SIPL SIGNAL & IMAGE PROCESSING LAB

Department of Electrical Engineering, Technion

Multi-Camera People Tracking and Re-Identification within a Crowd

Students : Ofer Geva & Matan Sela
Supervisor : Yair Moshe
In Association with: Mango DSP

Mango DSP
INTELLIGENT VIDEO SERVERS

Abstract

- Tracking systems are necessary for surveillance, traffic control and multiple computer vision applications

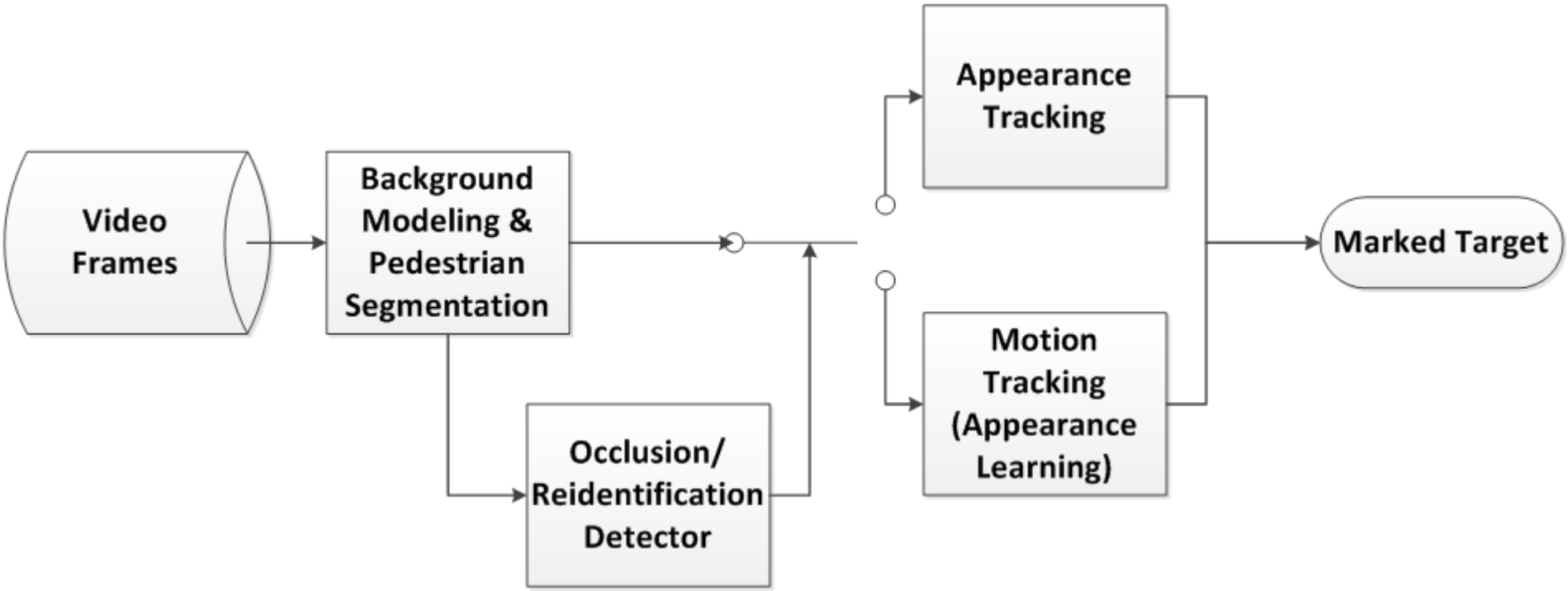
Problem Definition

Sparse Crowd

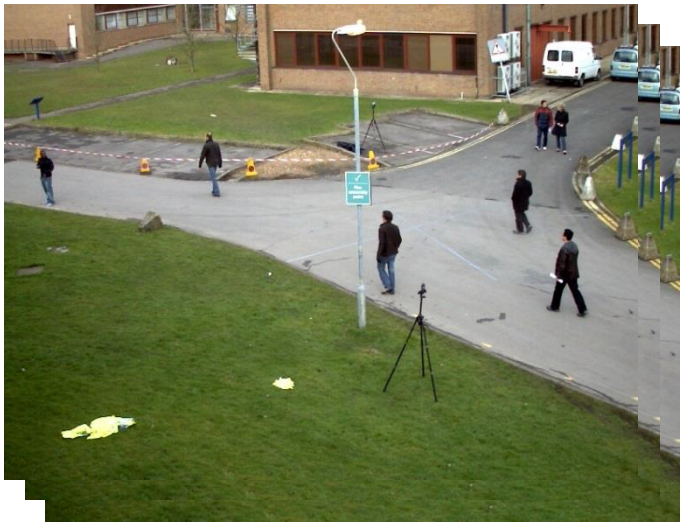
Medium Density Crowd

Dense Crowd

Part A - Single Camera Tracking Scheme



Background Modeling and Pedestrian Segmentation



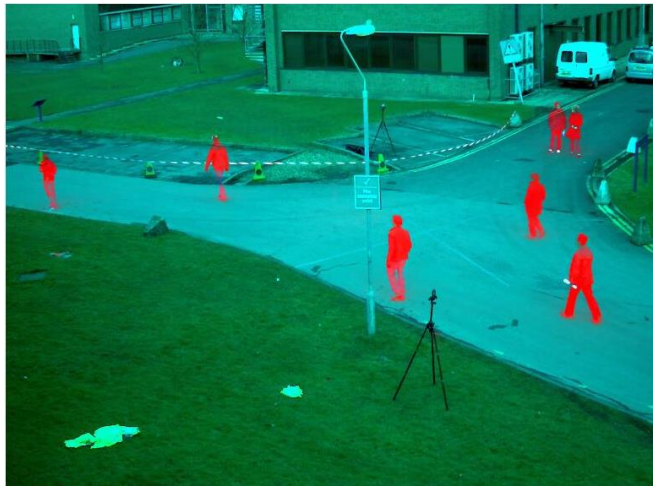
Frames

Median

Last Frame

HoG Detector

$C_1 \times$



Motion Foreground

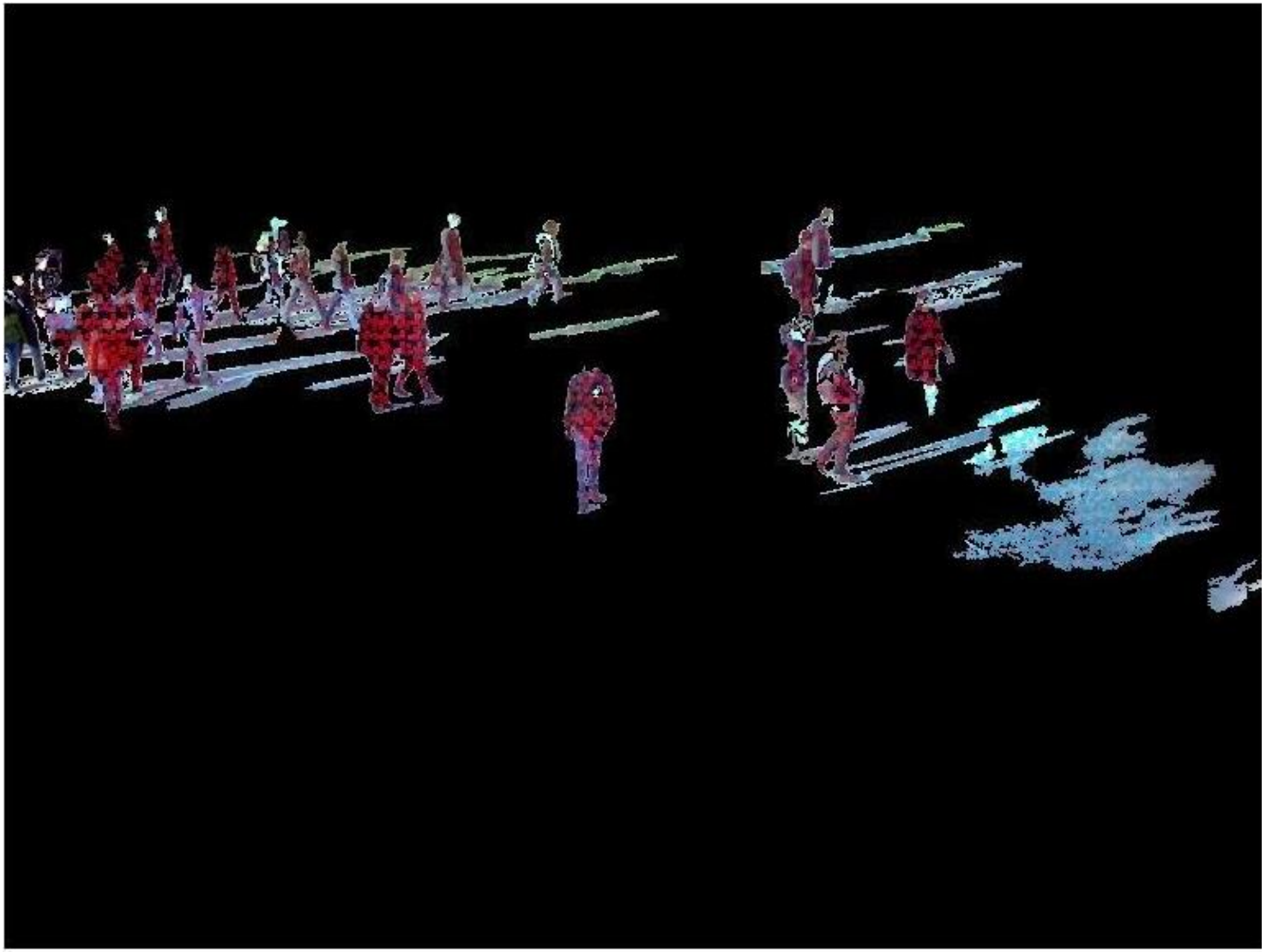
$C_2 \times$

HoG Confidence Function

Background Modeling and Pedestrian Segmentation

Improved Foreground Image

Background Modeling and Pedestrian Segmentation



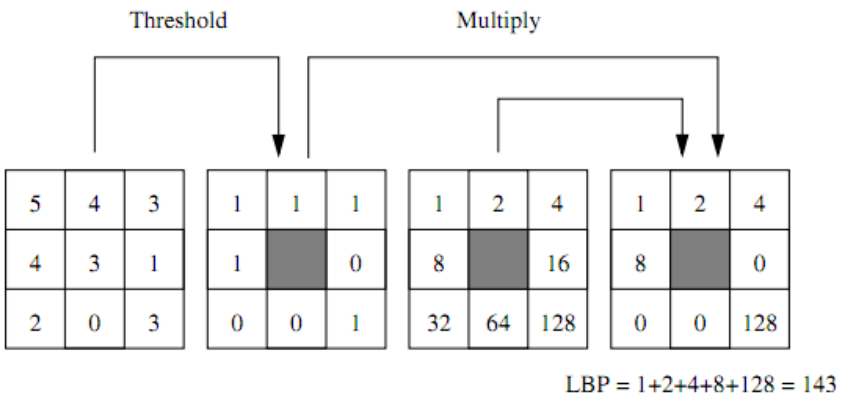
Foreground Detection in scenes with massive moving background and shades

Single Camera Tracking
Motion Mode – Kalman Filter

Tracker Data – Appearance Model

- **Sampling the target uniformly**

- **Extracting features from samples**
  - Color features
  - Texture features
  - Spatial features



Tracker Data – Appearance Model

- **Multi-Variable Kernel Density Estimation** – Estimating probability density function in the feature space:

$$\hat{p}(\underline{z}) = \frac{1}{N_p \sigma_1 \ldots \sigma_n} \sum_{i=1}^{N_p} \prod_{j=1}^{n} \kappa\left(\frac{z_j - s_{ij}}{\sigma_j}\right)$$

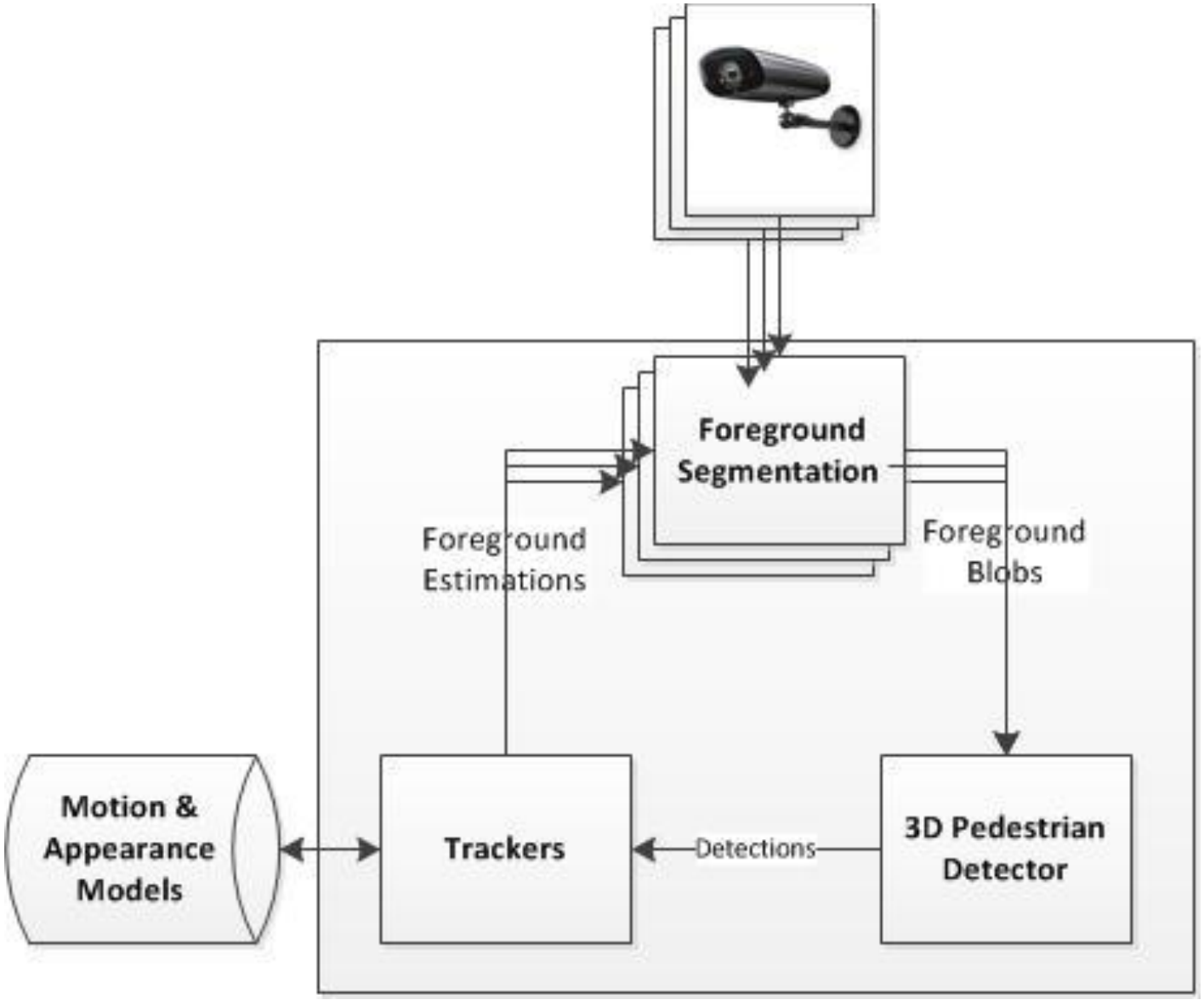
- **KL Distance** – Calculating the similarity between a new detection and the model :

$$D_{KL}\left(\hat{p}_b \mid \hat{p}_a\right) = \int \hat{p}_b(\underline{z}) \log \frac{\hat{p}_b(\underline{z})}{\hat{p}_a(\underline{z})} d\underline{z}$$

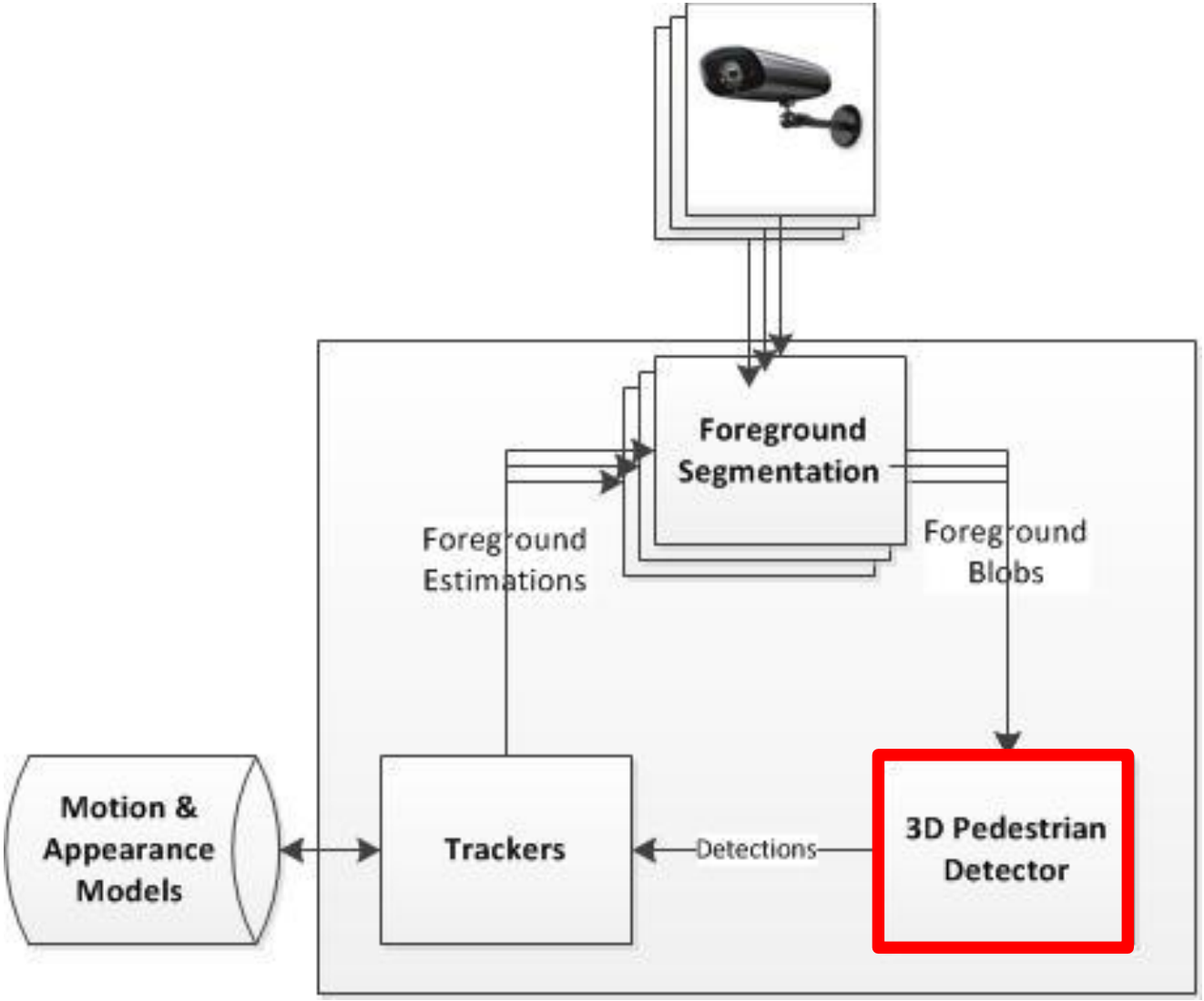
Part A - Single Camera
Appearance Mode

Part A - Tracking Demonstration

Part B - Multiple Cameras Tracking Scheme



Part B - Multiple Cameras Tracking Scheme



3D Pedestrian Detector

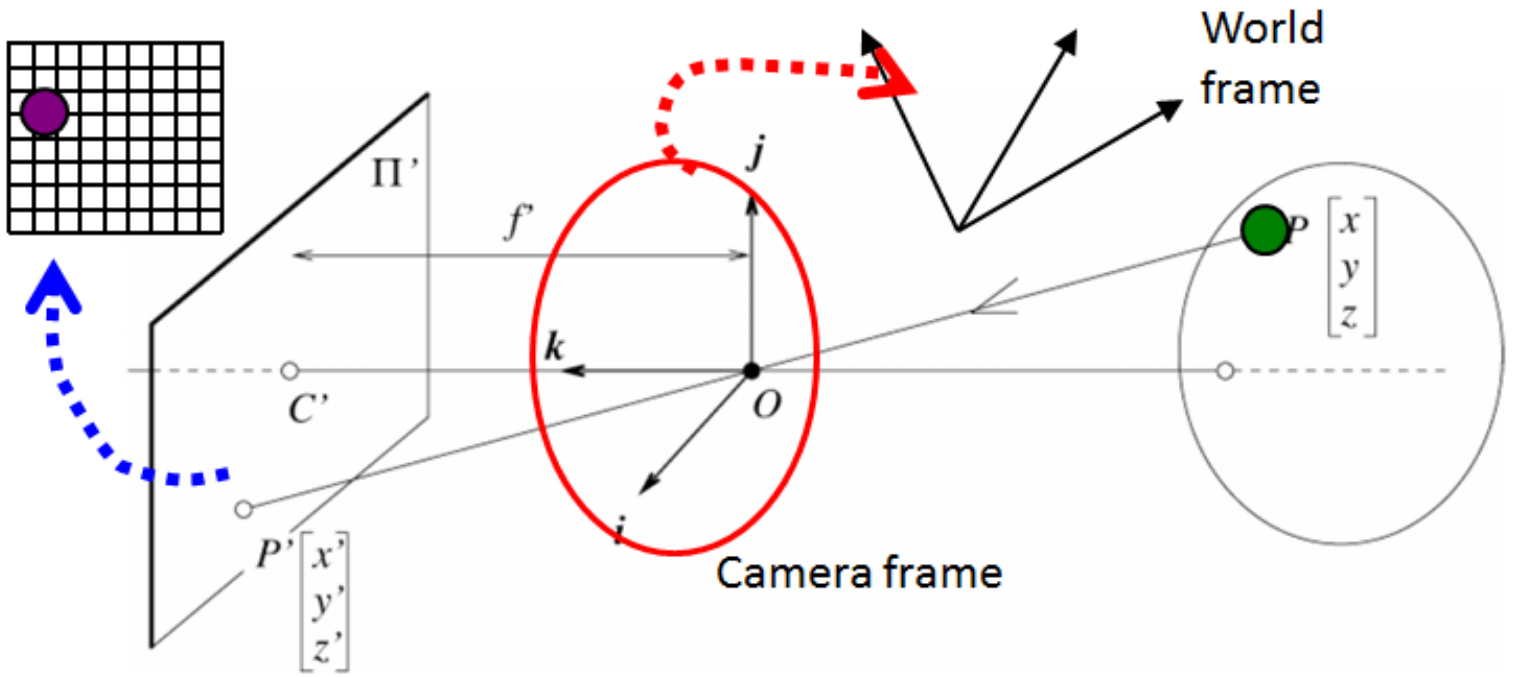
- Blobs Analysis :
  - In each view, **image coordinates** of each blob's lowest point are converted to **world coordinates**
  - blobs' lowest point is assumed to be in $Z = 0$ plane

Image 2D-Coordinates to Top-View World 3D-Coordinates

3D Pedestrian Detector

Camera Projection

- **Goal:** Estimating the 3D geometry of the scene from the 2D images



3D Pedestrian Detector

Camera Projection

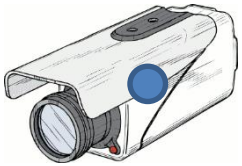
$$\begin{bmatrix} X_{im} \\ Y_{im} \\ 1 \end{bmatrix} = K \, [R \mid t] \times \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$$

Intrinsic:
Focal lengths
Sensor axes skew
Optical center

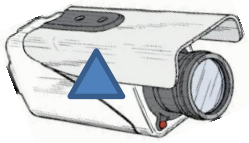
Extrinsic:
Rotation matrix
Translation vector

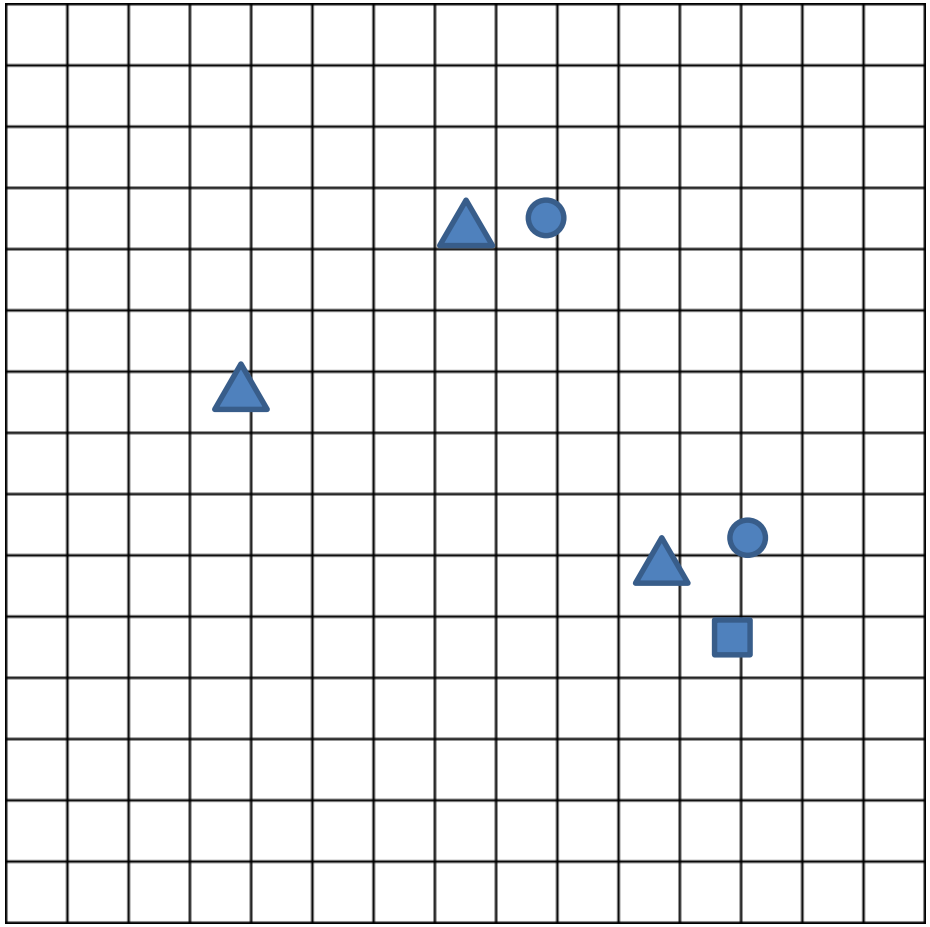
3D Pedestrian Detector

View 2

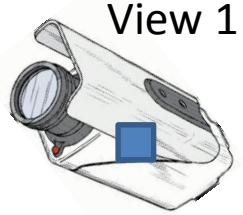


View 3



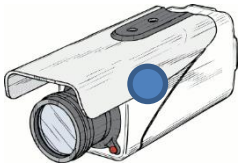


View 1

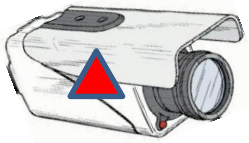


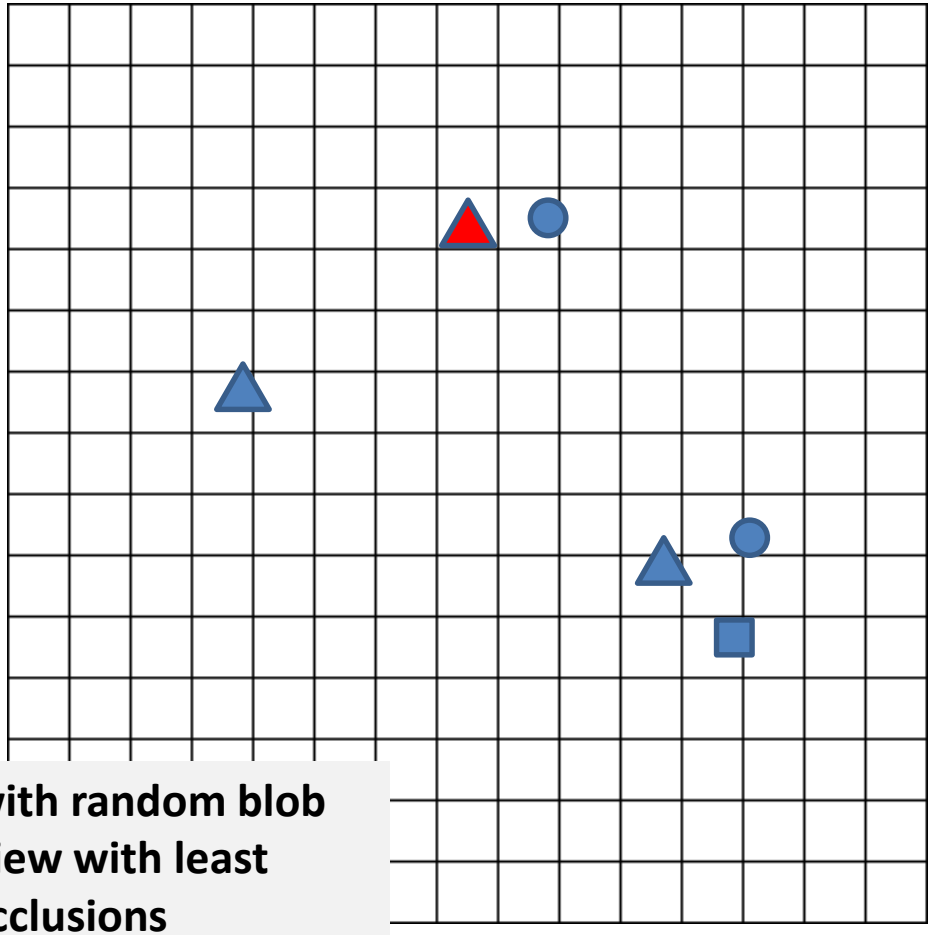
3D Pedestrian Detector

View 2



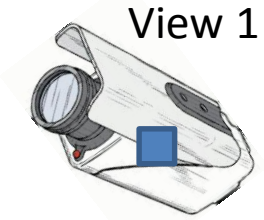
View 3





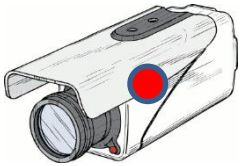
Starting with random blob from view with least occlusions

View 1

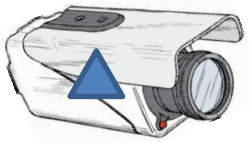


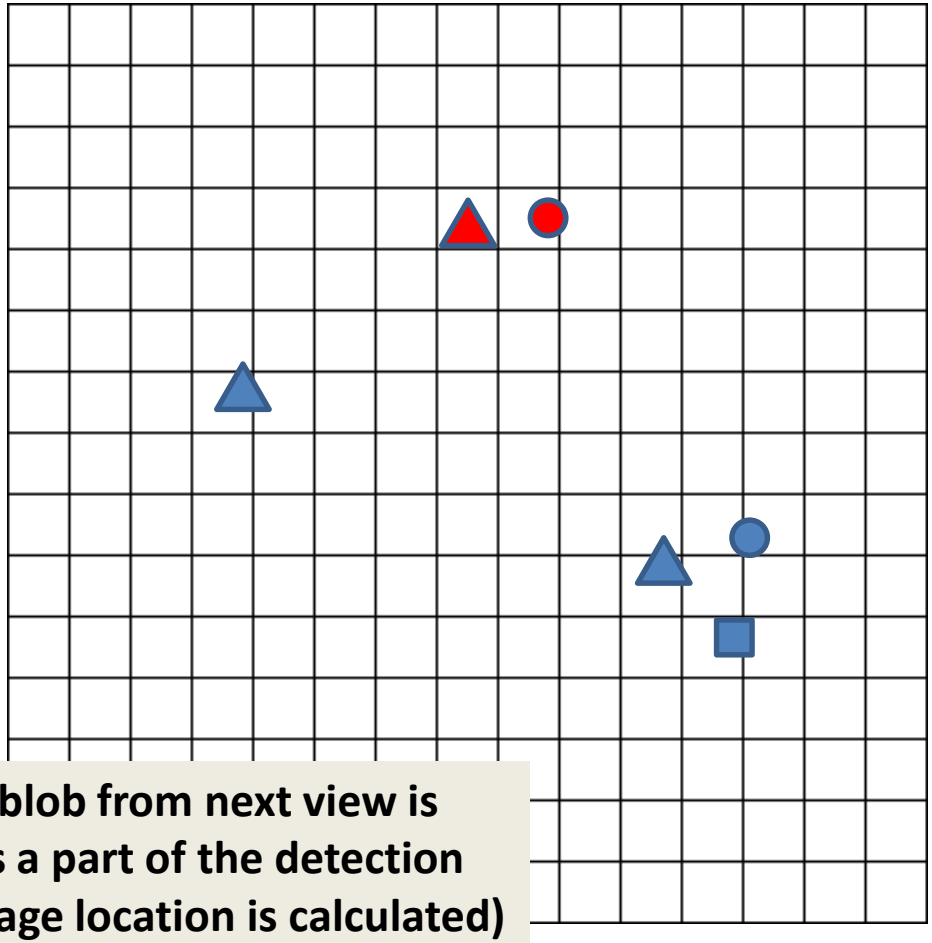
3D Pedestrian Detector

View 2



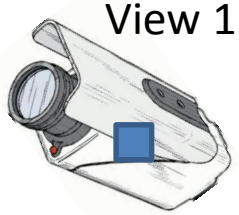
View 3





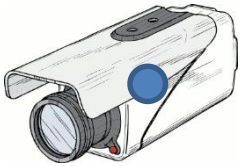
Nearest blob from next view is labeled as a part of the detection (group average location is calculated)

View 1

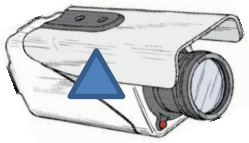


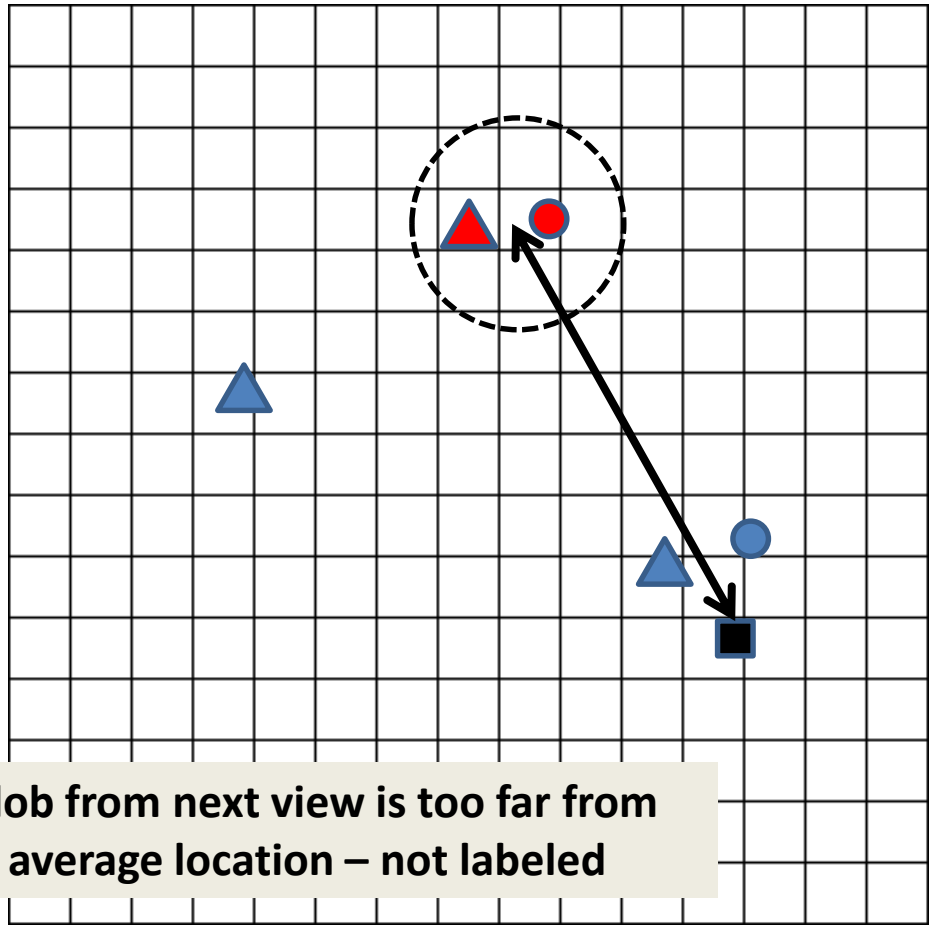
3D Pedestrian Detector

View 2



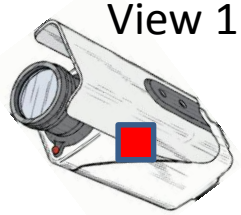
View 3





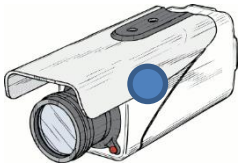
Nearest blob from next view is too far from current average location – not labeled

View 1

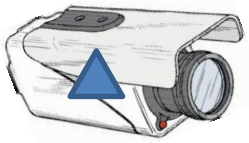


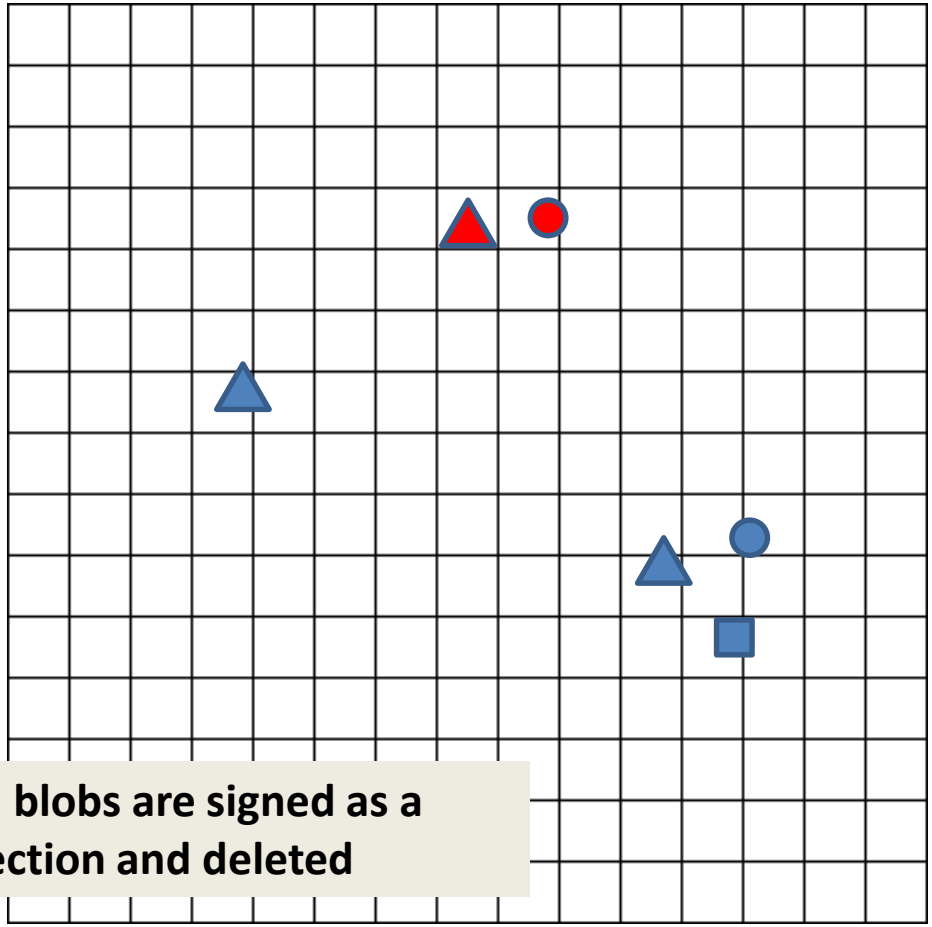
3D Pedestrian Detector

View 2



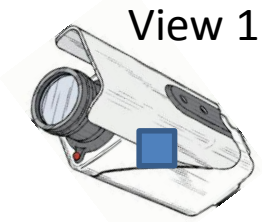
View 3





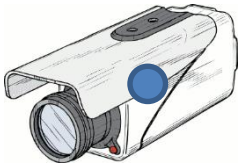
Labeled blobs are signed as a detection and deleted

View 1

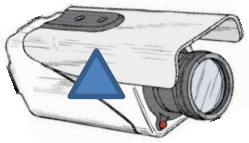


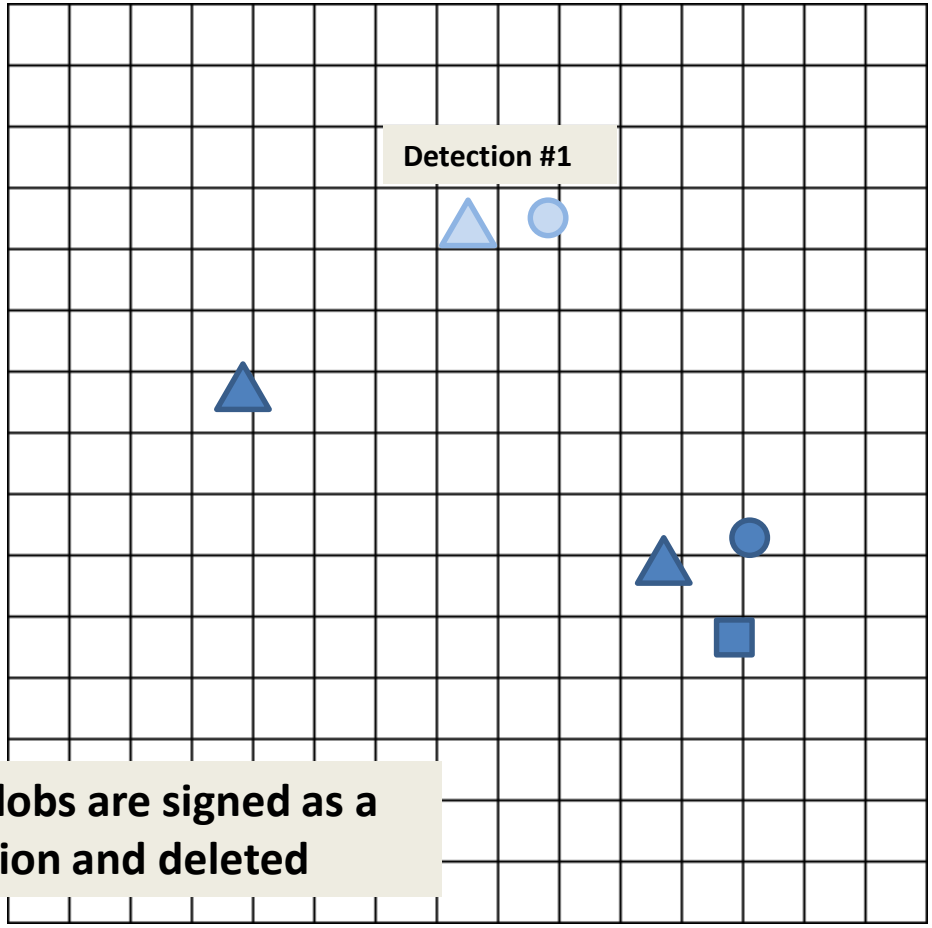
3D Pedestrian Detector

View 2



View 3



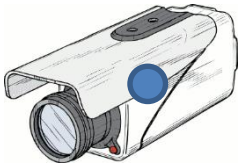


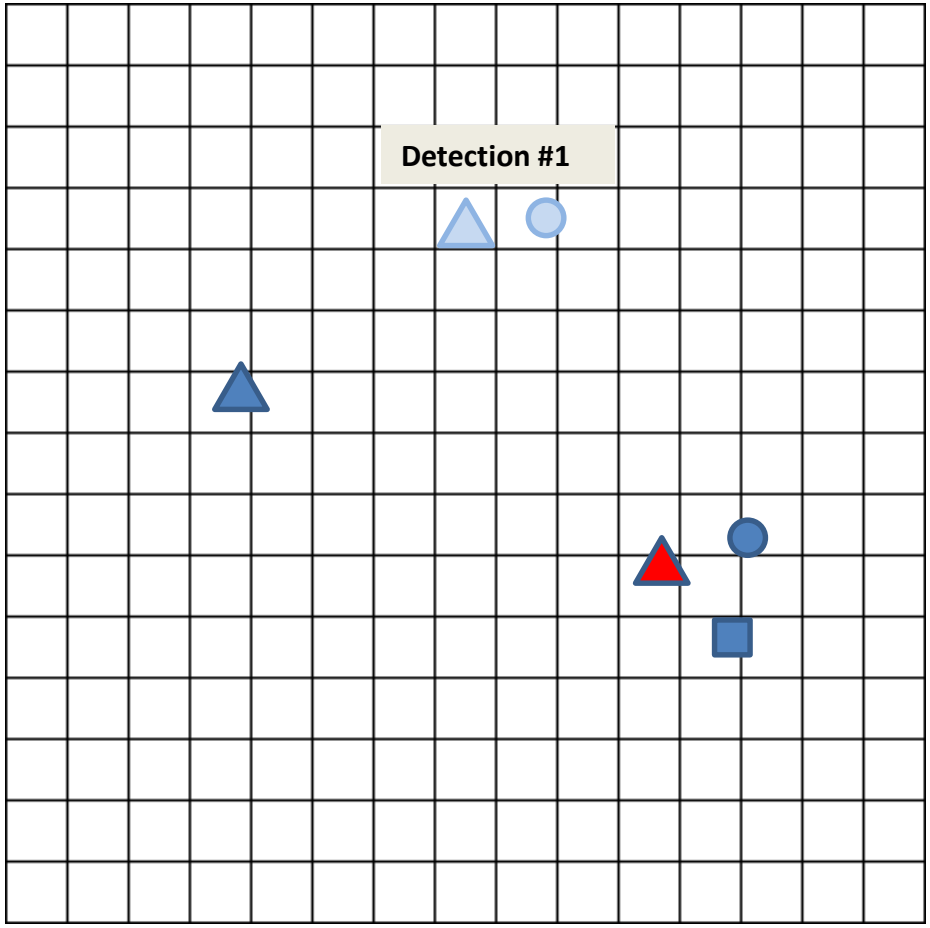
Labeled blobs are signed as a detection and deleted

View 1

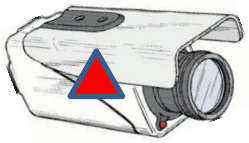
3D Pedestrian Detector

View 2

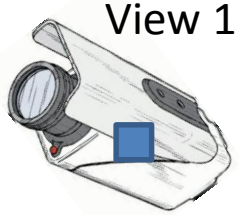




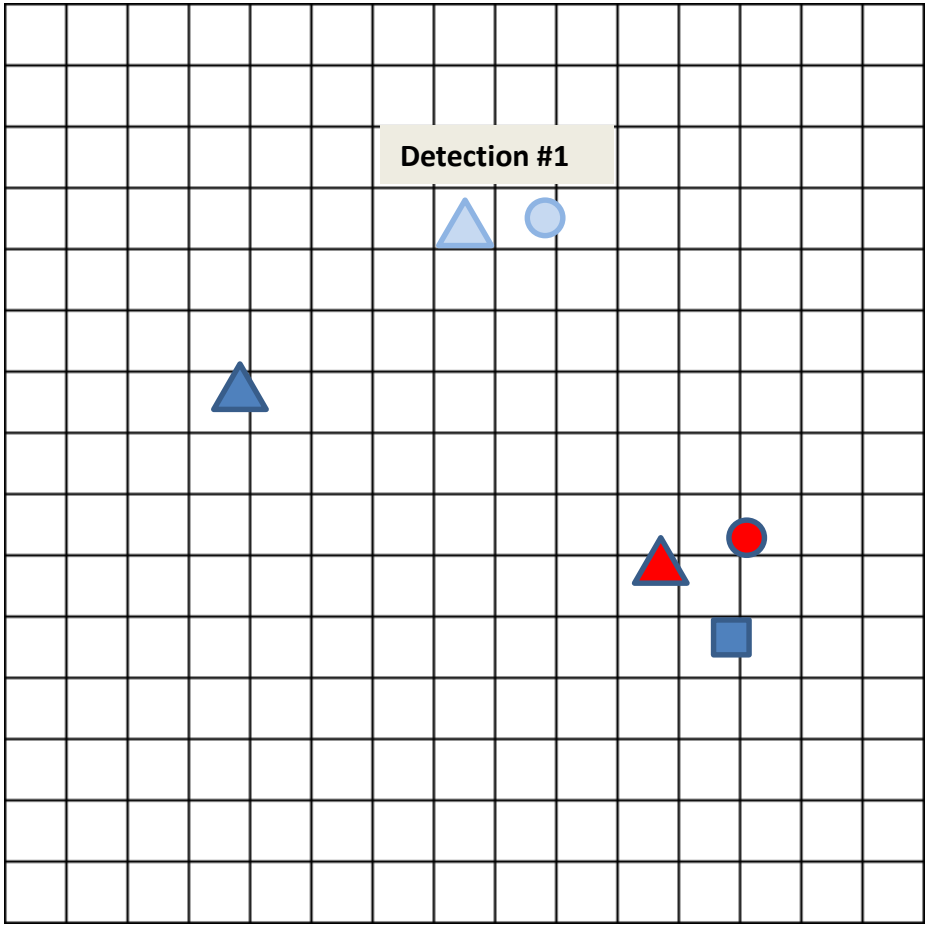
View 3



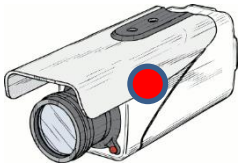
View 1



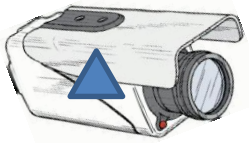
3D Pedestrian Detector



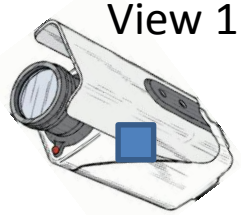
View 2



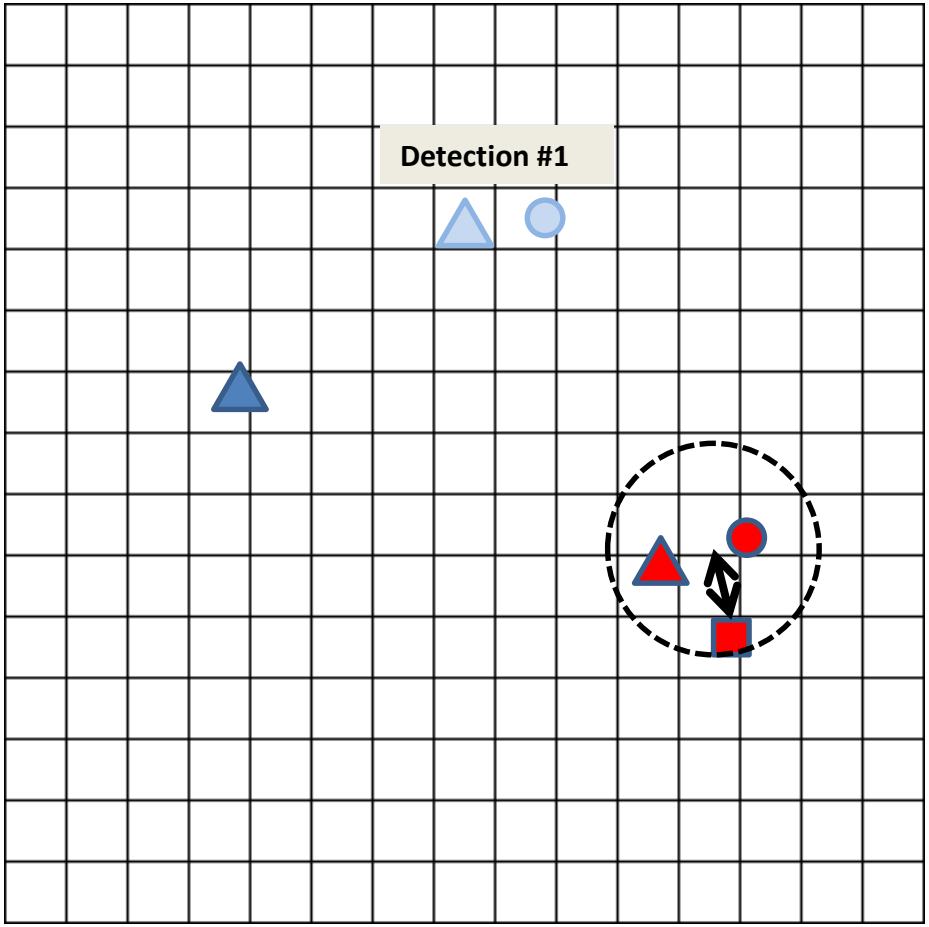
View 3



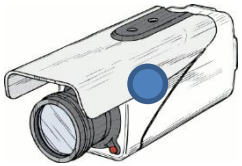
View 1



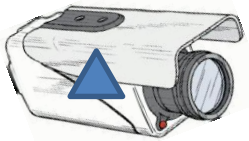
3D Pedestrian Detector



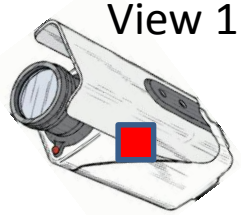
View 2



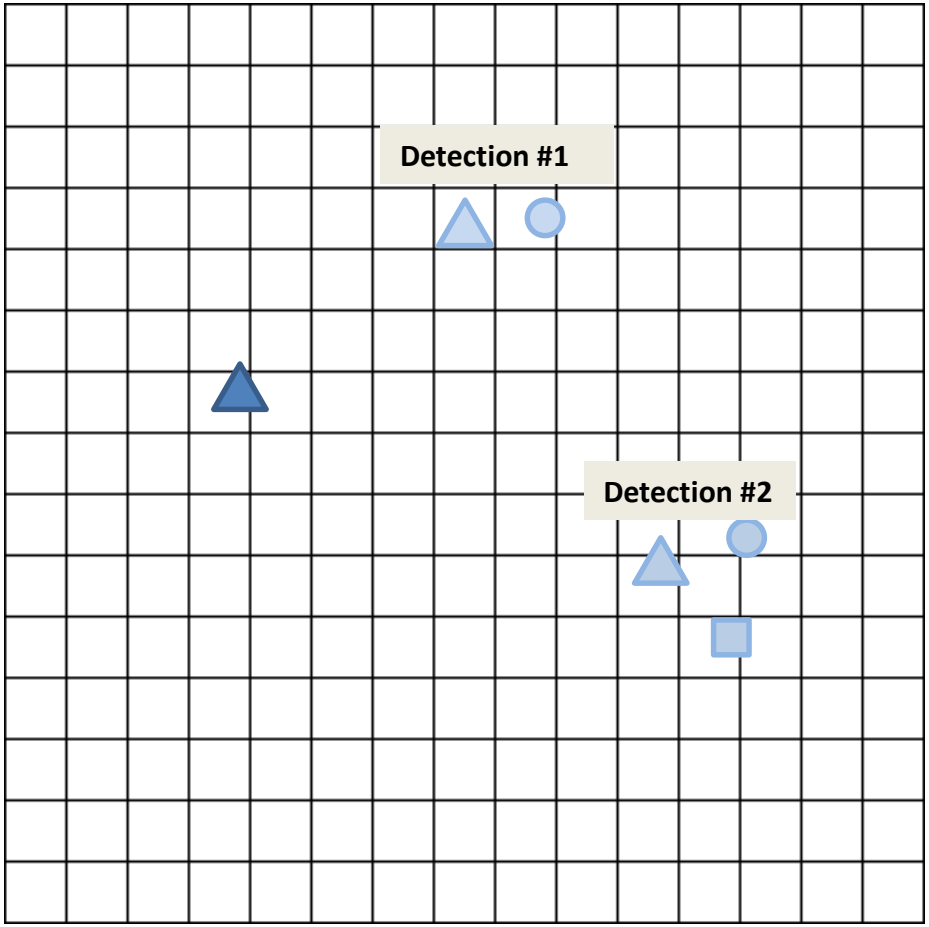
View 3



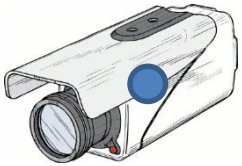
View 1



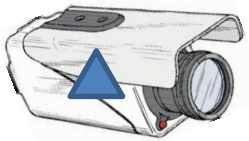
3D Pedestrian Detector



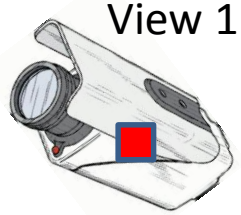
View 2



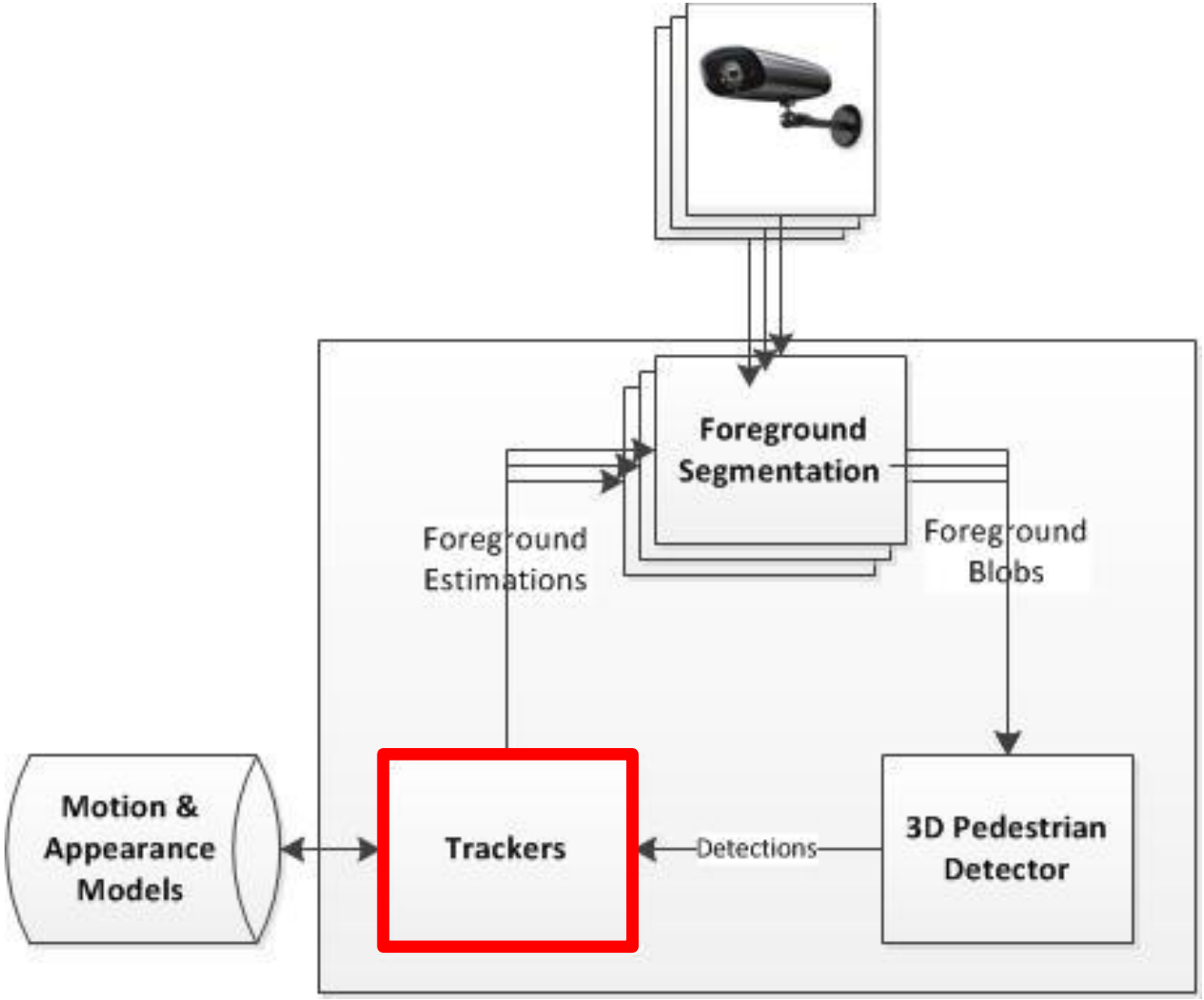
View 3



View 1



Part B - Multiple Cameras Tracking Scheme



Greedy Data Association

Example:

	Tracker 1	Tracker 2	Tracker 3
Detection 1	0.70	0.20	0.60
Detection 2	0.65	0.80	0.75
Detection 3	0.15	0.70	0.55

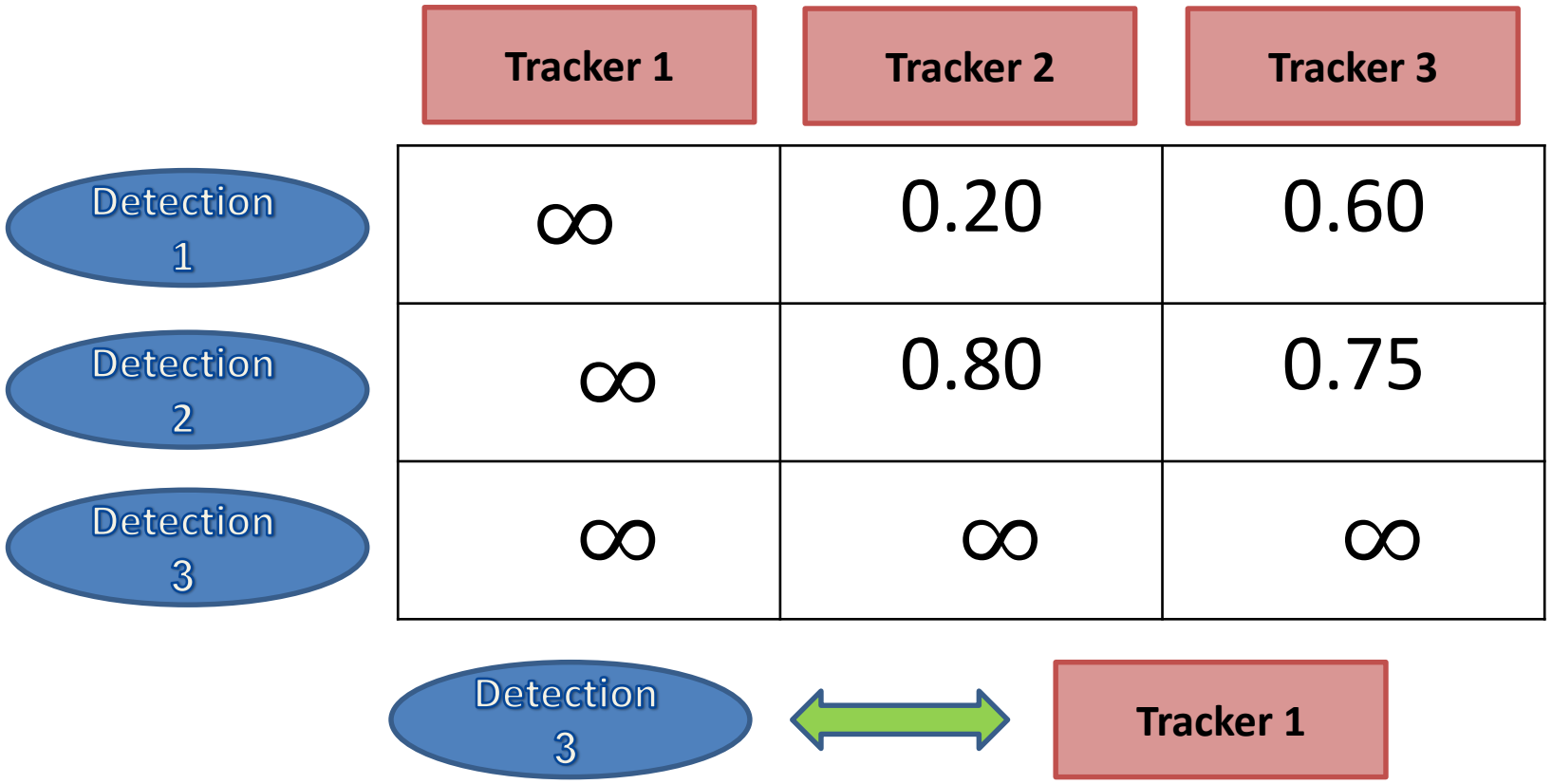
World distance and **KL-Distance** are weighted for creating a **score function** for each **Detection-Tracker** pair

Greedy Data Association

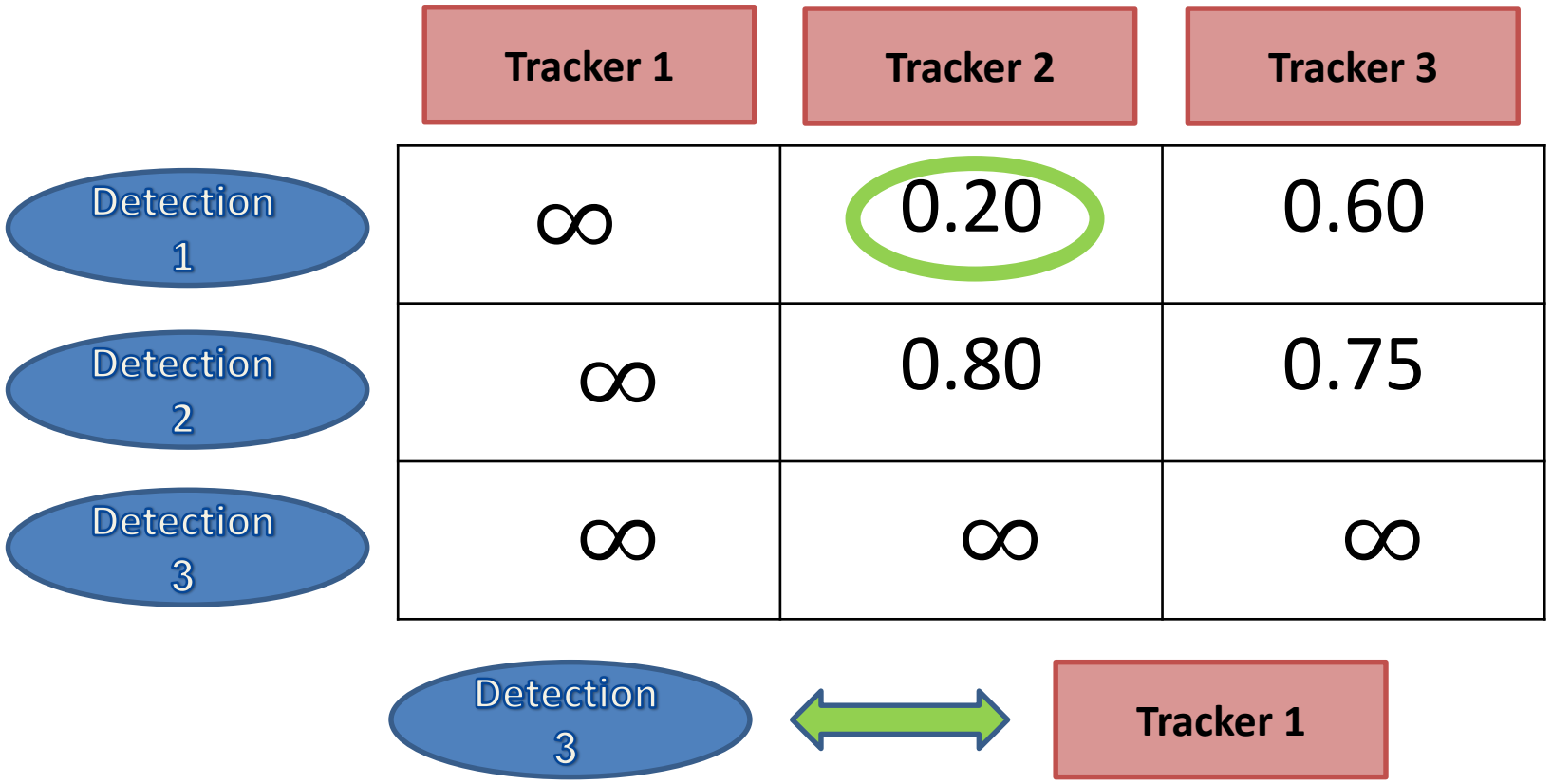
	Tracker 1	Tracker 2	Tracker 3
Detection 1	0.70	0.20	0.60
Detection 2	0.65	0.80	0.75
Detection 3	0.15	0.70	0.55

Greedy Data Association

Greedy Data Association



Greedy Data Association



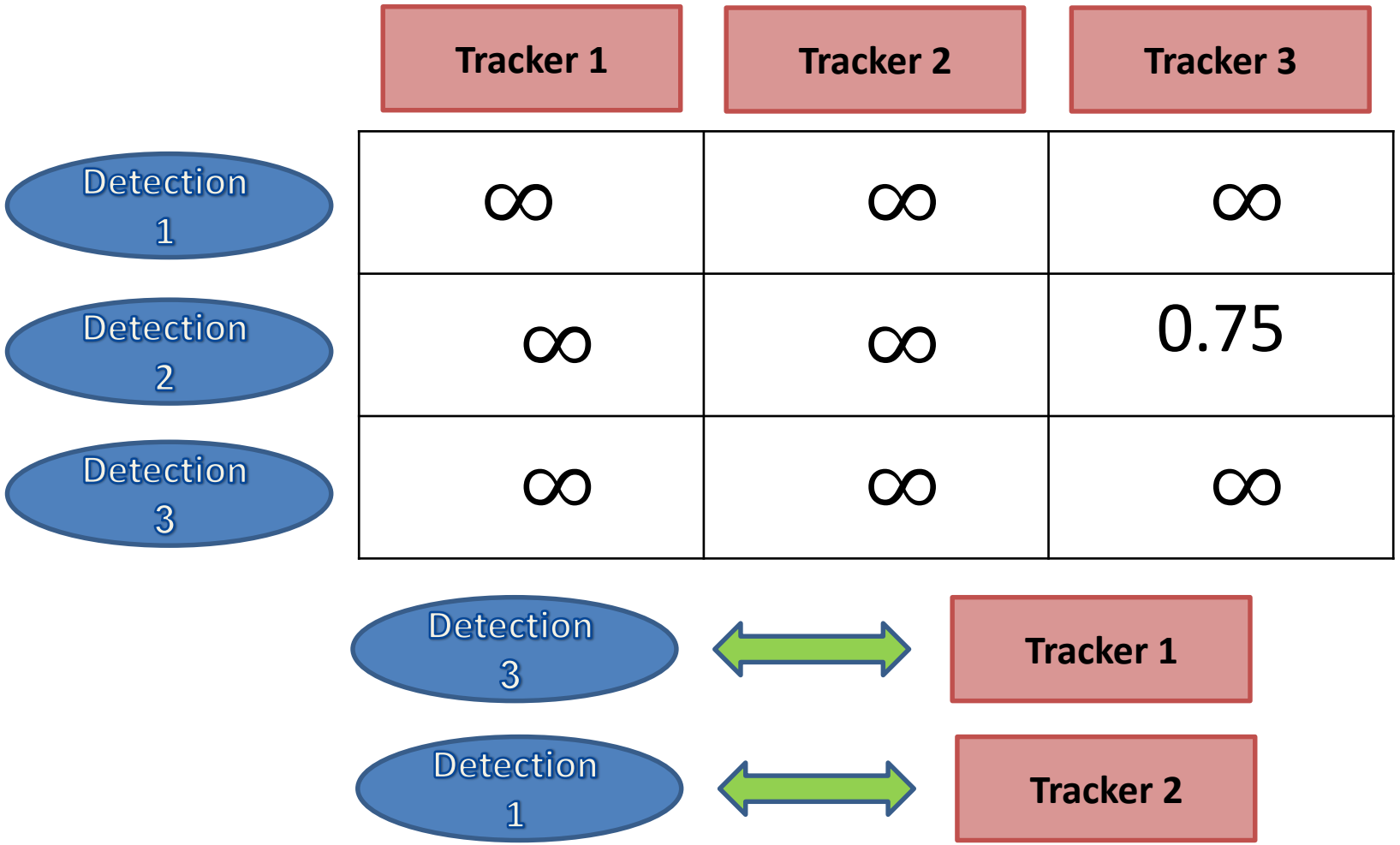
Greedy Data Association

	Tracker 1	Tracker 2	Tracker 3
Detection 1	∞	0.20	0.60
Detection 2	∞	0.80	0.75
Detection 3	∞	∞	∞

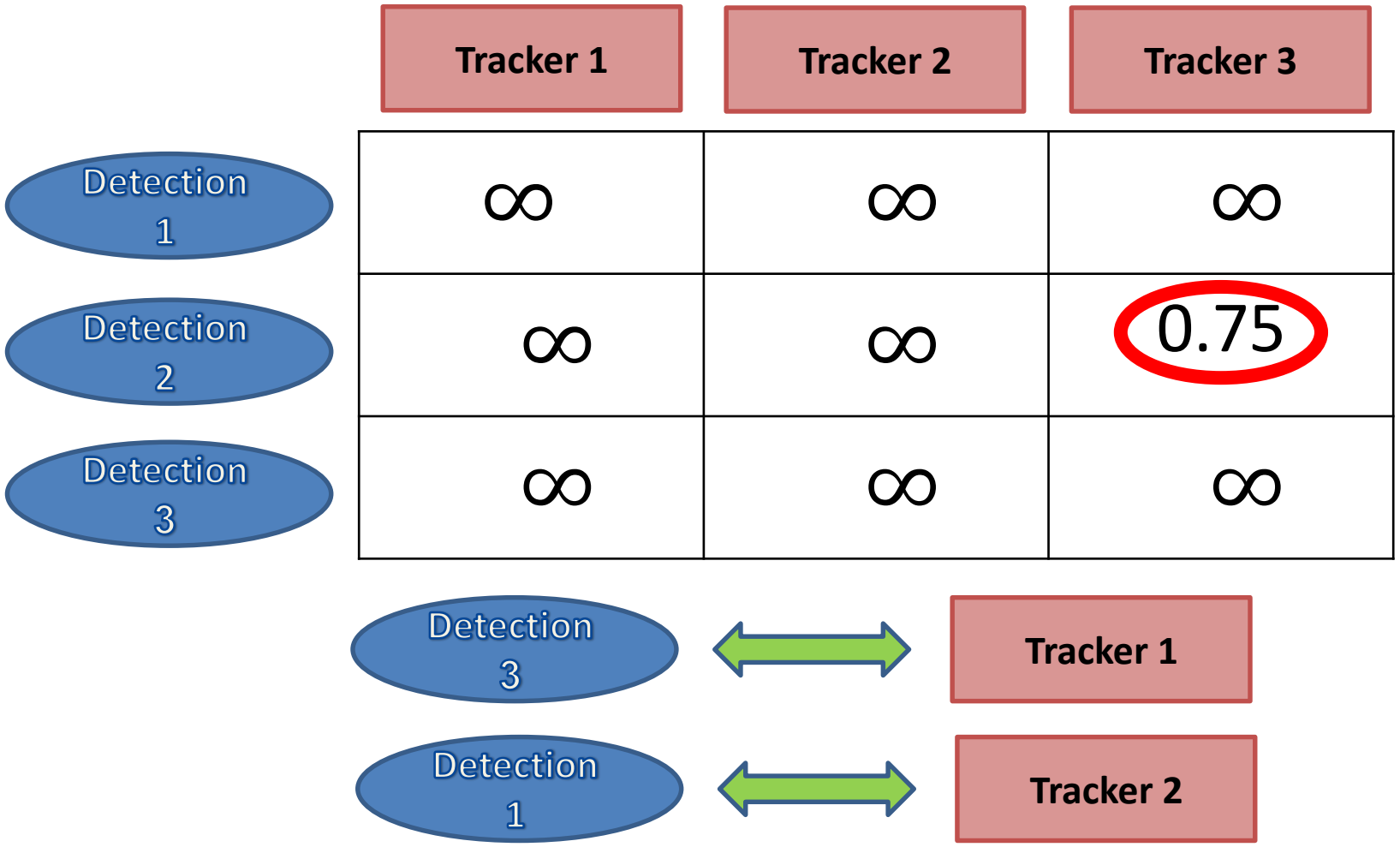
Detection 3 ⟷ Tracker 1

Detection 1 ⟷ Tracker 2

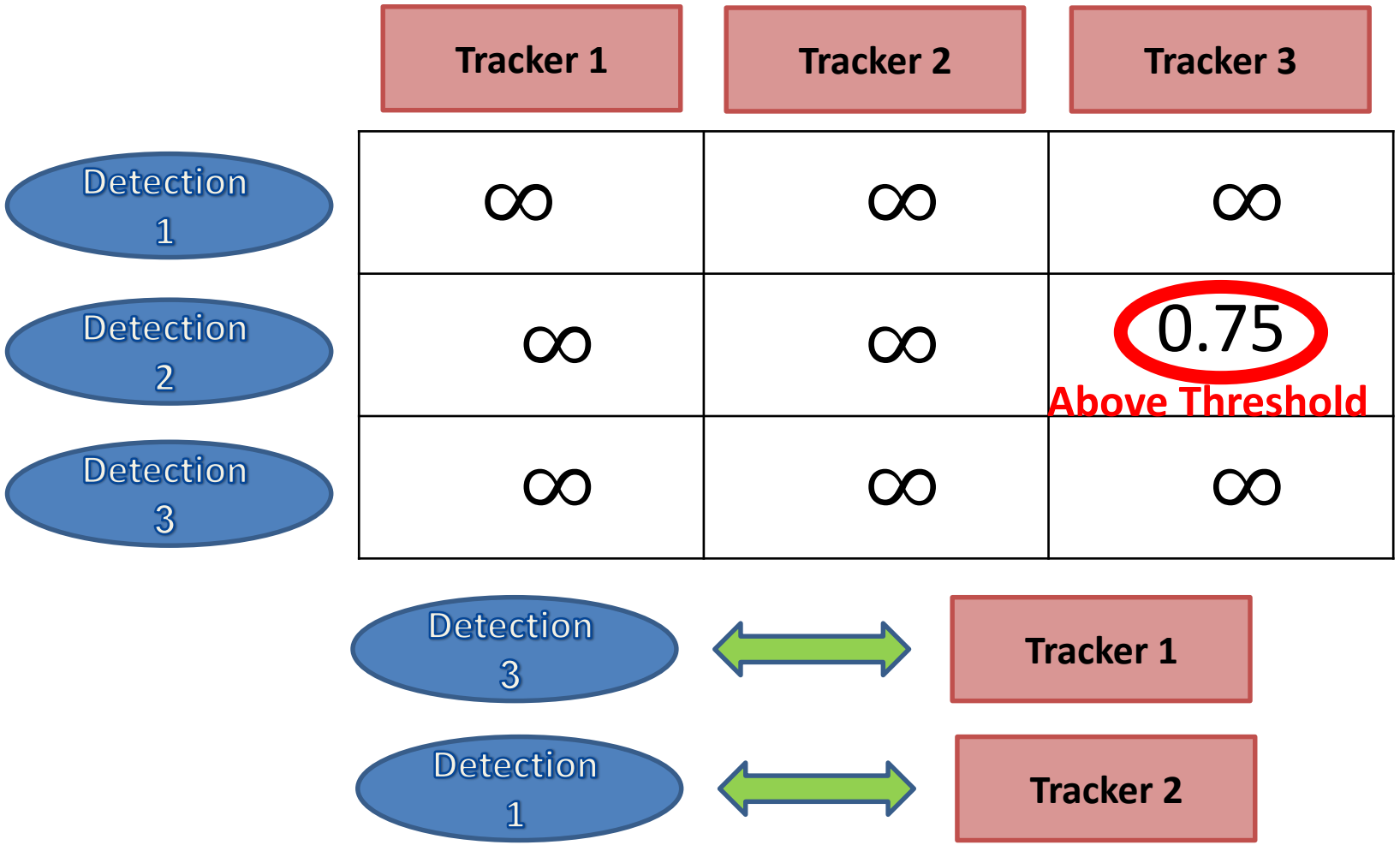
Greedy Data Association



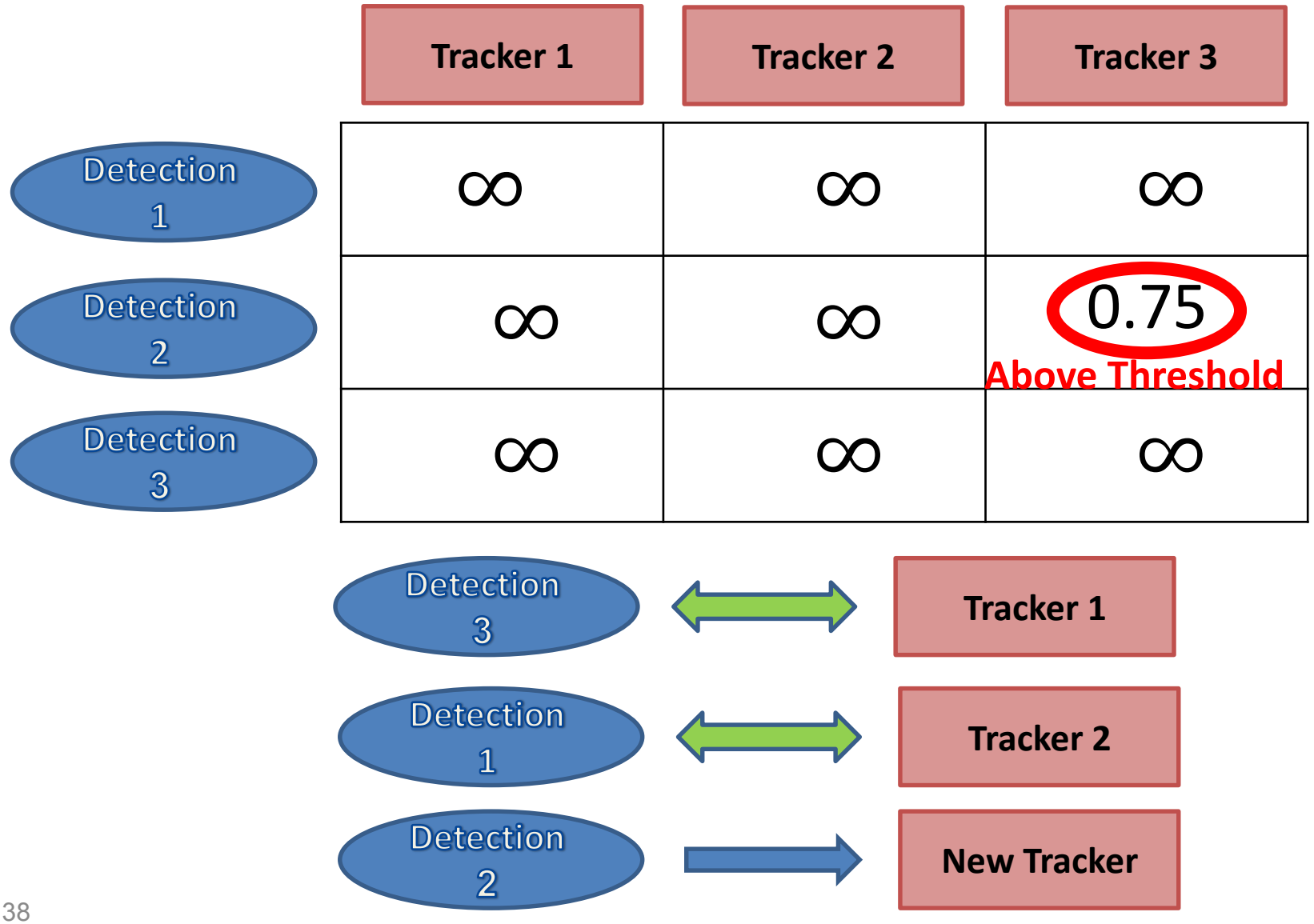
Greedy Data Association



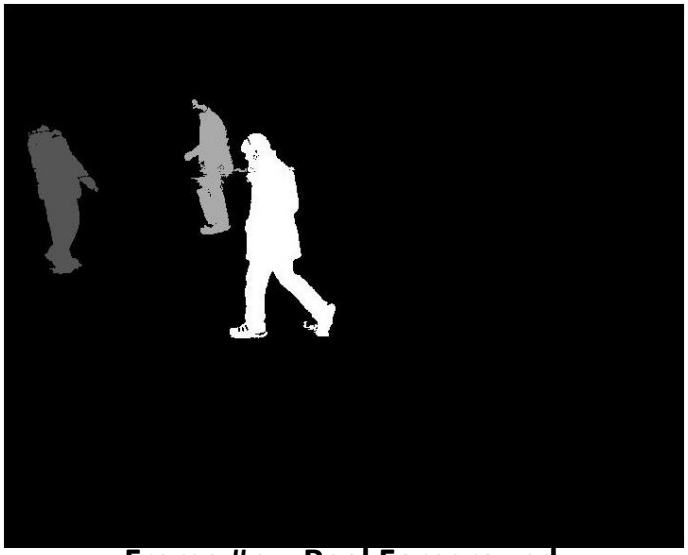
Greedy Data Association



Greedy Data Association

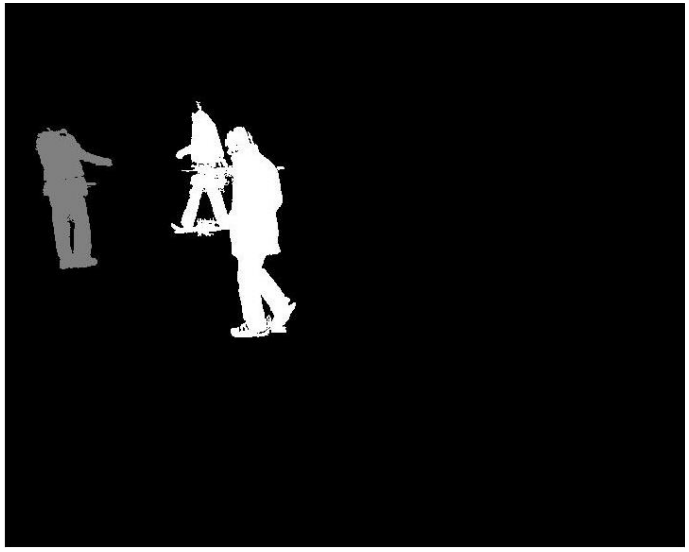


Occlusions Segmentation

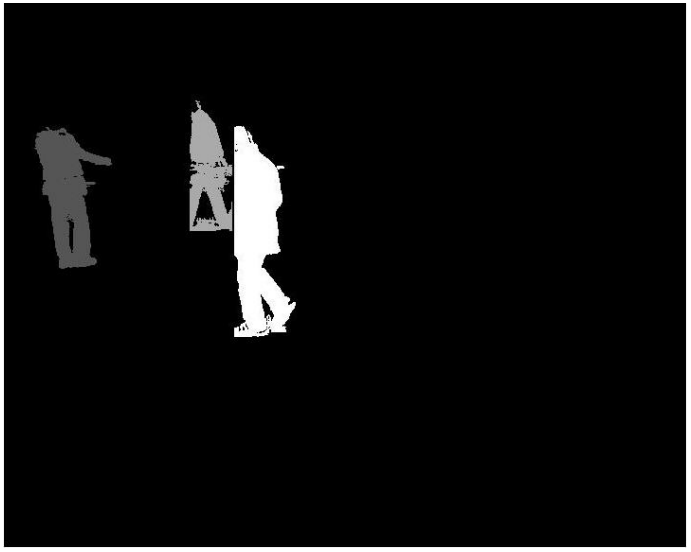


Frame #n – Real Foreground

Frame #n+1 – Estimated Foreground

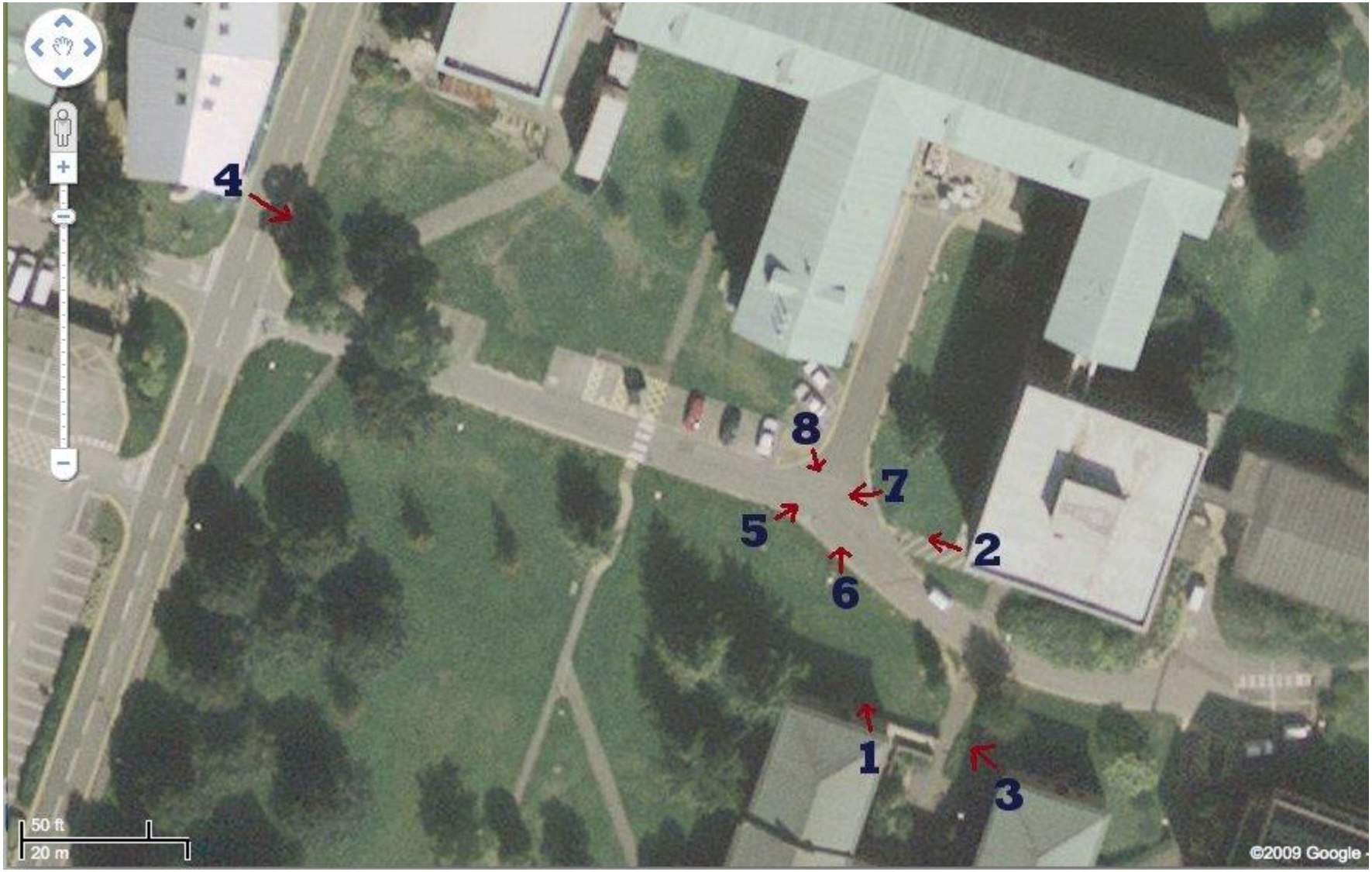


Frame #n+1 – Real Foreground



Frame #n+1 – Occlusion Segmentation

Tracking Demonstration

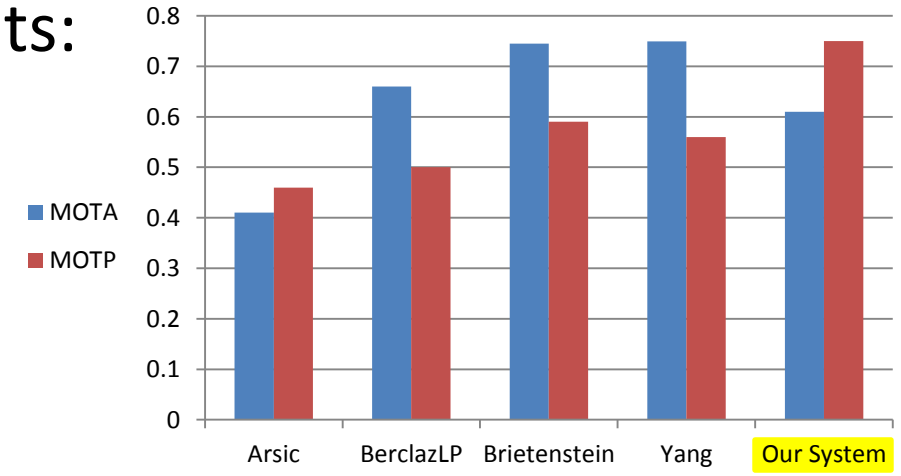


Part B - Tracking Demonstration

Experimental Results

- System was mainly tested on PETS 2010 S2.L1 dataset
- Results were evaluated against single view (view #1) **ground truth** data
- The evaluation used the following metrics[8]:
  - **MOTA (Multiple Object Tracking Accuracy)**
  - **MOTP (Multiple Object Tracking Precision)**

  Evaluation Results:



Conclusion

- A **multiple targets tracking system** has been developed and implemented
- **Novel pedestrian detection method** has been developed
- **Novel Tracking-by-Detection method** was developed – Trackers' data is used for next frames pedestrian detection
- The solution involved **multi camera** input handling, using **homography** tools
- The system includes a re-identification capability, using a **probabilistic appearance model**
- The Algorithm was tested and evaluated using ground truth data according to CLEAR-MOT metrics