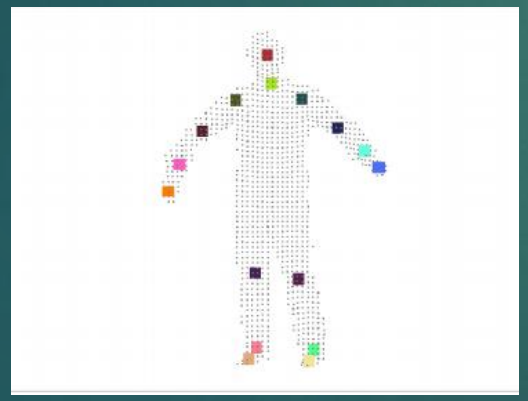
Real-Time Human Pose Recognition in Parts from Single Depth Images

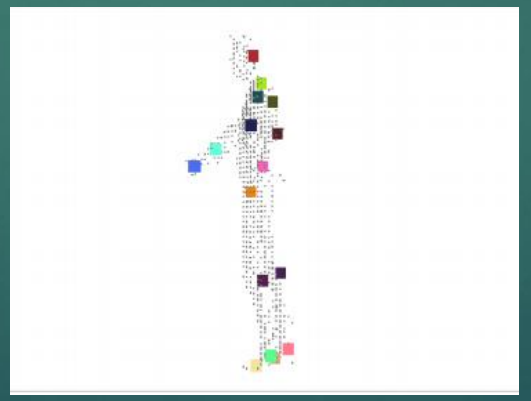
Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, Andrew Blake

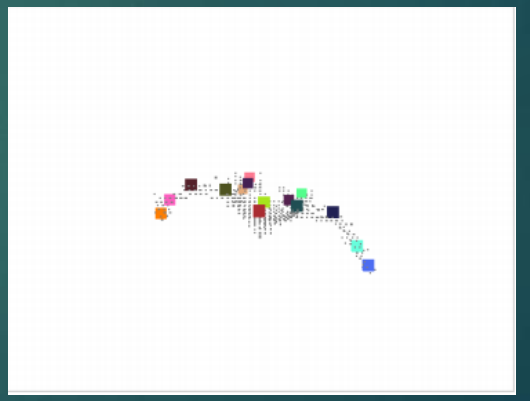
CVPR 2011

PRESENTER: AHSAN ABDULLAH

PROBLEM







APPROACH

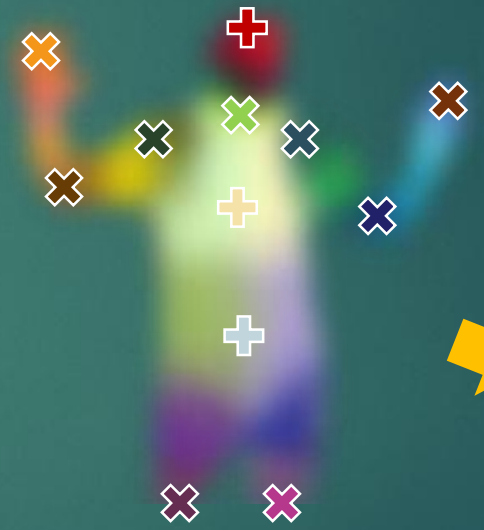
- Partitioning into body parts helps localizing the joints

Shotton et. al. CVPR 2011

PIPELINE

capture
depth image &
remove bg

infer
body parts
per pixel



cluster pixels to
hypothesize
body joint
positions

fit model &
track skeleton

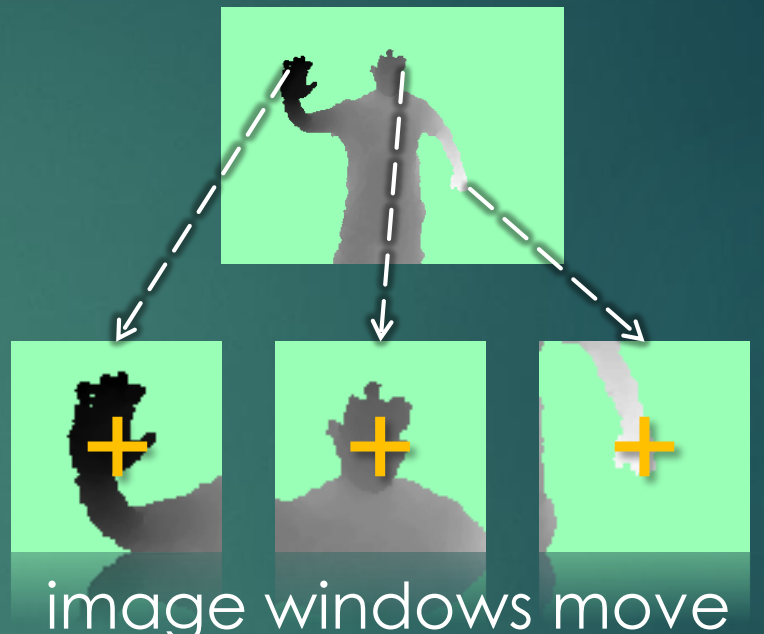
Shotton et. al. CVPR 2011

Design Goals

- Efficiency
- Robustness

BODY PART CLASSIFICATION

- Compute $\mathrm{P}(c_i \mid w_i)$
  - pixels $i = (x, y)$
  - body part c_i
  - image window w_i



- Discriminative approach
  - learn classifier $\mathrm{P}(c_i \mid w_i)$ from training data

Shotton et. al. CVPR 2011

LEARNING DATA

synthetic
(train & test)

real
(test)

Shotton et. al. CVPR 2011

LEARNING – DATA SYNTHESIS

Record MoCap
500k frames
distilled to 100k poses

Retarget to several models

Render (depth, body parts) pairs

Shotton et. al. CVPR 2011

FEATURE SET

- Depth comparisons
  - very fast to compute

image depth offset depth

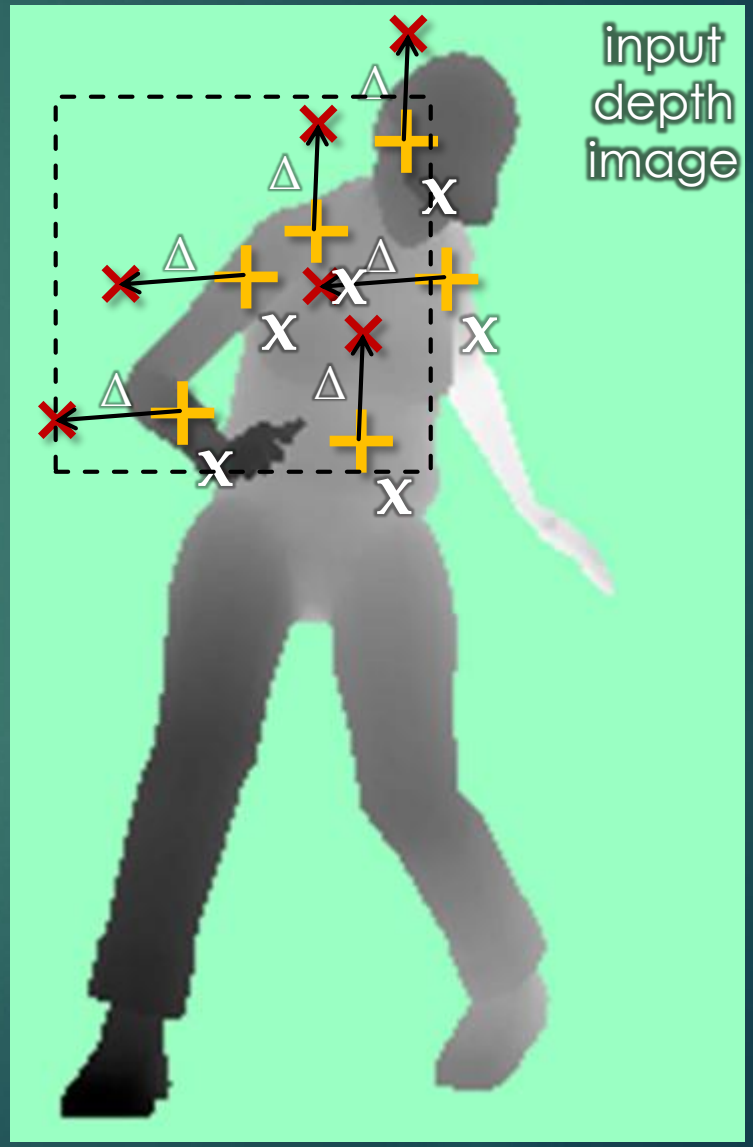
feature response $f(I, \mathbf{x}) = d_I(\mathbf{x}) - d_I(\mathbf{x} + \Delta)$

image coordinate

$$\Delta = \frac{\mathbf{v}}{d_I(\mathbf{x})}$$

scales inversely with depth

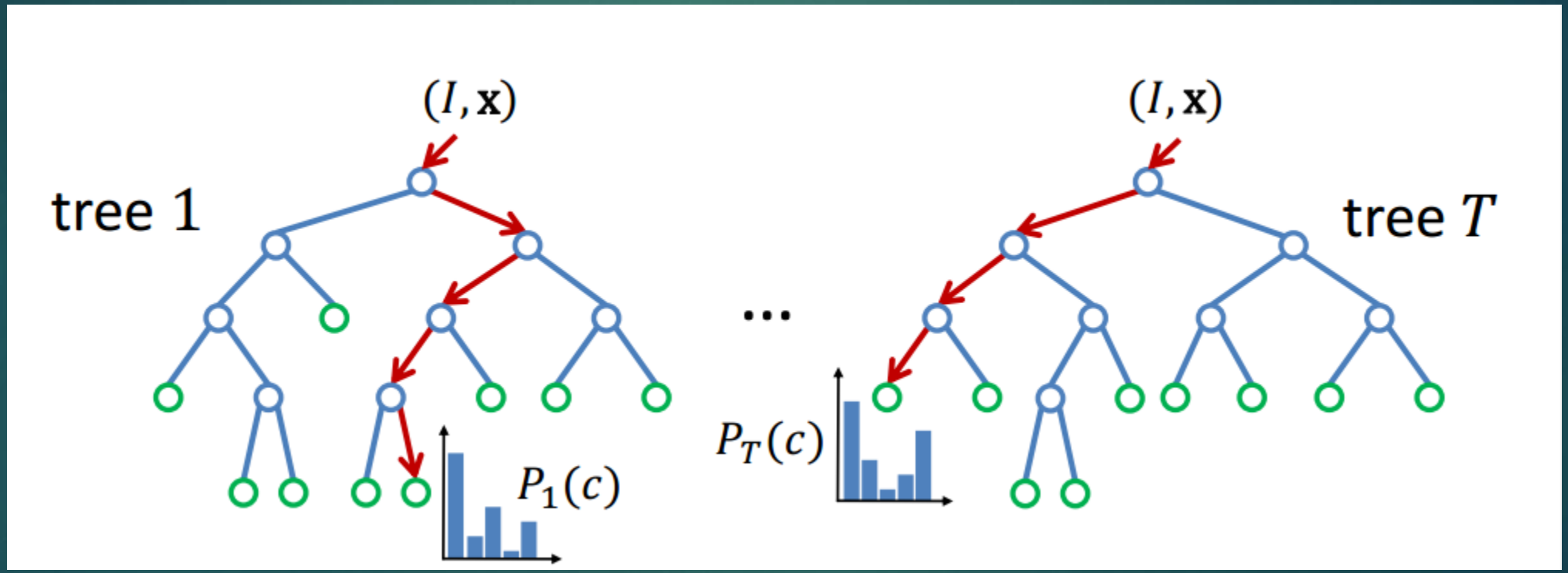
Background pixels
d = large constant



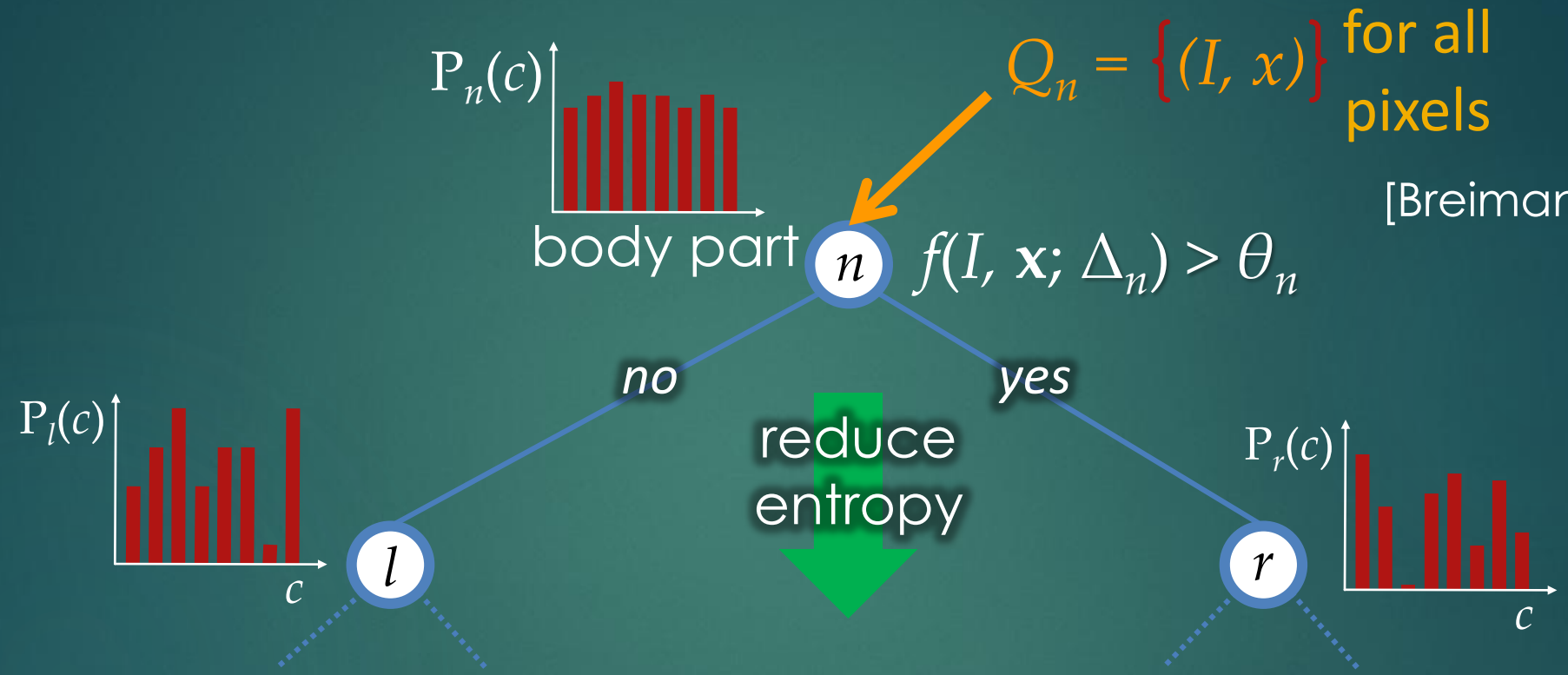
Shotton et. al. CVPR 2011

DECISION FORESTS

- Aggregation of decision trees



TRAINING DECISION TREES



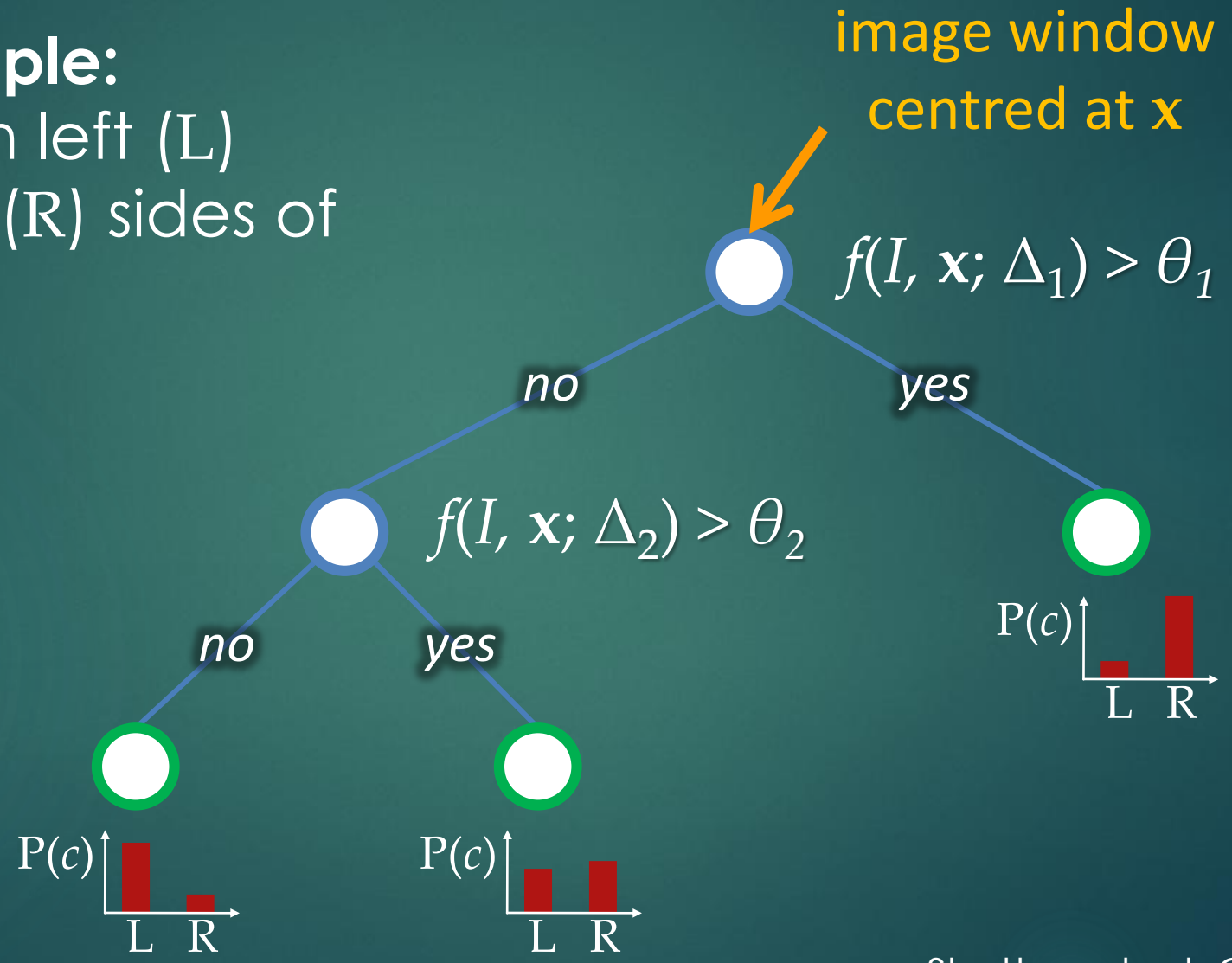
[Breiman *et al.* 84]

Take (Δ, θ) that maximises information gain

Shotton et. al. CVPR 2011

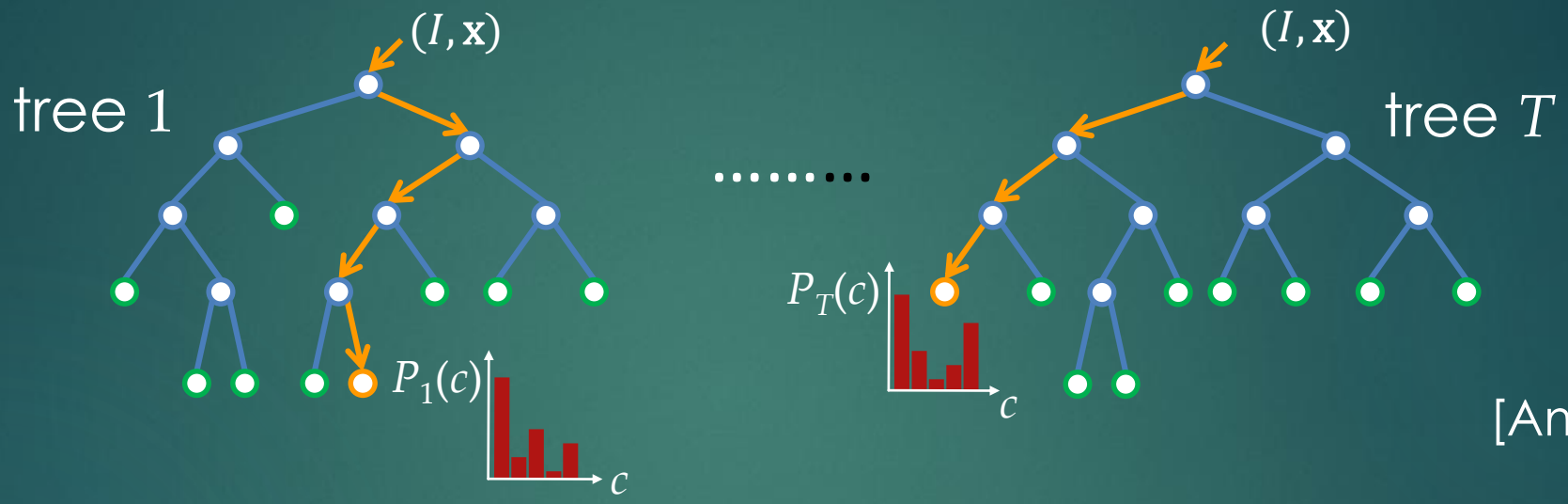
DECISION TREE CLASSIFICATION

Toy example:
Distinguish left (L) and right (R) sides of the body



Shotton et. al. CVPR 2011

DECISION FOREST CLASSIFIER



[Amit & Geman 97]
[Breiman 01]
[Geurts *et al.* 06]

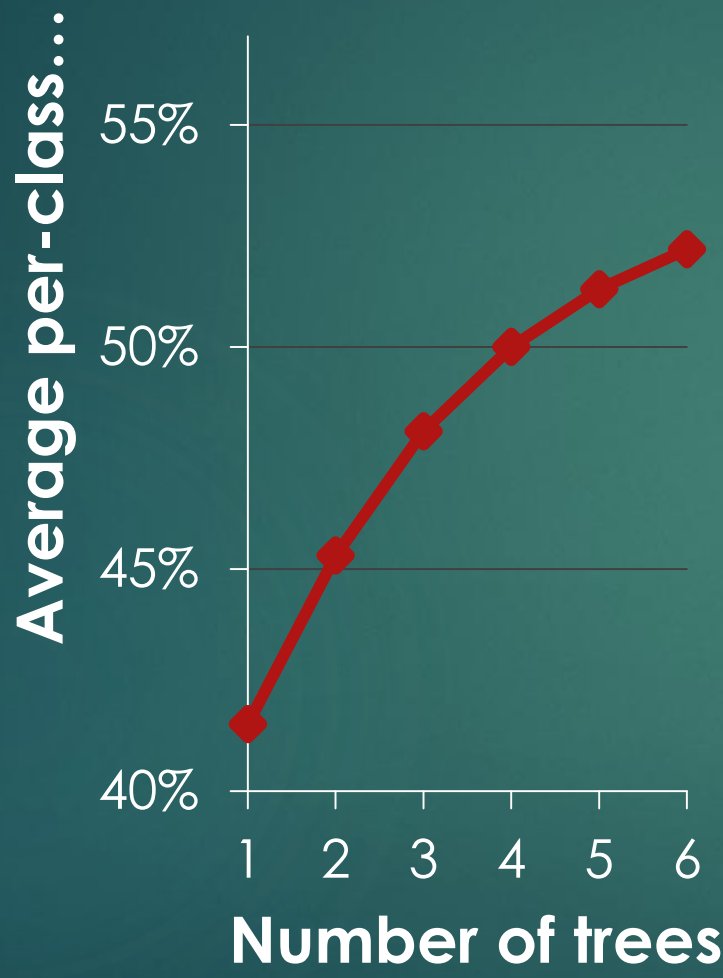
- Trained on different random subset of images
  - "bagging" helps avoid over-fitting

- Average tree posteriors

$$P(c|I,\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} P_t(c|I,\mathbf{x})$$

Shotton et. al. CVPR 2011

NUMBER OF TREES



ground truth

inferred body parts (most likely)

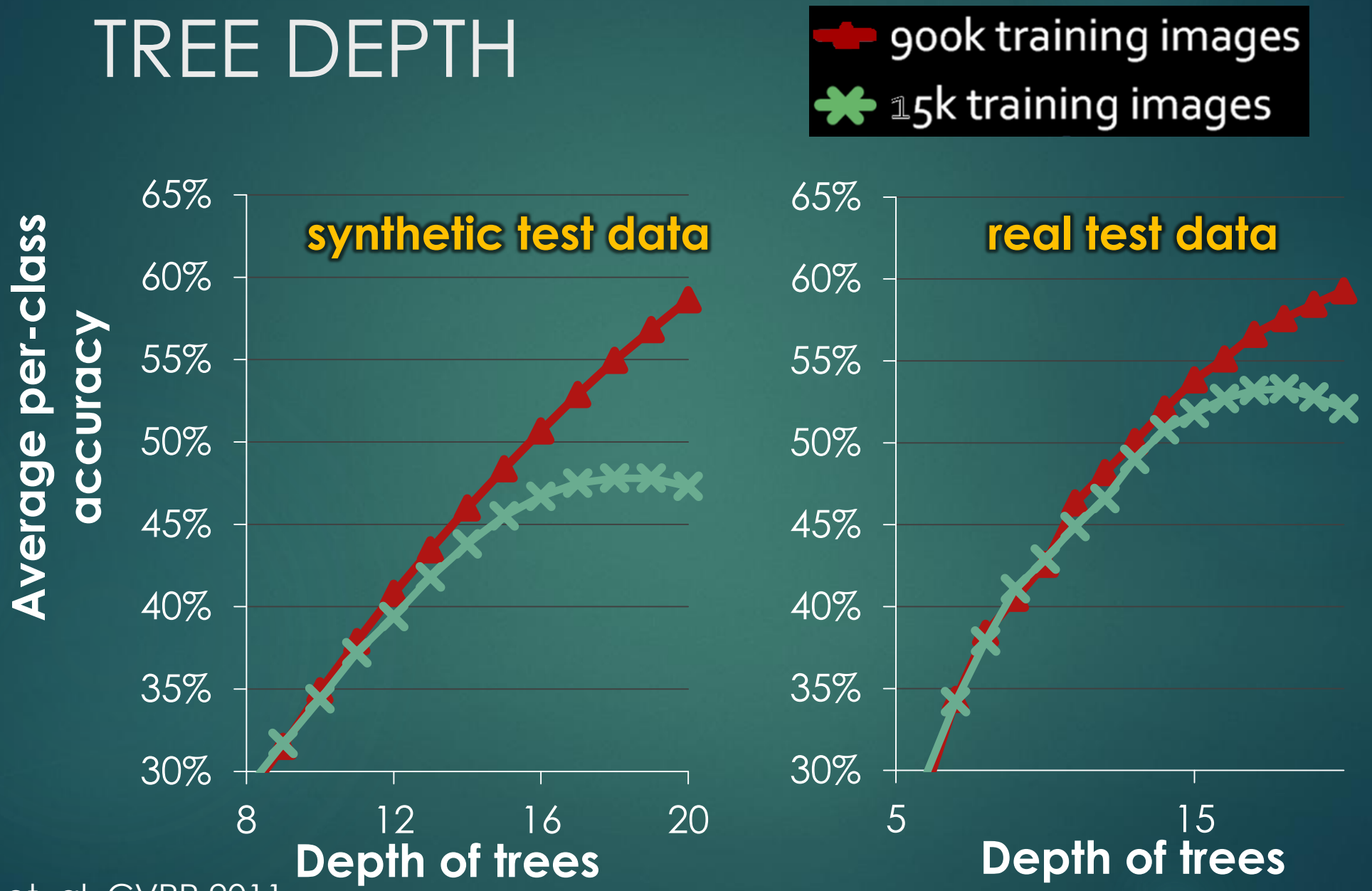
1 tree

3 trees

6 trees

Shotton et. al. CVPR 2011

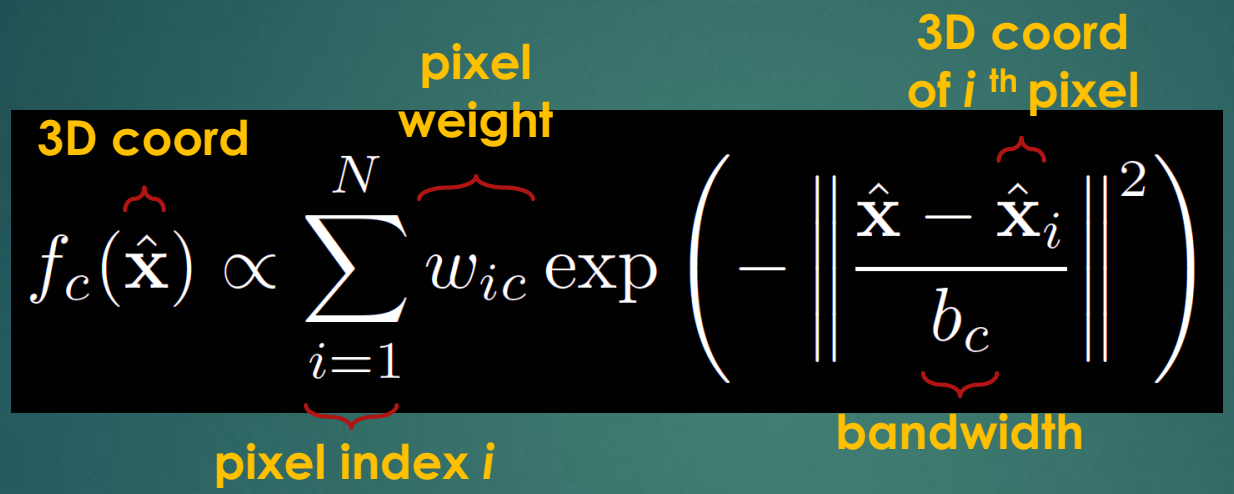
TREE DEPTH

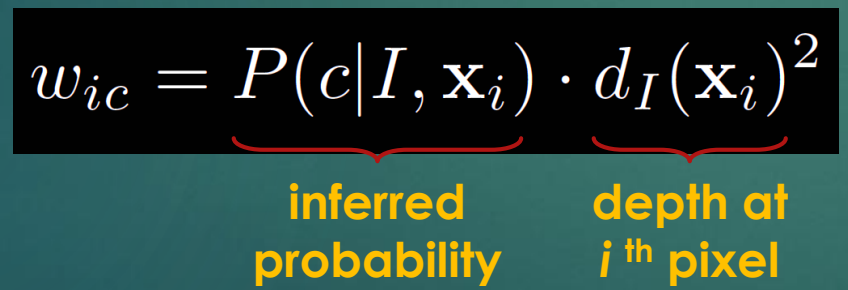


Shotton et. al. CVPR 2011

Body parts to joint hypotheses

- Define 3D world space density



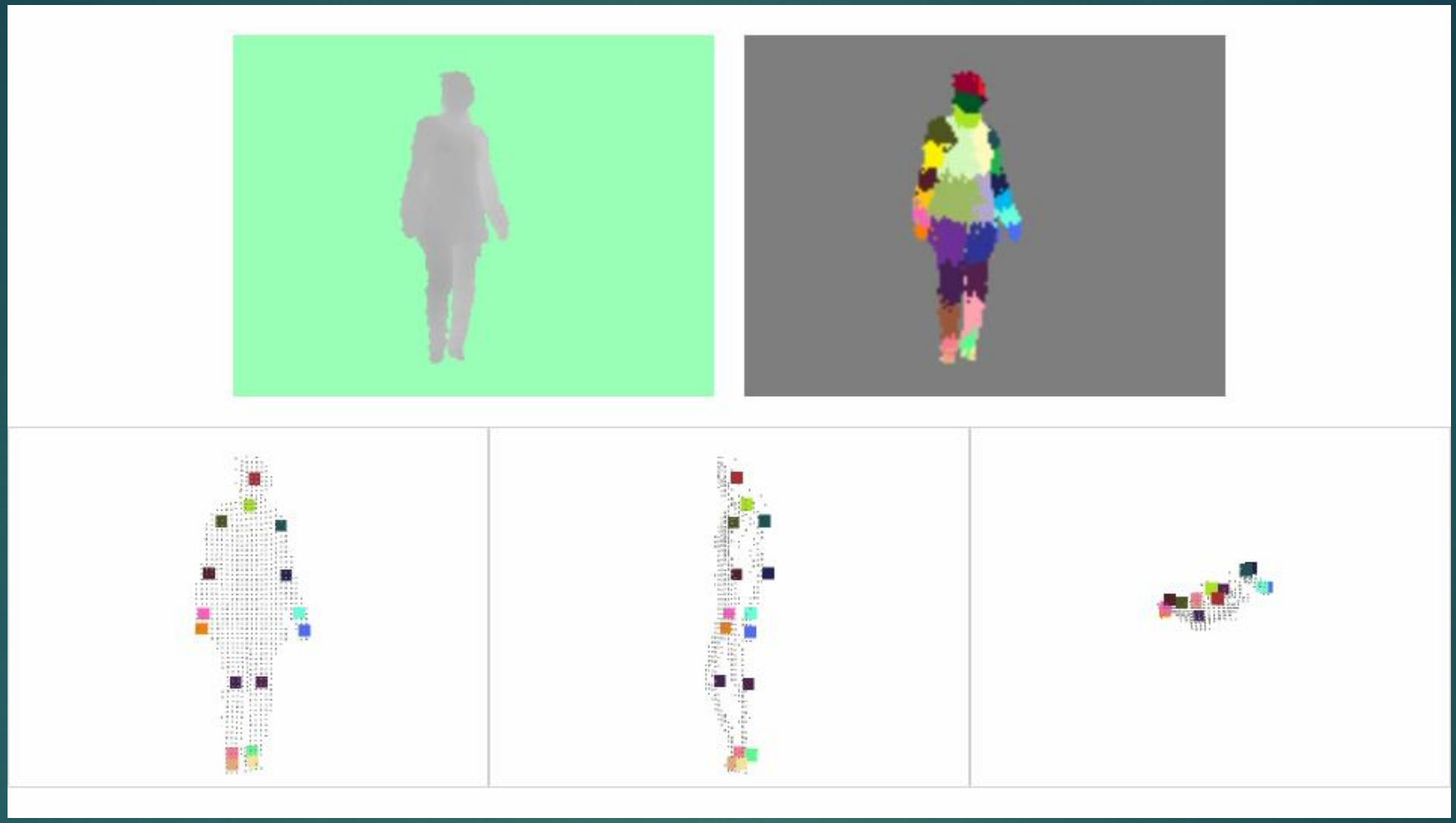


- Mean shift for mode detection

Shotton et. al. CVPR 2011

input depth

inferred body parts



front view

side view

top view

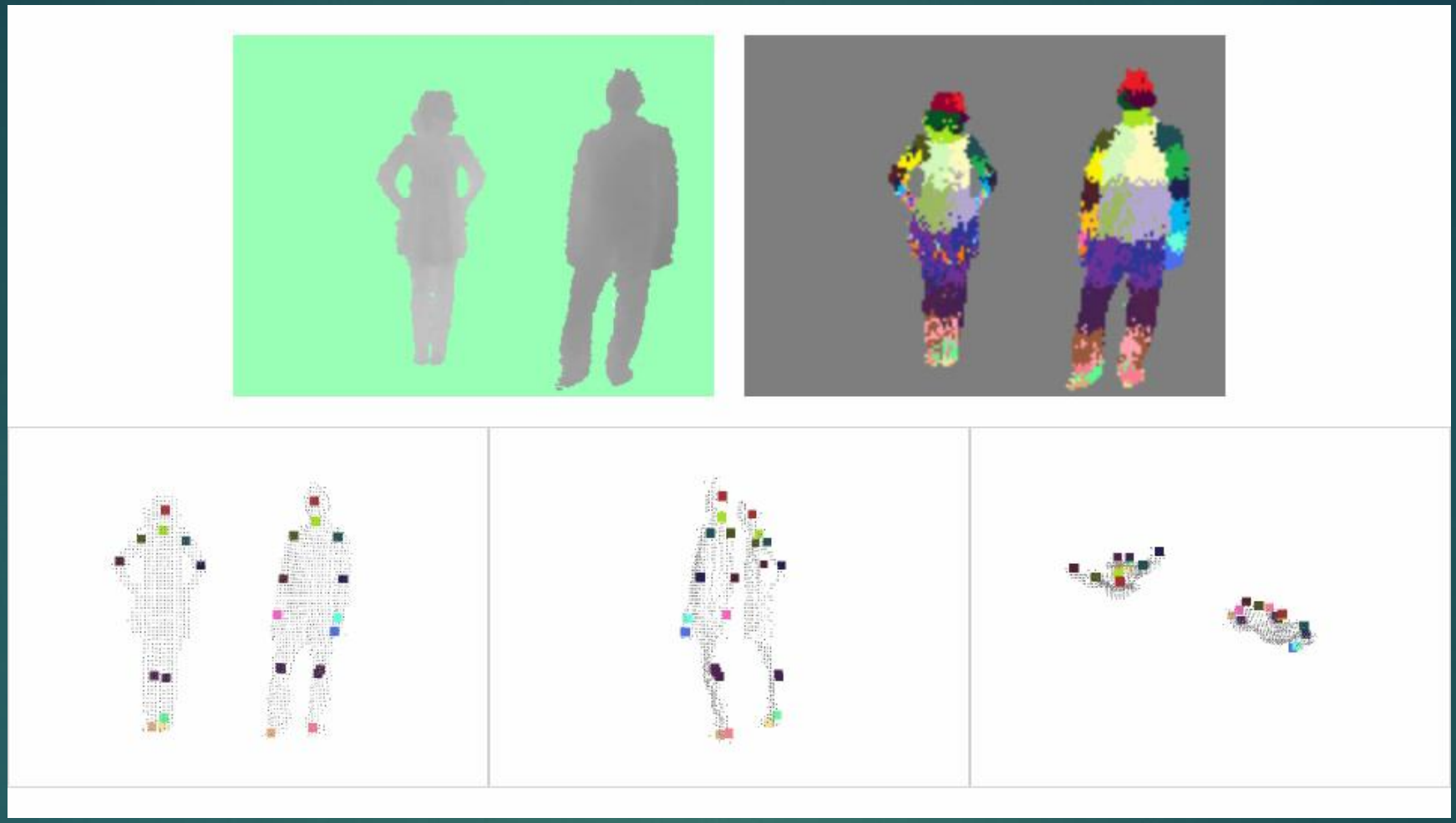
inferred joint positions

No tracking or smoothing

Shotton et. al. CVPR 2011

input depth

inferred body parts



front view

side view

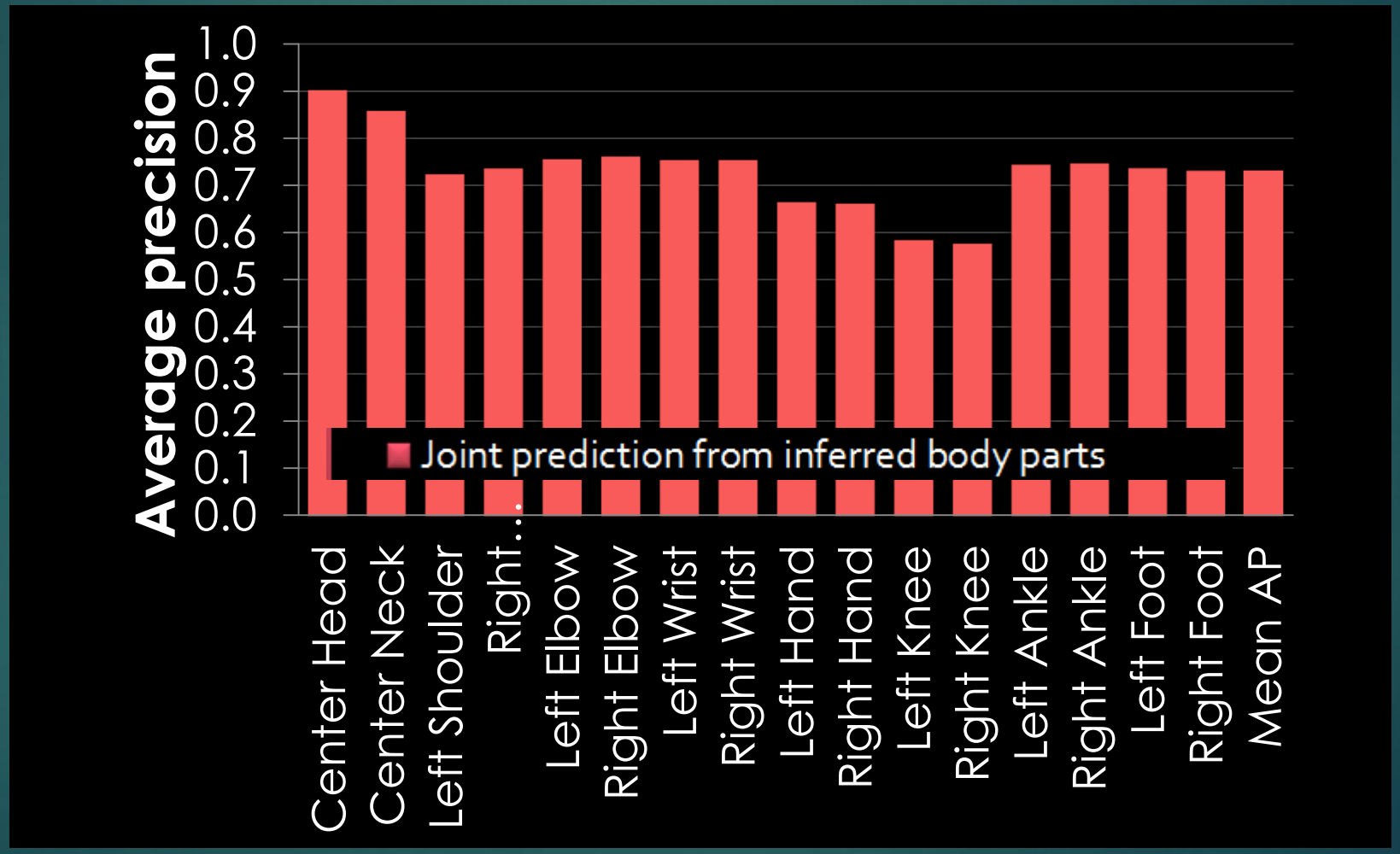
top view

inferred joint positions

No tracking or smoothing

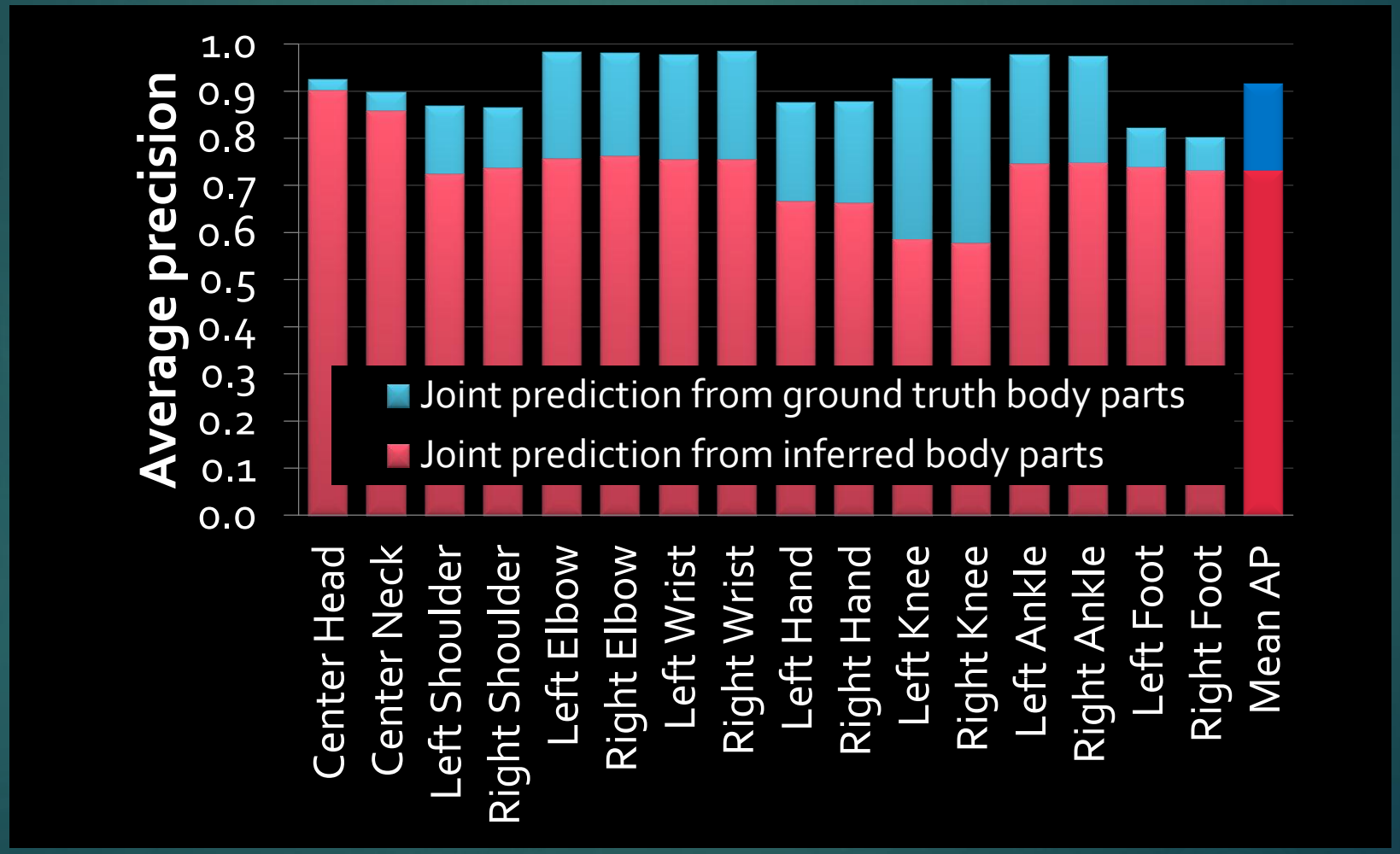
Shotton et. al. CVPR 2011

JOINT PREDICTION ACCURACY



Shotton et. al. CVPR 2011

JOINT PREDICTION ACCURACY



Shotton et. al. CVPR 2011

ANALYSIS

- No temporal information
  - frame-by-frame
- Very fast
  - simple depth image feature
  - parallel decision forest classifier

Shotton et. al. CVPR 2011

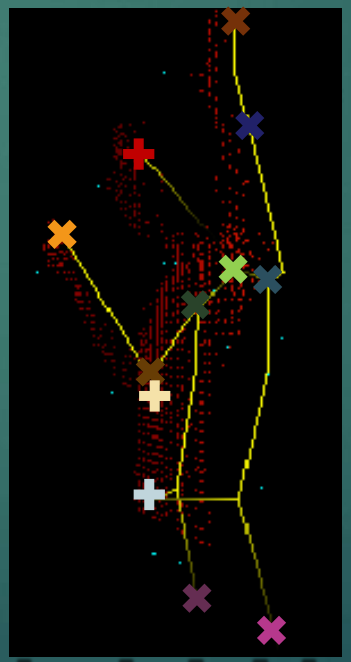
KINECT SYSTEM

Uses…

- 3D joint hypotheses
- kinematic constraints
- temporal coherence

… to give

- full skeleton
- higher accuracy
- invisible joints
- multi-player



4. track skeleton

SUMMARY

- Frame-by-frame gives robustness
- Body parts representation for efficiency
- Fast, simple machine learning
- Significant engineering to scale to a massive, varied training data set

Shotton et. al. CVPR 2011

QUESTIONS