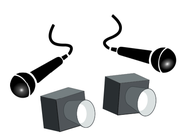
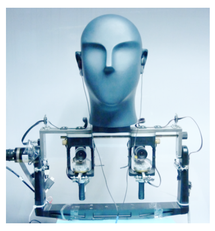
Alignment of Binocular-Binaural Data Using a Moving Audio-Visual Target

Vasil Khalidov, Florence Forbes, and Radu Horaud

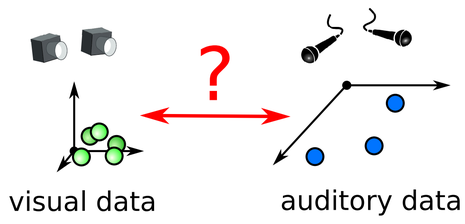
Motivation

Setup: Robotic head, 2 microphones, 2 cameras





Goal: Establish spatial alignment of auditory and visual modalities

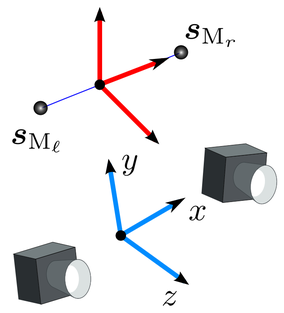


Approach

A/V data alignment through microphone location estimation

Visual observations model:

$$\boldsymbol{f} = (u, v, d) = \mathcal{F}(\boldsymbol{s}) = (x/z, y/z, 1/z)^\top$$

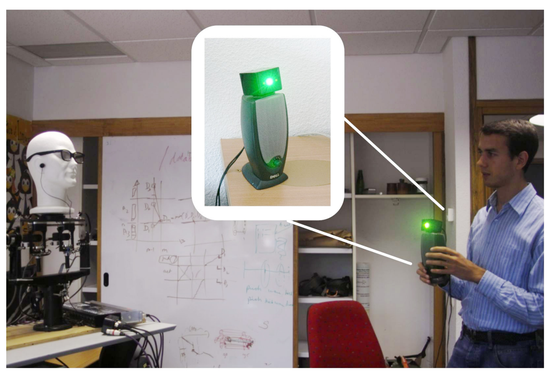


Auditory observations model:

$$g = \mathcal{G}(\boldsymbol{s};\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}) = c^{-1}\left(\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\ell}\| - \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}\|\right)$$

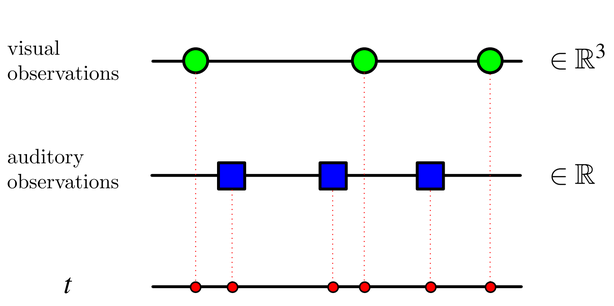
parameters to estimate

Experimental Setup



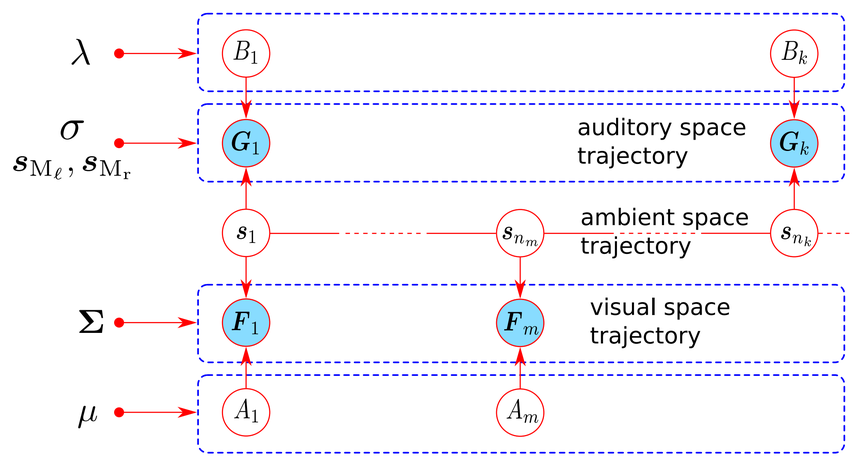
- Unconstrainted environment
- Calibration rig: a speaker with a mounted LED light bulb
- White noise is played for better auditory localisation

Challenges



- A&V observations are noisy and reside in different physical spaces
- A&V observations are not aligned
- Overall sampling rate is not constant

Multimodal Trajectory Matching



Observation likelihoods:

$$P(\boldsymbol{f}_m|\boldsymbol{s}_m) = \mu\mathcal{N}(\boldsymbol{f}_m|\mathcal{F}(\boldsymbol{s}_m), \Sigma) + (1-\mu)\mathcal{U}(V),$$
$$P(g_k|\boldsymbol{s}_k) = \lambda\mathcal{N}(g_k|\mathcal{G}(\boldsymbol{s}_k;\ \boldsymbol{s}_{\mathrm{M}_\ell}, \boldsymbol{s}_{\mathrm{M}_\mathrm{r}}), \sigma) + (1-\lambda)\mathcal{U}(U)$$

inliers outliers

A_m, B_k - auxiliary inlier/outlier indicator variables for V and A observations

Prior on trajectory regularity:

$$P(\mathbf{s}) \propto \exp\Big(-\gamma \sum_n \|\boldsymbol{s}_{n+1} - \boldsymbol{s}_n\|^2/(t_{n+1} - t_n)\Big)$$

Estimation: EM algorithm

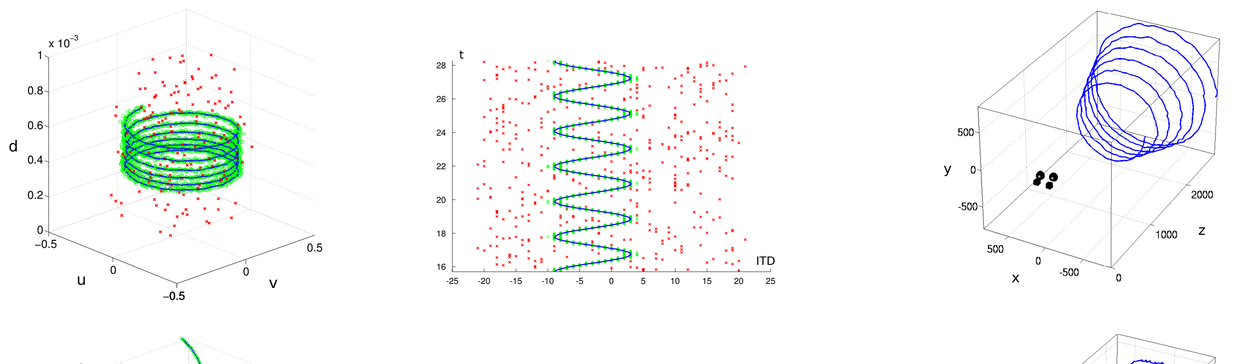
- **E**: compute posteriors
  $\alpha_m = P(A_m = inlier|\boldsymbol{f}_m)$
  $\beta_k = P(B_k = inlier|g_k)$
- **CM-1**: update $\boldsymbol{s}_{\mathrm{M}_\ell}$, $\boldsymbol{s}_{\mathrm{M}_\mathrm{r}}$
- **CM-2**: update $\{\boldsymbol{s}_n\}_{n=1}^N$
- **CM-3**: update $\mu, \lambda, \Sigma, \sigma$

Initialisation:
$\mathbf{s}^{(0)}$ from V observations (triangulation)
$\boldsymbol{s}^{(0)}_{\mathrm{M}_\ell}$, $\boldsymbol{s}^{(0)}_{\mathrm{M}_\mathrm{r}}$ from A observations

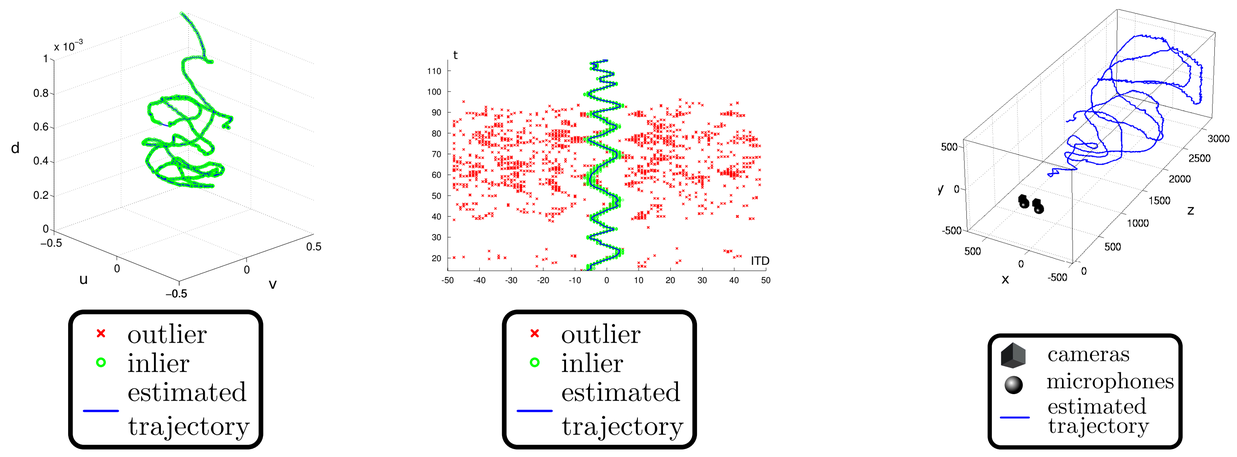
Experimental Results

Reconstructed V data | Reconstructed A data | Reconstructed 3D trajectory

Noise 2, R



Real 2



Algorithm	Scenario									
	N0	N1	N1, R	N2	N2, R	N3	N3, R	R1	R2	R3
Naive	0	0.13	0.23	0.21	0.3	3.46	3.52	0.32	0.6	0.41
Proposed	0	**0.05**	**0.14**	**0.13**	**0.21**	**3.32**	**3.4**	**0.27**	**0.28**	**0.33**

Table 2. Trajectory misalignment measures for the proposed and "naive" calibration algorithms.

Scenario	Acronym	σ	Σ	Rounded
Noiseless	N0	0	$\mathrm{diag}(0,0,0)$	no
Noise 1	N1	0.05	$\mathrm{diag}(10^{-6}, 10^{-6}, 10^{-14})$	no
Noise 1, R	N1, R	0.05	$\mathrm{diag}(10^{-6}, 10^{-6}, 10^{-14})$	yes
Noise 2	N2	0.1	$\mathrm{diag}(10^{-4}, 10^{-4}, 10^{-11})$	no
Noise 2, R	N2, R	0.1	$\mathrm{diag}(10^{-4}, 10^{-4}, 10^{-11})$	yes
Noise 3	N3	0.5	$\mathrm{diag}(10^{-2}, 10^{-2}, 10^{-8})$	no
Noise 3, R	N3, R	0.5	$\mathrm{diag}(10^{-2}, 10^{-2}, 10^{-8})$	yes
Real 1	R1	narrow field of view		
Real 2	R2	medium field of view		
Real 3	R3	large field of view		

Table 1. Simulated and real data sets with the corresponding auditory (σ) and visual (Σ) (co-)variance values or setting description.

Scenario	Proposed				Naive			
	L	R	A	M	L	R	A	M
N0	1.3	1.3	0	5.9	**0.7**	**0.8**	0	0
N1	**19.2**	**19.6**	**2.28**	**27.91**	223.2	224.1	87.8	7899.9
N1, R	**40.0**	**40.4**	**2.73**	**31.04**	228	230.8	96	8328.8
N2	**57.7**	**57.8**	**12.77**	**35.2**	226.6	230.3	112.5	7830.1
N2, R	**32.6**	**32.7**	12.65	**32.27**	248.2	251.8	**8.3**	7973.2
N3	248.6	250.6	**215.22**	**406.73**	**239.3**	**242.7**	575.3	12013.1
N3, R	**212.8**	**211.5**	**118.11**	**346.98**	222.8	224.6	556	11192.1

Table 3. Estimated left (L) and right (R) microphone location errors and average (A) and maximum (M) distances between points of the estimated and ground truth trajectory in simulated data experiments.

Conclusions

- Achieves accurate alignment of A&V modalities, microphone locations are not precise
- Works in unconstrainted environments with significant noise levels
- Allows for extrapolation ("turn head towards the sound" task)

Contacts and Funding

Vasil.Khalidov@idiap.ch,
{Florence.Forbes,Radu.Horaud}@inria.fr

This work was supported by the Humavips project, EU grant FP7-ICT-2010-247525.