AN ADAPTIVE VIDEO SURVEILLANCE ARCHITECTURE FOR BEHAVIOR UNDERSTANDING

Zini L., Noceti N., Odone F.

DISI, Università di Genova – Via Dodecaneso 35,16146 Genova, ITALY

{zini,noceti,odone}@disi.unige.it

Adaptivity to scene changes is a main requirement for video analysis. The interpretation of video streams can be dealt by triggering different techniques depending on the scene properties. We design a video surveillance architecture where different tasks in the context of behavior analysis are addressed, depending on the crowd level. An estimation of the scene occupancy allows us to focus on single person or groups, adopting appropriate strategies to model the dynamic information.

Final goal

Designing an **adaptive** video surveillance architecture for behavior understanding

IDEA: the adopted techniques should be **dynamically** influenced by the **scene complexity**

Motivations

- The diffusion of video surveillance cameras caused an increasing interest in applications where analyzing and interpreting videos are central topics.
- Videos are highly complex data due to variations over time of illumination, layout properties, **scene occupancy**

$\Downarrow$

- Adaptive methods are required

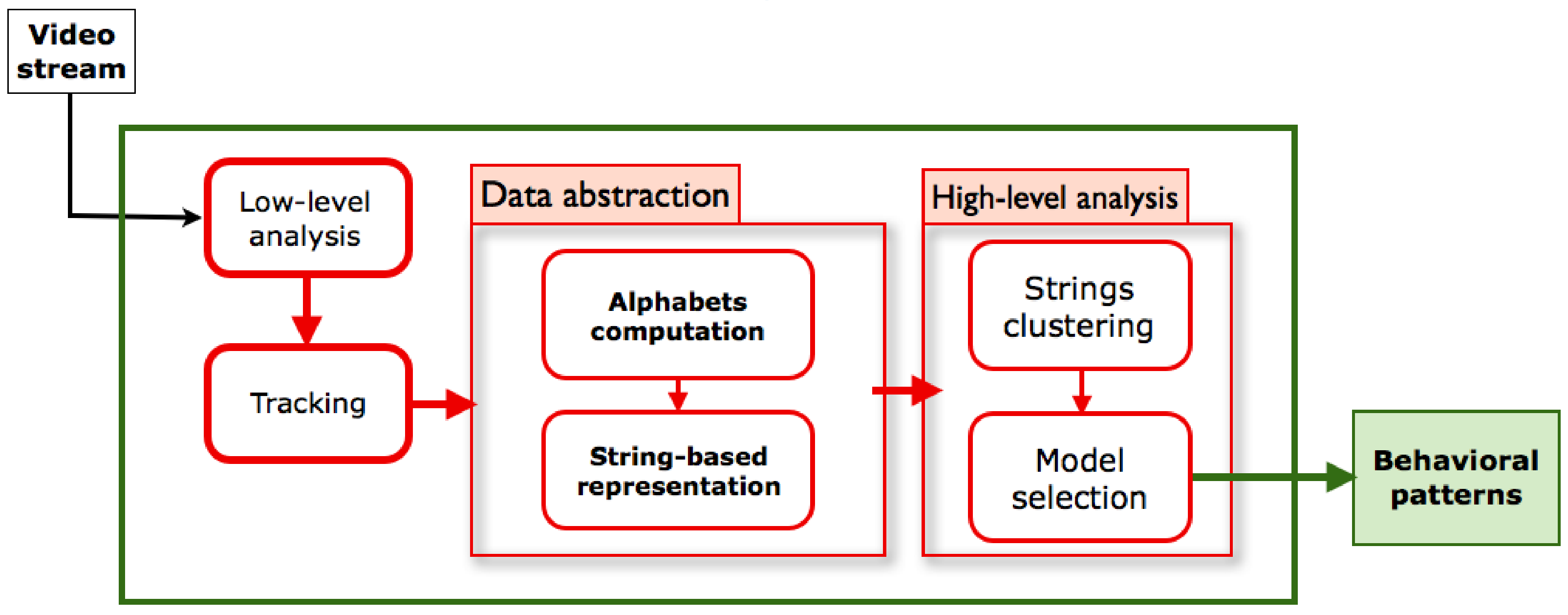
Tasks

- **Low occupancy:** study of the dynamic of single person or small groups
- **Crowded scenes:** estimation of the global motion of the crowd

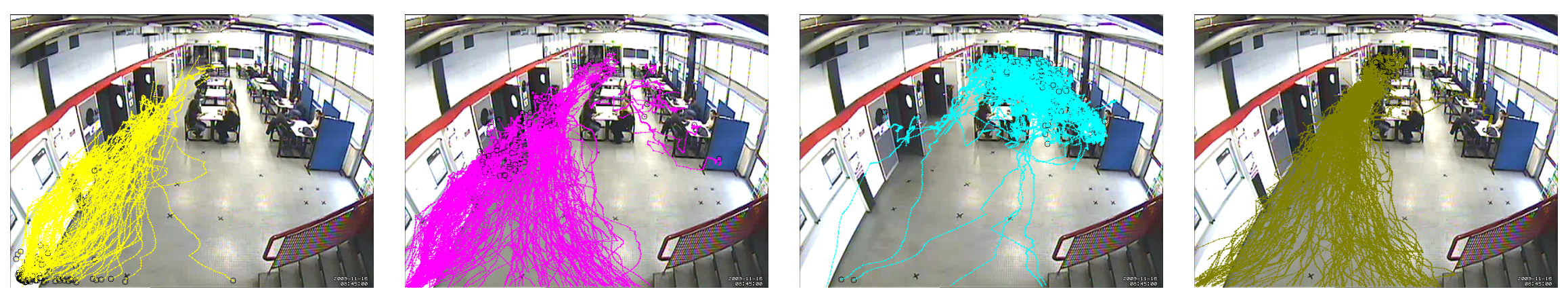
Aim: Integrating into an existing video surveillance pipeline for people behavior understanding [2] a module to cope with highly crowded scenes

How to trigger different techniques?

The existing architecture



Examples of estimated behavioral patterns



Experimental evaluations

- Two weeks of observations
- #Training set =1200 dynamic events
- #Test set = 5700 dynamic events
- Correct events association = 76,2%

What if the number of people increases too much? In this case the tracking fails to compute reiliable descriptions of the scene dynamic

A module for people counting

Camera calibration

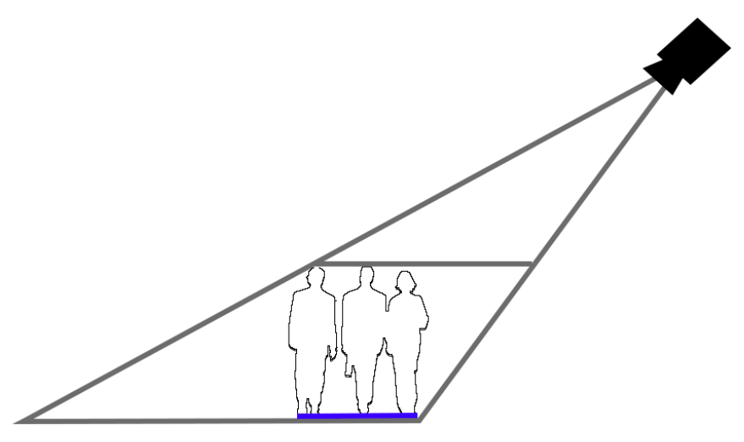
If a full camera calibration is not available we may simply estimate [3]:

- Ground plane Π_g homography $P = H_{ground}P_{ground}$
- Head plane Π_h homography $P = H_{head}P_{head}$

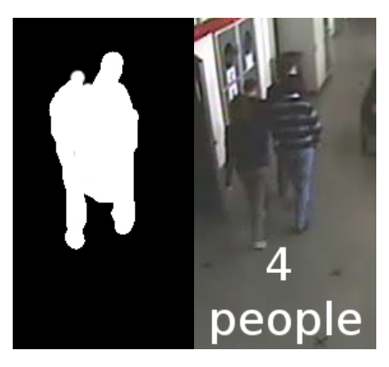
P is in pixel coordinates, P_{head}, P_{ground} are in world coordinates

Idea

- The area occupied by each person in the scene is the intersection of his/her projection onto the two fixed planes Π_g , Π_h



- In the case of groups the area will be proportional to the number of people in the group
- In our work the confidence in the estimate depends on the geometry: blobs of similar size may correspond to different people configurations

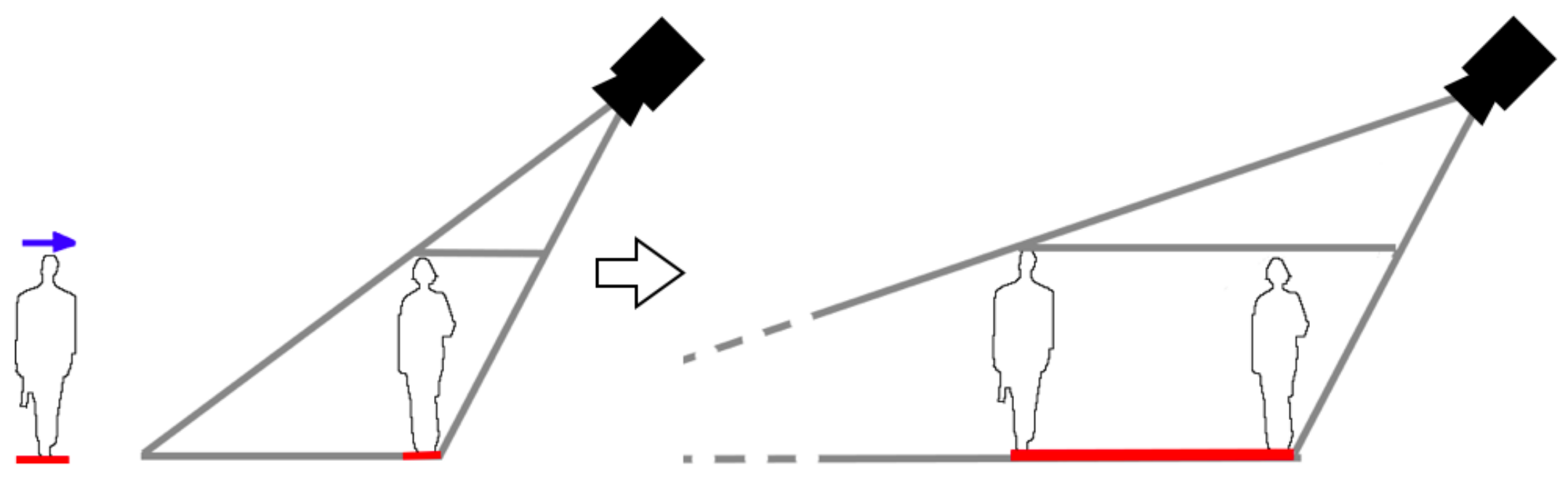


Algorithmic details

Coarse analysis as in [1]

Real time refinements based on temporal coherence:

- Make the people size uniform with respect to the position in the scene
- Deal with the occlusions



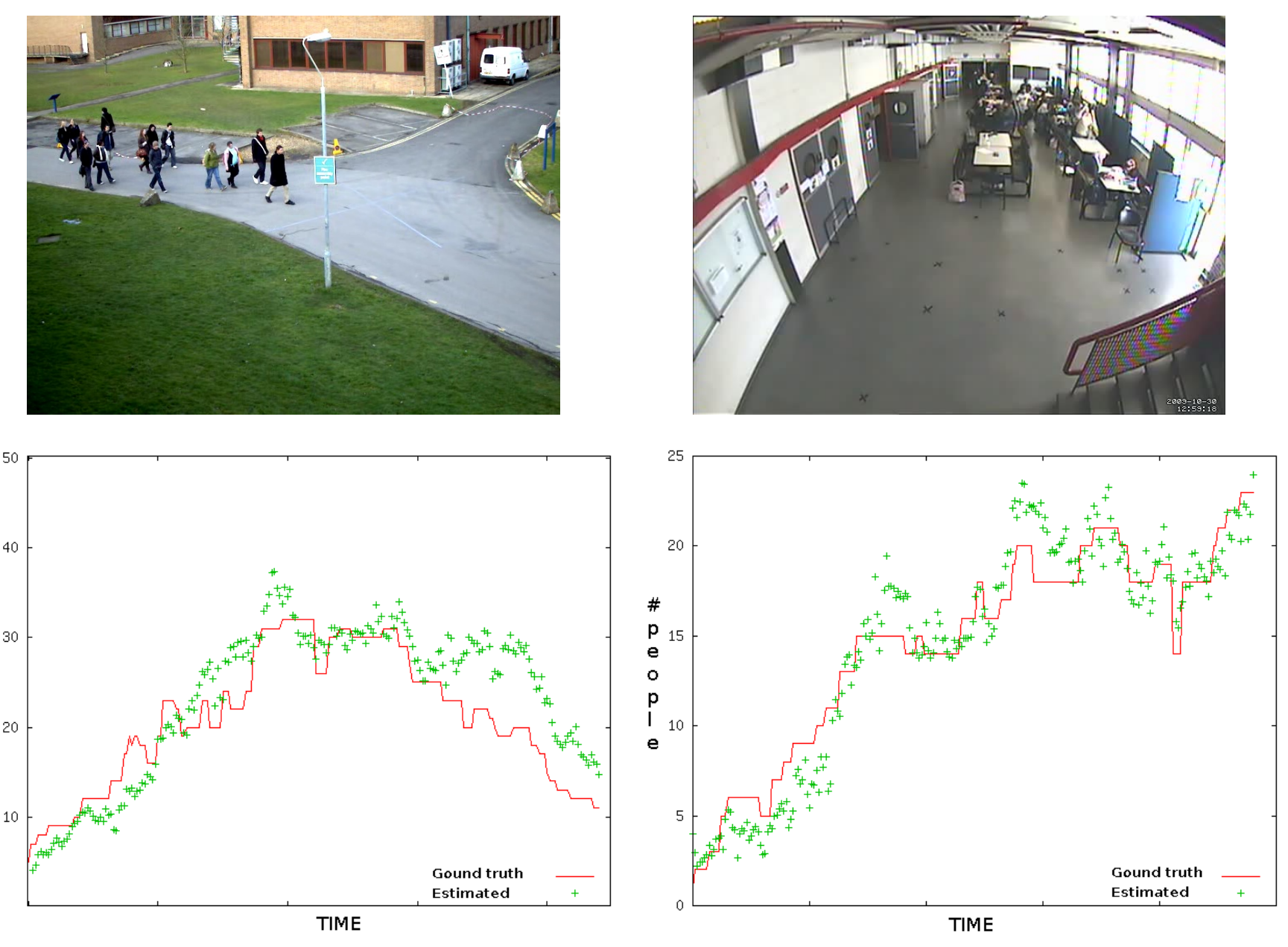
- Adjustments to make the method more robust to changes of the camera position

Results

Dataset: PETS 2009, DISI

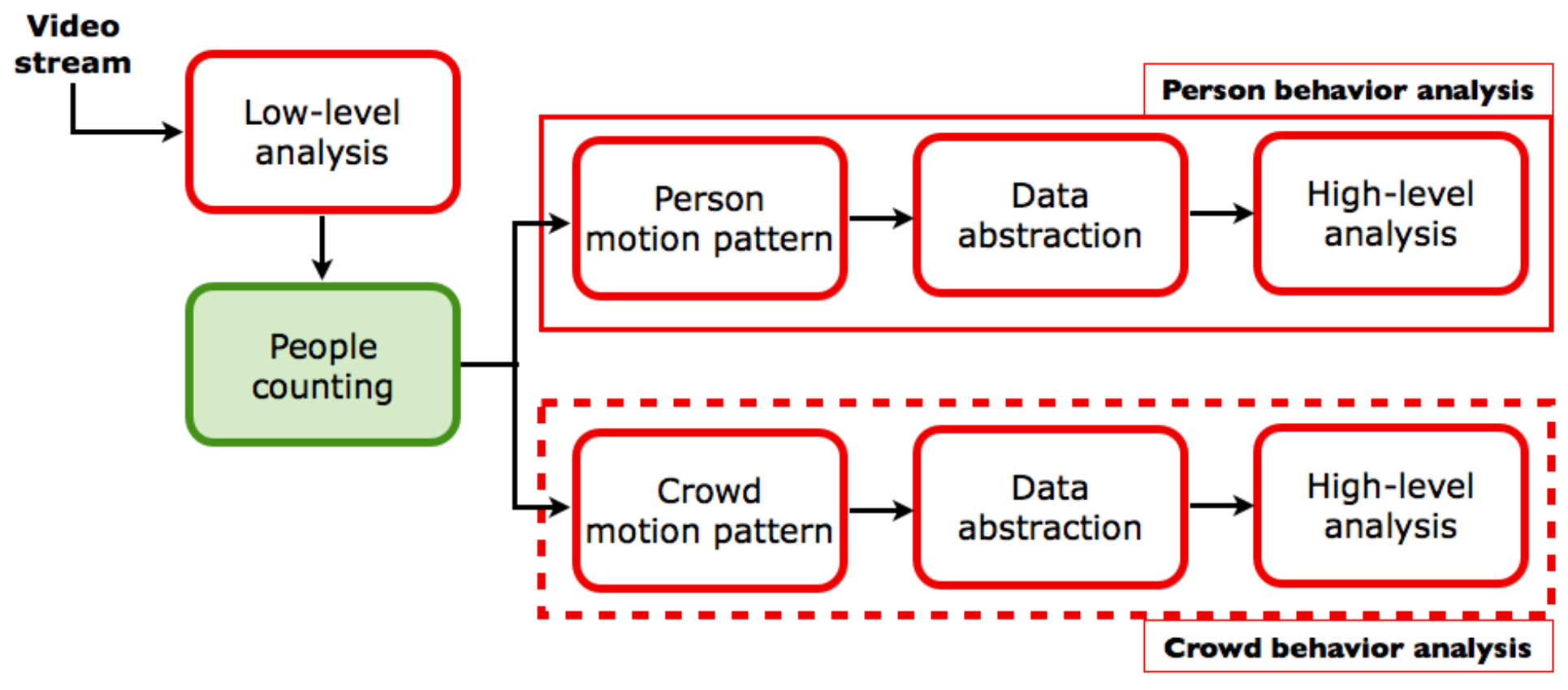
Accuracy

- #people$\leq$ 8: 97%
- $8 \leq$ #people ≤ 25: 85%
- #people ≥ 25: 98%



The plan of my Phd

The people counting module has been integrated into the existing architecture. It allows us to trigger different techniques depending on scene occupation level



The study and development of the dotted box content will be the focus of my work

[1] Kilambi,Ribnick,Joshi, Masoud and Panikolopoulos. Estimating pedestrian counts in groups. Computer Vision and Image Understanding, 2008

[2] Noceti, Learning to classify visual dynamic cues. University of Genoa. Phd Thesis 2010

[3] Hartley, Zisserman, Multiple view computer vision. Cambridge University Press. 2003