Learning from Disagreements: Discriminative Performance Evaluation

Christina Pavlopoulou, David Martin, Stella Yu, Hao Jiang

Boston College

PETS 09

Past Work

- Reliable evaluation requires ground truth.
- Precision-recall curves, ROC curves, statistical significance testing…

Contribution

- Evaluate algorithms on test cases they perform differently.
- Test case selection.

Agenda

- Motivation
- Method
- Case Study
- Extensions
- Conclusions

Motivation

- Algorithm A: 70%
- Algorithm B: 75%
- How significant is the difference?

Statistical Significance

- Performance measures become population samples
- Hypothesis testing: do the populations have the same means?

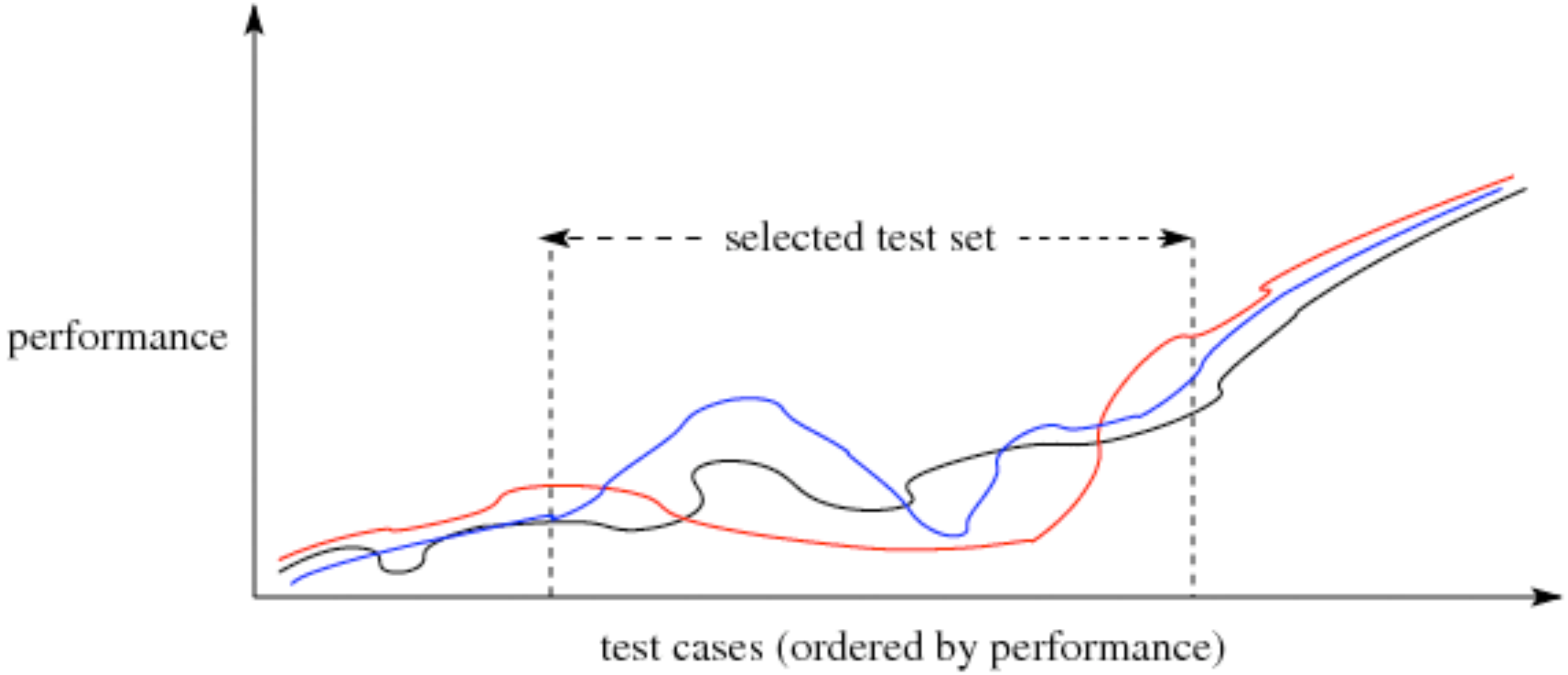
Problem

- If methods have similar performances on many cases, statistical tests are not powerful.

↓

- Select test cases
    - not too difficult
    - not too easy

Discriminative Performance Evaluation

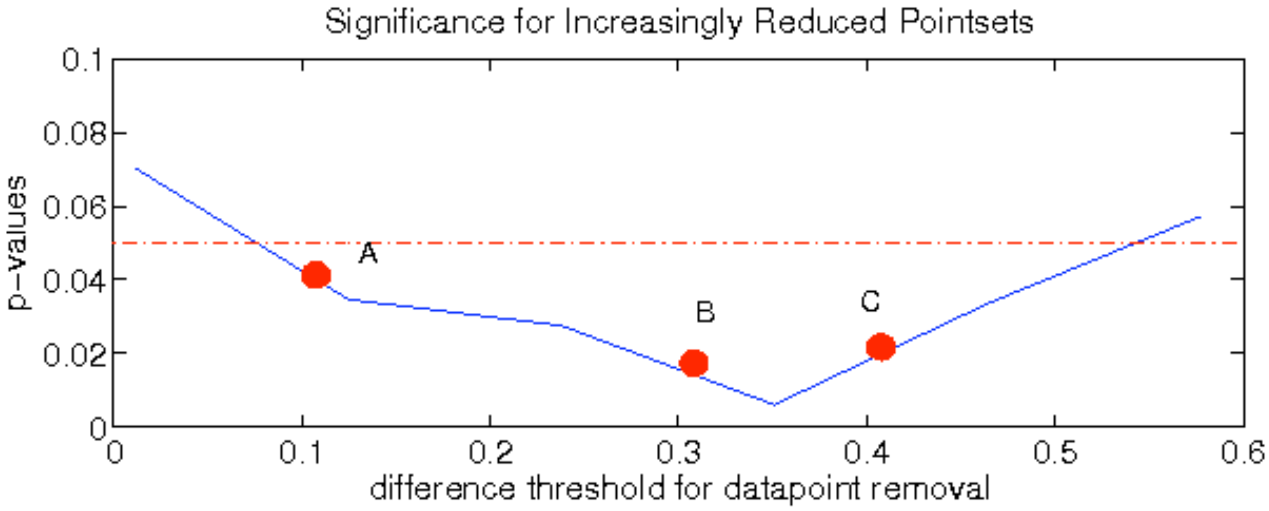


Method

- for a sequence of increasing thresholds
  - remove test cases with *performance similarity* below threshold
  - compute p-value

Decision Rule

- There is t_0 for which $p_0 < a$
- Subsequent p values remain under confidence value.
- t_0 is small.



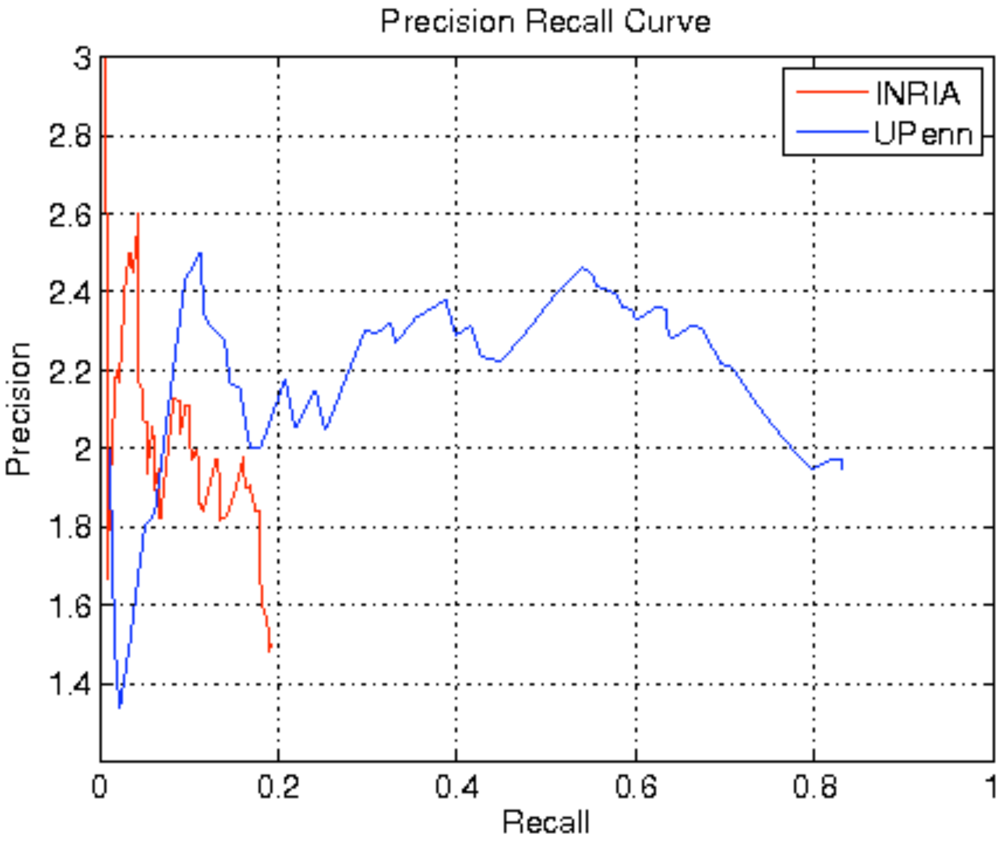
Case Study

- Compared two object detection methods.
- INRIA: features + SVM (appearance)
- Penn: shape context + matching (geometric)
- Test set: 78 images (PASCAL 2005)

Performance Similarity Measure

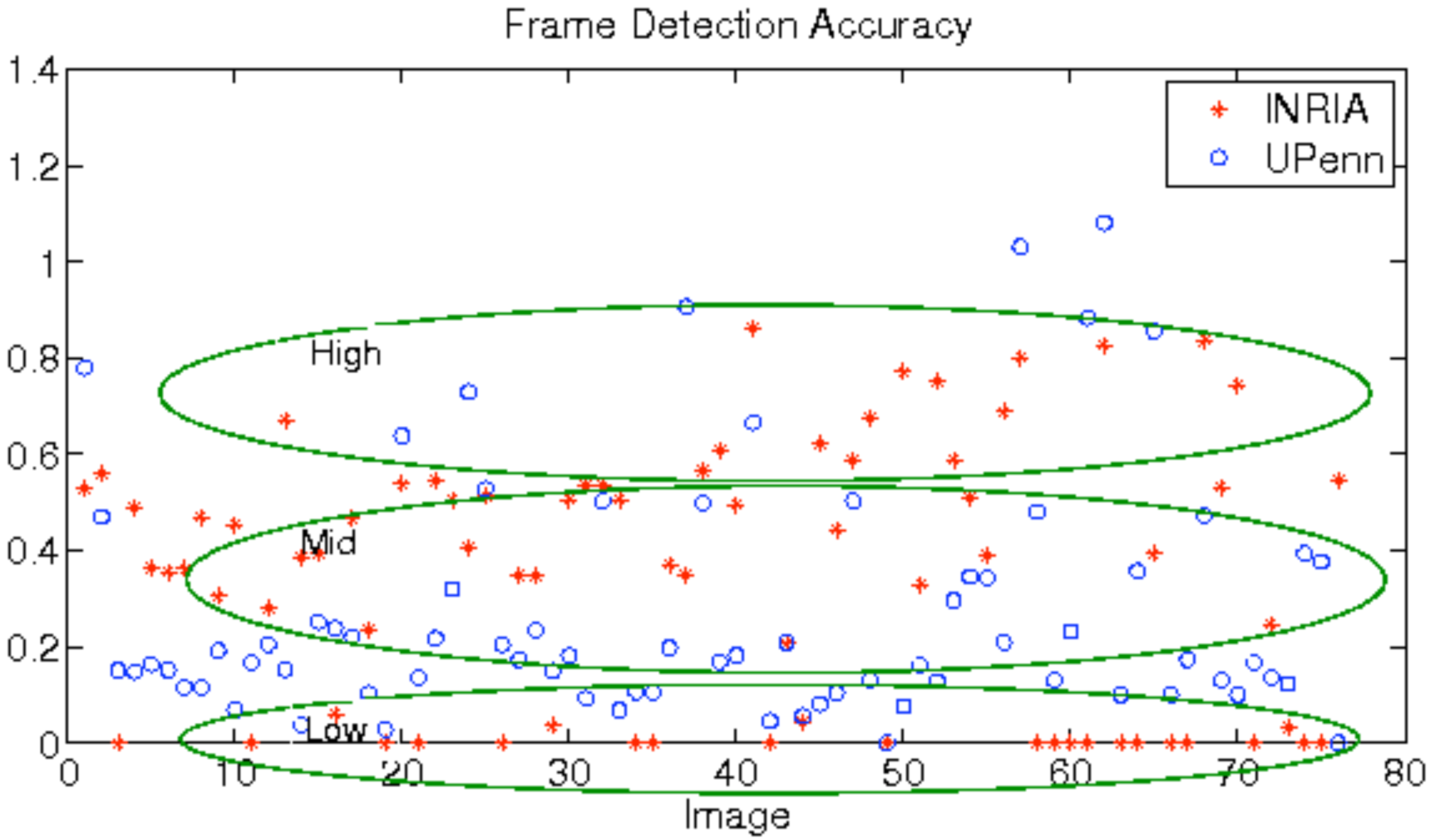
Area of overlap between detected person and ground truth.

Common Measures

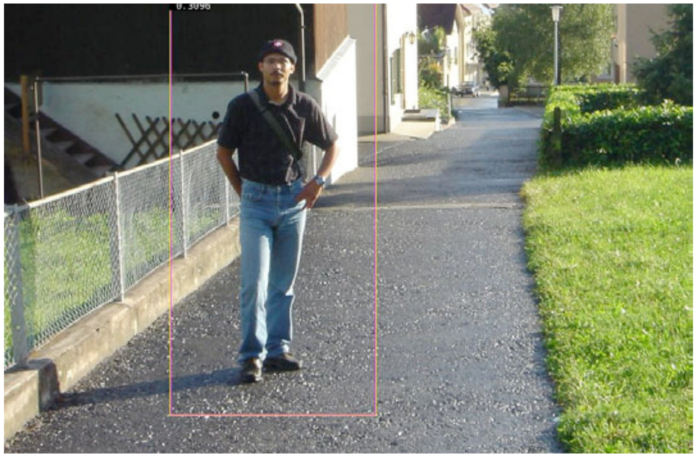


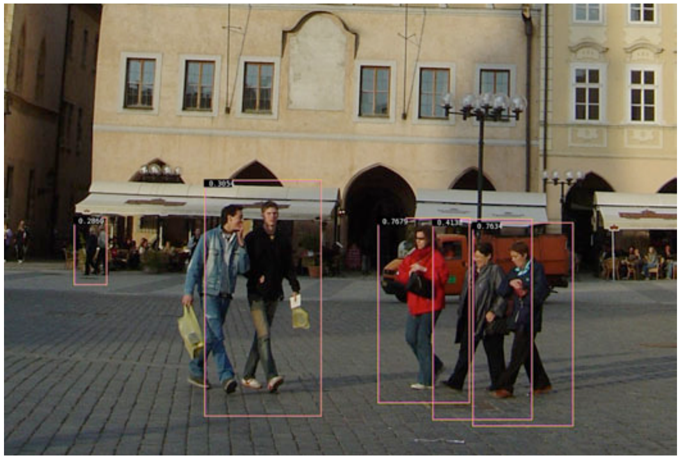
Avg. correctness:

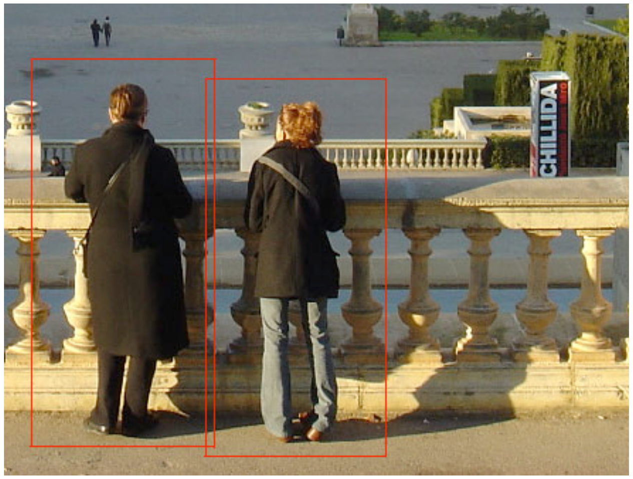
INRIA: 35%
Penn: 27%


Frame Detection Accuracy
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Image
INRIA
UPenn
High
Mid
Low

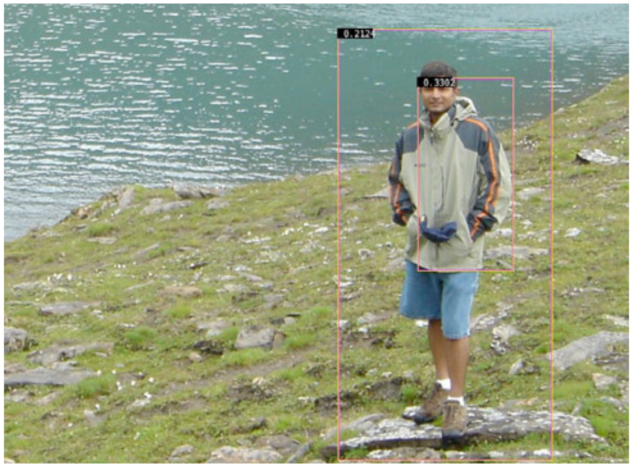
High Accuracy

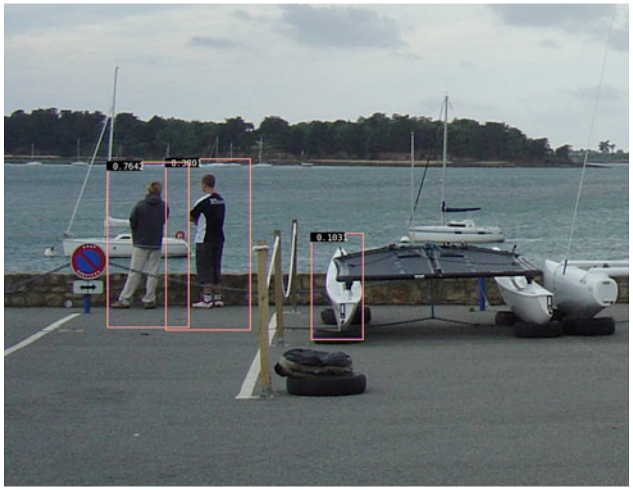


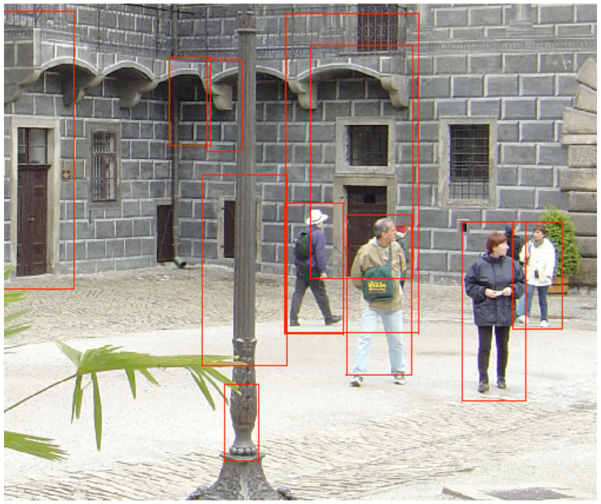


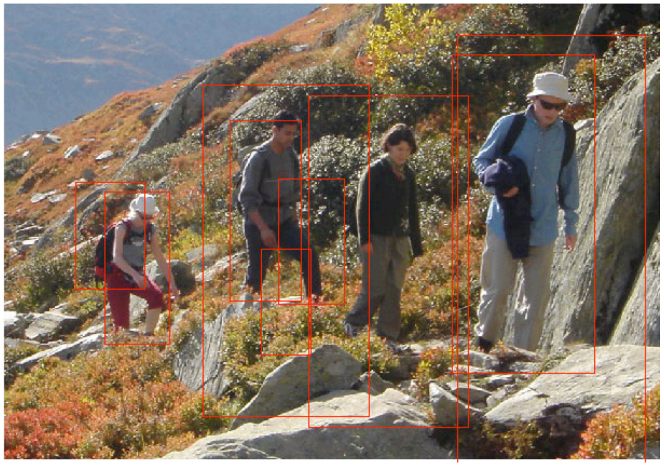


Medium Accuracy

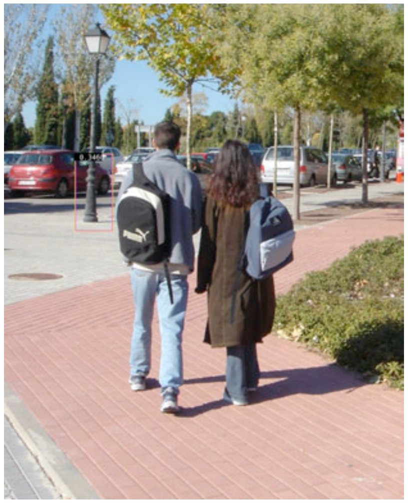


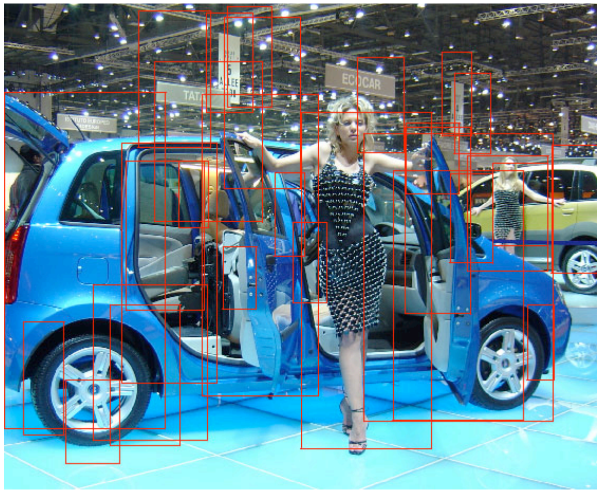


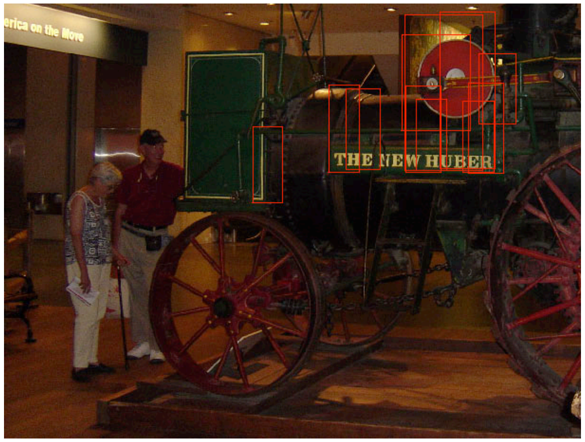


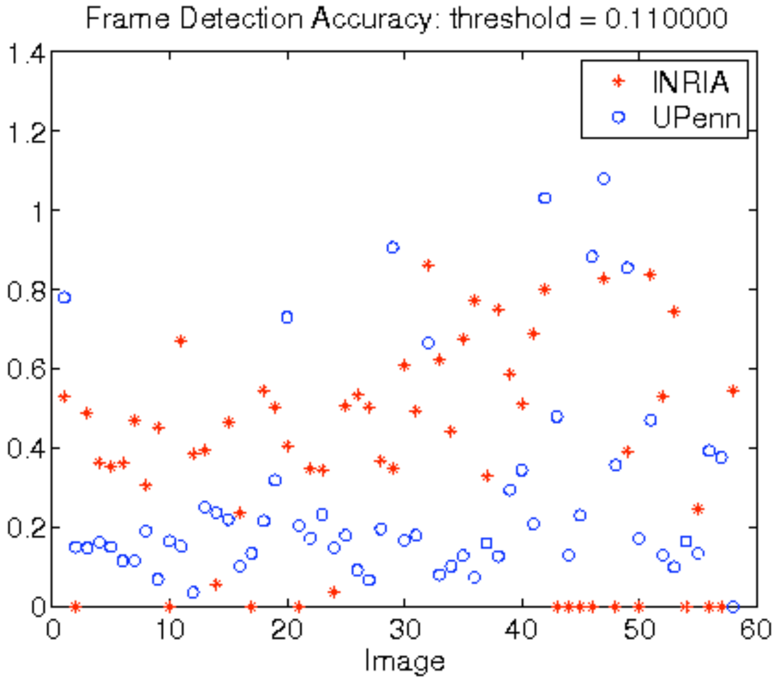


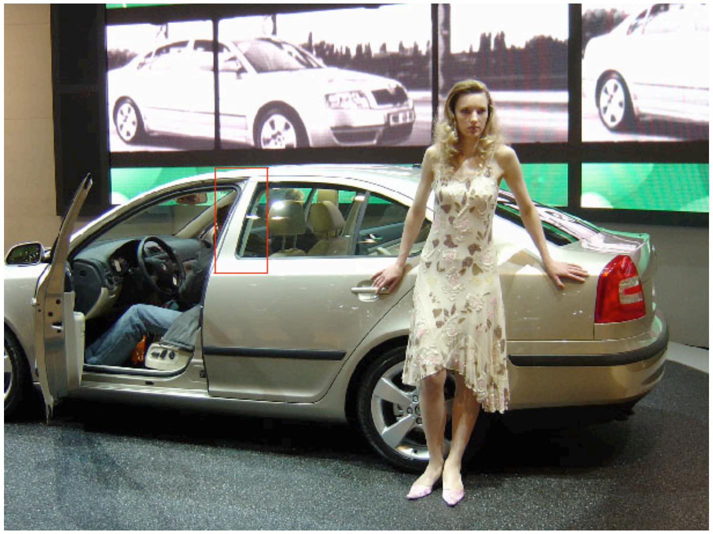
Low Accuracy

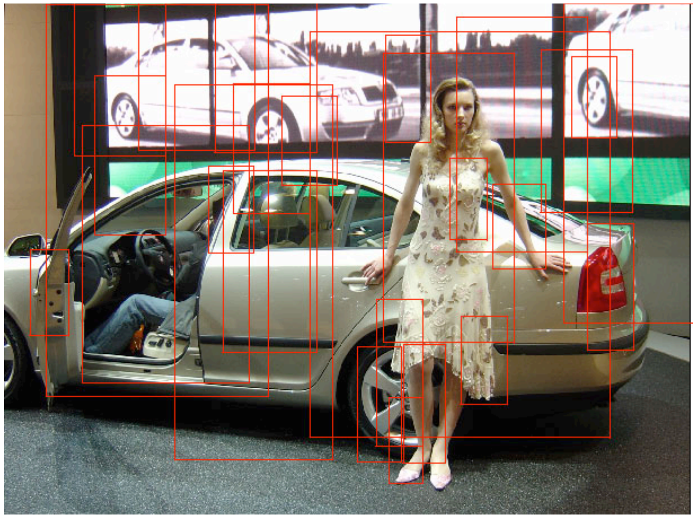


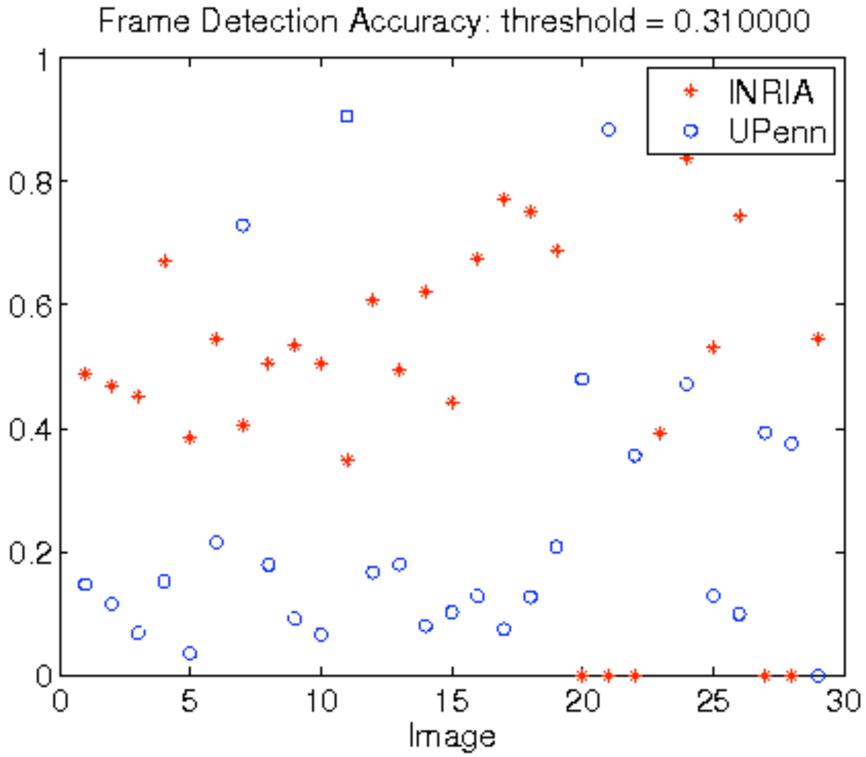


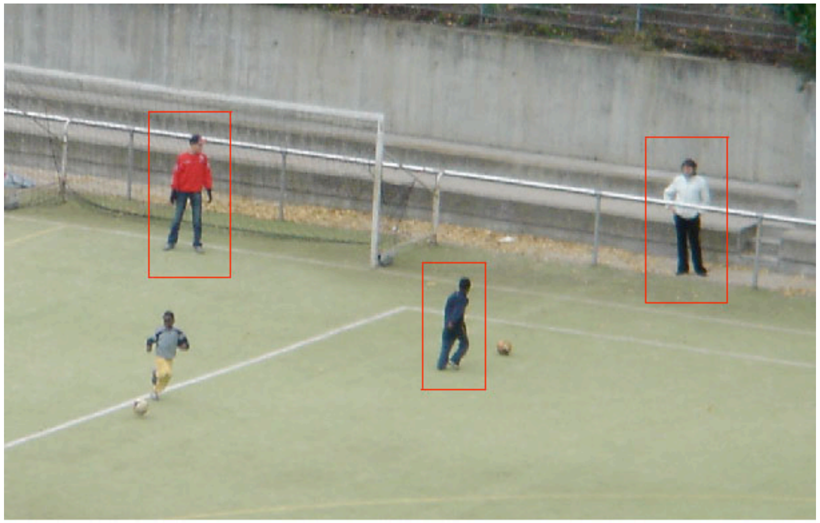


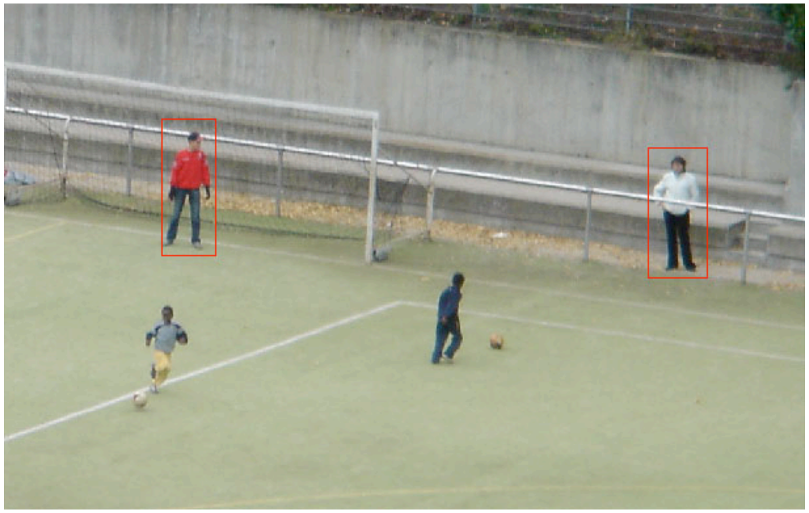

Frame Detection Accuracy: threshold = 0.110000
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
Image
INRIA
UPenn

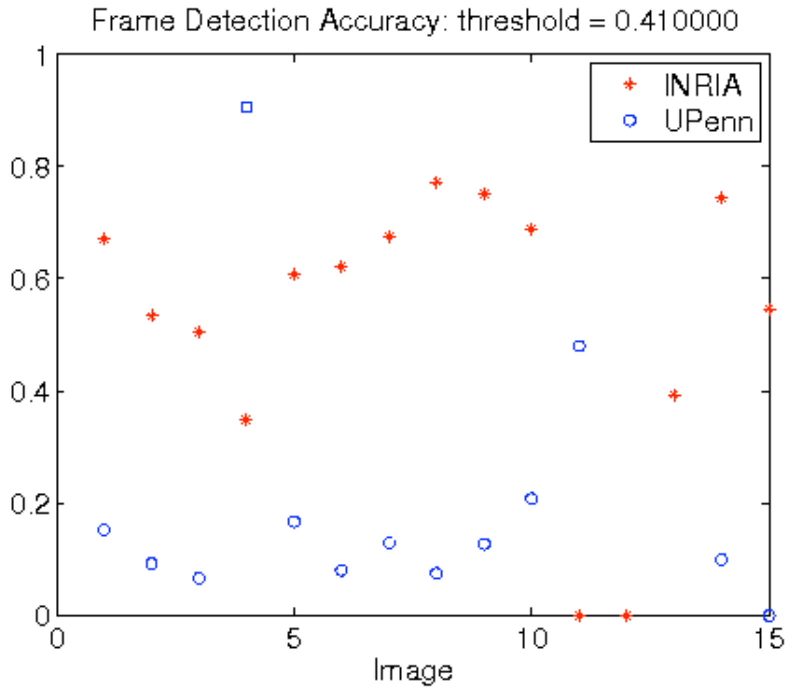


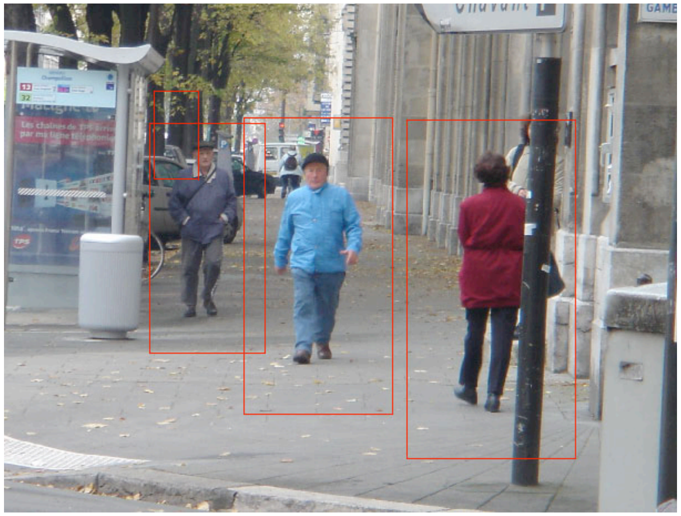


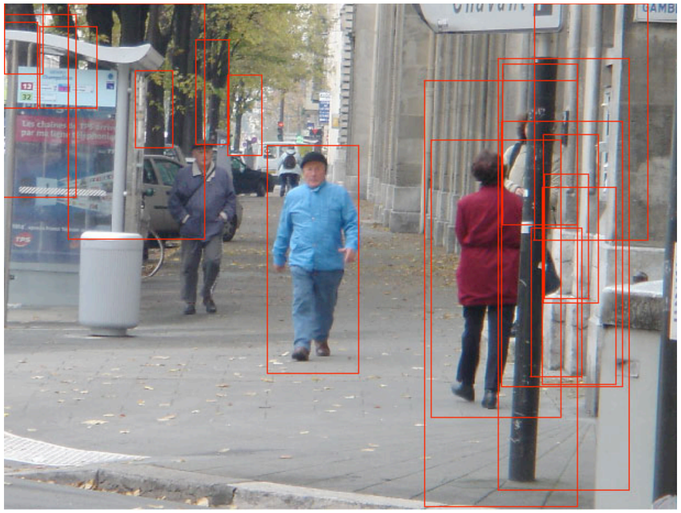

Frame Detection Accuracy: threshold = 0.310000
INRIA
UPenn
1
0.8
0.6
0.4
0.2
0
0
5
10
15
20
25
30
Image

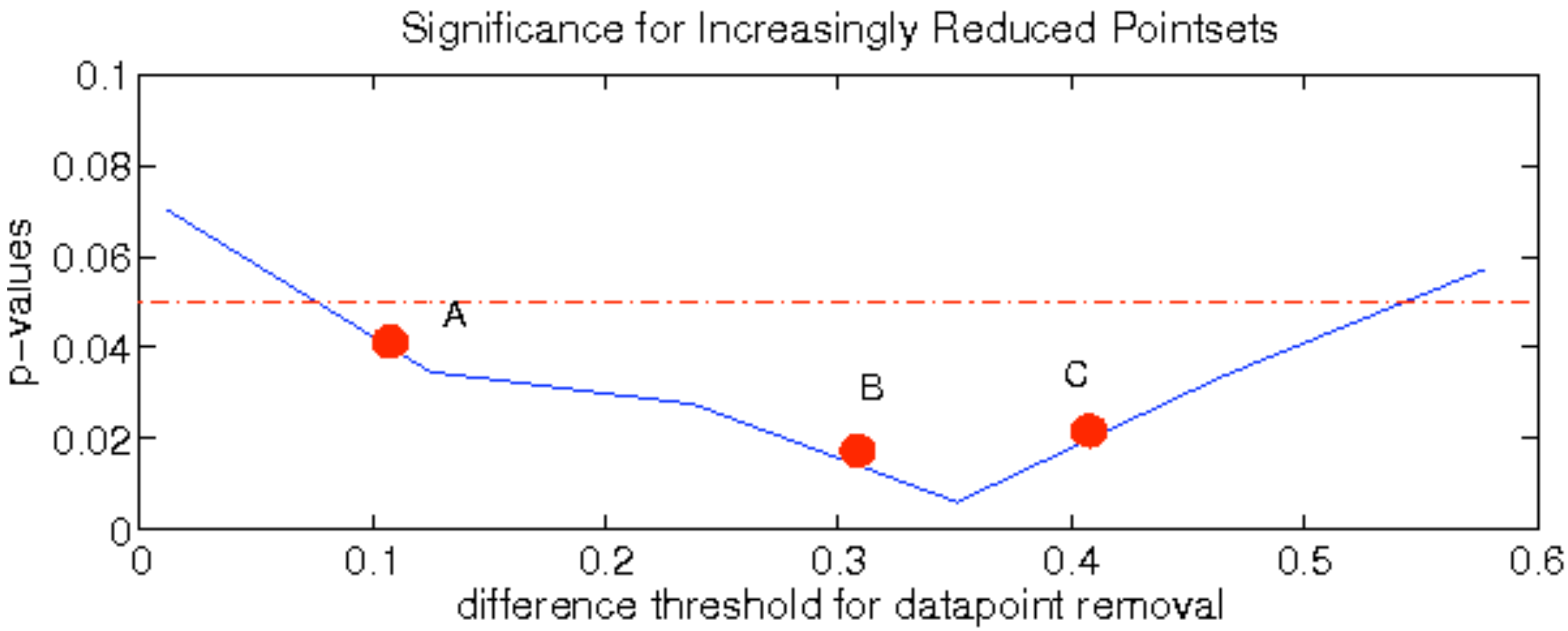





Frame Detection Accuracy: threshold = 0.410000
INRIA
UPenn
1
0.8
0.6
0.4
0.2
0
0
5
10
15
Image





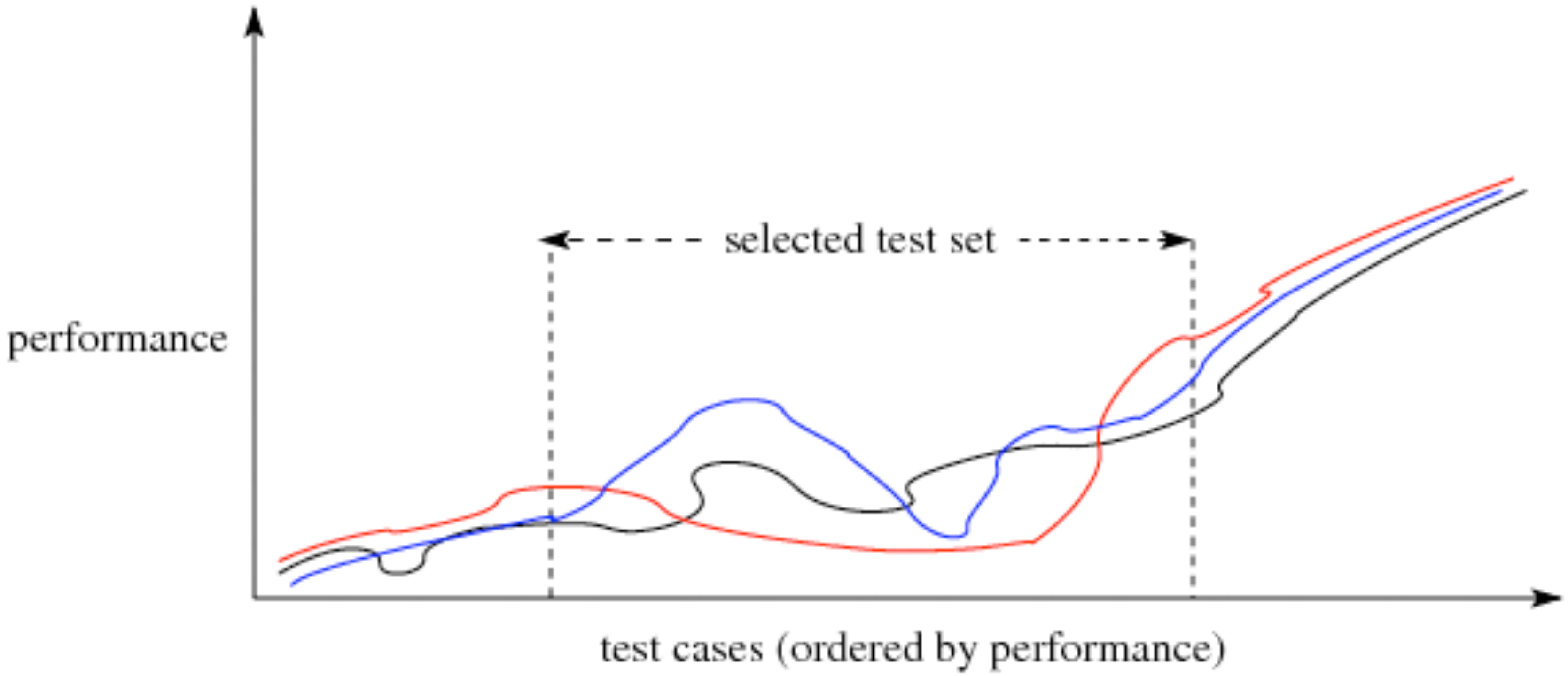

Significance for Increasingly Reduced Pointsets
p−values
difference threshold for datapoint removal
0.1
0.08
0.06
0.04
0.02
0
0
0.1
0.2
0.3
0.4
0.5
0.6
A
B
C

What about t small?

- Should be human-determined
  - – Are the results below t similar enough?

When ground truth is needed?

- Similarity directly among methods' results (omit ground truth)
- Provide ground truth for dissimilar results.


selected test set
performance
test cases (ordered by performance)

Conclusions

- Evaluation based on dissimilar results
- Decision framework for assessing statistical significance
- Selection of ground-truth data