Detection, Localization and Tracking of 3D Audio-Visual Objects

Vassil Khalidov, Florence Forbes, Miles Hansard, Elise Arnaud,
Radu Horaud
INRIA Grenoble, France

In collaboration with Heidi Christensen, Sheffield University

Work done under the POP project:
`http://perception.inrialpes.fr/POP/`

Outline

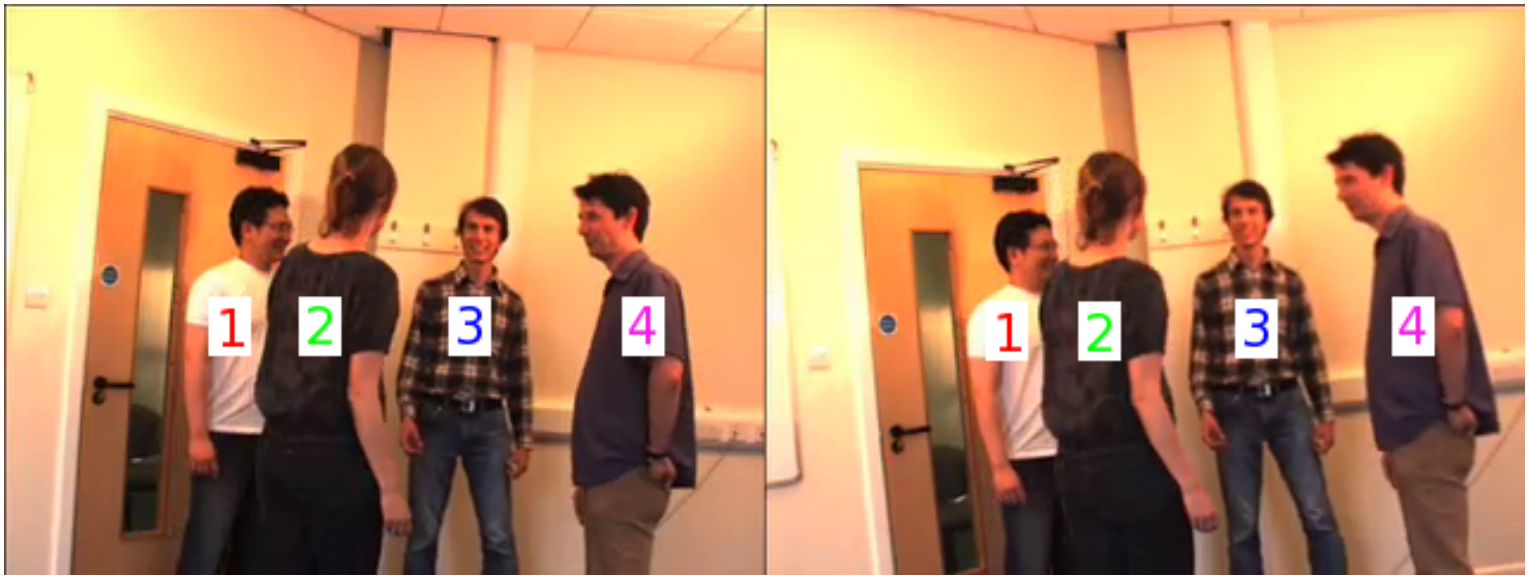
- Introduction
- Audio-Visual Features Integration
- Multimodal Clustering Framework
  - Conjugate Mixture Models
  - Conjugate Particle Tracking
- Experimental Results
- Conclusions and future work

Interpreting the scene...

Several problems to be addressed

- Estimate the number of audio-visual objects
- Localize and track every object
- Select the ones that are involved in auditory activity

Several problems to be addressed



- Estimate the number of audio-visual objects
- Localize and track every object
- Select the ones that are involved in auditory activity

Several problems to be addressed

- Estimate the number of audio-visual objects
- Localize and track every object
- Select the ones that are involved in auditory activity

Several problems to be addressed

- Estimate the number of audio-visual objects
- Localize and track every object
- Select the ones that are involved in auditory activity

Main difficulties

- Steps towards audio-visual perception: How can we recognize an object that is both seen and heard?
- In most natural situations it is difficult to extract unambiguous information from a single modality
- Visual data is dense, light sources are not relevant
- Auditory data is sparse, acoustic sources must be detected in the presence of reverberations
- Integration required !

Main difficulties

- Steps towards audio-visual perception: How can we recognize an object that is both seen and heard?
- In most natural situations it is difficult to extract unambiguous information from a single modality
- Visual data is dense, light sources are not relevant
- Auditory data is sparse, acoustic sources must be detected in the presence of reverberations
- Integration required !

Audio-Visual integration types

- Integration is based on features' coherence
- Two coherence types:
  - static: features related to the same location in ambient space
  - dynamic: correlated evolution through time
- Which features to choose?

Experimental setup

- Stereoscopic camera pair
- Binaural microphone pair

Monocular observations (1)

Right camera image:

Left camera image:

- Detect 'interest points' in the left and right images (Harris et al., 1988)
- $\mathbf{f}^l = \{\boldsymbol{f}^l_1, \ldots, \boldsymbol{f}^l_{m^l}, \ldots, \boldsymbol{f}^l_{M^l}\} \in$ 2D
- $\mathbf{f}^r = \{\boldsymbol{f}^r_1, \ldots, \boldsymbol{f}^r_{m^r}, \ldots, \boldsymbol{f}^r_{M^r}\} \in$ 2D

Monocular observations (2)

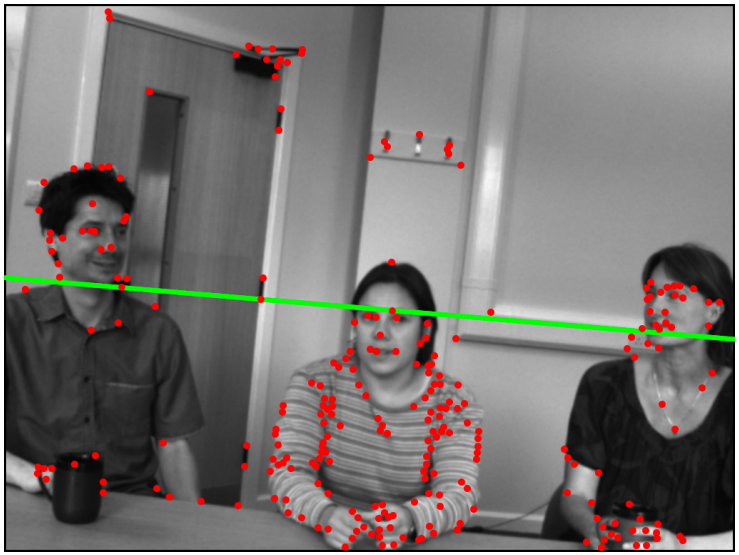
Right camera image:

Left camera image:

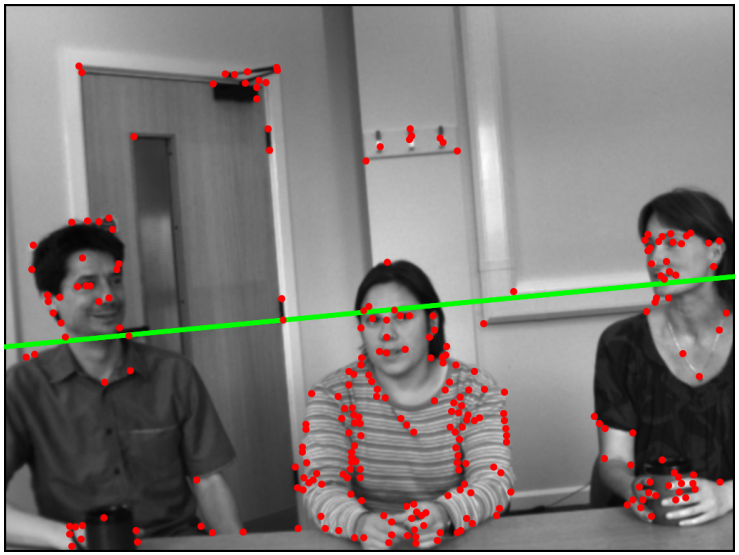
- Detect monocular motion in the left and right images (Lucas et al., 1981)
- $\mathbf{w}^l = \{\boldsymbol{w}^l_1, \ldots, \boldsymbol{w}^l_{m^l}, \ldots, \boldsymbol{w}^l_{M^l}\} \in$ 2D
- $\mathbf{w}^r = \{\boldsymbol{w}^r_1, \ldots, \boldsymbol{w}^r_{m^r}, \ldots, \boldsymbol{w}^r_{M^r}\} \in$ 2D

Binocular observations (1)

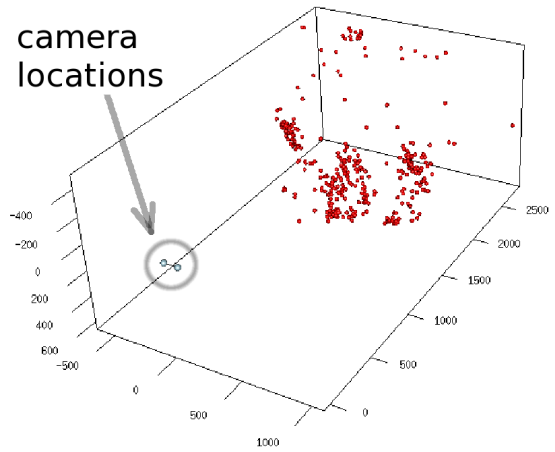
Right camera image:



Left camera image:



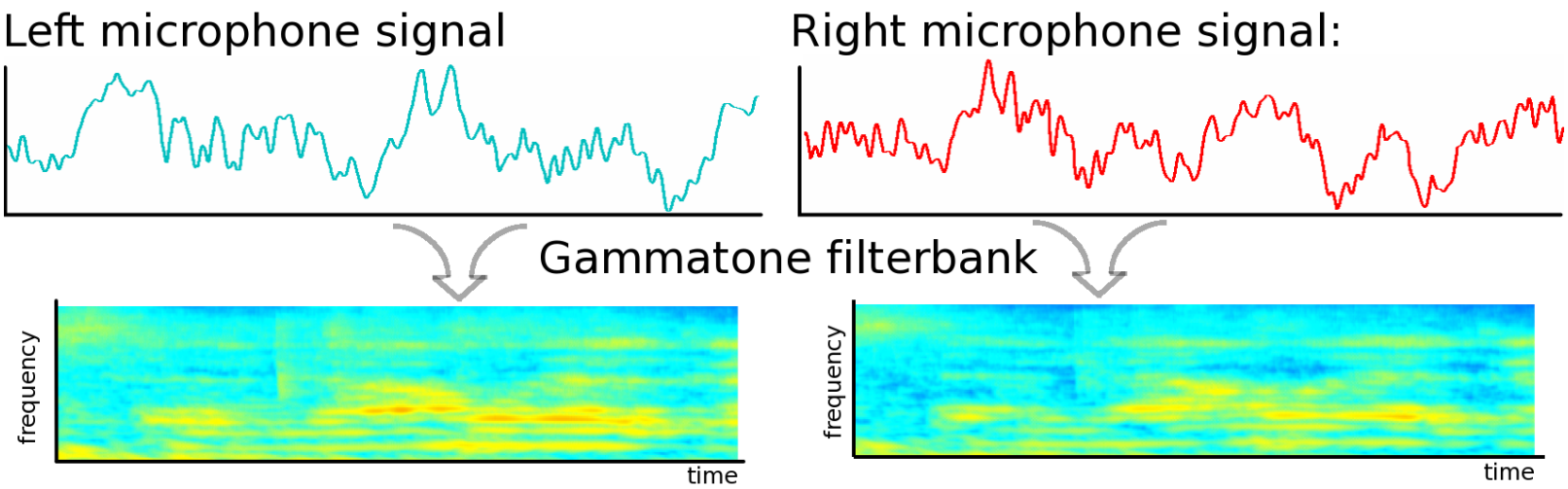
Observations – matched points:



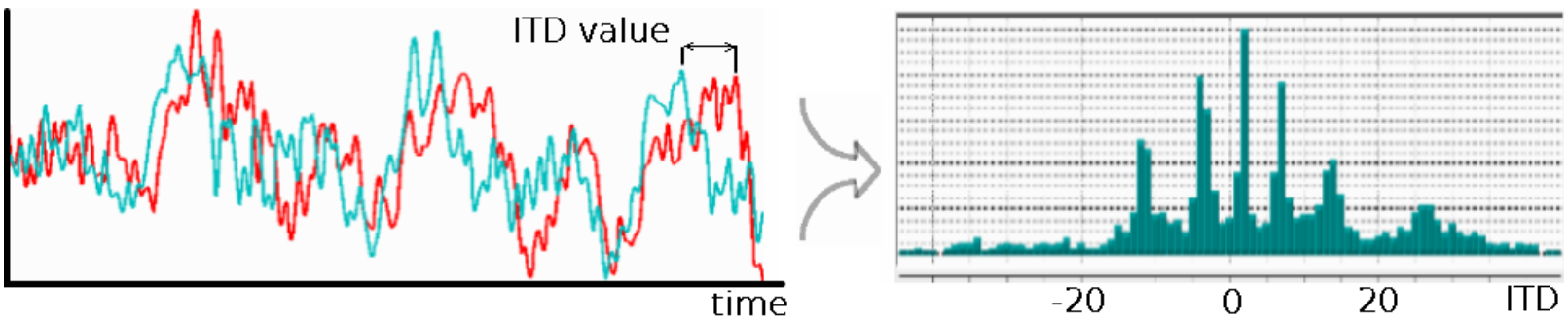
Binocular observations (2)

- $\mathbf{f} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m, \ldots, \boldsymbol{f}_M\} \in$ 3D
- $\boldsymbol{f}_m = (u, v, d)$: $(u, v) =$ pixel, $d =$ disparity
- $\boldsymbol{s} = (x, y, z)$: 3D coordinates of an audio-visual object
- 3D visual disparity is linked to 3D coordinates :
  $\boldsymbol{f}_m = (u, v, d) = \mathcal{F}(\boldsymbol{s}) = \left(\frac{x}{z}, \frac{y}{z}, \frac{B}{z}\right)$

Binaural observations (1)

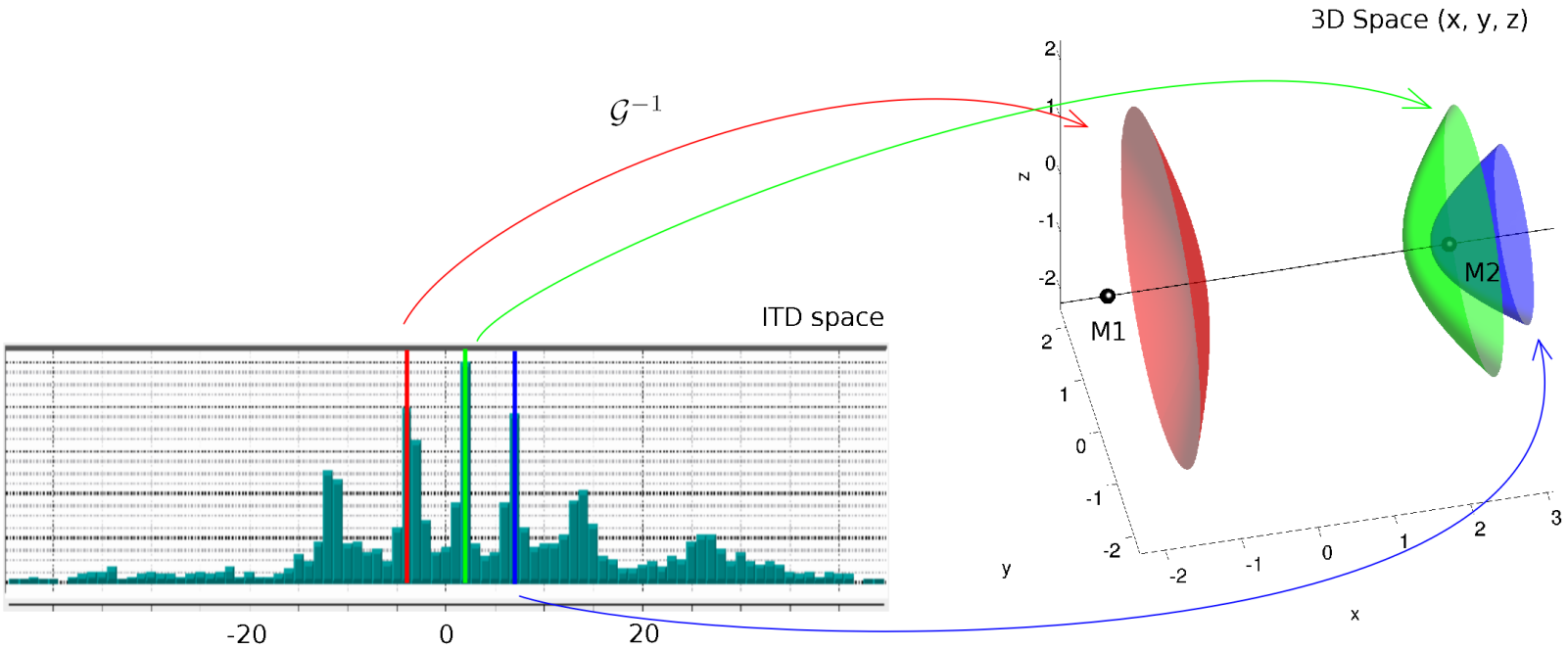


Interaural time differences (ITDs) are obtained through the analysis of cross-correlogram of the filtered signals



Binaural observations (2)

- $\boldsymbol{g} = \{g_1, \ldots, g_k, \ldots, g_K\} \in$ 1D
- $g_k =$ interaural time difference (ITD)
- 1D auditory disparity (ITD) is related to 3D coordinates:
  $g_k = \mathcal{G}(\boldsymbol{s}) = \frac{1}{c}\left(\|\boldsymbol{s} - \boldsymbol{s}_{M_1}\| - \|\boldsymbol{s} - \boldsymbol{s}_{M_2}\|\right)$



The geometry of 3D audio-visual fusion

- $s = (x, y, z)$: 3D coordinates of an audio-visual object
- 1D auditory disparity (ITD):
  $g_k = \mathcal{G}(\boldsymbol{s}) = \frac{1}{c}\left(\|\boldsymbol{s} - \boldsymbol{s}_{M_1}\| - \|\boldsymbol{s} - \boldsymbol{s}_{M_2}\|\right)$
- 3D visual disparity (projective space):
  $\boldsymbol{f}_m = (u, v, d) = \mathcal{F}(\boldsymbol{s}) = \left(\frac{x}{z}, \frac{y}{z}, \frac{B}{z}\right)$
- Auditory and visual data are put on an equal footing

Approach

- Exploit temporal and spatial relations between binocular and binaural observations
- Cast the audio-visual localization problem in the framework of maximum likelihood with missing data
- Maximize the expected complete-data log-likelihood:

$$E[\log P(\underbrace{\text{audio}, \text{vision}}_{\text{observed data}}, \underbrace{\text{audio-visual objects}}_{\text{missing data}})]$$

Associating observations with objects

- Objects $1, 2, \ldots, N$; outlier class $N+1$
- Two sets of assignment variables (the hidden data):
  $\boldsymbol{A} = \{A_1, ..., A_M\}$ for video, and
  $\boldsymbol{A}' = \{A'_1, ..., A'_K\}$ for audio
- The notation $A_m = n$, $n = 1, \ldots, N, N+1$ means that the visual observation m is assigned to object n or it is assigned to an outlier class $N+1$.
- Similarly, $A'_k = n$, $n = 1, \ldots, N, N+1$ means that the audio observation k is assigned to object n or it is assigned to an outlier class $N+1$.
- Both $\boldsymbol{A}$ and $\boldsymbol{A}'$ are missing: the *observation-to-object assignments* are unknowns.

Clustering with conjugate mixture models

- $\boldsymbol{s} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n, \ldots, \boldsymbol{s}_N\}$ - object locations in 3D
- Probability of an observation to belong to an object (inlier):
  - $P(\boldsymbol{f}_m \mid A_m = n) = \mathcal{N}\left(\boldsymbol{f}_m \mid \mathcal{F}(\boldsymbol{s}_n), \Sigma_n\right)$
  - $P(g_k \mid A'_k = n) = \mathcal{N}\left(g_k \mid \mathcal{G}(\boldsymbol{s}_n), \sigma_n^2\right)$
  - Alternatively, we can also use the t-distribution.
  - The methodology is independent of the normal/t-distribution choice.
- Likelihood of an observation to be an outlier:
  - $P(\boldsymbol{f}_m \mid A_m = N+1) = \mathcal{U}\left(V_{3D}\right) = \frac{1}{V}$
  - $P(g_k \mid A'_k = N+1) = \mathcal{U}\left(U_{1D}\right) = \frac{1}{U}$
- The model's parameters:

$$\boldsymbol{\theta} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N, \Sigma_1, \ldots, \Sigma_N, \sigma_1, \ldots, \sigma_N\}$$

The observed-data log-likelihood

We assume that the variables are independent and identically distributed, then:

$$\log P(\mathbf{f}, \boldsymbol{g}) = \sum_{n=1}^{N+1} \log \left(\pi_n \sum_{m=1}^{M} P(\boldsymbol{f}_m | A_m = n) + \pi'_n \sum_{k=1}^{K} P(\boldsymbol{g}_k | A'_k = n) \right)$$

where π_n and π'_n are the prior probabilities.

The maximization of the log-likelihood is not tractable because of the presence of the hidden variables.

Maximum likelihood with hidden data

- The expectation-maximizaton (EM) algorithm finds a local maximum of the likelihood

- Maximize the expected complete-data log-likelihood w.r.t. $\boldsymbol{\theta}$:

$$
\begin{aligned}
& E_{\boldsymbol{A},\boldsymbol{A}'}[\log P(\mathbf{f},\boldsymbol{g},\boldsymbol{a},\boldsymbol{a}')|\mathbf{f},\boldsymbol{g}] = \\
& -\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\alpha_{mn}\left(\left(\boldsymbol{f}_m-\mathcal{F}(\boldsymbol{s}_n)\right)^{\top}\Sigma_n^{-1}\left(\boldsymbol{f}_m-\mathcal{F}(\boldsymbol{s}_n)\right)+\log|\Sigma_n|\right) \\
& -\frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N}\alpha'_{kn}\left((g_k-\mathcal{G}(\boldsymbol{s}_n))^2/\sigma_n^2+\log\sigma_n^2\right) \\
& + \text{ other constant terms}
\end{aligned}
$$

Posterior probabilities

- α_{mn} and α'_{kn} are the posterior probabilities of a visual observation m and of an auditory observation k to belong to the speaker n (or to be an outlier), given the observations:

$$\alpha_{mn} = P(A_m = n|\boldsymbol{f}_m) \quad \text{and} \quad \alpha'_{mn} = P(A'_k = n|g_k)$$

- Can be computed explicitly using Bayes' formula

Maximizing the expectation

- Dependency of the target function on θ is expressed through non-linear mappings $\mathcal{F}$ and $\mathcal{G}$
- Maximization problem does not yield a closed-form solution as in the standard EM algorithm
- Conjugate EM algorithm: method of generations combining local ascent and global search
- Several possibilities to speed-up
  - compute local Lipschitz constants for gradient for local ascent
  - sample manifolds $\mathcal{F}^{-1}(\cdot)$ and $\mathcal{G}^{-1}(\cdot)$ for global search

Interpreting the algorithm's output

- Object locations and their audio-visual characteristics:

$$\boldsymbol{\theta} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N, \Sigma_1, \ldots, \Sigma_N, \sigma_1, \ldots, \sigma_N\}$$

- Auditory activity estimation:
  - assign every audio observation to an object or to an outlier class using the MAP estimate:

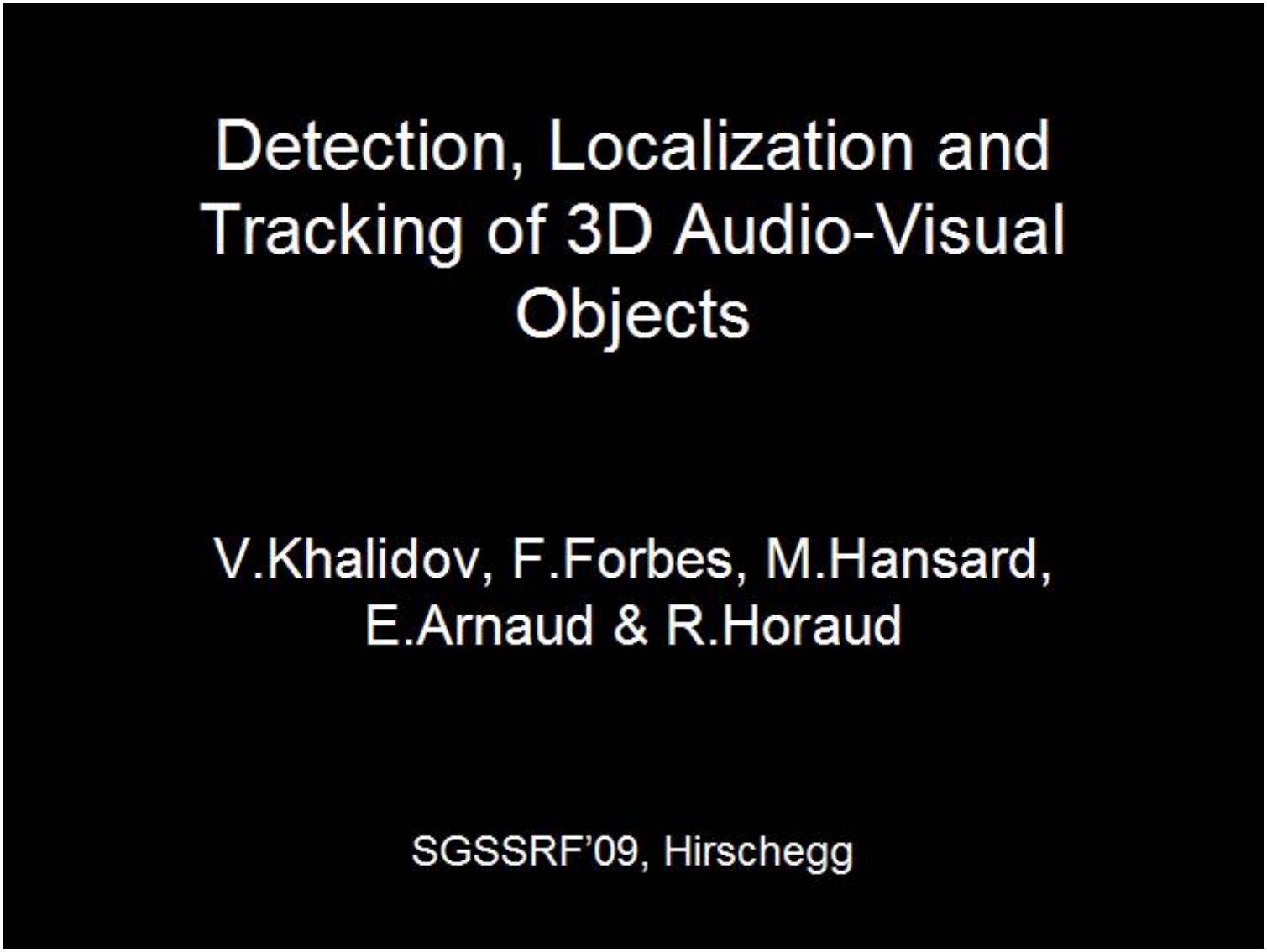
$$g_k \text{ was generated by } \eta'_k, \text{ where } \eta'_k = \underset{n=1,\ldots,N+1}{\arg\max}\ \alpha'_{kn}$$

  - for each object make a decision on auditory activity based on the number of assigned observations

Implementation

- Verified on CAVA database: `http://perception.inrialpes.fr/CAVA_Dataset`
- Each sequence of input data is split into time-intervals, each interval is 1/8 seconds long
- There are 1000 visual observations and 10 auditory observations per time-interval
- Next interval EM iteration is initialized with the results of the previous iteration

Examples

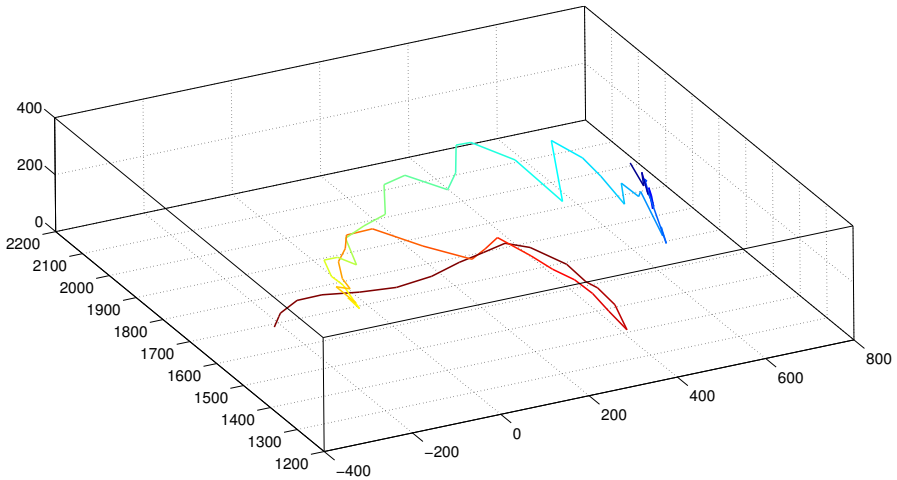

Detection, Localization and Tracking of 3D Audio-Visual Objects
V.Khalidov, F.Forbes, M.Hansard, E.Arnaud & R.Horaud
SGSSRF'09, Hirschegg

Qualitative results

- Auditory activity estimation statistics:

	time-int	*AV-int*	*AV-OK*	*AV-missed*	*AV-false*
M1	166	89	75	0.16	0.14
TTOS1	76	69	60	0.13	0.43

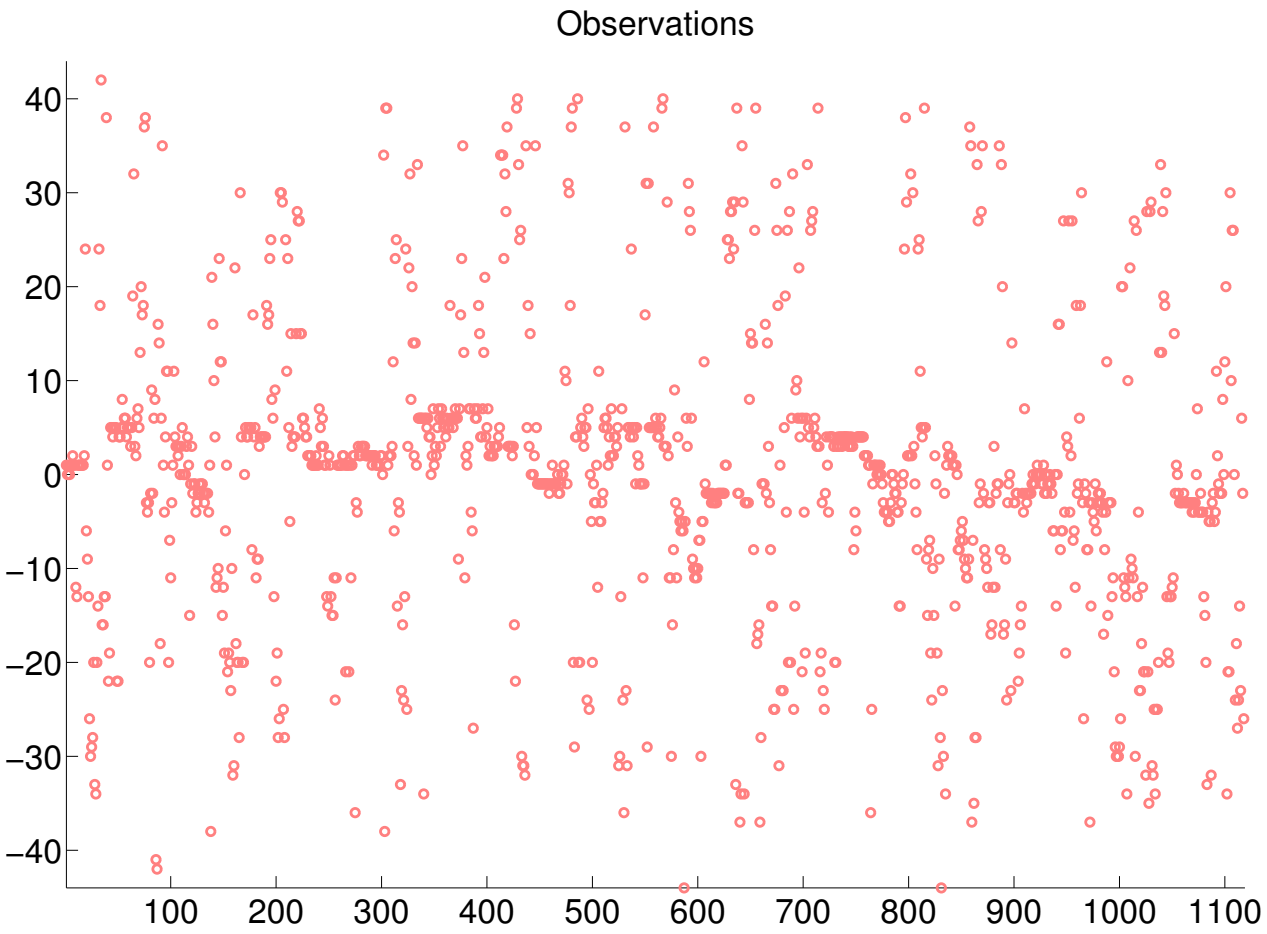
- Localization accuracy is ± 10cm



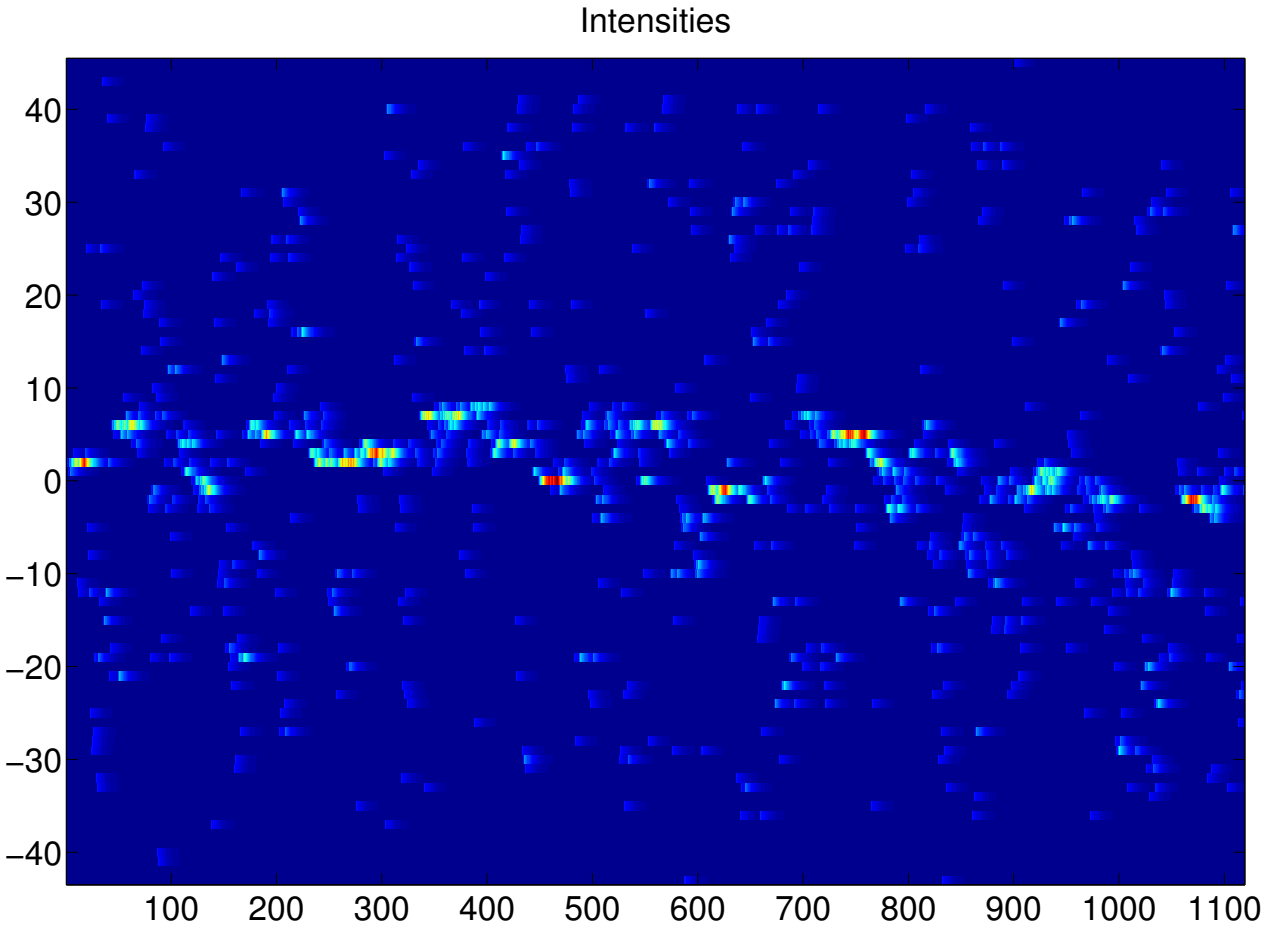
Major drawbacks of the approach

- Dynamic cues (motion vectors) are not included into the model
- Temporal properties of observation streams are ignored
- Possibilities to follow dynamic configuration changes are limited
- Background model is weak, assignments are incorrect

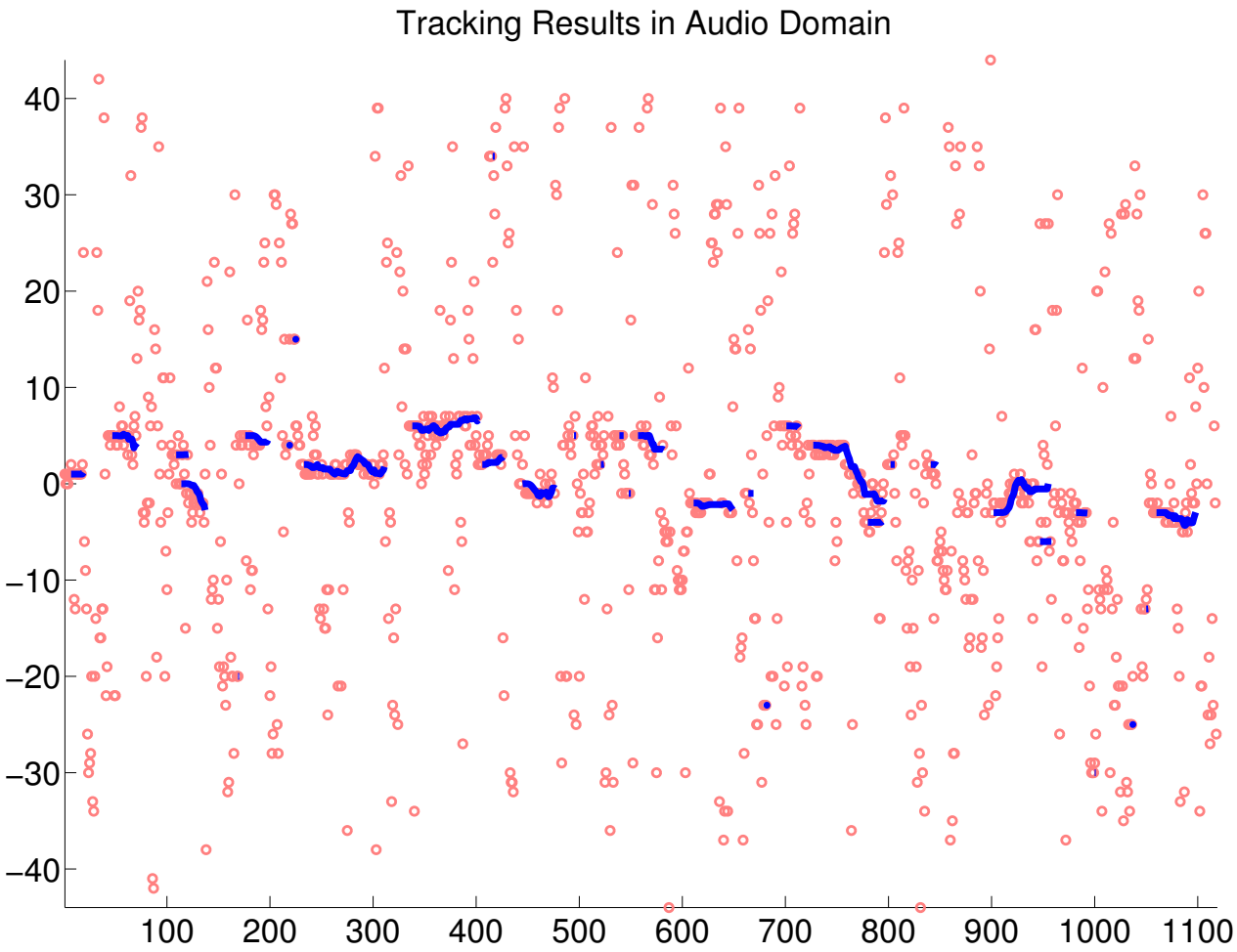
Audio stream intensity



Audio stream intensity



Audio stream intensity



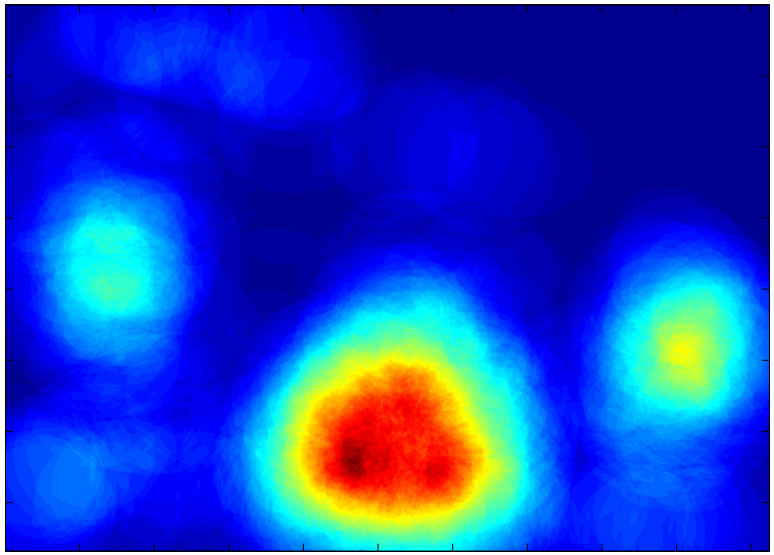
Video observations density

Right camera image:

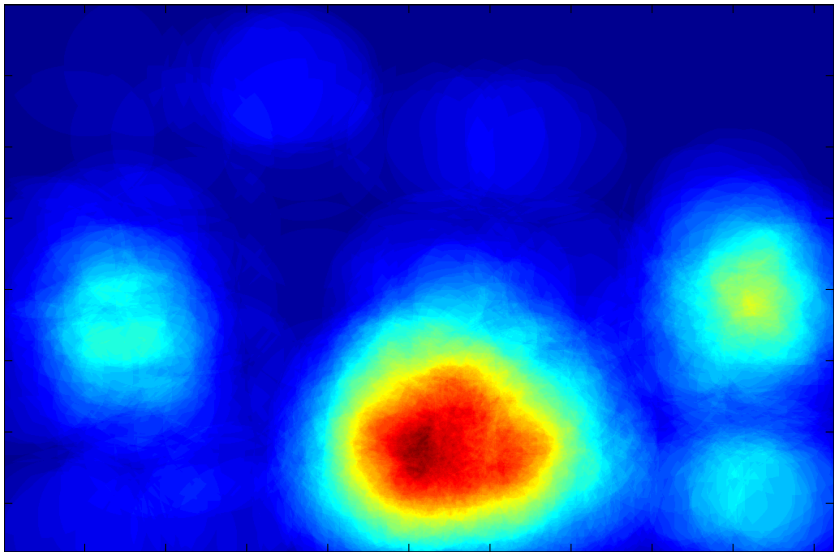
Left camera image:

Video observations density

Right camera image:



Left camera image:



Ideas to improve the model

- Audio observations as a point process in ITD space
- Monocular video observations as point processes in image spaces
- Audio-visual object dynamics: diffusion in the parameter space
- Background is stable, motion vector is $\mathbf{0}$
- Learning background depth map

Object dynamics

- State evolution

$$\begin{aligned} \mathrm{d}\boldsymbol{s}_t &= \boldsymbol{w}(t)\mathrm{d}t + \sigma_{\mathrm{s}}(t)\mathrm{d}B_t \\ \mathrm{d}\boldsymbol{w}_t &= \sigma_{\mathrm{w}}(t)\mathrm{d}B_t \end{aligned}$$

- Binocular observations at time t

$$\boldsymbol{f}_m = \begin{cases} \mathcal{F}(\boldsymbol{s}_m) + \sigma_{\mathrm{f}} W, & \text{if computed correctly} \\ \mathcal{U}(\mathbb{F}), & \text{otherwise} \end{cases}$$

- Audio observations at time t

$$g_k = \begin{cases} \mathcal{G}(\boldsymbol{s}) + \sigma_{\mathrm{g}} W, & \text{if computed correctly} \\ \mathcal{U}(\mathbb{G}), & \text{otherwise} \end{cases}$$

Image space processes

- Image points should follow projected dynamics equations

$$\begin{aligned} \mathrm{d}\boldsymbol{f}_t^l &= \boldsymbol{w}^l(t)\mathrm{d}t + \sigma_{\mathrm{f}^l}(t)\mathrm{d}B_t \\ \mathrm{d}\boldsymbol{w}_t^l &= \sigma_{\mathrm{w}^l}(t)\mathrm{d}B_t \end{aligned}$$

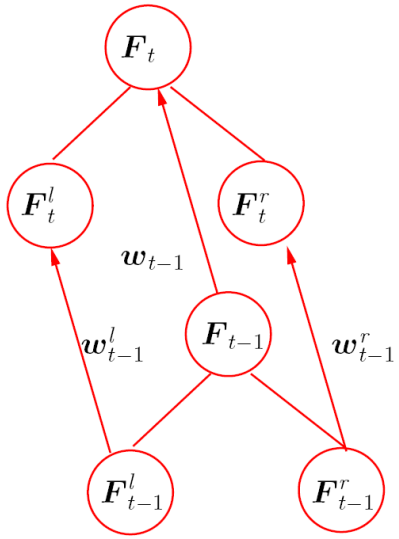
- Points configuration is defined by a pairwise interaction Markov point process:

$$P(\mathbf{f}^l|\boldsymbol{a}^l;\,\nu) \quad \propto \quad \mu^{M^l}\prod_{i \underset{a^l}{\sim} j} h(\|\boldsymbol{f}_i^l - \boldsymbol{f}_j^l\|)$$

$$i \underset{a^l}{\sim} j \quad \Longleftrightarrow \quad a_i^l = a_j^l \text{ and } \|\boldsymbol{f}_i^l - \boldsymbol{f}_j^l\| \leq r$$

Model consistency conditions (1)

- Redundancy in process definition



- Set of *error flags* $\{\boldsymbol{e}, \boldsymbol{e}^l, \boldsymbol{e}^r\}$ to define whether to use 3D motion model, projected motion model or the point process

- One more set of hidden variables to estimate

Model consistency conditions (2)

- Joint distributions of observation assignments $\{\boldsymbol{a}, \boldsymbol{a}^l, \boldsymbol{a}^r\}$ and error flags $\{\boldsymbol{e}, \boldsymbol{e}^l, \boldsymbol{e}^r\}$ are highly important

- Examples of rules:

$$P\left(e_m \,|\, a^l_{\tau^l(m)},\, a^r_{\tau^r(m)}\right) = \begin{cases} p^{\mathrm{MA}} 1_{\{e_m=1\}} + (1-p^{\mathrm{MA}}) 1_{\{e_m=0\}}, & \text{if } a^l_{\tau^l(m)} = a^r_{\tau^r(m)} \\ 1_{\{e_m=1\}}, & \text{otherwise.} \end{cases}$$

$$P\left(A_m = n \,|\, e_m,\, a^l_{\tau^l(m)},\, a^r_{\tau^r(m)}\right) = \begin{cases} 1_{\{a^l_{\tau^l(m)}=n\}}, & \text{if } a^l_{\tau^l(m)} = a^r_{\tau^r(m)} \text{ and } e_m = 0 \\ \frac{1}{N+1}, & \text{otherwise.} \end{cases}$$

Inference for conjugate particle tracking model

- Observations
$$\mathfrak{o} = \{\mathbf{f}, \mathbf{f}^l, \mathbf{f}^r, \mathbf{w}, \mathbf{w}^l, \mathbf{w}^r, g\}$$
- Model parameters
$$\boldsymbol{\theta} = \{\text{positions, motion vectors, (co)variances, priors}\}$$
- Hidden variables
$$\mathcal{A} = \{\boldsymbol{A}, \boldsymbol{A}^l, \boldsymbol{A}^r, \boldsymbol{A}'\} \text{ and } \mathcal{E} = \{\boldsymbol{E}, \boldsymbol{E}^l, \boldsymbol{E}^r\}$$
- Cannot apply the EM algorithm directly, neither of the steps has a closed form solution

Variational EM approximation

- Consider restricted class of distributions $\tilde{\mathcal{D}}$:

$$q_{A^l}(\boldsymbol{a}^l|\mathbf{f}^l) = \prod_{m^l=1}^{M^l} q_{A^l_{m^l}}(a^l_{m^l}|\mathbf{f}^l) \quad \text{and} \quad q_{A^r}(\boldsymbol{a}^r|\mathbf{f}^r) = \prod_{m^r=1}^{M^r} q_{A^r_{m^r}}(a^r_{m^r}|\mathbf{f}^r)$$

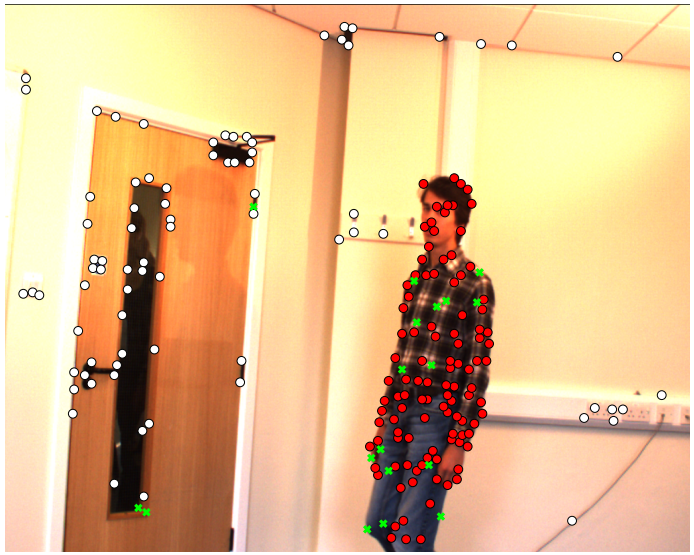
- VE-step: find distribution $q^{\eta+1}_{\mathcal{A},\mathcal{E}}$ such that

$$q^{\eta+1}_{\mathcal{A},\mathcal{E}} = \underset{q_{\mathcal{A},\mathcal{E}} \in \tilde{\mathcal{D}}}{\arg\max}\, \mathbb{E}_{q_{\mathcal{A},\mathcal{E}}}\{\log P(\mathfrak{o}, \mathcal{A}, \mathcal{E}; \boldsymbol{\theta}^\eta)\} + I[q_{\mathcal{A},\mathcal{E}}],$$

- M-step: use the method of generations to find position parameters $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$

Experimental results (1) - TTOS1

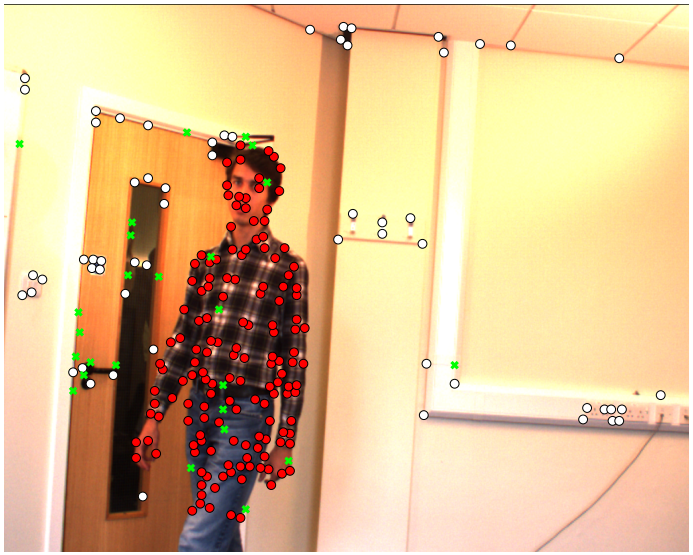
Right camera image:



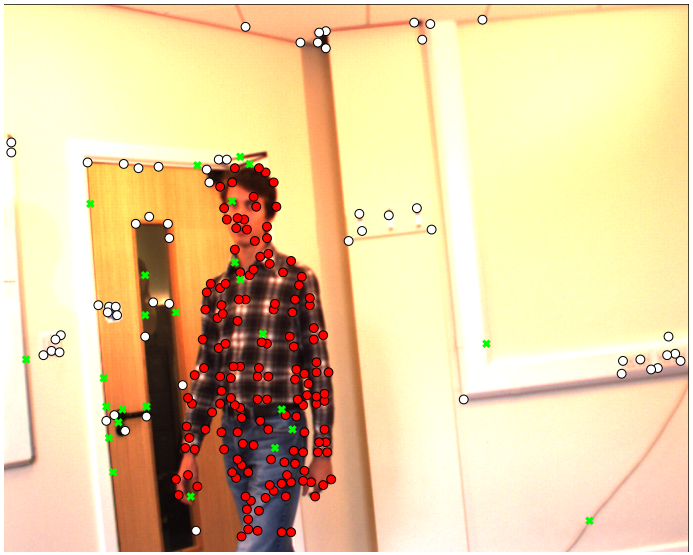
Left camera image:

Experimental results (1) - TTOS1

Right camera image:

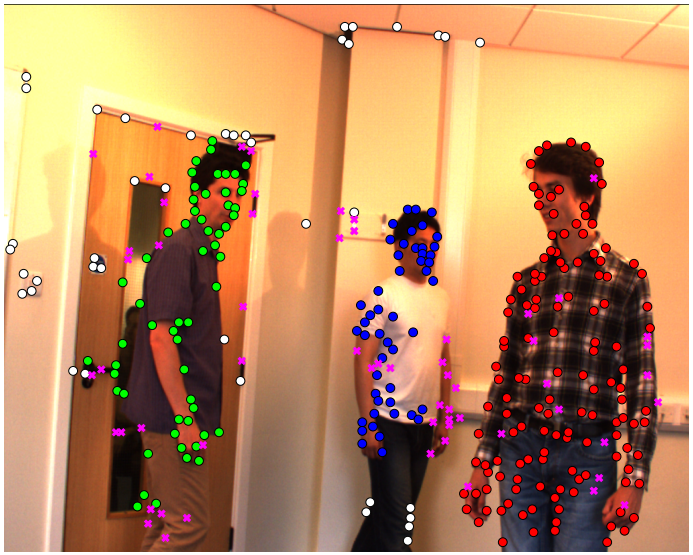


Left camera image:

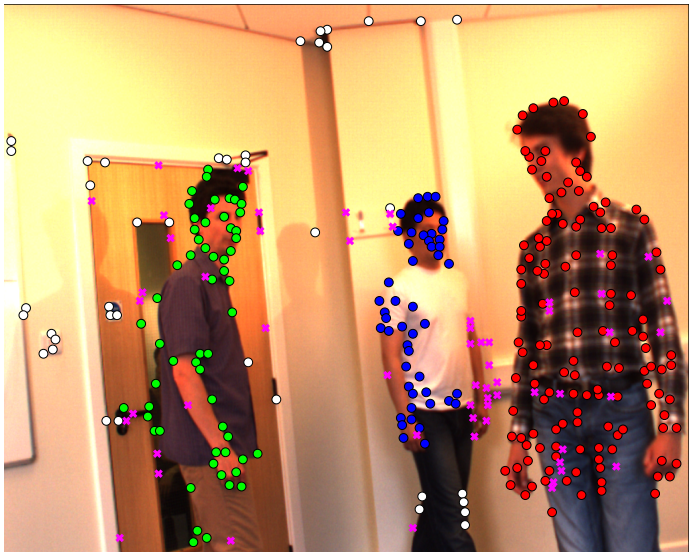


Experimental results (2) - CTMS3

Right camera image:

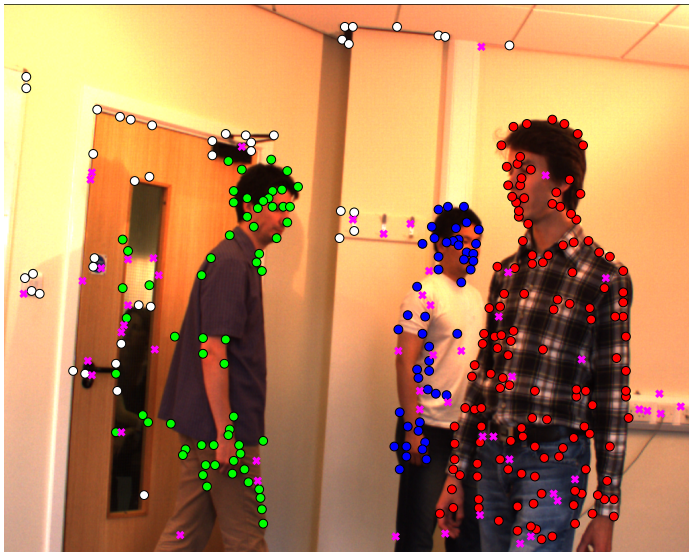


Left camera image:

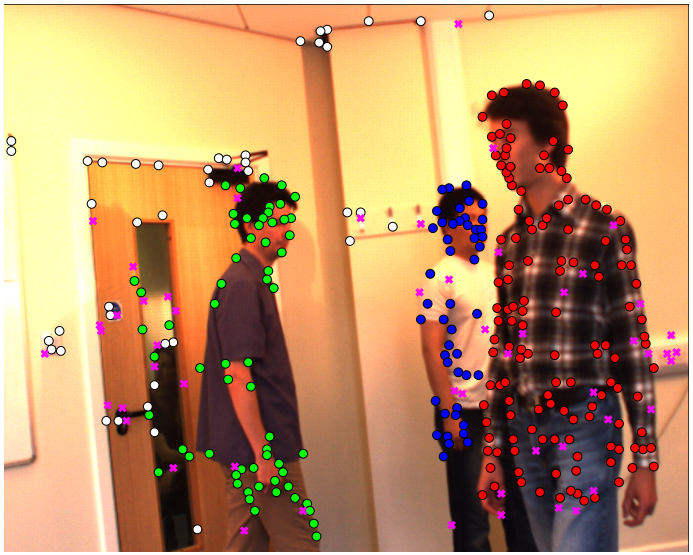


Experimental results (2) - CTMS3

Right camera image:



Left camera image:



Conclusions and future work

- We introduced a framework for audio-visual fusion using spatial and temporal coherence.
- The method is based on unsupervised clustering using the variational EM algorithm.
- Unified approach: the same model for detection, localization and auditory activity estimation.
- Possibility to automatically correct erroneous observations.
- A more sophisticated model should include eye and head motions, i.e., active perception and attention.

Thank you!

Vassil Khalidov <vasil.khalidov@inrialpes.fr>