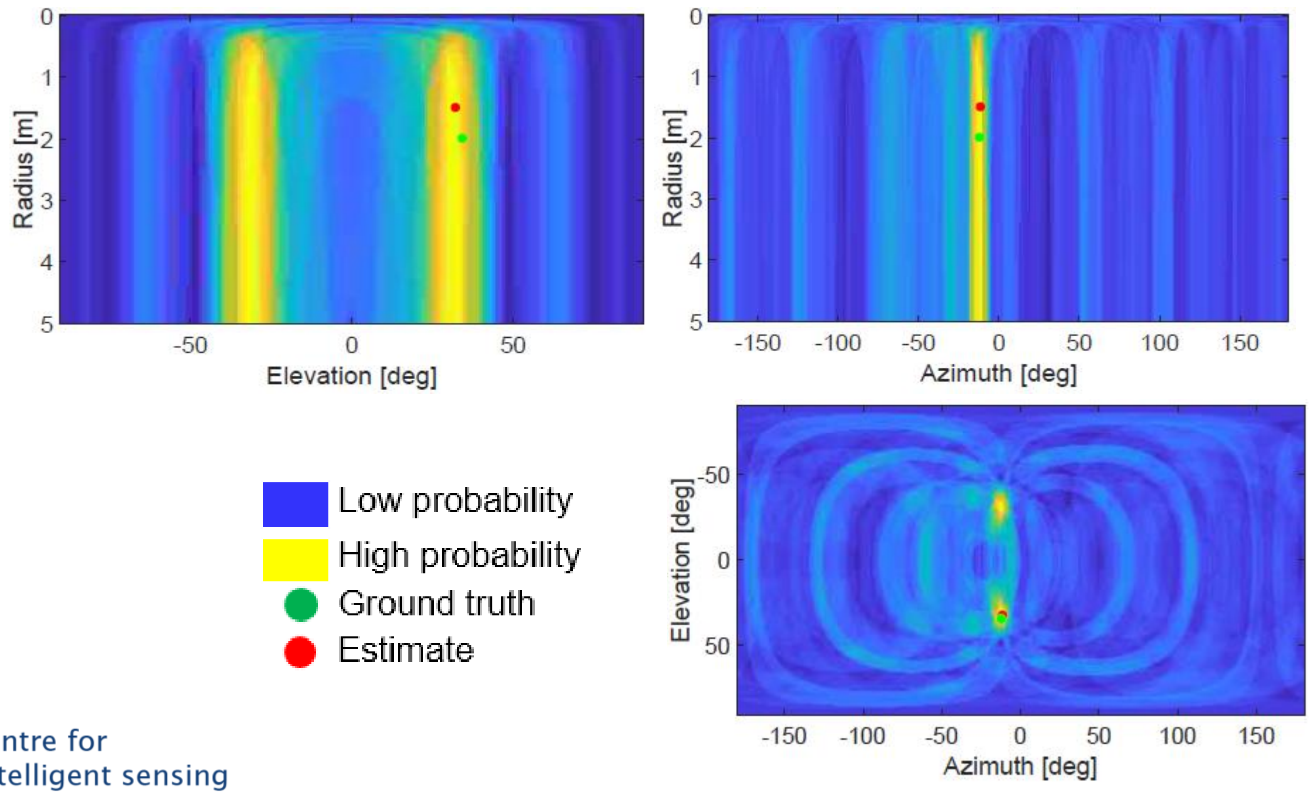
Low probability High probability

# Audio likelihood

- Video-assisted acoustic map

$$g_v(\mathbf{p}, t) = \frac{1}{M} \sum_{m=1}^{M} C_m \left( \tau_m(\mathbf{p} | o_{t',i}^{d,z}), t \right)$$

Speaker height suggested from video 3D estimate



# Experiments

- AV16.3 dataset (public)
  - 8-element circular microphone array (20 cm diameter)
  - Standard RGB camera
  - Ground truth
    - sensor calibration information
    - target 3D location
- CAV3D dataset (self-collected)
  - All above
  - Co-located audio-visual sensing platform

# CAV3D dataset

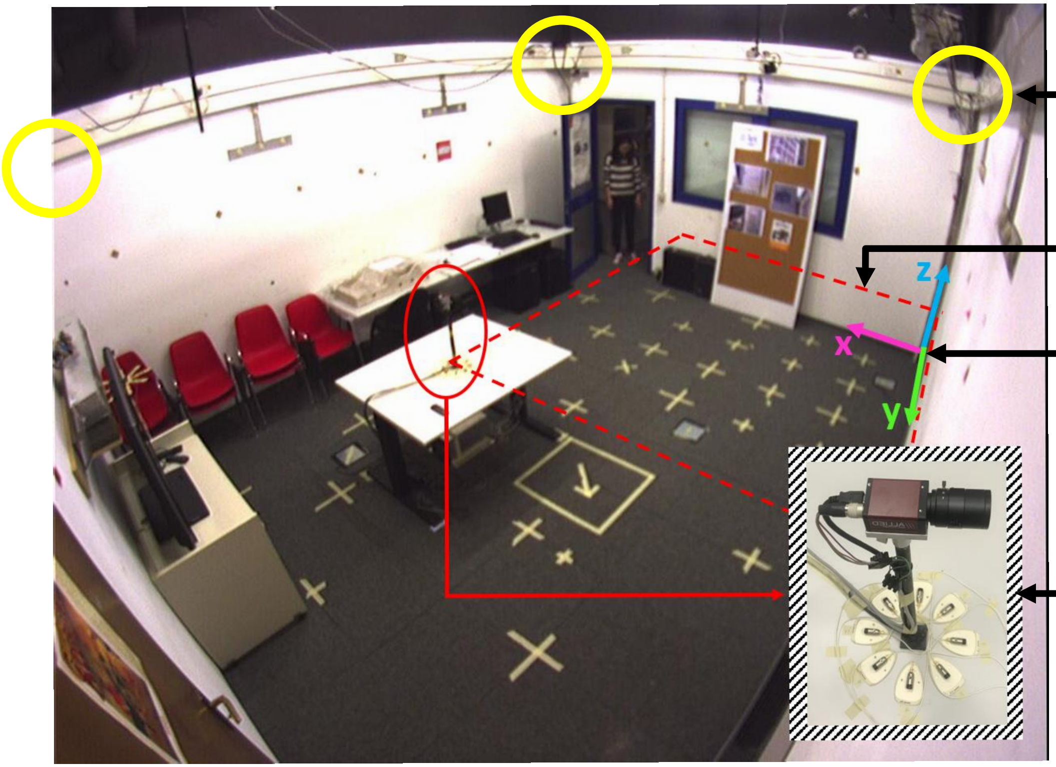
A novel audio-visual dataset for MOT in 3D!

- Calibrated sensors
- Synchronized audio-visual signals
- Ground Truth (GT)
  - Image location, 3D location, voice activities

- Three sub-sets (20), durations from $15s$ to $80s$ :
  - CAV3D-SOT (9), 1 speaker
  - CAV3D-SOT2 (6), 2 targets take turns to talk
  - CAV3D-MOT (5), 3 concurrent speakers

# CAV3D dataset

A novel audio-visual dataset for MOT in 3D!

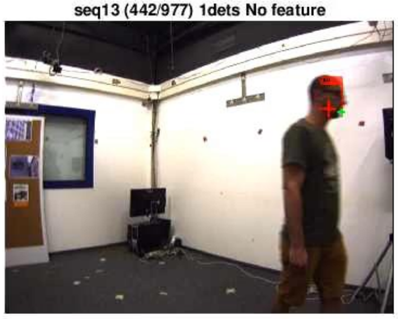



- Corner cameras
- Camera's Field of View
- Origin of the 3D world coordinates
- Audio-visual sensing platform

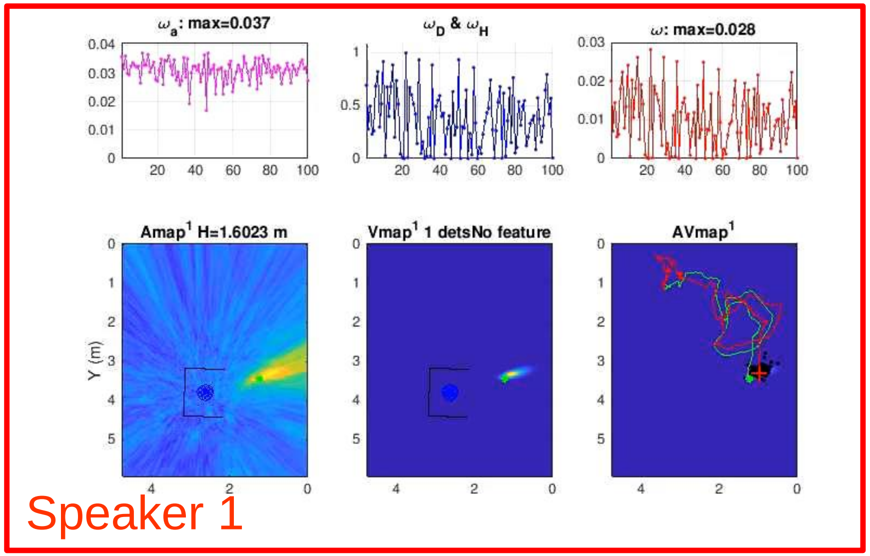
Recording environment (view of ◯ cam #1)

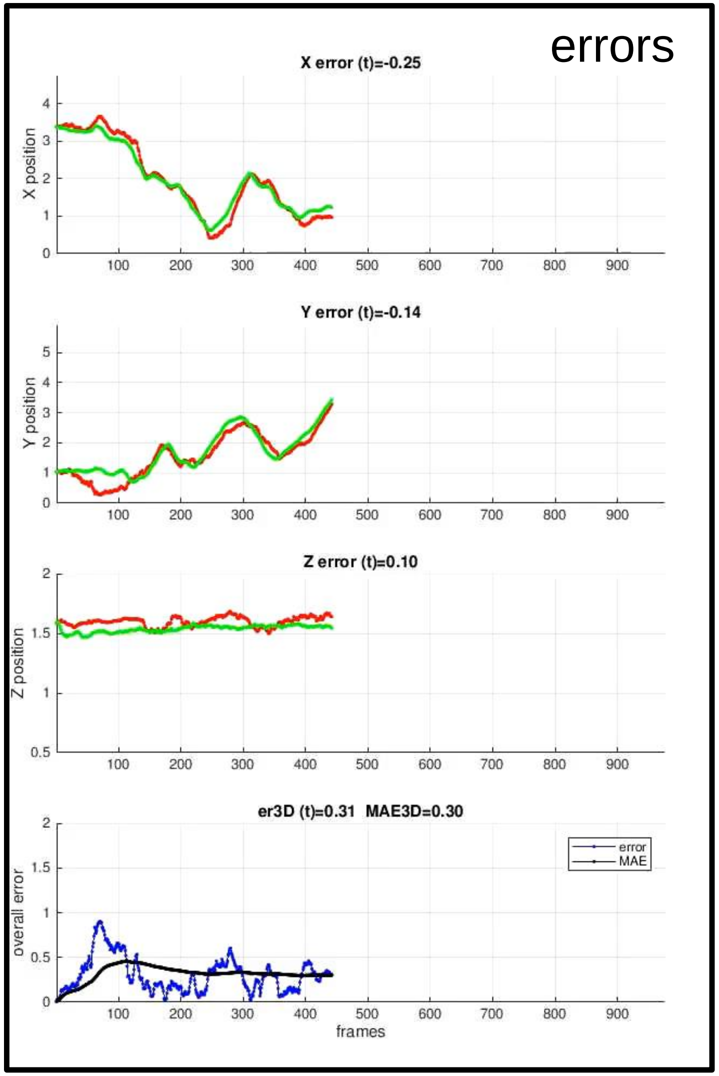
# Tracking result – demo1

**EX1**. seq13 (SOT)



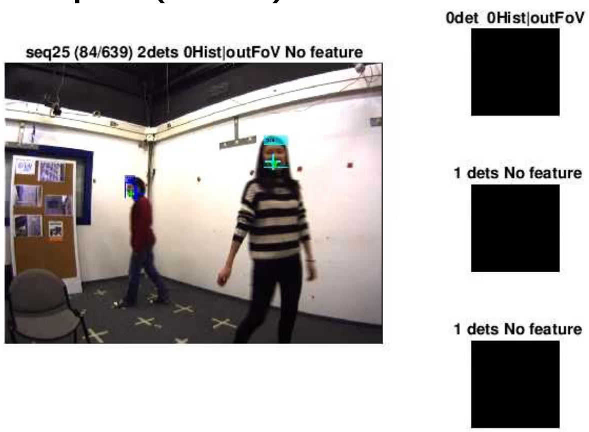


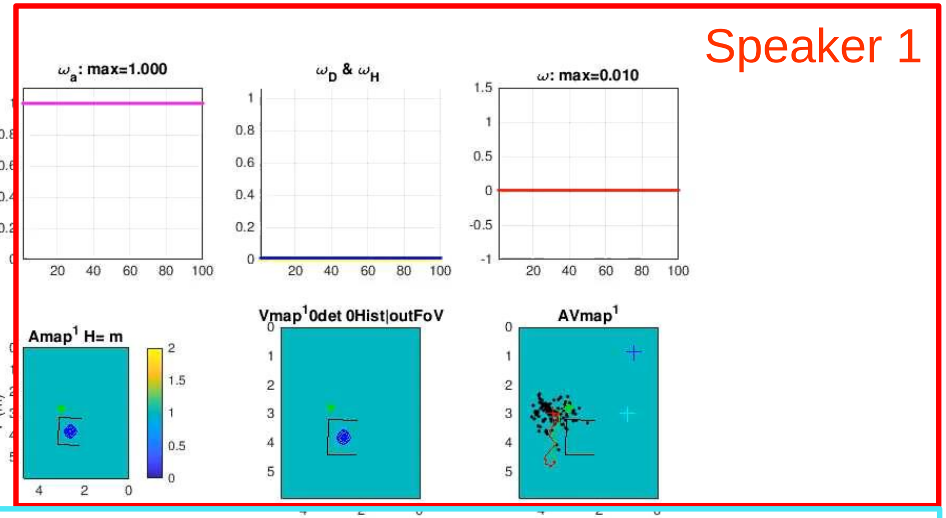


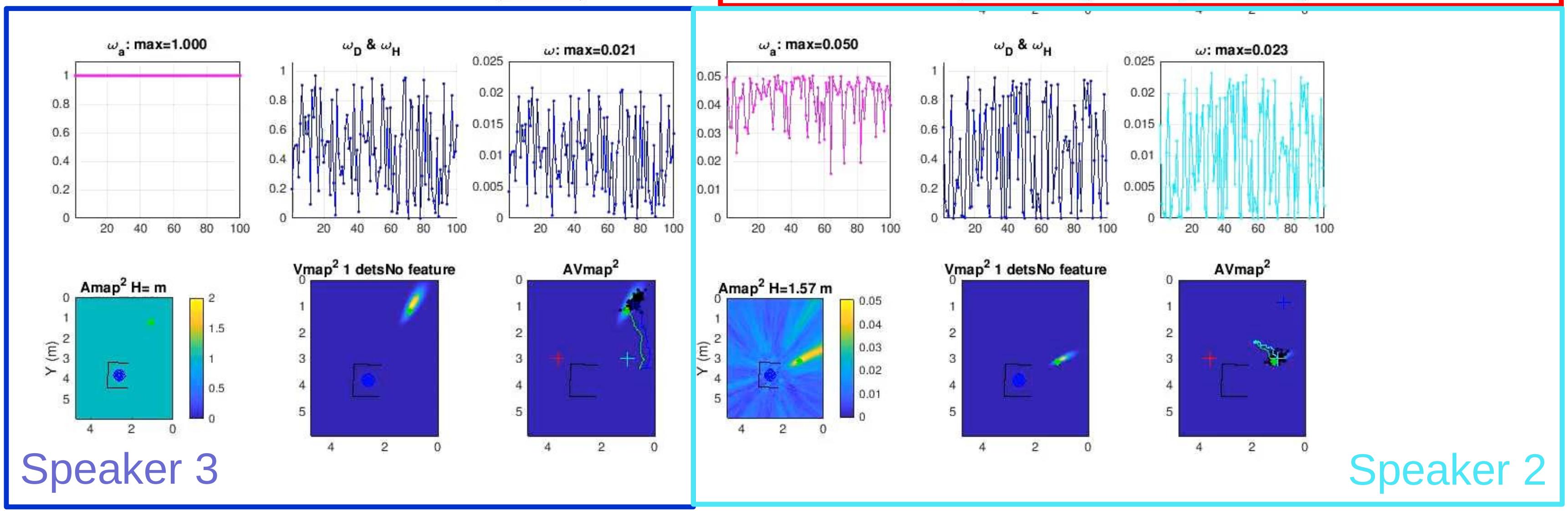



# Tracking result – demo2

**EX2**. seq25 (MOT)




CIS centre for intelligent sensing

Queen Mary University of London

# Results

| | | Image plane | | | | | 3-D | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Kilic2015 | Qian2018 | AO (2-D) | VO | AV3T | Zotkin2002 | Qian2018 | AO (2-D) | VO | AV3T |
| SOT | TLR | 29.5±12.4 | 25.0±1.2 | 52.2 ± 4.7 | 38.4 ± 17.5 | **7.0 ± 3.6** | 84.8±5.4 | 68.7±2.9 | 56.5±4.4 | 47.3±13.5 | **31.8 ± 3.5** |
| | $\varepsilon$ | 60.0±34.1 | 38.2±2.3 | 60.3 ± 6.9 | 80.2±103.0 | **16.5 ± 8.6** | .84 ± | .50 ± .02 | .52 ± .08 | .76 ± .34 | **.30 ± .05** |
| | $\varepsilon'$ | 24.5±30.5 | 15.5 ± .4 | 27.7 ± 1.2 | 12.7 ± 1.1 | **12.2 ± .3** | .17 ± .02 | .20 ± .01 | .17 ± .01 | .16 ± .01 | **.16 ± .01** |
| SOT2 | TLR | 33.0±18.5 | 23.0 ± .9 | 38.3 ± 3.9 | 13.4 ± 7.6 | **4.0 ± 1.6** | 85.2±4.5 | 62.9±2.8 | 43.6±4.9 | 20.1 ± 7.1 | **11.1 ± 3.1** |
| | $\varepsilon$ | 81.7±73.5 | 53.4±2.6 | 48.0 ± 6.0 | 36.5 ± 27.2 | **20.8 ± 5.4** | .75 ± .07 | .47 ± .02 | .37 ± .07 | .31 ± .12 | **.18 ± .02** |
| | $\varepsilon'$ | 23.7±64.5 | 13.3 ± .3 | 25.0 ± .6 | 12.0 ± .2 | **11.7 ± .2** | .17 ± .02 | .20 ± .01 | .15 ± .01 | .14 ± .01 | **.14 ± .00** |
| MOT | TLR | 16.0±10.0 | - | 59.4 ± 11.5 | 37.1 ± 7.1 | **11.2 ± 5.9** | 77.7±8.1 | - | 70.2±9.0 | 56.6 ± 6.2 | **35.7 ± 6.6** |
| | $\varepsilon$ | 59.3±33.9 | - | 155.7±60.6 | 127.9±60.1 | **24.8±23.7** | .92 ± .23 | - | 1.03±.27 | 1.05 ± .22 | **.43 ± .12** |
| | $\varepsilon'$ | 17.6±27.4 | - | 19.9 ± 2.1 | 12.2 ± 1.3 | **10.1 ± .6** | .16 ± .02 | - | .16 ± .02 | **.14 ± .02** | .15 ± .01 |

Tracking results comparison on CAV3D dataset

| | | Image plane | | | | | 3-D | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Kilic2015 | Qian2018 | AO (2-D) | VO | AV3T | Zotkin2002 | Qian2018 | AO (2-D) | VO | AV3T |
| SOT | TLR | - | 48.2±3.8 | 48.1±6.0 | 9.0 ± 1.9 | **8.5 ± 2.6** | **10.4 ± 3.4** | 29.2 ± 3.7 | 34.9 ± 8.9 | 52.7 ± 5.5 | 13.3 ± 4.3 |
| | $\varepsilon$ | 11.8±.2 | 19.9±1.6 | 24.1±5.7 | 8.2 ± 1.1 | **7.7 ± 1.3** | **.15 ± .01** | .25 ± .02 | .28 ± .07 | .41 ± .05 | .16 ± .02 |
| | $\varepsilon'$ | - | 8.5 ± .3 | 7.6 ± .5 | **5.3 ± .1** | **5.3 ± .1** | .12 ± .01 | .14 ± .01 | .15 ± .01 | .16 ± .01 | **.11 ± .01** |
| MOT | TLR | - | - | 56.6±9.4 | 15.5±9.0 | **9.2 ± 6.0** | 37.7 ± 5.6 | - | 44.9 ± 1.2 | 56.3 ± 9.8 | **15.8 ± 8.9** |
| | $\varepsilon$ | 11.2±.1 | - | 38.4±9.2 | 17.9±8.8 | **10.1 ± 3.7** | .31 ± .03 | - | .48 ± .12 | .52 ± .11 | **.21 ± .07** |
| | $\varepsilon'$ | - | - | 7.7 ± .9 | 5.1 ± .4 | **4.9 ± .3** | .14 ± .01 | - | .15 ± .02 | .15 ± .02 | **.11 ± .01** |

Tracking results comparison on AV16.3 dataset

http://cis.eecs.qmul.ac.uk/AV3T.html

# References

[Qian2018] X. Qian et al., “3-D mouth tracking from a compact microphone array co-located with a camera,” in Proc. Int. Conf. Acoust., Speech, Signal Process., Calgary, AB, Canada, Apr. 2018, pp. 3071–3075.

[Kilic2015] V. Kılıc¸, M. Barnard, W. Wang, and J. Kittler, “Audio assisted robust visual tracking with adaptive particle filtering,” IEEE Trans. Multimedia, vol. 17, no. 2, pp. 186–200, Feb. 2015.

[Zotkin2002] D.N. Zotkin, R.Duraiswami, and L. S.Davis, “Joint audio–visual tracking using particle filters,” EURASIP J. Advances Signal Process., vol. 2002, no. 1, pp. 1154–1164, Dec. 2002.

[AV16.3] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, “AV16.3: An audio-visual corpus for speaker localization and tracking,” in Machine Learning for Multimodal Interaction. Martigny, Switzerland: Springer, Jun. 2004.

Thanks for your listening!

Any questions?

Xinyuan Qian

Centre for Intelligent Sensing

Queen Mary University of London