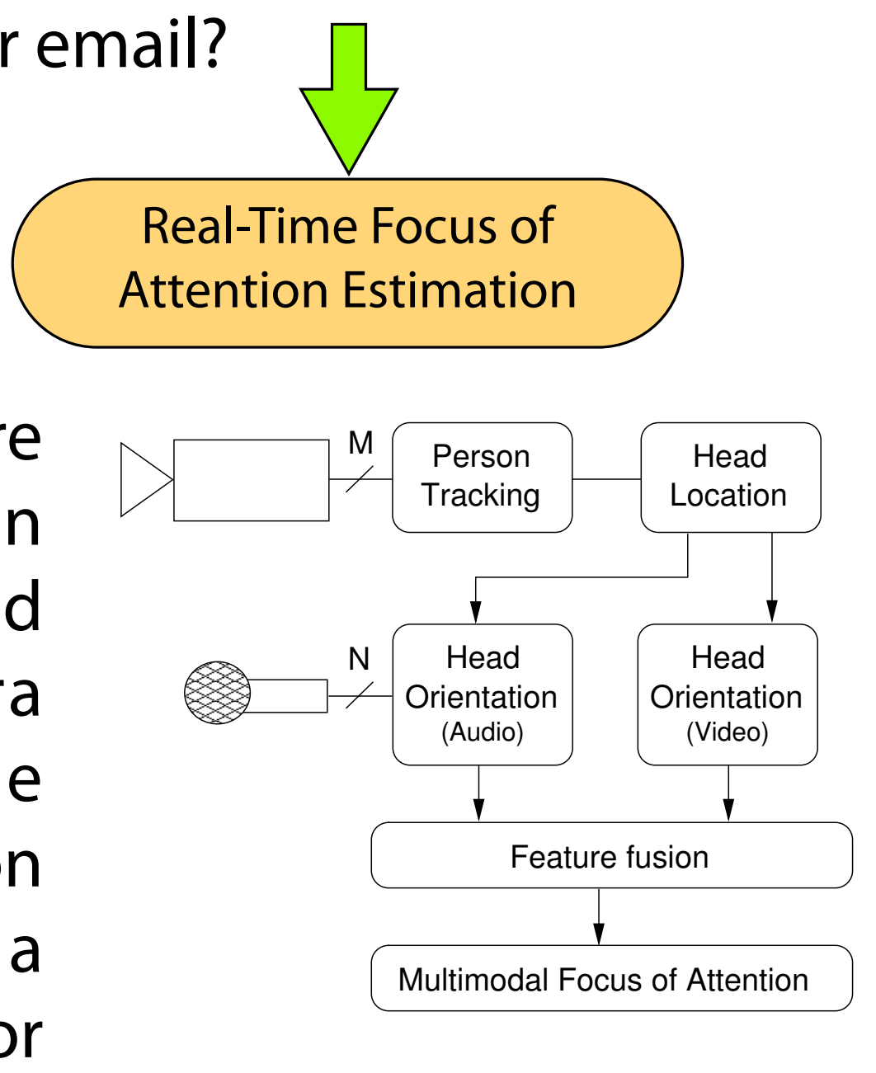
CVPR 2008
June 24 - 26
Anchorage Alaska

Multimodal Real-Time Focus of Attention Estimation in SmartRooms

C. Canton-Ferrer, C. Segura, M. Pardàs, J.R. Casas, J. Hernando
{ccanton, csegura, montse, josep, javier}@gps.tsc.upc.edu
Technical University of Catalonia, Barcelona - Spain

1. Objectives

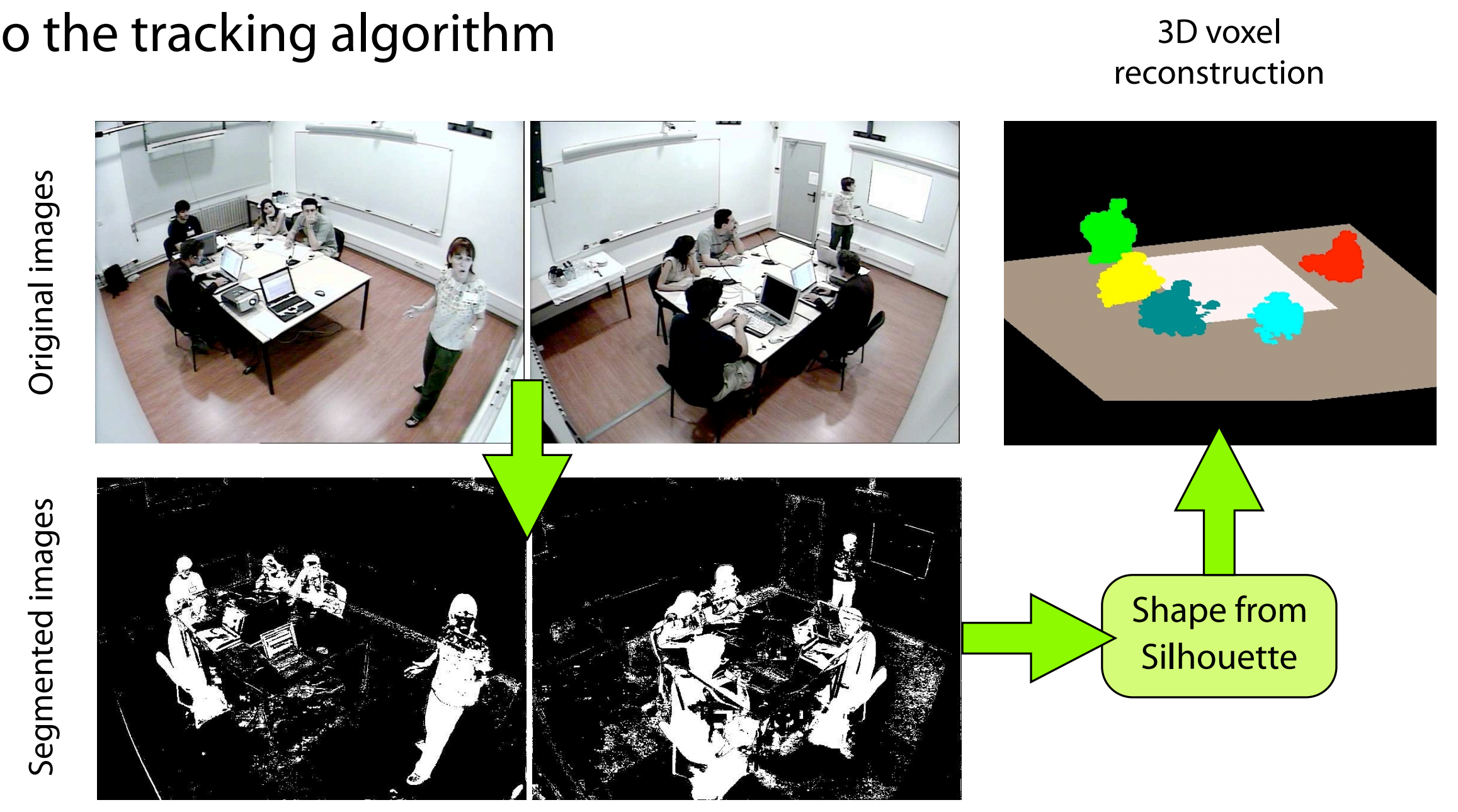
Question: Who is distracted looking through the window or checking his/her email?



- Combine two already mature technologies such as person tracking and multi-modal head orientation in a multi-camera environment towards real-time Focus of Attention (FoA) estimation in SmartRooms. Develop and test a distributed pipeline architecture for this task.

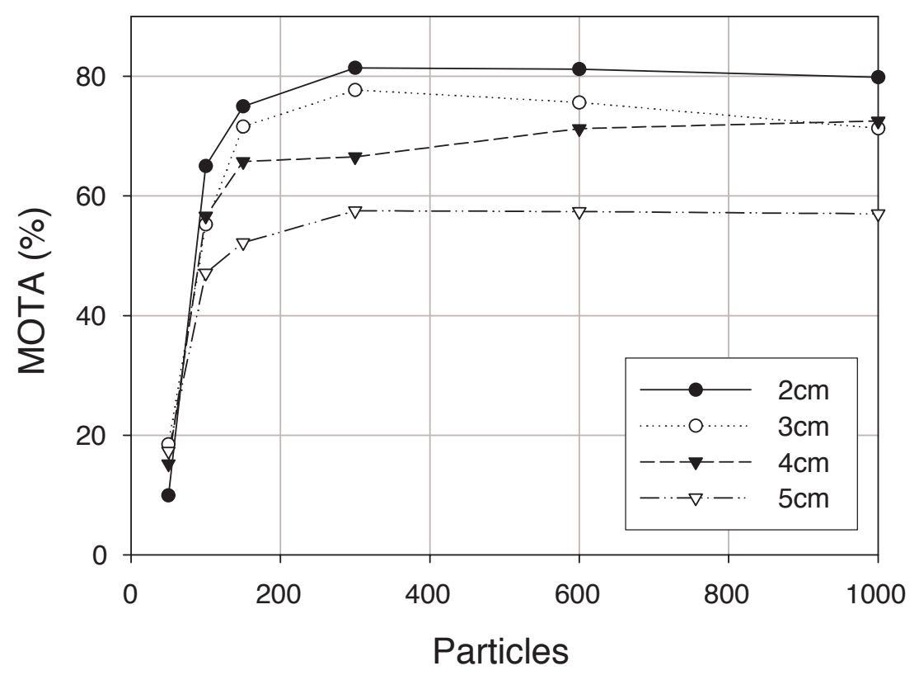
2. Multi-Camera Person Tracking

- Position of people inside the room provides information to locate the head of the targets, as well as contextual information.
- A 3D voxel based reconstruction of the space is the input data to the tracking algorithm

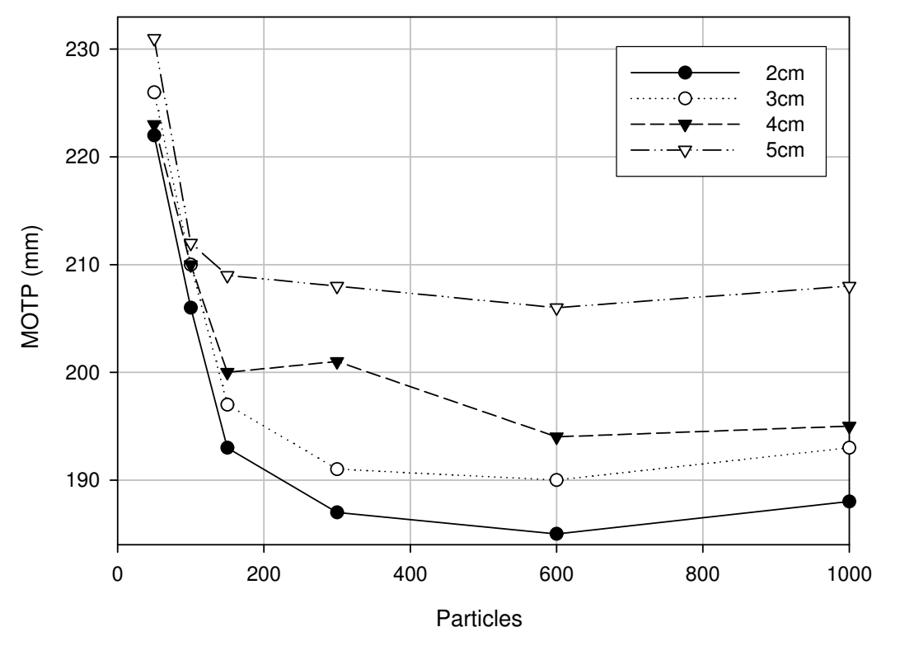


- Sparse Sampling algorithm is presented as an alternative to Particle Filtering (PF) in the context of 3D person tracking.
- A set of samples placed at the surface voxels allows estimating the centroid of the target. The position of samples evolve with time following the PF steps: re-sampling, propagation, evaluation and estimation.
- Likelihood of a sample belonging to the surface of the object attains its maximum value when a sample has the half of its neighbors empty and the half occupied.
- Blocking algorithm allows assigning a separate tracker for each target.
- Performance: Evaluation over the **CLEAR 2007** database (5 calibrated cameras, 25 fps, 768x576, ~3 hours of data)

MOTA: Tracking accuracy, i.e., fraction of time tracking all targets correctly



MOTP: Tracking precision, i.e., metric error in estimating target's position

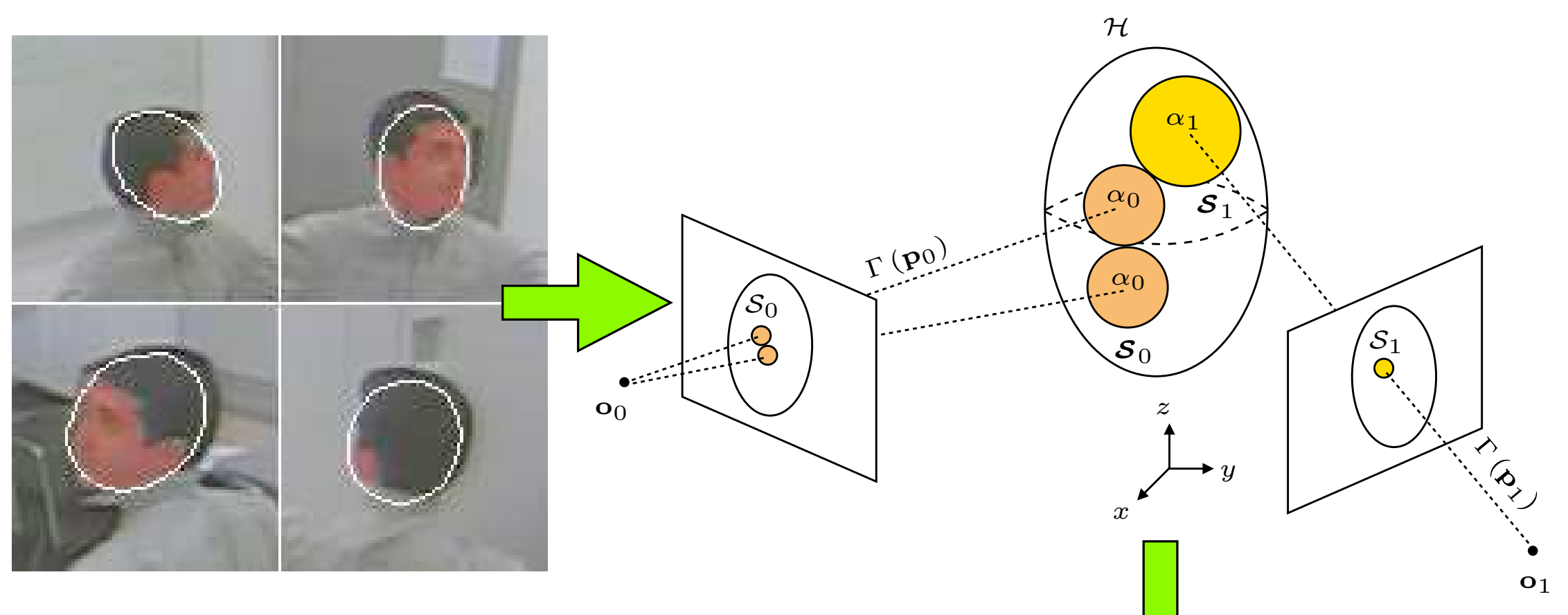


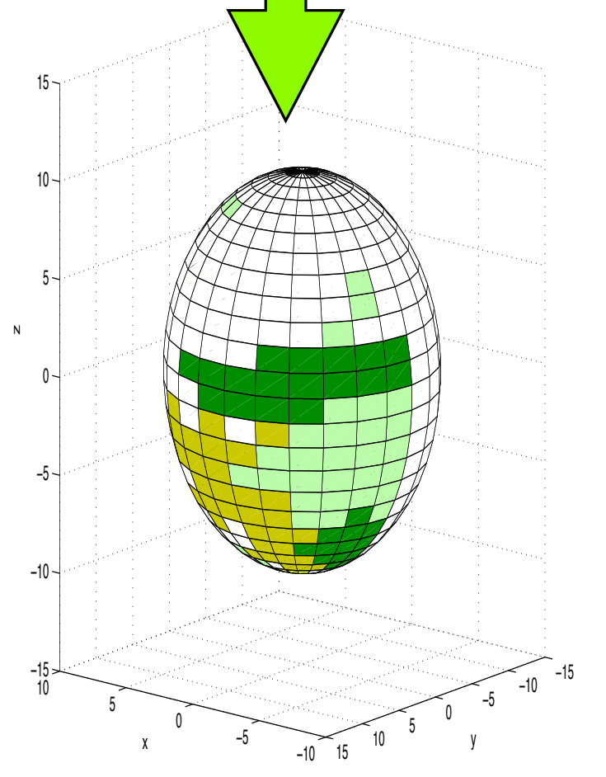
- Bibliography: A. López, C. Canton-Ferrer, J. R. Casas. **"Multi-Person 3D Tracking with Particle Filters on Voxels"**. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu (Hawaii, USA), April 2007.

3. Head Orientation Estimation

Multi-Camera Estimation

- Exploit spatial redundancy among multiple views of the same scene to generate a robust video based head orientation estimation.
- Position of people inside the room estimated from the tracking module together with 3D voxel reconstruction of the person allow estimating their head position. This information is used to extract skin colored patches in the head region in each camera.
- Calibration information is employed to back-project these skin patches onto an ellipsoid shaped estimation of the human head. In this way, a synthetic representation of the skin appearance on the head of the target can be extracted.

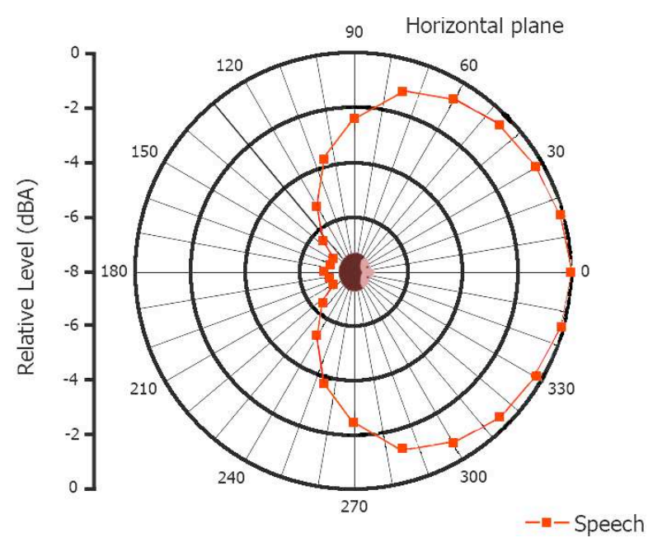




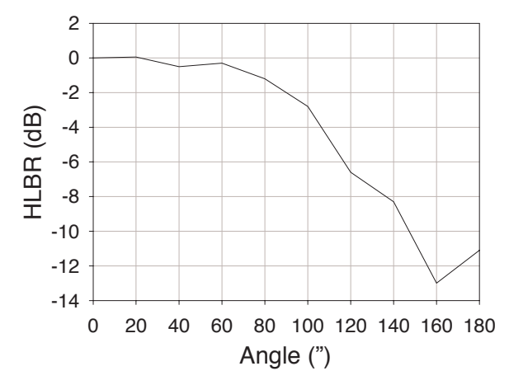
- Head orientation is computed as the weighted average of the angle obtained at each back-projected skin patch.

Multi-Microphone Estimation

- Human speakers do not radiate speech isotropically. The knowledge of human radiation pattern may be used to estimate head orientation from a set of 4 T-Shaped 4-channel microphone clusters.



- The High/Low Band Ration (HLBR) is defined as the ration between high and low bands of frequencies of the radiation pattern. This allows estimating the head orientation of an speaker. This normalization process allows comparison among different microphones circumventing bad calibration and propagation losses.



- The estimated speaker orientation can be computed by searching the angle that maximizes the correlation between the HLBR among microphones.

Multi-Modal Estimation

- A decentralized Kalman filter is used to combine information coming from the audio and video sources.
- Head orientation results from **CLEAR 2006** database:

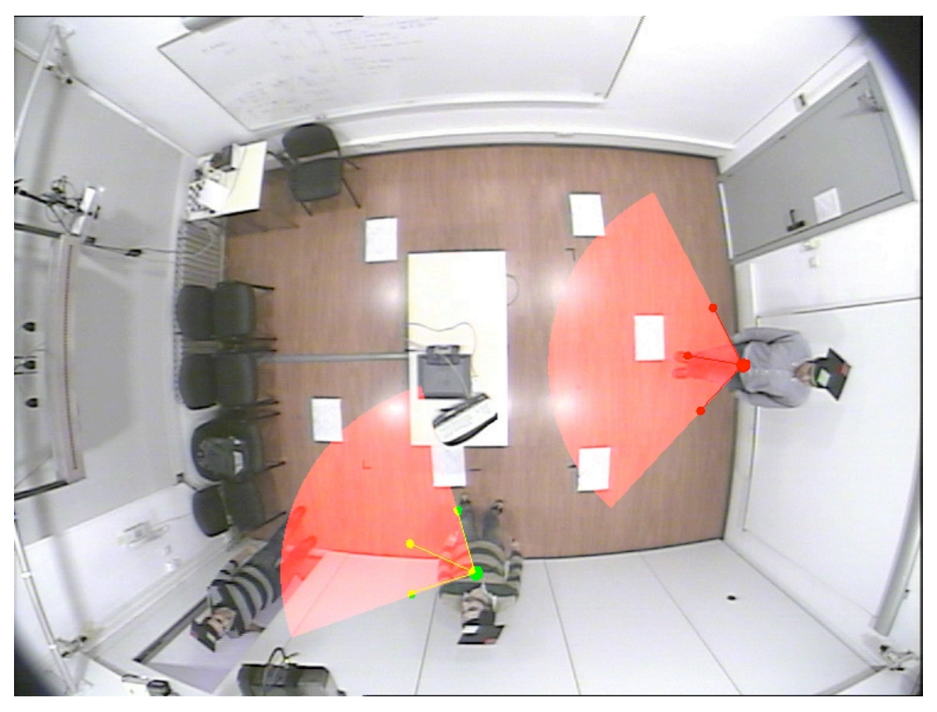
Method	PMAE (°)	PCC (%)	PCCR (%)
Video	57.23	32.88	71.39
Audio	53.14	28.47	69.17
Multimodal	48.53	38.19	73.47

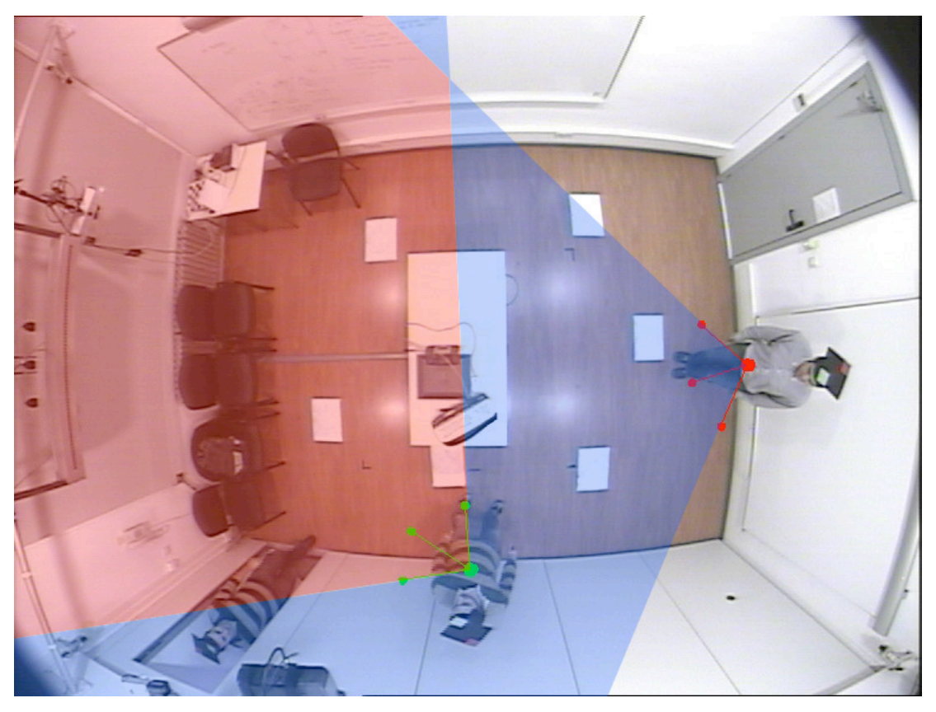
- Multimodality allows compensating errors occurred in separate modalities:

4. Focus of Attention Estimation

- The spatial region where the attention of a person is drawn is tightly correlated with the orientation of our head and the horizontal and vertical span of our eyes. In this paper, only attention in the *xy* plane is considered.
- In order to keep computational complexity low, focus of attention is addressed through two easily computable descriptors:

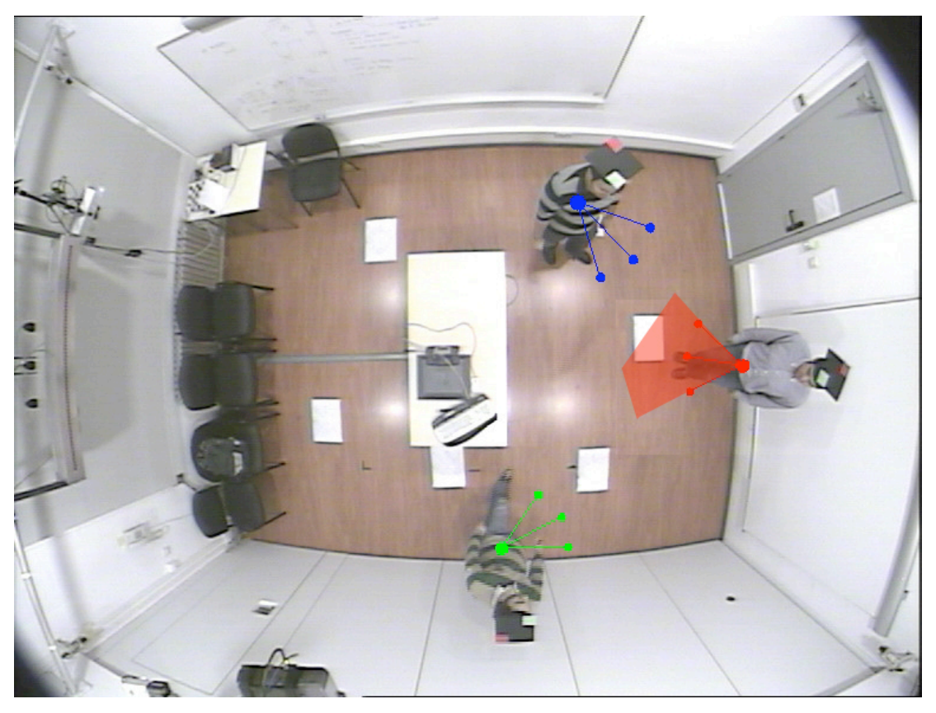
Attention cone: the cone generated in the center of the head with an opening angle *d=30º*.





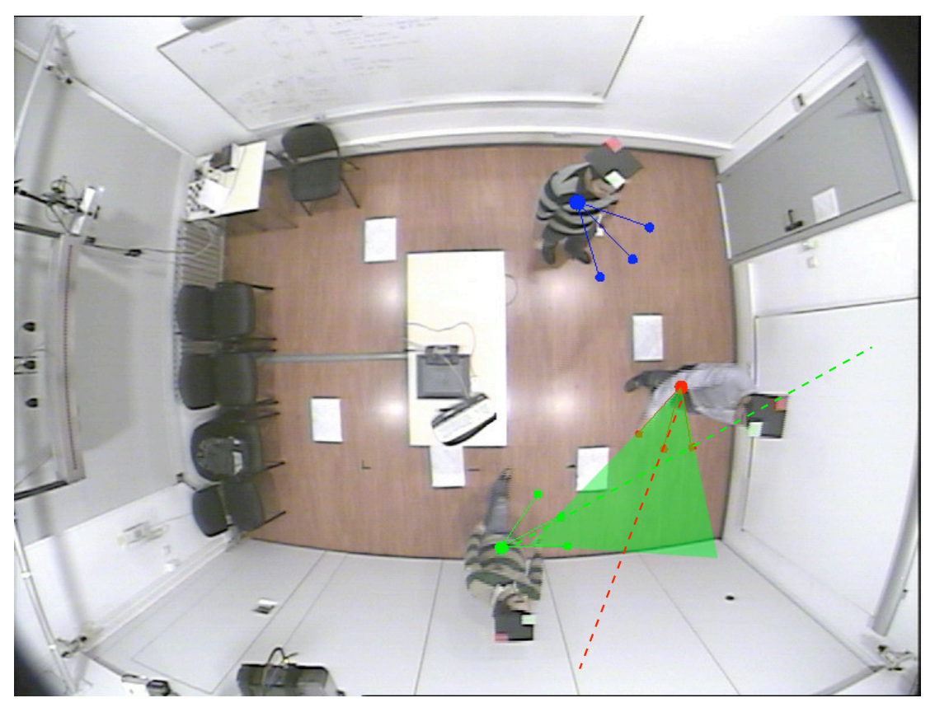
Attention map: the cumulative intersection of all the attention cones in the *z*-plane.

- Although being two simple descriptors they capture the underlying information regarding group attention.
- Two particular cases are can be detected:



Region of interest detection: analysis of the intersections of attention cones allow locating *hot areas* of interest. For example, detecting the meeting center of attention.

Interaction detection: Two highly overlapped attention cones and orientation vectors close to the origin of these cones denote interaction between two participants.



- Results obtained over an annotated database yielded to an 85% of correctly detected events.
- Near real-time performance (6-15 fps) is attained with a distributed processing system consisting in 5 off-the-shelf machines with a 2.2GHz processor.

5. Conclusions

- This paper presented a multi-person focus of attention tracking system that combine two technologies: person tracking and head orientation estimation.
- Real-time operation is obtained by applying some considerations such as the sparse sampling technique employed in person tracking.
- Future research include:
  - Defining more sophisticate automatic human behavior analysis techniques
  - Combination of the proposed output with audio event detection and recognition algorithms towards a multimodal scene analysis system
  - Further validation of the proposed algorithm on larger databases