Multi-pose people detection in 3D point clouds

Stefano Zanella - March 2017

Problem at hand

Recognise people in RGB-D 3D point clouds
in various poses

Existing approaches

Existing approaches

- pedestrian recognition
  - *A Layered Approach to People Detection in 3D Range Data* - ***Spinello et al.***
  - *Tracking people within groups with RGB-D data* - ***Munaro, Basso, Menegatti***

Existing approaches

- pedestrian recognition
  - *A Layered Approach to People Detection in 3D Range Data* - ***Spinello et al.***
  - *Tracking people within groups with RGB-D data* - ***Munaro, Basso, Menegatti***
- pose estimation
  - *A Multi-view RGB-D Approach for Human Pose Estimation in Operating Rooms* - ***Kadkhodamohammadi et al.***

Existing approaches

- pedestrian recognition
  - *A Layered Approach to People Detection in 3D Range Data* - ***Spinello et al.***
  - *Tracking people within groups with RGB-D data* - ***Munaro, Basso, Menegatti***
- pose estimation
  - *A Multi-view RGB-D Approach for Human Pose Estimation in Operating Rooms* - ***Kadkhodamohammadi et al.***
- supervised ML on full or partial body
  - *People Detection in 3d Point Clouds using Local Surface Normals* - ***Hegger et al.***

Existing approaches

- pedestrian recognition
  - *A Layered Approach to People Detection in 3D Range Data* - ***Spinello et al.***
  - *Tracking people within groups with RGB-D data* - ***Munaro, Basso, Menegatti***
- pose estimation
  - *A Multi-view RGB-D Approach for Human Pose Estimation in Operating Rooms* - ***Kadkhodamohammadi et al.***
- supervised ML on full or partial body
  - *People Detection in 3d Point Clouds using Local Surface Normals* - ***Hegger et al.***
- mixed approaches
  - *Detecting and Tracking People using an RGB-D Camera via Multiple Detector Fusion* - ***Choi, Pantofaru, Savarese***

Drawbacks of existing approaches

Drawbacks of existing approaches

- pedestrian recognition
  - **not general enough**

Drawbacks of existing approaches

- pedestrian recognition
  - **not general enough**
- pose estimation
  - **additional complexity not always necessary**

Drawbacks of existing approaches

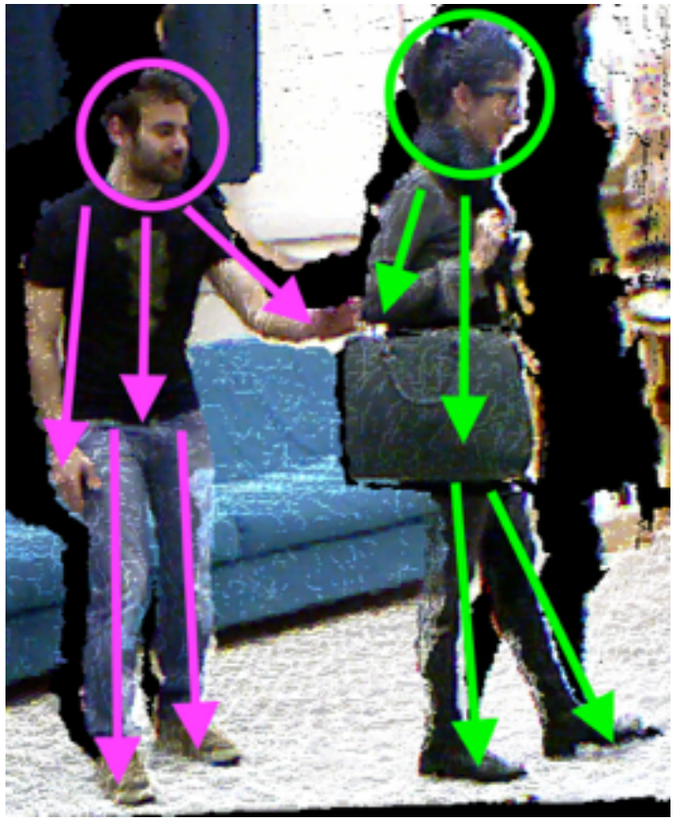
- pedestrian recognition
  - **not general enough**
- pose estimation
  - **additional complexity not always necessary**
- supervised ML on full or partial body
  - **not general enough, slower than pedestrian recognition**

Drawbacks of existing approaches

- pedestrian recognition
  - **not general enough**
- pose estimation
  - **additional complexity not always necessary**
- supervised ML on full or partial body
  - **not general enough, slower than pedestrian recognition**
- mixed approaches
  - **additional complexity for the merging of partial results**

Intuition

Head detection + body segmentation



Why?

- Head is a good predictor of human presence
- Head is rigid - low number of possible poses
- Body is compact so segmentation should be easy once a head is found

Benefits

- Robust to very different body poses
- Robust to partial, occluded bodies
- Leverages body compactness: no need for complex ML algorithms
- Doesn’t require complex pose estimation
- Divide et impera
- Lot of literature on head (face) detection
- Lot of literature on segmentation

Project Architecture

- Head detection: adapted Viola-Jones algorithm

- Body segmentation: RGB region growing from PCL

Project Architecture

Key contributions

- Flexible people detector that works on any pose
- Viola-Jones implementation (training + detection)
- Optimisation of window scaling based on 3D info
- Optimisation of training algorithm based on 3D info

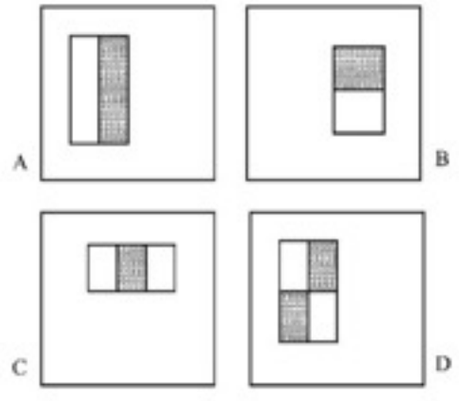
Viola-Jones

Key features

- (Really) Fast
- Fairly robust to head rotation (±15° in plane, ±45° out of plane)
- Guarantees asymptotic “perfection”

Viola-Jones

Features



Weak classifier

$$h(x, f, p, \theta) = \begin{cases} 1, & \text{if } pf(x) < p\theta \\ 0, & \text{otherwise} \end{cases}$$

Viola-Jones

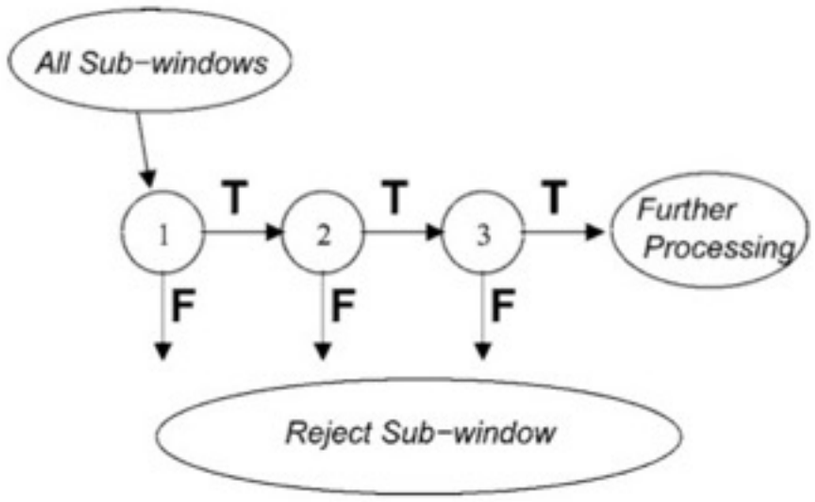
Weak classifier $$h(x, f, p, \theta) = \begin{cases} 1, & \text{if } pf(x) < p\theta \\ 0, & \text{otherwise} \end{cases}$$

Strong classifier $$C(X) = \begin{cases} 1 & \sum_{t=1}^{T} \log \frac{1-\epsilon_t}{\epsilon_t} h_t(X) \geq \frac{1}{2} \sum_{t=1}^{T} \log \frac{1-\epsilon_t}{\epsilon_t} \\ 0 & \text{otherwise} \end{cases}$$

Viola-Jones

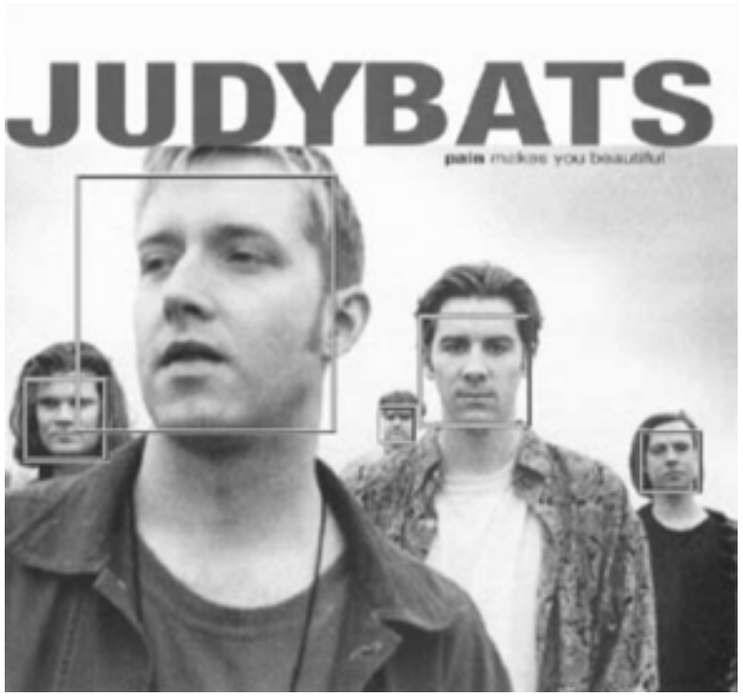
Strong classifier $$C(X) = \begin{cases} 1 & \sum_{t=1}^{T} \log \frac{1-\epsilon_t}{\epsilon_t} h_t(X) \geq \frac{1}{2} \sum_{t=1}^{T} \log \frac{1-\epsilon_t}{\epsilon_t} \\ 0 & \text{otherwise} \end{cases}$$

Cascaded classifier



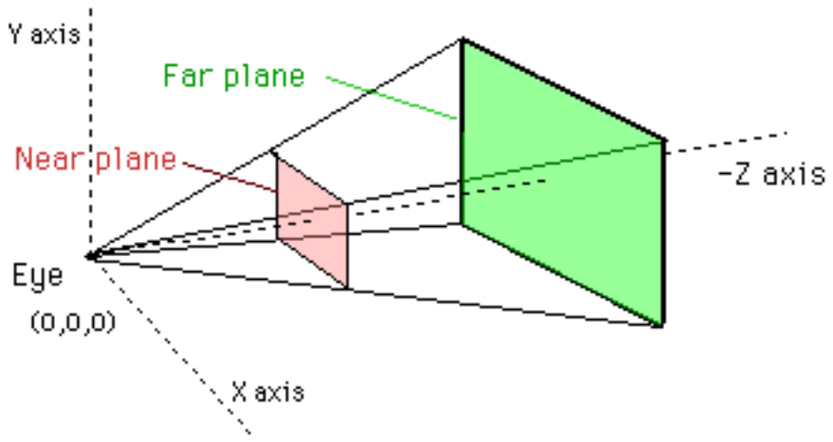
Searching for faces using depth information

Multi-scale scanning

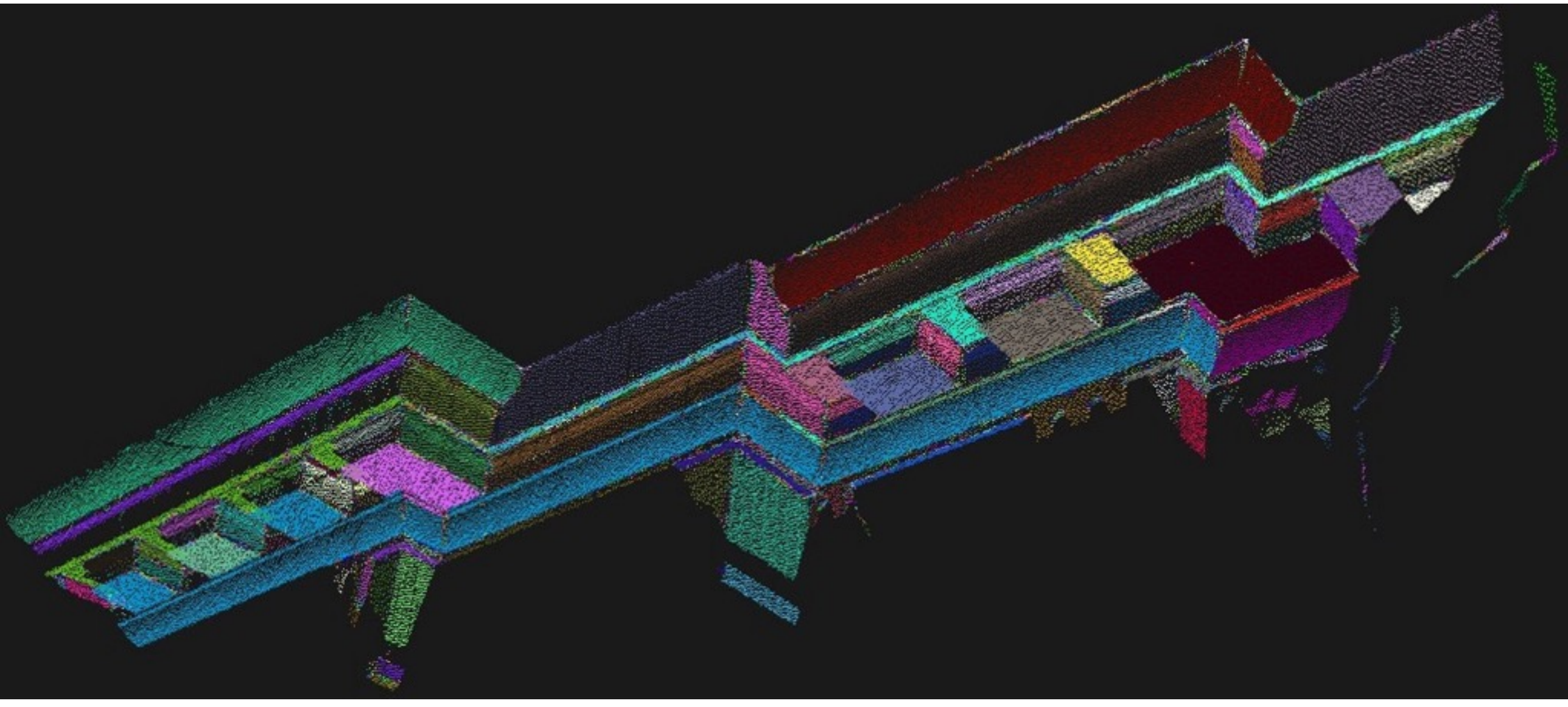


vs

Scaling the sub-window depending on current depth



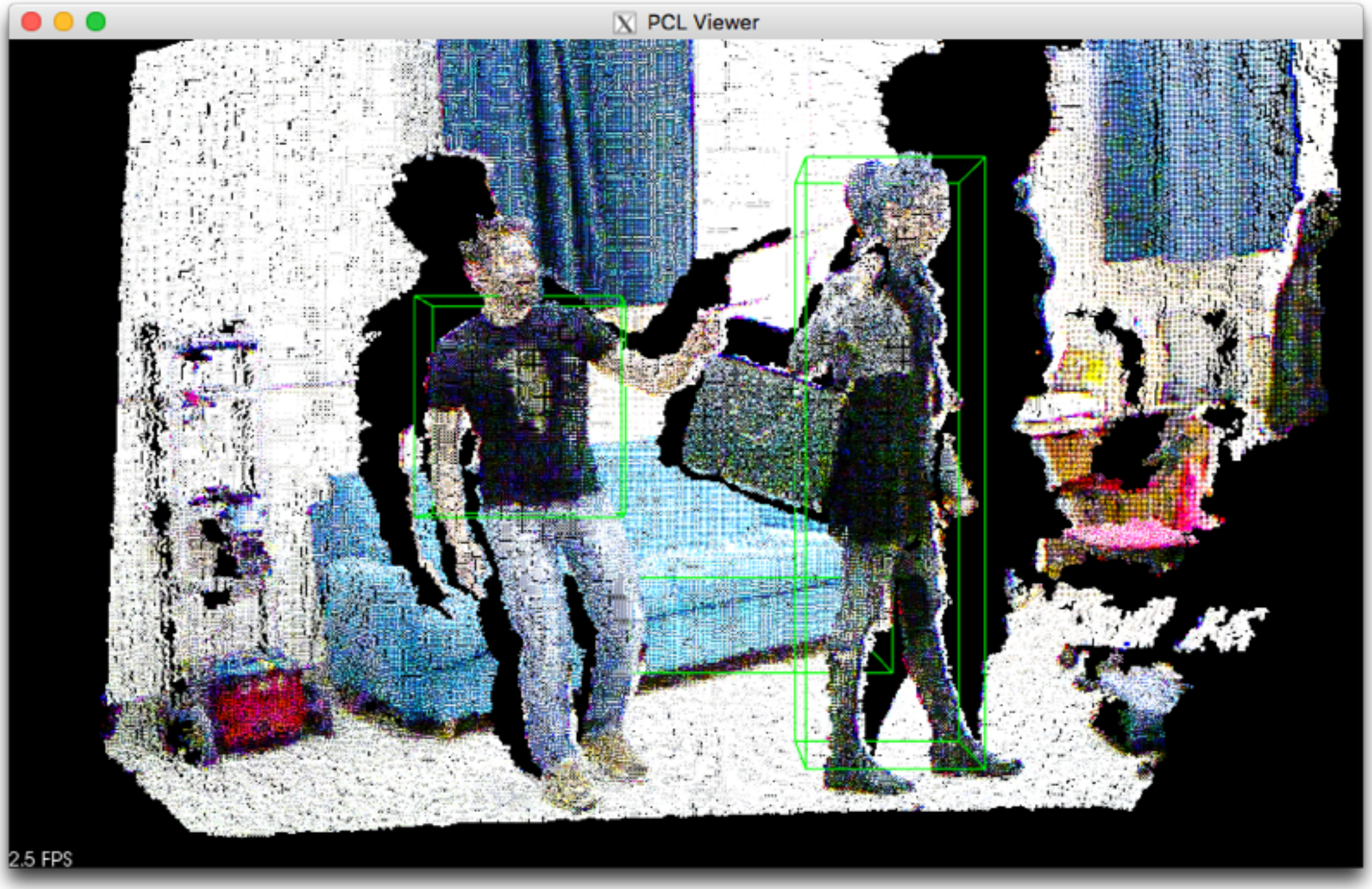
RGB Region Growing



Training caveats

- Small training set (<3k vs >15k in original paper)
- Trained for heads in different positions
- Upscaling > downscaling features
- Exiting the training loop when performance saturates

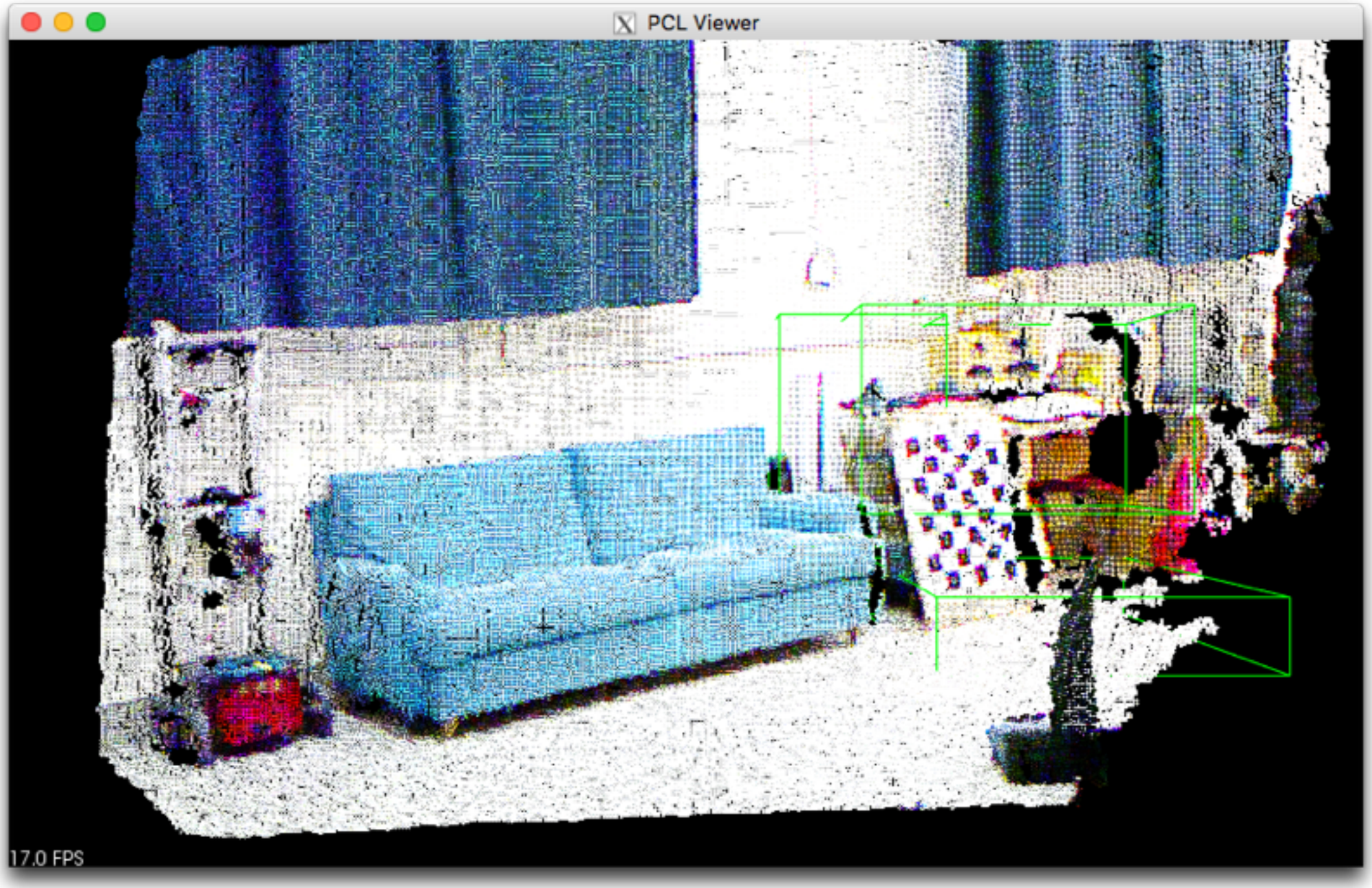
Results (the good)



Results (the good)

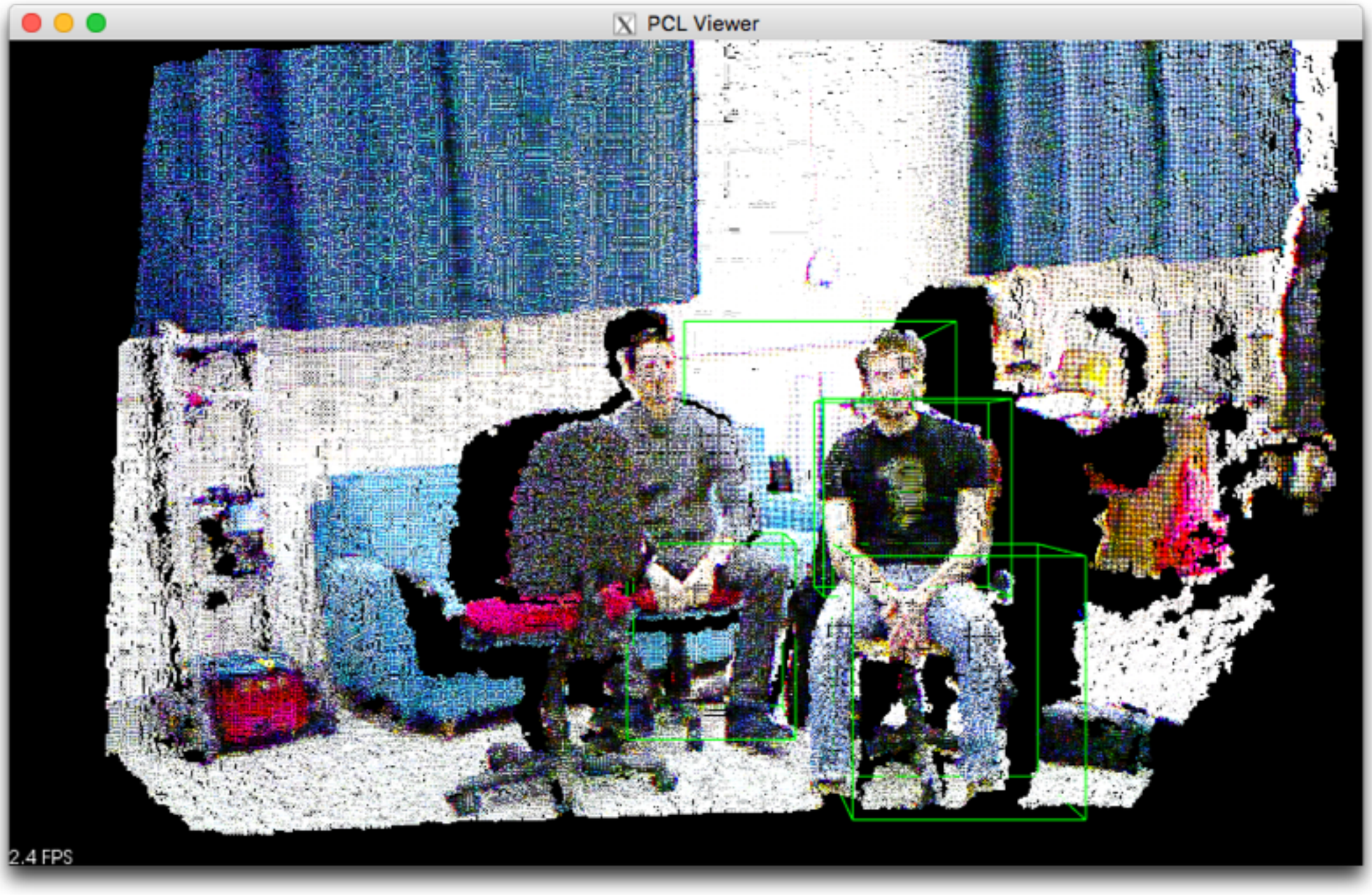
Results (the good)

Results (the bad)



Results (the bad)

Results (the bad)



Results

- DR: 0% - 100% (avg 58.57%)
- FPR: 0% - 0.06%
- Accuracy qualitatively less than desired
- Head detection vs body segmentation: 1:200 speed ratio

Future developments

- Train with more data (and more general)
- Parallel, specialized detectors for heads in different positions
- Use depth map for head detection rather than intensity information
- Improvements of region growing algorithm to avoid full-scene segmentation

Thanks