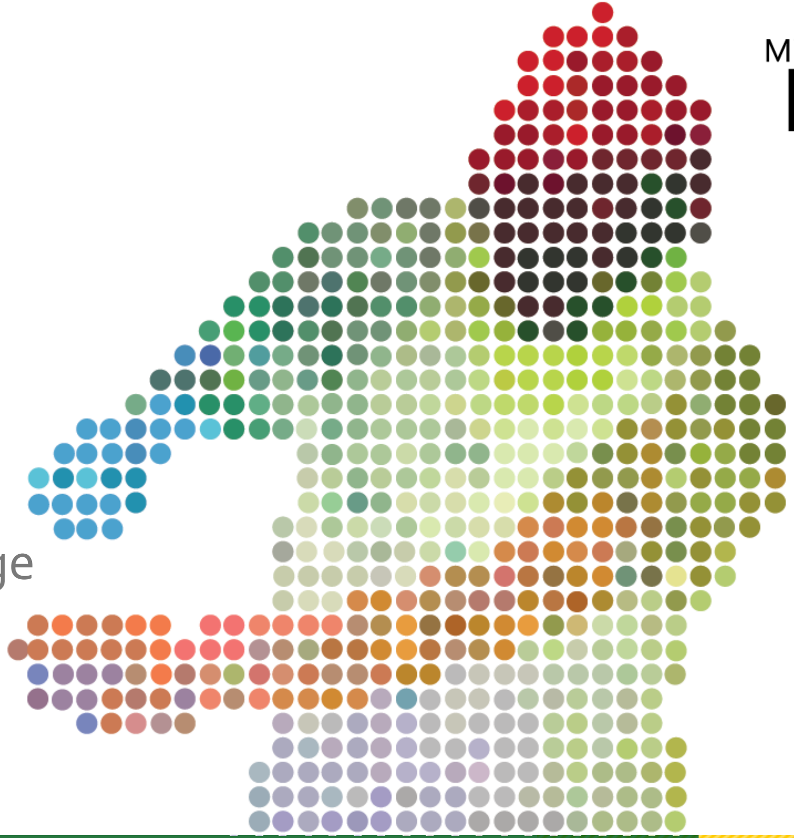


Presenter:
Andrew Fitzgibbon
Principal Researcher
Microsoft Research Cambridge

BLUE SKIES, GROUND TRUTH: MACHINE LEARNING FOR KINECT

MSR OUTPUTS

PATENTS & PAPERS

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
Microsoft Research Cambridge & Xbox Incubation

Abstract

We propose a new method to quickly and accurately predict 3D positions of body joints from a single depth image, using no temporal information. We take an object recognition approach, designing an intermediate body parts representation that maps the difficult pose estimation problem into a simpler per-pixel classification problem. Our large and highly varied training dataset allows the classifier to estimate body parts invariant to pose, body shape, clothing, etc. Finally we generate confidence-scored 3D proposals of several body joints by reprojecting the classification result and finding local modes.

The system runs at 200 frames per second on consumer hardware. Our evaluation shows high accuracy on both synthetic and real test sets, and investigates the effect of several training parameters. We achieve state of the art accuracy in our comparison with related work and demonstrate improved generalization over exact whole-skeleton nearest neighbor matching.

depth image → body parts → 3D joint proposals

Figure 1. **Overview.** From an single input depth image, a per-pixel body part distribution is inferred. (Colors indicate the most likely part labels at each pixel, and correspond in the joint proposals). Local modes of this signal are estimated to give high-quality proposals for the 3D locations of body joints, even for multiple users.

1. Introduction

Robust interactive human body tracking has applications including gaming, human-computer interaction, security, telepresence, and even health-care. The task has recently been greatly simplified by the introduction of real-time depth cameras [16, 19, 44, 37, 28, 13]. However, even the best existing systems still exhibit limitations. In particular, until the launch of Kinect [21], none ran at interactive rates on consumer hardware while handling a full range of human body shapes and sizes undergoing general body motions. Some systems achieve high speeds by tracking from frame to frame but struggle to re-initialize quickly and so are not robust. In this paper, we focus on pose recognition in parts: detecting from a single depth image a small set of 3D position candidates for each skeletal joint. Our focus on per-frame initialization and recovery is designed to complement any appropriate tracking algorithm [7, 39, 16, 42, 13] that might further incorporate temporal and kinematic coherence. The algorithm presented here forms a core component of the Kinect gaming platform [21].

Illustrated in Fig. 1 and inspired by recent object recognition work that divides objects into parts (*e.g.* [12, 43]), our approach is driven by two key design goals: computational efficiency and robustness. A single input depth image is segmented into a dense probabilistic body part labeling, with the parts defined to be spatially localized near skeletal joints of interest. Reprojecting the inferred parts into world space, we localize spatial modes of each part distribution and thus generate (possibly several) confidence-weighted proposals for the 3D locations of each skeletal joint.

We treat the segmentation into body parts as a per-pixel classification task (no pairwise terms or CRF have proved necessary). Evaluating each pixel separately avoids a combinatorial search over the different body joints, although within a single part there are of course still dramatic differences in the contextual appearance. For training data, we generate realistic synthetic depth images of humans of many shapes and sizes in highly varied poses sampled from a large motion capture database. We train a deep randomized decision forest classifier which avoids overfitting by using hundreds of thousands of training images. Simple, discriminative depth comparison image features yield 3D translation invariance while maintaining high computational efficiency. For further speed, the classifier can be run in parallel on each pixel on a GPU [34]. Finally, spatial modes of the inferred per-pixel distributions are computed using mean shift [10] resulting in the 3D joint proposals.

An optimized implementation of our algorithm runs in under 5ms per frame (200 frames per second) on the Xbox 360 GPU, at least one order of magnitude faster than existing approaches. It works frame-by-frame across dramatically differing body shapes and sizes, and the learned discriminative approach naturally handles self-occlusions and

1

TECHNOLOGY TRANSFERS

Innovation into Products

IP Licensing to 3rd parties

Expertise for Products

Microsoft®

"GrabCut" — Interactive Foreground Extraction using Iterated Graph Cuts

Carsten Rother* Vladimir Kolmogorov† Andrew Blake‡
Microsoft Research Cambridge, UK

Figure 1: **Three examples of GrabCut .** The user drags a rectangle loosely around an object. The object is then extracted automatically.

Abstract

The problem of efficient, interactive foreground/background segmentation in still images is of great practical importance in image editing. Classical image segmentation tools use either texture (colour) information, e.g. Magic Wand, or edge (contrast) information, e.g. Intelligent Scissors. Recently, an approach based on optimization by graph-cut has been developed which successfully combines both types of information. In this paper we extend the

free of colour blee
degrees of interacti
the labour-intensive
background in a fe

1.1 Previous a

In the following we describe briefly and compare several state of the art interactive tools for segmentation: Magic Wand, Intelligent

SIGGRAPH2004

Computer Vision at MSRC

Microsoft

Microsoft® Office® 2010

COMPUTER V ... MSRC

Microsoft®

unwrap mosaics

Rav-Acha | Kohli | Rother | Fitzgibbon
http://research.microsoft.com/unwrap

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

[Shotton, Johnson, Cipolla 08]

COMPUTER VISION AT MSRC

Ground truth | Entangled | Conventional

A. Montillo, J. Shotton, J. Winn, J. E. Iglesias, D. Metaxas, and A. Criminisi, *Entangled Decision Forests and their Application for Semantic Segmentation of CT Images*, in *Information Processing in Medical Imaging (IPMI)*, July 2011

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
Microsoft Research Cambridge & Xbox Incubation

Abstract

We propose a new method to quickly and accurately predict 3D positions of body joints from a single depth image, using no temporal information. We take an object recognition approach, designing an intermediate body parts representation that maps the difficult pose estimation problem into a simpler per-pixel classification problem. Our large ... dataset allows the classifier to ... nt to pose, body shape, clothing, ... nfidence-scored 3D proposals of ... ojecting the classification result ... frames per second on consumer ... shows high accuracy on both synthetic and real test sets, and investigates the effect of several training parameters. We achieve state of the art accu-

CVPR 11
20 COLORADO 11

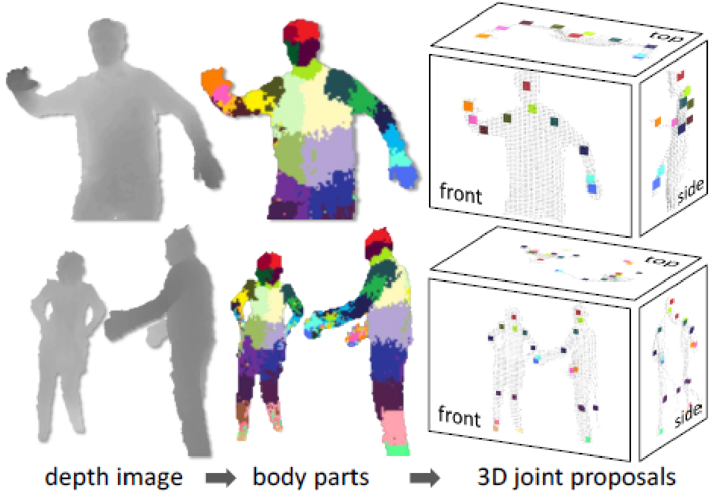


Figure 1. **Overview.** From an single input depth image, a per-pixel body part distribution is inferred. (Colors indicate the most likely part labels at each pixel, and correspond in the joint proposals).

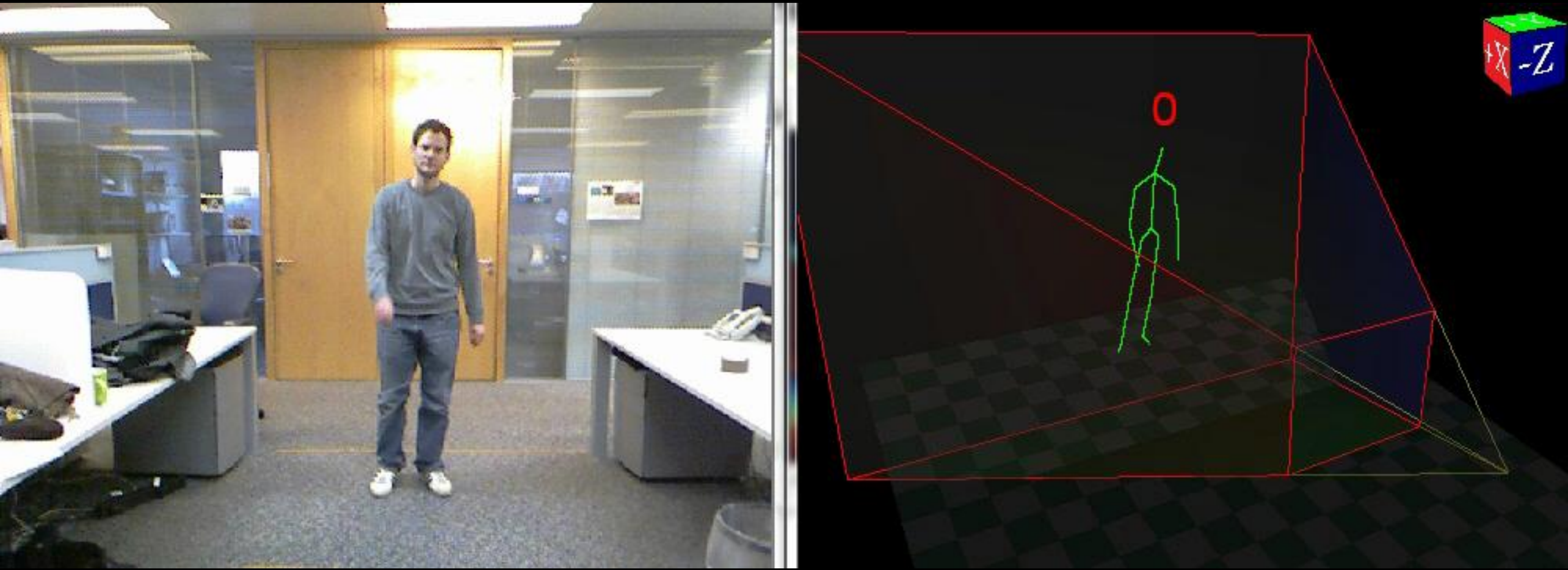
From: Mark Finocchio
To: Jamie Shotton
Date: 11 Sept 2008
Subject: Your computer vision expertise

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



THE SCENARIO

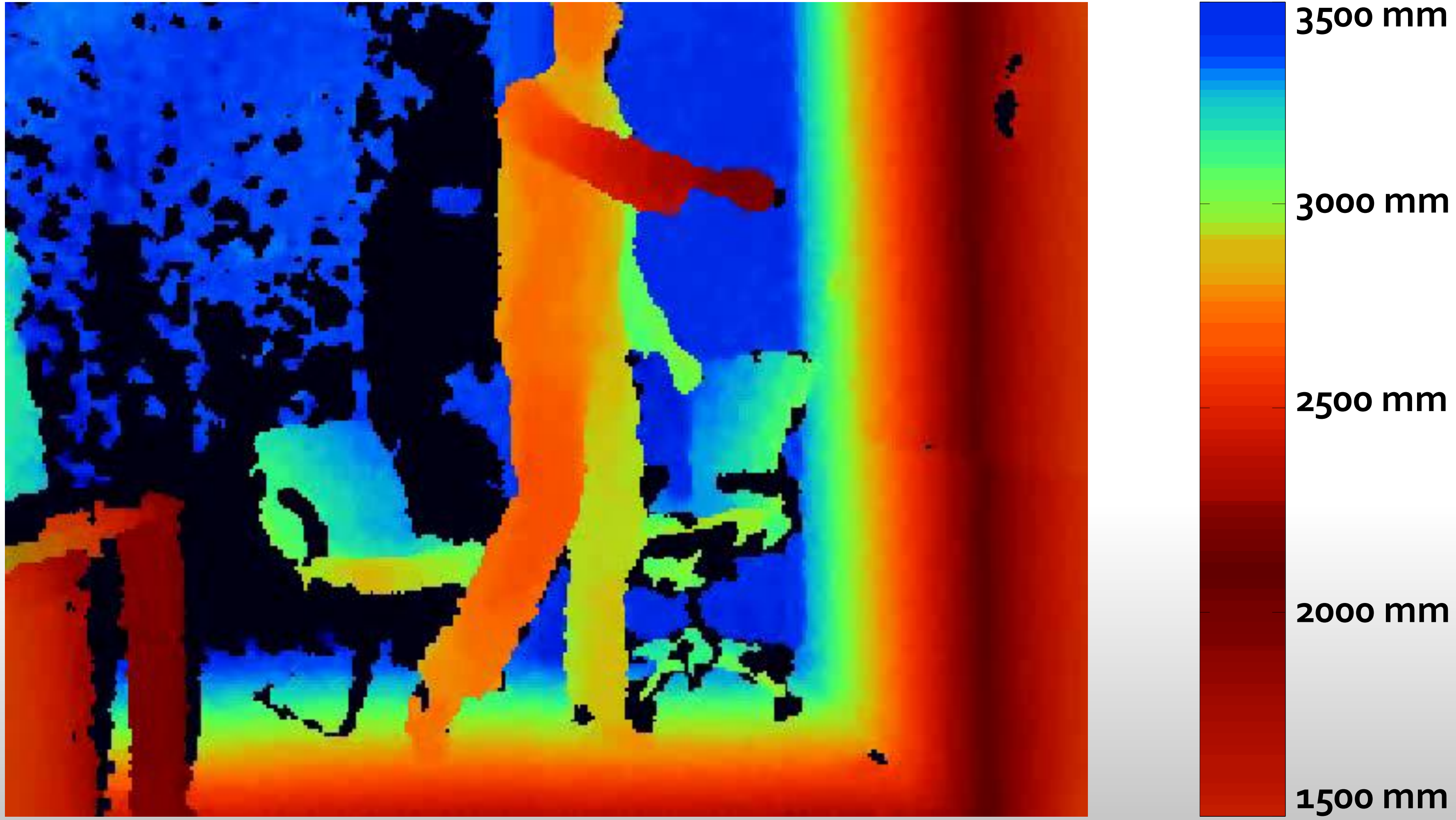
Microsoft

Okada & Stenger 2008

Navaratnam *et al.* 2007

STATE OF THE ART

Microsoft

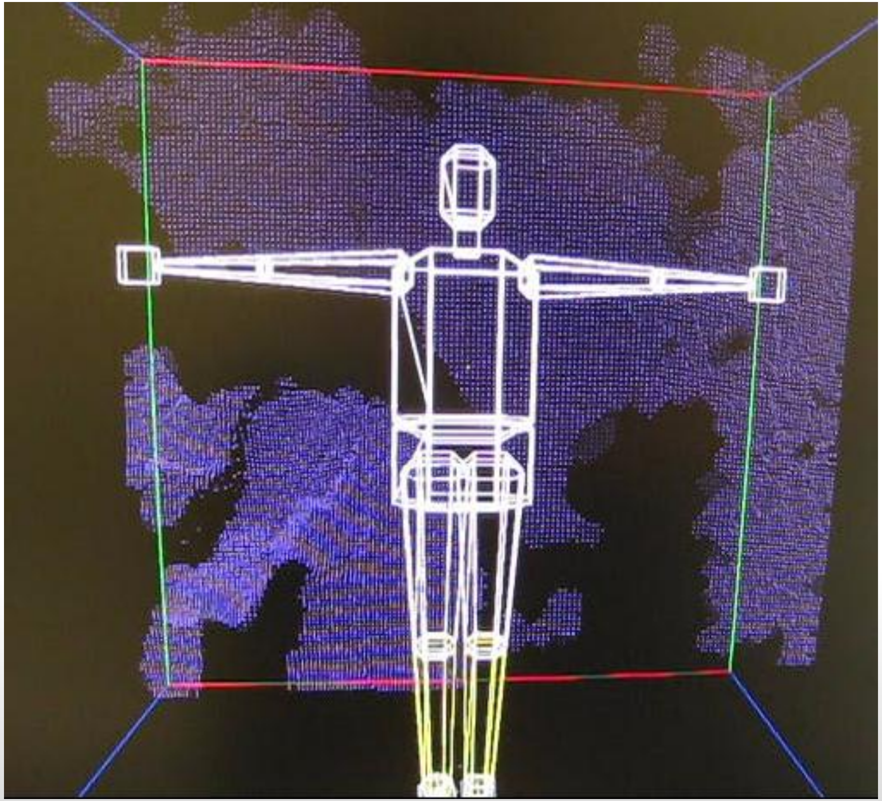

3500 mm
3000 mm
2500 mm
2000 mm
1500 mm

XBox prototype, Sept 2008

- Real time
- Accurate
- General poses

But…

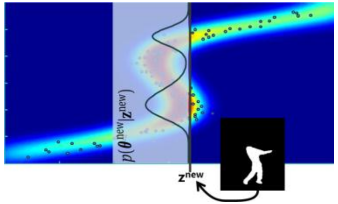
- **Needs initialization**
- Limited body types
- Limited agility



[Hogg 1982]

Generative/ Model-based

Discriminative/ Regression

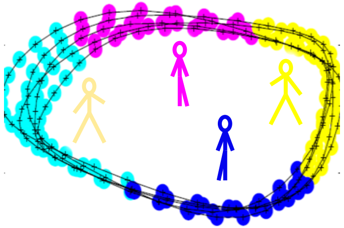


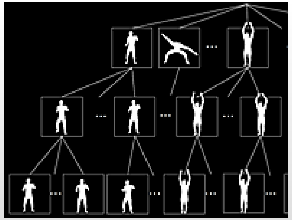
[Agarwal & Triggs 2004]
[Navaratnam & al 2007]

[Bourdev & Malik 09]

Detection

Tracking

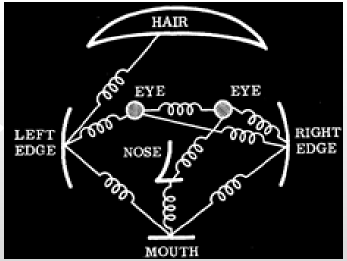




[Gavrila 2000]

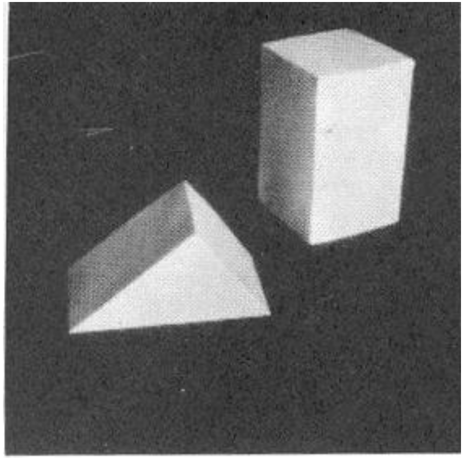
Whole

Parts

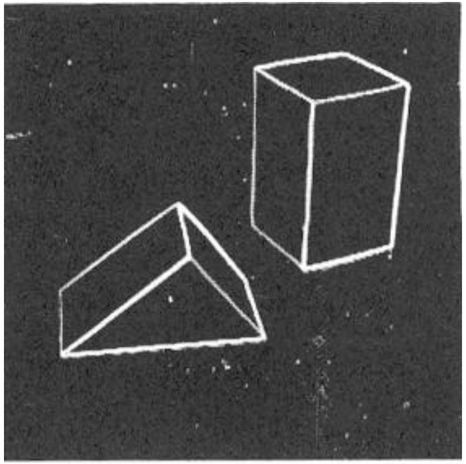


[Fischler & Elschlager 1973]

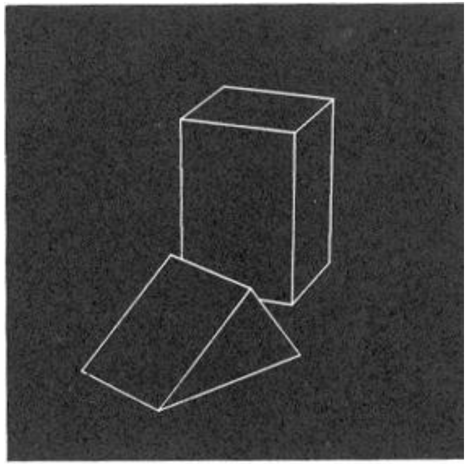
STATE OF THE ART



(a) Original picture.



(b) Differentiated picture.

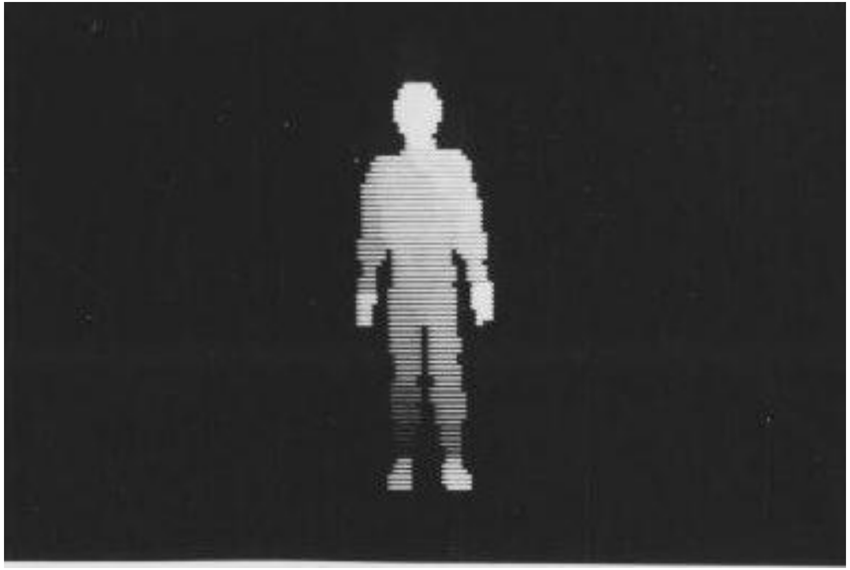


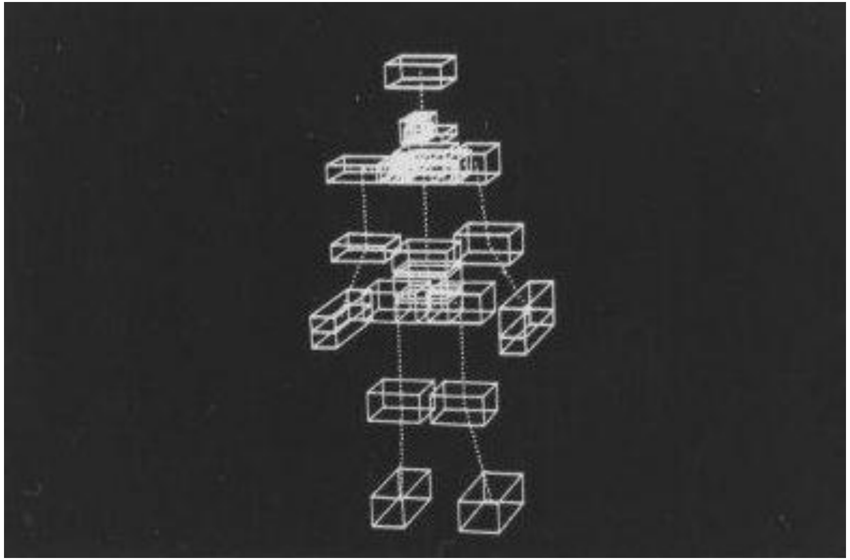
(d) Rotated view.

1965. L. G. Roberts, **Machine Perception of Three Dimensional Solids**, in *Optical and electro-optical information processing*, J. T. Tippett (ed.), MIT Press.

MODEL-BASED VISION

Microsoft





1980. J. O'Rourke and N. Badler. **Model-based image analysis of human motion using constraint propagation.** IEEE Trans. on Pattern Analysis and Machine Intelligence.

MODEL-BASED VISION

Model-based vision: a program to see a walking person

David Hogg

For a machine to be able to 'see', it must know something about the object it is 'looking' at. A common method in machine vision is to provide the machine with general rather than specific knowledge about the object. An alternative technique, and the one used in this paper, is a model-based approach in which particulars about the object are given and this drives the analysis. The computer program described here, the WALKER model, maps images into a description in which a person is represented by the series of hierarchical levels, i.e. a person has an arm which has a lower-arm which has a hand. The performance of the program is illustrated by superimposing the machine-generated picture over the original photographic images.

Keywords: vision, machine perception, WALKER model

INTRODUCTION

Vision systems, both natural and artificial, require knowledge about the perceived objects, although the role played by this knowledge in the analytical process is unclear. Many techniques of machine vision seek to generate 3D structural descriptions without invoking object specific knowledge. An alternative is to adopt the 'model-based' approach wherein particular knowledge about the objects being sought drives the analysis.

This paper is concerned with a computer program that understands TV image sequences depicting a person walking through an arbitrary environment (Figure 1). The program maps given image sequences into a description in which the human body is represented by a collection of connected cylinders corresponding to its parts. It is supposed that such a 3D structural description would be both necessary and sufficient for many everyday tasks to be performed effectively. For example, touching someone's arm or deciding whether several people are marching in step all appear to require a grasp of 3D structure whether perceived visually or otherwise. Each output description is an instance of an abstract 3D model for a class of human walkers, henceforth called the WALKER model, itself an input to the program (Figure 2).

Descriptions generated by the program are sufficiently detailed to determine a pictorial reconstruction of the person from the perspective of the original imaging device. By superimposing these reconstructions over the original images a clear indication of the program's performance is visible to the human observer. When presented with the sequence depicted in Figure 1, the program generates as part of its output the sequence shown in Figure 3. The program copes with the enormous local ambiguity in an image by weighing evidence from across the image in support of a large number of possible interpretations. As a consequence, the program's performance should degrade gracefully for increasingly difficult image sequences in which the walker may be obscured or occuluded to the camera.

Visual problem

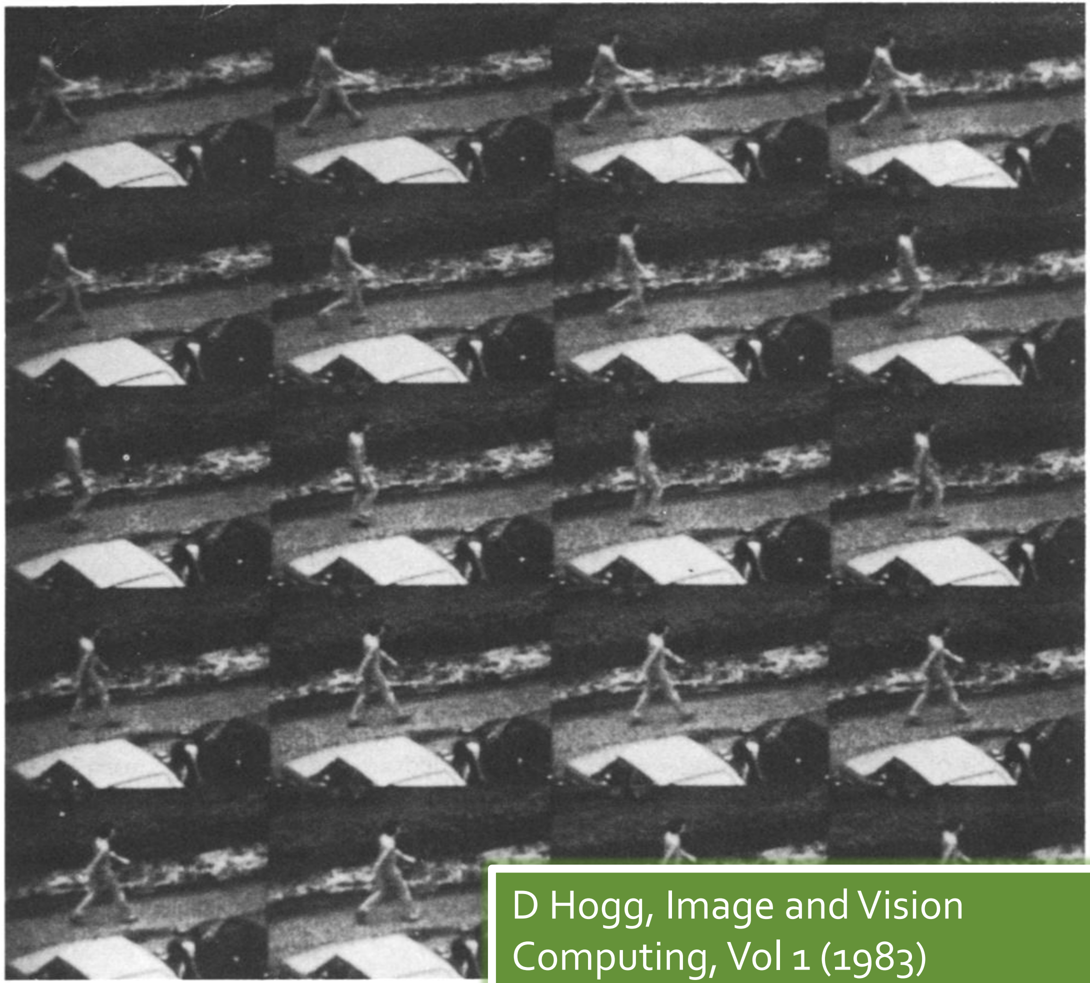
The visual problem can be divided broadly into two parts; namely, what should be described and how can such descriptions be derived from a time-varying 2D image. It is impossible to divorce these two issues from one another since the difficulty of deriving a description from an image is bound to depend on the things being described. Moreover, certain representations may be required solely as intermediate descriptions for the interpretative process itself.

The question of what should be represented must depend on the visual system's function within a cognitive machine whose ultimate goal may be far removed from the visual world[1]. This paper takes a noncontroversial stand in accpeting the usefulness of 3D structural descriptions as an interface to a larger system and instead concentrates on the second issue of how to generate such a description from an image.

General-knowledge inference

Much of the current work in computer vision is concerned with the generation of 3D descriptions using only general-

School of Engineering and Applied Sciences, University of Sussex, Brighton, Sussex, UK
The research reported in this paper was carried out while the author was an SERC funded research student in the Cognitive Studies Programme at the University of Sussex



D Hogg, Image and Vision Computing, Vol 1 (1983)

Model-based vision: a program to see a walking person

David Hogg

For a machine to be able to 'see', it must know something about the object it is 'looking' at. A common method in machine vision is to provide the machine with general rather than specific knowledge about the object. An alternative technique, and the one used in this paper, is a model-based approach in which particulars about the object are given and this drives the analysis. The computer program described here, the WALKER model, maps images into a description in which a person is represented by the series of hierarchical levels, i.e. a person has an arm which has a lower-arm which has a hand. The performance of the program is illustrated by superimposing the machine-generated picture over the original photographic images.

Keywords: vision, machine perception, WALKER model

INTRODUCTION

Vision systems, both natural and artificial, require knowledge about the perceived objects, although the role played by this knowledge in the analytical process is unclear. Many techniques of machine vision seek to generate 3D structural descriptions without invoking object specific knowledge. An alternative is to adopt the 'model-based' approach wherein particular knowledge about the objects being sought drives the analysis.

This paper is concerned with a computer program that understands TV image sequences depicting a person walking through an arbitrary environment (Figure 1). The program maps given image sequences into a description in which the human body is represented by a collection of connected cylinders corresponding to its parts. It is supposed that such a 3D structural description would be both necessary and sufficient for many everyday tasks to be performed effectively. For example, touching someone's arm or deciding whether several people are marching in step all appear to require a grasp of 3D structure whether perceived visually or otherwise. Each output description is an instance of an abstract 3D model for a class of human walkers, henceforth called the WALKER model, itself an input to the program (Figure 2).

Descriptions generated by the program are sufficiently detailed to determine a pictorial reconstruction of the person from the perspective of the original imaging device. By superimposing these reconstructions over the original images a clear indication of the program's performance is visible to the human observer. When presented with the sequence depicted in Figure 1, the program generates as part of its output the sequence shown in Figure 3. The program copes with the enormous local ambiguity in an image by weighing evidence from across the image in support of a large number of possible interpretations. As a consequence, the program's performance should degrade gracefully for increasingly difficult image sequences in which the walker may be obscured or occuluded to the camera.

Visual problem

The visual problem can be divided broadly into two parts; namely, what should be described and how can such descriptions be derived from a time-varying 2D image. It is impossible to divorce these two issues from one another since the difficulty of deriving a description from an image is bound to depend on the things being described. Moreover, certain representations may be required solely as intermediate descriptions for the interpretative process itself.

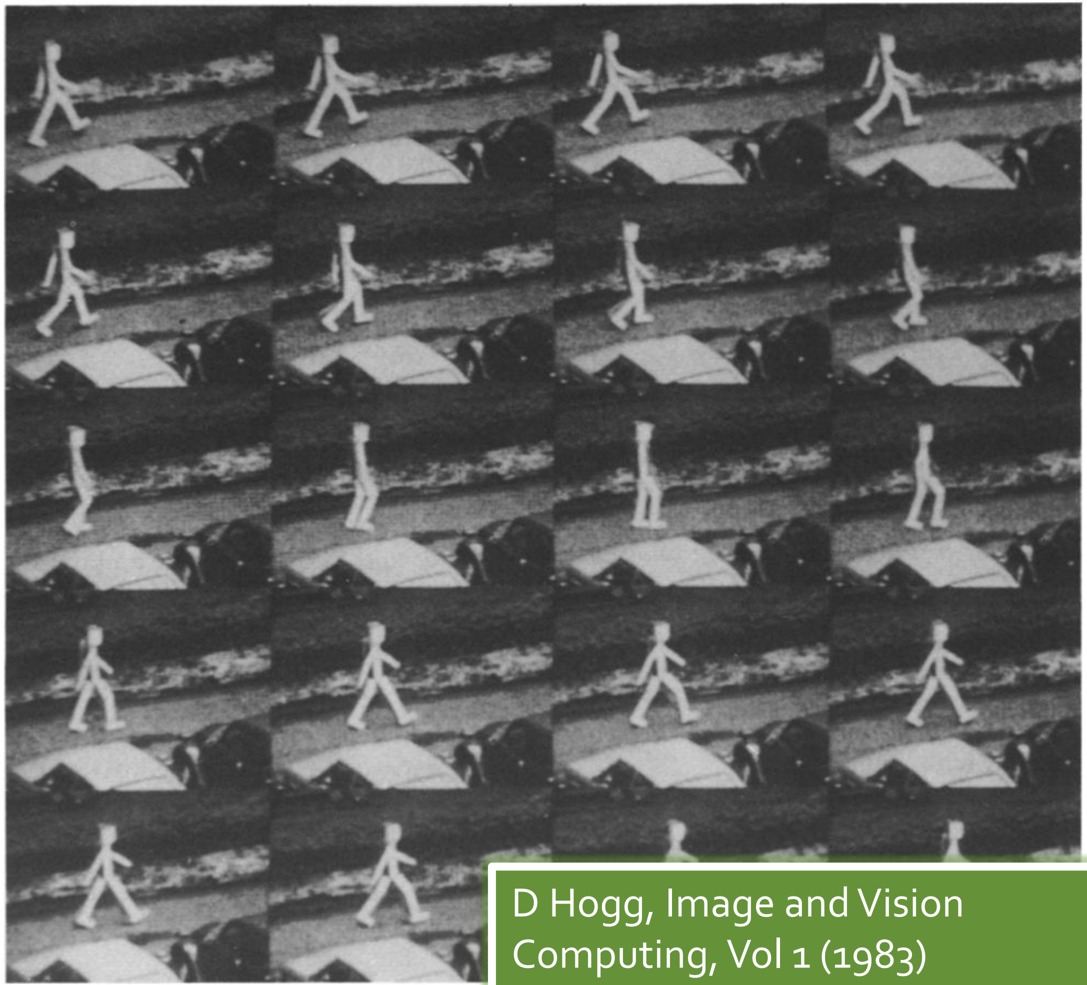
The question of what should be represented must depend on the visual system's function within a cognitive machine whose ultimate goal may be far removed from the visual world[1]. This paper takes a noncontroversial stand in accpeting the usefulness of 3D structural descriptions as an interface to a larger system and instead concentrates on the second issue of how to generate such a description from an image.

General-knowledge inference

Much of the current work in computer vision is concerned with the generation of 3D descriptions using only general-

School of Engineering and Applied Sciences, University of Sussex, Brighton, Sussex, UK

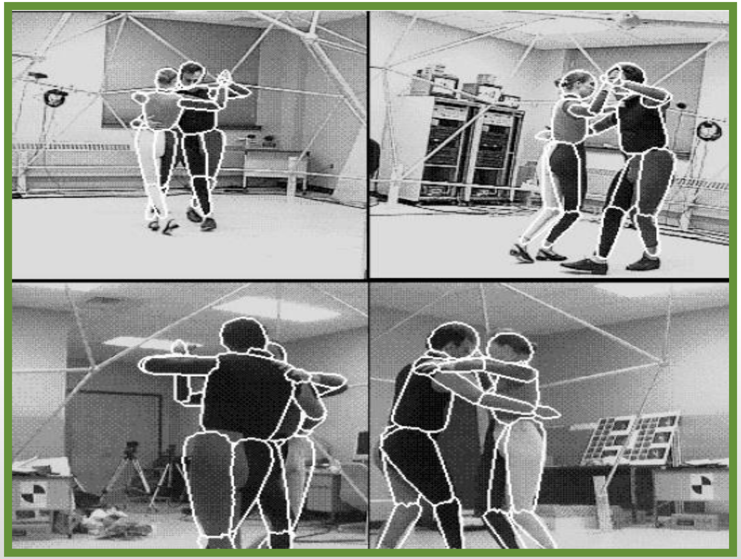
The research reported in this paper was carried out while the author was an SERC funded research student in the Cognitive Studies Programme at the University of Sussex



D Hogg, Image and Vision Computing, Vol 1 (1983) in Proc. IEEE Computer Vision and Pattern Recognition, San Francisco, 1996

3-D model-based tracking of humans in action: a multi-view approach

D.M. Gavrila and L.S. Davis
Computer Vision Laboratory, CfAR,
University of Maryland



MODEL-BASED VISION

Microsoft

- The “model” is *Forward/Generative/Graphical*
- Requiring *search* in many dimensions
  - say 10^{13} for the body

Resolved using

- (a) clever search: gradient descent and better
- (b) *temporal coherence*
  - Assume we were right in the *previous* frame
  - And search only “nearby” configurations in this

Exponential likelihood of failure

Assume 0.1% failure rate per frame

- After n frames, chance of success = 0.999^n
- At 30 frames per second, that's:
  - 3.0% chance of failure after 1 second
  - 83.5% chance of failure after 1 minute
  - 99.99% chance of failure after 5 minutes

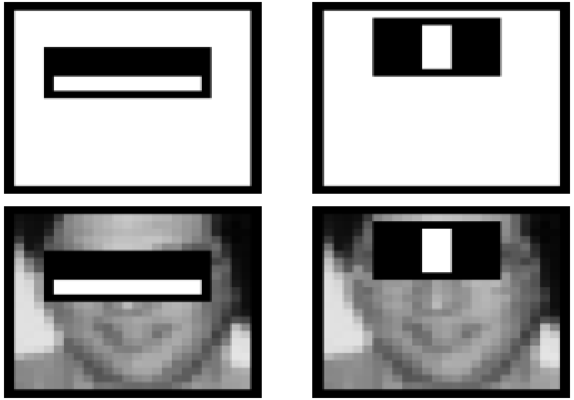
Exponential likelihood of failure

Assume 0.01% failure rate per frame

- After n frames, chance of success = 0.9999^n
- At 30 frames per second, that's:
  - 0.3% chance of failure after 1 second
  - 16.5% chance of failure after 1 minute
  - 59.3% chance of failure after 5 minutes

- Need a method which works on a single frame
  - Single-frame methods all based on machine learning
  - So we'll need training data
  - Lots of training data
  - And will need to represent multiple hypotheses

SO WE CAN'T USE TEMPORAL COHERENCE.

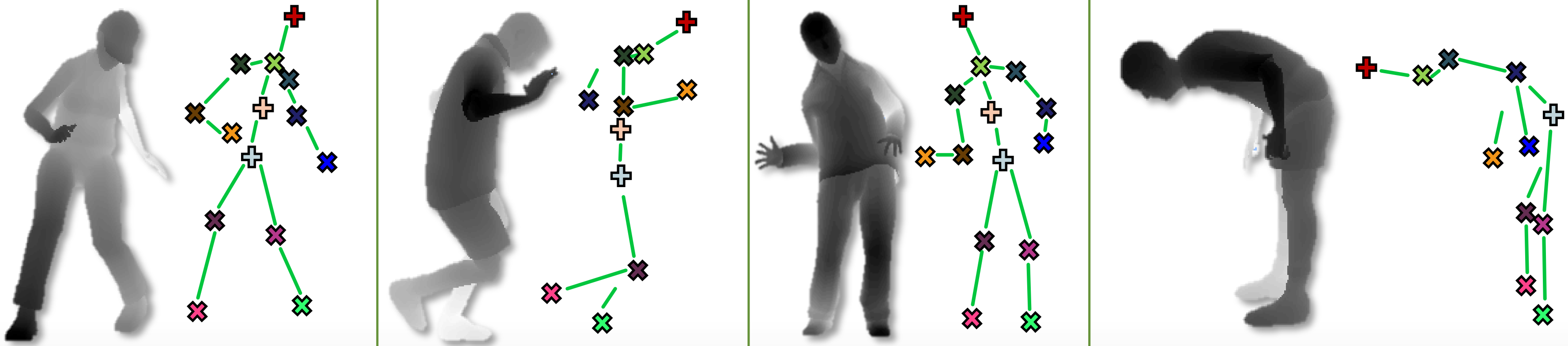


Paul A. Viola, Michael J. Jones
Robust Real-Time Face Detection
IEEE International Conference on Computer Vision, 2001

LEARNING A FACE DETECTOR

Microsoft

Step Zero: Training data



$$(\mathbf{z}^1, \boldsymbol{\theta}^1) \quad (\mathbf{z}^2, \boldsymbol{\theta}^2) \quad \cdots \quad (\mathbf{z}^i, \boldsymbol{\theta}^i) \quad \cdots \quad (\mathbf{z}^N, \boldsymbol{\theta}^N)$$

- Real home visits

- **Pose:** Motion capture
  - Standard “CMU” database
  - Custom database

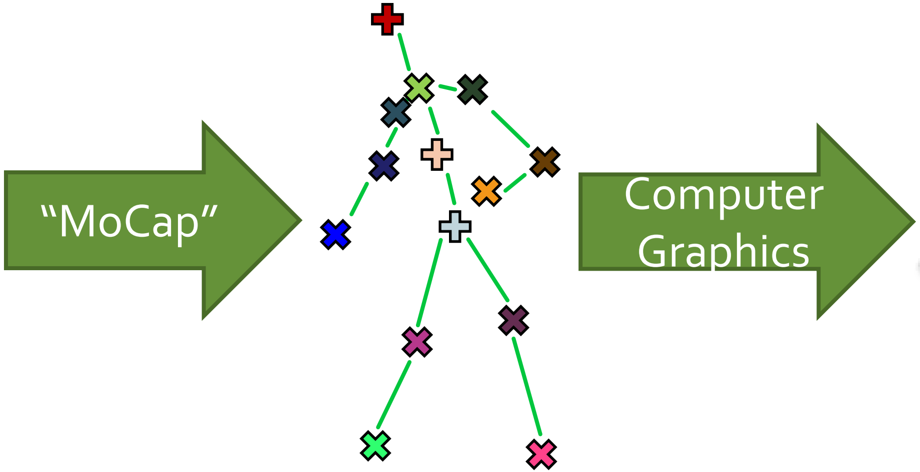
- **Body size & shape:** Retargeting
  - Effects/Games industry tool: MotionBuilder

SOURCES OF VARIED DATA

Microsoft

Actor wearing spherical markers

Observed by multiple cameras



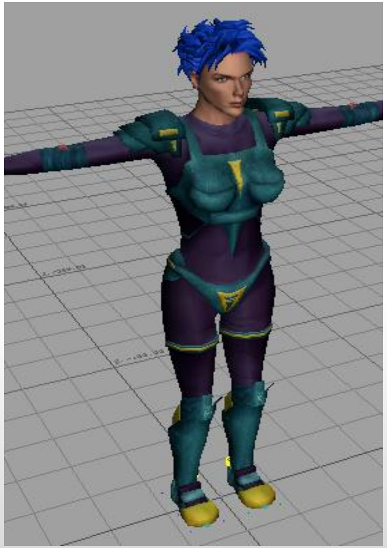
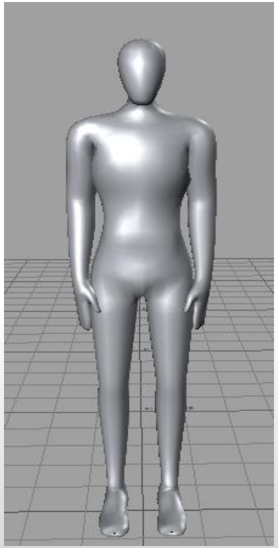
3D joint positions

Synthetic Depth Image

MOTION CAPTURE

- Standard motion capture datasets on the web
- Feed to *MotionBuilder* to generate 3D images
- Limited range of body types



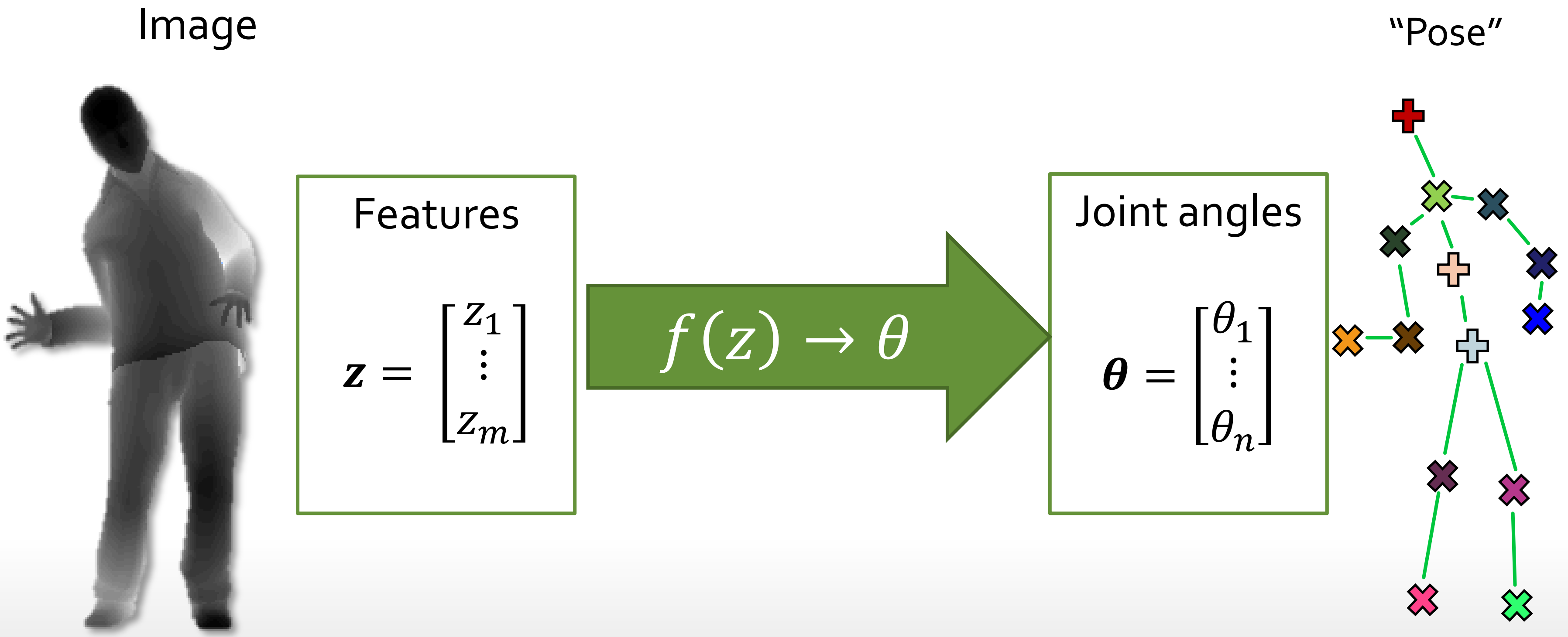
INITIAL EXPERIMENTS

Synthetic data:
realistic, but too clean

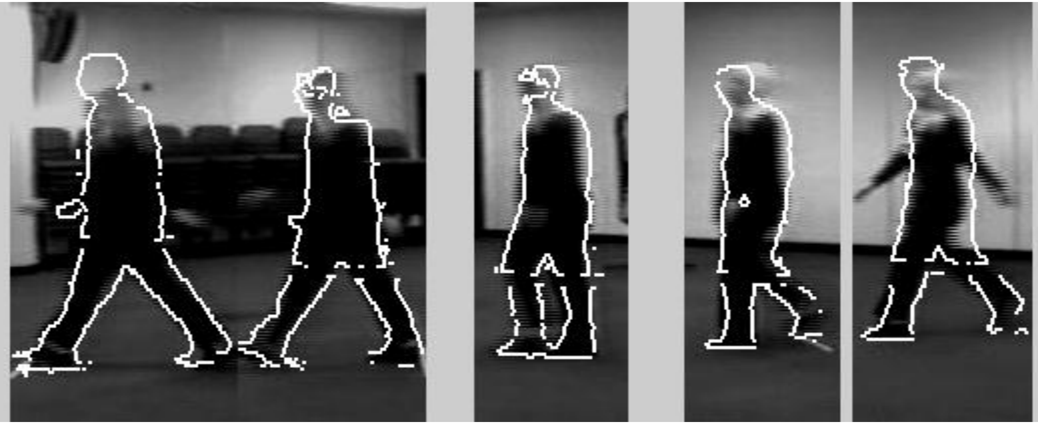
Artificially corrupted data

- depth resolution & noise
- rough edges
- missing pixels: hair/beards
- cropping & occlusions

SIMULATING CAMERA ARTEFACTS

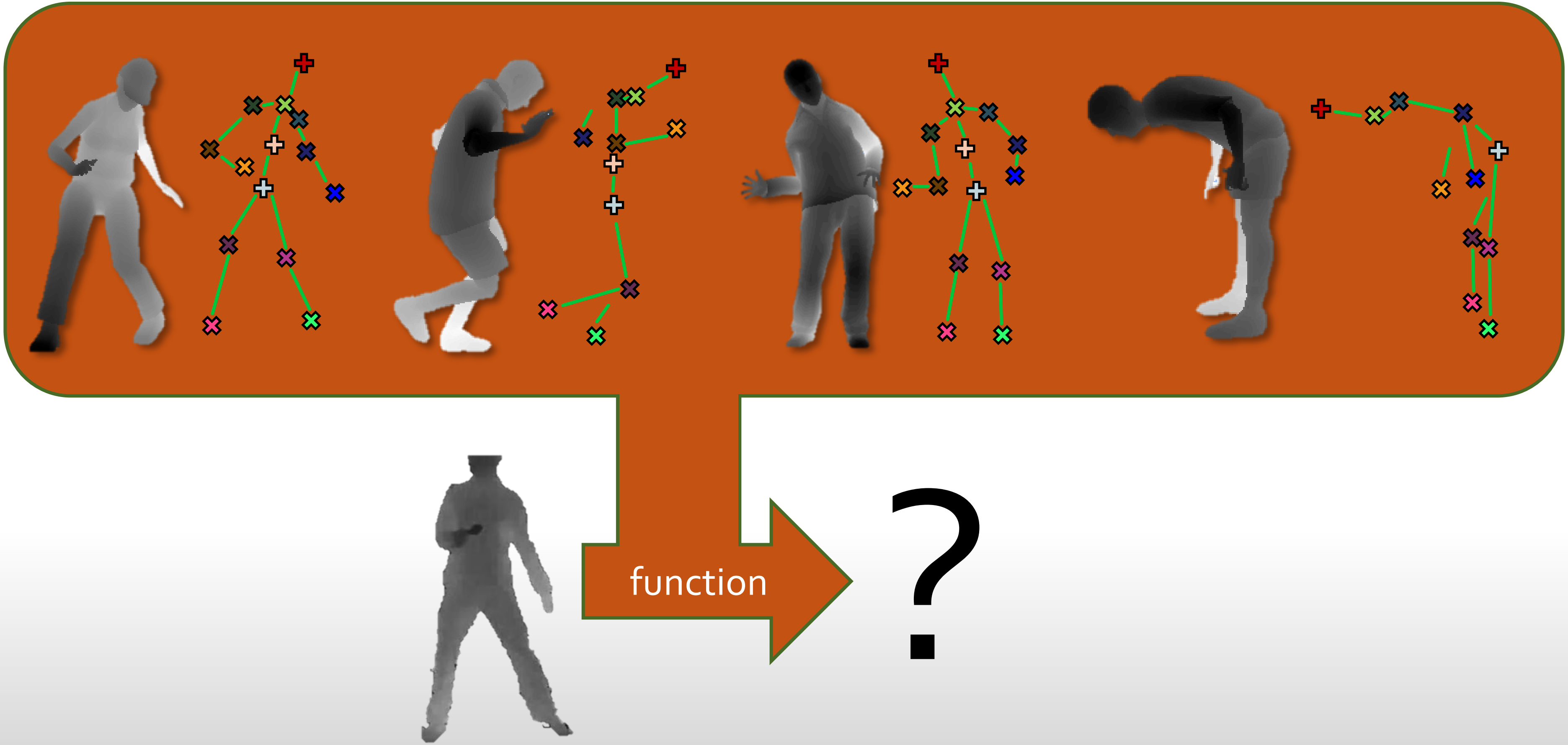


TRACKING BY REGRESSION



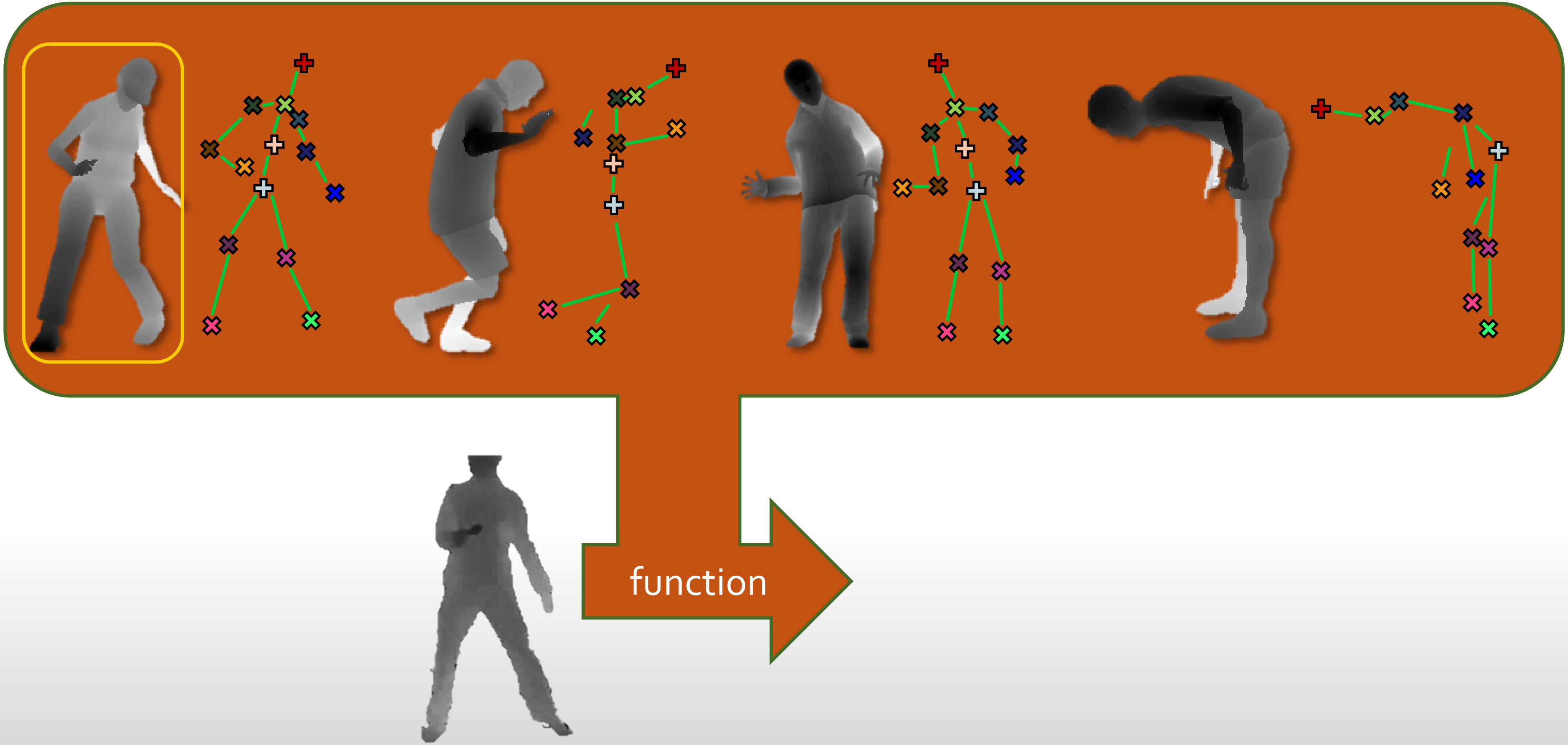
Andrew Blake, Kentaro Toyama,
Probablistic tracking in a metric space
IEEE International Conference on Computer Vision, 2001

DETECTION VS. TRACKING

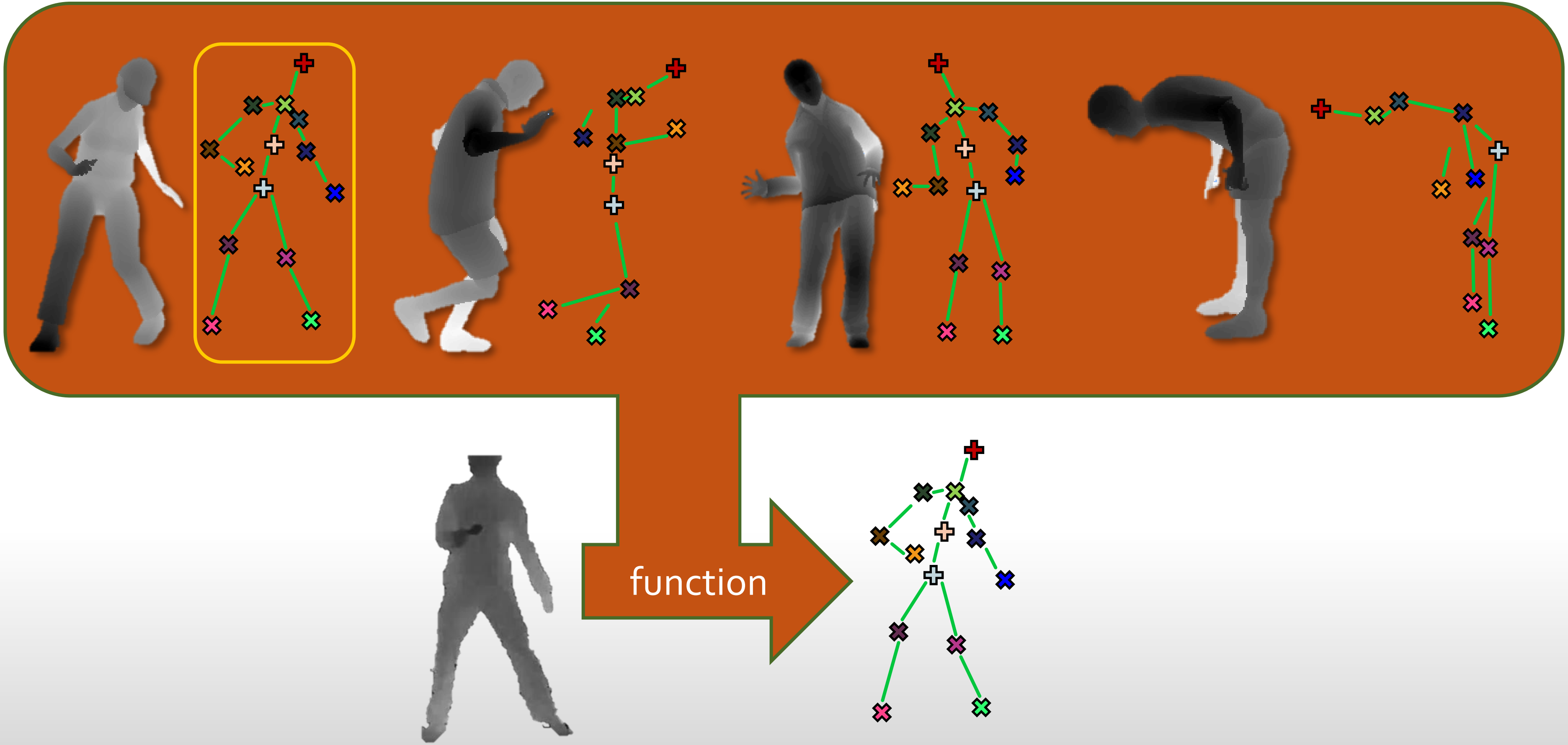


"LEARN" FUNCTION FROM DATA

Microsoft

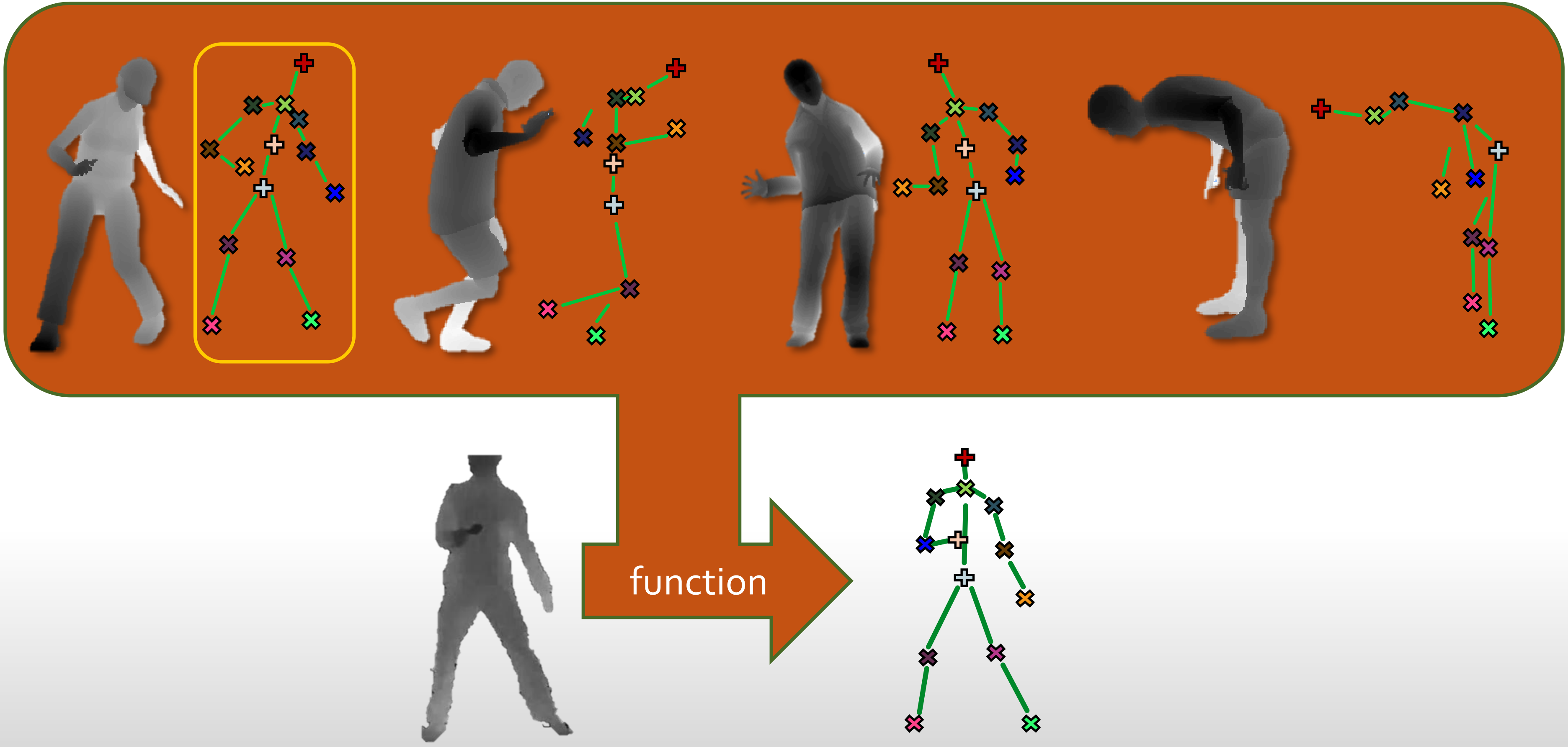


Nearest Neighbour

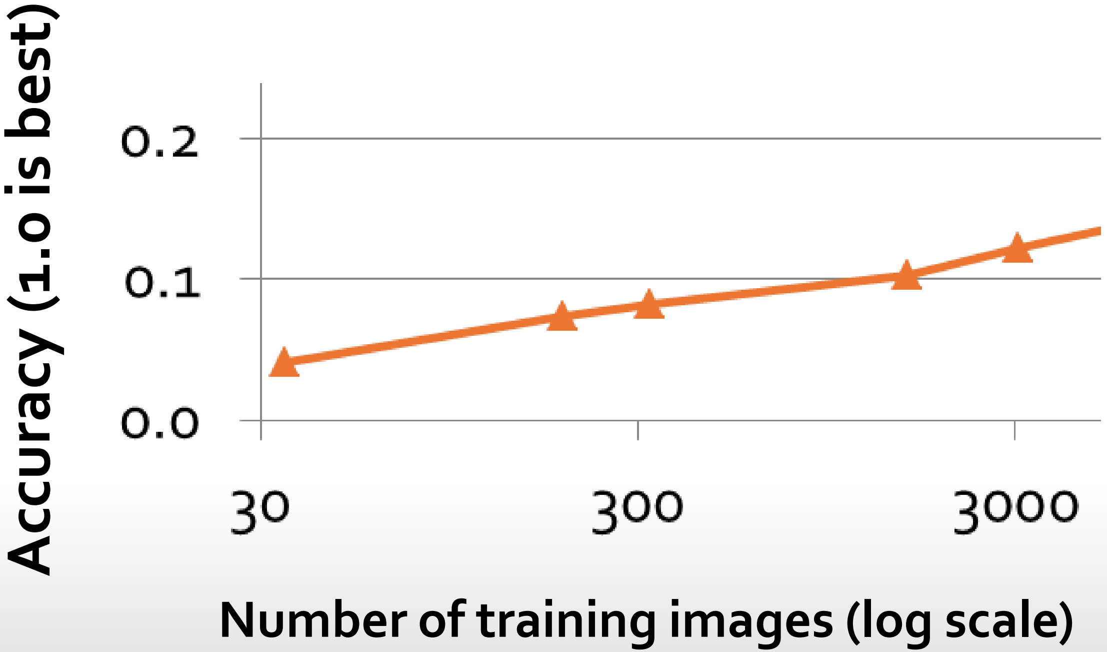


Nearest Neighbour

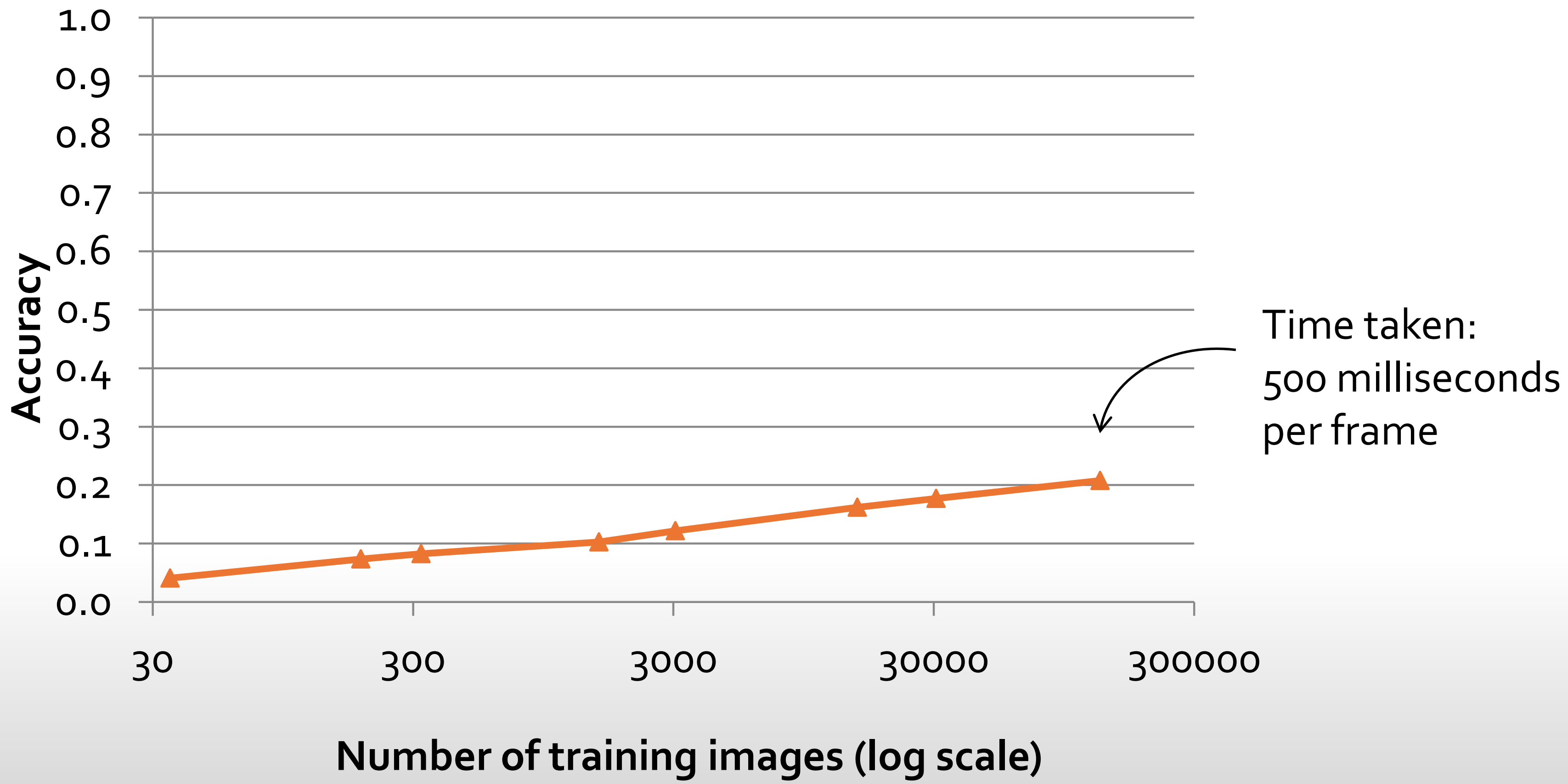
Microsoft



Nearest Neighbour



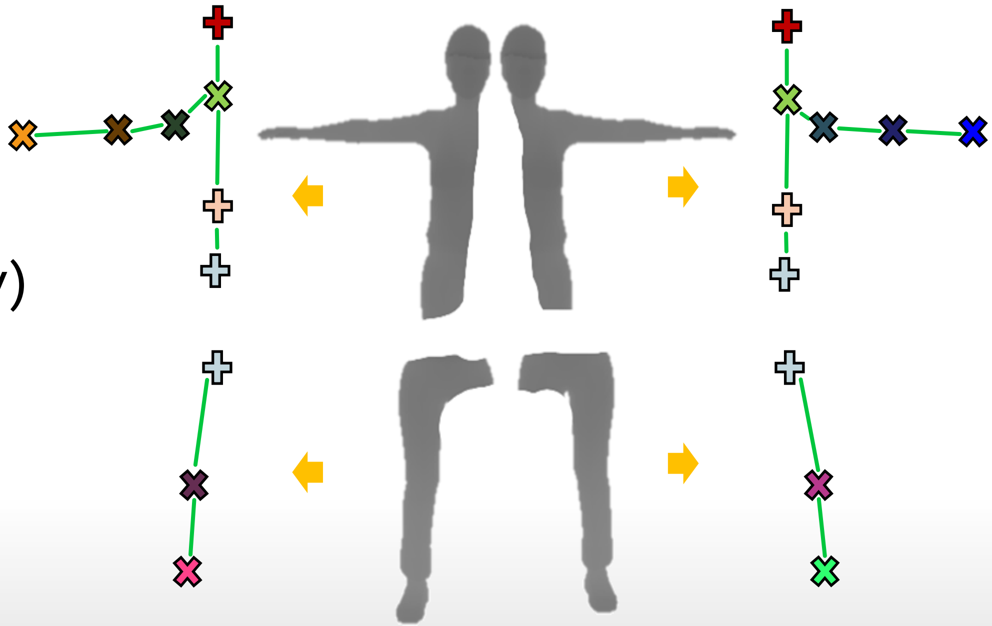
ALWAYS TRY NEAREST NEIGHBOUR FIRST



NEAREST NEIGHBOUR DIDN'T SCALE

Microsoft

- Whole body 10^{12} poses (say)
- Four parts $4 \times 10^{3+\epsilon}$ poses
- But ambiguity increases

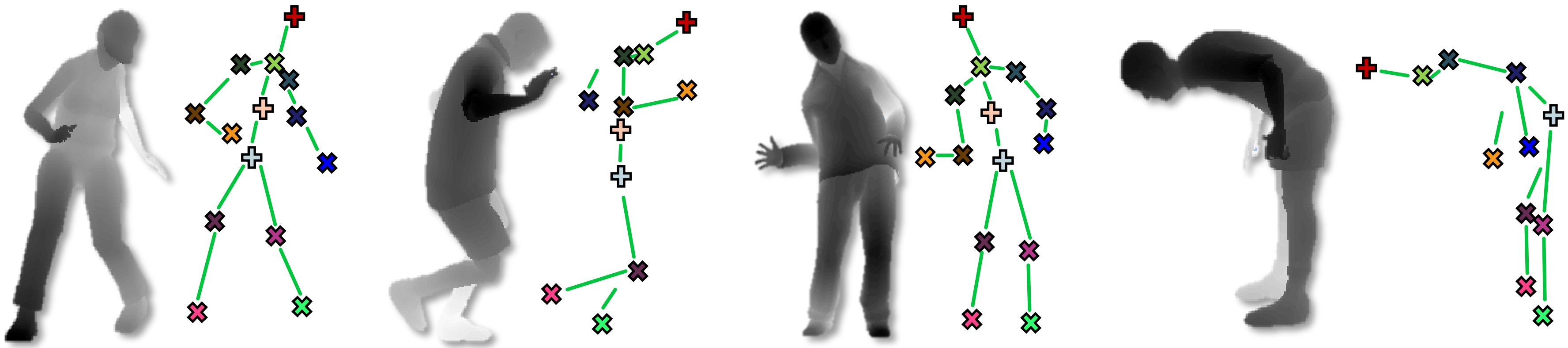


BODY PARTS

Microsoft

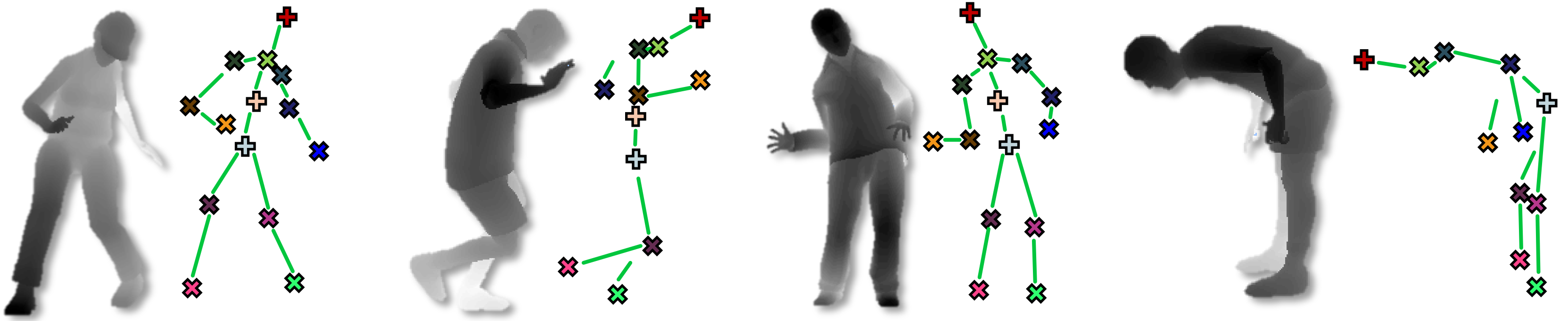
FOUR PARTS GOOD, 32 PARTS BETTER

Microsoft



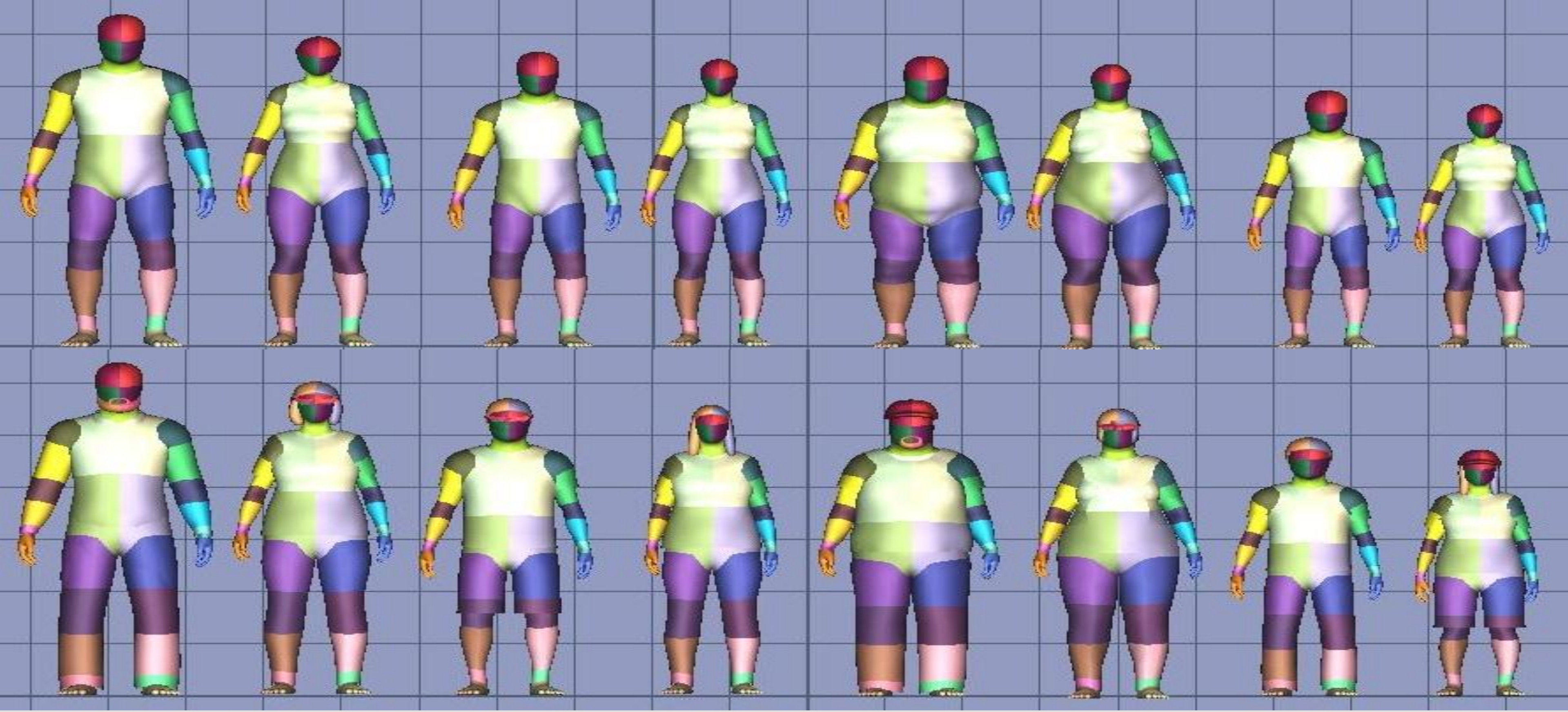
TRAINING DATA

Old (holistic) approach

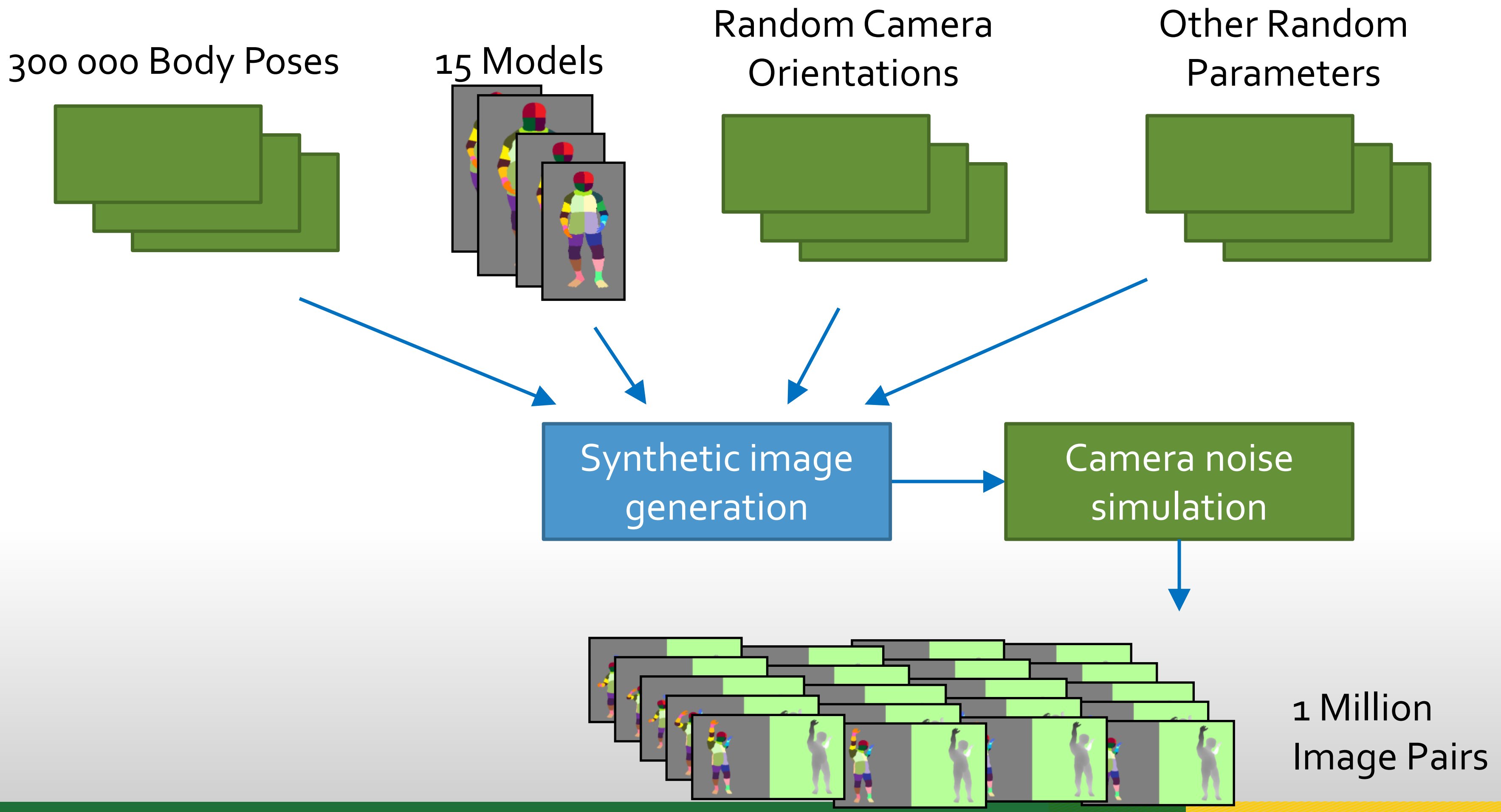


New (parts) approach

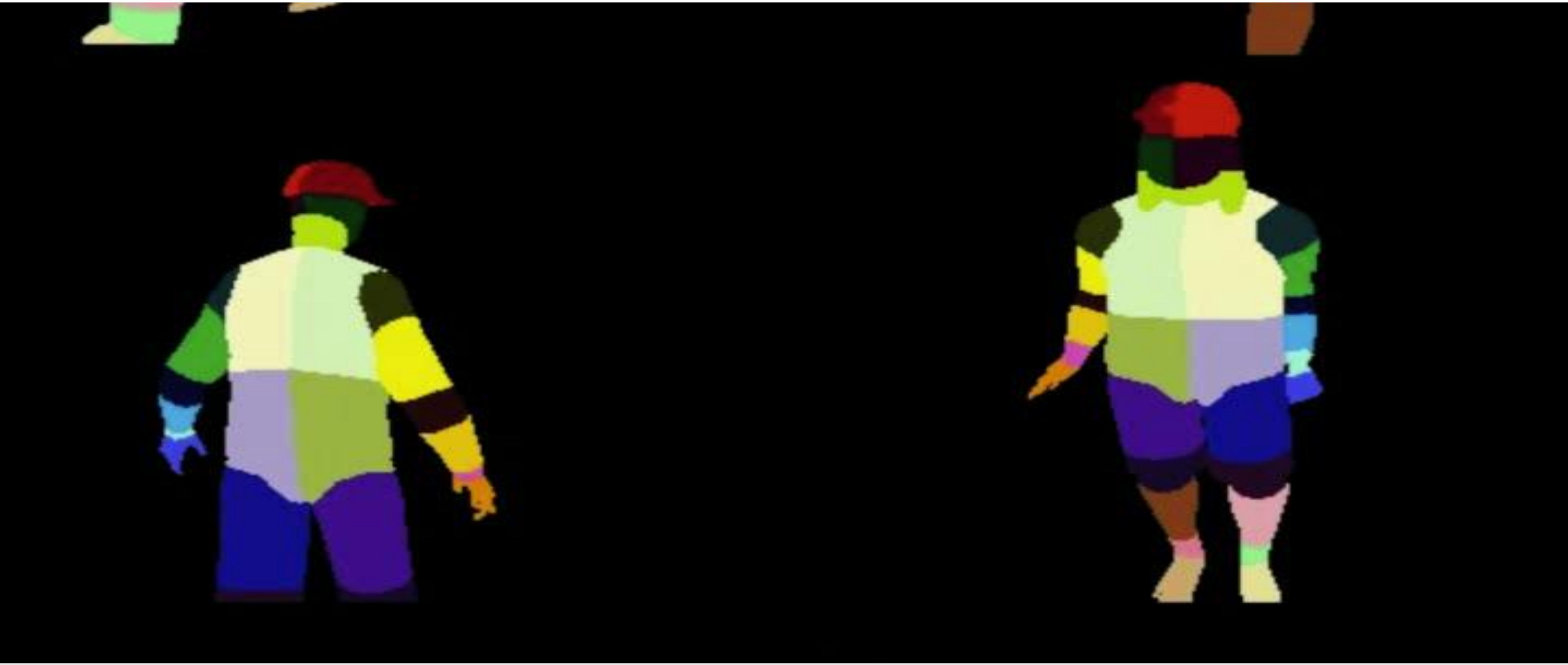
VIRTUAL HARLEQUIN SUIT



EXPANDING THE REPERTOIRE


300 000 Body Poses
15 Models
Random Camera Orientations
Other Random Parameters
Synthetic image generation
Camera noise simulation
1 Million Image Pairs

Microsoft

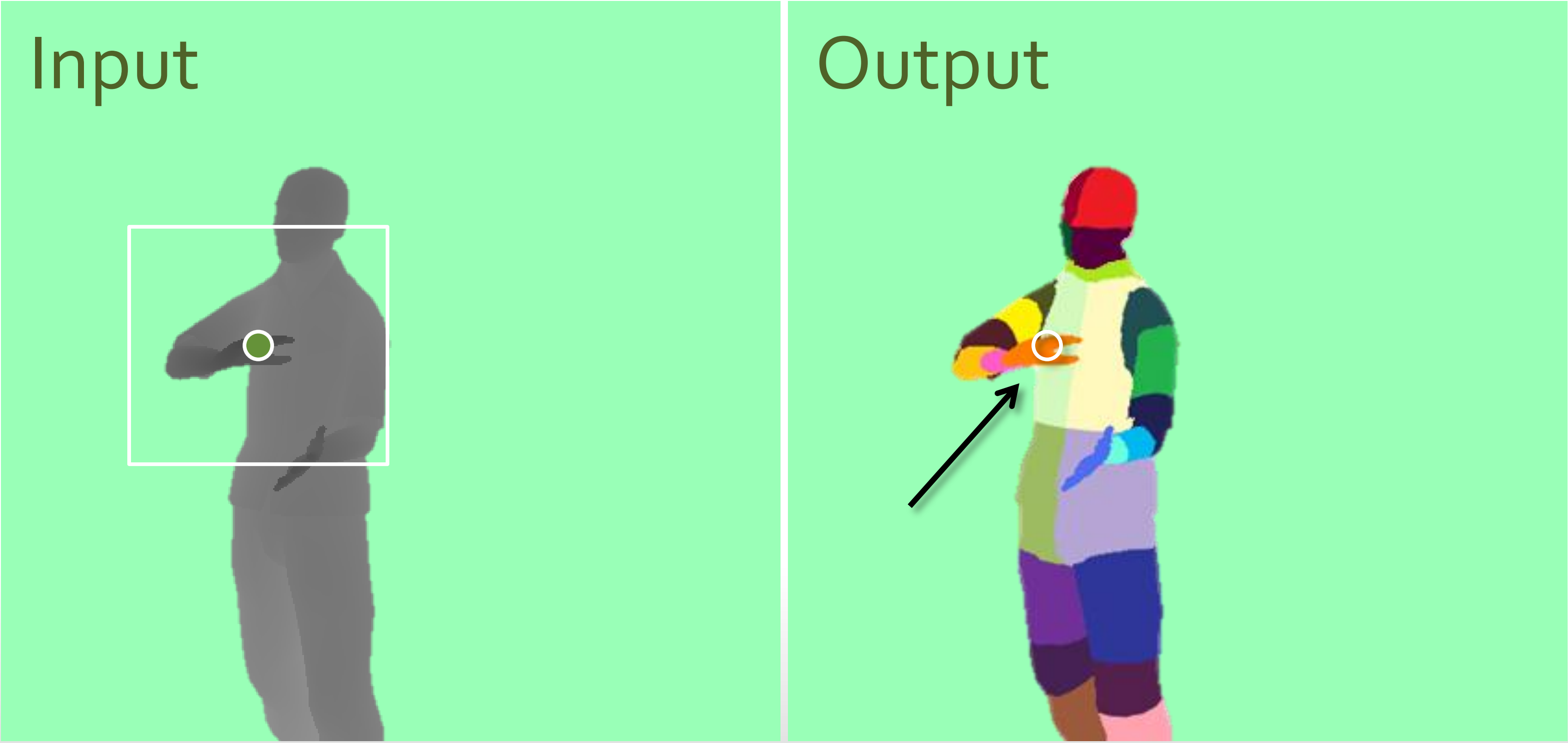


synthetic (held-out mocap poses)

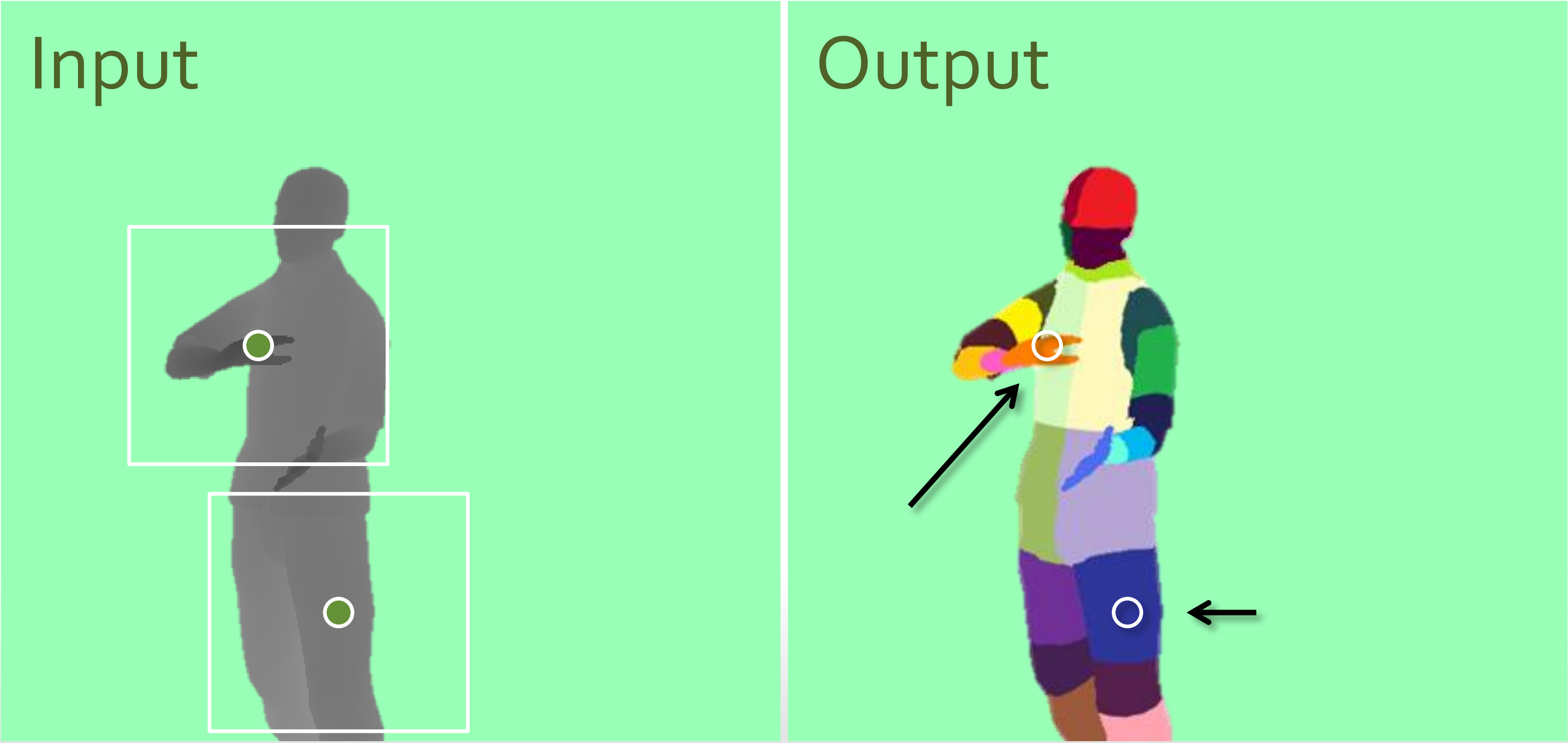
real (from home visits)

TEST DATA

EXAMPLE INPUTS & OUTPUTS

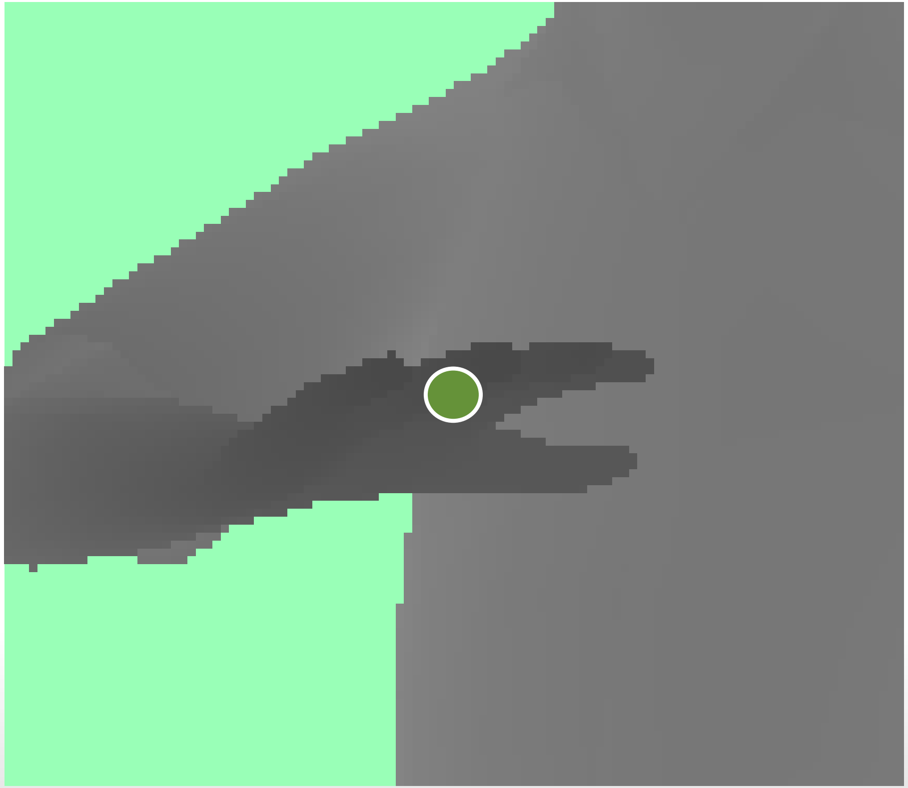


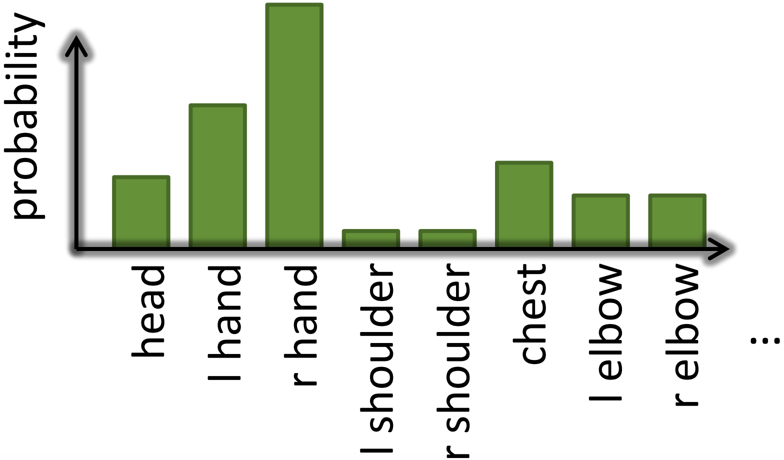
SLIDING WINDOW CLASSIFIER



SLIDING WINDOW CLASSIFIER

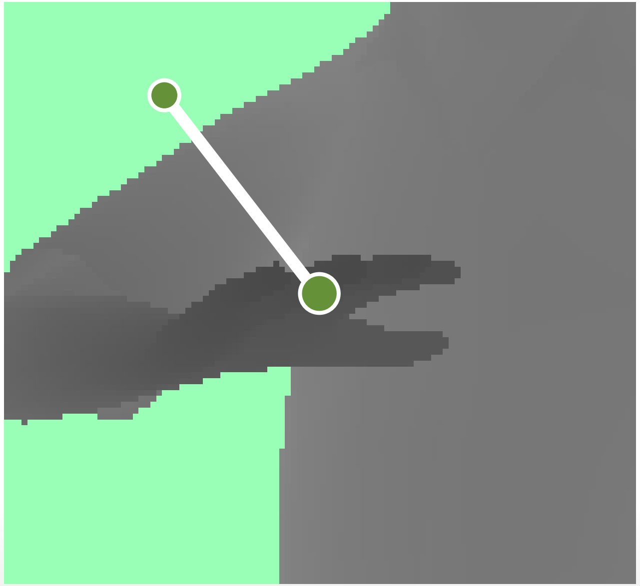
Microsoft

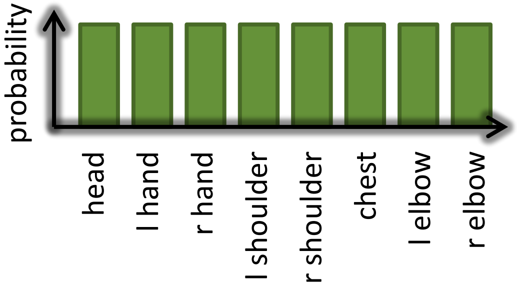




- Learn Prob(*body part* | *window*) from training data

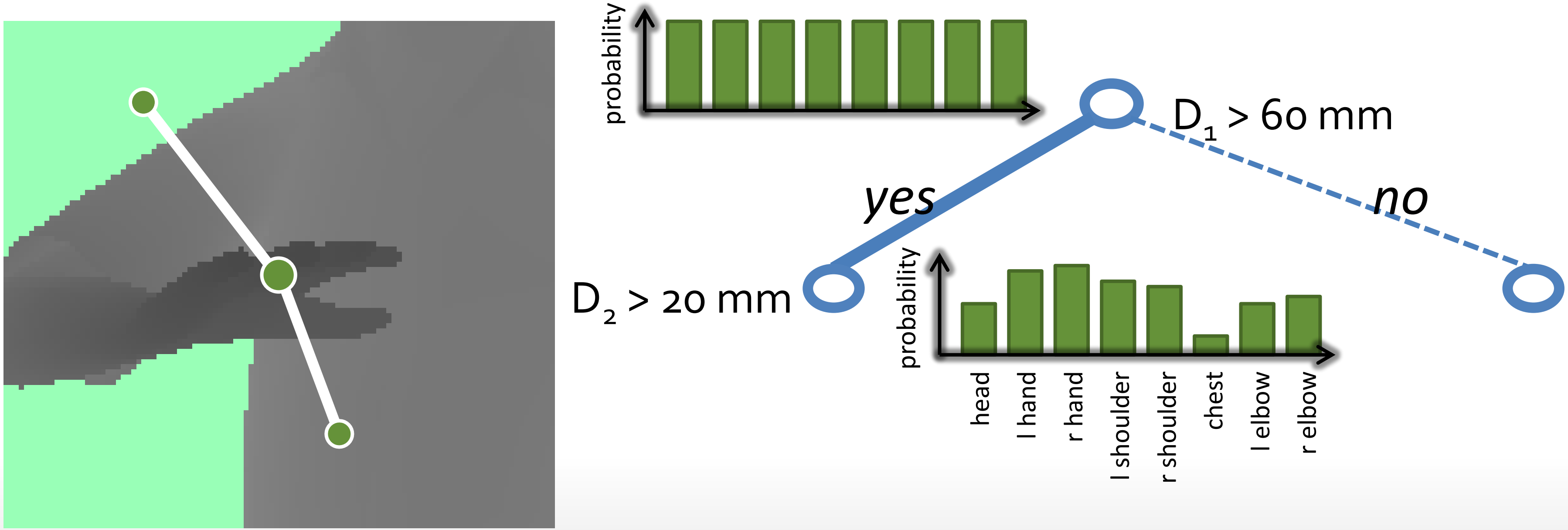
Focus on a single pixel: What part am I?



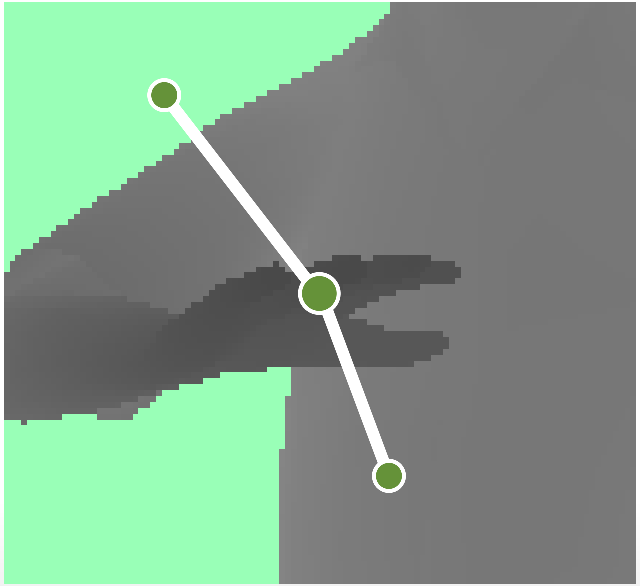


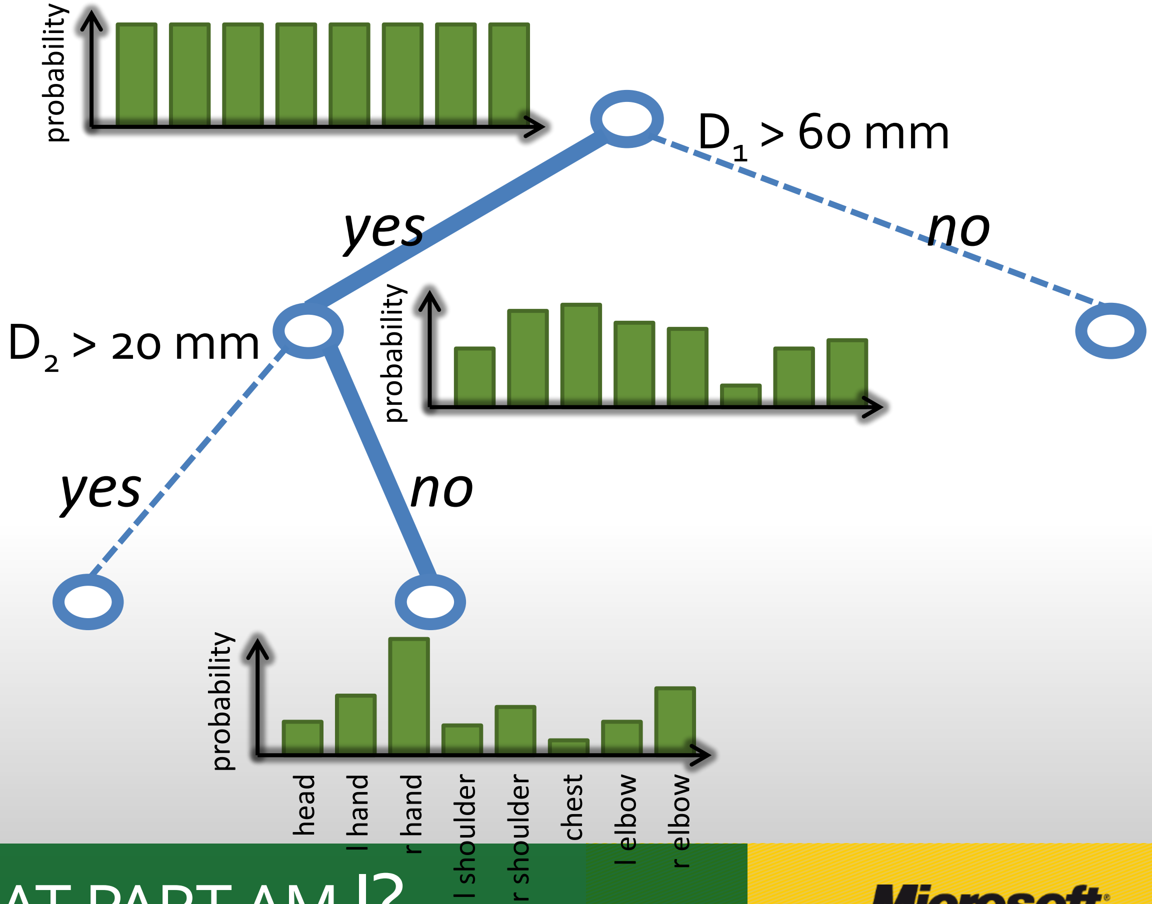
$D_1 > 60$ mm

Example pixel 1: What part am I?

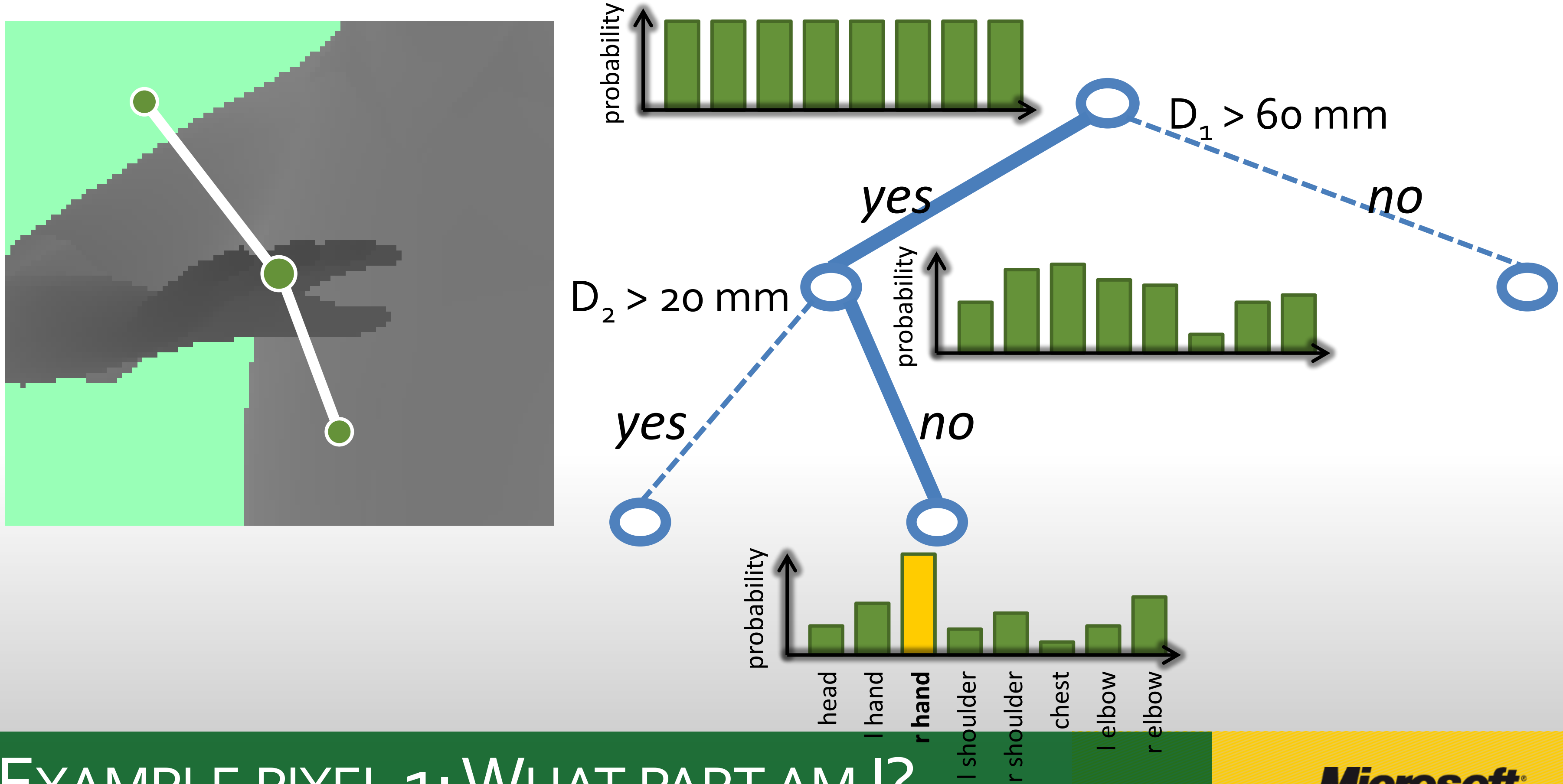


EXAMPLE PIXEL 1: WHAT PART AM I?

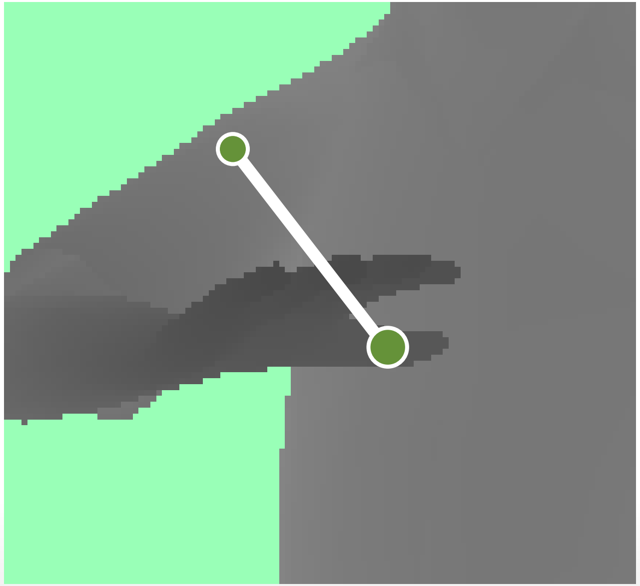


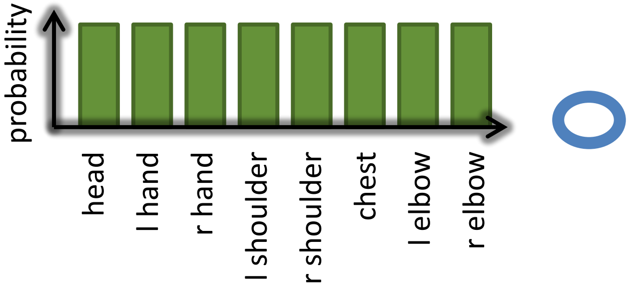


Example pixel 1: What part am I?



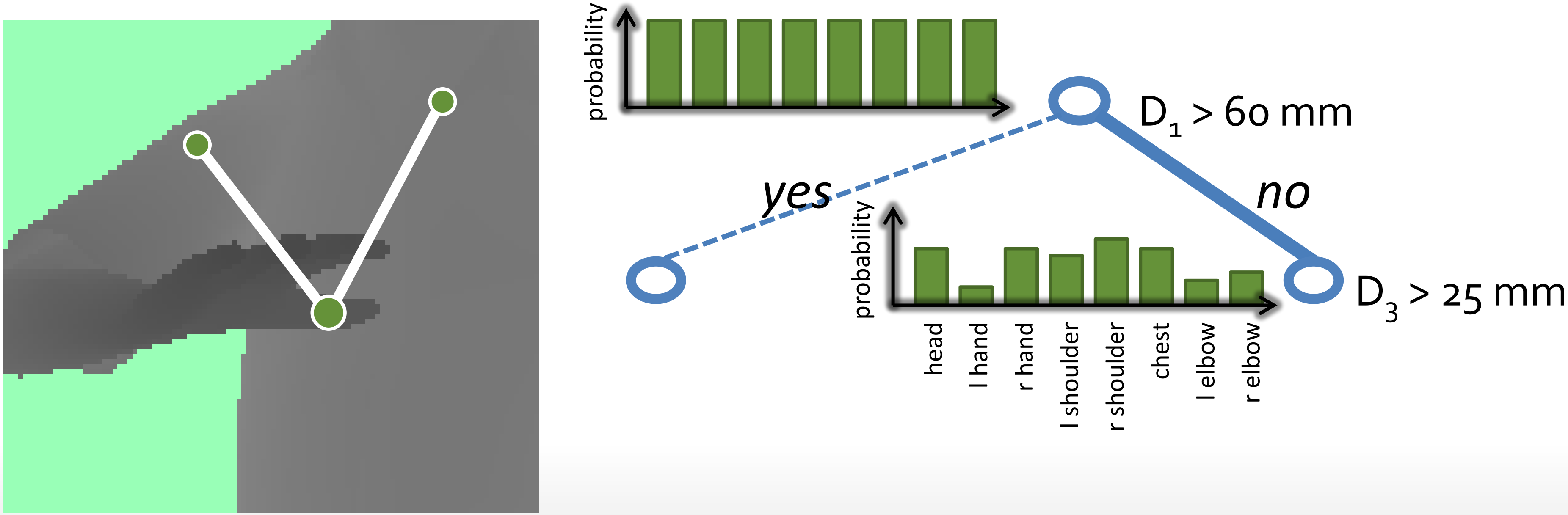
Example pixel 1: What part am I?





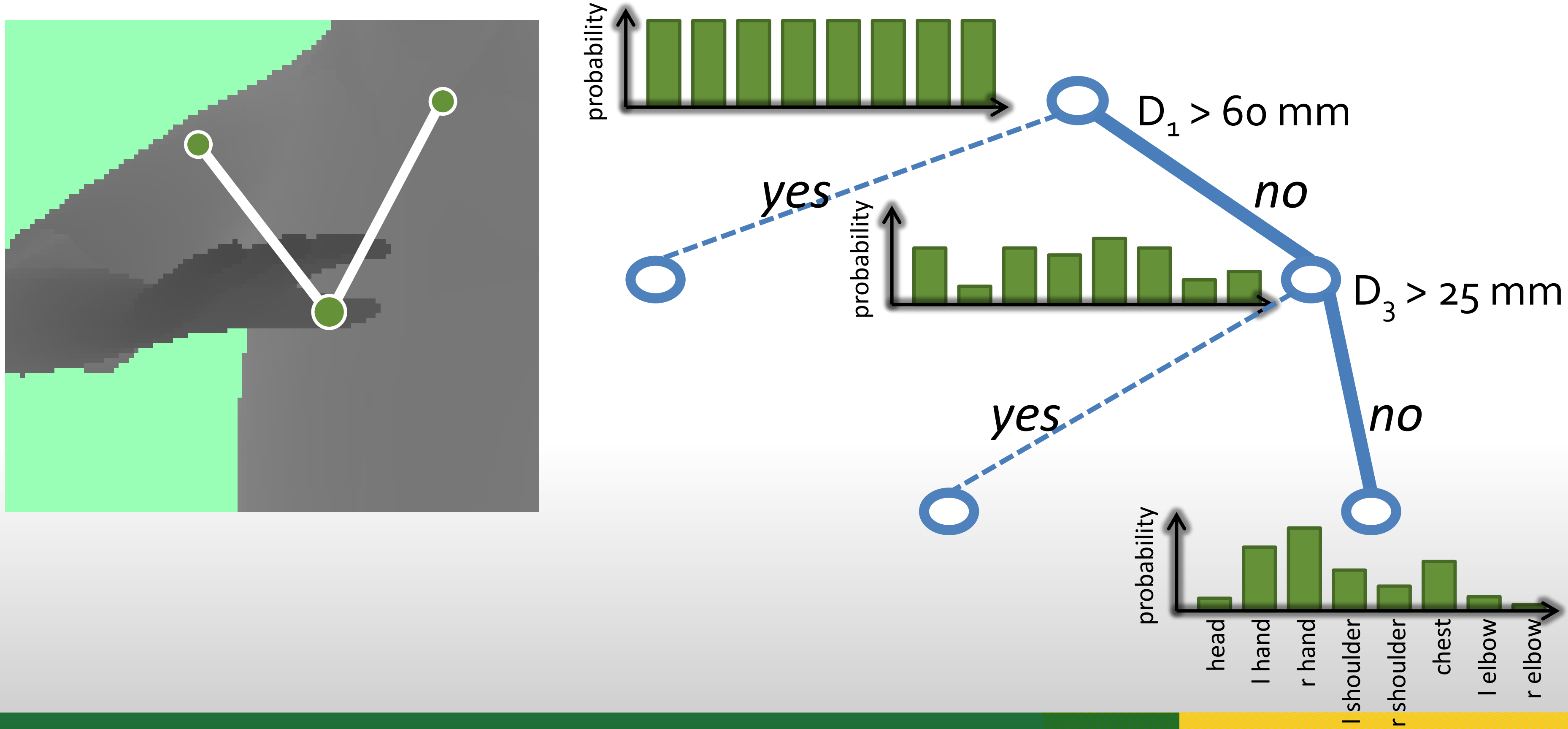
$D_1 > 60$ mm

Example pixel 2: What part am I?



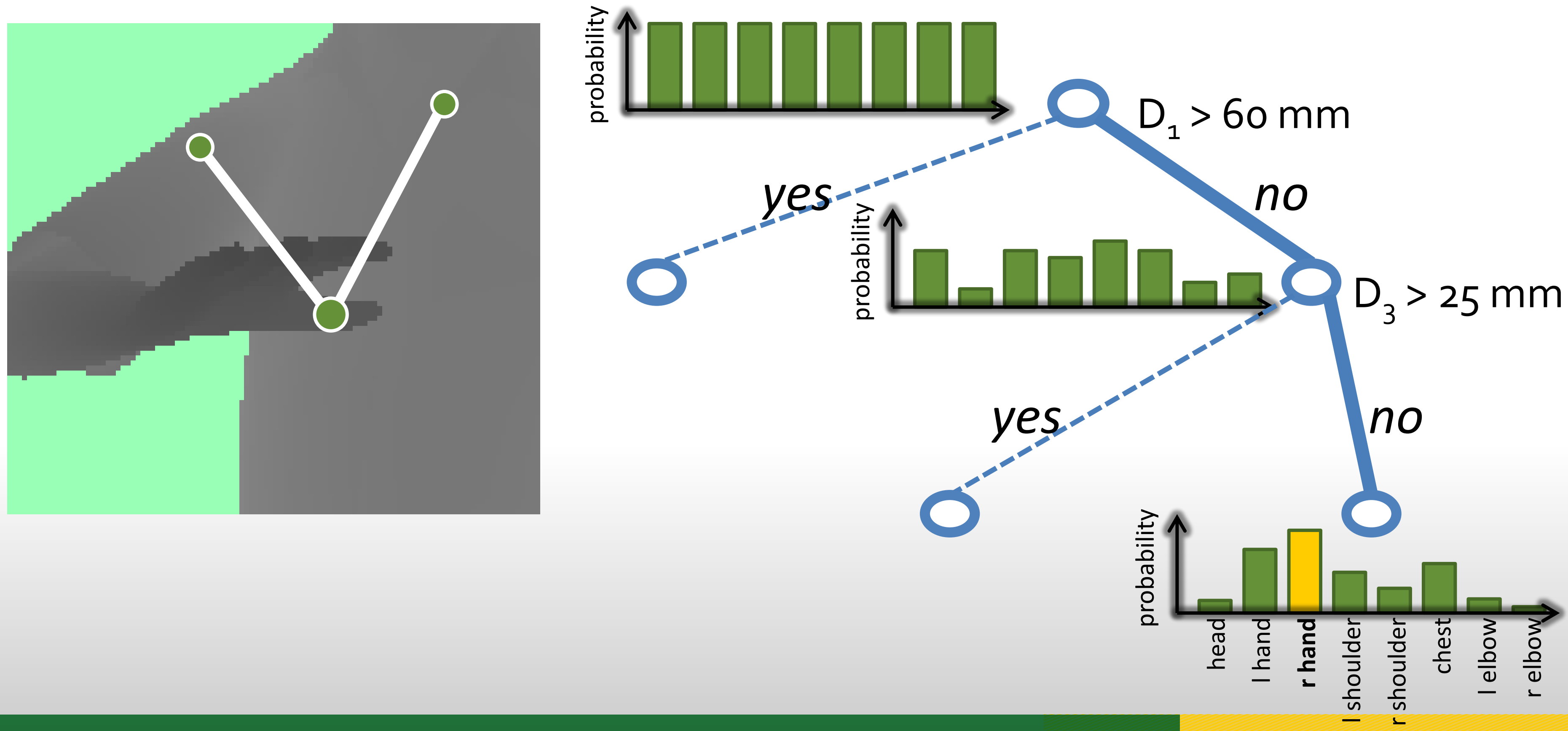
EXAMPLE PIXEL 2: WHAT PART AM I?

Microsoft



EXAMPLE PIXEL 2: WHAT PART AM I?

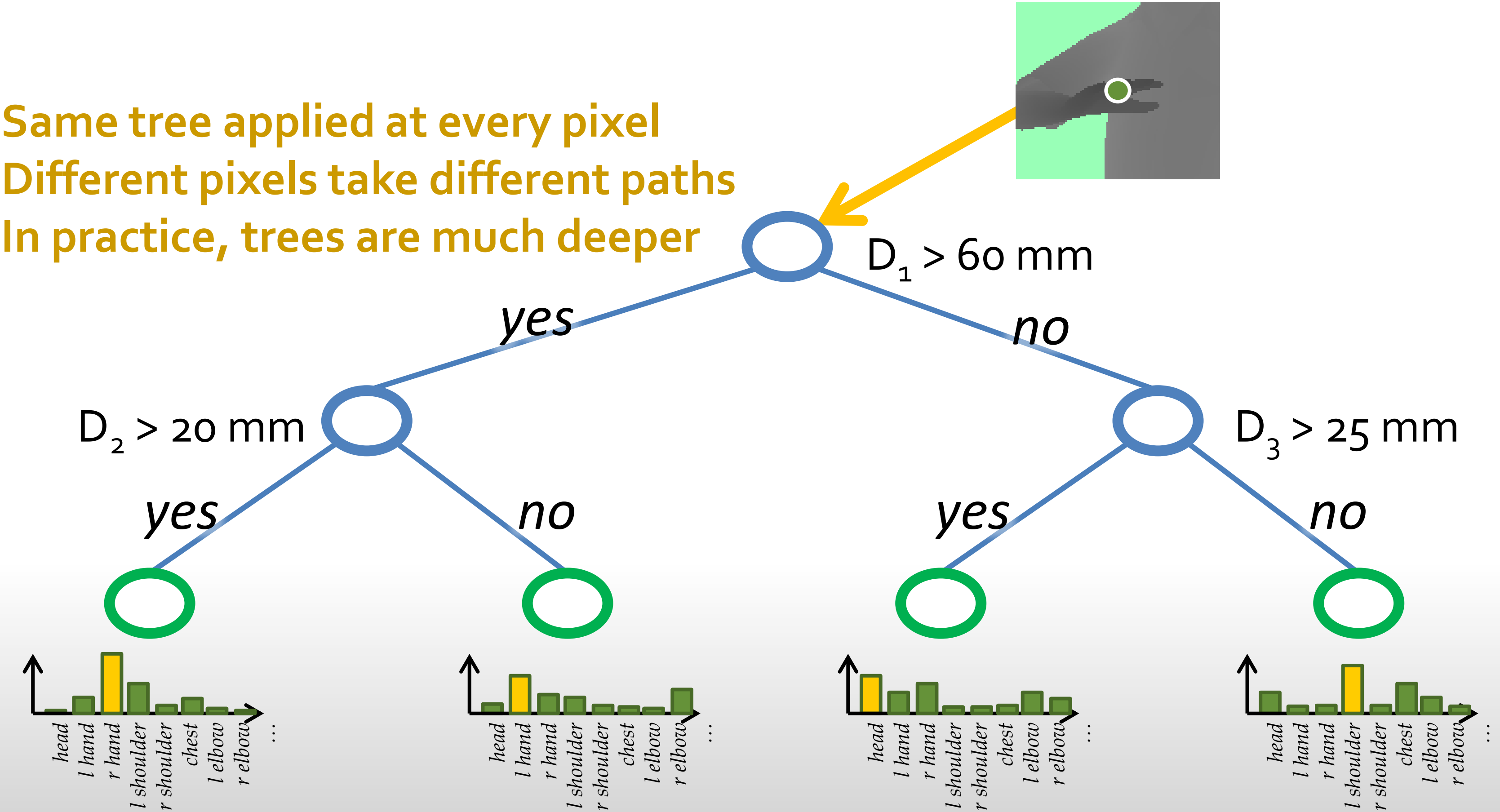
Microsoft



Example pixel 2: What part am I?

Microsoft

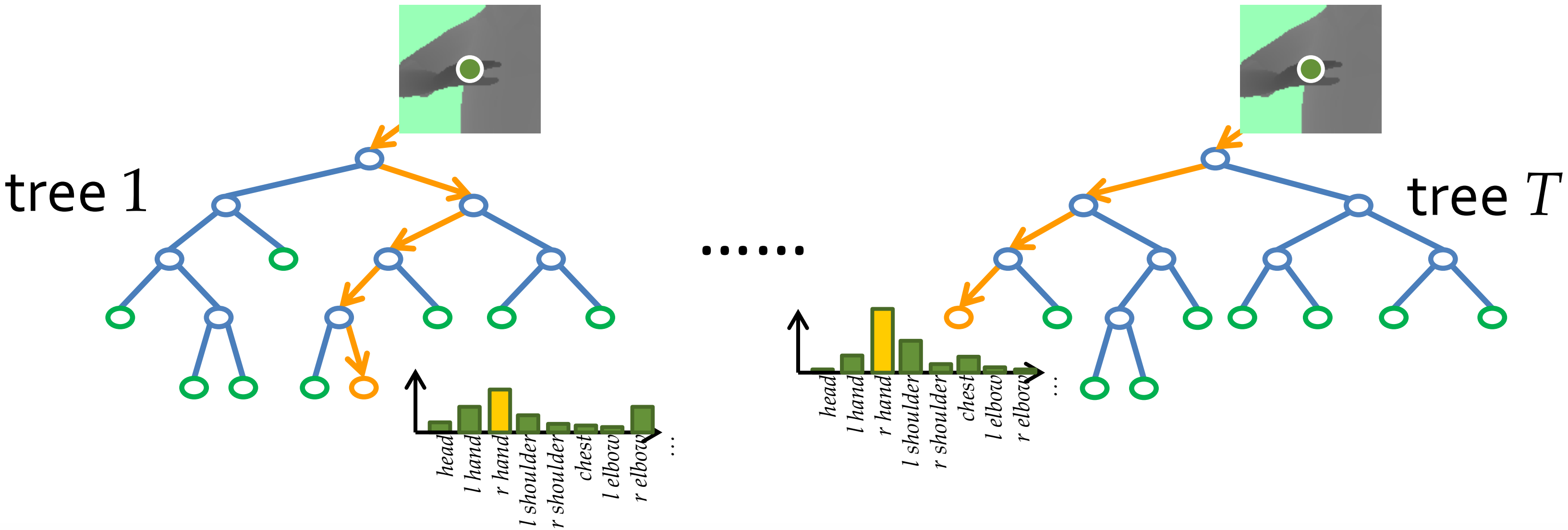
- Same tree applied at every pixel
- Different pixels take different paths
- In practice, trees are much deeper



DECISION TREE CLASSIFICATION

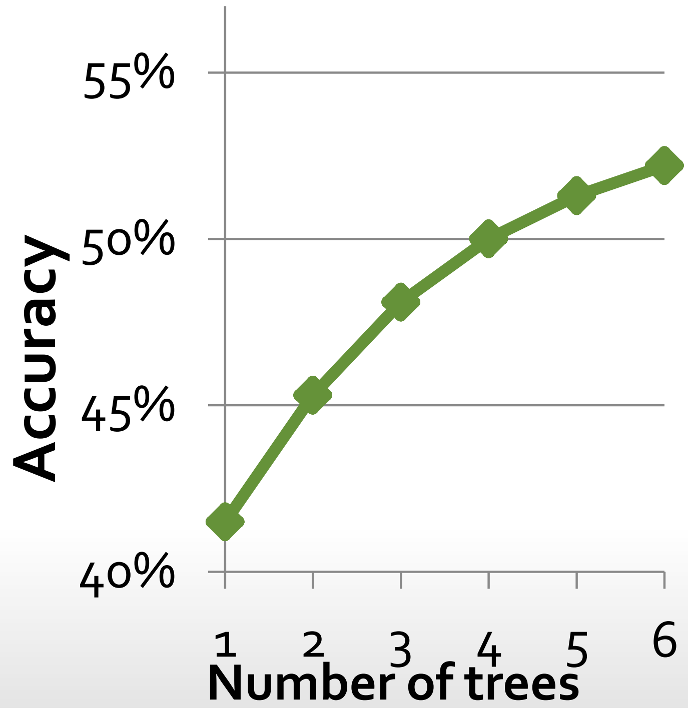
Microsoft

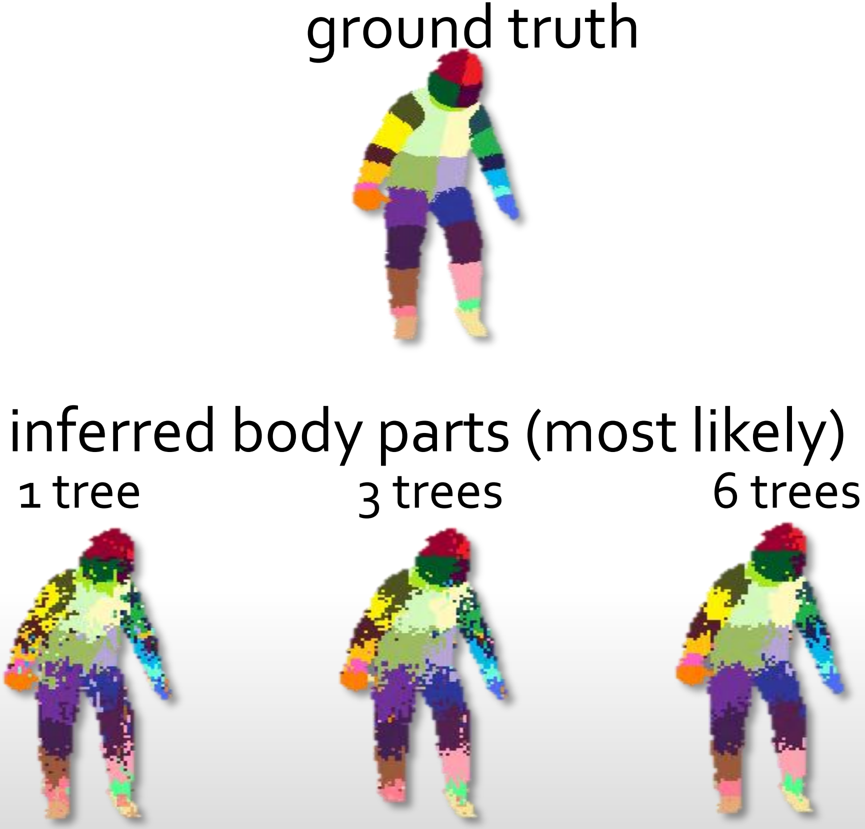
- A forest is an ensemble of trees:



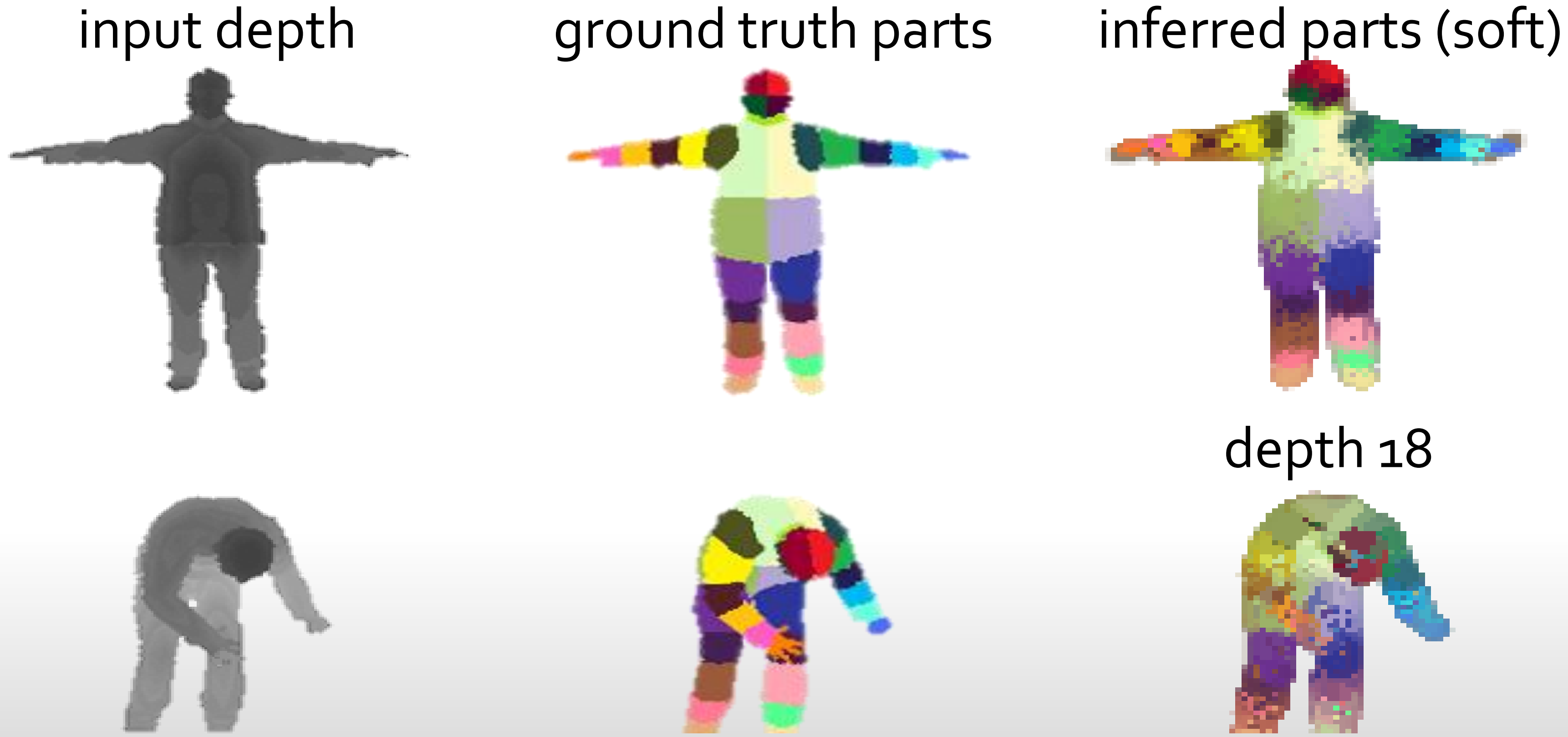
- Helps avoid over-fitting during training
- Testing takes average of leaf nodes distributions

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

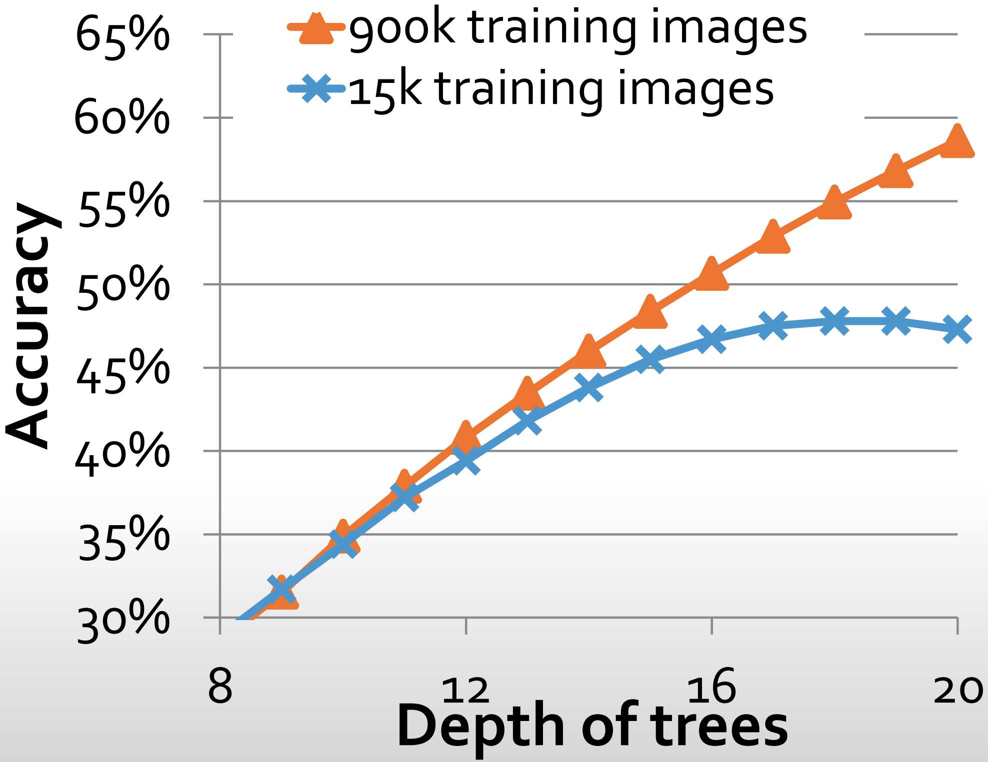




NUMBER OF TREES



DEPTH OF TREES



DEPTH OF TREES

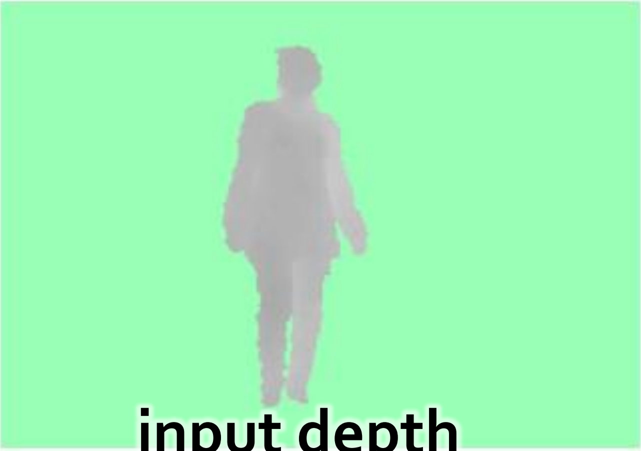
Microsoft

- Given
  - depth image
  - inferred body part probabilities

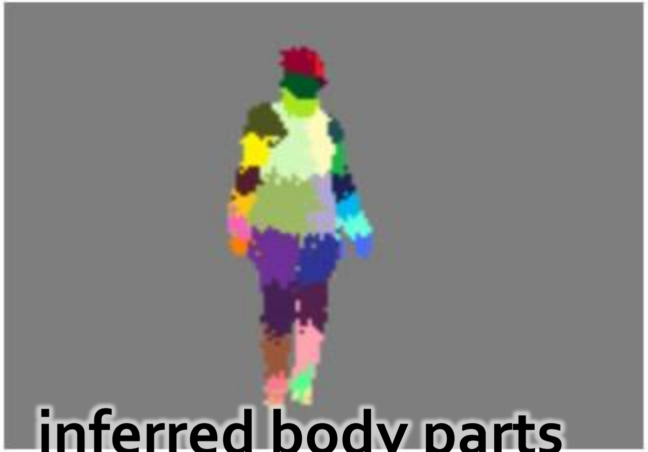
- Cluster high probability parts in 3D

hypothesized body joints

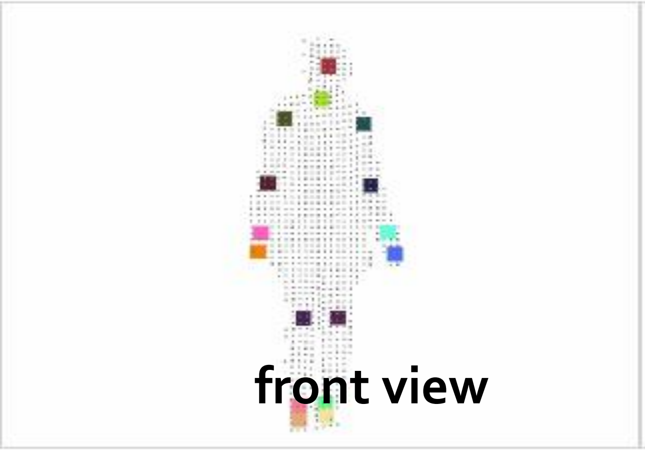
BODY PARTS TO JOINT POSITIONS



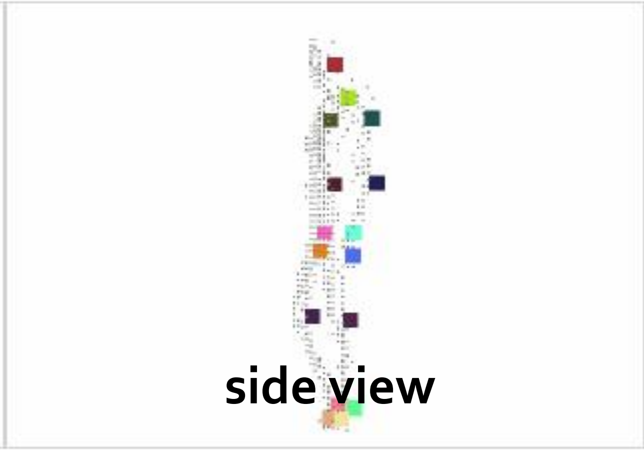
input depth



inferred body parts



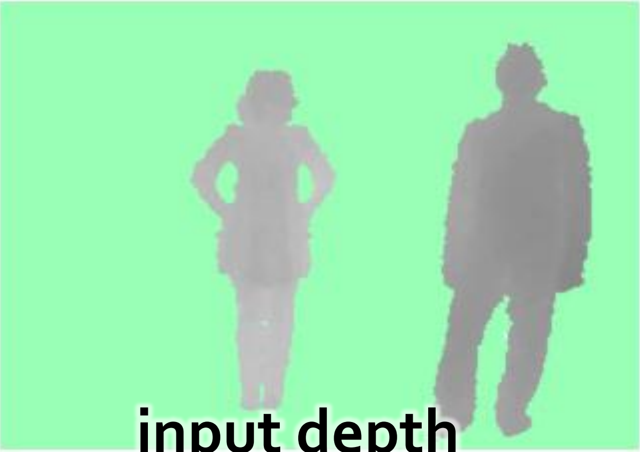
front view

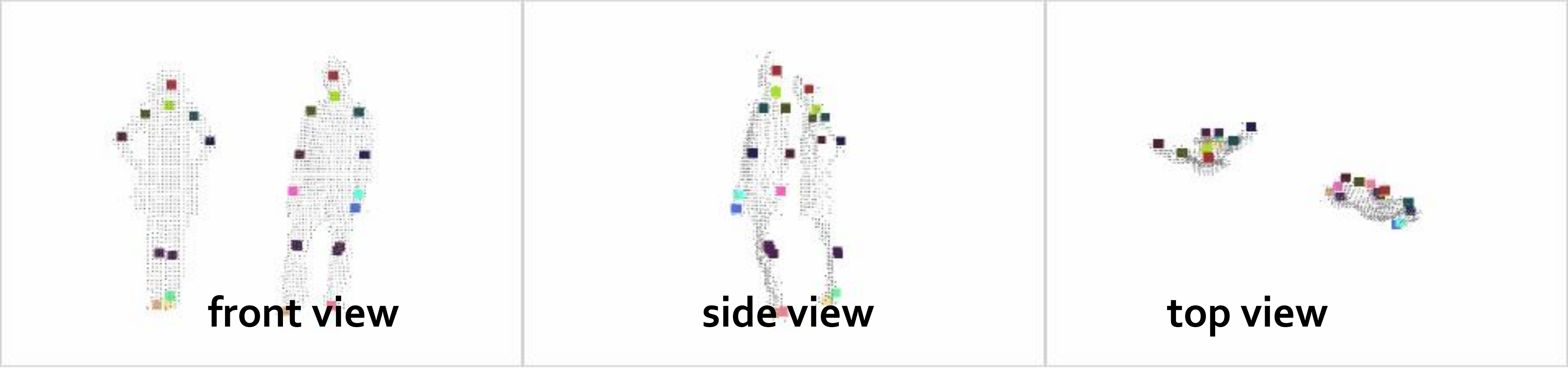


side view

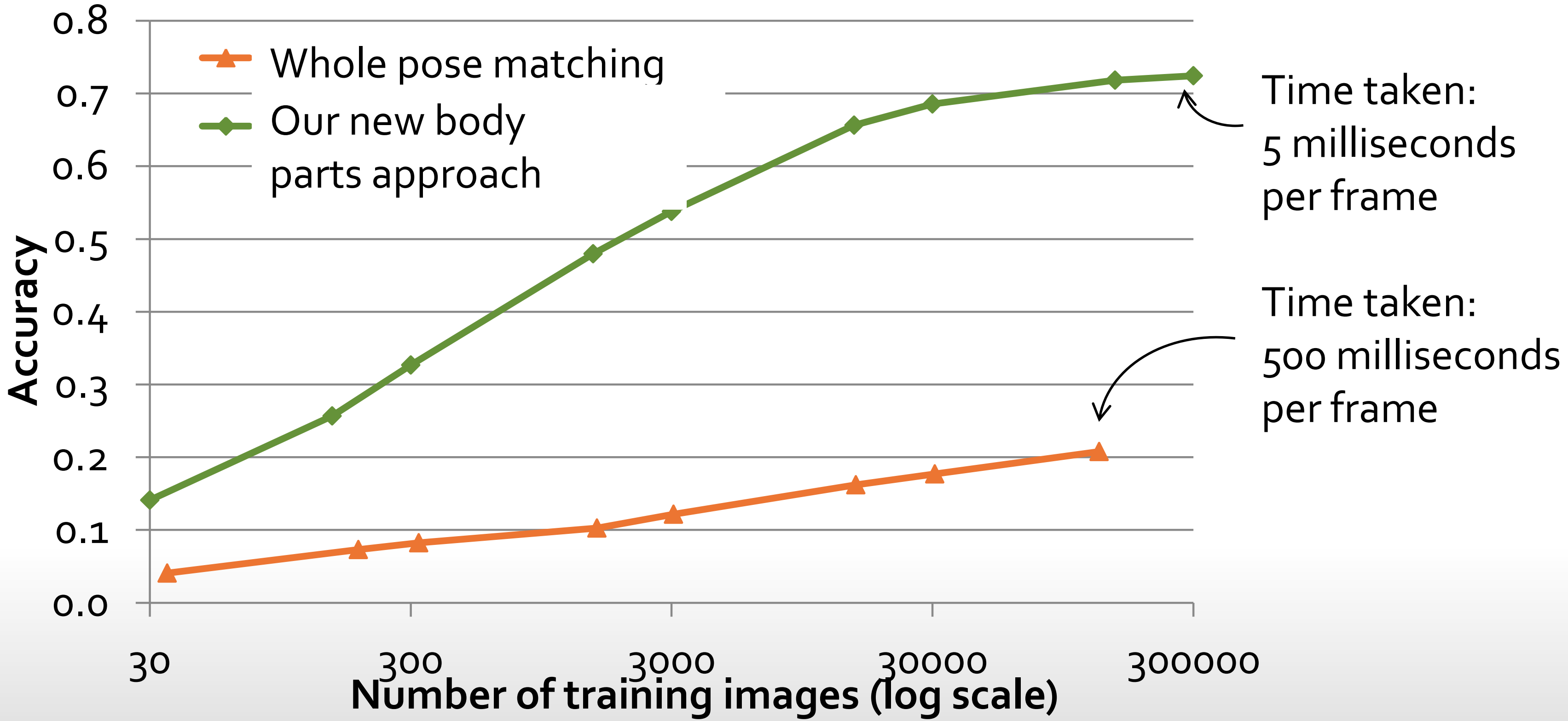
top view

inferred joint positions: no tracking or smoothing





inferred joint positions: no tracking or smoothing



MATCHING BODY *PARTS* IS BETTER

Microsoft

Microsoft Research

Joint position hypotheses are not the whole story

Follow up with skeleton fitting incorporating

- Kinematic constraints (limb lengths etc)
- Temporal coherence (it's back!)

And of course the incredible imagination of games designers…

WRAPPING UP

Microsoft

Microsoft Research

Joint position hypotheses are not the whole story

Follow up with skeleton fitting incorporating

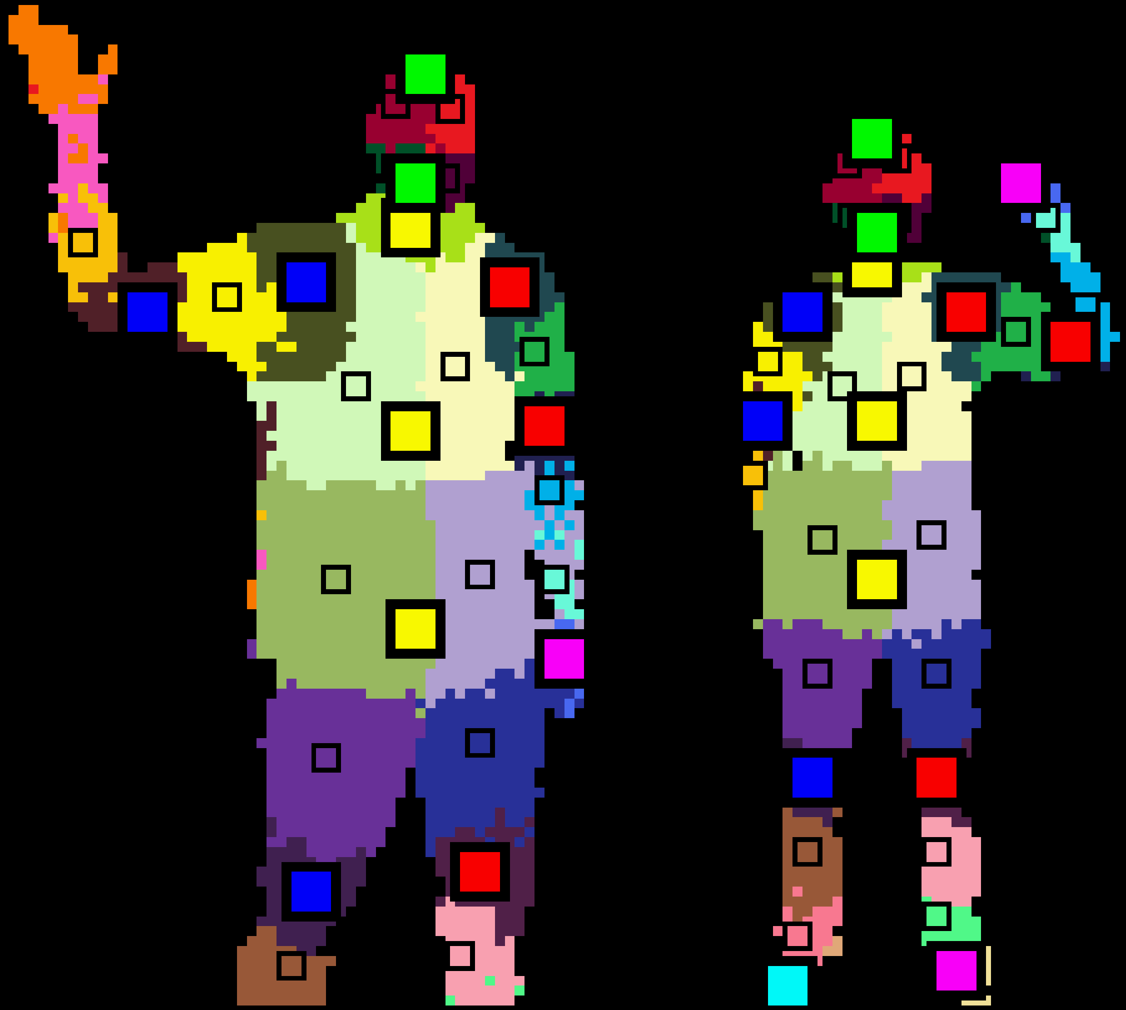
- Kinematic constraints (limb lengths etc)
- Temporal coherence (it's back!)

And of course the incredible imagination of games designers… and

YOU!

WRAPPING UP

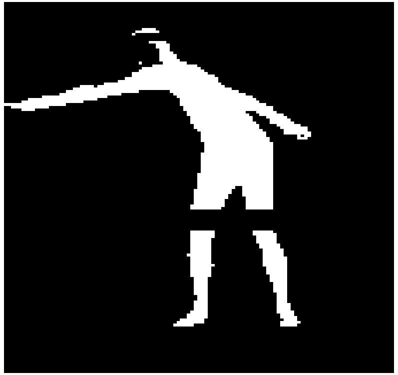
Microsoft



[Aside]

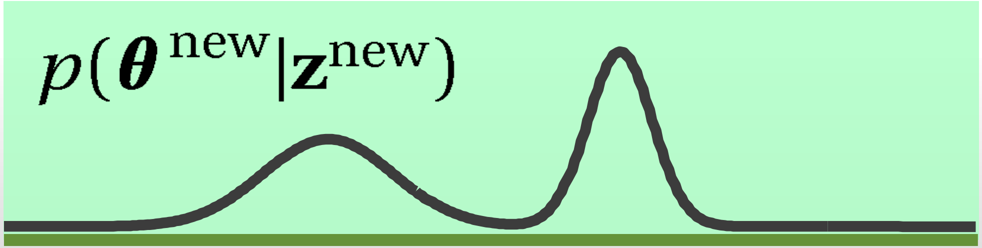
MULTIPLE HYPOTHESES AND REWRITING HISTORY

MULTIVALUED F:



or

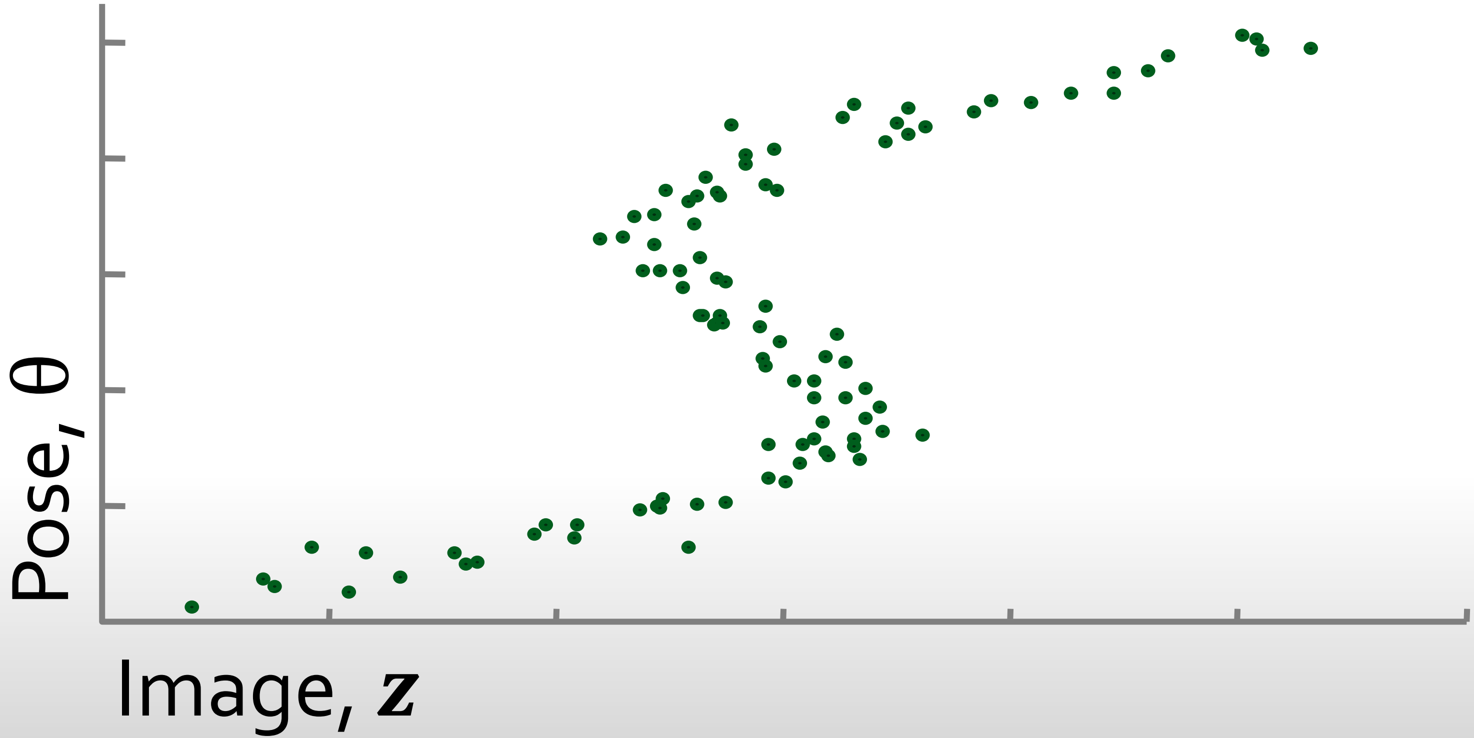
?



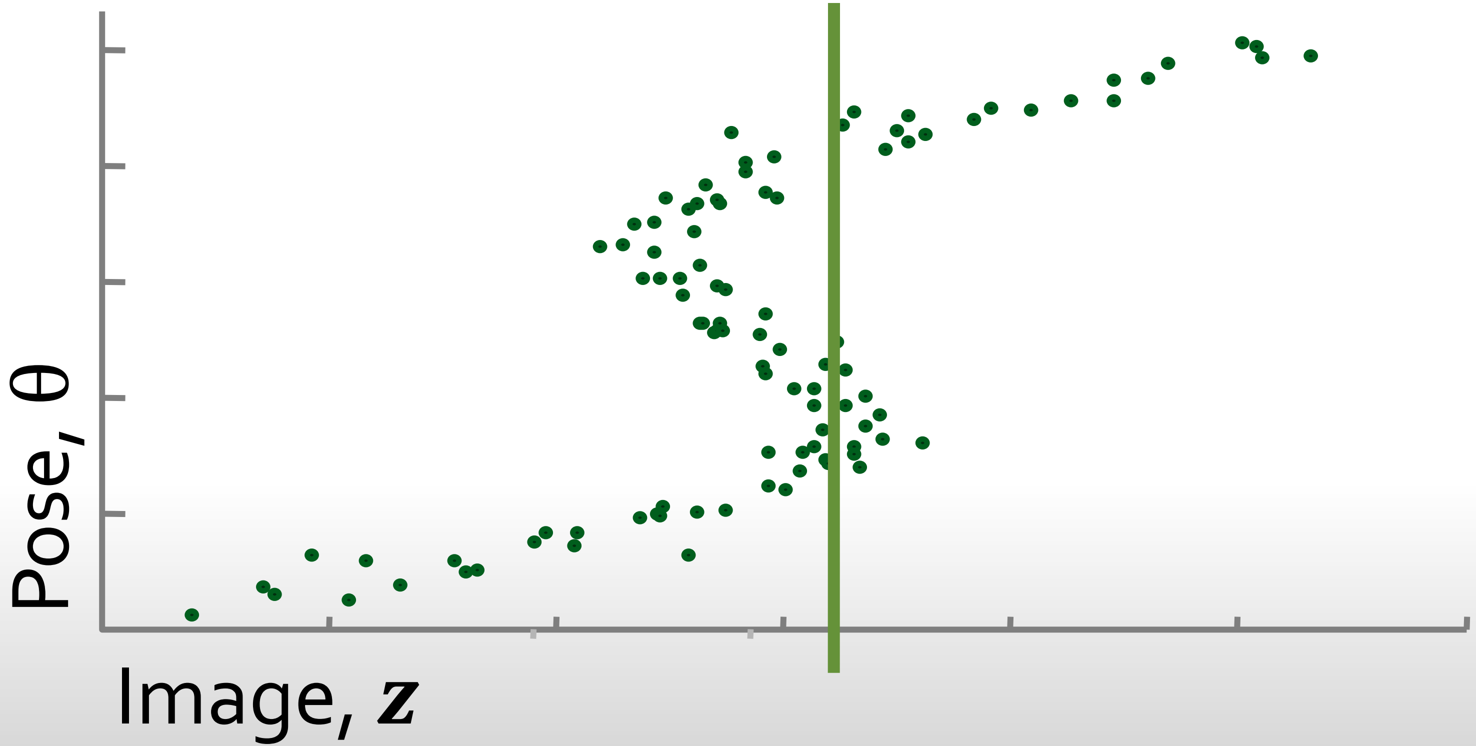
MULTIVALUED F:

62	10	16	36	30	52	49	44	32
67	87	24	7	81	5	27	39	85
41	0	13	22	33	21	67	40	79
8	70	12	11	10	32	17	28	50
21	74	78	17	70	74	62	46	56
82	78	52	22	35	1	17	46	53

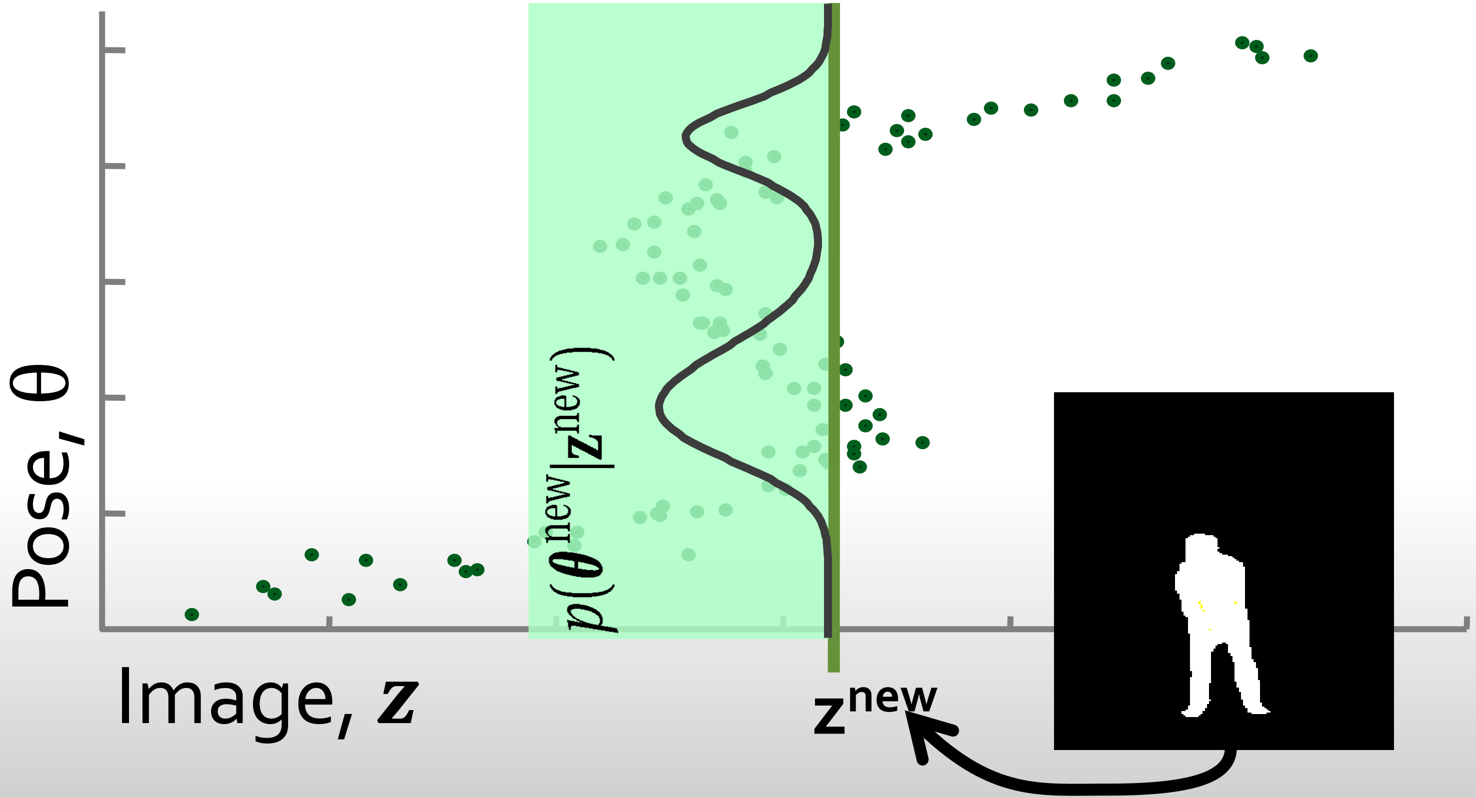
Training Data



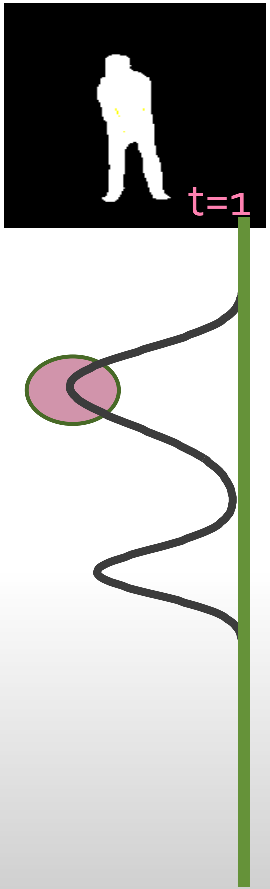
TRACKING BY REGRESSION



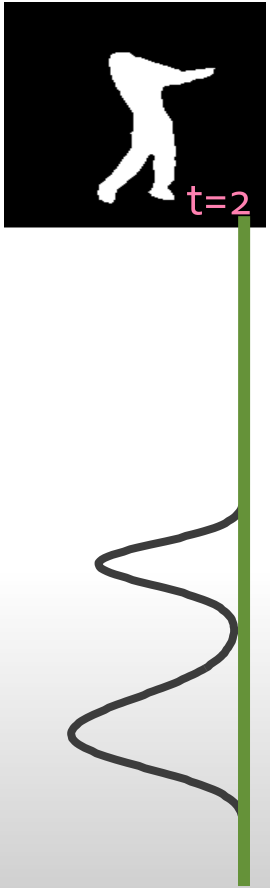
Tracking by Regression


Pose, θ
Image, z
$p(\boldsymbol{\theta}^{\text{new}}|\mathbf{z}^{\text{new}})$
$\mathbf{z}^{\text{new}}$

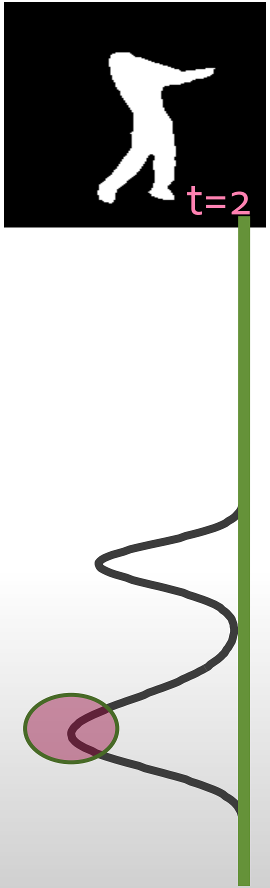

t=1



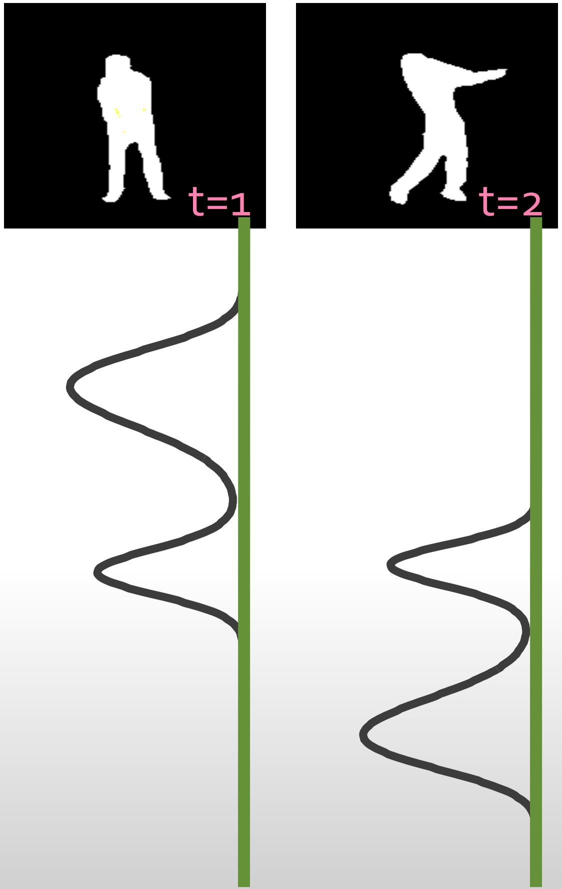
TEMPORAL FILTERING



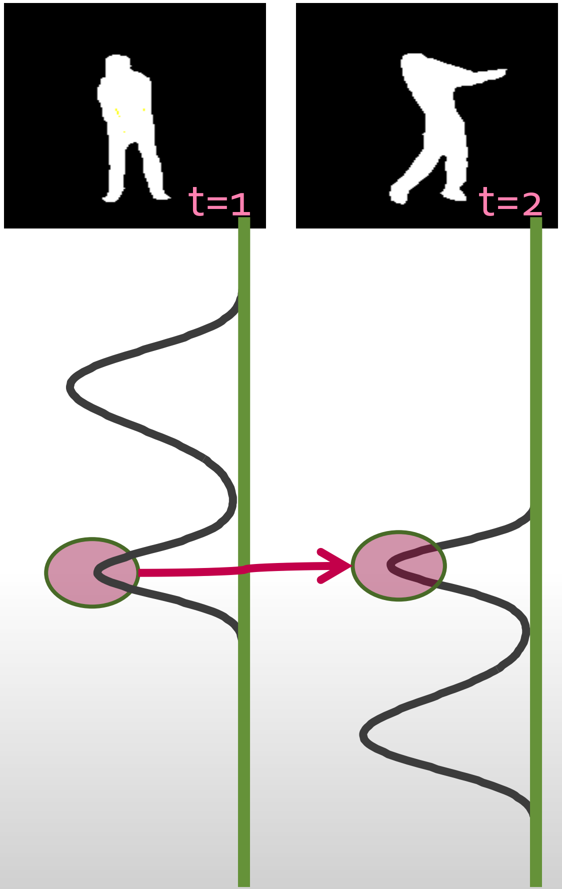
Temporal Filtering



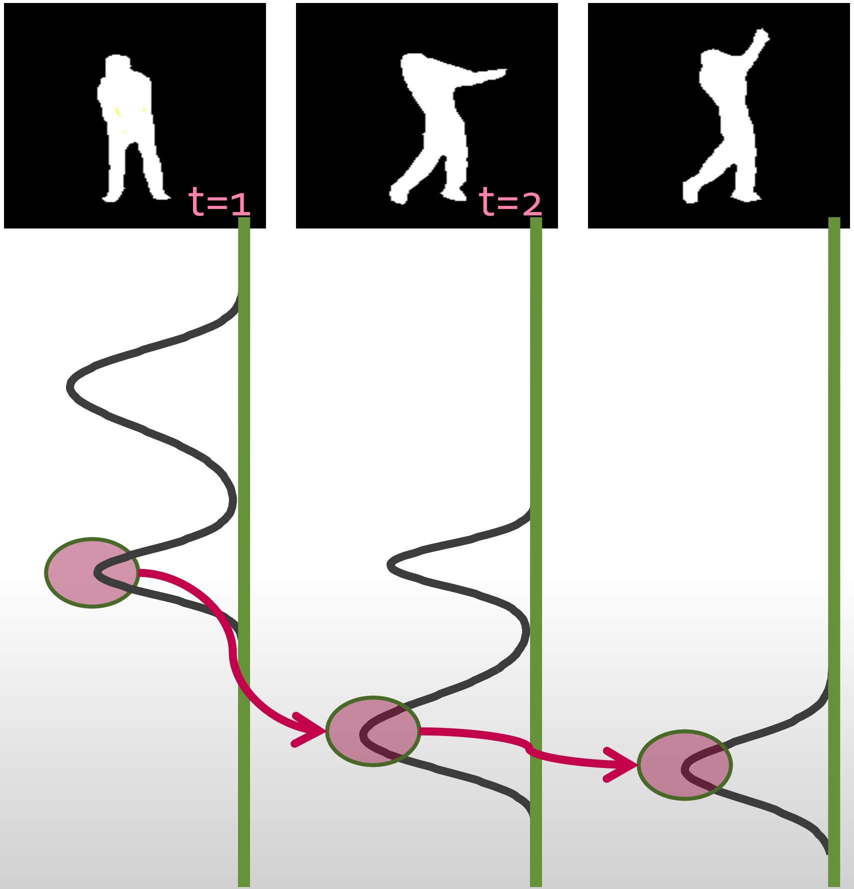
Temporal Filtering



Temporal Filtering



Temporal Filtering



Temporal Filtering

R Navaratnam, A Fitzgibbon, R Cipolla
The Joint Manifold Model for Semi-supervised Multi-valued Regression
IEEE International Conference on Computer Vision, 2007