Imperial College London

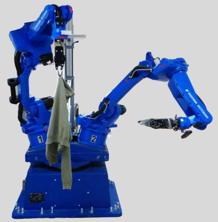
Active and Interactive Vision

T-K Kim
Computer Vision and Learning Lab
EEE, ICL
http://www.iis.ee.ic.ac.uk/ComputerVision/

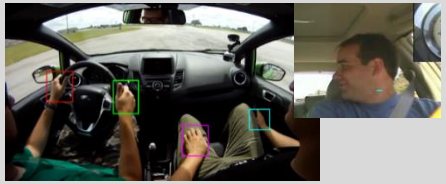
Imperial College London

- We have tackled pose estimation of
  – Hands, Face, Body as structured label estimation problems
  – 6D Object Pose
- Active and interactive Vision
  – Interaction among Human-Computer-Object

Pham et al. CVPR’15



Kim et al. ICRA14/ECCV14

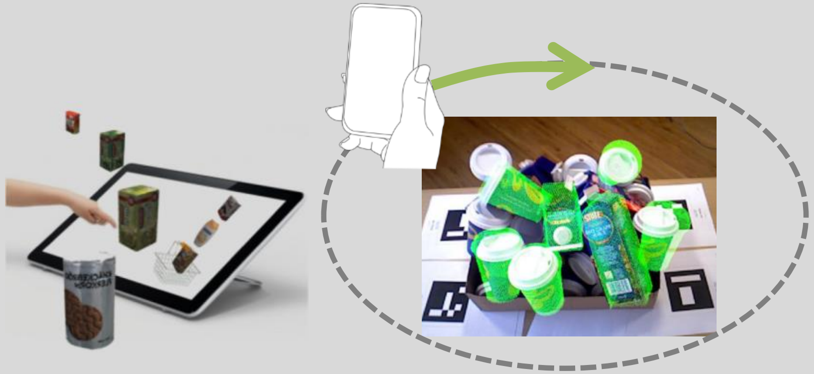


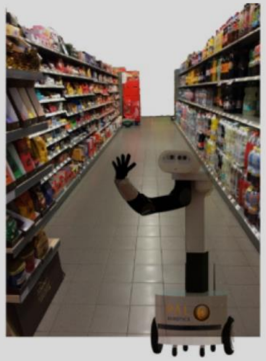
Trivedi et al. UCSD

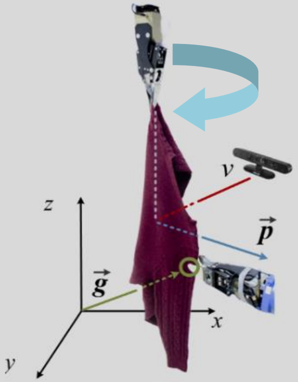
Object Pose and Next-Best-View Estimation

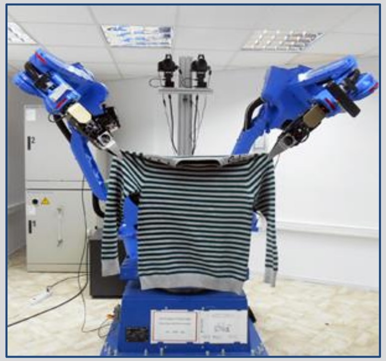
- **Problem** - estimating objects’ 3D location and pose
- **Application** - e.g. picking and placing for logistics
- **Challenge** - highly crowded scenes, active camera planning

- **Problem** - estimating clothes types, grasp points and pose
- **Application** - autonomously unfolding clothes
- **Challenge** - highly deformed objects, multi-view solution, active planning









Object Pose and Next-Best-View Estimation

- Estimating objects’ 3D location and pose

Latent Hough Forest (ECCV14) :

- novel template-matching based splitting, one-class learning

6D Object Detection and Next-Best-View Prediction in the Crowd (ongoing):

- deep-features, a novel active solution on Hough Forests, joint registration

- Estimating clothes types, grasp points and pose

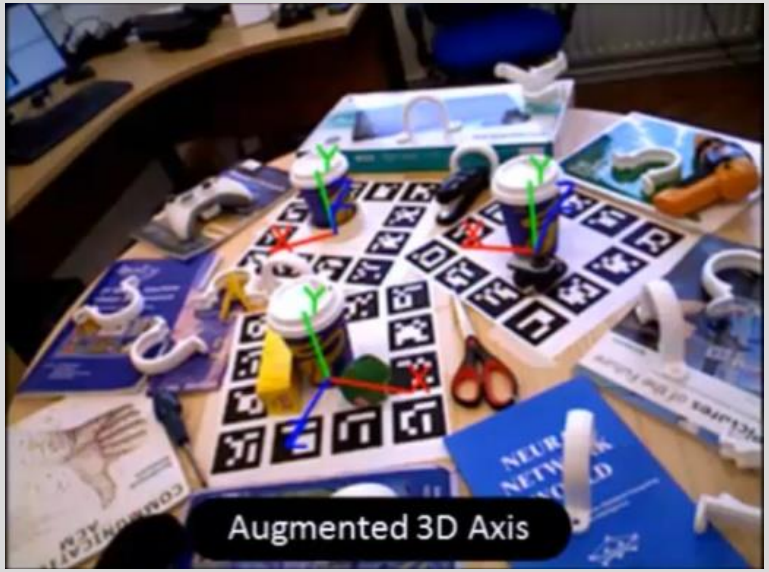
Autonomous unfolding clothes (ICRA14, best paper award):

- regression forests, probabilistic active planning

Active Forest (ECCV14) :

- multi-task learning, next-best view learning in RF

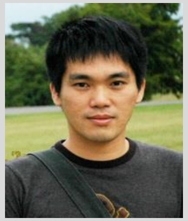
Imperial College London



Latent-Class Hough Forests for 3D Object Detection and Pose Estimation


Alykhan Tejani


Rigas Kouskouridas

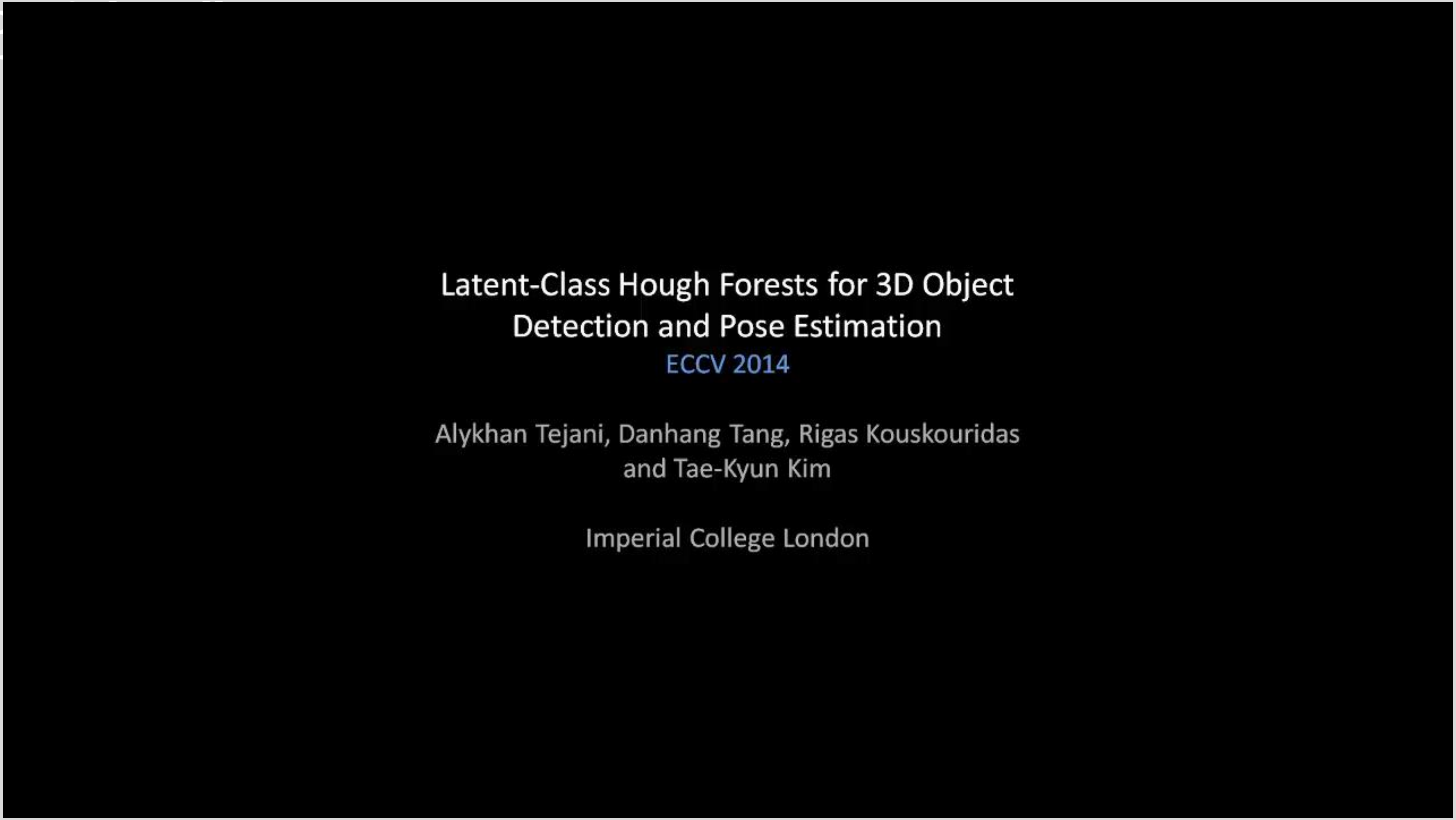

Danhang Tang


Andreas Doumanoglou


T-K Kim

Imperial College London

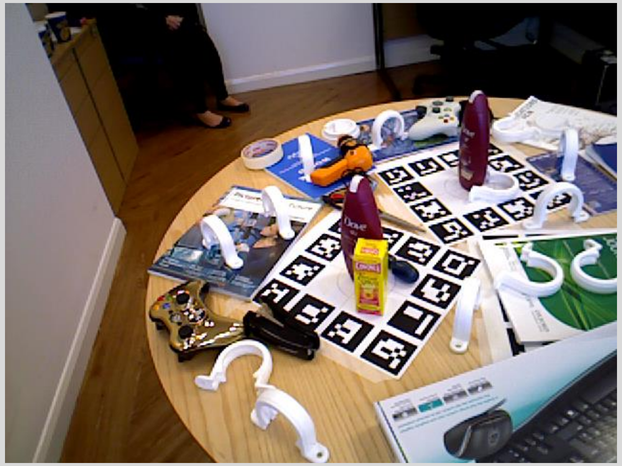
ECCV 2014



https://www.youtube.com/watch?v=idY3Q7wg5rk

Challenges and Proposed Ideas

- Challenges
  - Foreground occlusions, Multi-instances, Large scale changes



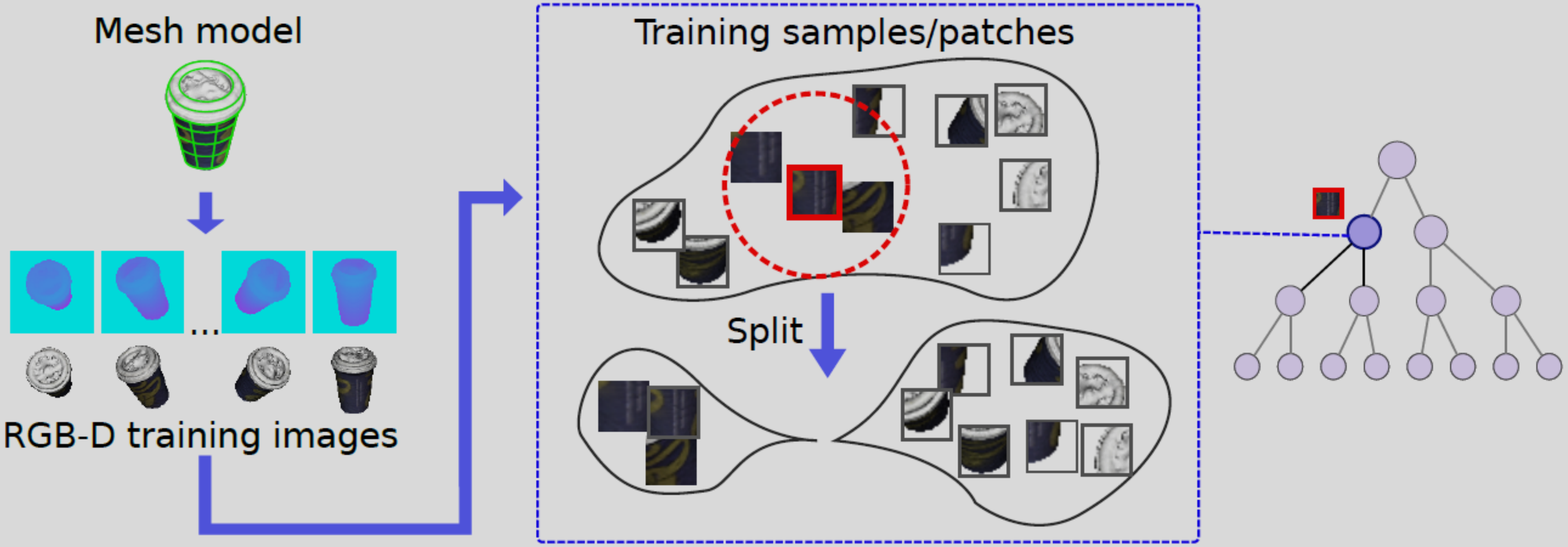

- Main ideas
  - Integration of LINEMOD *(S. Hinterstoisser, et al. PAMI12)* Template Matching into Hough Forests *(J. Gall, et al. PAMI11)* : Efficient data split at node levels
  - Making LINEMOD scale-invariant
  - Inference of occlusion masks: Iteratively updating class distributions (latent variable, one-class learning)

Template-matching Split Functions

- A random patch T (with red frame) is chosen. All other patches are compared with T.

$$S(\mathcal{X}, \mathbb{T}) = \sum_{r \in \mathcal{P}} g(\mathbf{ori}(\mathcal{X}, r), \mathbf{ori}(\mathcal{O}, r))$$

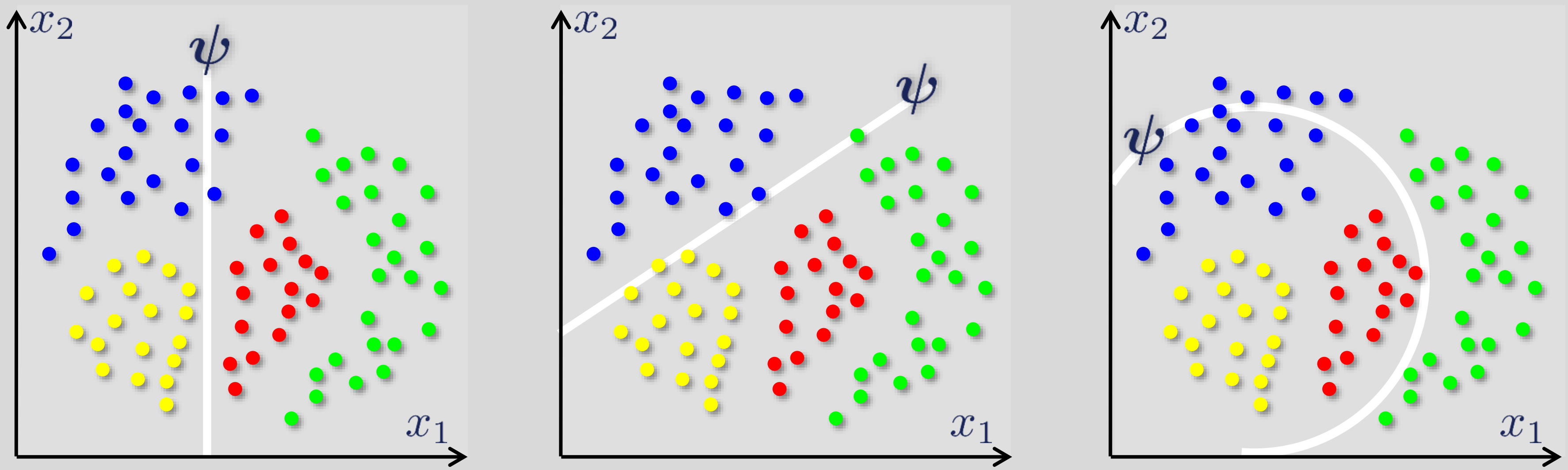
- They go to e.g. a right child node if the similarity is greater than a threshold, otherwise to a left child node.
- This achieves more discriminative (nonlinear) yet fast splits.



Imperial College London

Split function model in Decision Forests

Examples of split functions



Split ftn: axis aligned

Split ftn: oriented line

Split ftn: conic section

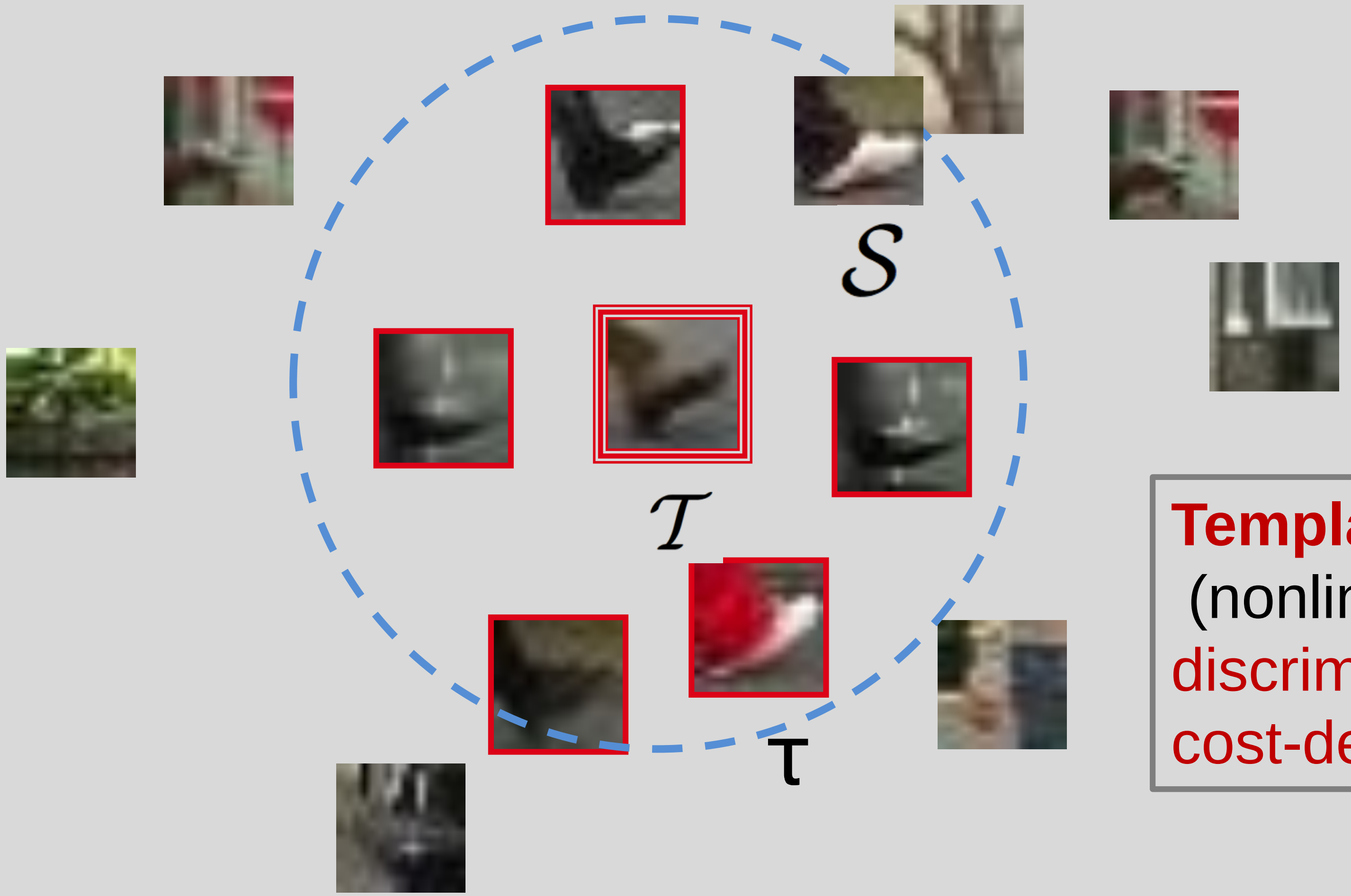
Slide credit to Criminisi and Shotton, ICCV11 tutorial

Template-matching Split Function

2 pixel test
(axis aligned splits):
efficient but less discriminative

Examples from Pedestrian Detection

Template-matching Split Function

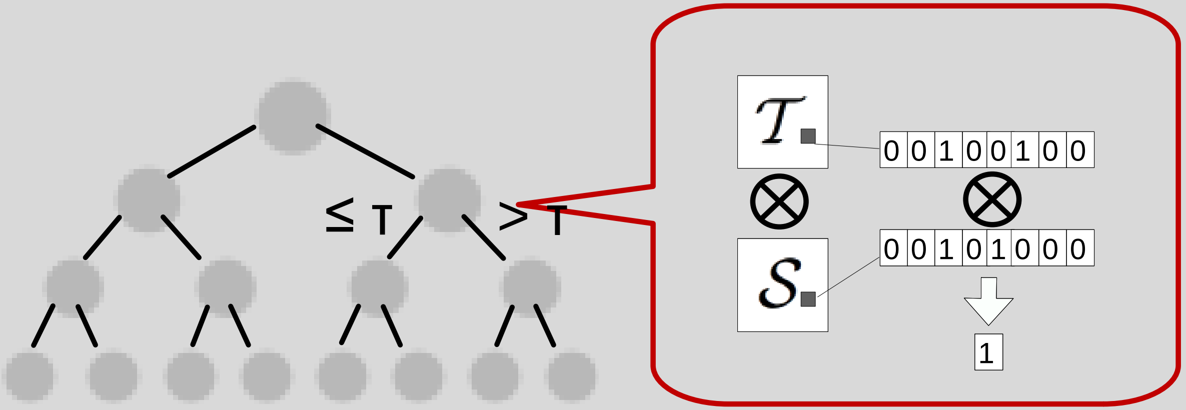


Template matching (nonlinear splits): discriminative but cost-demanding?

Examples from Pedestrian Detection

Imperial College London

Template-matching Split Function using Binary Bit Operations



$$F(\mathcal{S},\mathcal{T}) = \sum_{\substack{P_d^{\mathcal{S}} \in \mathcal{S} \\ P_d^{\mathcal{T}} \in \mathcal{T}}} \delta(P_d^{\mathcal{S}} \otimes P_d^{\mathcal{T}} \neq 0), d = 1, \ldots, n$$

$$h_i(\mathcal{S}) = \begin{cases} 0, & F(\mathcal{S},\mathcal{T}_i) \leq \tau_i \\ 1, & F(\mathcal{S},\mathcal{T}_i) > \tau_i \end{cases}$$

Template matching split is highly accelerated by binary bit operations.

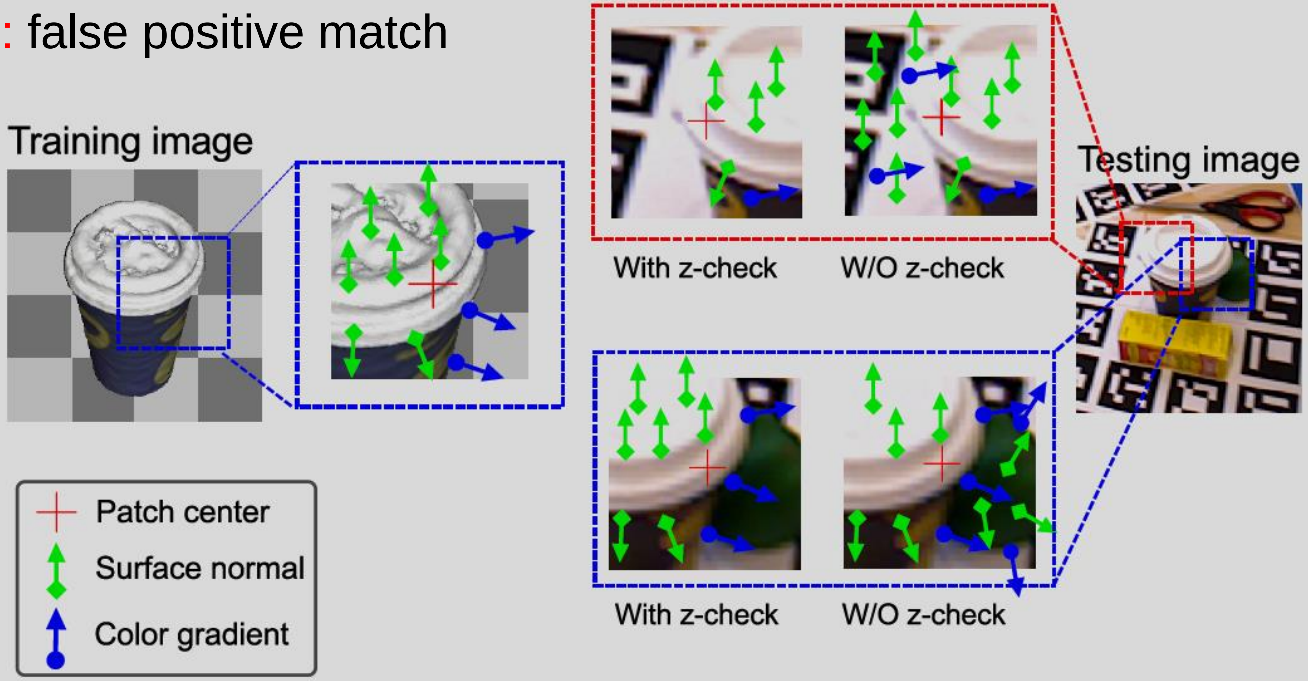
Split Function Properties

- The split function with an efficient z-value check:

$$\begin{cases} S(\mathcal{X},\mathbb{T}) & = \sum_{r\in\mathcal{P}} f(\mathcal{X},\mathcal{O},c,r)\,g(\mathbf{ori}(\mathcal{X},r),\mathbf{ori}(\mathcal{O},r)), \\ f(\mathcal{X},\mathcal{O},c,r) & = \delta(|(D(\mathcal{X},c)-D(\mathcal{X},r))-(D(\mathcal{O},c)-D(\mathcal{O},r))|<\tau) \end{cases}$$

Blue patch: true positive match

Red patch: false positive match

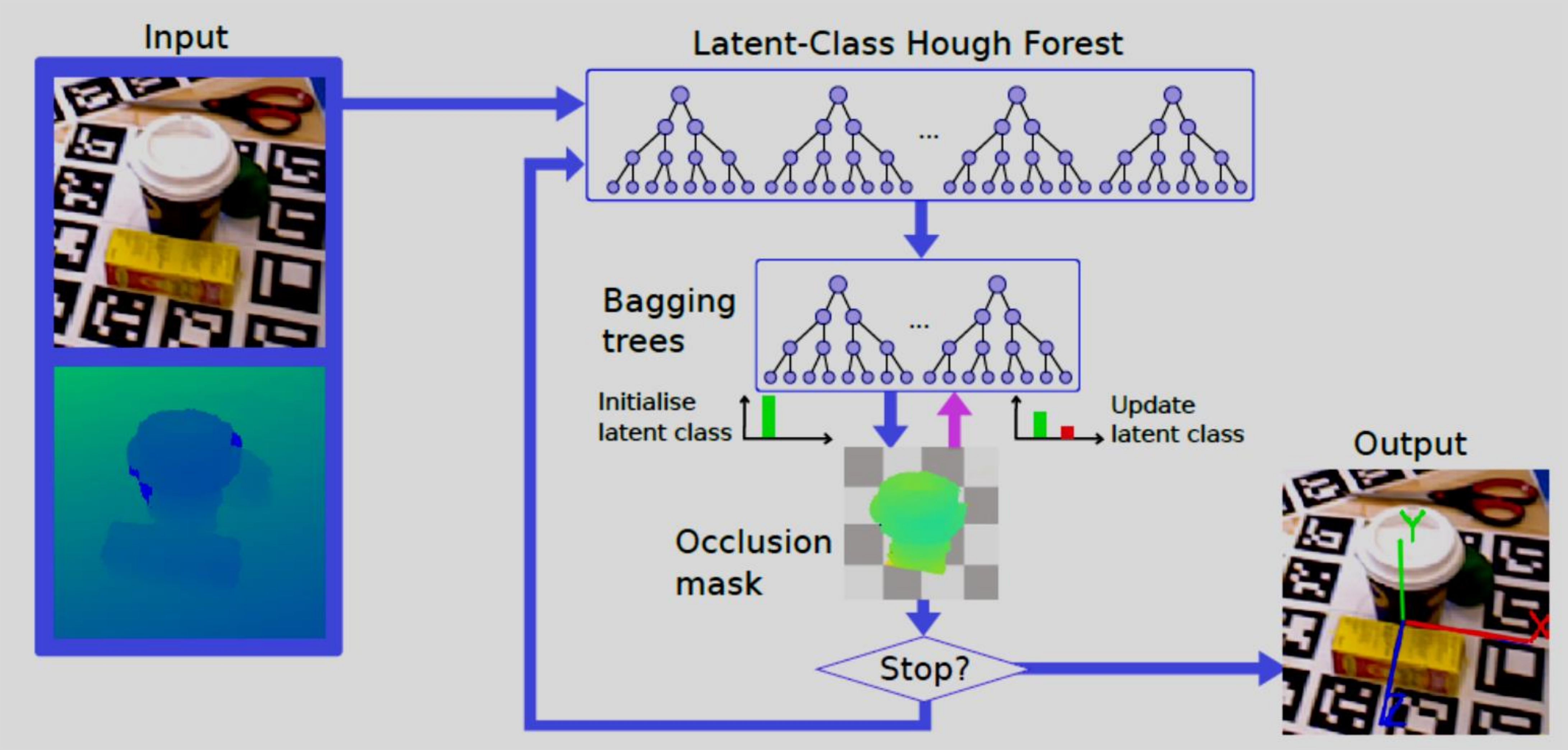


- Scale invariance:

$$\begin{cases} S(\mathcal{X},\mathbb{T}) & = \sum_{r\in\mathcal{P}} f(\mathcal{X},\mathcal{O},c,r)\,g(\mathbf{ori}(\mathcal{X},\frac{r}{D(\mathcal{X},c)}),\mathbf{ori}(\mathcal{O},\frac{r}{D(\mathcal{O},c)})), \\ f(\mathcal{X},\mathcal{O},c,r) & = \delta(|(D(\mathcal{X},c)-D(\mathcal{X},\frac{r}{D(\mathcal{X},c)}))-(D(\mathcal{O},c)-D(\mathcal{O},\frac{r}{D(\mathcal{O},c)}))|<\tau) \end{cases}$$

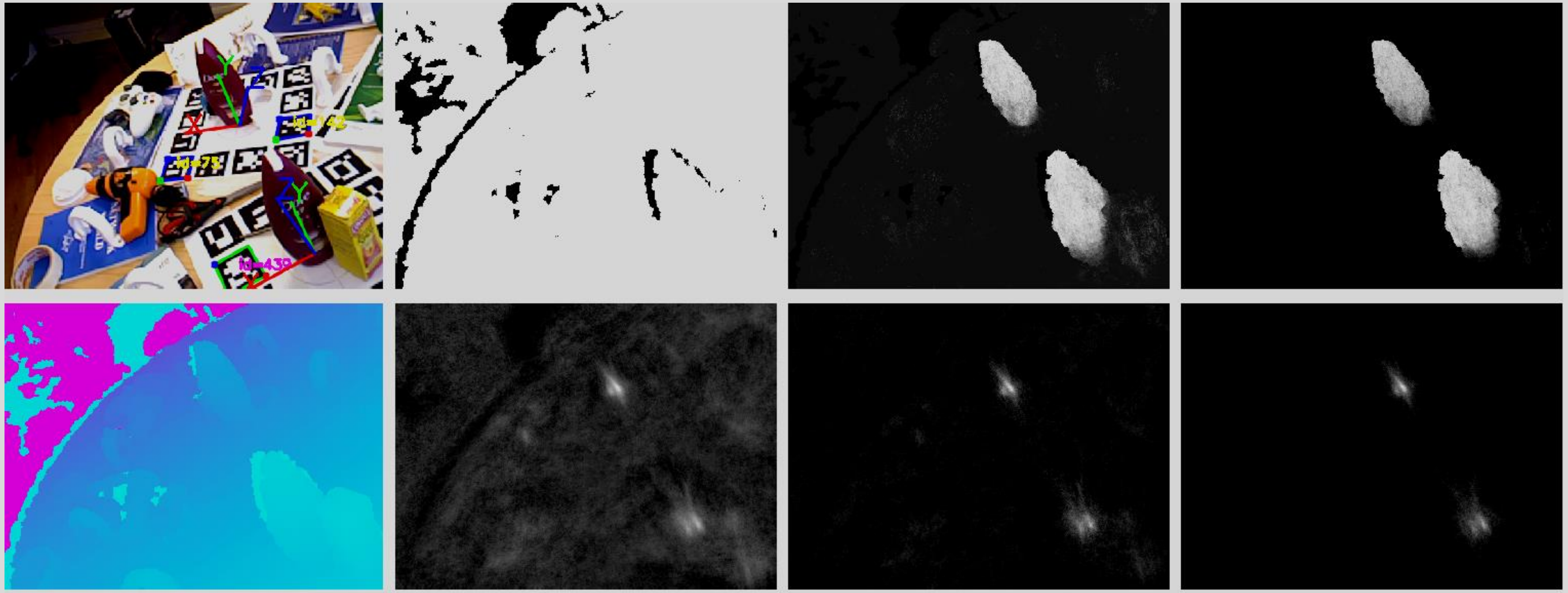
Imperial College London

Inference with Iterative Refinement



Imperial College London

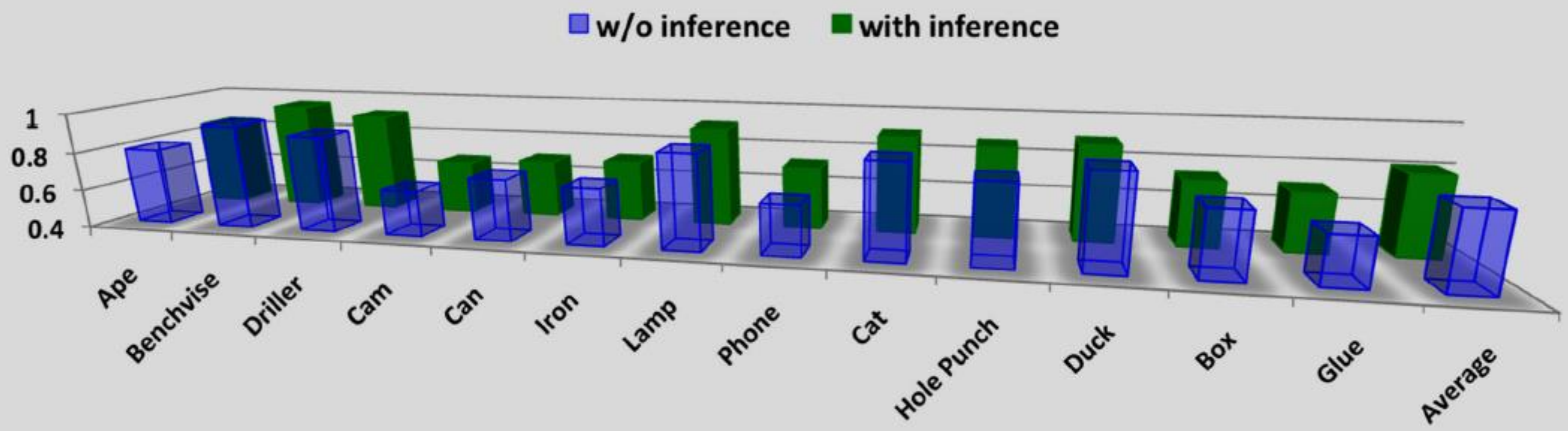
Inference with Iterative Refinement

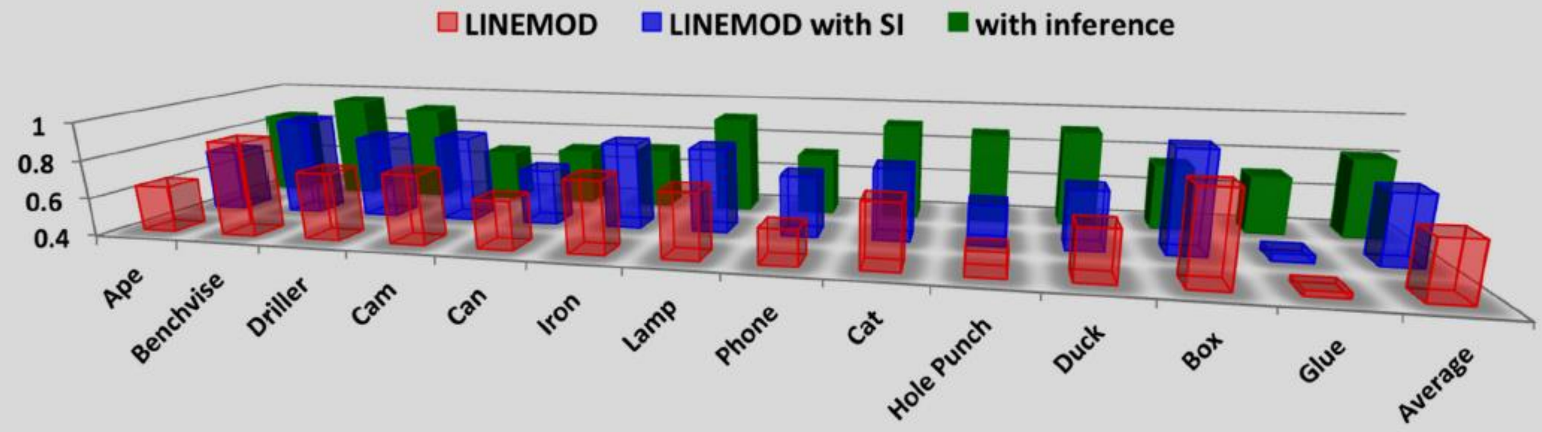


RGB-D images | #iter=1 | #iter=5 | #iter=10

Imperial College London

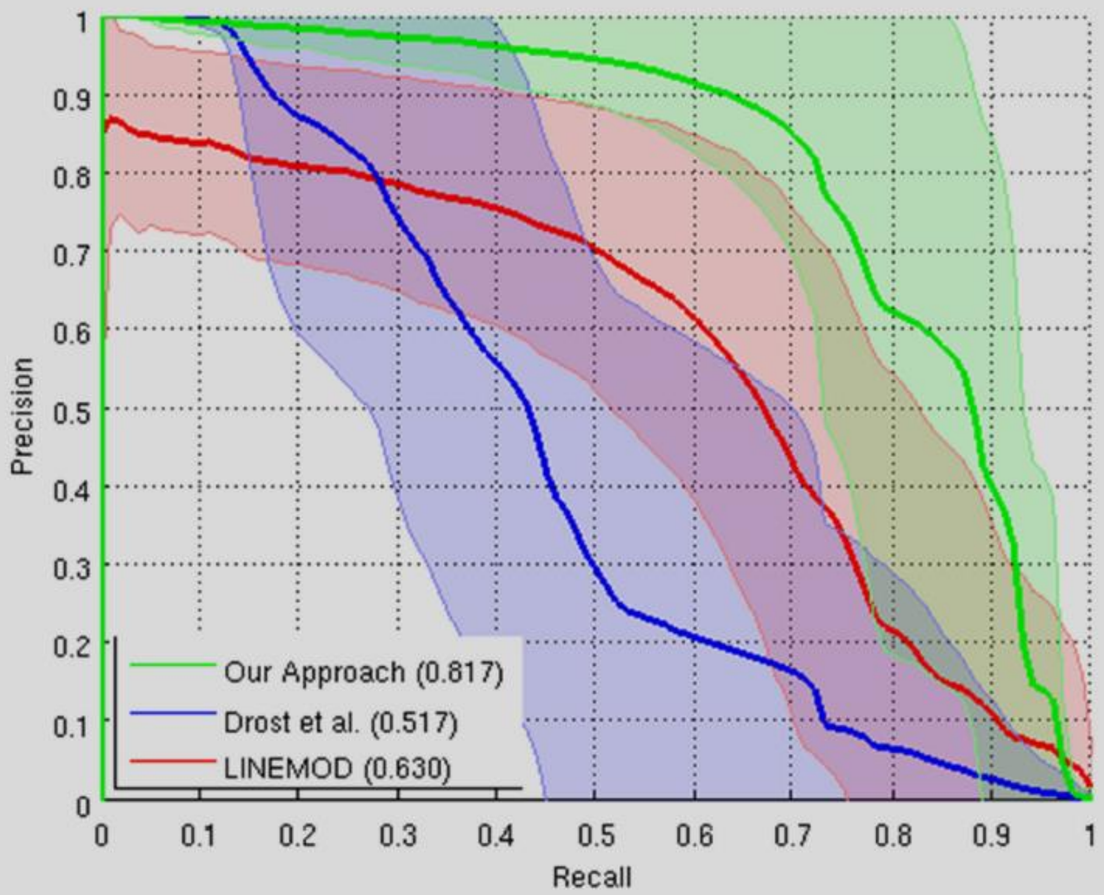
Results





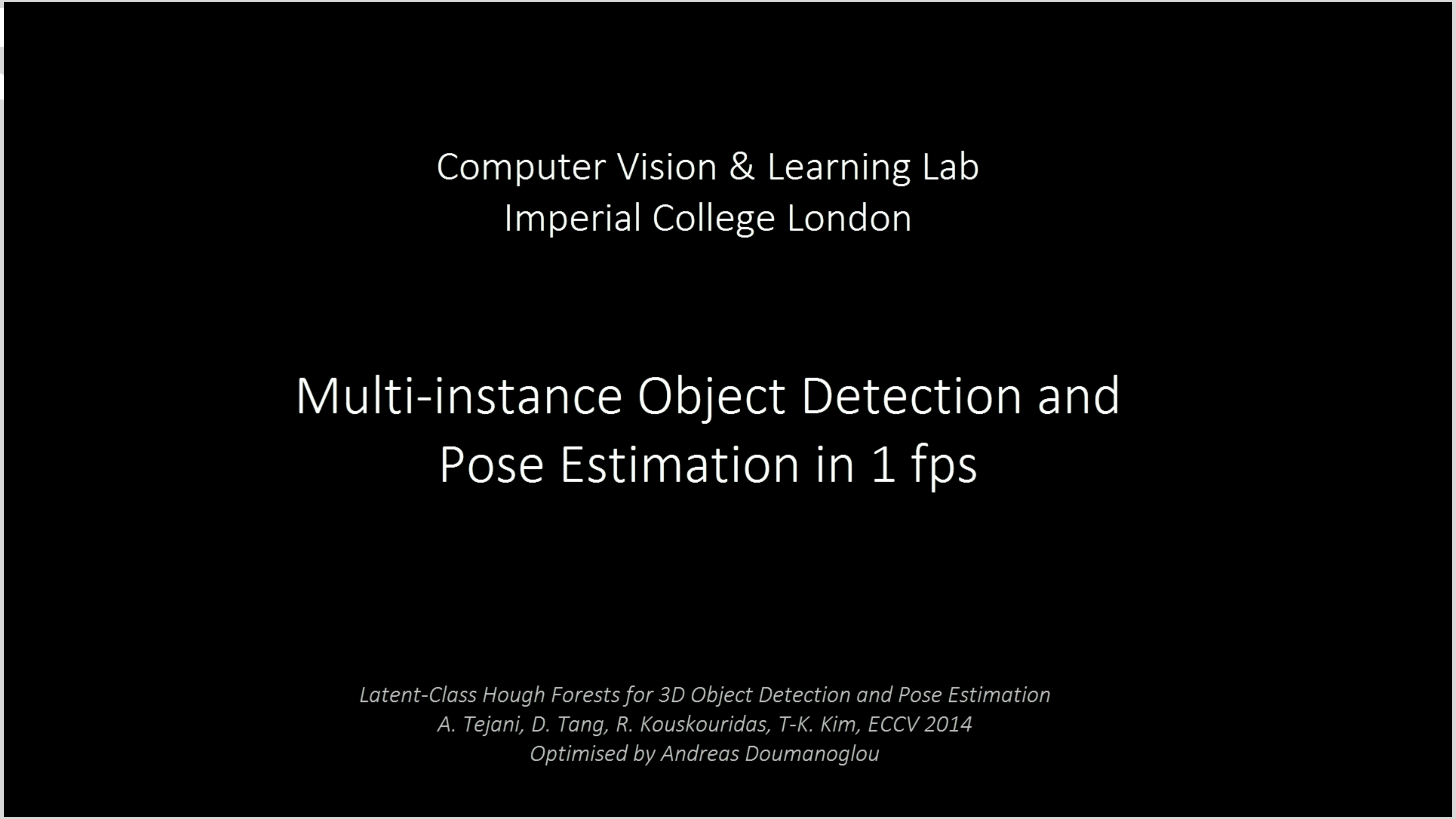
F1-Scores for the 13 objects in the dataset of Hinterstoisser et al. (1,100 RGBD images)

Results



Average Precision-Recall curve over all objects in the dataset of Hinterstoisser et al.

Imperial London



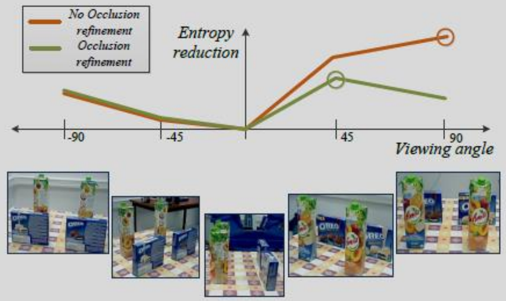
Demonstrated at Imperial College Science Festival in May 2015

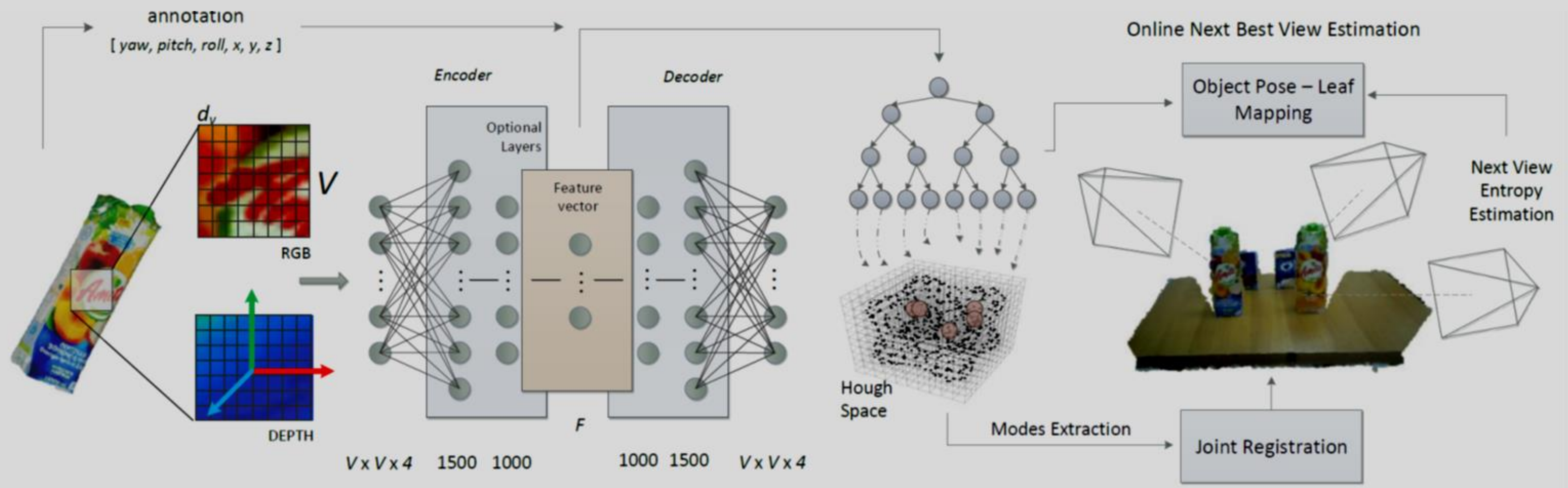
https://www.youtube.com/watch?v=dh2VtnnsGuY

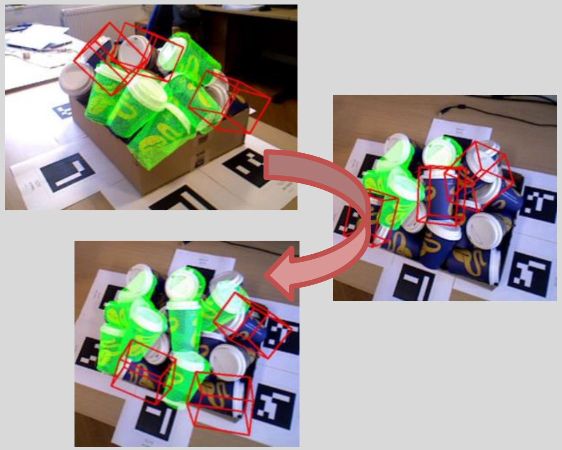
Imperial College London

Directions

- Object pose in the crowd (or bin-picking)
  - Better Feature Learning (deep convolutional networks)
  - Active vision (moving cameras, manipulators interacting objects)
  - Joint multiple object pose estimation (global optimization)







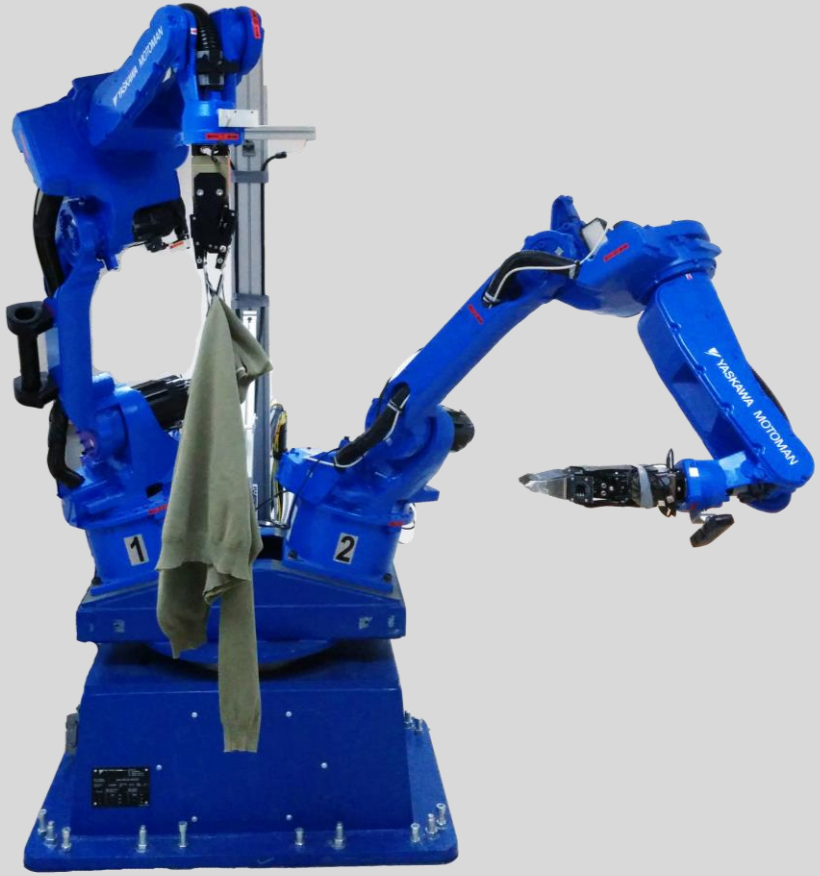
A complete pipeline including, sparse autoencoders, 6D hough voting, a novel next-best-view estimation based on Hough Forests (ongoing work)

Imperial College London

Autonomous Active Recognition and Unfolding of Clothes using Decision Forests

__A. Doumanoglou__, A. Kargakos, T-K. Kim, S. Malassiotis
ICRA 2014 (best service robotics paper award)

__A. Doumanoglou,__ T-K. Kim, X. Zhao, S. Malassiotis
ECCV 2014



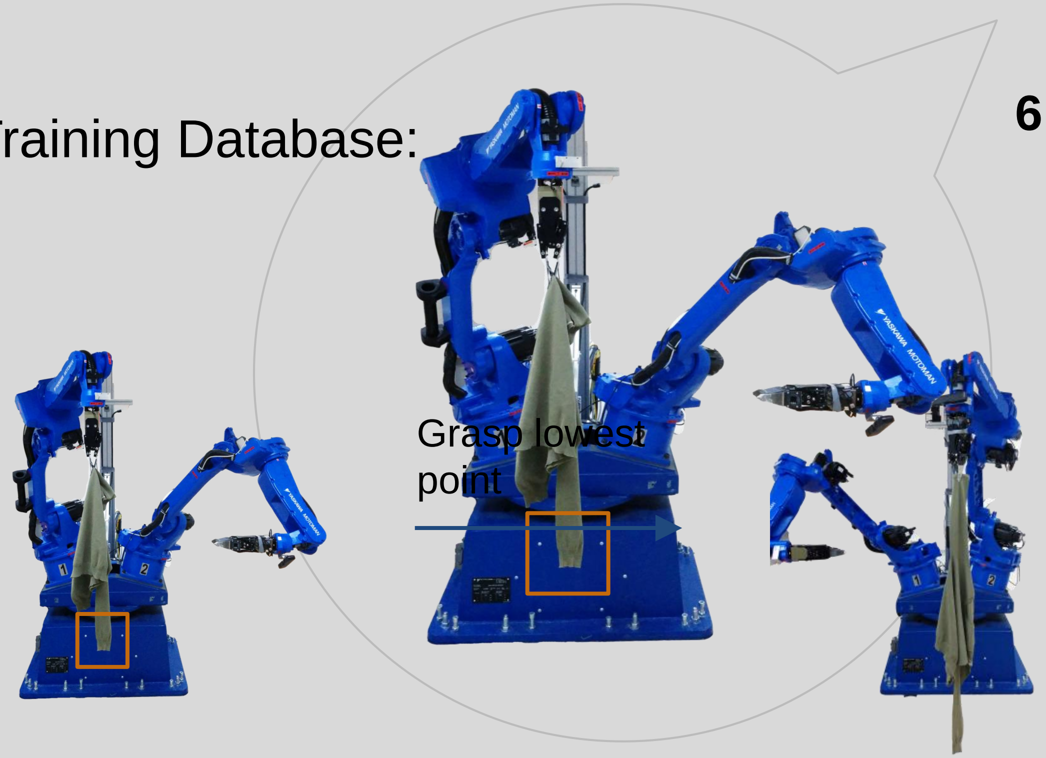
Clothes Recognition

- How to reduce the large configuration space ?

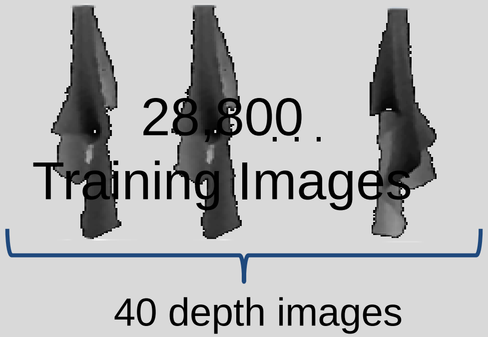
- Grasp Lowest Hanging Point First:

6 non-symmetric lowest points → **6 Classes** Total

- Training Database:



Rotate 360°



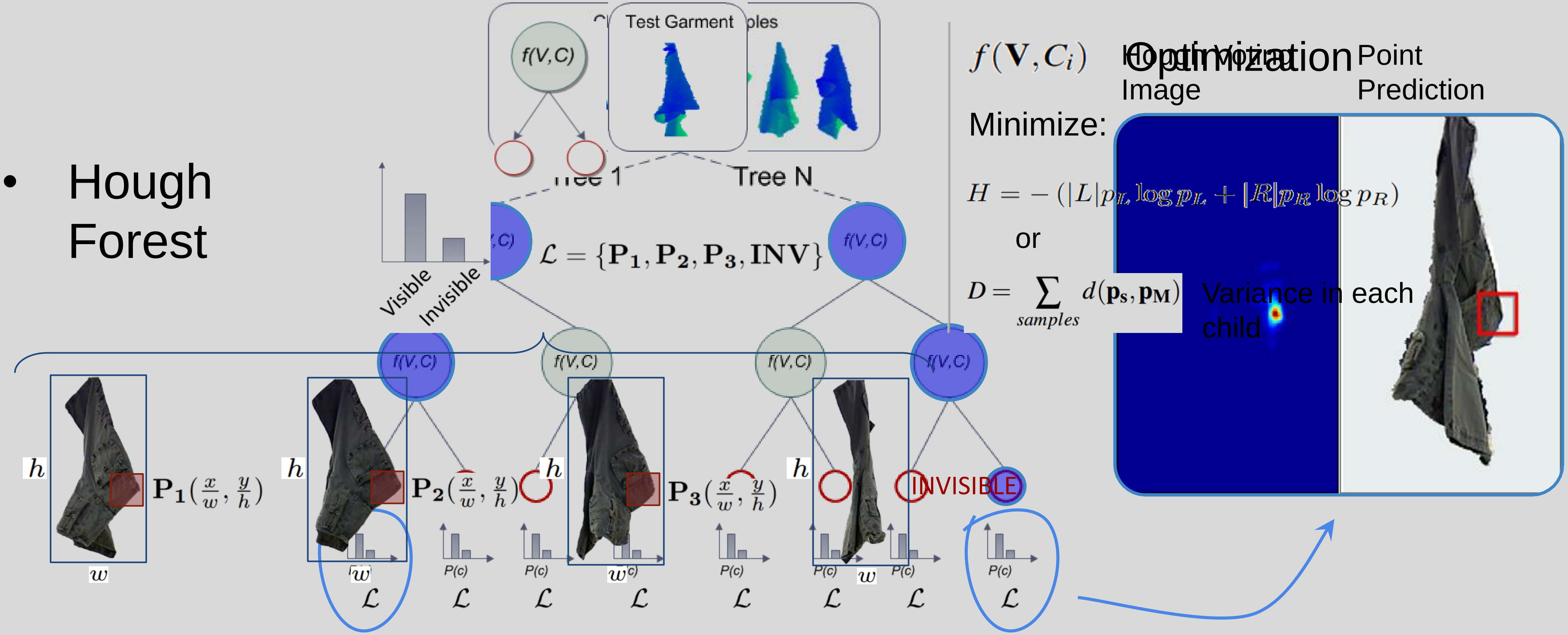
- RF training by pixel-tests in depth/curvature channels, and class entropy

Imperial College London

Grasp Point Detection

- Desired grasp Points:

- Hough Forest



Imperial College London

Active Planning

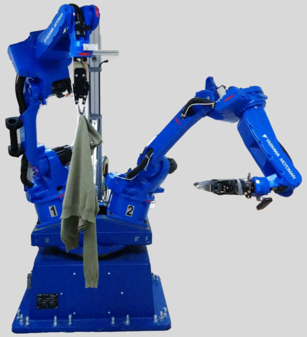
Single view

success ~ 90%

Crucial Decisions

How can other views help ?

Approach



Keep looking sequential views

Until we reach a certain **degree of confidence**

Imperial College London

Active Planning

POMDP (Partially Observable Markov Decision Processes) solution

Active Recognition

Actions (**A**): Rotate Cloth /
Take Final Decision

States (**S**): Clothes Classes

Observation Probabilities: **P (O | S, A)**
Measured Experimentally

Active Point Estimation

(i, j)

Actions (**A**): Rotate Cloth /
Grasp Garment at (i, j)

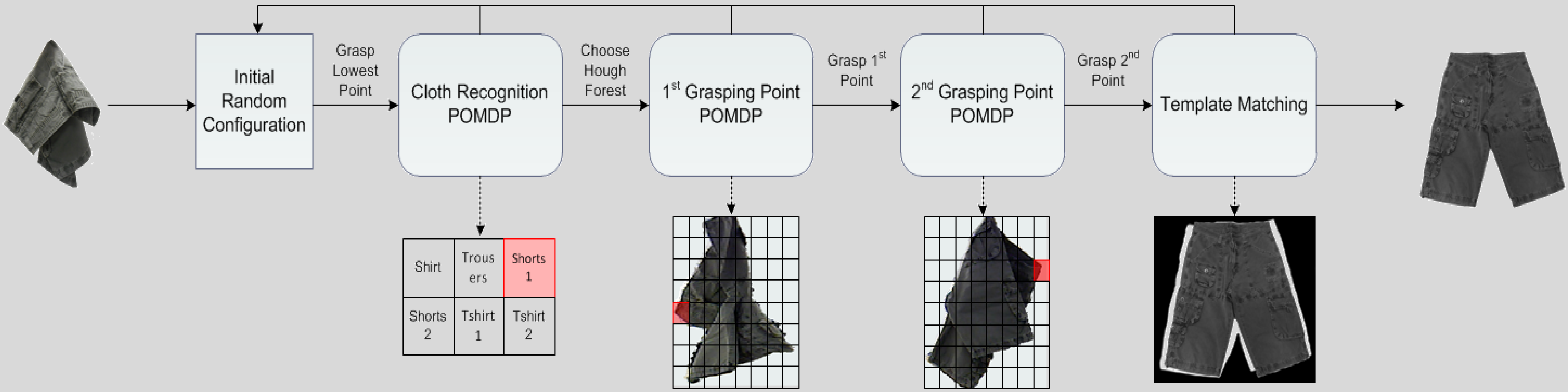
States (**S**): 65 — 8x8 grid quantization, or (INV)

Observation Probabilities: **P (O | S, A)**
Measured Experimentally

POMDP solution policy: ***A***(current belief state) → Optimal Action

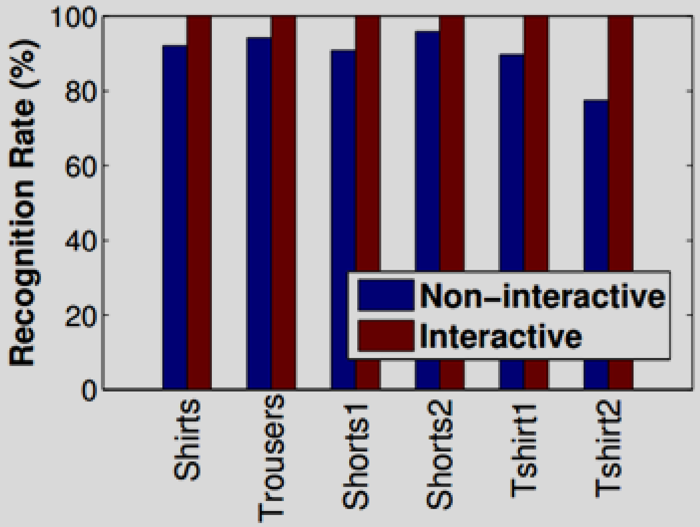
Block Diagram

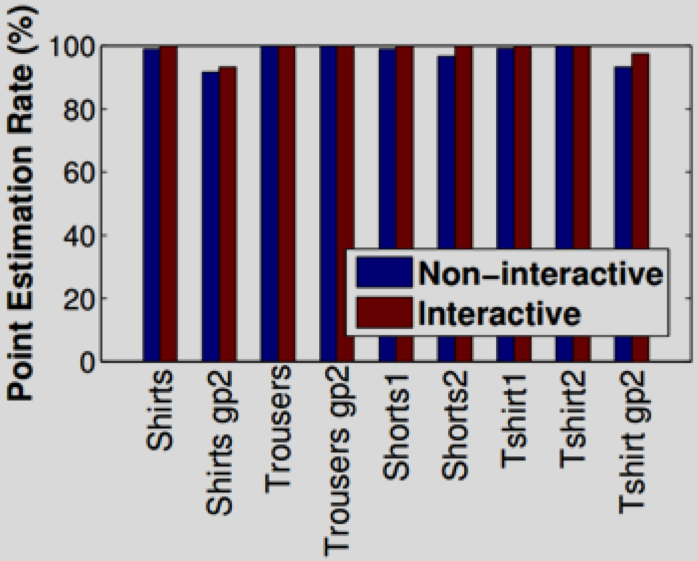
Unfolding Process



Results

- 28,800 training images and 1,440 testing images, captured with Xtion





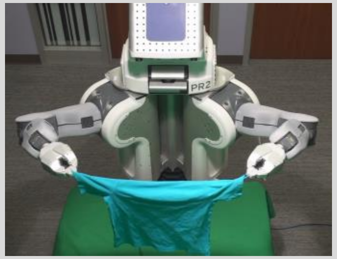
positive examples

negative examples

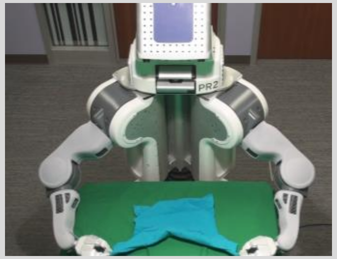
Imperial College London

Comparison with State-of-the-Art

Bringing clothing into desired configurations with limited perception, ICRA 2011 — M. Cusumano-Towner *et. al*



grasp
lowest point
twice



unfolding
using table
(slow)

baby clothes

grasp
lowest point
once

unfolding
in the air
(fast)

regular-sized
clothes

https://www.youtube.com/watch?v=YpD-ip6g5lY

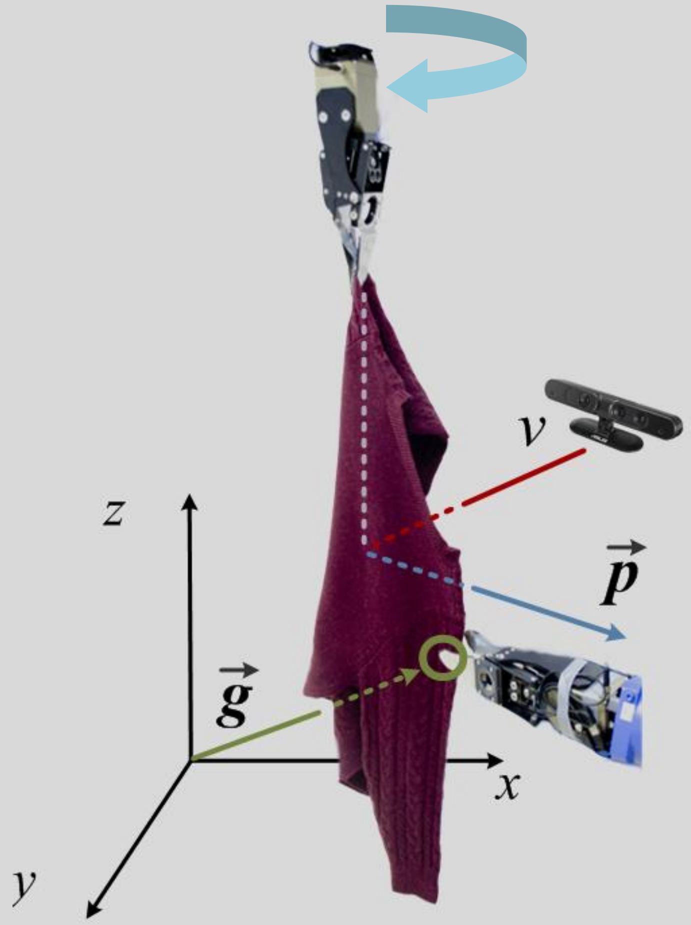
Imperial College London

Active Random Forests

Improve POMDP solution

Create a *Generic Active Vision Framework*

Extend objectives – Estimate Garment Pose



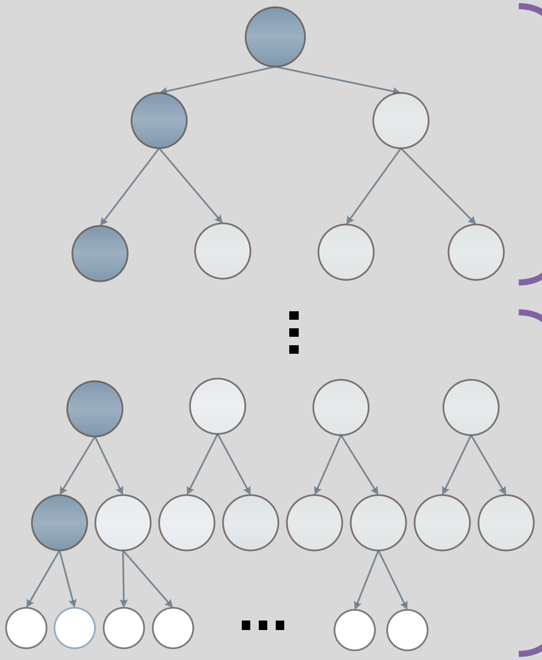
g, grasp point
v, viewpoint
p, pose

Imperial College London

Active Random Forests

One Forest for all objectives (*Classification, Regression, Pose Estimation*)

$$(\mathbf{I}(v), c, \mathbf{g}(v), \mathbf{p}(v))$$



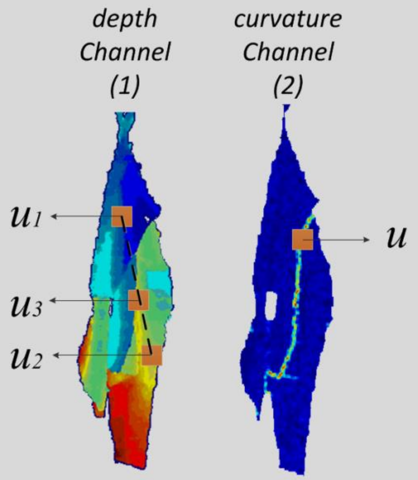
Hierarchical coarse to fine quality function *Q*

Classification (Q_c)

$$Q = \alpha Q_c + (1 - \alpha) Q_r$$

Regression (Q_r)

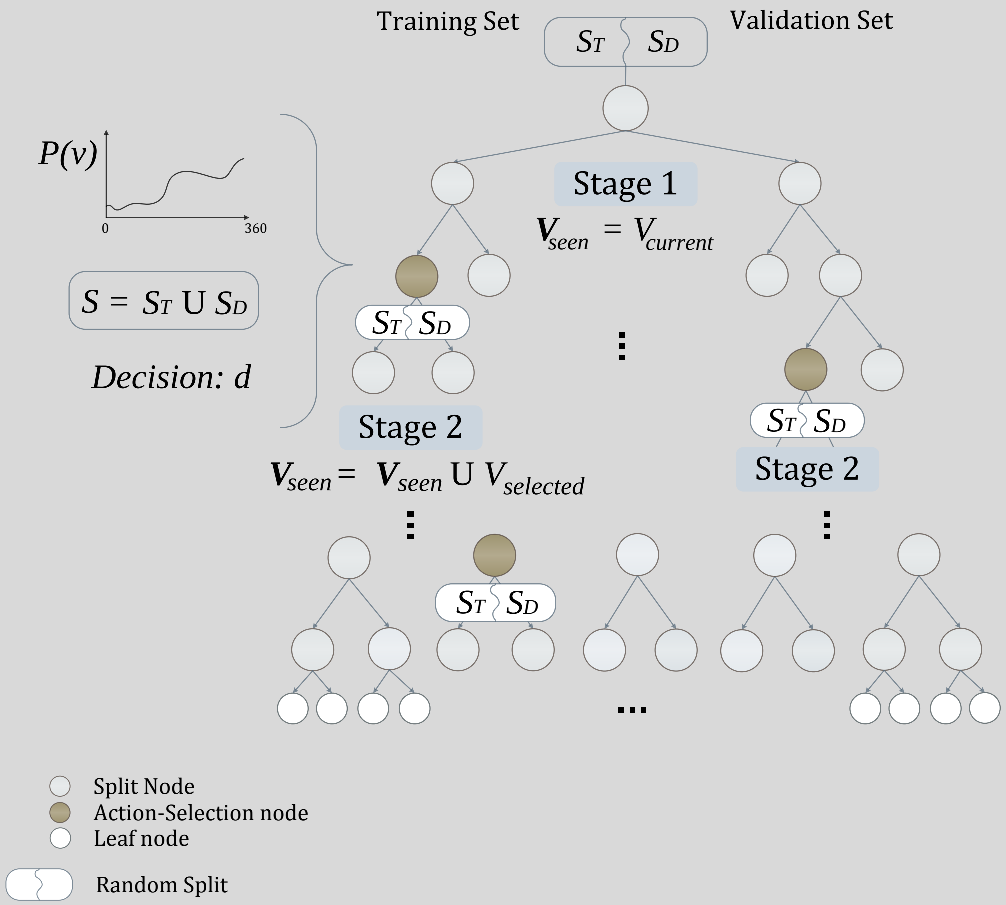
(Desired Grasp point, Pose)



Pixel tests

ARF Training

Training



‘Action-Selection’ Node Insertion Criteria

a) Hellinger Distance

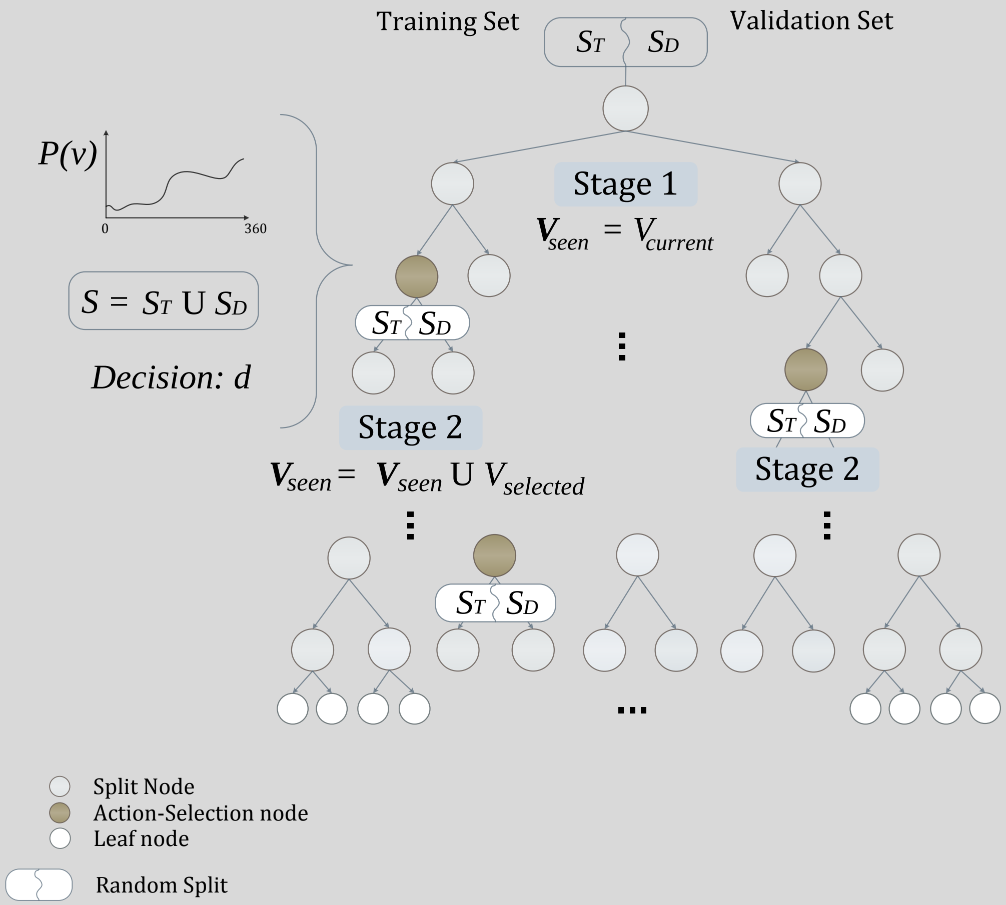
$$HL(S_T^j \| S_D^j) = \frac{1}{\sqrt{2}} \sqrt{\sum_{c=1}^{C} \left(\sqrt{P_{S_T^j}(c)} - \sqrt{P_{S_D^j}(c)} \right)^2} > t_\Delta$$

b) Jeffrey Divergence

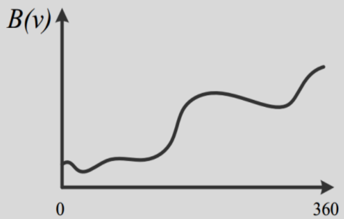
$$JS(S_T^j \| S_D^j) = \frac{1}{C} \sum_{c=1}^{C} P_{S_T^j}(c) \log \frac{P_{S_T^j}(c)}{P_m(c)} + P_{S_D^j}(c) \log \frac{P_{S_D^j}(c)}{P_m(c)} > t_\Delta$$

ARF Training

Training

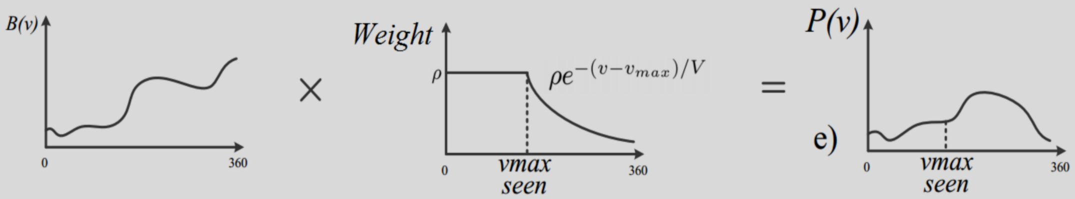


Grasp Point Visibility



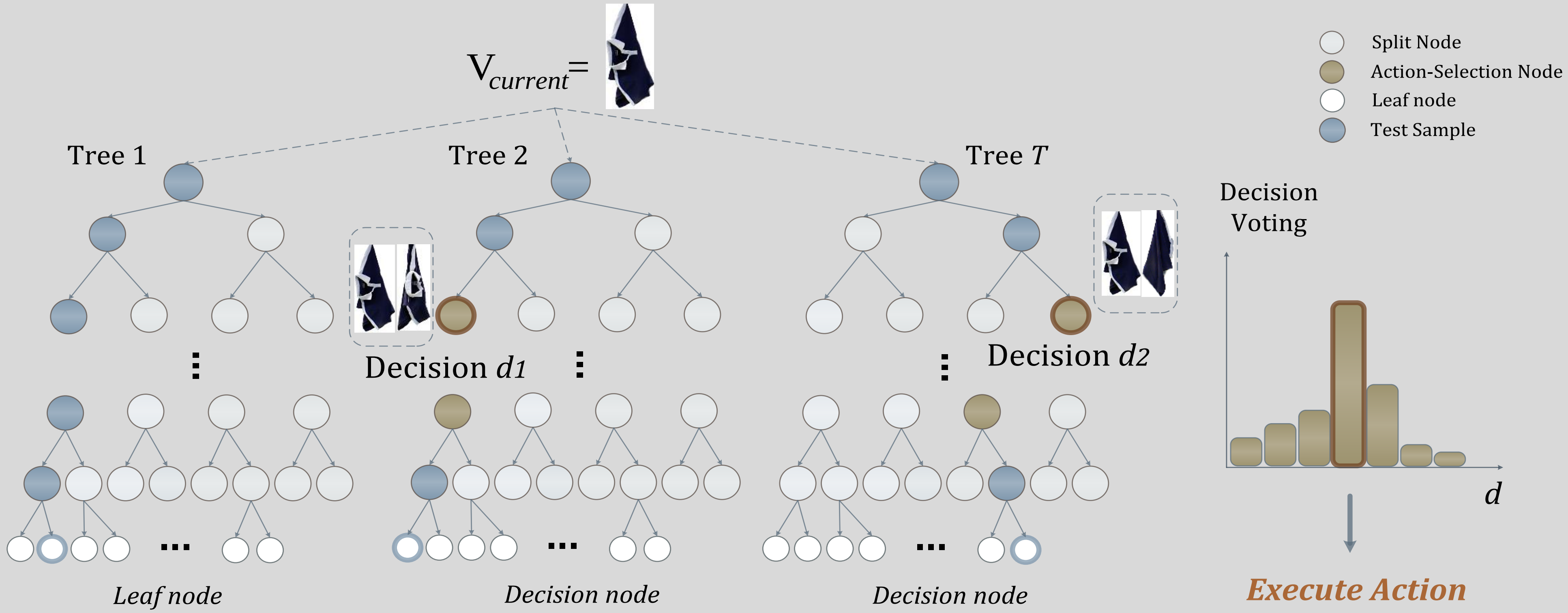
- *Calculated from training*
- *Random sampling for next best view in action-selection nodes*

Cost of actions



Imperial College London

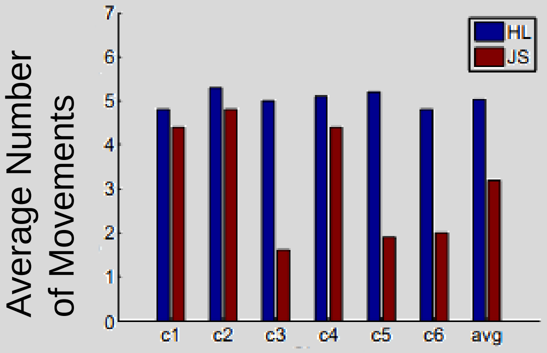
ARF Testing

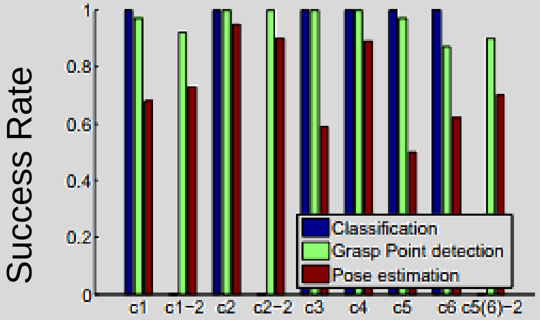


Imperial College London

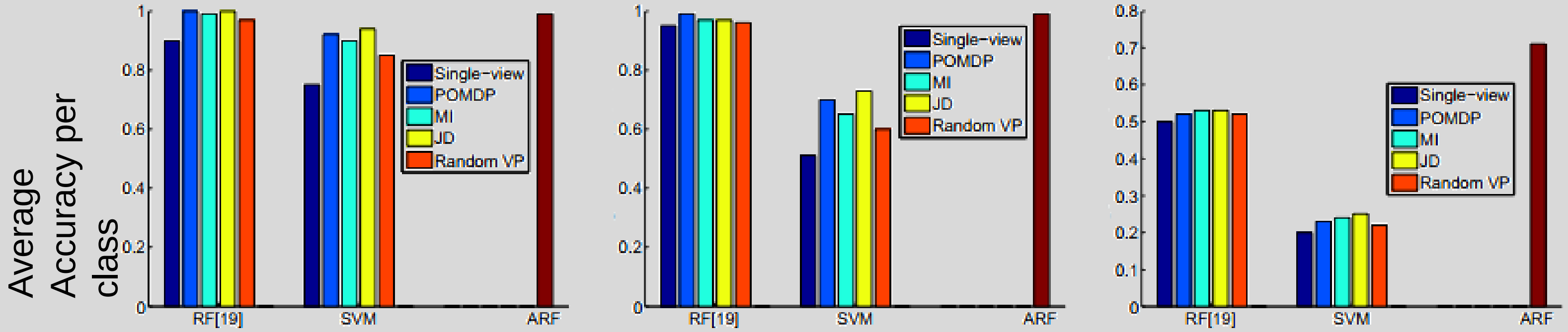
ARF Results

Self Comparisons





Comparison with state of the art

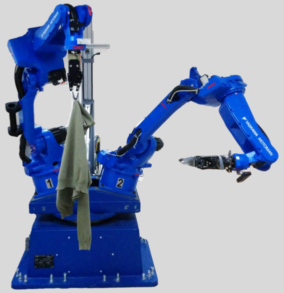


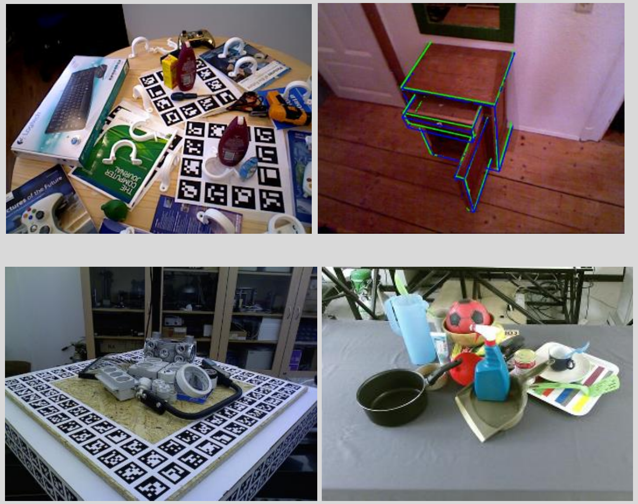
Qualitative Results

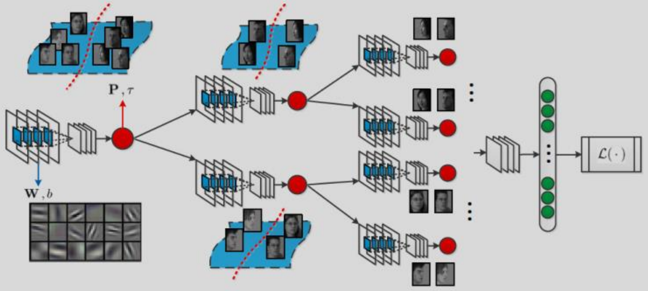
Imperial College London

Directions

- Various benchmarks/methods have been collected.
- A comparative study (using the challenge results) will be done.
  - Feature comparison, active vision, multi-object registration, multi-view registration, real-time performance, texture-less, articulated objects, highly occluded scenarios, etc.
- Deep learning + RF
  - learning representation, conditional computing, efficiency
- Active RF classifiers
  - action as a learning parameter







Chao et al. ICCV15