# Attentional Object Detection

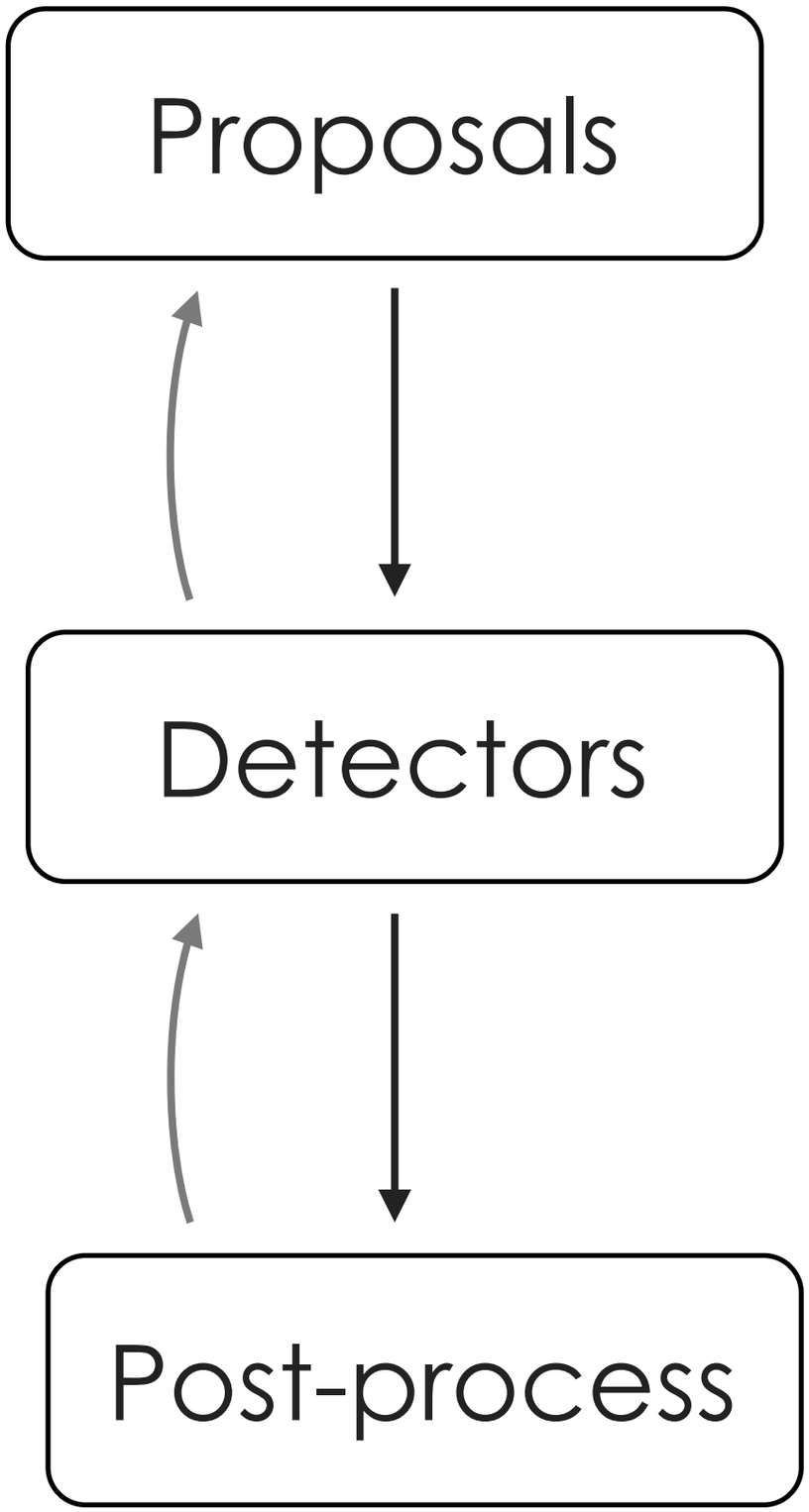
## Why look for everything everywhere?

Sergey Karayev

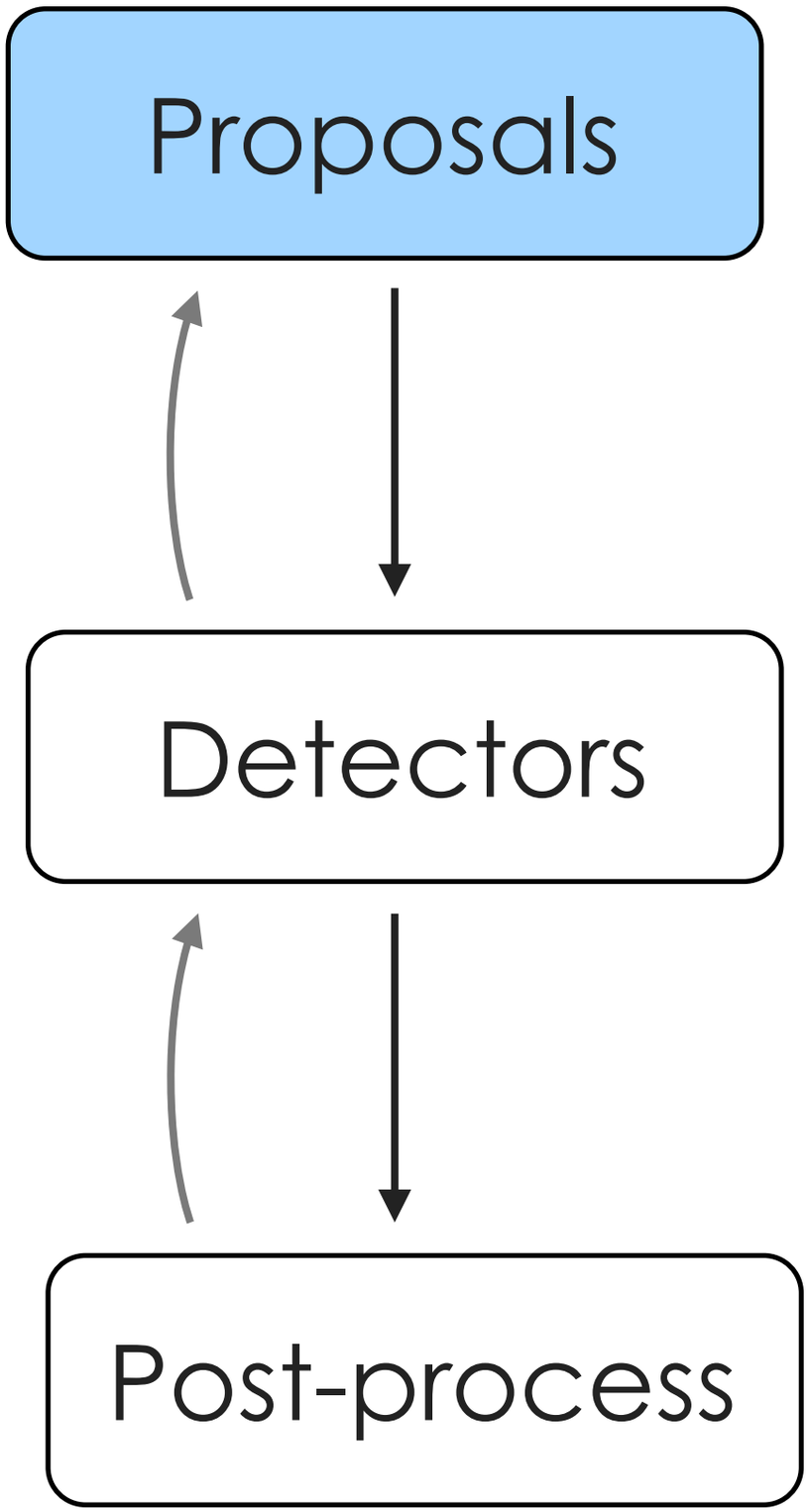
for UC Berkeley Computer Vision Retreat 2011

Problem:

Recognition and localization of objects of multiple classes in cluttered scenes.




# Object Detection



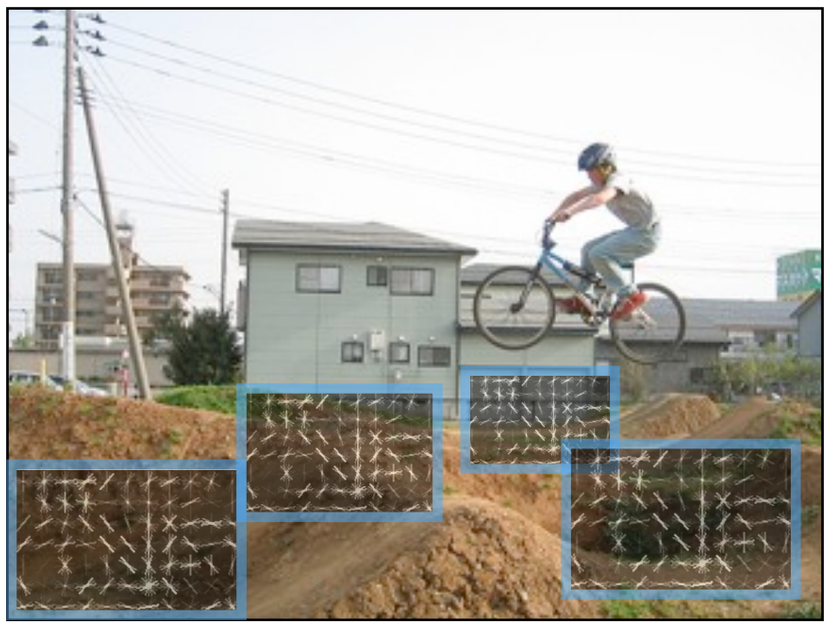

# Object Detection


Sliding window

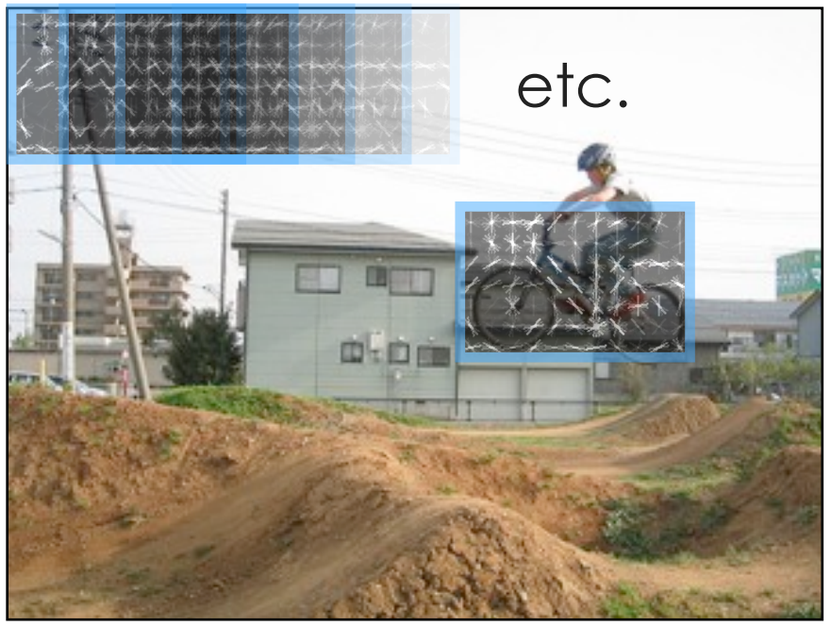
**Proposals**

...with priors/ pruning

Voting



Efficient search



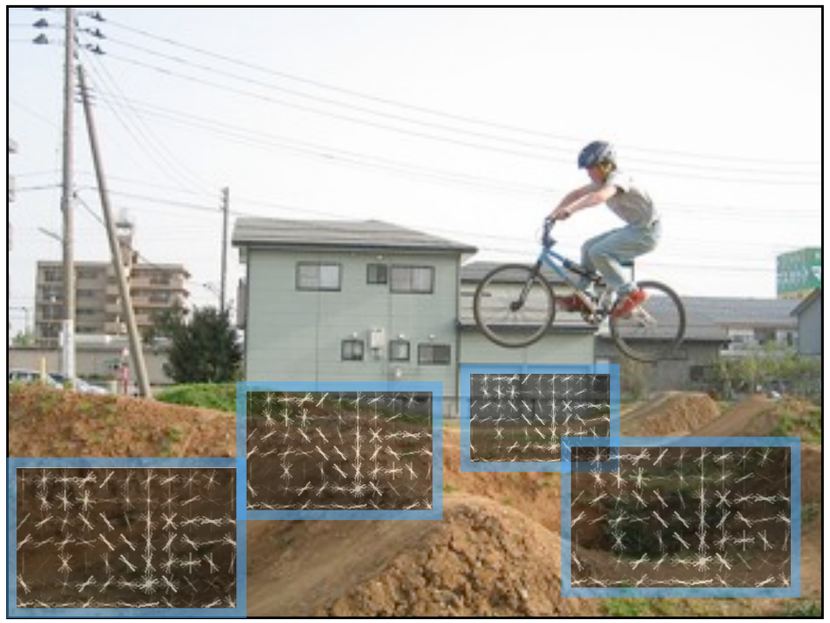

Sliding window

**Proposals**

- **Too slow**: quadratic in number of search dimensions (x,y,scale,class).
- **Speed-ups**:
  - Parallelization.
  - ★Priors/Pruning with non-detector features.
  - ★Algorithmic efficiency.

# Proposals

Priors/pruning



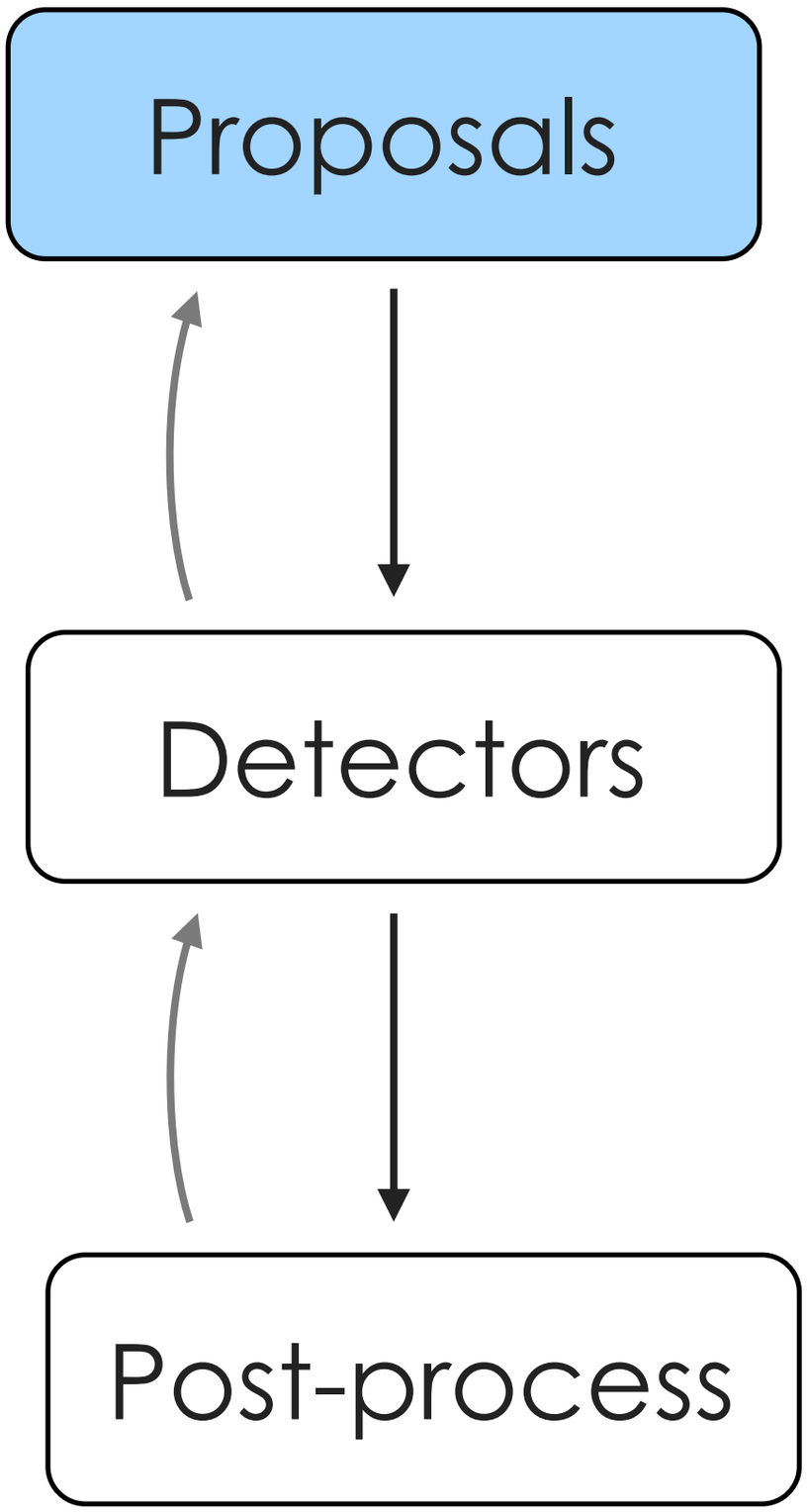
- Uses non-detector features (location, geometry, context, depth, "objectness")
- **Often done in post-processing.**

# Proposals

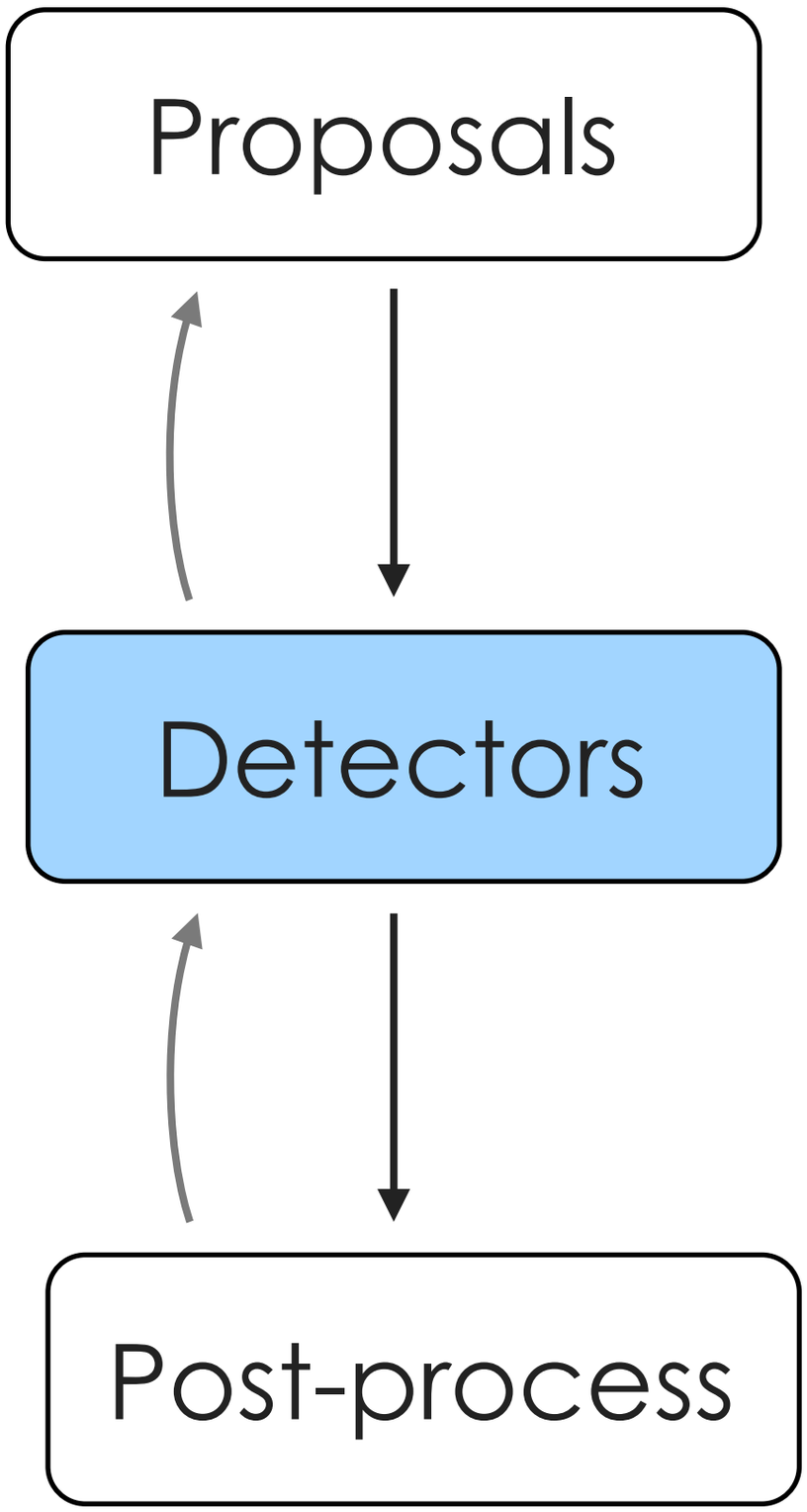
Currently only works for local features.

Voting

Efficient subwindow search

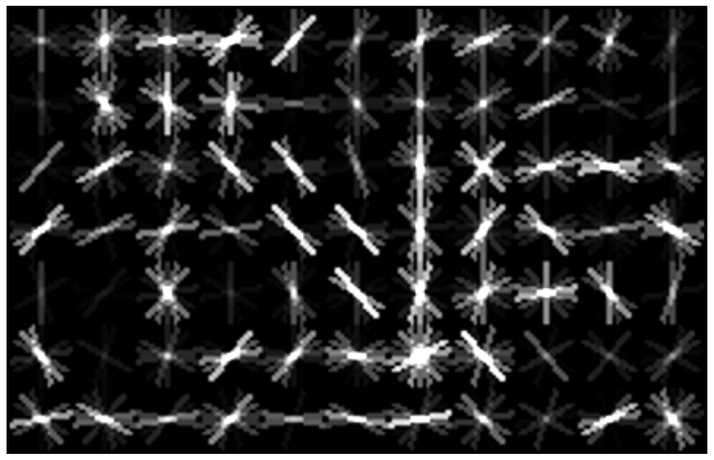



- Priority ordered? How?
- Pruned / Exhaustive?
- Class-specific?




- Priority ordered? How?
- Pruned / Exhaustive?
- Class-specific?

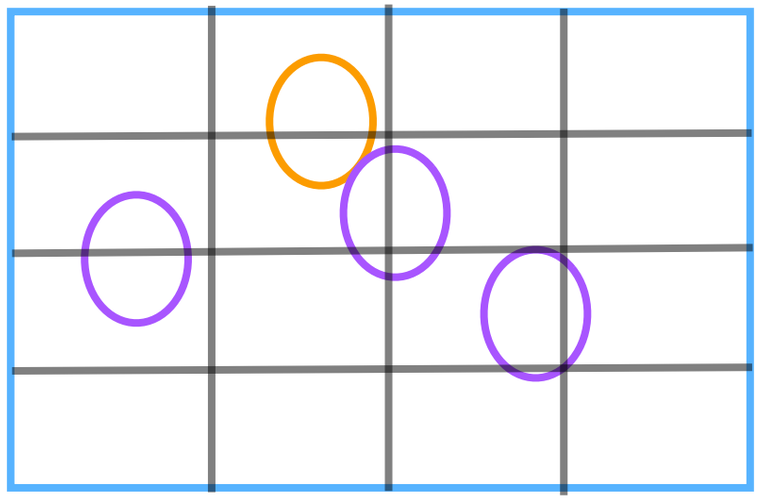
# Detector



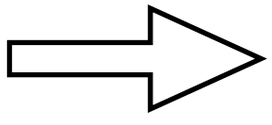
Template/Parts

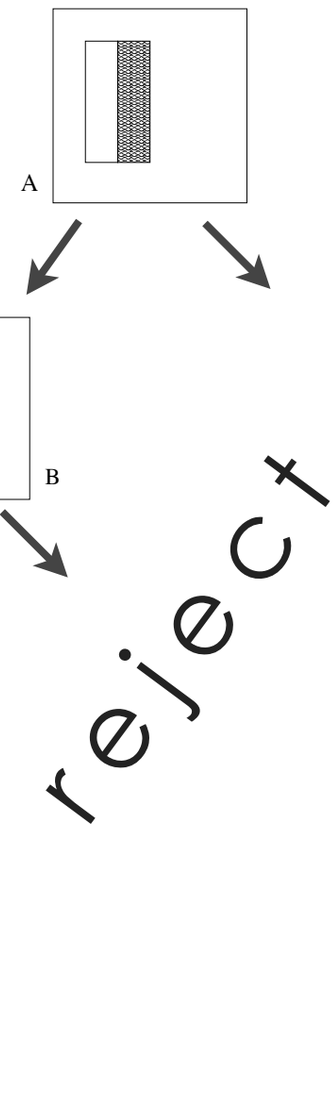
Local features

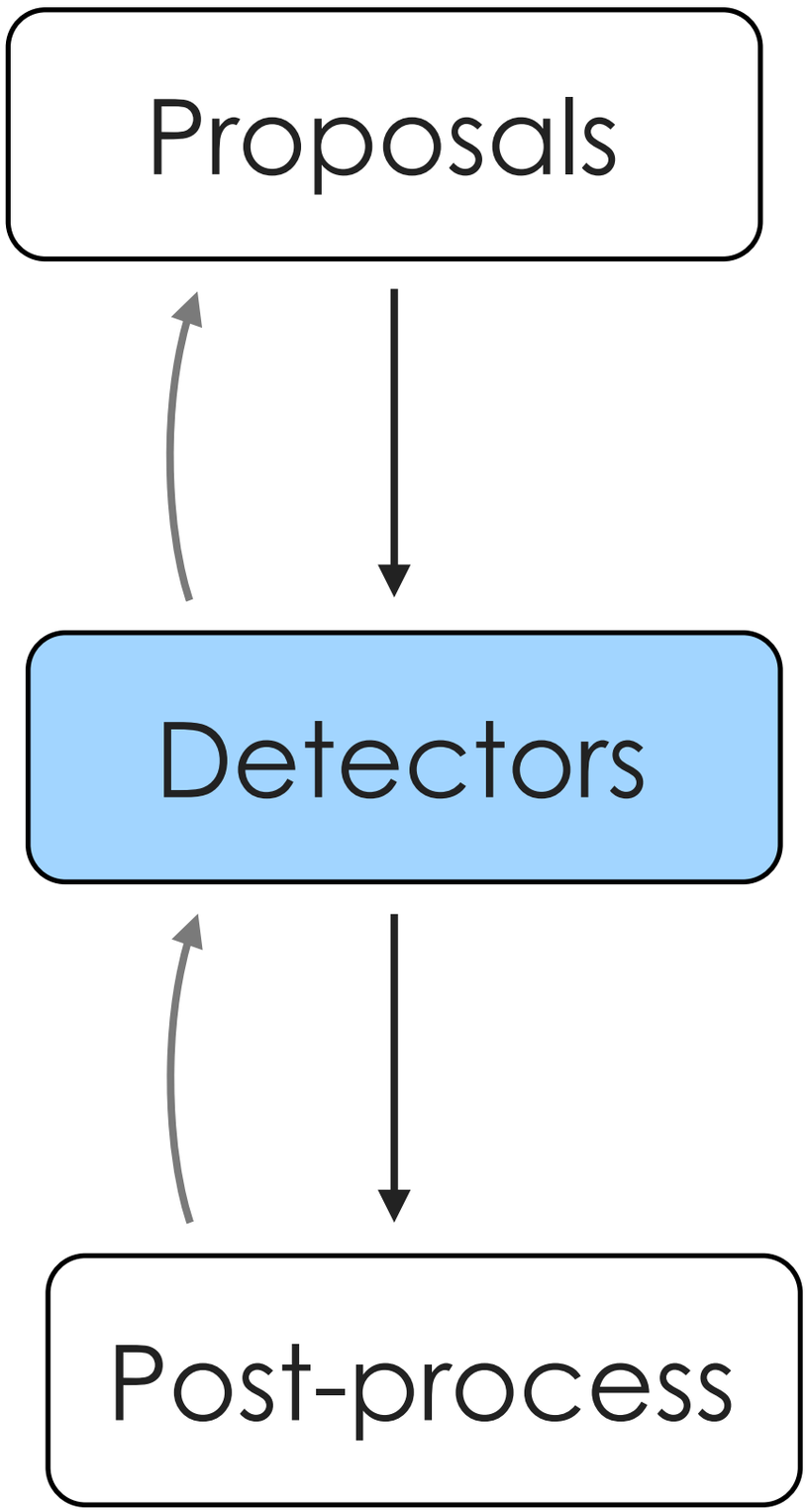




Decision stumps

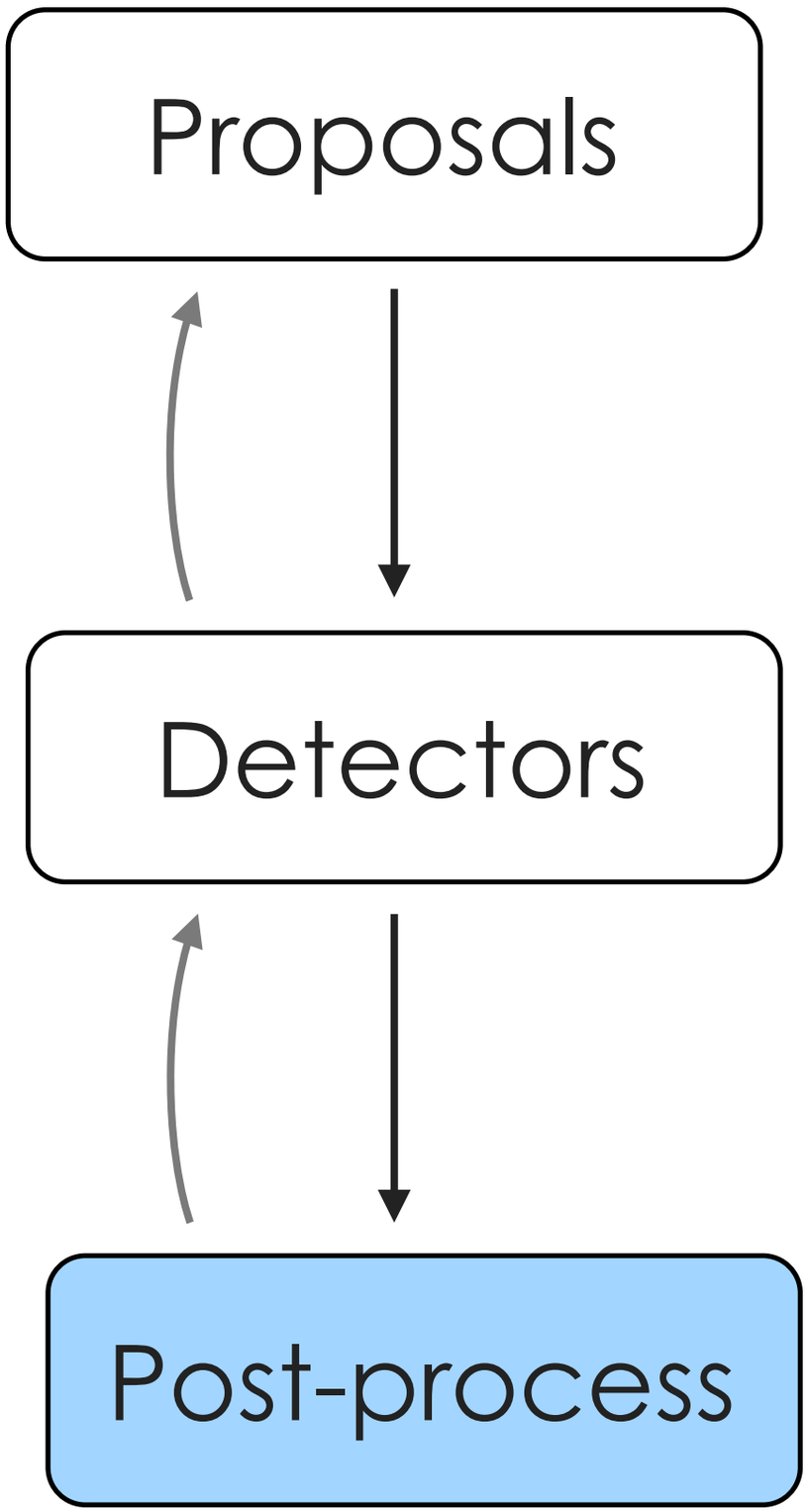







- Priority ordered? How?
- Pruned / Exhaustive?
- Class-specific?

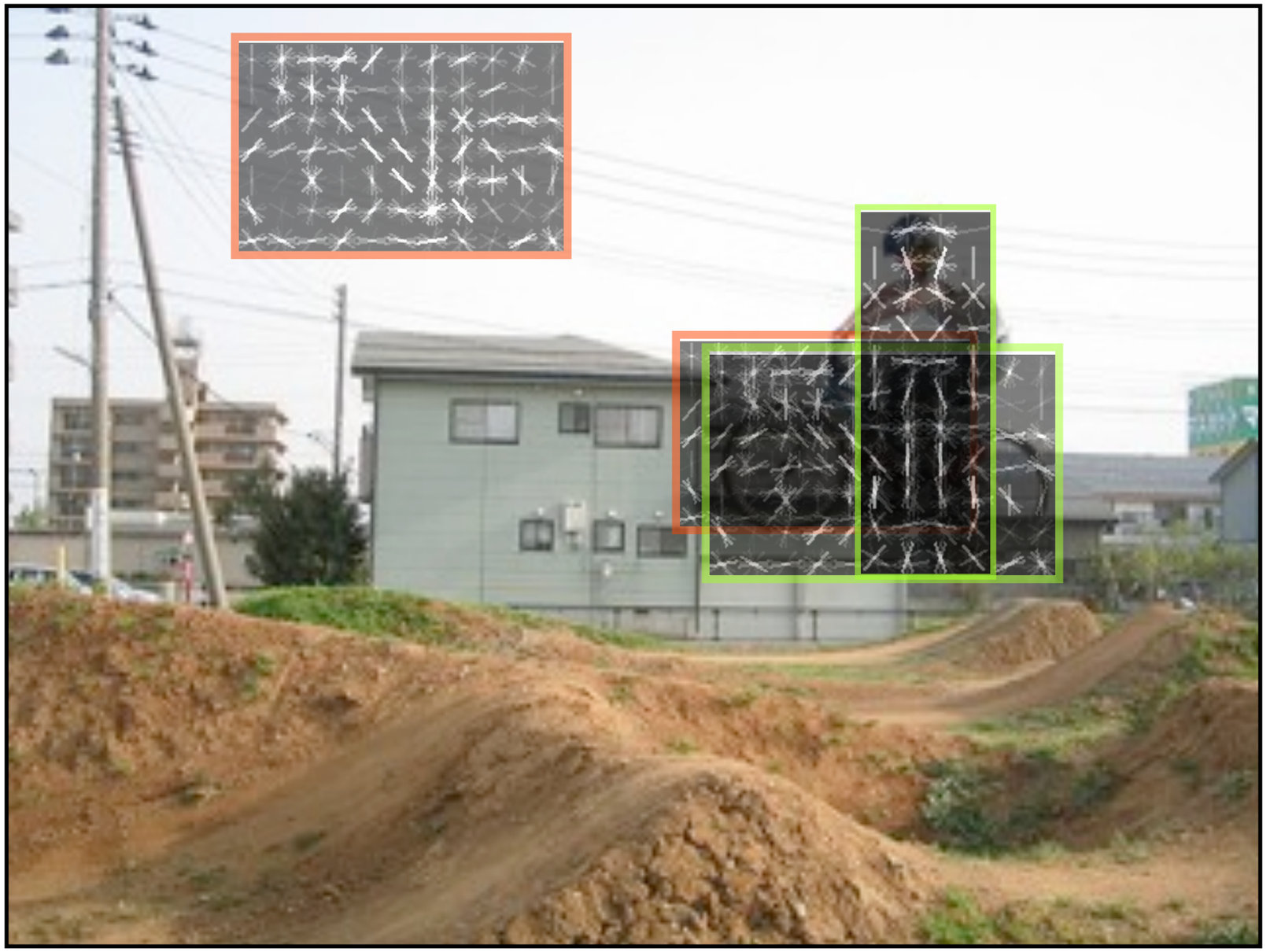
- Local or global feature?
- Shared parts across classes?
- Cascaded?
- Confidence ≈ likelihood?

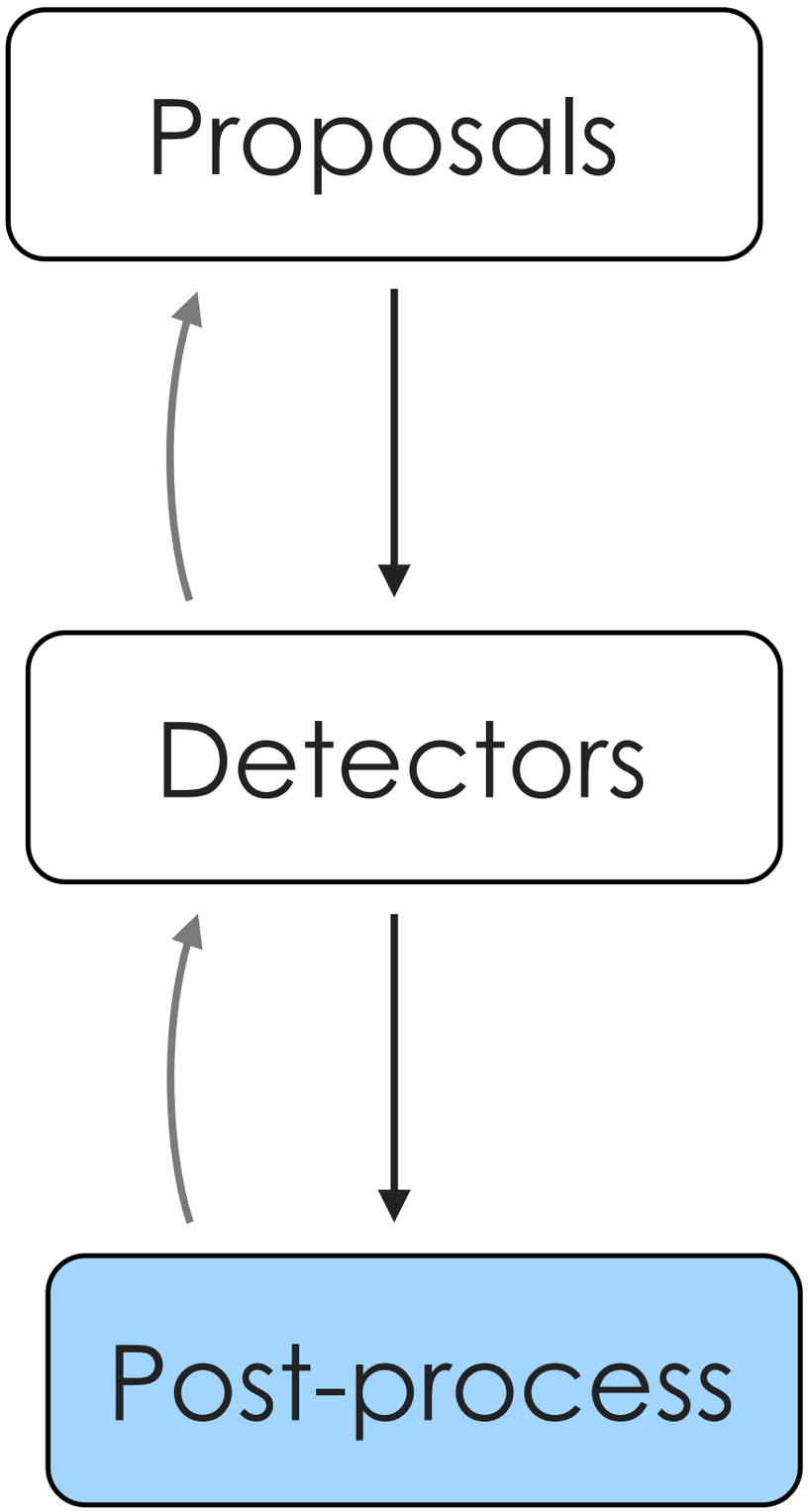



- Priority ordered? How?
- Pruned / Exhaustive?
- Class-specific?

- Local or global feature?
- Shared parts across classes?
- Cascaded?
- Confidence ≈ likelihood?

# Post-process

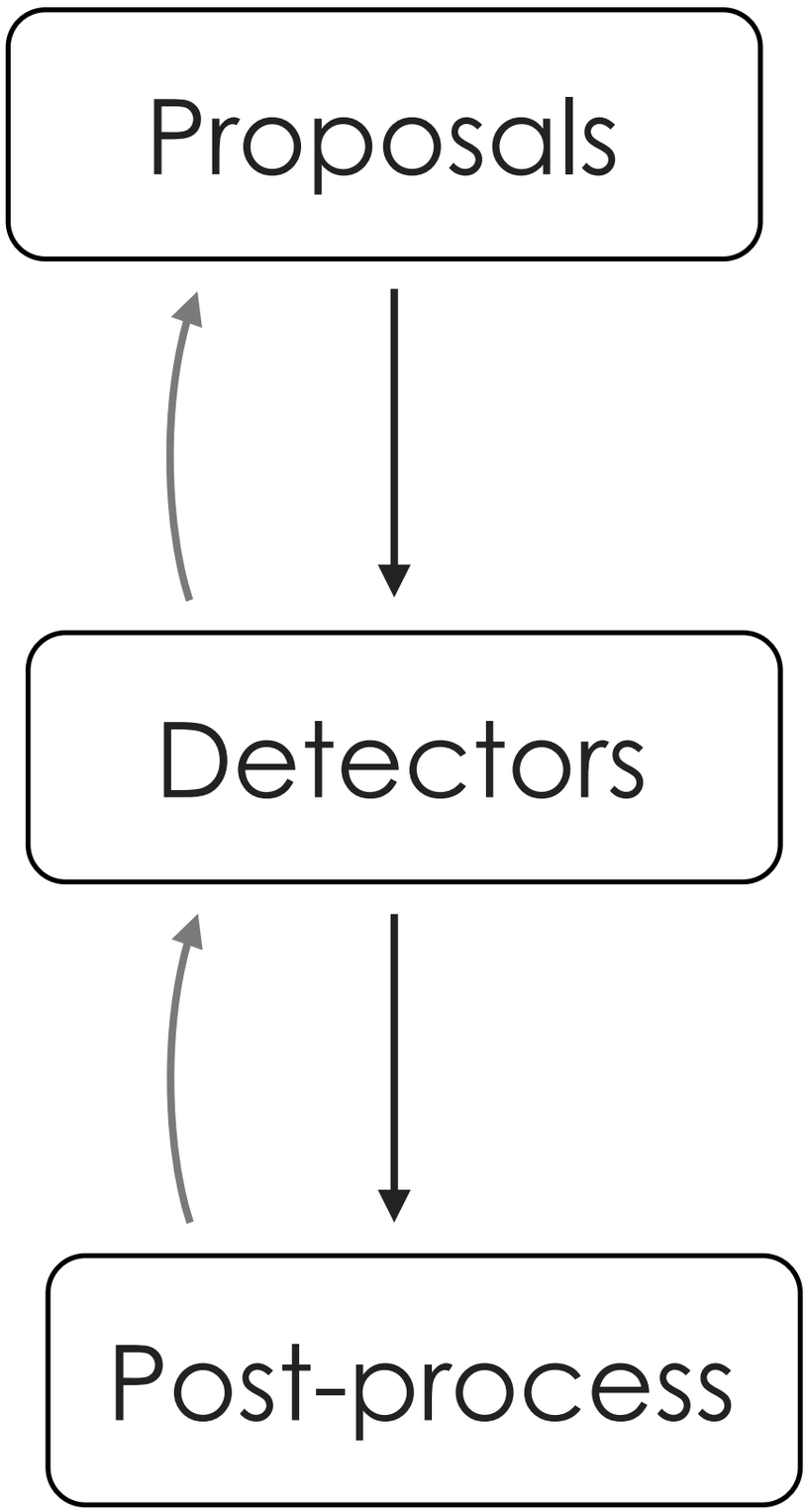





- Priority ordered? How?
- Pruned / Exhaustive?
- Class-specific?

- Local or global feature?
- Shared parts across classes?
- Cascaded?
- Confidence ≈ likelihood?

- NMS/Meanshift?
- Context? (Inter-object?)




- Priority ordered? How?
- Pruned / Exhaustive?
- Class-specific?

- Local or global feature?
- Shared parts across classes?
- Cascaded?
- Confidence ≈ likelihood?

- NMS/Meanshift?
- Context? (Inter-object?)

# Where we are

Cascaded Deformable Part Models.
Per class, ~1 sec / medium-sized image.

# Where we are

- **PASCAL**: ~5K test images, 20 classes. 28 hours to process.
- **ImageNet '11**: ~450K test images, 3000 classes. 375,000 hours to process.

# Where we are

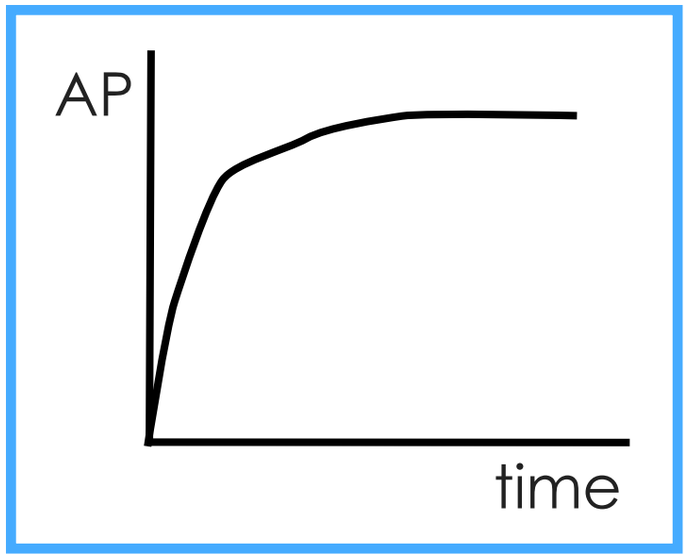
- **Standard movie**: ~130K frames. 36 hours per object class.

So what can we do?

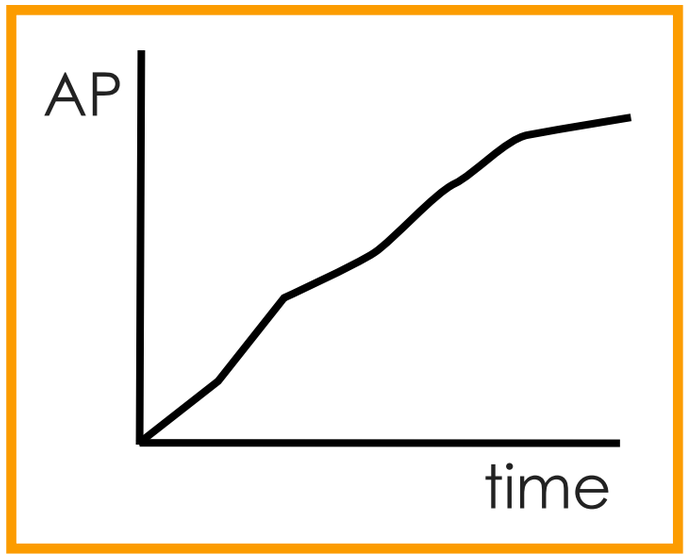
Not look for everything everywhere!

# New Performance Evaluation

- Goal: Be able to stop detection and have the most correct detections and the fewest incorrect detections **at any time**.




vs.



# How?

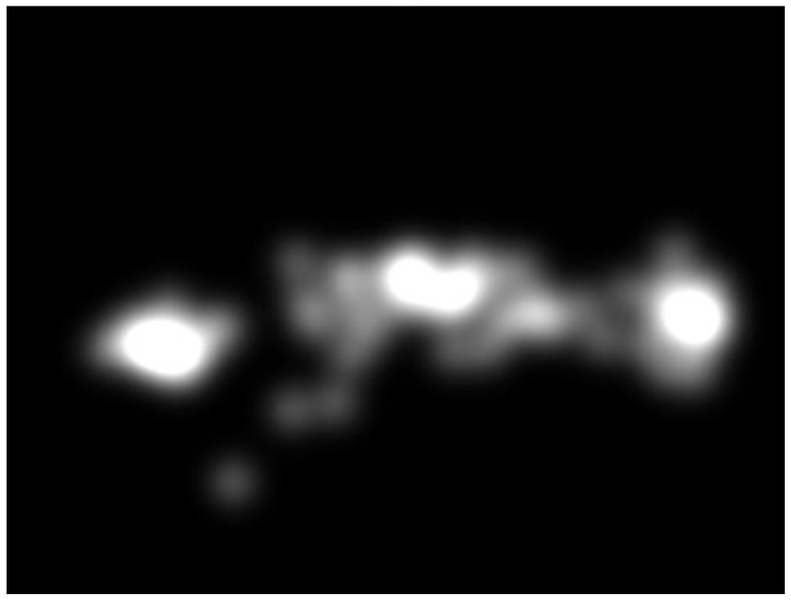
# Attention

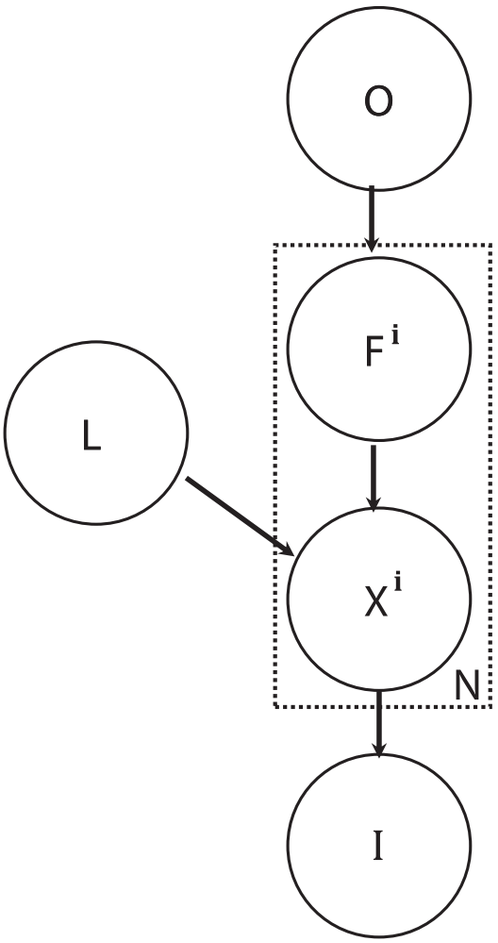
- Natural bottleneck in animal vision.
- Two kinds:
  - Bottom-up: rapid, driven by featurization.
  - Top-down: secondary, driven by task.
- Eye fixations are a good proxy for implicit attention. Necessary because of the fovea.

# Basic ideas

- Single saliency map from which foci of attention are selected.

- Sequential selection due to “inhibition of return,” or information maximization.
- Influenced from the top.






Free viewing
(uniform priors)

Search for cars and pedestrians
(learned priors)

car search

pedestrian search

Fixations

Model posterior

Fixations

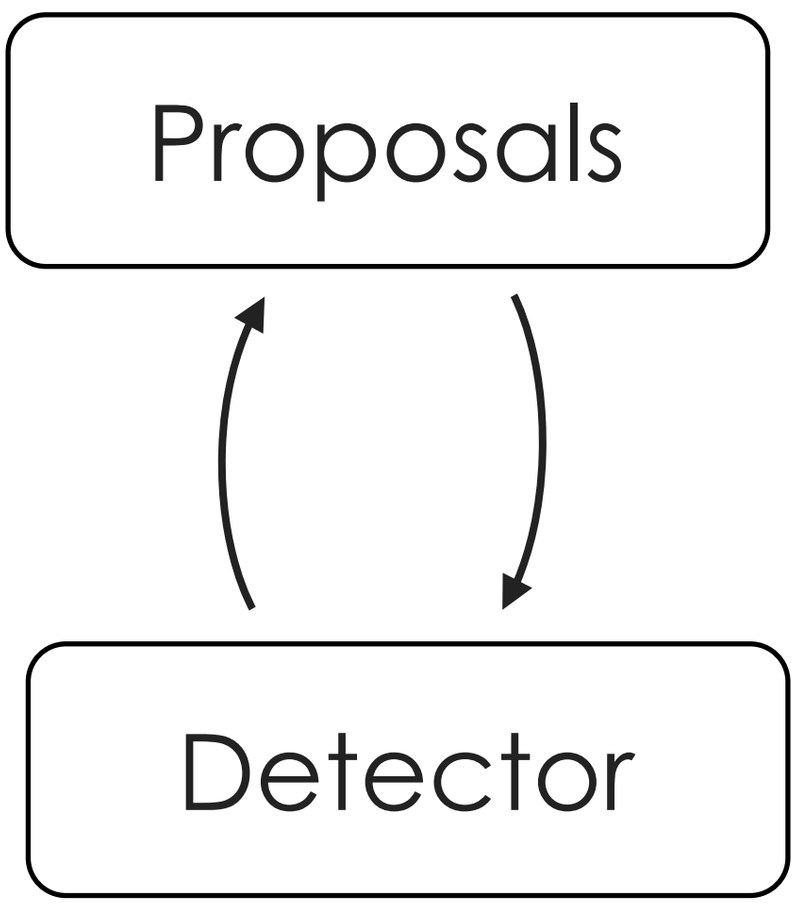
Model posterior

# Attentional Object Detector

Assume we have a powerful but expensive per-class classifier.

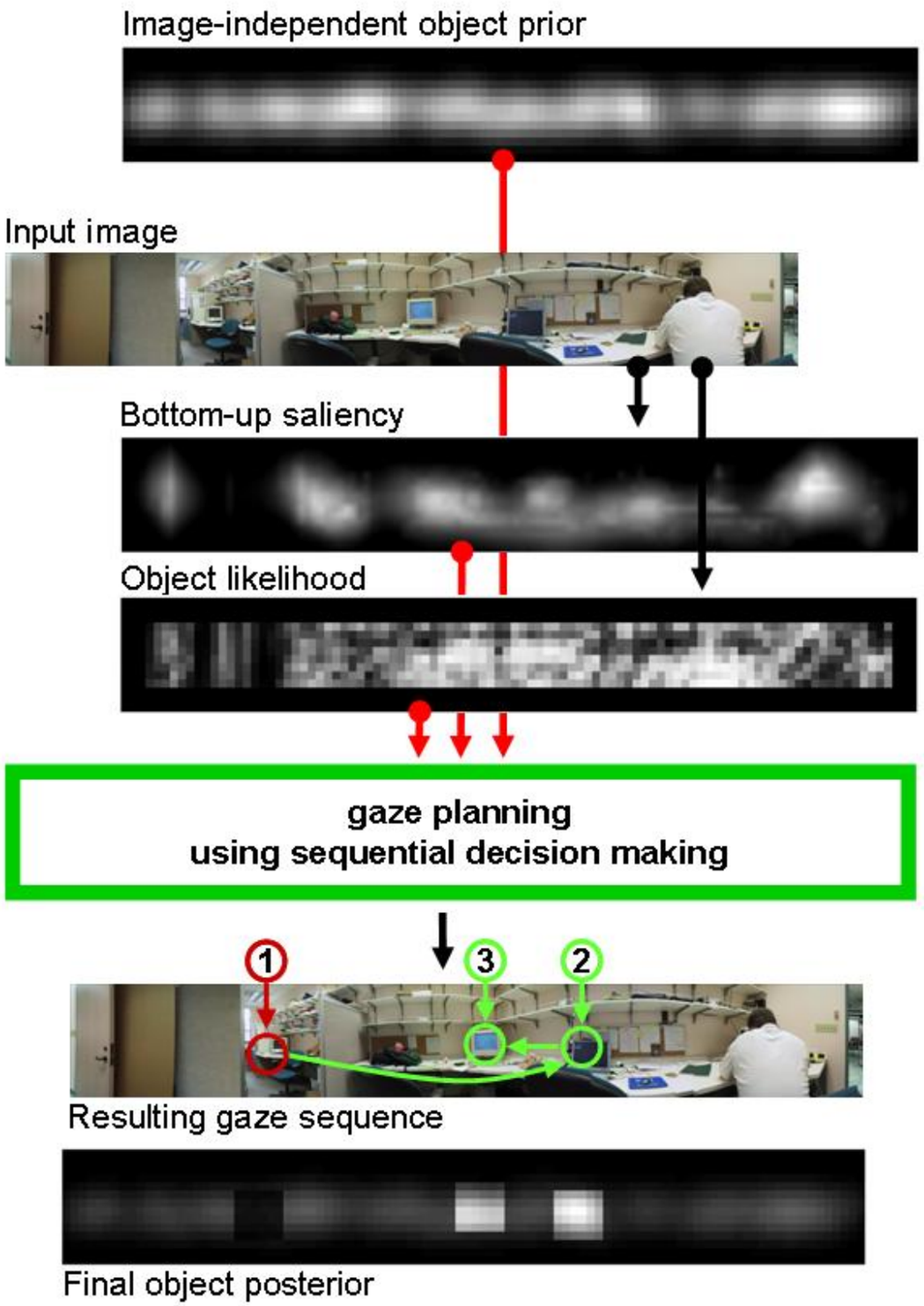
- How should we pick locations to consider?
- What should we look for at a location?

# Attentional Object Detector



# Some related work

# Vogel and Freitas. **Target-directed attention: Sequential decision-making for gaze planning**. ICRA 2008.

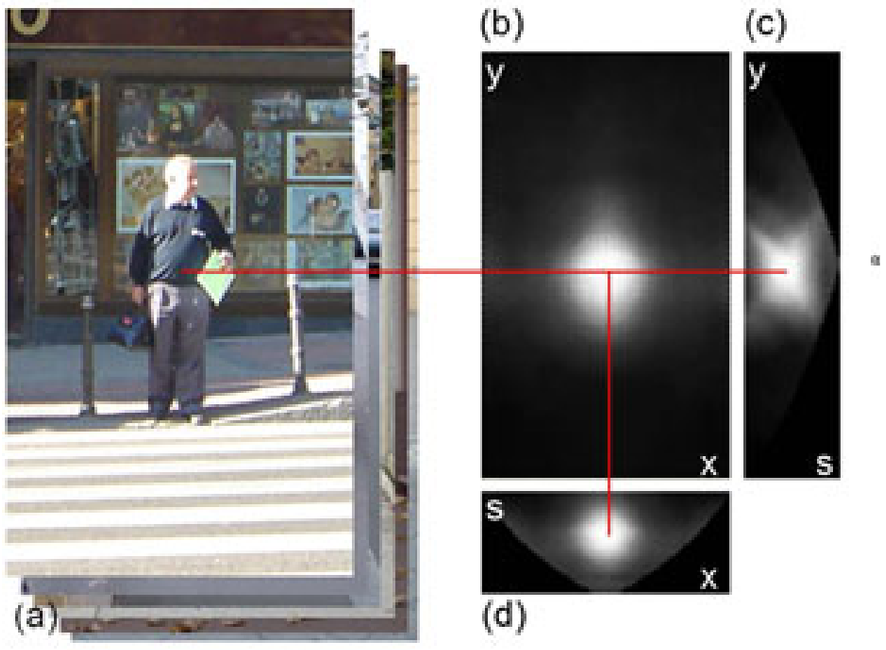



- GIST and a simple regressor to compute likelihood map.
- Reinforcement learning to find best gaze sequence.
- "Heavier" feature and regressor to evaluate the fixation locations.

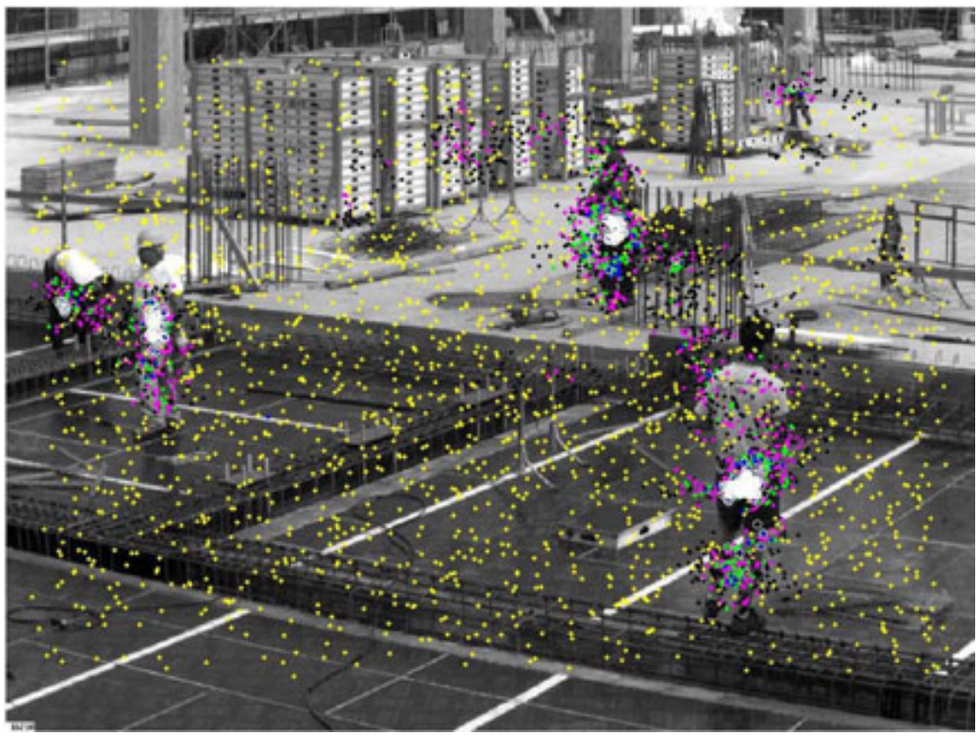
Vogel and Freitas. **Target-directed attention: Sequential decision-making for gaze planning**. ICRA 2008.

- Evaluated only on Caltech Office scenes.
- Gaze planning improves over just using bottom-up saliency while being only slightly slower.
- Detection rate is lower than full image, but maximum precision is higher.

# Gualdi et al. **Multi-stage Sampling with Boosting Cascades for Pedestrian Detection in Images and Videos**. ECCV 2010.




- LogitBoost classifier with covariance descriptors.
- Score falls off over some region of support.

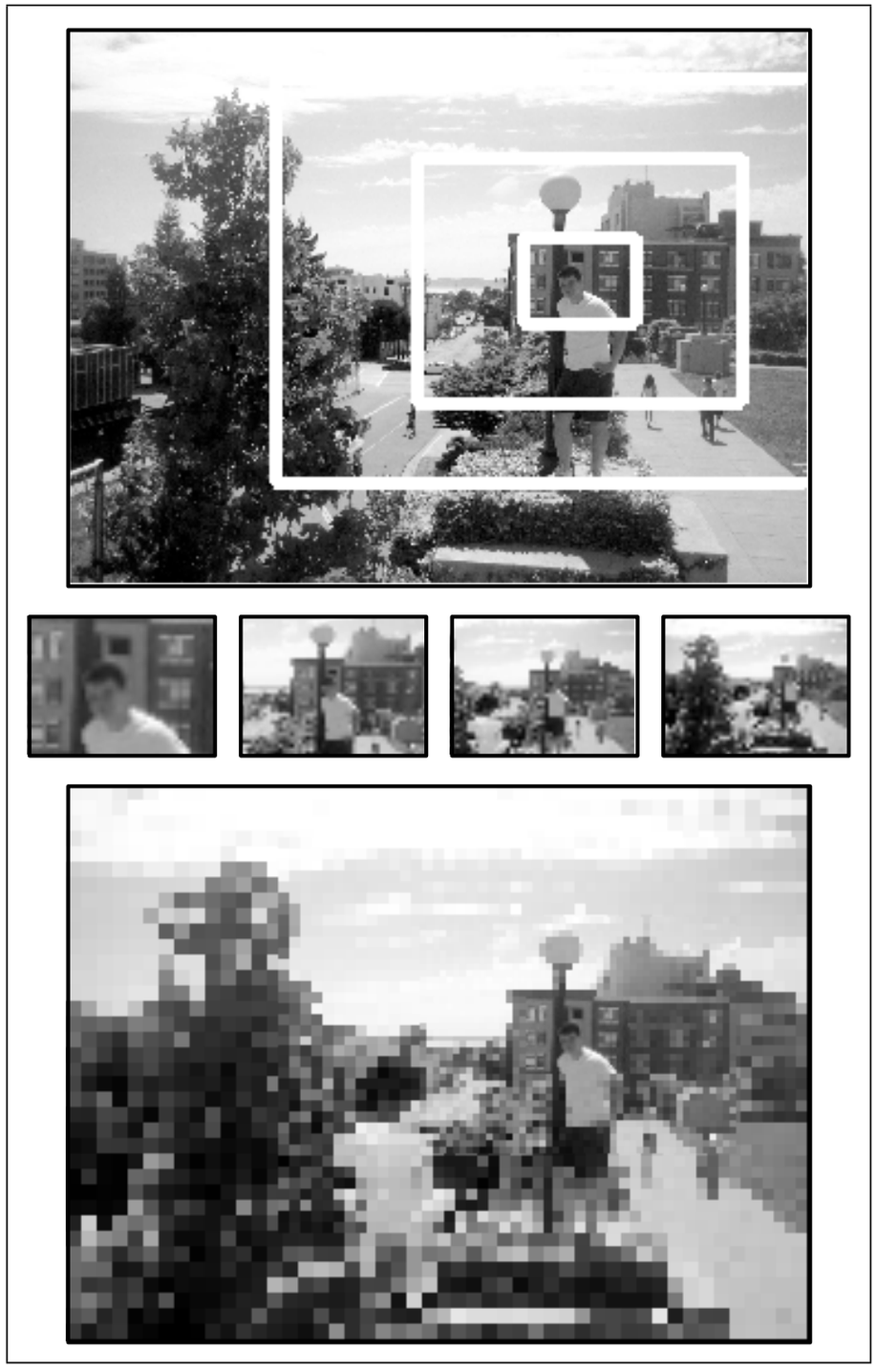


- Sample points in image to estimate *P(O | I)*. **Resample close to promising points.**
- Result: multimodal Gaussian mixture.

Gualdi et al. **Multi-stage Sampling with Boosting Cascades for Pedestrian Detection in Images and Videos**. ECCV 2010.

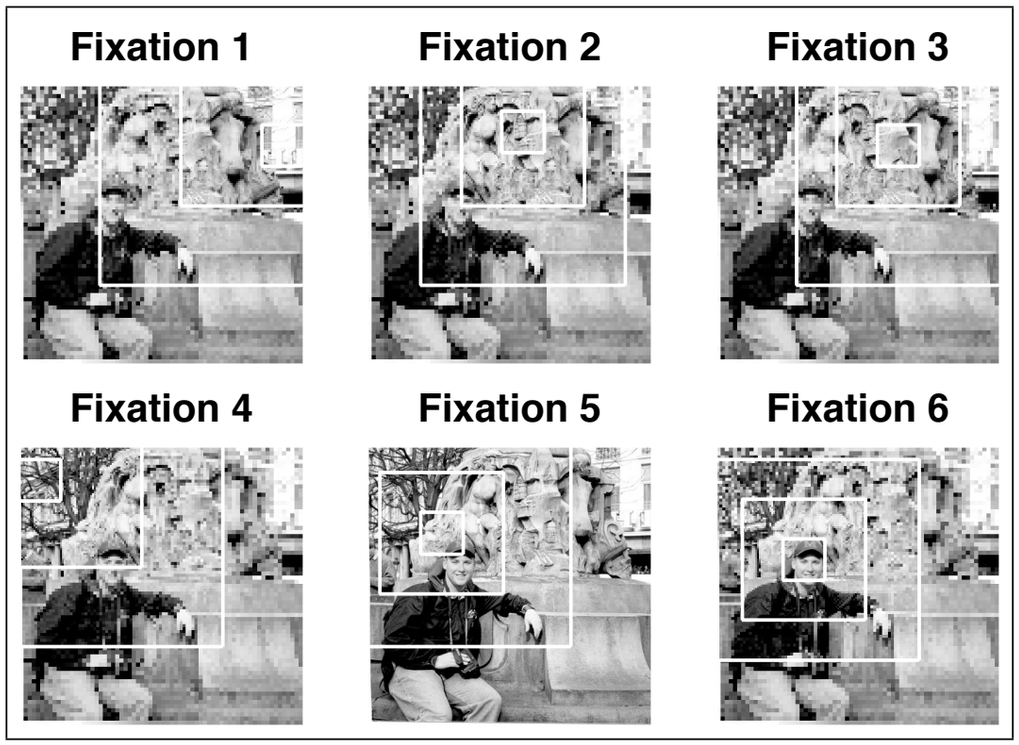
- Evaluated on INRIA Pedestrians, Graz02, and some videos.
- Always reduces miss rate over sliding window, while being 2-6x faster.

# Butko and Movellan. **Optimal Scanning for Faster Object Detection**. CVPR 2009.



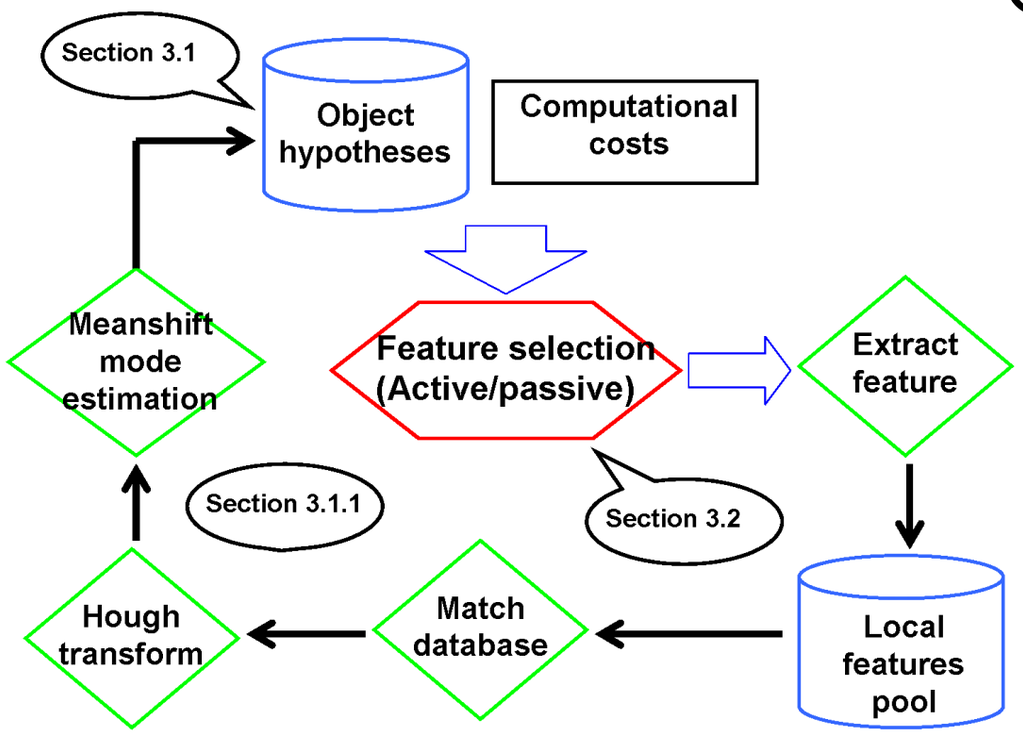
- Digital fovea placed sequentially to maximize expected information gain.
- Liken it to stochastic optimal control, and use a "multinomial infomax POMDP" to pick the sequence.

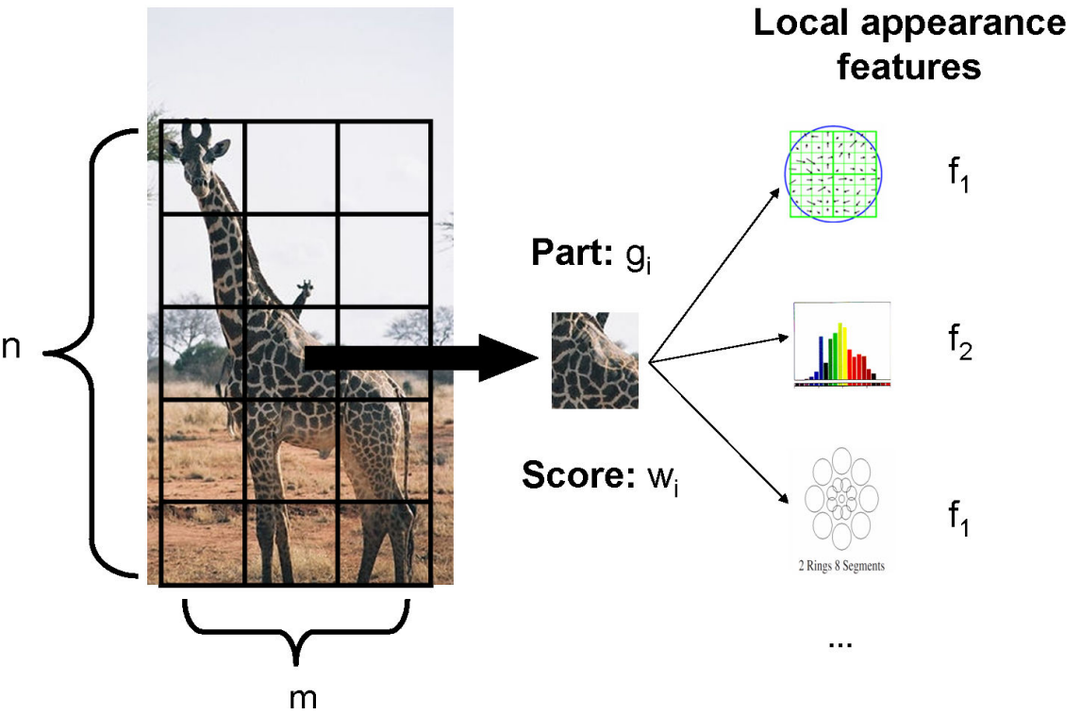
Butko and Movellan. **Optimal Scanning for Faster Object Detection**. CVPR 2009.




- Evaluate on own faces dataset against V-J. 2x speedup, but small decrease in accuracy.

# Vijayanarasimhan and Kapoor. **Visual Recognition and Detection Under Bounded Computational Resources**. CVPR 2010.




| Feature | Channel | Dim | Computation time (ms) |
|---|---|---|---|
| SIFT | R, G, B, Gray | 128 | 0.21 |
| P64_T1a_S2_17 | Gray | 68 | 1.2 |
| P18_T2_S2_9 | Gray | 36 | 0.09 |

- Hough voting with multiple feature types.
- Uses Value of Information to pick region to look at and the best feature to extract.
- Active approach extracts less features, takes less time, and has higher accuracy on ETHZ and Horses.

# Image Attributions

- Girschick et al. - Cascaded deformable part models.
- Viola & Jones - Rapid object detection.
- Judd et al. - Learning to predict where humans looks.
- Chikkerur et al. - What and where? A Bayesian theory of attention.
- ...and the papers reviewed.