

Human Skeletal Tracking, and the Development of KINECT

KINECT

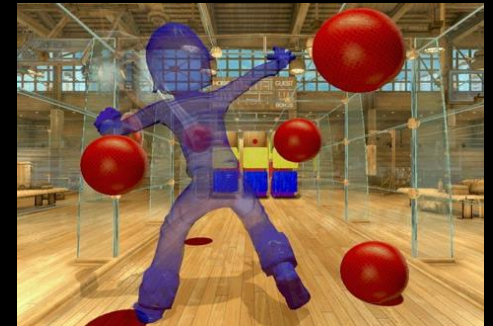
Jamie Shotton
Senior Researcher

Microsoft Research

KINECT™



- Plugs into your Xbox or PC
- Combines new technologies:
 - depth sensing camera
 - real time human skeletal tracking
 - face and voice recognition
- Applications in both gaming and much more





Kinect: Human Motion Tracking

Talk roadmap

- Background: visual recognition
- “Project Natal”:
a call to action
- Machine learning for
body part recognition
- Other applications



Visual object recognition



0.2, 0.1, 0.1	0.1, 0.1, 0.4	0.0, 0.2, 0.7	0.1, 0.4, 0.3	1.0, 0.3, 0.3
0.2, 0.9, 0.4	0.9, 0.8, 0.7	0.3, 0.2, 0.1	0.5, 0.0, 0.2	0.5, 0.5, 0.0
0.1, 0.3, 0.5	0.3, 0.6, 0.1	0.1, 1.0, 0.5	0.9, 0.8, 0.7	0.2, 0.6, 0.7
0.3, 0.7, 0.3	0.2, 0.9, 0.5	0.1, 0.5, 0.0	1.0, 0.0, 0.9	0.1, 0.8, 0.8
0.1, 0.2, 0.7	0.6, 0.7, 0.0	0.7, 0.7, 0.9	0.2, 0.1, 0.0	0.8, 0.4, 0.5
0.6, 0.4, 0.2	0.0, 0.7, 0.2	1.0, 0.1, 1.0	0.0, 0.6, 1.0	0.3, 0.1, 0.4
0.9, 0.0, 0.0	0.7, 0.8, 0.7	0.5, 0.6, 0.6	0.6, 0.6, 0.3	0.3, 1.0, 1.0
0.9, 0.4, 1.0	0.5, 0.0, 0.7	0.6, 0.0, 0.0	0.8, 0.9, 0.2	0.4, 0.9, 0.0
0.0, 0.3, 0.5	0.2, 0.4, 0.0	0.2, 0.3, 0.1	0.5, 0.9, 0.5	1.0, 0.6, 0.6
0.9, 0.8, 1.0	0.8, 0.7, 0.5	0.2, 0.8, 0.4	0.6, 0.9, 0.8	0.9, 0.8, 0.9
0.5, 0.9, 0.7	1.0, 0.1, 0.0	0.5, 0.6, 0.7	0.4, 0.5, 0.8	0.4, 0.8, 0.7
0.7, 0.6, 0.6	0.6, 0.2, 0.1	0.4, 0.7, 0.3	0.1, 0.2, 0.4	0.9, 1.0, 0.2
0.3, 0.3, 0.1	1.0, 0.3, 0.1	0.6, 0.4, 0.9	0.3, 0.7, 0.2	0.8, 0.1, 0.0
0.1, 0.4, 1.0	0.9, 0.9, 0.2	1.0, 0.4, 0.4	1.0, 1.0, 0.5	0.4, 0.8, 0.5
0.0, 0.7, 0.6	0.6, 0.1, 0.9	0.0, 1.0, 0.8	0.2, 0.7, 0.8	0.7, 0.0, 0.2
0.3, 0.0, 0.4	0.9, 0.7, 0.5	0.0, 0.1, 1.0	0.6, 0.2, 0.4	0.6, 0.4, 0.6
0.0, 0.2, 0.4	1.0, 0.9, 0.3	0.8, 0.2, 0.3	1.0, 0.0, 0.6	0.2, 0.1, 0.1
0.4, 0.6, 0.0	0.7, 0.8, 0.8	0.5, 0.7, 0.5	0.7, 0.4, 0.5	0.2, 0.3, 0.4
0.5, 0.1, 0.1	1.0, 0.9, 0.4	0.6, 0.6, 0.3	0.5, 1.0, 0.6	0.8, 0.9, 0.1
0.0, 0.5, 0.3	0.4, 0.5, 0.1	0.6, 0.1, 0.2	0.5, 0.6, 0.7	1.0, 0.9, 0.5
1.0, 0.1, 0.5	0.4, 0.1, 0.6	1.0, 0.4, 0.4	0.9, 0.3, 0.5	0.6, 0.8, 0.0
0.7, 0.9, 0.3	0.6, 0.5, 0.0	0.3, 0.9, 0.6	0.3, 0.6, 0.7	0.0, 0.0, 0.5
1.0, 0.8, 0.1	0.0, 0.3, 0.0	0.4, 1.0, 0.3	0.7, 0.3, 1.0	0.4, 0.6, 1.0
0.1, 0.5, 0.3	0.8, 0.1, 0.1	0.5, 0.1, 0.2	0.1, 0.4, 1.0	0.6, 0.4, 0.7
0.3, 0.2, 0.3	0.9, 0.9, 0.4	0.7, 0.8, 0.1	0.6, 0.8, 0.0	0.4, 0.0, 0.2
0.1, 0.3, 0.0	0.2, 0.7, 0.9	0.8, 0.4, 0.8	0.0, 0.1, 0.0	0.6, 0.7, 0.8
0.7, 0.5, 0.3	0.0, 0.2, 0.6	0.5, 0.5, 0.2	0.3, 0.9, 1.0	0.8, 0.2, 0.7
1.0, 0.6, 0.8	0.5, 0.5, 0.0	0.8, 0.4, 1.0	0.1, 0.4, 0.8	1.0, 0.2, 0.9
0.1, 0.5, 0.7	0.2, 0.0, 0.6	0.3, 0.6, 0.8	0.8, 1.0, 0.1	0.5, 0.0, 0.5
1.0, 0.5, 0.4	0.3, 0.2, 1.0	0.4, 0.0, 0.5	0.8, 0.9, 0.7	0.5, 0.1, 0.7

viewing angle



object pose

lighting



occlusion



scale



environment



seat

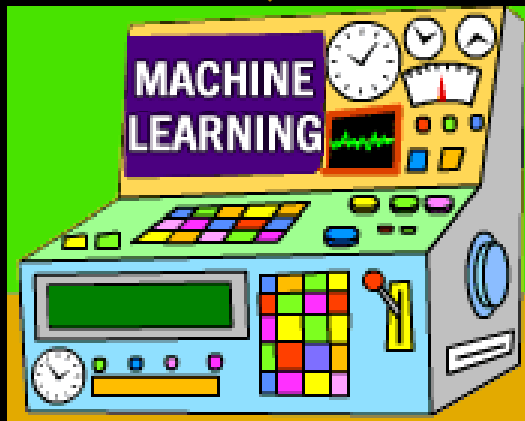


not a seat



training set

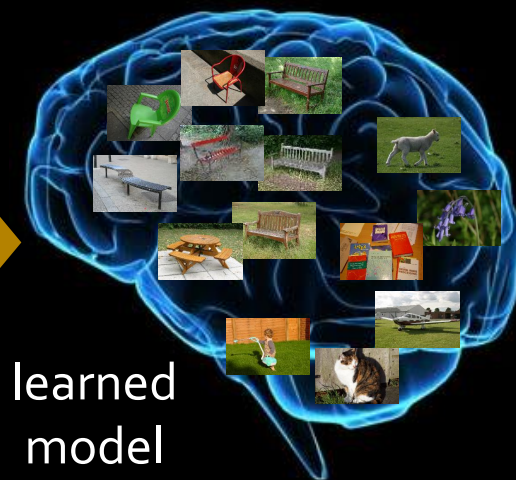
Image
Features



test image



test image



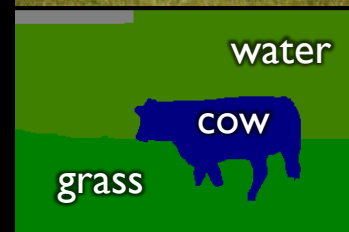
learned
model

seat

not a
seat



[Shotton, Blake, Cipolla 05]



[Shotton, Winn, Rother, Criminisi 06 + 08]
[Winn & Shotton 06]



[Shotton, Johnson, Cipolla 08]

A call to action

Thu 11/09/2008 20:19

Hi Jamie,

I work on Xbox Incubation and I noticed some work you've done on visual recognition using contours (<http://jamie.shotton.org/work/research.html>). I was hoping to be able to discuss an important scenario we are trying to solve with you. Would you be able to chat?

Thanks,

- Mark

A call to action

Thu 11/09/2008 21:50

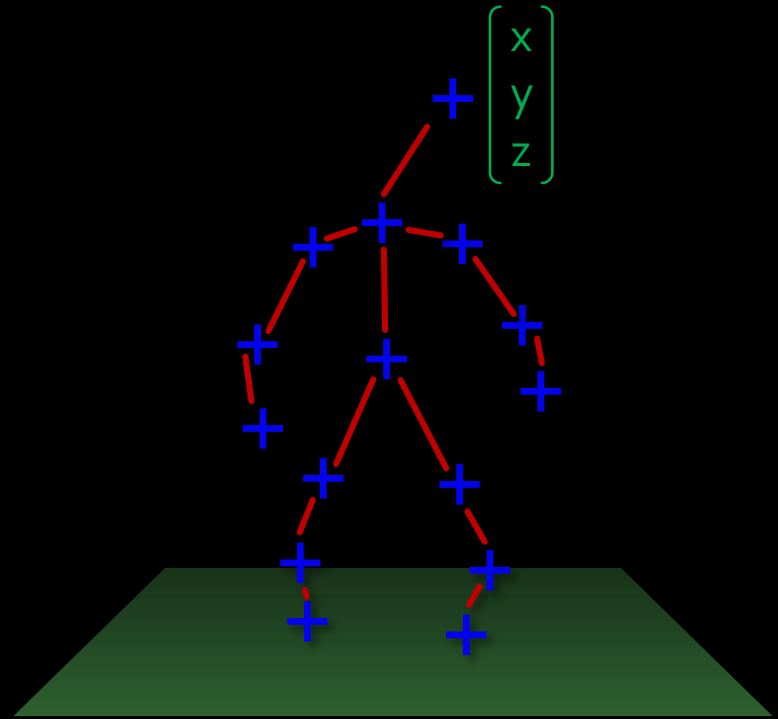
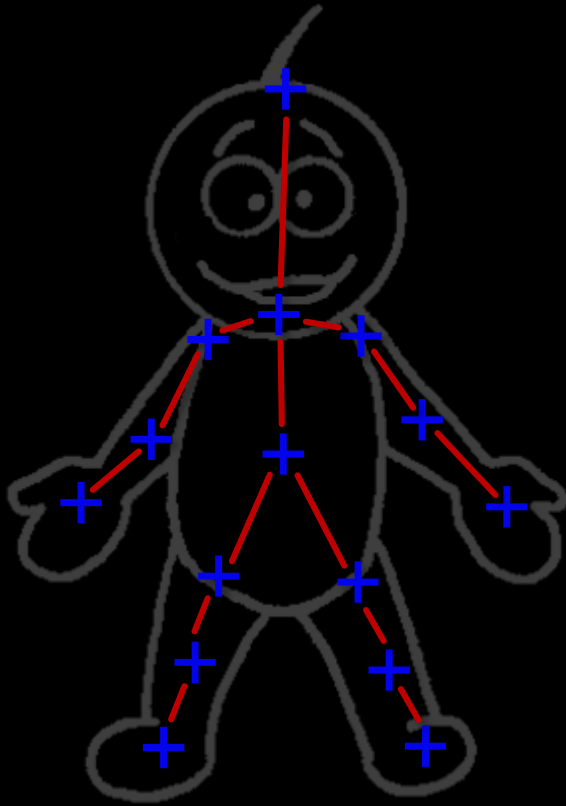
Hey Jamie,

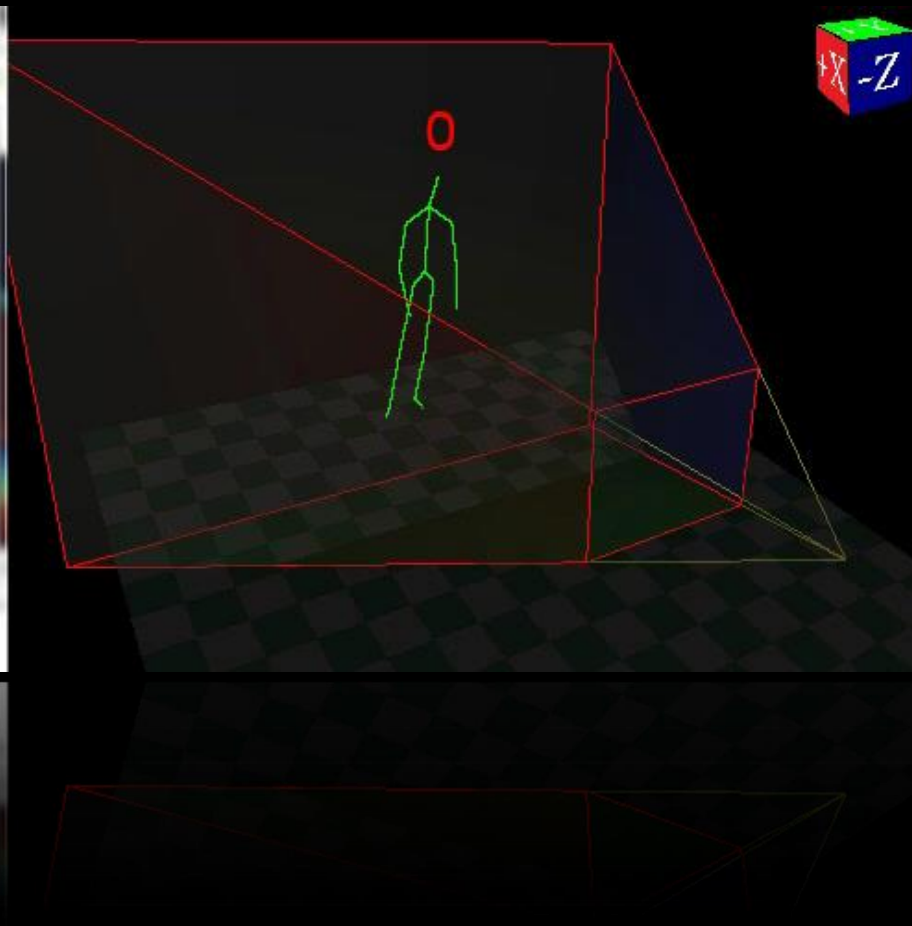
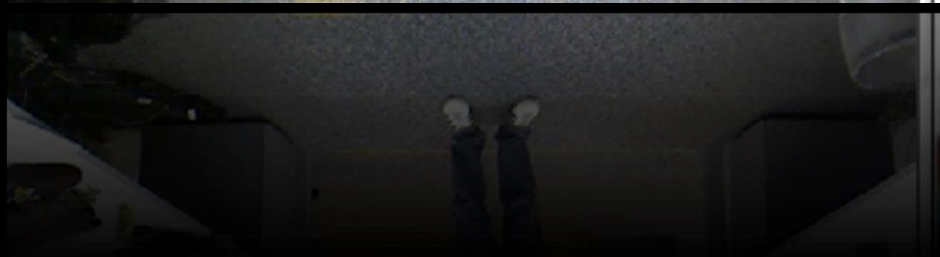
Can you talk right now? 😊

- Mark

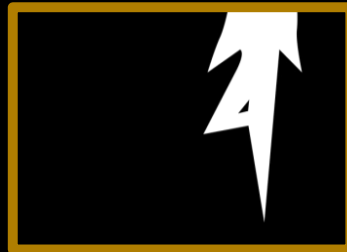
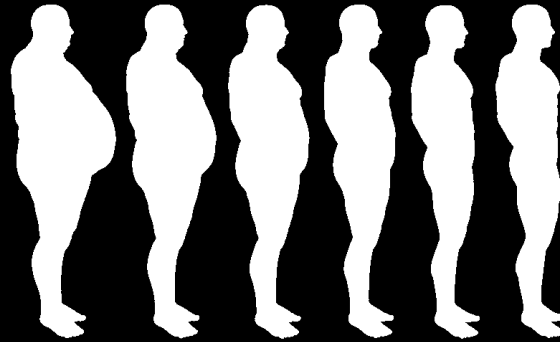


Human pose estimation





Why is it hard?



Motion capture



[Vicon]



[Xsens]

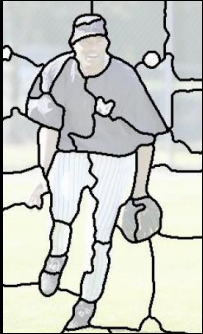
- ✓ very accurate
- ✓ high frame rate

- ✗ suit / sensors
- ✗ expensive

- ✗ large space
- ✗ calibration

Computer vision approaches

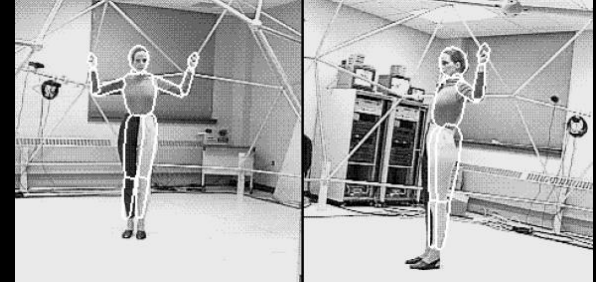
[Mori *et al.* 04]



Monocular,
natural
images

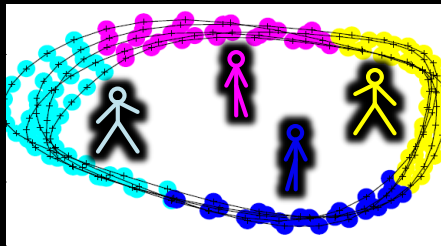


Stereo &
3D
images



[Gavrila & Davis 96]

[Agarwal & Triggs 04]



Tracking
motion

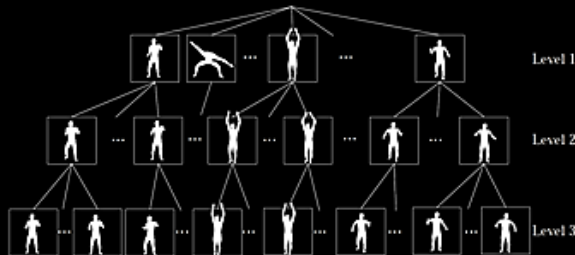


Frame-
by-frame



[Bourdev & Malik 09]

[Okada & Stenger 08]



Whole
body



Parts
models



[Fischler & Elschlager 73]

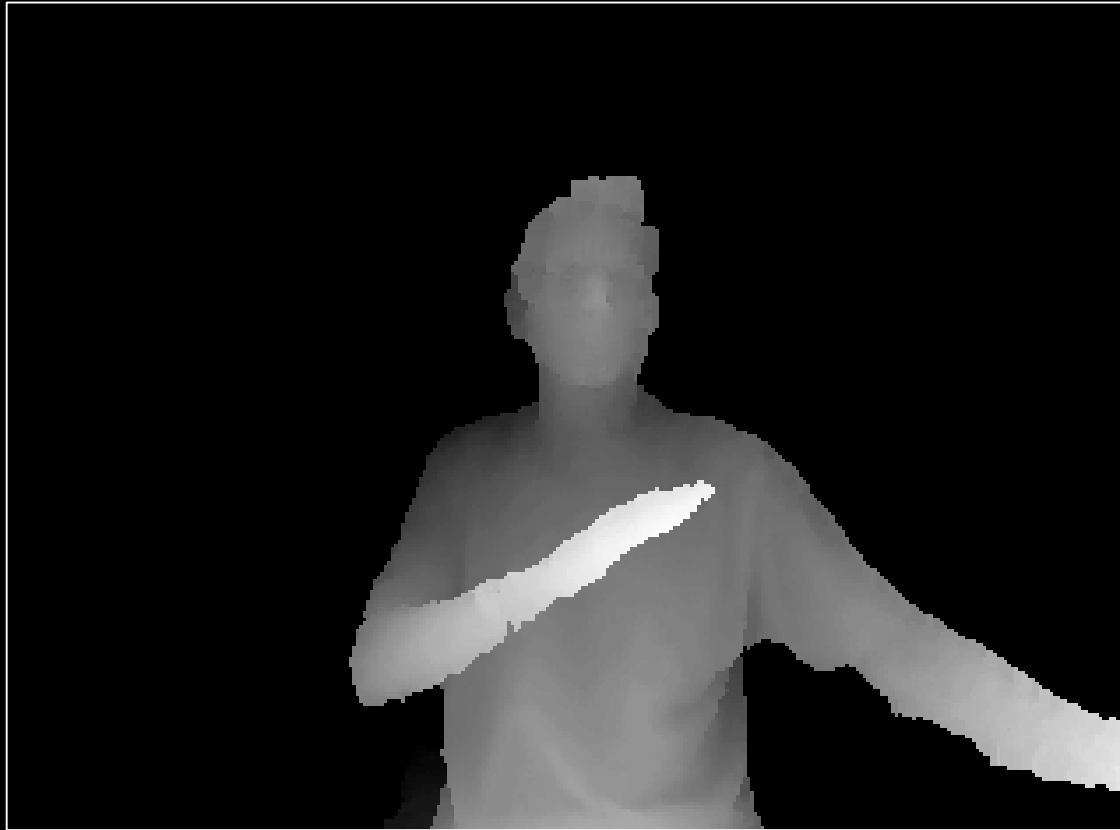
Requirements



- Human pose estimation
 - any pose
 - any body shape & size
- No calibration or instrumentation of the user
- Must “never fail”
- Must run at real time in 10% of Xbox 360 (2005 era hardware)
- Must ship “Holiday 2010”



The depth camera



depth sensor

infrared
emitter

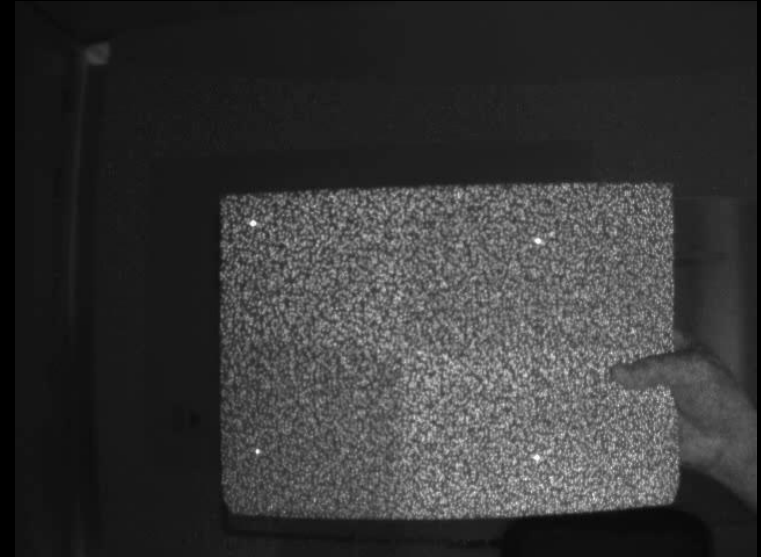
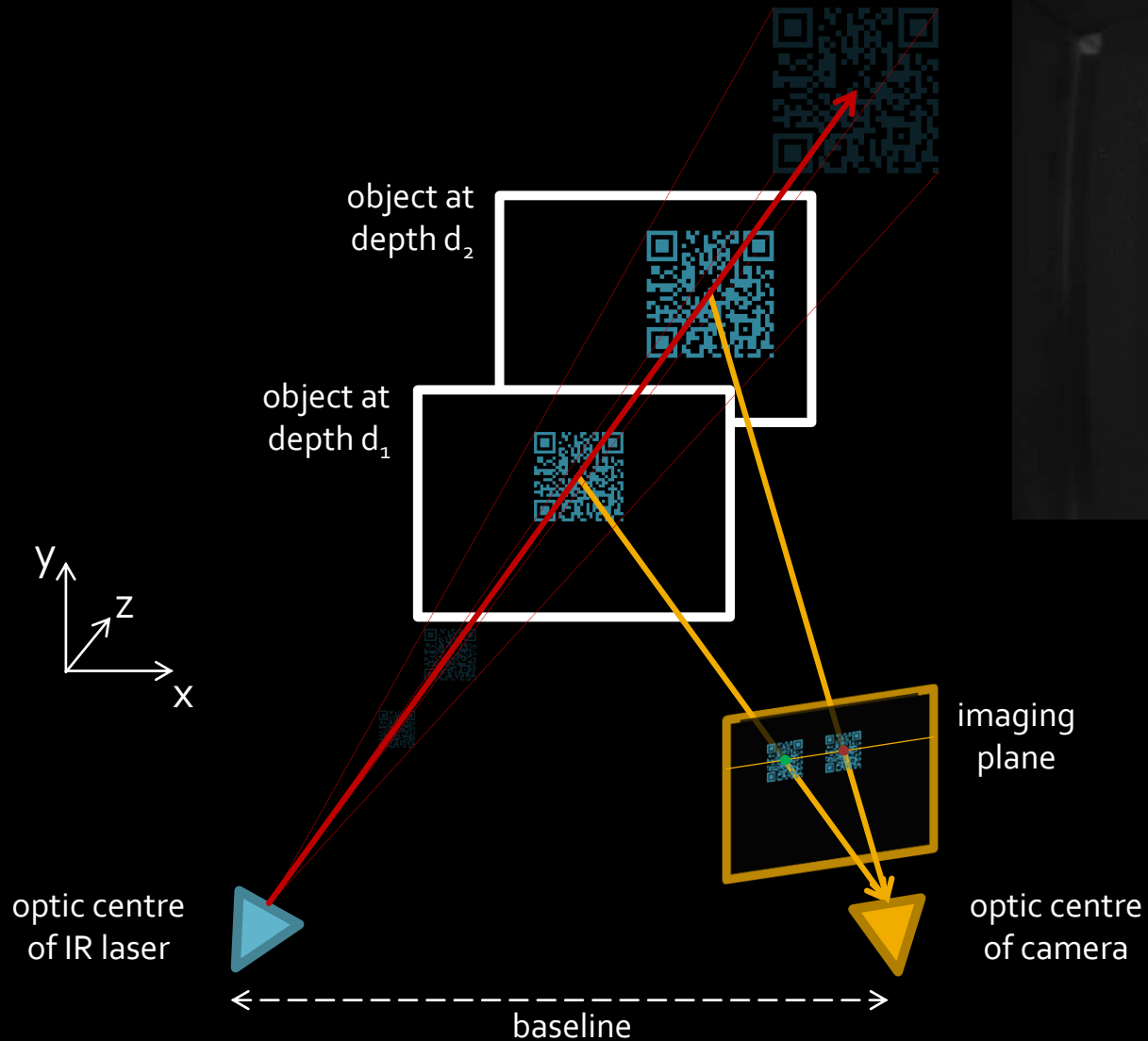
infrared
camera

XBOX 360

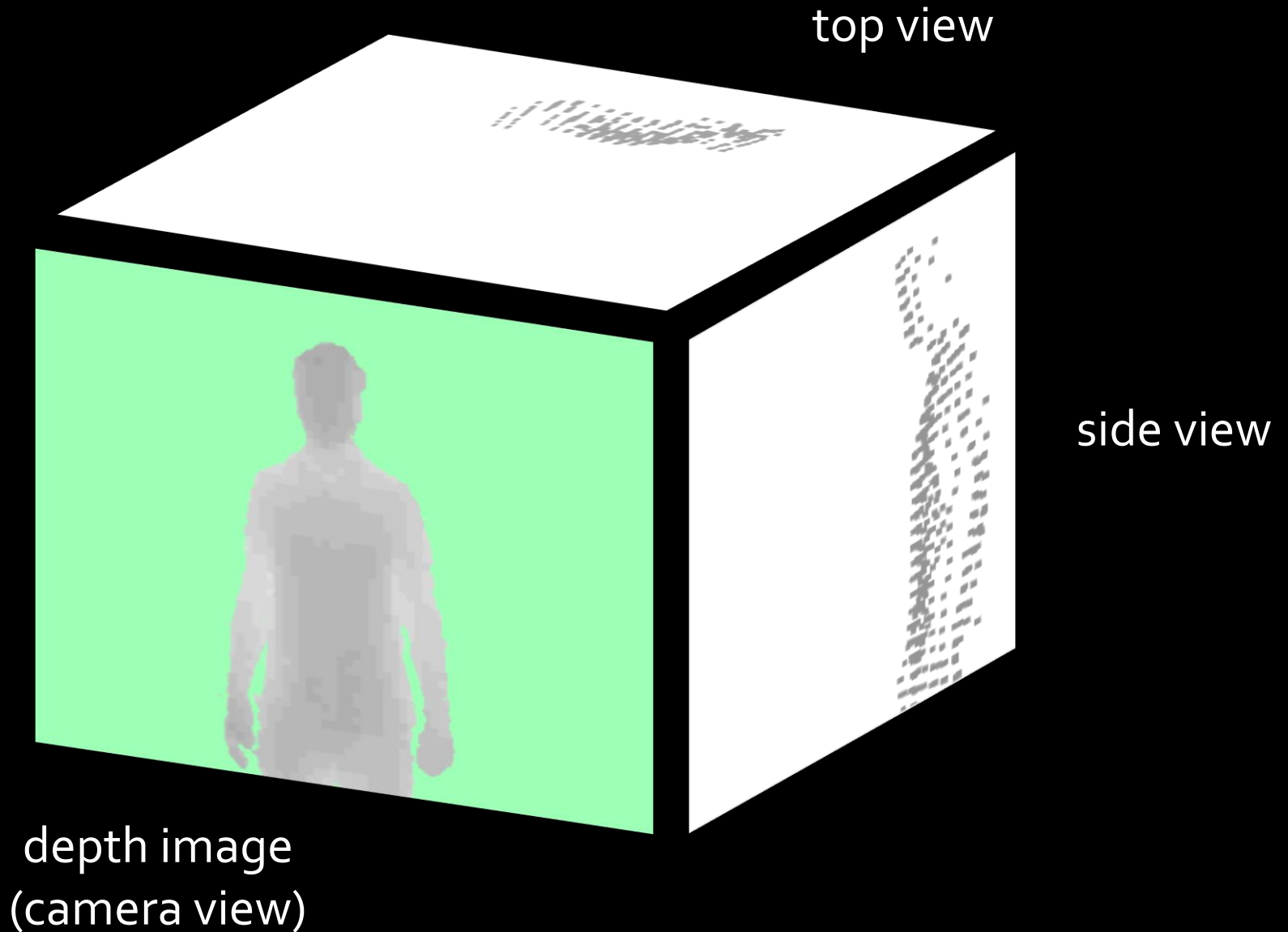
RGB
camera



Structured light



The Kinect camera



RGB vs depth for pose estimation

RGB

- ✗ Only works well lit
- ✗ Background clutter
- ✗ Scale unknown
- ✗ Clothing & skin colour

DEPTH

- ✓ Works in low light
- ✓ Person 'pops' out from bg
- ✓ Scale known
- ✓ No colour or texture variation

Related work using 3D input



[Anguelov *et al.* 05]



[Grest *et al.* 05]



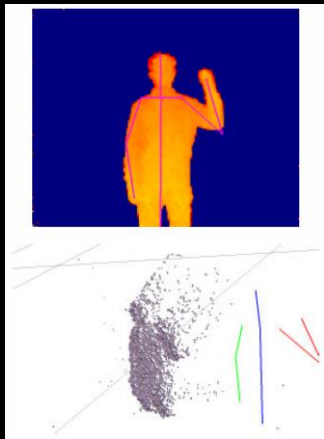
[Knoop *et al.* 06]



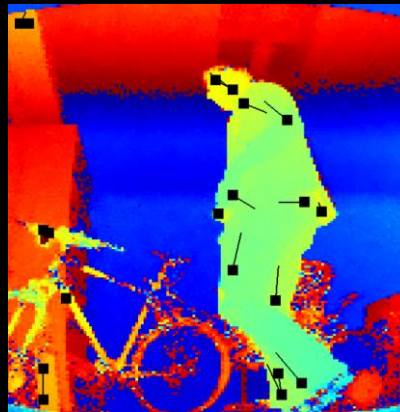
[Zhu & Fujimura 07]



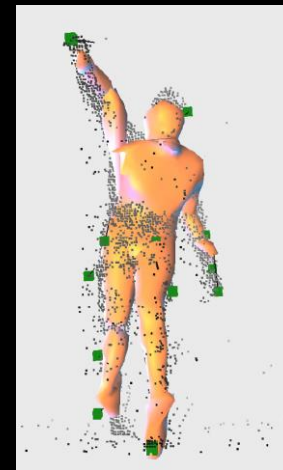
[Kalogerakis
et al. 10]



[Siddiqui &
Medioni 10]



[Plagemann *et al.* 10]

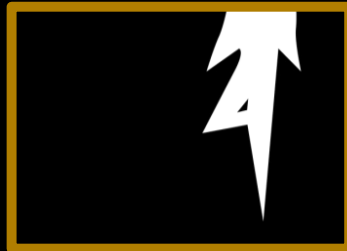
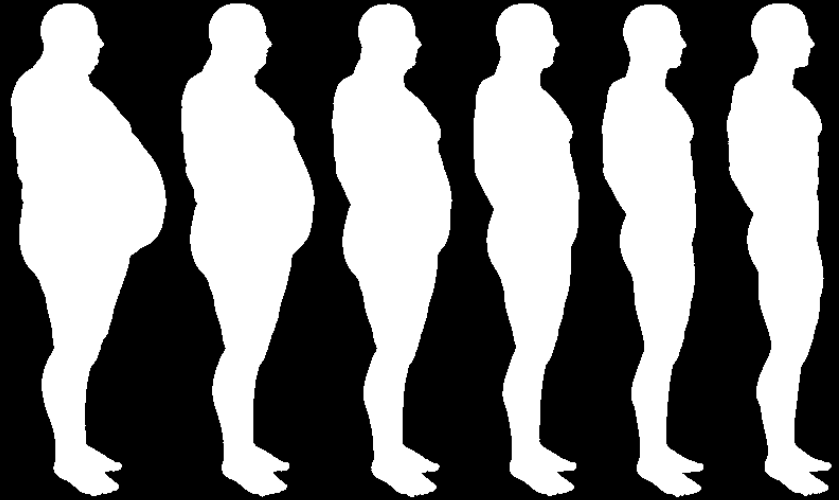


[Ganapathi
et al. 10]

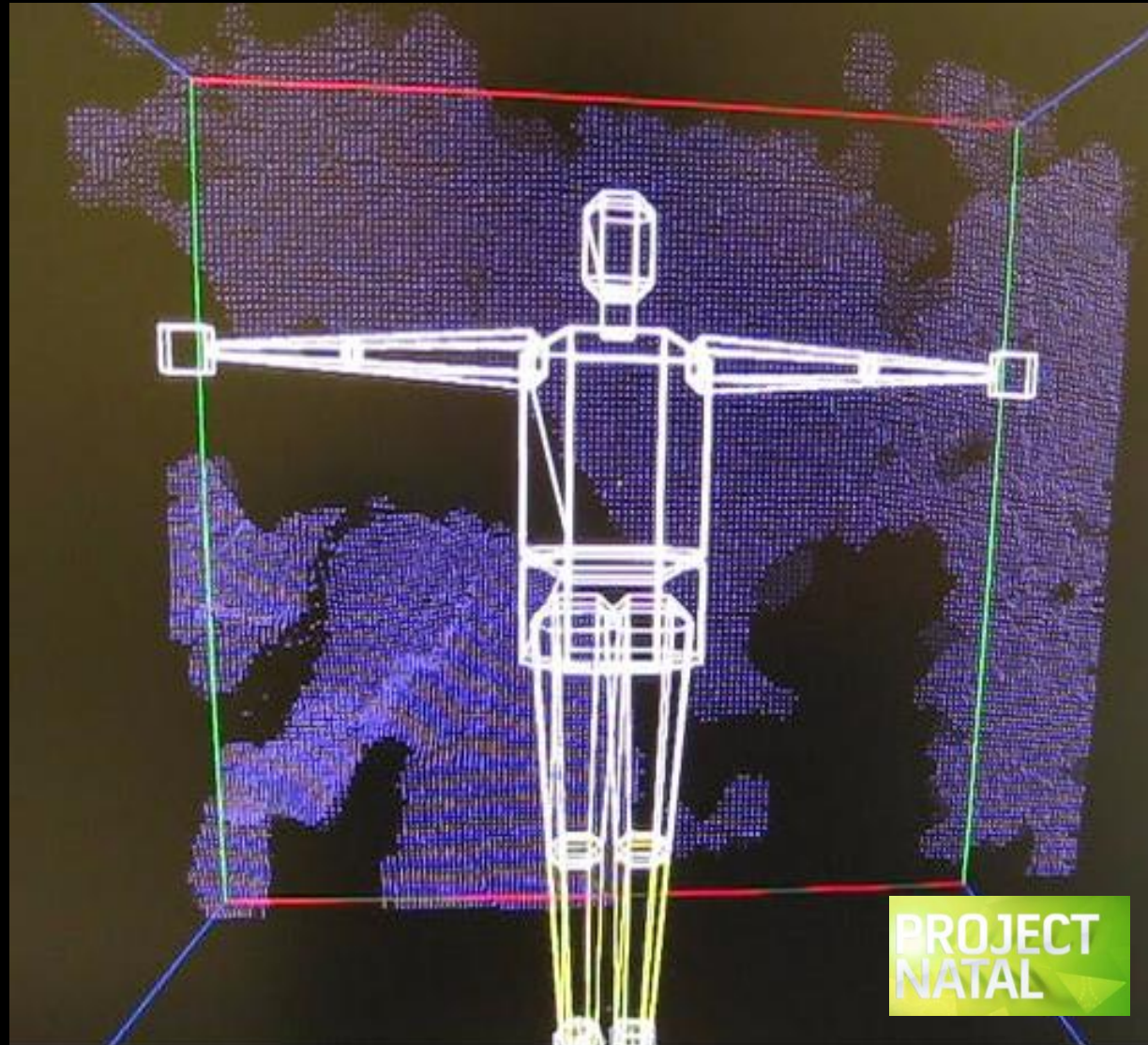


[Baak
et al. 11]

Problems remaining

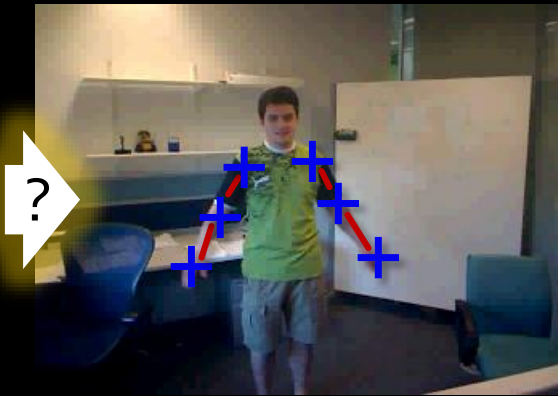


Tracking

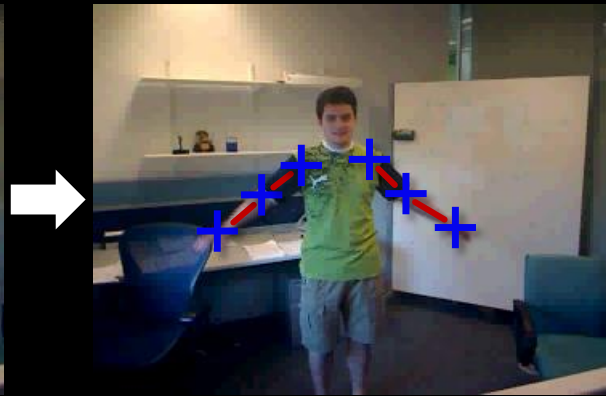


Tracking

time t



time $(t+1)$



slow movement

time $(t+2)$



faster movement

✓ Smooth, realistic output

✗ Initialisation

✗ Prone to catastrophic failure

Our mission

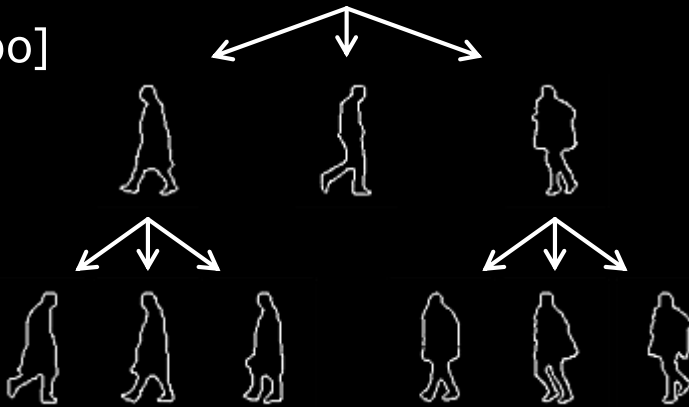


- Auto-initialise tracking algorithm
- Detect and recover from failures

Matching whole poses

- ✓ An image of a *whole* person is very indicative of the person's pose

[Gavrila 00]



[Okada & Stenger 08]



- ✗ Massive search space of whole poses
 - exponential in number of joints
 - hard to scale up

Matching whole poses

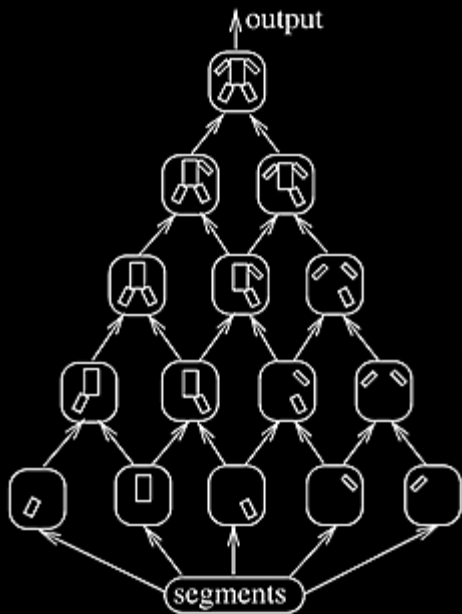


- ❌ Massive search space of whole poses
 - exponential in number of joints
 - hard to scale up

Parts-based models

- Find parts of the body separately
- Stitch them together efficiently

[Ioffe & Forsyth 01]



[Mori *et al.* 04]

[Fischler & Elschlager 73]

[Felzenszwalb & Huttenlocher 05]

[Ferrari *et al.* 08]



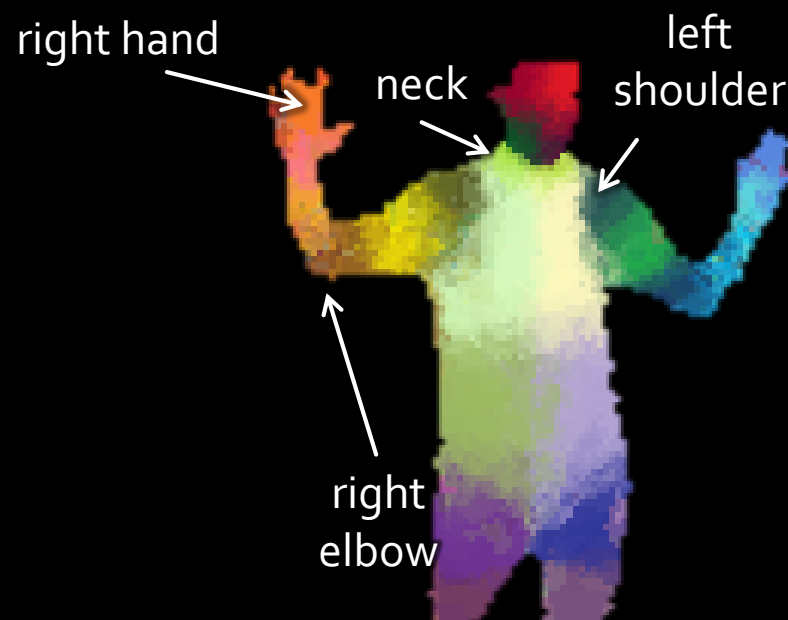
[Bourdev & Malik 09]

Our solution: body part recognition

- Local pose estimate of parts
 - each pixel & each body joint treated independently
 - reduced training data and computation time

- No temporal information
 - frame-by-frame

- Very fast
 - simple depth image features
 - parallel decision forest classifier



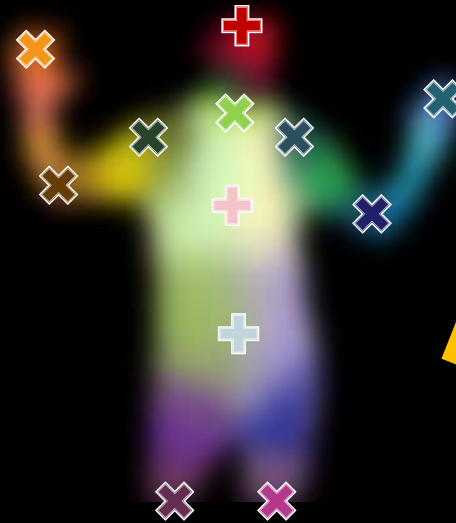
The Kinect pose estimation pipeline



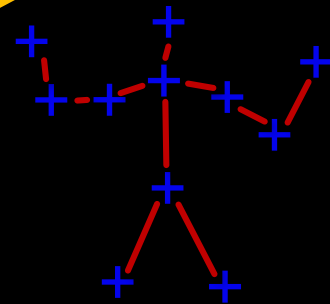
capture
depth image



infer
body parts
per pixel



cluster pixels
predictions
into body joint
hypotheses



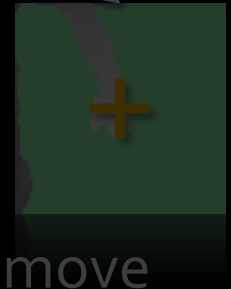
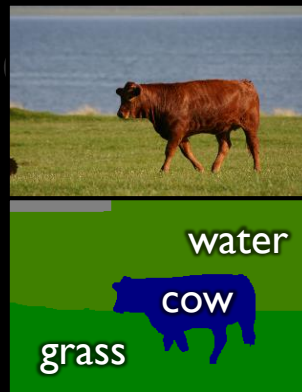
fit skeleton

Classifying pixels

- Compute $P(c|w)$

- body part c

- in



Train by example to be invariant to:



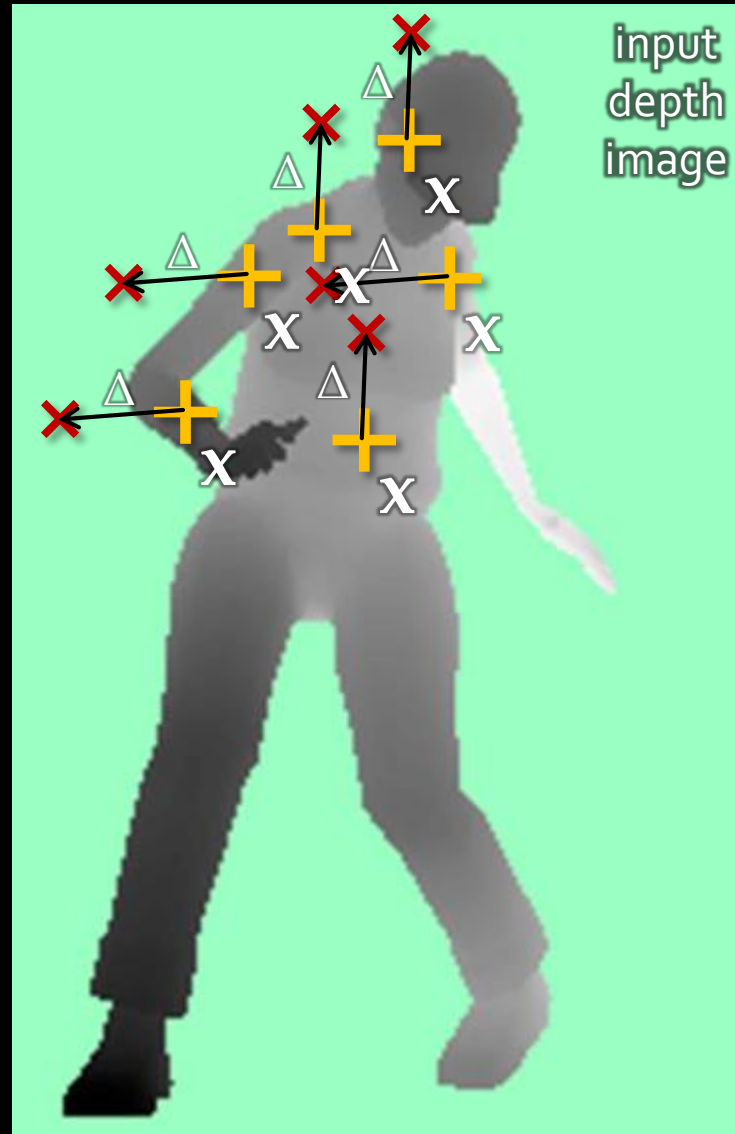
Training data



Fast depth image features

- Depth comparisons
 - very fast to compute

Background pixels
 $d = \text{large constant}$

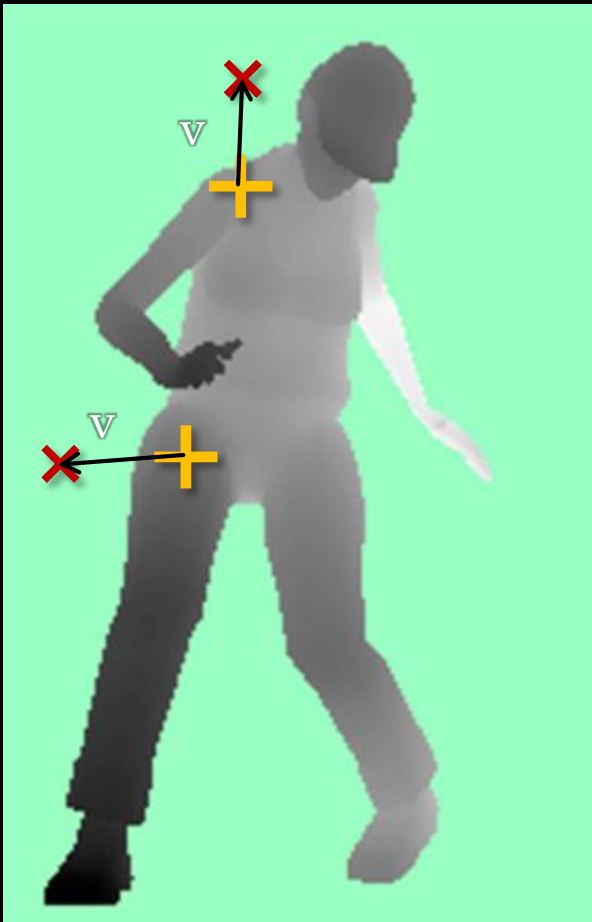


Depth invariance

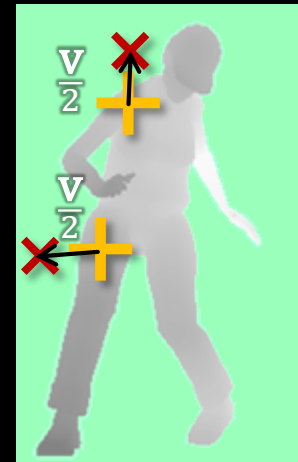
$$f(I, \mathbf{x}) = d_I(\mathbf{x}) - d_I(\mathbf{x} + \Delta)$$

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

scales inversely with depth

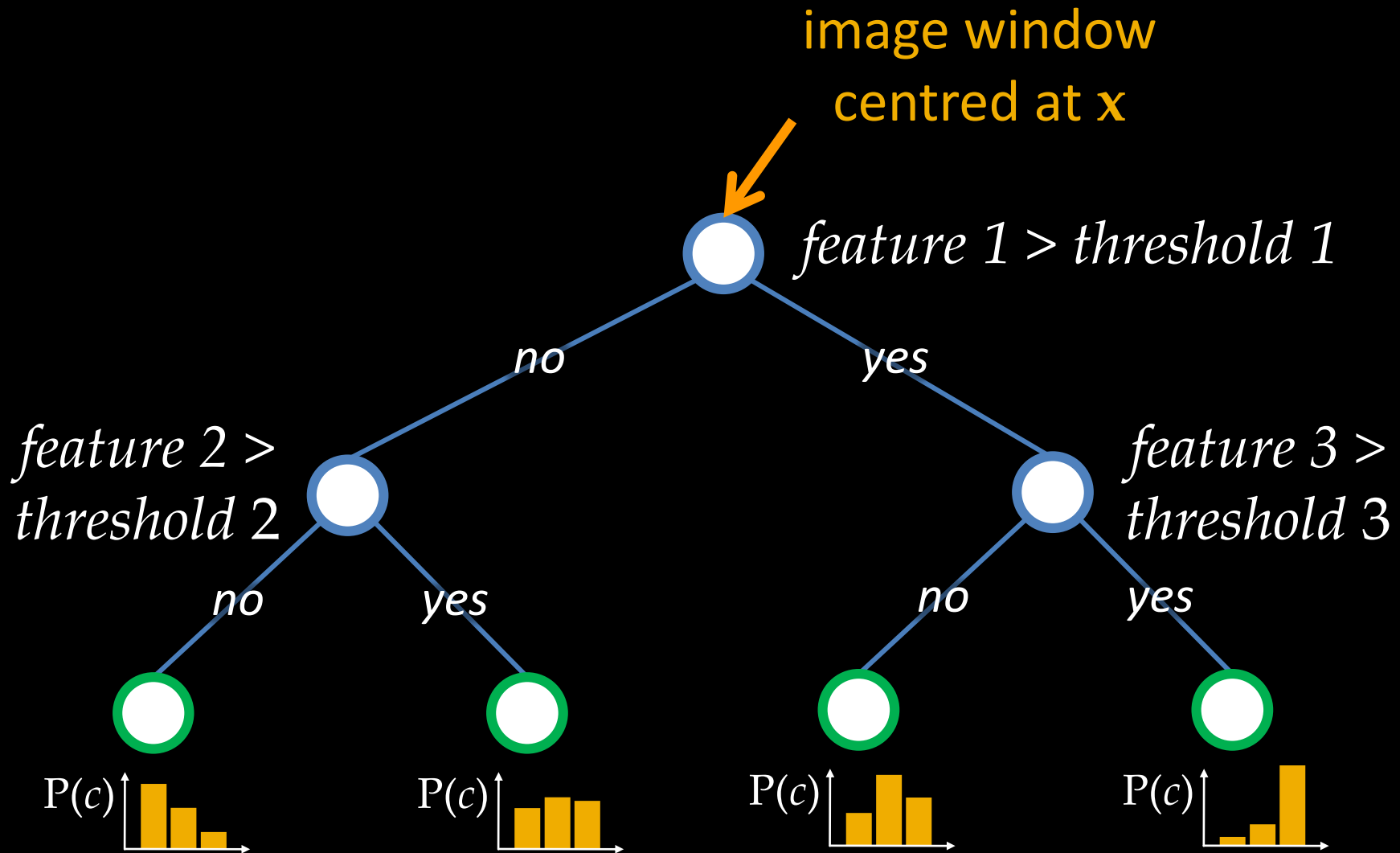


Average
depth 1m



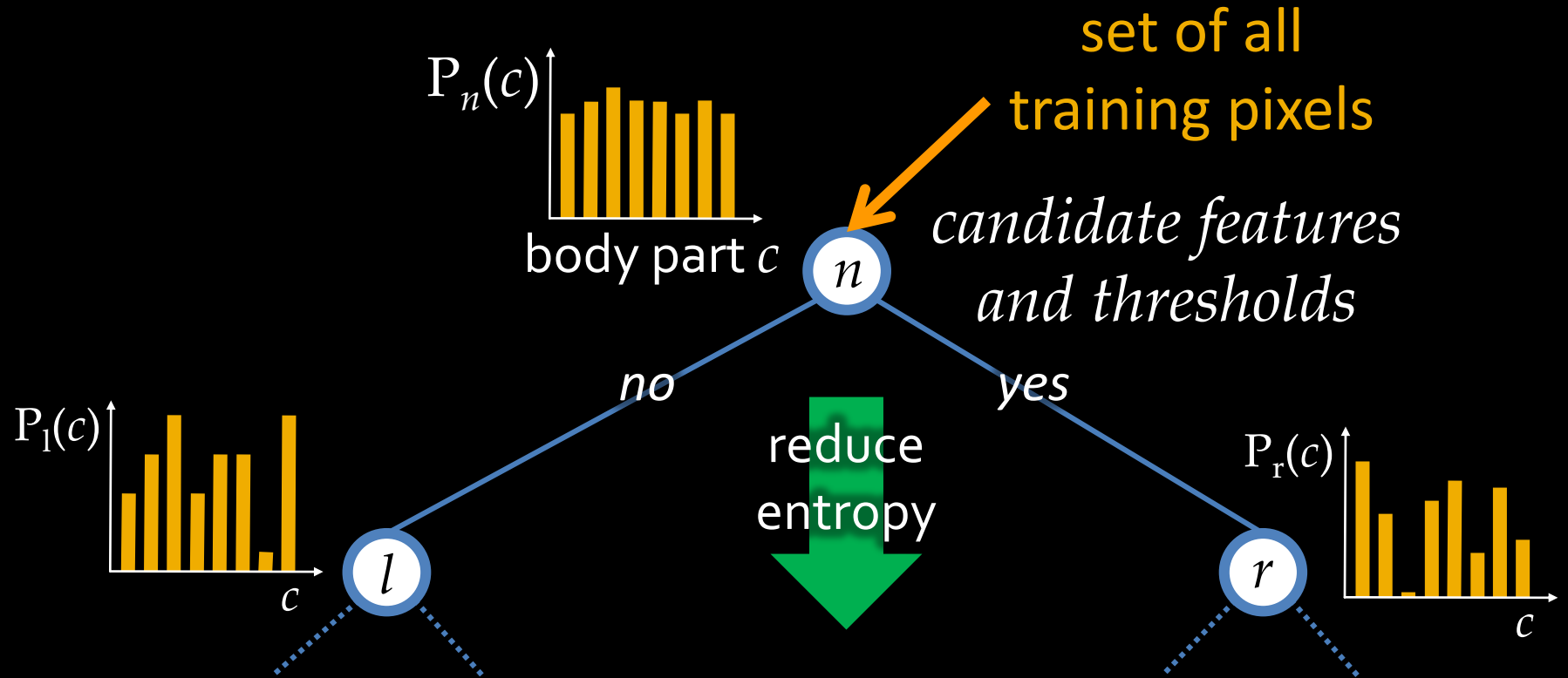
Average
depth 2m

Decision tree classification



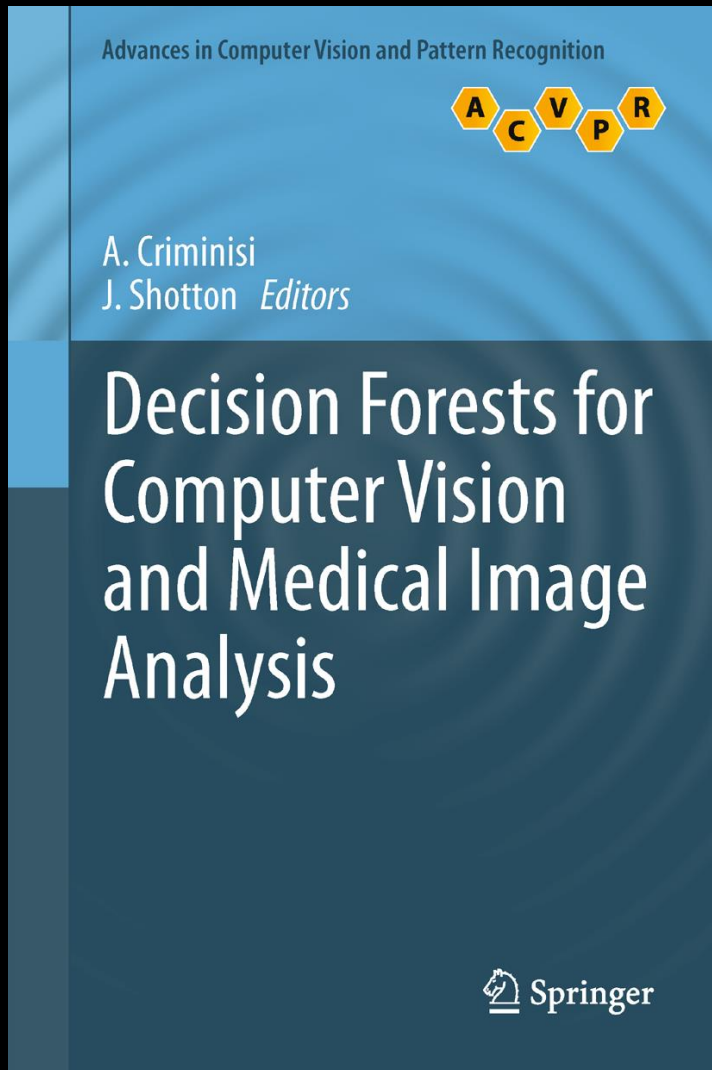
Training decision trees

[Breiman *et al.* 84]



Take (*feature, threshold*) that most reduces Shannon Entropy

Goal: drive entropy at leaf nodes to zero



See our new book!

- Theory – Tutorial & Reference
- Practice – Invited Chapters
 - Software and Exercises
 - Tricks of the Trade

A. Criminisi & J. Shotton
Springer, 2013

Early results



Scaling up: synthetic training data

Record mocap
100k poses



Retarget to several models



Render (depth, body parts) pairs





Depth of trees

input depth



ground truth parts



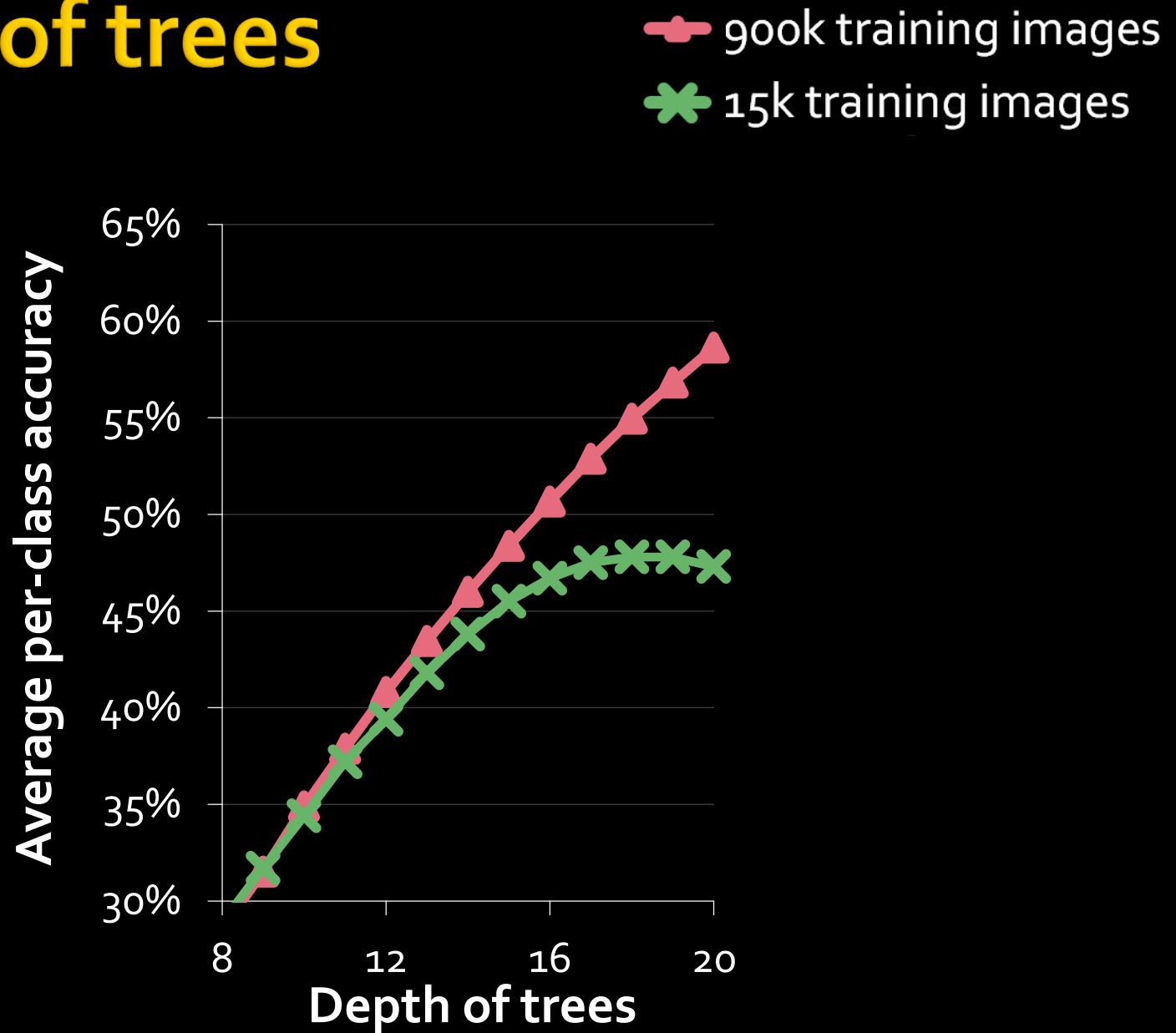
inferred parts (soft)



depth 18



Depth of trees



Scaling up

- 31 body parts
- 3 Trees to depth 20
 - $\sim 3 \times 2^{20}$ nodes
- Training
 - $\sim 1,000,000$ training images
 - $\sim 2,000$ pixels per image
 - $\sim 10,000$ features tested per node
- Very fast at test time
 - only ~ 60 image feature evaluations per pixel
 - readily parallelisable for GPU [Sharp 08]



input depth



inferred body parts



no tracking or smoothing

The Kinect pose estimation pipeline



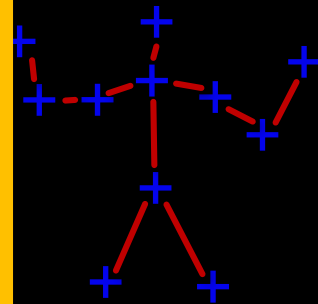
capture
depth image



infer
body parts
per pixel



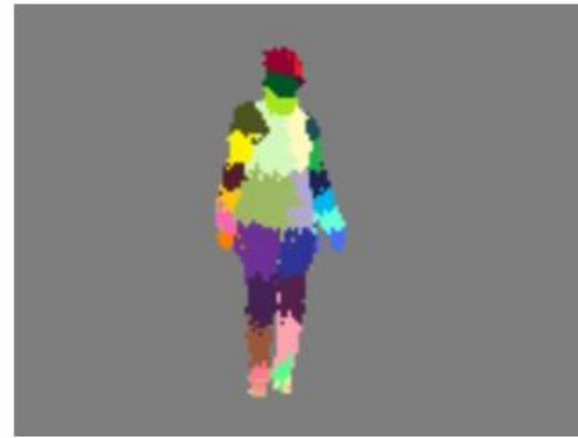
cluster pixel
predictions
into body joint
hypotheses



fit skeleton

input depth

inferred body parts



front view

side view

top view

inferred joint positions

no tracking or smoothing

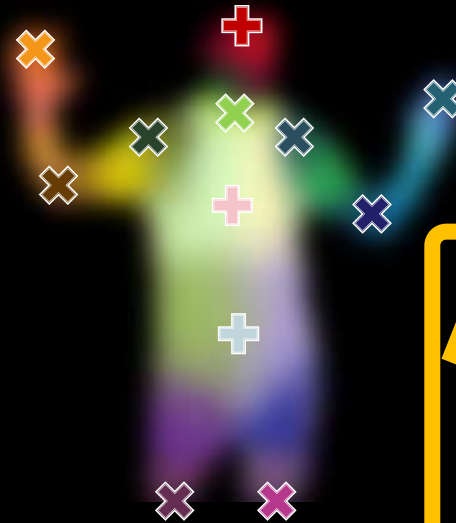
The Kinect pose estimation pipeline



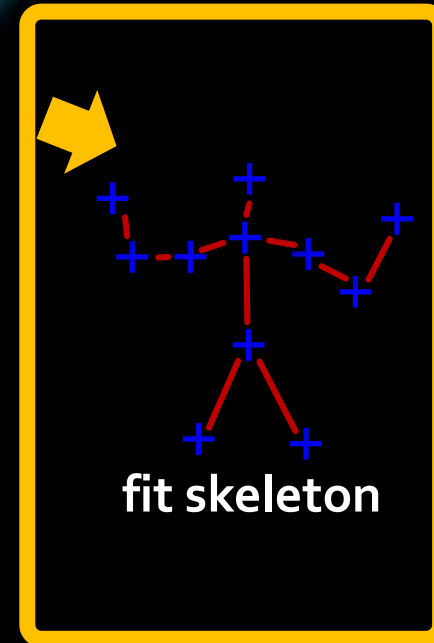
capture
depth image



infer
body parts
per pixel



cluster pixel
predictions
into body joint
hypotheses



KINECT launch



Microsoft Kinect 'fastest-selling device on record'

Microsoft has sold more than 10 million Kinect sensor systems since launch on 4 November, and - according to Guinness World Records - is the fastest-selling consumer electronics device on record.

The sales figures outstrip those of both Apple's iPhone and iPad when launched, Guinness said.

Kinect is an infrared camera add-on for Microsoft's Xbox 360 games console that allows it to track body movements.



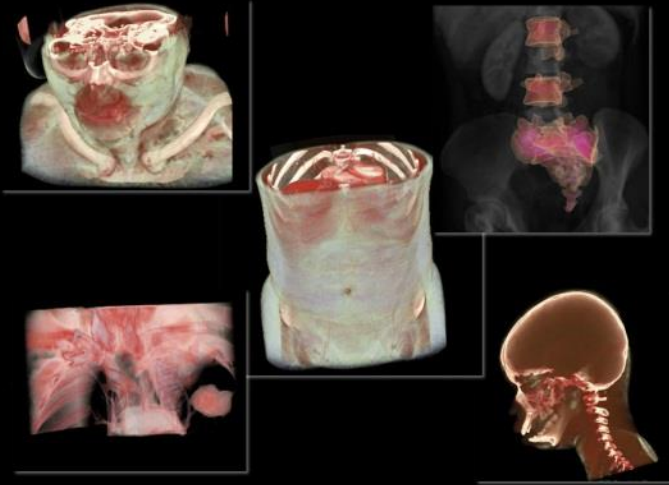
The popularity of the Kinect has helped to boost sales of games, Microsoft says





KINECT™ for Windows®

A new world for research



Home US World Politics Business Sports Entertainment Health Tech


Digital Home on msnbc.com

Child with autism connects with Kinect

When Kyle's father got Xbox's motion control system, he had no idea it would be a breakthrough for his boy

[Jump to video: 'Kinect' surprisingly fun](#)

Below: [Video](#) [Discuss](#) [Related](#)

 **By Wilson Rothman**

msnbc.com
updated 11/11/2010 7:19:30 PM ET

[Print](#) | [Font: A A + -](#)

John Yan reviews games for a site called [Gaming Nexus](#), so despite his initial lack of enthusiasm in the Xbox 360 Kinect motion controller, he knew he'd have to buy one when they came out. After all, it wouldn't be fair to dump all the Kinect reviews on

BBC News Sport Weather Travel Future

NEWS HEALTH

Home UK Africa Asia Europe Latin America Mid-East US & Canada Business Health Sci/Environ

31 May 2012 Last updated at 06:39 GMT [Share](#) [f](#) [t](#) [e](#) [m](#)

Trial of "touchless" gaming technology in surgery

By Adam Brimelow
Health Correspondent, BBC News

Doctors in London are trialling "touchless" technology, often used in TV games, to help them carry out delicate keyhole surgery.

The system allows them to manipulate images with their voice and hand-gestures rather than using a keyboard and mouse.

Surgeons say it gives them more control and avoids disruption.



The technology could be a valuable aid to surgery



KINECTFusion



Joint work with Shahram Izadi, Richard Newcombe, David Kim, Otmar Hilliges, David Molyneaux, Pushmeet Kohli, Steve Hodges, Andrew Davison, Andrew Fitzgibbon. SIGGRAPH, UIST and ISMAR 2011.

Hand Grip/Release Detection



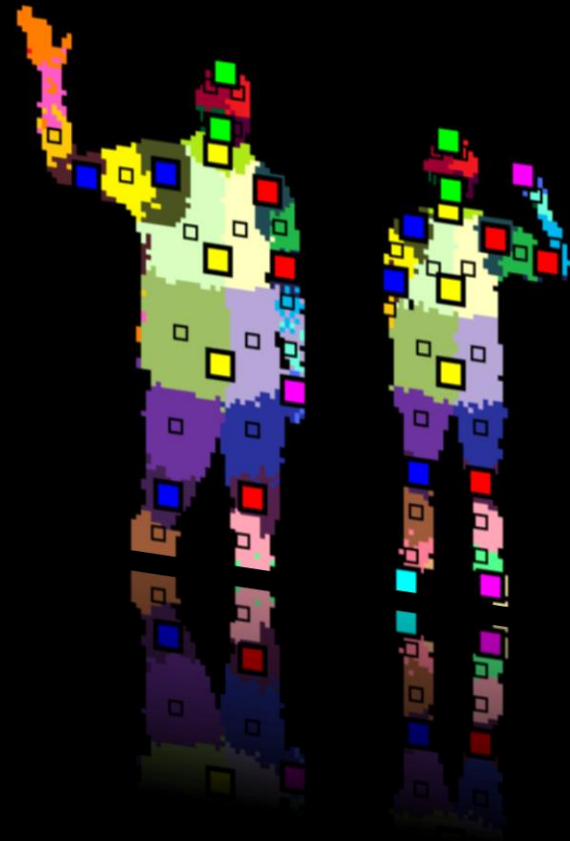
Hand Grip/Release Detection



Take home thoughts

- Blue skies PhD research contributed heavily to Kinect's success
- Machine learning can solve hard problems through big data
- Kinect opening up myriad applications

KINECT



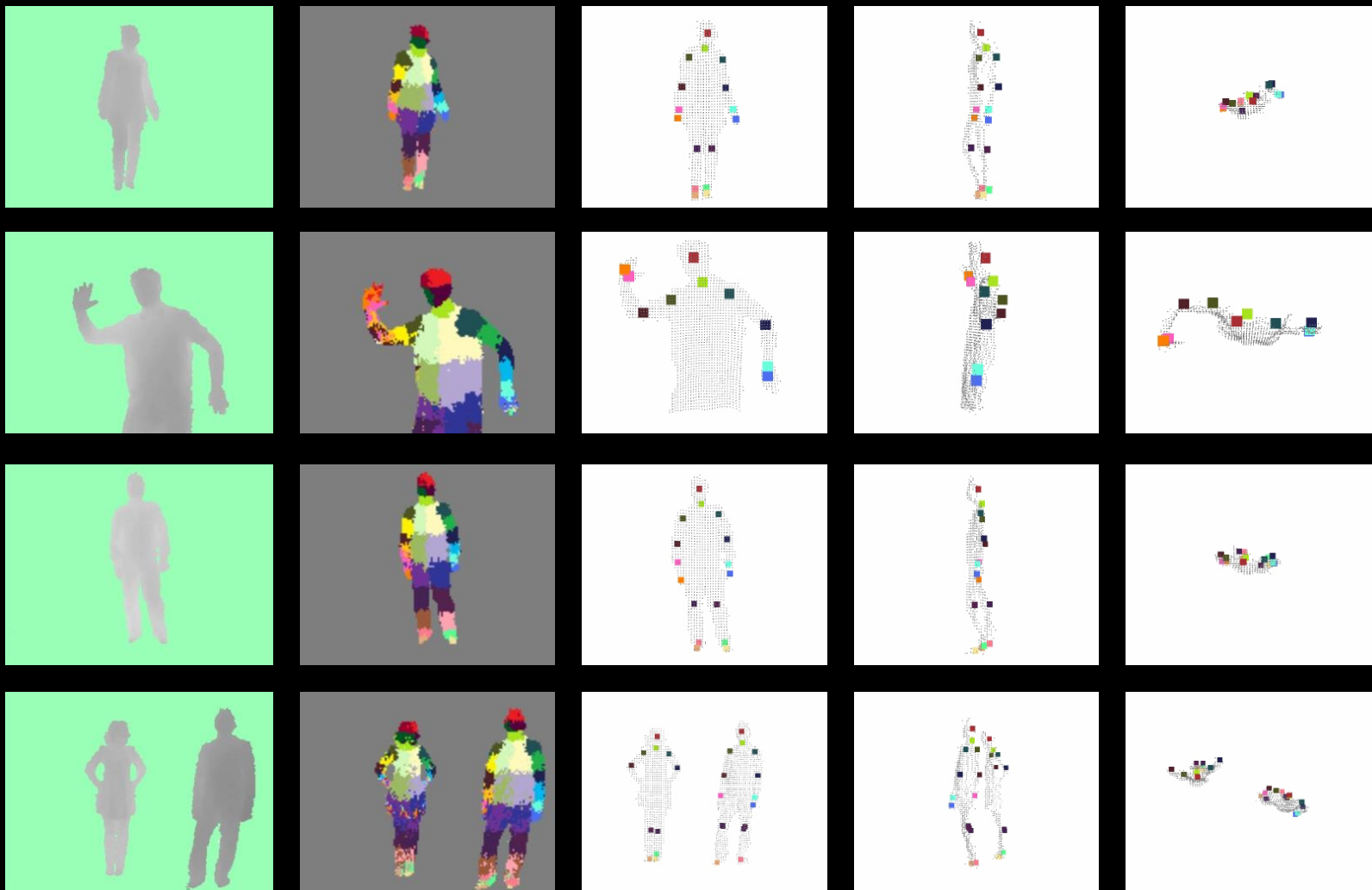
With thanks to:

Microsoft®
Research

Andrew Fitzgibbon, Mat Cook, Andrew Blake, Toby Sharp, Ollie Williams, Sebastian Nowozin, Antonio Criminisi, Mihai Budiu, Ross Girshick, Duncan Robertson, John Winn, Shahram Izadi, Pushmeet Kohli



The whole Kinect team, especially: Mark Finocchio, Alex Kipman, Ryan Geiss, Richard Moore, Robert Craig, Momin Al-Ghosien, Matt Bronder, Craig Peeper



<http://www.microsoft.com/en-us/kinectforwindows/>

Microsoft®
Research

KINECT