Paper ID 801

Audio-Video detection of the active speaker in meetings

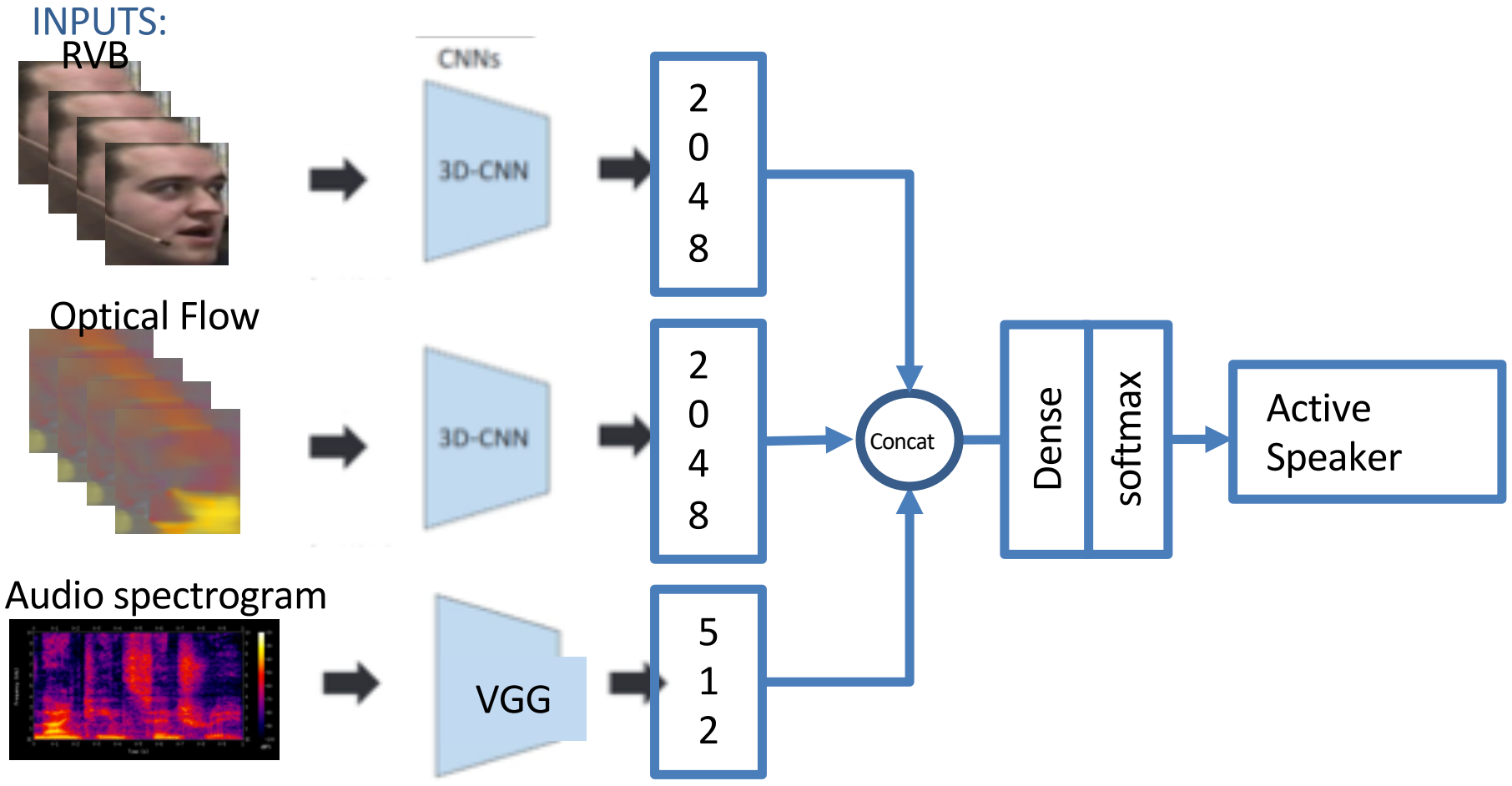
25th International Conference on Pattern Recognition.

Francisco Madrigal, Frédéric Lerasle,Lionel Pibre and Isabelle Ferrané
LAAS-CNRS, IRIT, Université de Toulouse,
Toulouse, France

January 2021

Motivation

- Fusion audio and video cues with contextual information to estimate speaker
  - Audio and video cues :



Contextual information Framework

Objectives :

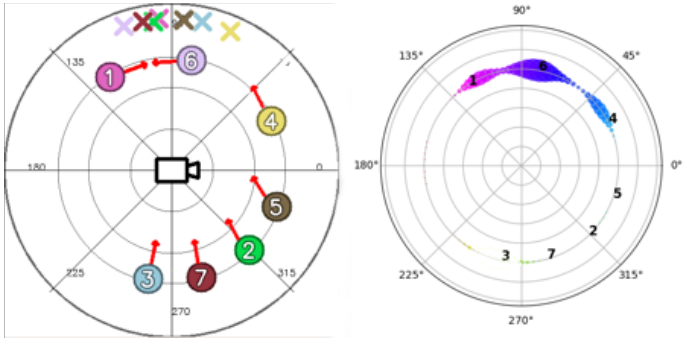
- Identify speakers during speaking turns in meetings

Conditions :

- Participants look globally towards the main speaker
- → Convergence of gazes towards the speaker
- Orientation of gaze = orientation of the head
- Few public multi-person audio-video datasets…

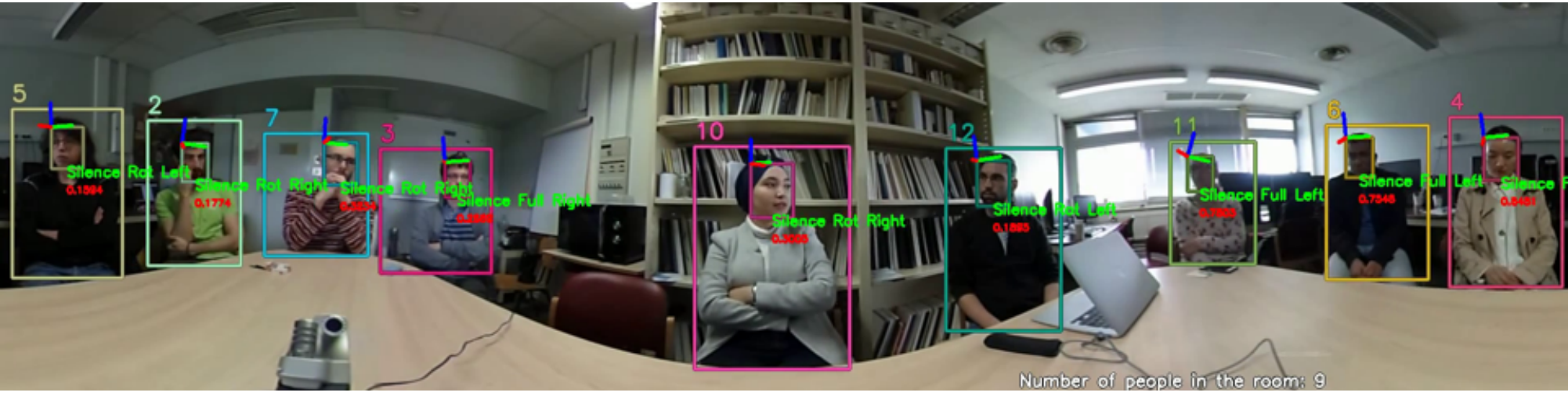
Steps :

- Face orientation characterization / detection
- Orientation projection to a topological space
- Estimate the distribution of the active speaker



Visual speaker detection

Step 1 - "HyperFace" face orientation detection



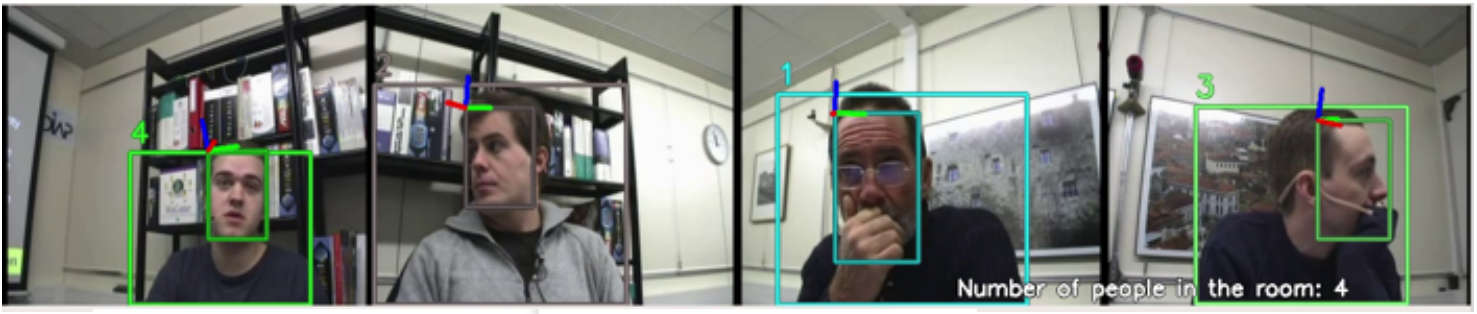
For each face detected:

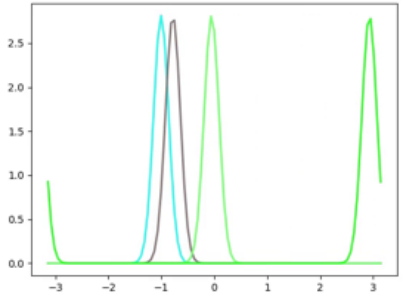
- HyperFace (opensource): Upper left corner, the RGB lines represent the rotation

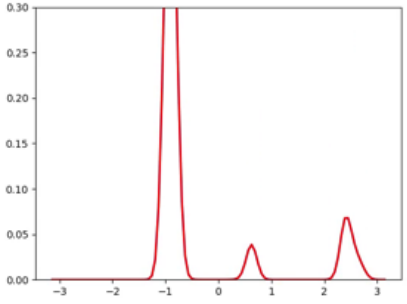
Visual speaker detection

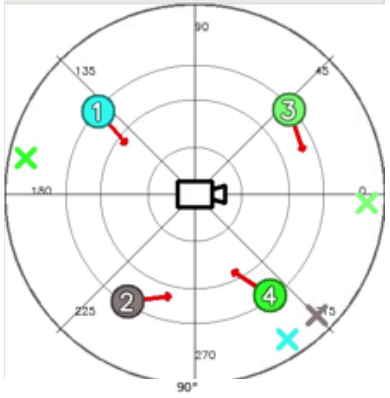
Step 2 - Projection of the orientation of the face onto the topological space

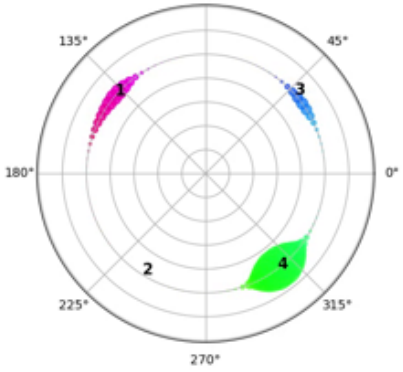
- Project a point due to the detected orientation angle (red arrows).
- Calculate a “von Mises” distribution on the inferred position.
- Combine distributions and project them onto the topology.







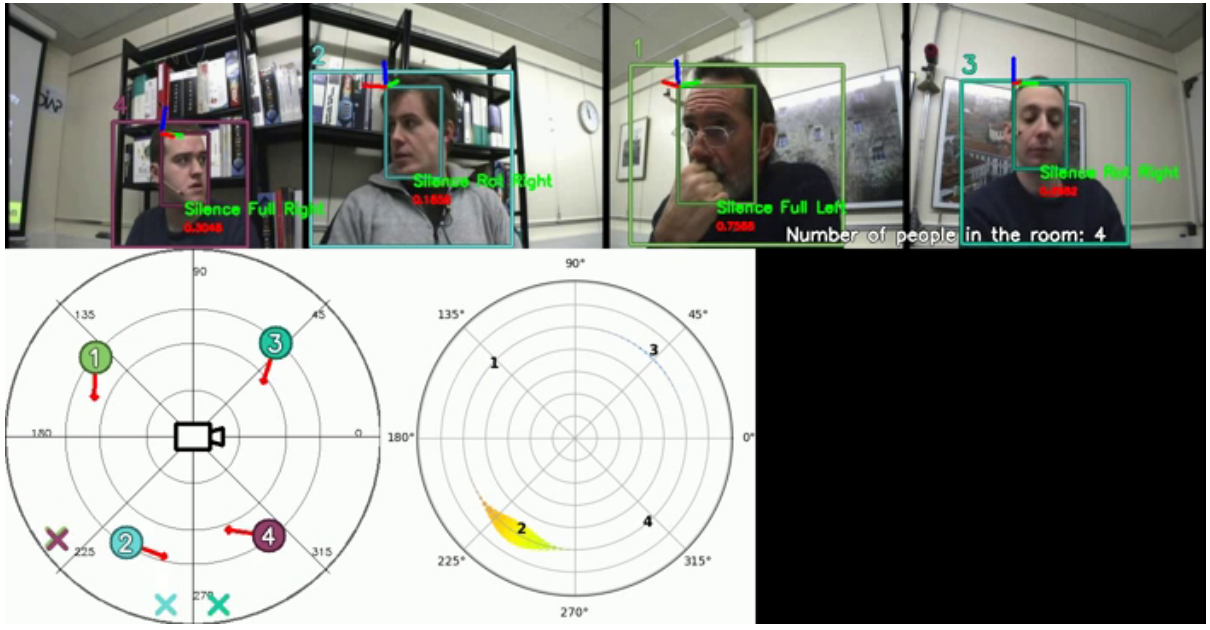




Visual speaker detection

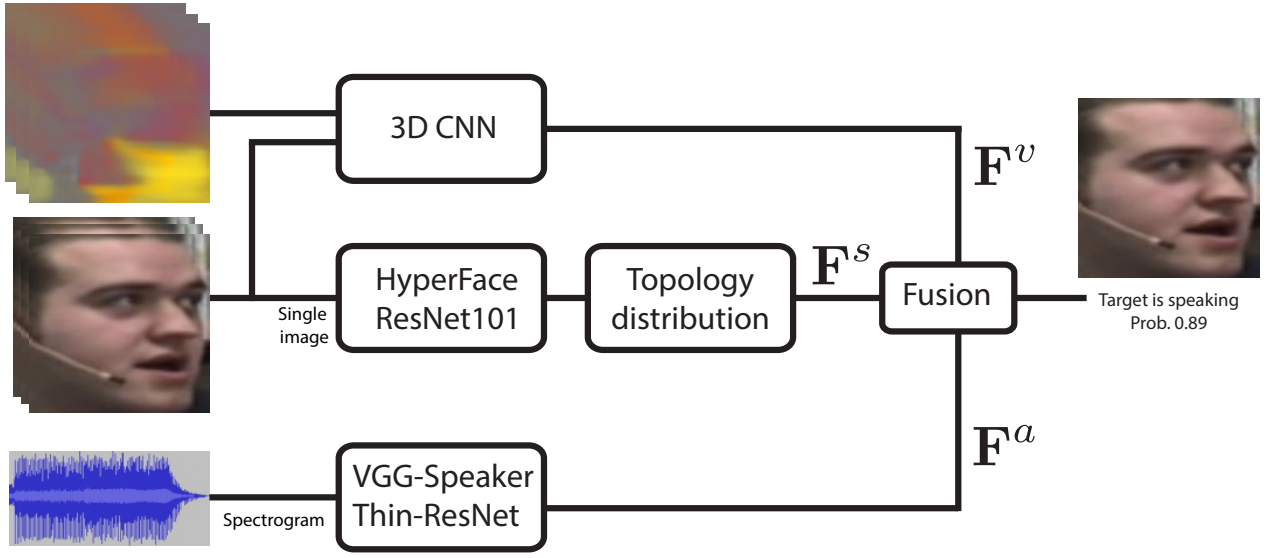
Step 3 - Fusion of all the probabilities

Video : https://drive.google.com/file/d/1P4hYdfmq61TEOdpYdMOpCG0IML167aDd/view?usp=sharing



Visual speaker detection

- Multi-cue fusion



Probability distribution of each cue is fusion according to:

$$\mathbf{F}_i = \begin{cases} (\mathbf{F}_i^a + \mathbf{F}_i^v + \mathbf{F}_i^s)/3, & \text{if } \forall \mathbf{F}_i^* > 0.5 \text{ or } \forall \mathbf{F}_i^* < 0.5 \\ (\mathbf{F}_i^\alpha + \mathbf{F}_i^\beta)/2, & \text{if } \forall \mathbf{F}_i^\alpha > 0.5 \text{ and } \mathbf{F}_i^\beta > 0.5 | \alpha \in \{a, v, s\}, \beta \in \{a, v, s\} \text{ and } \alpha \neq \beta \end{cases}$$

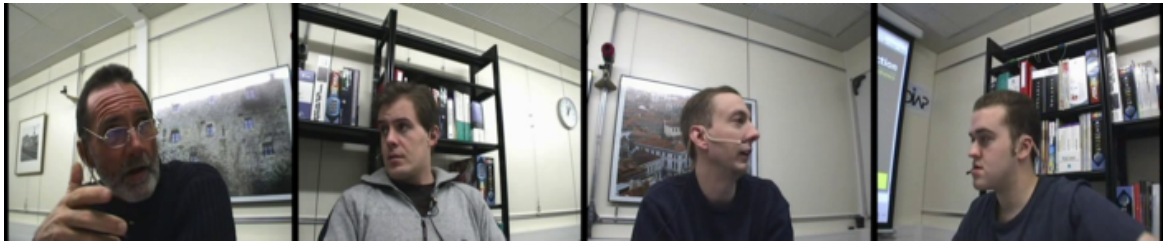
Audio + video database

Ami corpus

- Multi-perspective cameras, single target
- +70 participants
- 4 cameras: focused on each participant
- Audio recorded through individual headphones
- Environment simulating a real meeting
- Audio-based ground truth

- The data is divided into 5 batches we call **CV**, one for evaluating and the rest for training (Cross-validation).
- Creation of a synthetic 360º image by concatenating the four videos.

Views



VT: Speaker

Results

- Ami Corpus meeting sequences are randomly divided into 5 groups, 4 are used to train the visual-based networks, and one group for testing.
- We call each groups CV.

	C3D		ResNet3D-18		ResNet3D-34	
Fold	RGB	RGB-OF	RGB	RGB-OF	RGB	RGB-OF
CV1	0.5	0.55	0.7	0.75	0.7	0.78
CV2	0.7	0.63	0.71	0.81	0.77	0.82
CV3	0.65	0.5	0.79	0.82	0.79	0.85
CV4	0.5	0.77	0.76	0.85	0.78	0.84
CV5	0.67	0.5	0.68	0.76	0.68	0.76
Mean	0.60	0.59	0.73	0.80	0.74	0.81

Table I. RESULTS OF MACRO AREA UNDER CURVE (AUC) FOR ALL FOLDS OF AMI CORPS USING THE 3D CNNS.

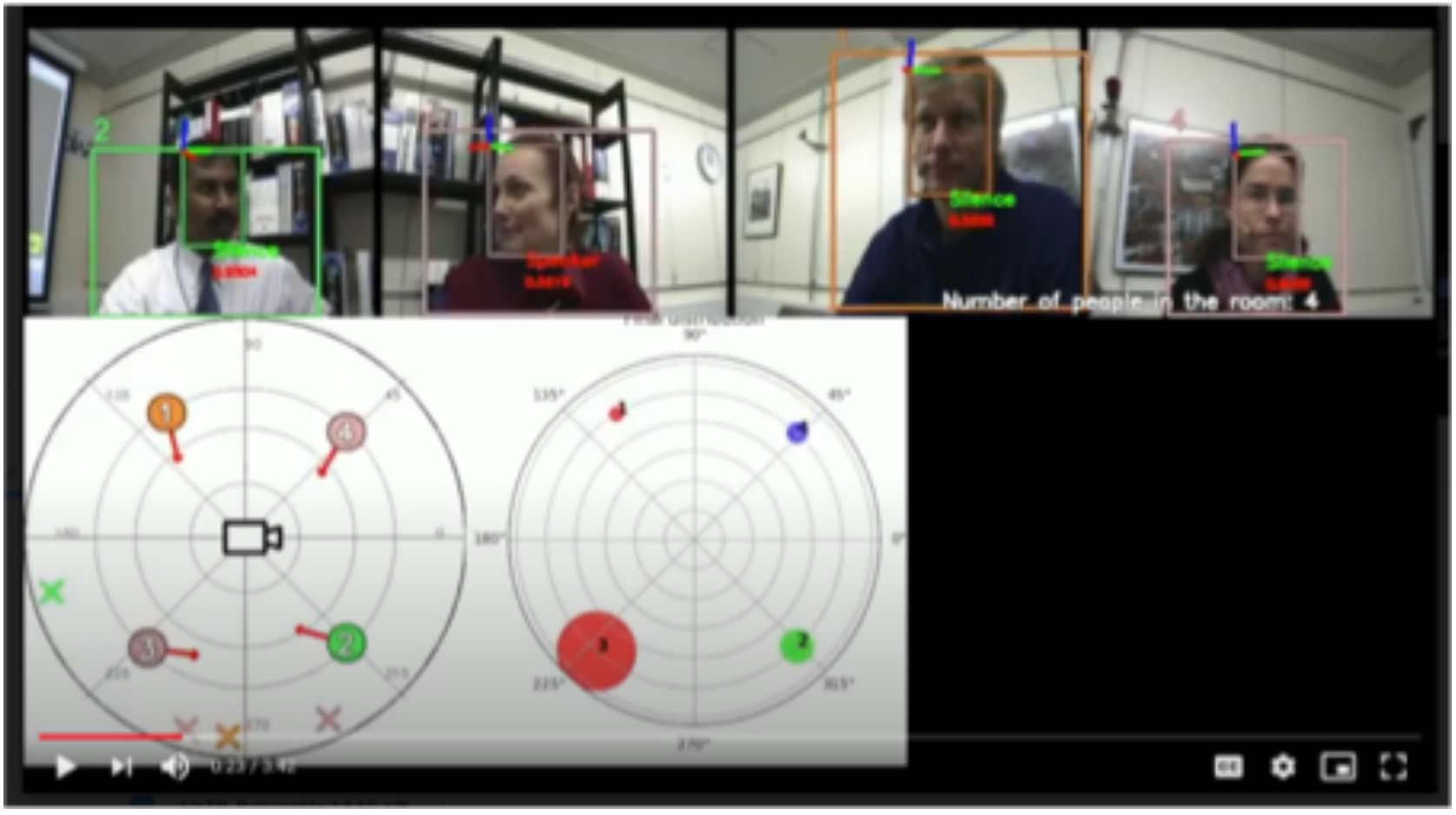
	C3D		ResNet3D-18		ResNet3D-34	
Fold	RGB	RGB-OF	RGB	RGB-OF	RGB	RGB-OF
CV1	0.48	0.55	0.69	0.75	0.7	0.78
CV2	0.69	0.63	0.71	0.8	0.77	0.82
CV3	0.64	0.5	0.76	0.82	0.79	0.85
CV4	0.5	0.73	0.76	0.84	0.77	0.84
CV5	0.64	0.5	0.68	0.75	0.68	0.76
Mean	0.59	0.59	0.72	0.79	0.74	0.81

Table I. RESULTS OF MICRO AREA UNDER CURVE (AUC) FOR ALL FOLDS OF AMI CORPS USING THE 3D CNNS.

Motivation

- Video

https://drive.google.com/file/d/11B-1f6EhvH3lPVV080KoY_ITdW483emZ/view?usp=sharing



Thank you!

See you at poster 801

This work was carried out in LAAS-CNRS.