15-869

Lecture 4
Game Motion Capture

Leonid Sigal
Human Motion Modeling and Analysis
Fall 2012

Projects

Capture Project

- Details are on the blog (questions?)
- 3D scanner demo during class today
- You should be starting this week

Projects

Capture Project

- Details are on the blog (questions?)
- 3D scanner demo during class today
- You should be starting this week

Upcoming Deadlines

- October 10 - Capture project in class presentations
- October 15 - Final project pitches (3 weeks)

Brief Review

Last Class: Motion capture more broadly

Practical systems for capturing motion
- Allow (some) editing of motion
- Can be used as measuring tools

This Class: Motion capture for gaming (and HCI)

[Minority Report, 2002]

Plan for Today's Class

- Review of some hardware devices for gaming
- Focus on Kinect
  - Structured light for depth estimation
  - Inference of 3D pose
- Discussion and some applications

Game Capture vs. Motion Capture

Technologies as you will see are very similar, but are tweaked for the HCI type of scenario

Usability:

Cost:

Computation:

Quality:

Game Capture vs. Motion Capture

Technologies as you will see are very similar, but are tweaked for the HCI type of scenario

Usability: Easy to use, put on and take off

Cost:

Computation:

Quality:

Game Capture vs. Motion Capture

Technologies as you will see are very similar, but are tweaked for the HCI type of scenario

Usability: 1 or less hand-held sensors

Cost:

Computation:

Quality:

Game Capture vs. Motion Capture

Technologies as you will see are very similar, but are tweaked for the HCI type of scenario

Usability: 1 or less hand-held sensors

Cost: < $100

Computation:

Quality:

Game Capture vs. Motion Capture

Technologies as you will see are very similar, but are tweaked for the HCI type of scenario

Usability: 1 or less hand-held sensors

Cost: < $100

Computation: Low (10% of CPU can be spent on sensing)

Quality:

Game Capture vs. Motion Capture

Technologies as you will see are very similar, but are tweaked for the HCI type of scenario

Usability: 1 or less hand-held sensors

Cost: < $100

Computation: Low (10% of CPU can be spent on sensing)

Quality: Low accuracy is OK, full body motion is not always unnecessary

Game Capture vs. Motion Capture

Technologies as you will see are very similar, but are tweaked for the HCI type of scenario

Usability: 1 or less hand-held sensors

Cost: < $100

Computation: Low (10% of CPU can be spent on sensing)

Quality: Low accuracy is OK, full body motion is not always unnecessary

Wiimote

- Wireless communication (Bluetooth)
- Sensors
    - Accelerometer for orientation (3 axis)
    - Optical sensor for pointing
- Supports two handed interaction
    - Can use 2 Wiimotes simultaneously

[Some content taken from Joseph LaViola]

Wiimote

- Wireless communication (Bluetooth)
- Sensors
  - Accelerometer for orientation (3 axis)
  - Optical sensor for pointing
- Supports two handed interaction
  - Can use 2 Wiimotes simultaneously

Simplified version of:

Inertial Suites

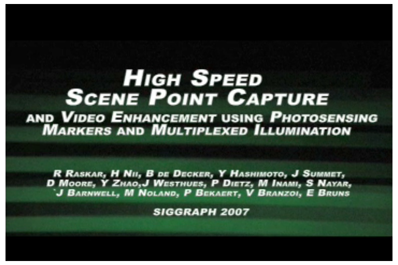
- Inertial sensors (gyros)
  - Accelerometer: measures acceleration
  - Gyroscope: measures orientation

[Xsens]

+

Semi-passive

- Multi-LED IR projectors in the environment emit spatially varying patterns
- Photo-sensitive marker tags decode the signals and estimate their position



[Some content taken from Joseph LaViola]

Wiimote

- Wireless communication (Bluetooth)
- Sensors
  - Accelerometer for orientation (3 axis)
  - Optical sensor for pointing
- Supports two handed interaction
  - Can use 2 Wiimotes simultaneously

Can be used by themselves or jointly

[Some content taken from Joseph LaViola]

Wiimote

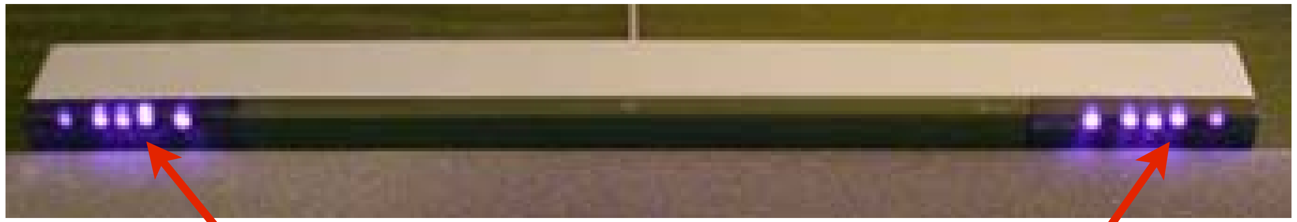
- Wireless communication (Bluetooth)
- Sensors
  - Accelerometer for orientation (3 axis)
  - Optical sensor for pointing
- Supports two handed interaction
  - Can use 2 Wiimotes simultaneously

Can be used by themselves or jointly

optical sensor is used for more accurate aiming control

[Some content taken from Joseph LaViola]

Optical Sensing with Wiimote



10 LED lights (5 on each side)

Use triangulation to determine depth

- Distance between imaged LEDs on sensor varies with depth
- Distance between LEDs on the sensor bar fixed
- Angle can be calculated from angle between imaged LEDs

[Some content taken from Joseph LaViola]

Wiimote Limitations

- Not quite 6 DOFs (orientation + depth)
- Only provides approximate depth
- Limited range (~ 5 meters)
- To triangulate depth requires line of sight to the bar

[Some content taken from Joseph LaViola]

Wiimote Limitations

- Not quite 6 DOFs (orientation + depth)
- Only provides approximate depth
- Limited range (~ 5 meters)
- To triangulate depth requires line of sight to the bar

Wii Motion Plus

- Adds a gyro for additional orientation quality
- Still unable to provide reliable 3D position information

[Some content taken from Joseph LaViola]

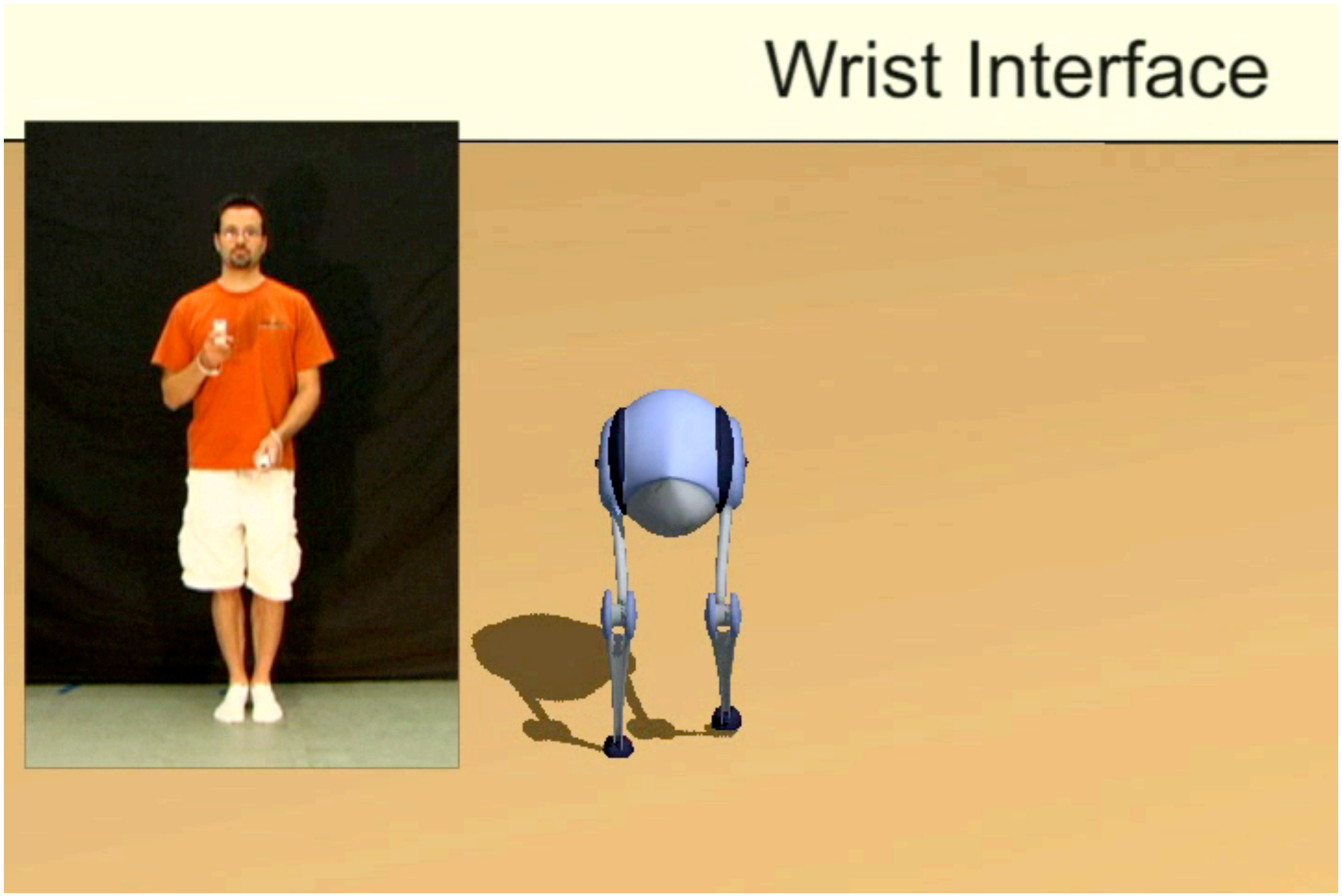
Prototype Systems

Use multiple Wiimotes and map them to motion of a character

[Shiratori and Hodgins, ACM SIGGRAPH Asia, 2008]

Prototype Systems

Use multiple Wiimotes and map them to motion of a character



[Shiratori and Hodgins, ACM SIGGRAPH Asia, 2008]

PlayStation Move

- Wireless communication
- Sensors
    - Optical camera tracking (absolute 3D position)
    - 3 axis accelerometer
    - 3 axis gyroscope
    - magnetometer (helps with drift)
- Can use up to 4 controllers simultaneously

[Some content taken from Joseph LaViola]

PlayStation Move

- Wireless communication
- Sensors
    - Optical camera tracking (absolute 3D position)
    - 3 axis accelerometer
    - 3 axis gyroscope
    - magnetometer (helps with drift)
- Can use up to 4 controllers simultaneously
- PlayStation Eye
    - 640 x 480 (60 Hz)
    - 320 x 240 (120 Hz)

[Some content taken from Joseph LaViola]

PlayStation Move

- Wireless communication
- Sensors
  - Optical camera tracking (absolute 3D position)
  - 3 axis accelerometer
  - 3 axis gyroscope
  - magnetometer (helps with drift)

Simplified version of:

Inertial Suites

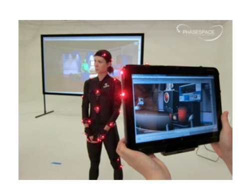
- Inertial sensors (gyros)
  - Accelerometer: measures acceleration
  - Gyroscope: measures orientation

[Xsens]

+

Active Marker-based Systems

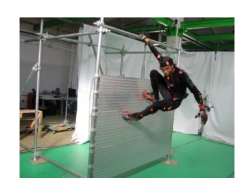
- Resolve correspondence by activating one LED marker at a time (very quickly)
- LEDs can be tuned to be easily picked up by cameras



[PhaseSpace]

Weta used for "Rise of the Planet of the Apes"

ILM used for "Van Helsing"



[Some content taken from Joseph LaViola]

PlayStation Move: Optical Tracking

44mm sphere serves as an active LED marker
(with controlled color)

Controllable color simplifies

- Correspondences (immediately know id of controller)
- Segmentation of the marker from background
  (remember active optical markers)

[Some content taken from Joseph LaViola]

PlayStation Move: Optical Tracking

44mm sphere serves as an active LED marker
(with controlled color)

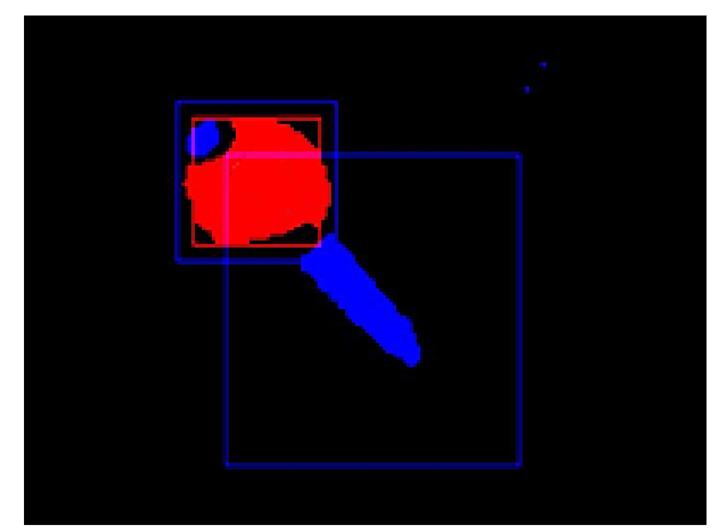
Under perspective projection spherical marker images
as an ellipsoid

[Some content taken from Joseph LaViola]

PlayStation Move: Optical Tracking

44mm sphere serves as an active LED marker
(with controlled color)

Under perspective projection spherical marker images
as an ellipsoid



- Detect marker pixels
- Fit ellipsoid to them
- Ellipsoid + calibration = 3D position
  - Ray through centroid gives a line in space
  - Size and orientation of the ellipsoid give depth along the line

[Some content taken from Joseph LaViola]

Playstation Move Limitations

- 6 DOFs (orientation + position in 3D)
- Limited range (~ 5 meters)
- Requires line of sight to the camera

Kinect

KINECT for XBOX 360

Two key contributions:

- Inexpensive and accurate depth camera / sensor
- 3D pose estimation

Color Image | Depth Image | Body Part Segmentation | 3D Joint Estimation

Kinect

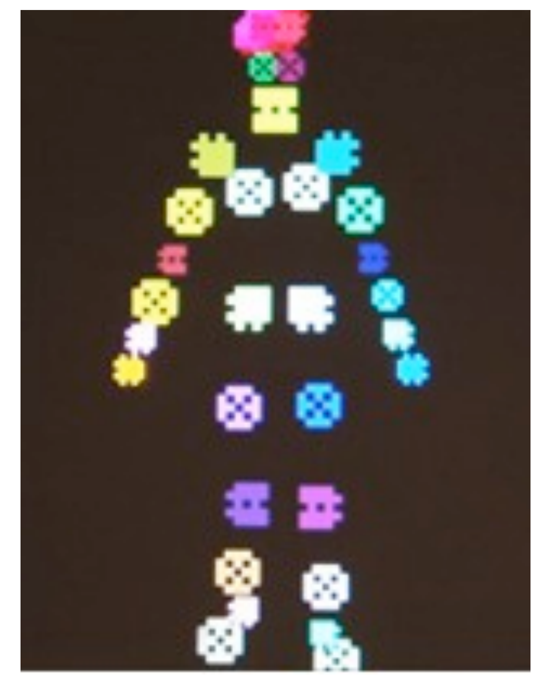
Two key contributions:

- Inexpensive and accurate depth camera / sensor
- 3D pose estimation

Color Image

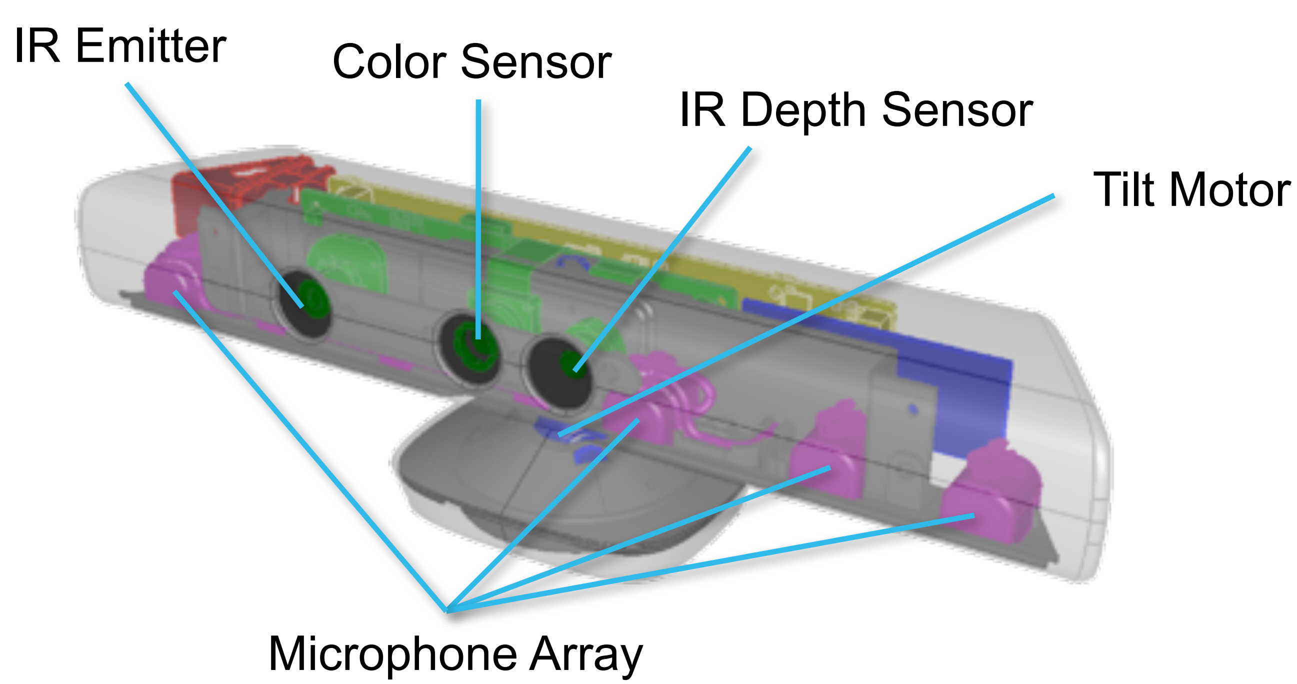
Depth Image

Body Part Segmentation



3D Joint Estimation

Structure of the Sensor



[Src: Kinect for Windows SDK]

Depth Map Construction

Kinect combines structured light with two other computer vision techniques: depth from focus and depth from stereo

[Slide after John MacCormick]

Depth Map Construction

Kinect combines structured light with two other computer vision techniques: depth from focus and depth from stereo

(Everything I will tell you is a speculation, taken from PrimeSense patent and notes from John MacCormick, Microsoft)

[Slide after John MacCormick]

Depth Map Construction

Kinect combines structured light with two other computer vision techniques: depth from focus and depth from stereo

(Everything I will tell you is a speculation, taken from PrimeSense patent and notes from John MacCormick, Microsoft)

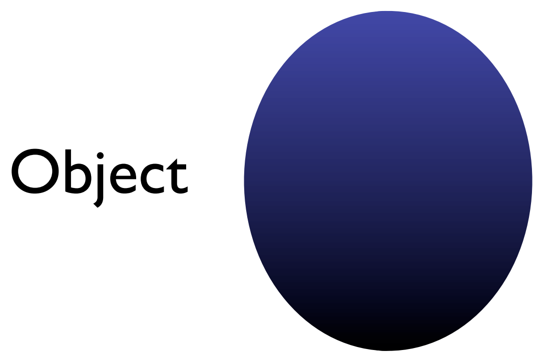
Structured light has a long history in vision:

[Zhang et al, 3DPVT, 2002]

Cleverly projected pattern of light observed from the camera can tell us a lot about 3D structure of the scene

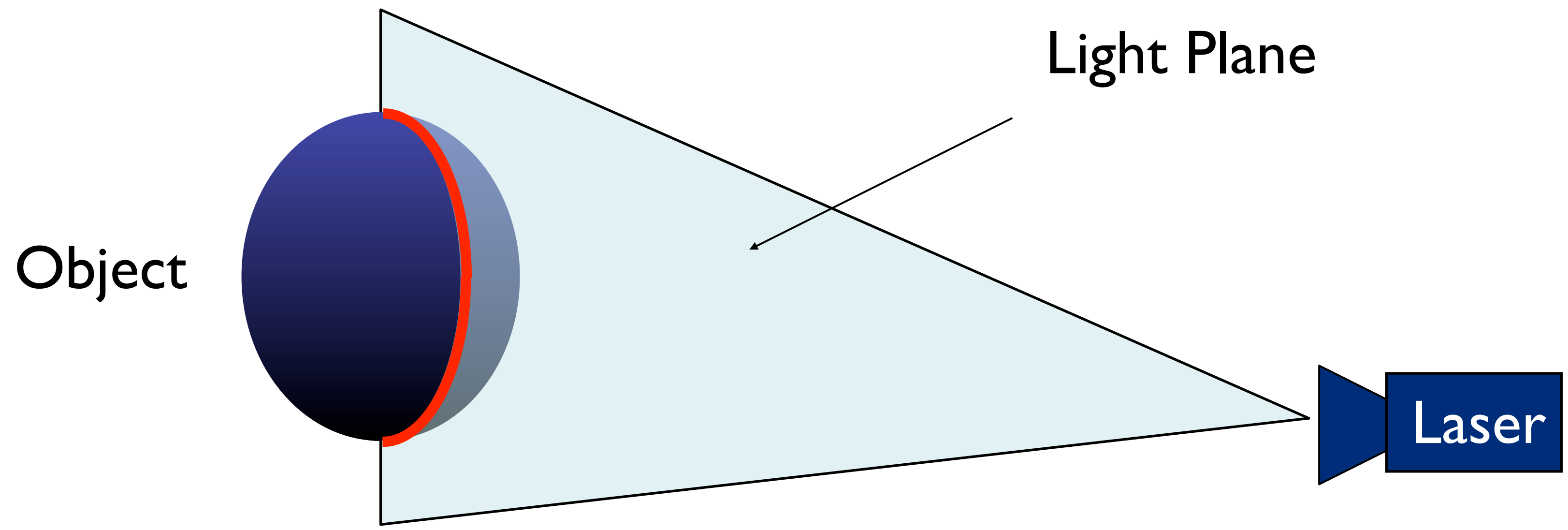
[Slide after John MacCormick]

Example: Line as a “the structure”



[Slide from S. Narasimhan]

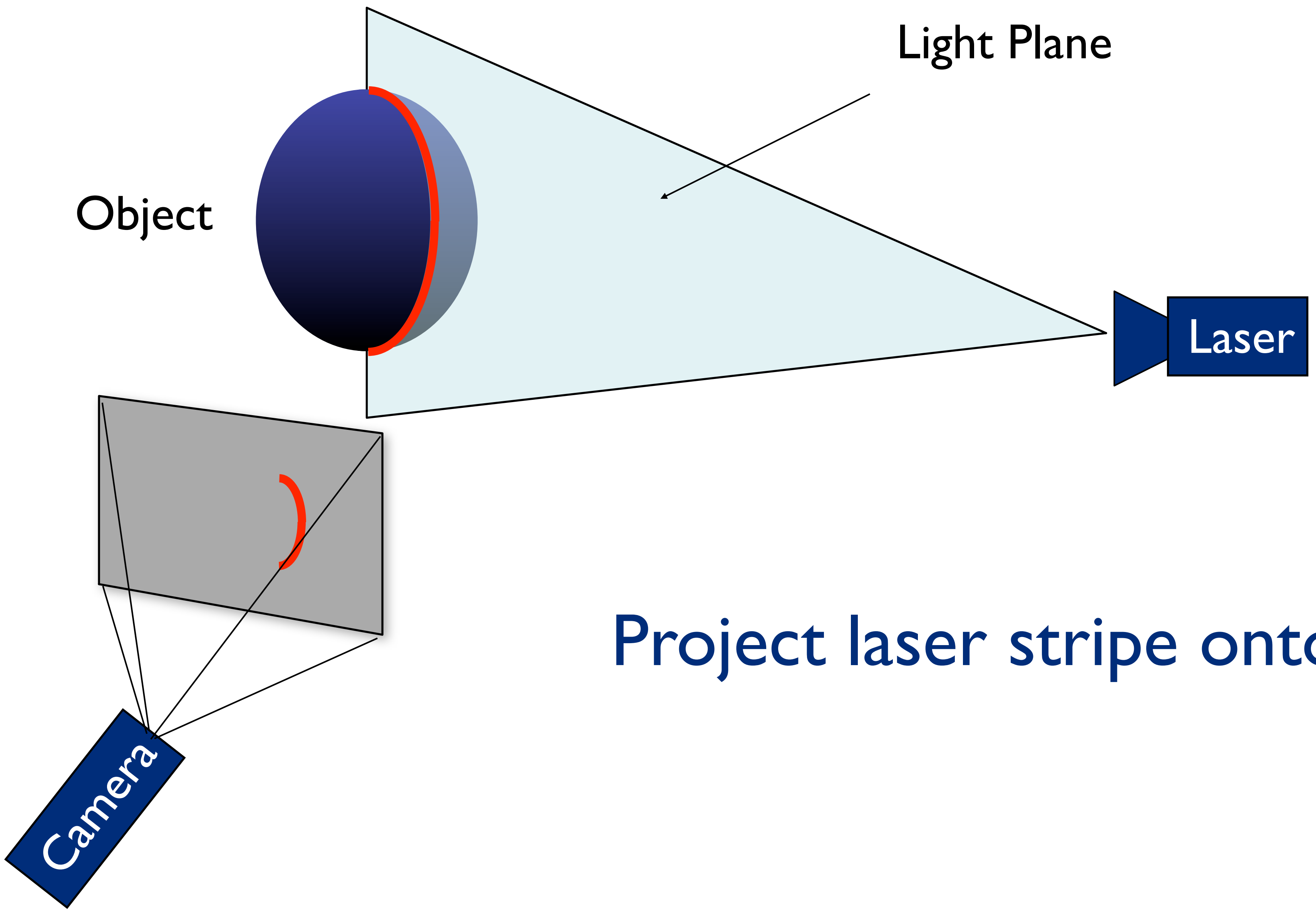
Example: Line as a “the structure”



Project laser stripe onto object

[Slide from S. Narasimhan]

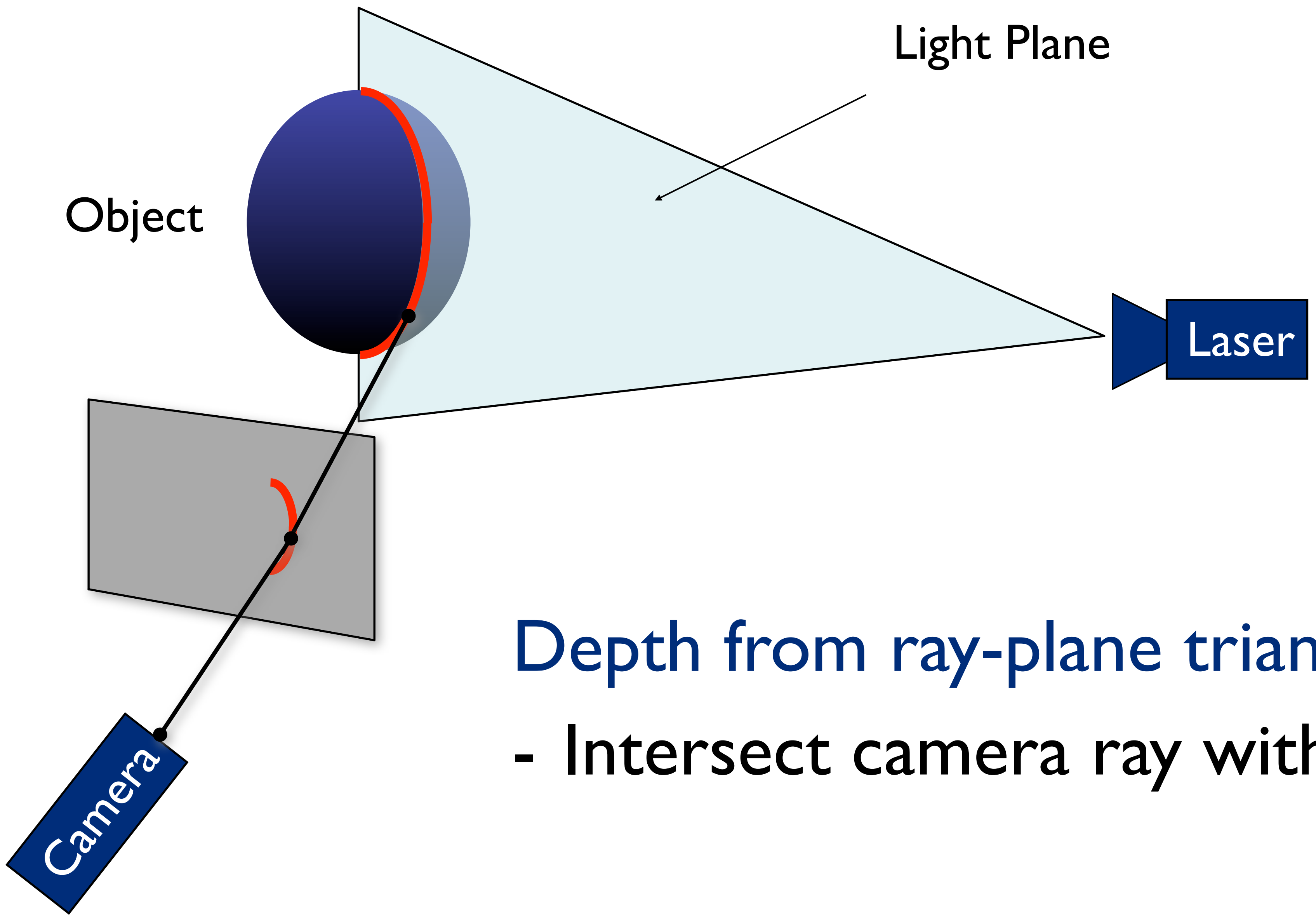
Example: Line as a “the structure”



Project laser stripe onto object

[Slide from S. Narasimhan]

Example: Line as a “the structure”

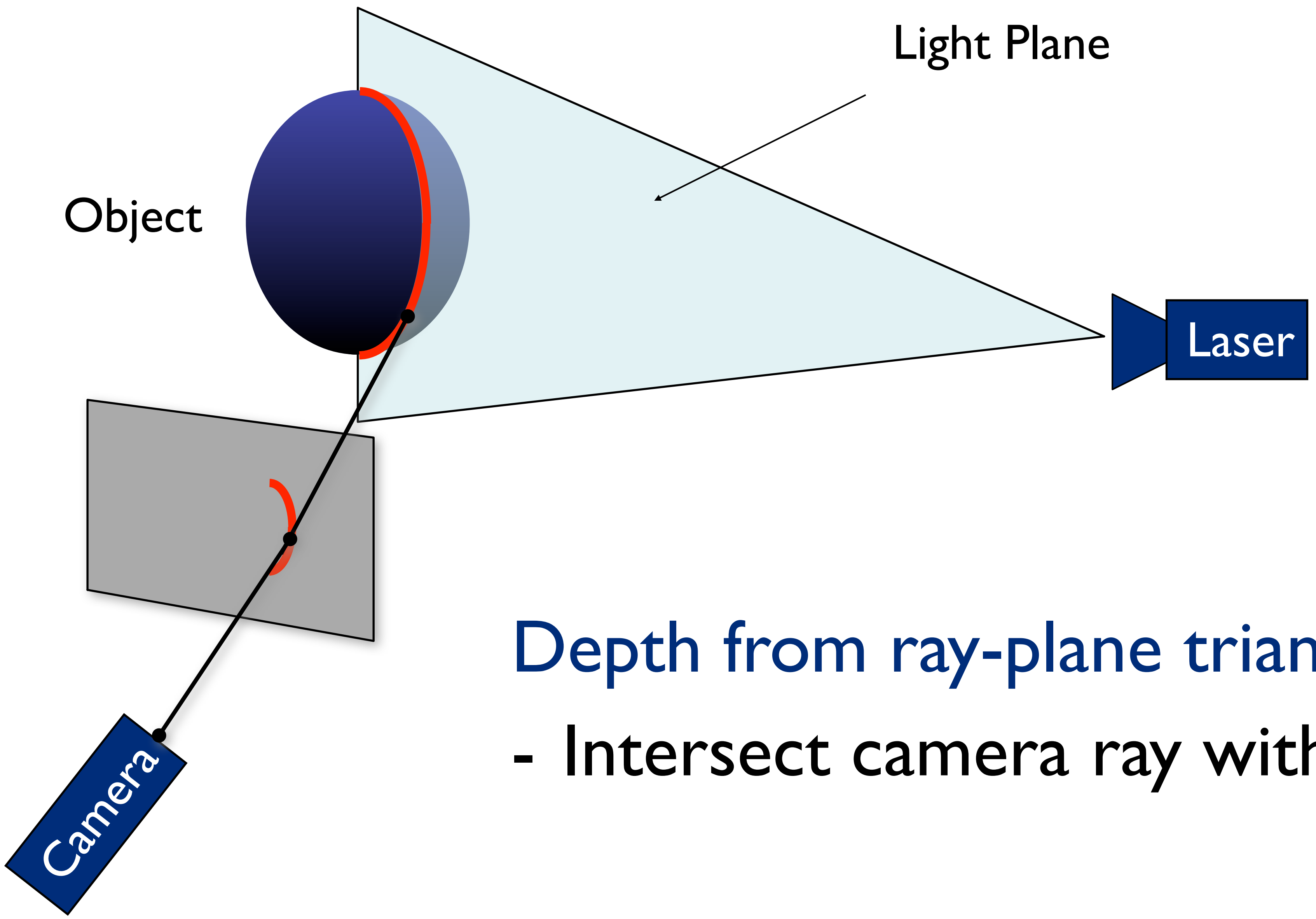


Depth from ray-plane triangulation:

- Intersect camera ray with light plane

[Slide from S. Narasimhan]

Example: Line as a “the structure”



Depth from ray-plane triangulation:

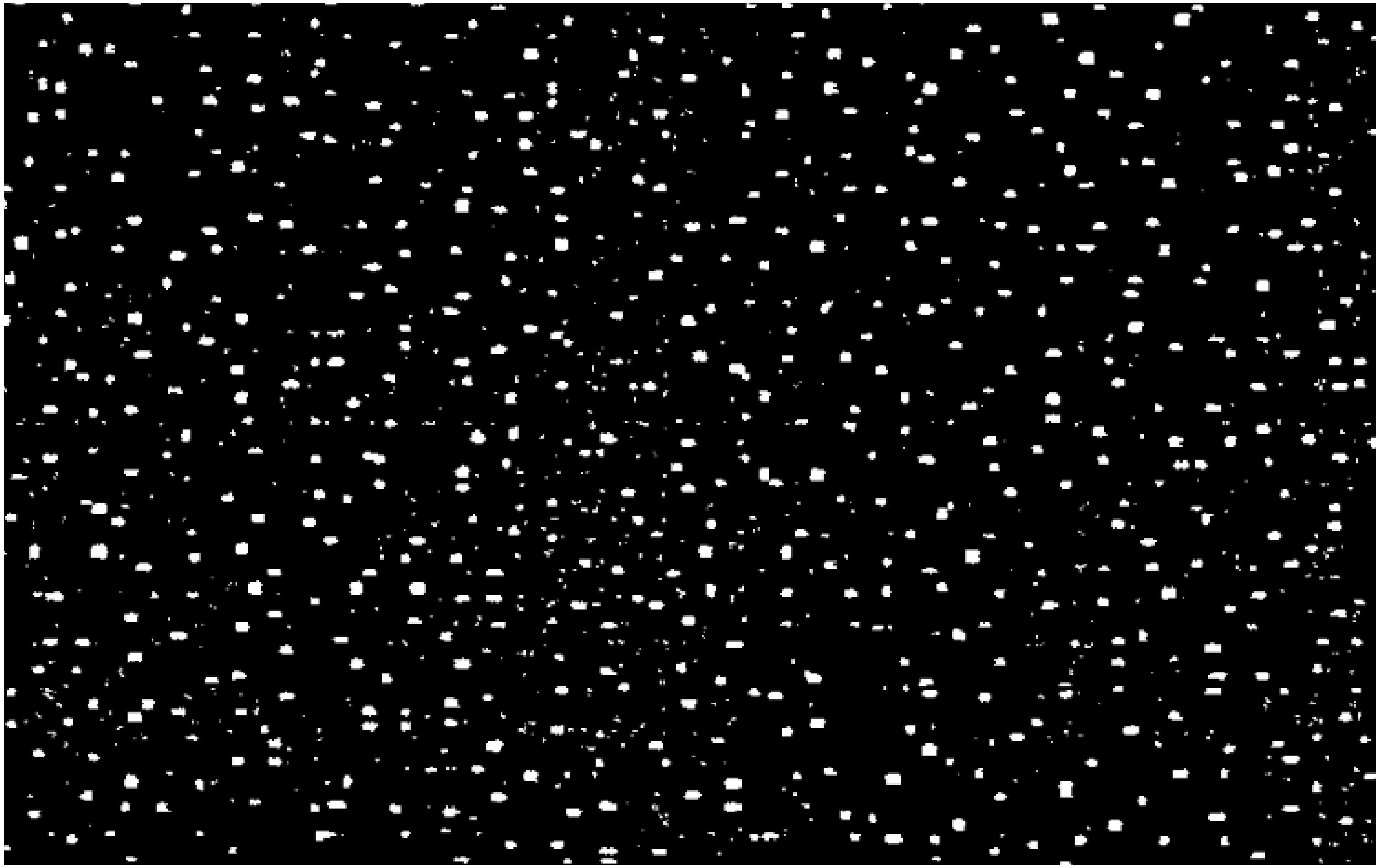
- Intersect camera ray with light plane

Accurate but slow

[Slide from S. Narasimhan]

Kinect's Structured Light

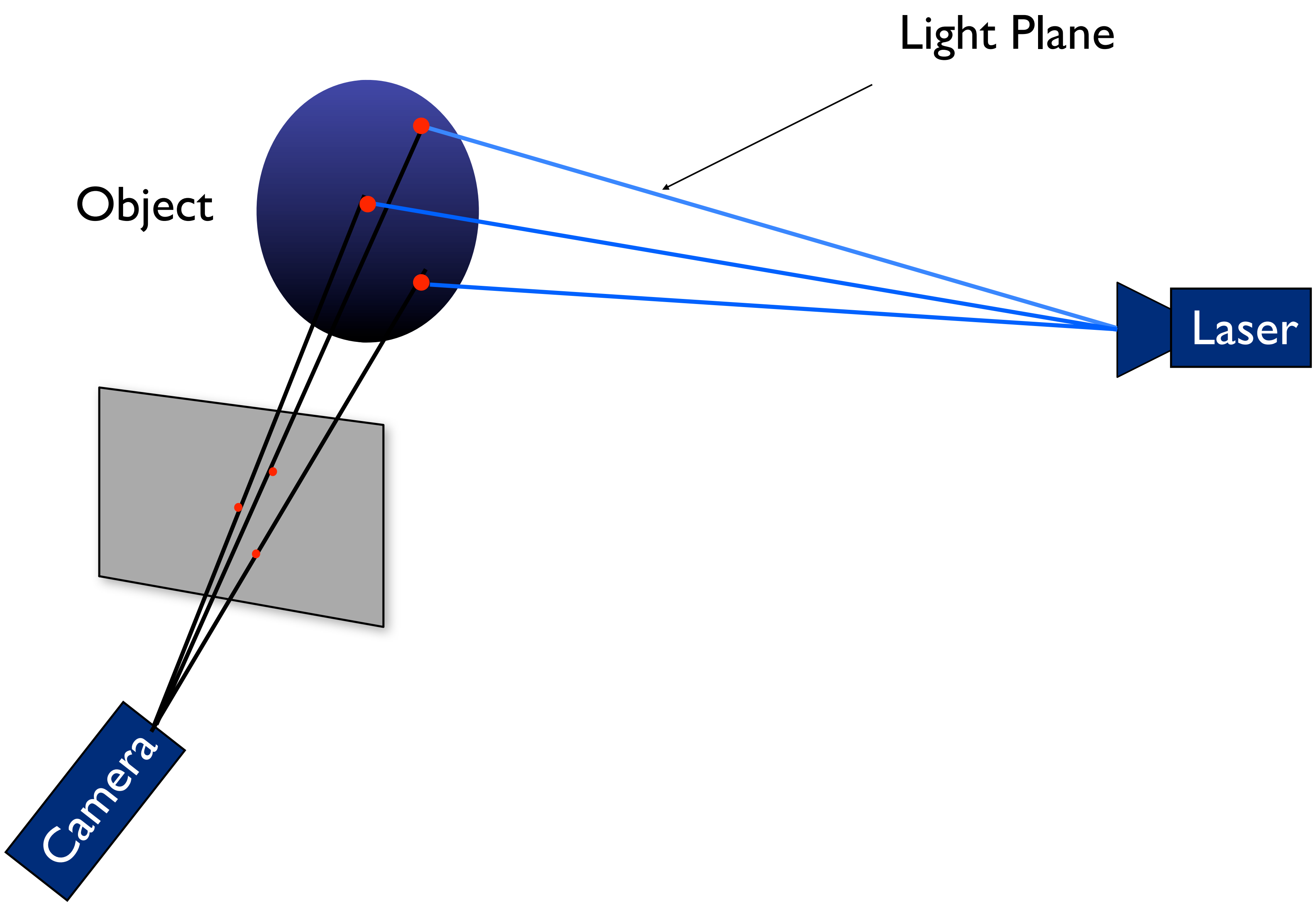
Speckle patterns using infrared light



[Shpunt et al., PrimeSense patent application US 2008/0106746]

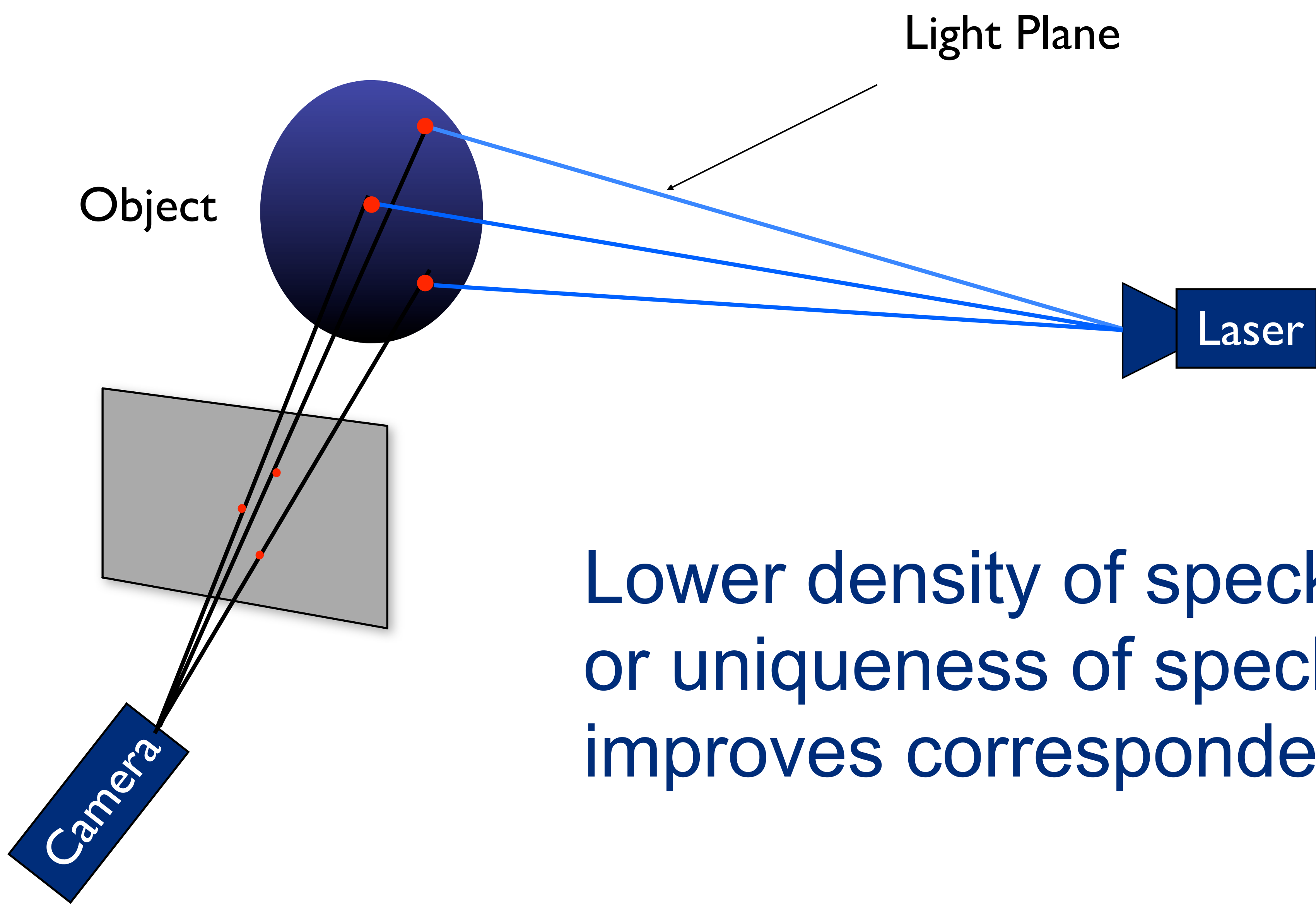
[Slide after John MacCormick]

Dealing with Correspondences



Correspondences can be computed based on closest distance between intersection of 3D lines (from camera and laser)

Dealing with Correspondences



Lower density of speckles and/ or uniqueness of speckles improves correspondence

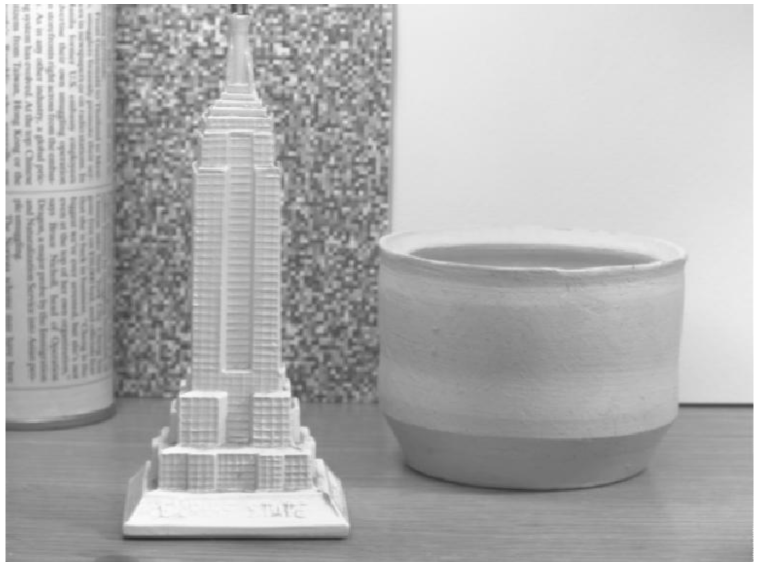
Correspondences can be computed based on closest distance between intersection of 3D lines (from camera and laser)

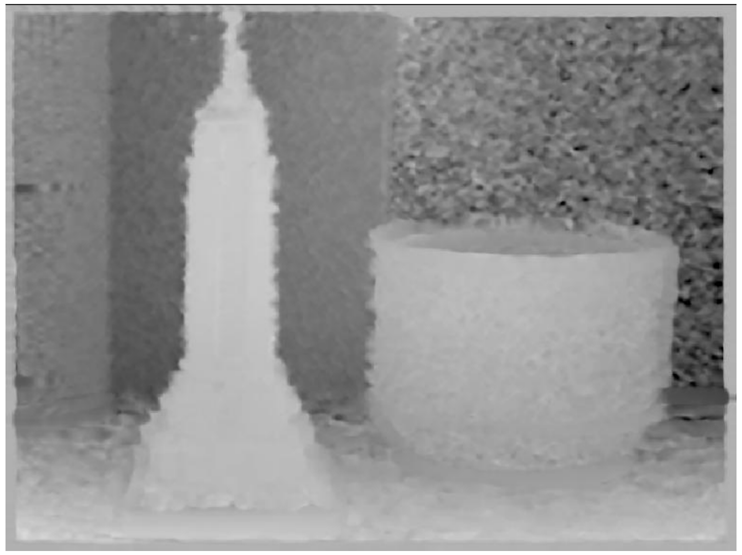
Depth From Focus

- What is in focus depends on the depth
- For a given lens there is a nominal depth where everything is in focus, otherwise object will be out of focus

[Slide after John MacCormick]

Depth From Focus





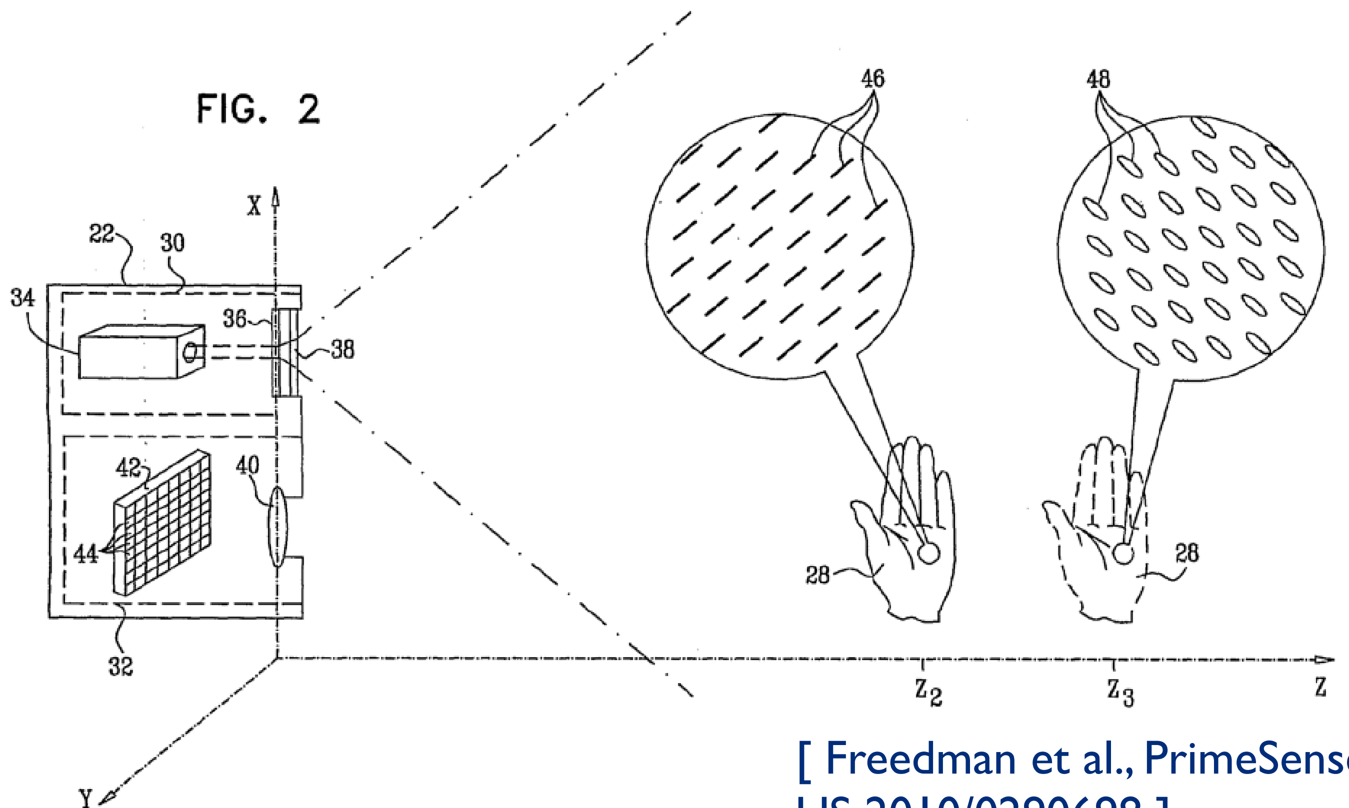
[Watanabe and Nayar, 1998]

- What is in focus depends on the depth
- For a given lens there is a nominal depth where everything is in focus, otherwise object will be out of focus

[Slide after John MacCormick]

Depth From Focus

FIG. 2



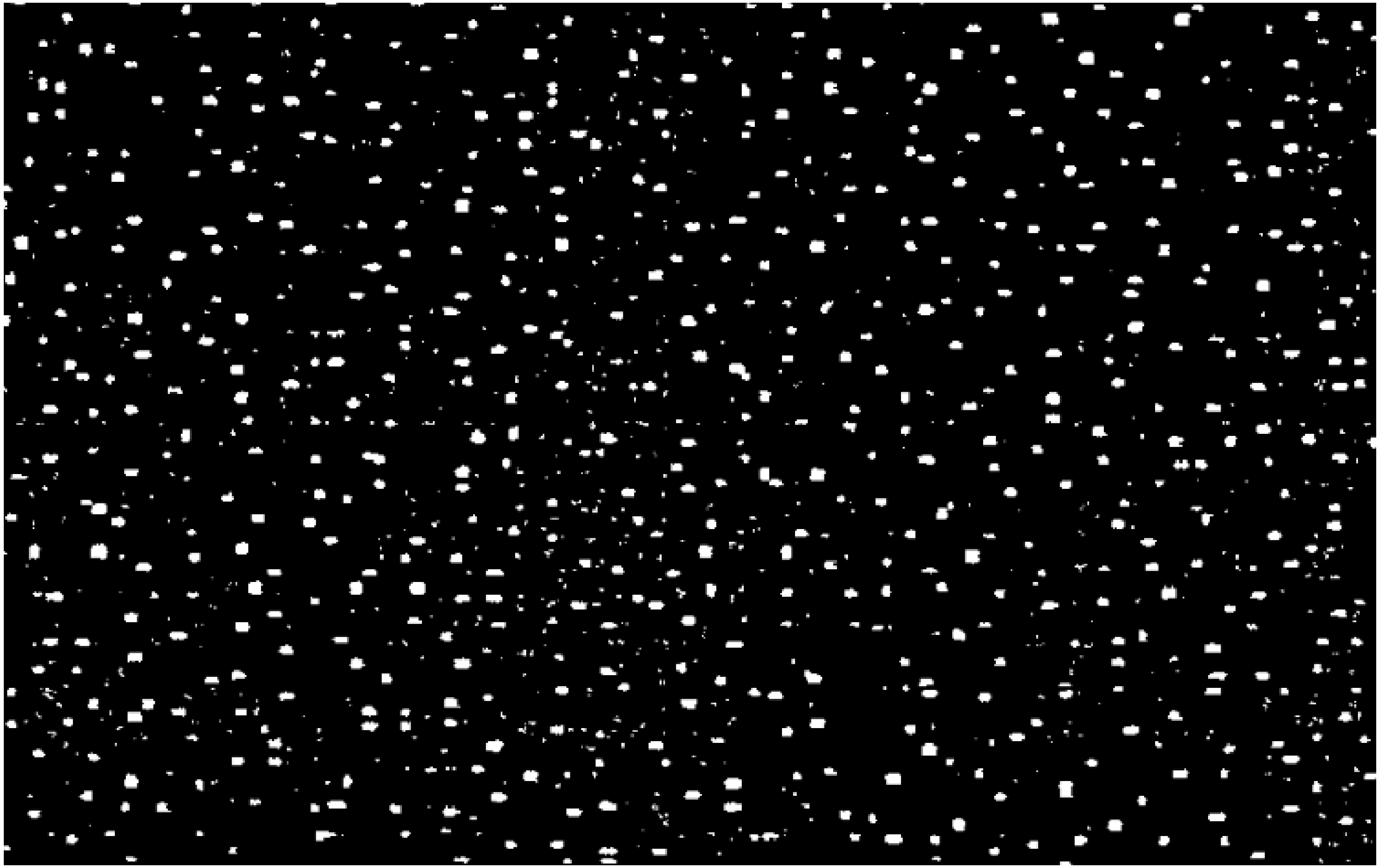
[Freedman et al., PrimeSense patent application US 2010/0290698]

- Kinect improves the accuracy of traditional depth from focus
- Kinect uses a spatial "astigmatic" lens with different focal length in x and y dimensions
- A projected circle becomes an ellipse whose orientation depends on depth

[Slide after John MacCormick]

Depth From Focus

Speckles are the “circles”



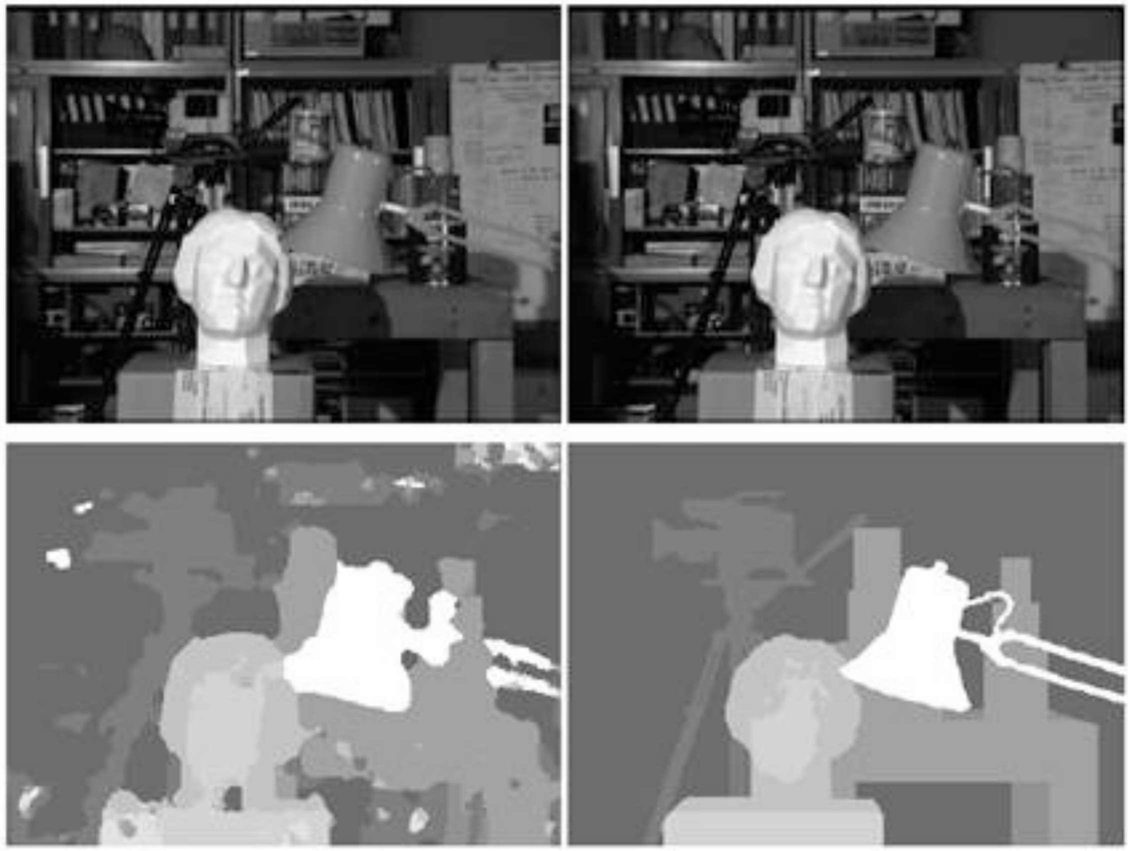
[Shpunt et al., PrimeSense patent application US 2008/0106746]

[Slide after John MacCormick]

Depth From Stereo

- Looking at the scene from 2 different angles, pixels that correspond to closer objects move more than pixels that correspond to further objects

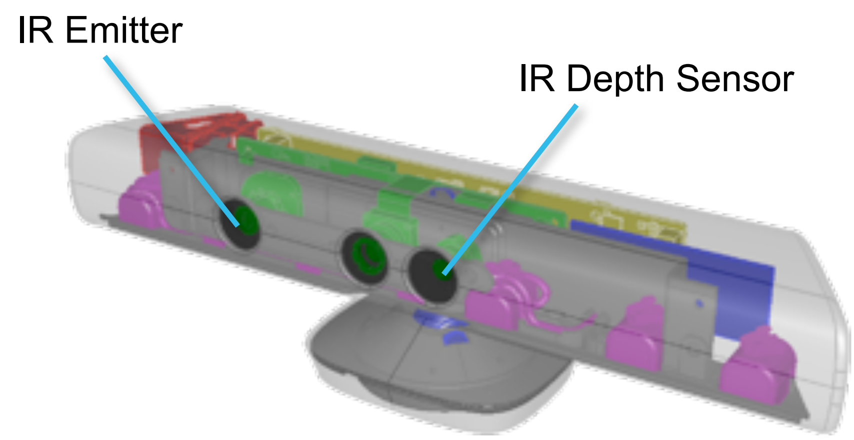
This is how many depth cameras work



[M. Domínguez-Morales, A. Jiménez-Fernández, R. Paz-Vicente, A. Linares-Barranco, G. Jiménez-Moreno, 2012]

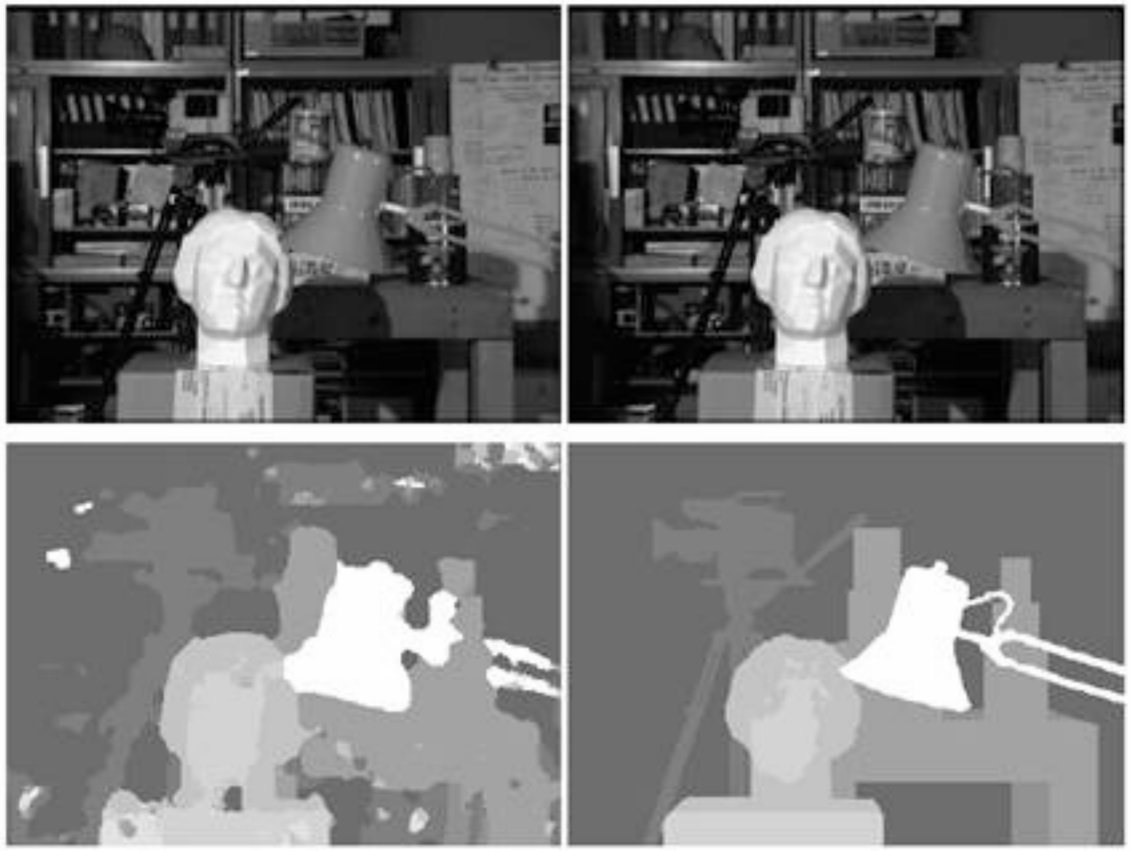
Depth From Stereo

- Looking at the scene from 2 different angles, pixels that correspond to closer objects move more than pixels that correspond to further objects



This is how many depth cameras work

- Kinect analyzes shift of the speckle pattern by projecting from one location and observing from another



[M. Domínguez-Morales, A. Jiménez-Fernández, R. Paz-Vicente, A. Linares-Barranco, G. Jiménez-Moreno, 2012]

Kinect

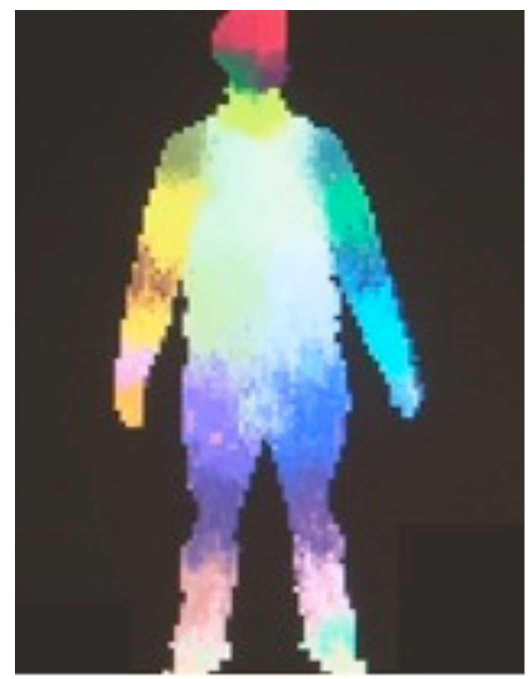
KINECT

Two key contributions:

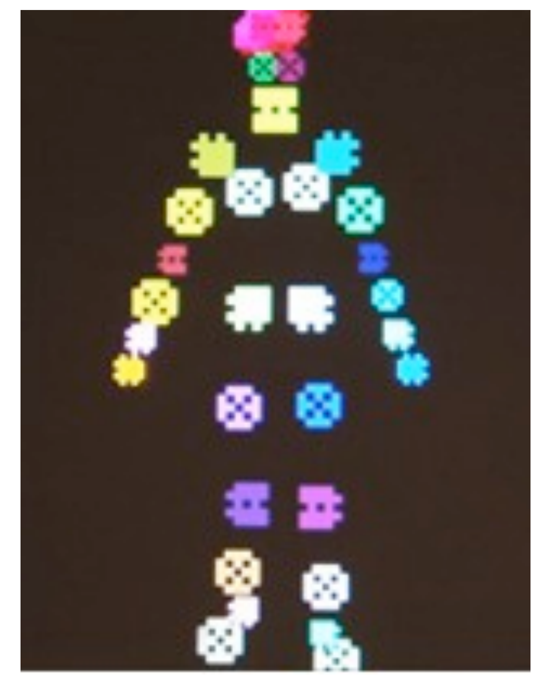
- Inexpensive and accurate depth camera / sensor
- 3D Pose estimation

Color Image

Depth Image



Body Part Segmentation



3D Joint Estimation

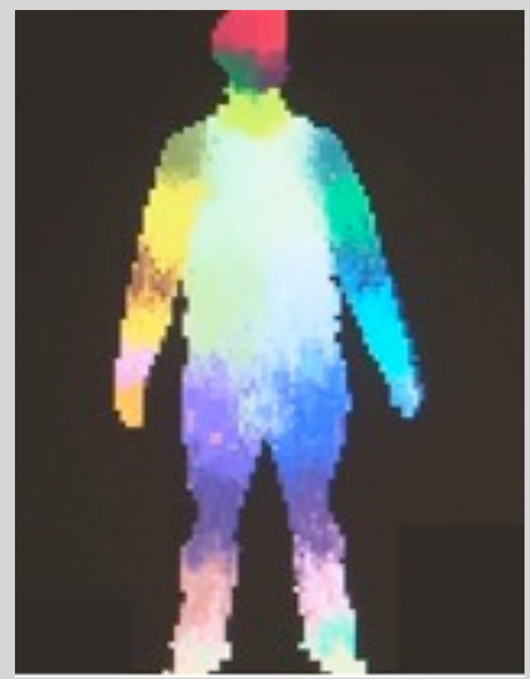
Kinect

Two key contributions

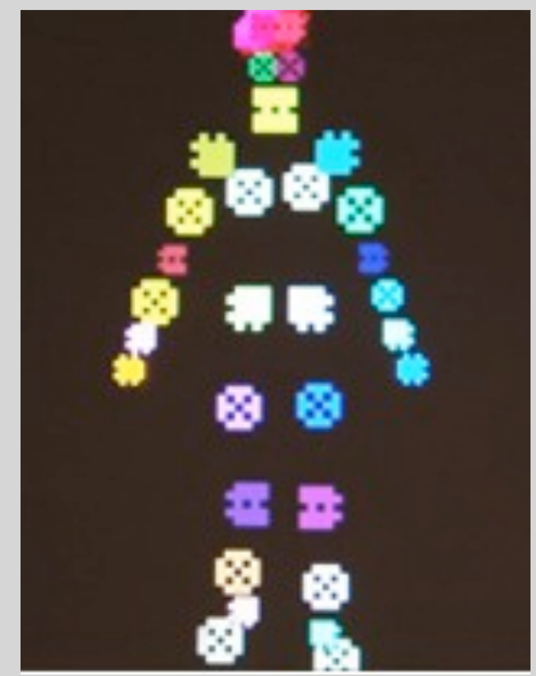
- Inexpensive and accurate depth camera / sensor
- 3D Pose estimation

Color Image

Depth Image



Body Part Segmentation



3D Joint Estimation

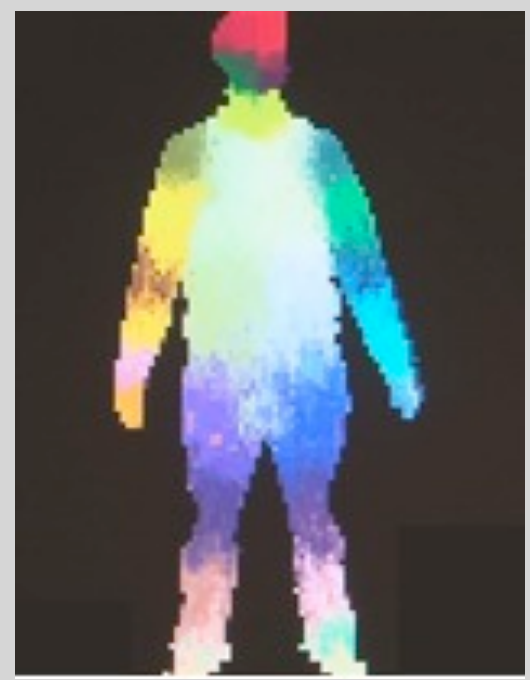
Kinect

Two key contributions

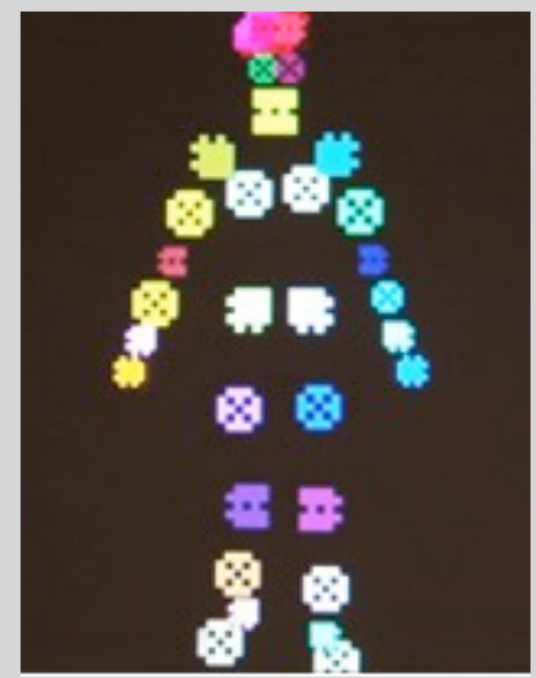
- Inexpensive and accurate depth camera / sensor
- 3D Pose estimation

Color Image

Depth Image



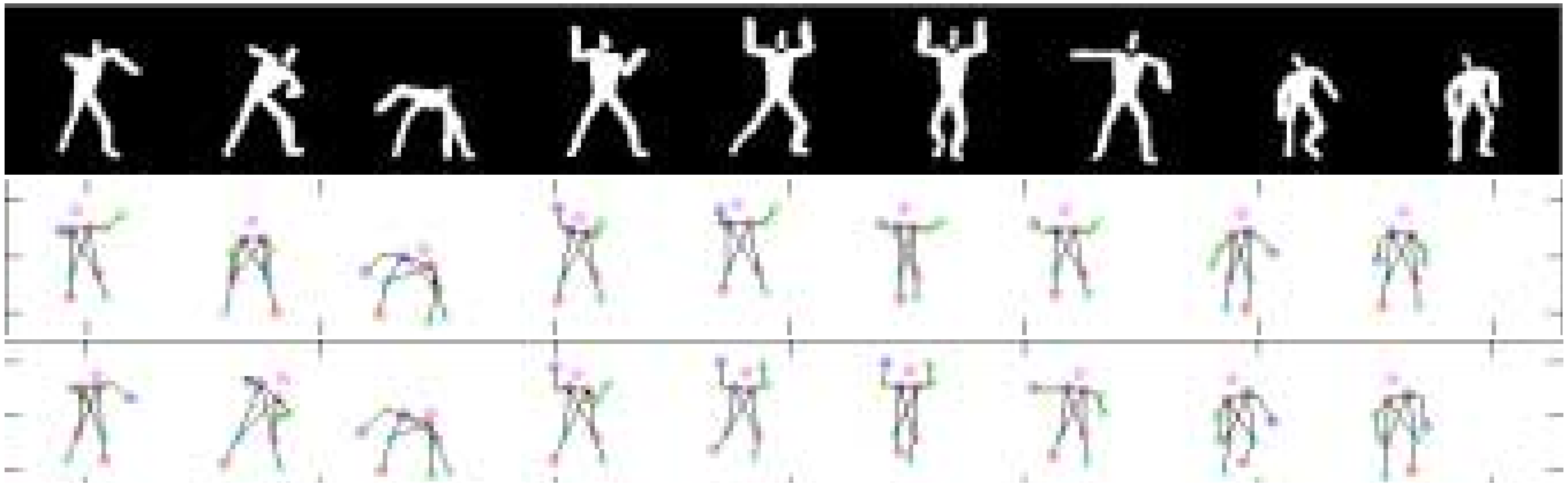
Body Part Segmentation



3D Joint Estimation

Again, correspondences are difficult

3D Pose Estimation
(without correspondences)



[Inferring body pose without tracking body parts, Rosales & Sclaroff, CVPR, 2000]

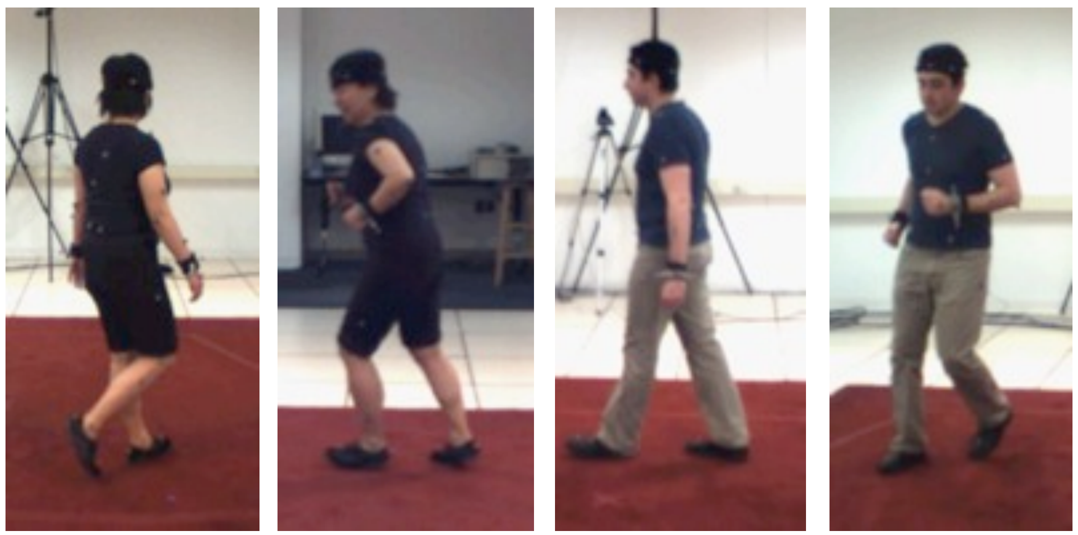
- Generate synthetic examples of "images" from MoCap
- This creates a database of image-pose pairs
- Learn a function that takes image features as input and outputs 3D pose

Simplest Regression-based Method

[Shakhnarovich, Viola, Darrell, ICCV'03]

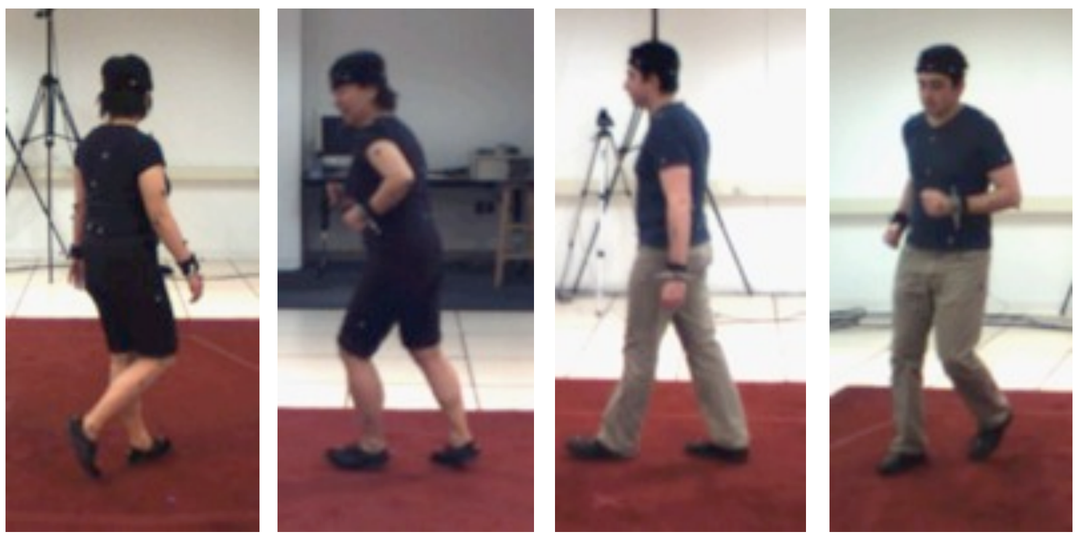
Simplest Regression-based Method

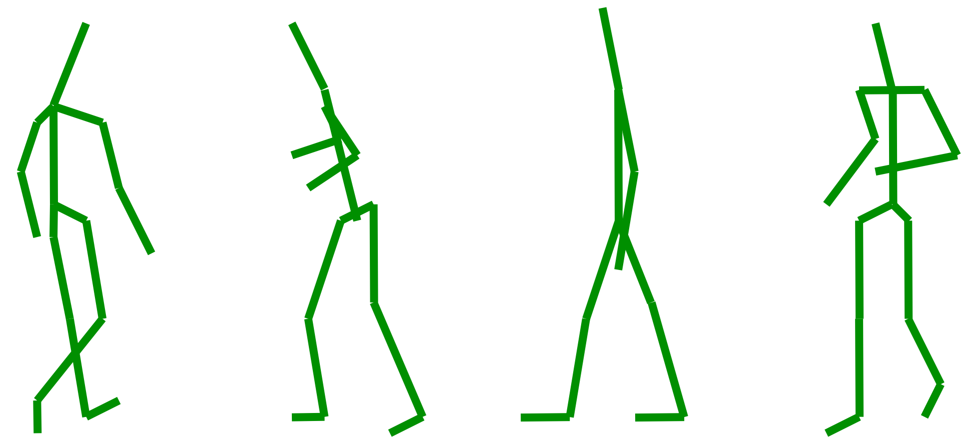
[Shakhnarovich, Viola, Darrell, ICCV'03]



Simplest Regression-based Method

[Shakhnarovich, Viola, Darrell, ICCV'03]

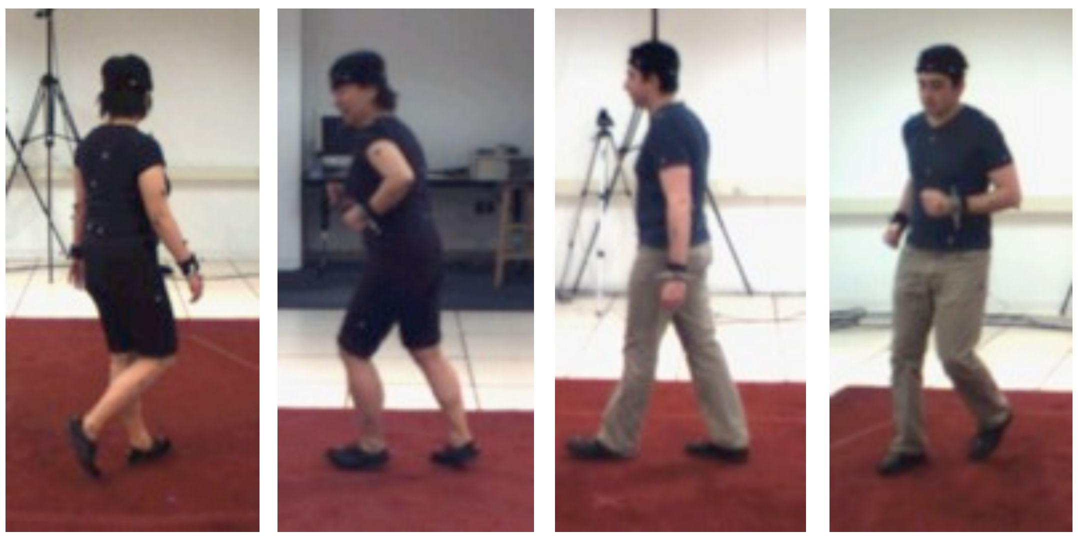


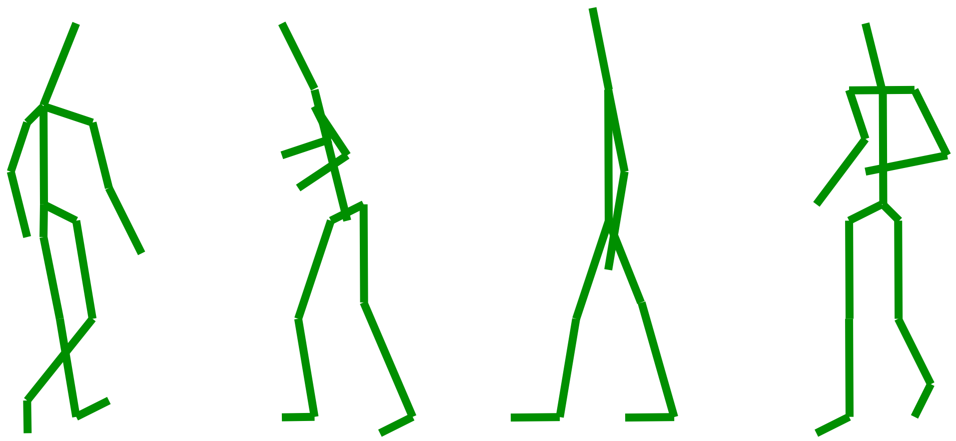


Simplest Regression-based Method

[Shakhnarovich, Viola, Darrell, ICCV'03]

Given: large database of image-pose pairs

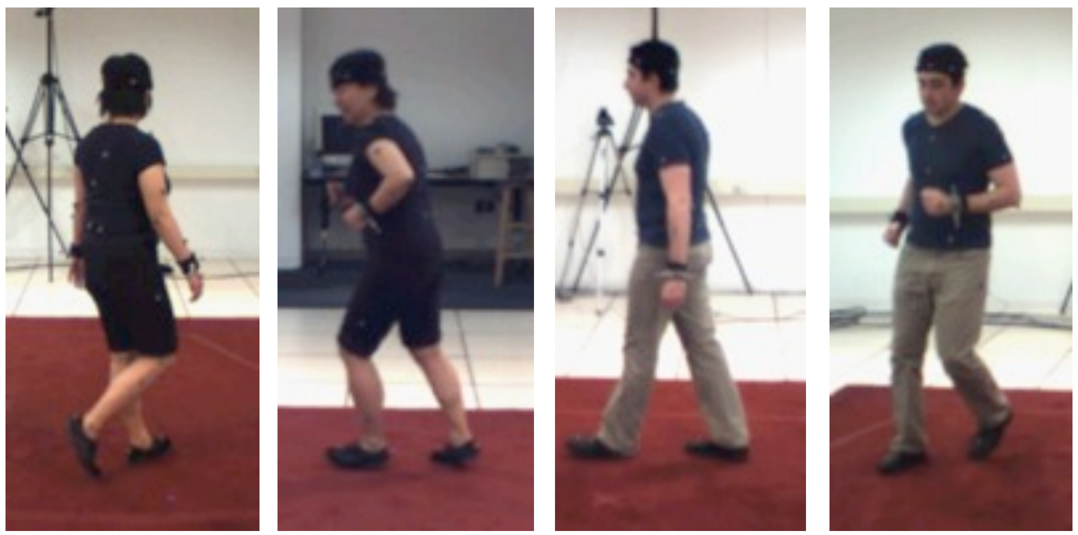


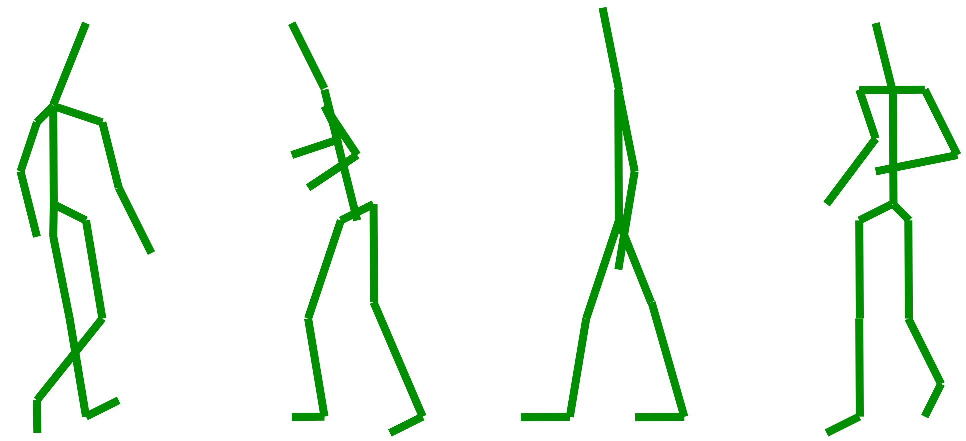


Simplest Regression-based Method

[Shakhnarovich, Viola, Darrell, ICCV'03]

Given: large database of image-pose pairs



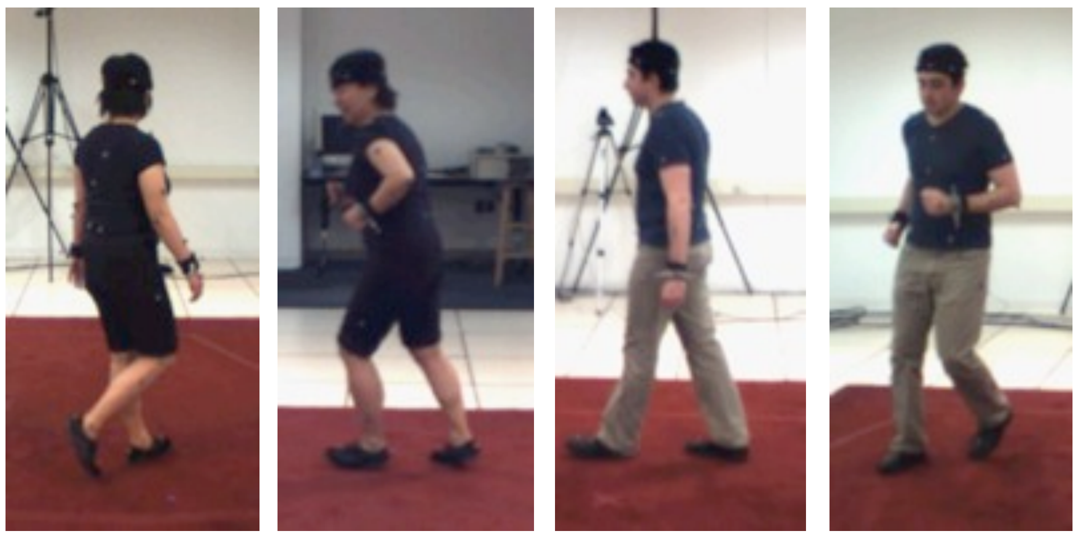


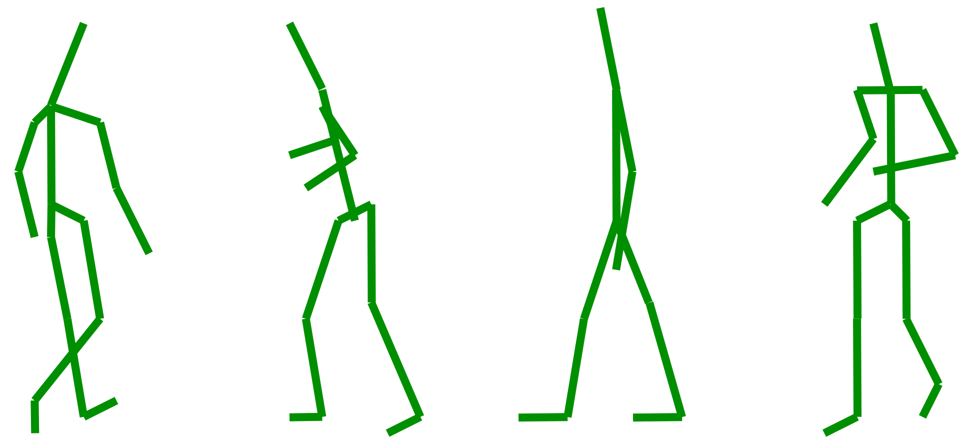
Pose inference: trivial using a NN approach

Simplest Regression-based Method

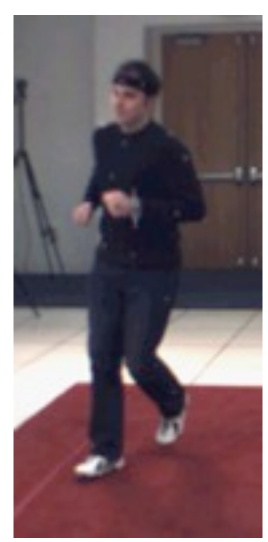
[Shakhnarovich, Viola, Darrell, ICCV'03]

Given: large database of image-pose pairs





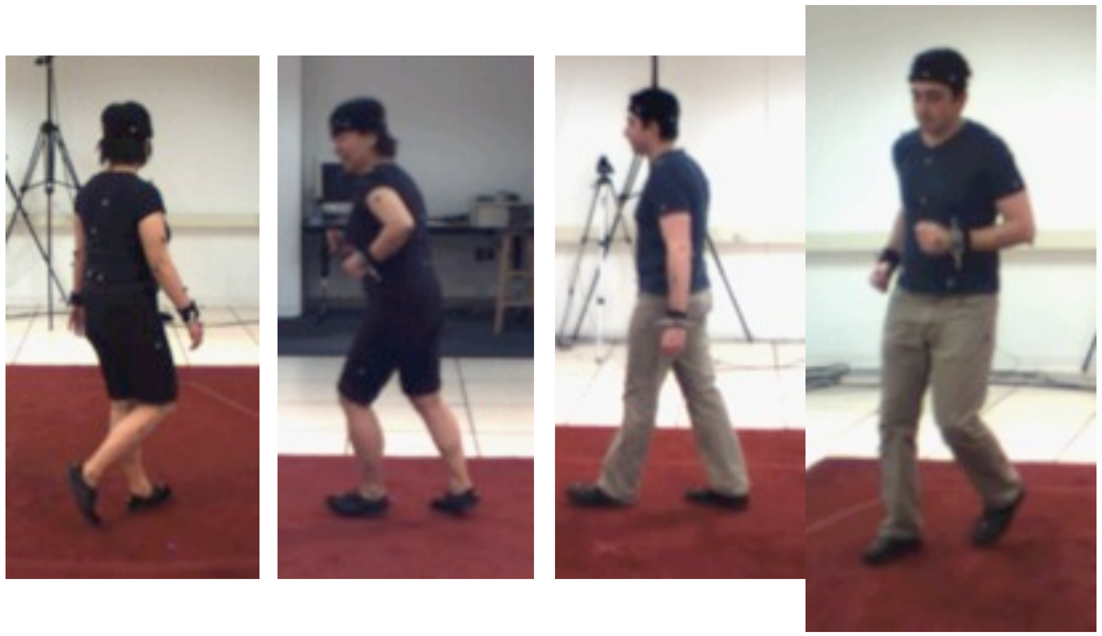
Pose inference: trivial using a NN approach

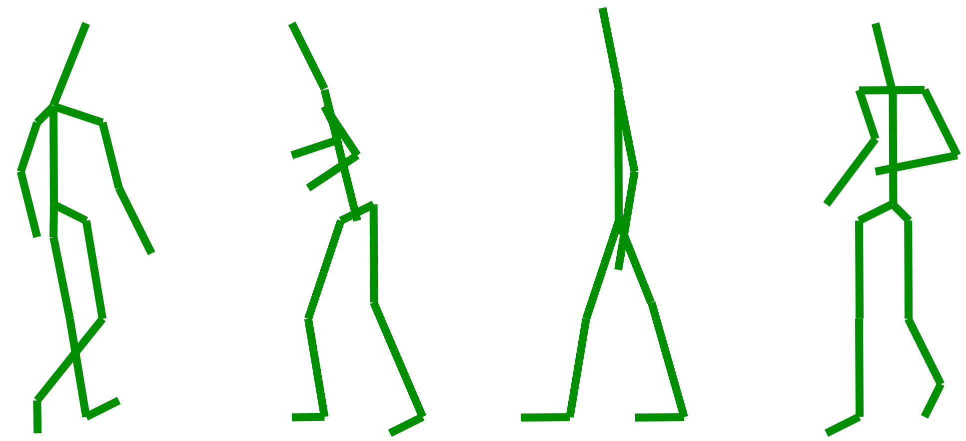


Simplest Regression-based Method

[Shakhnarovich, Viola, Darrell, ICCV'03]

Given: large database of image-pose pairs



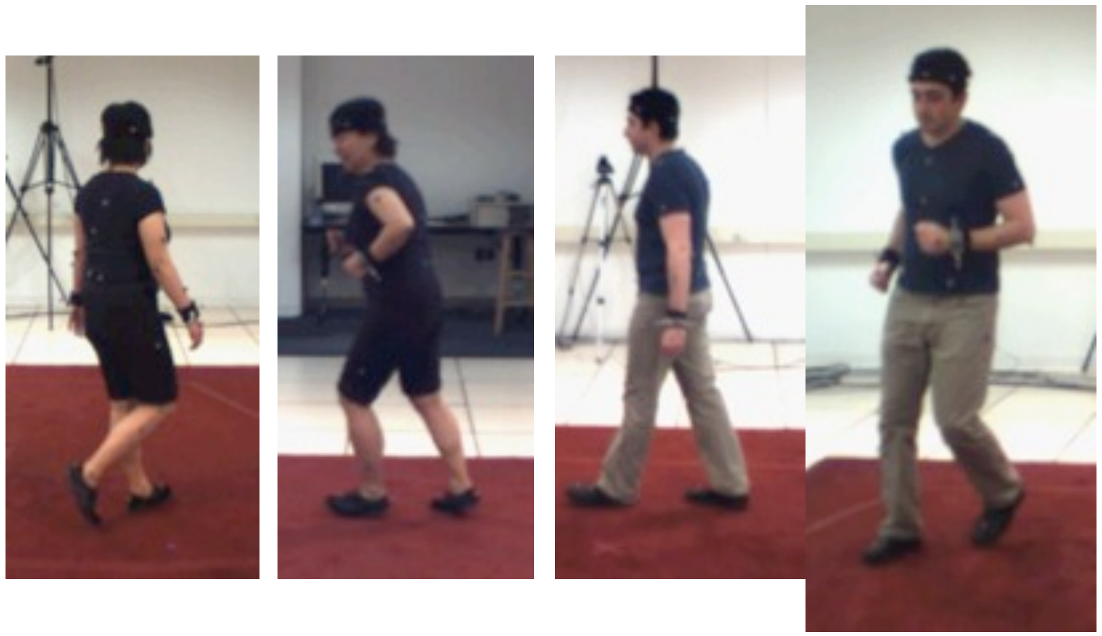


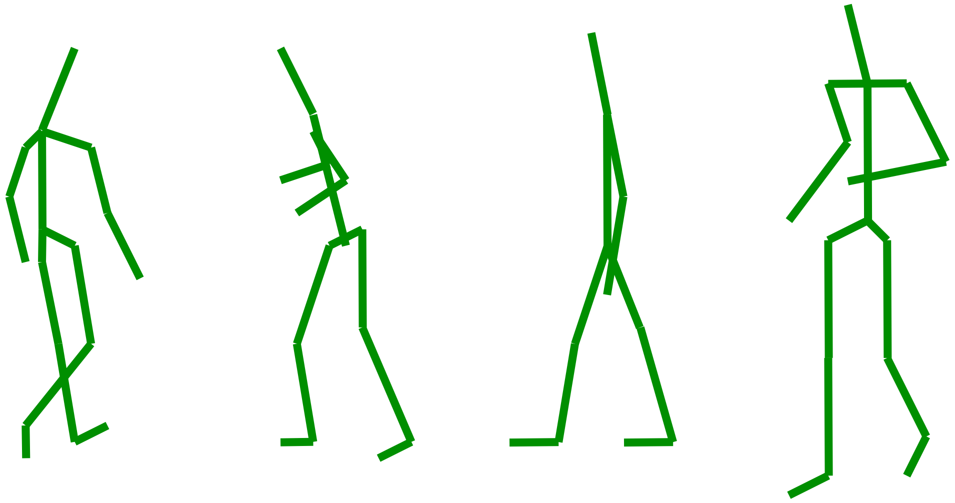
Pose inference: trivial using a NN approach

Simplest Regression-based Method

[Shakhnarovich, Viola, Darrell, ICCV'03]

Given: large database of image-pose pairs



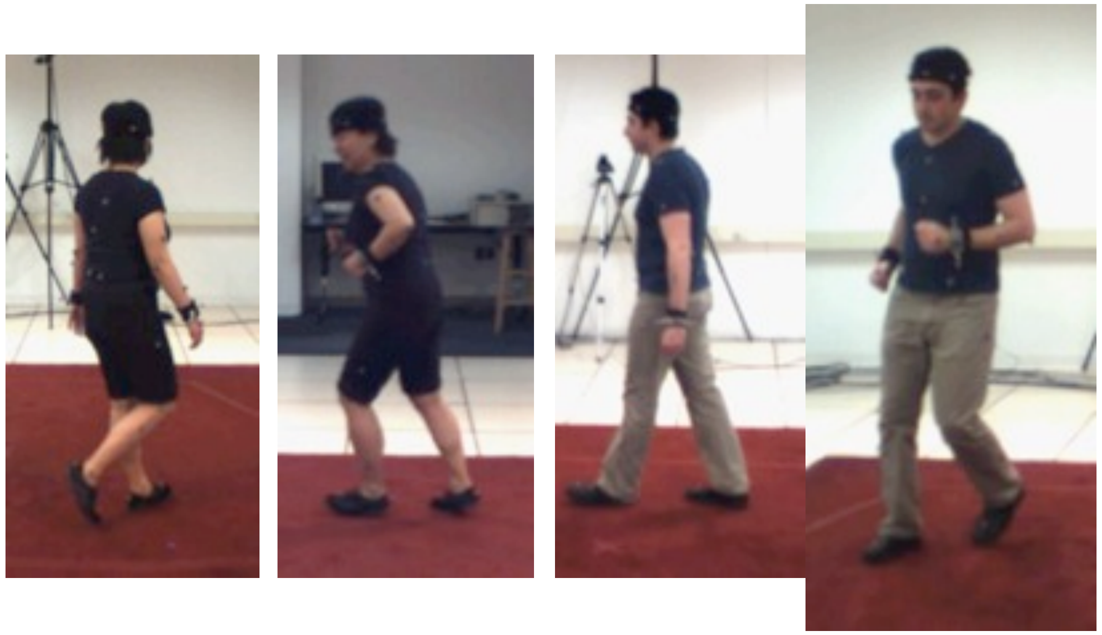


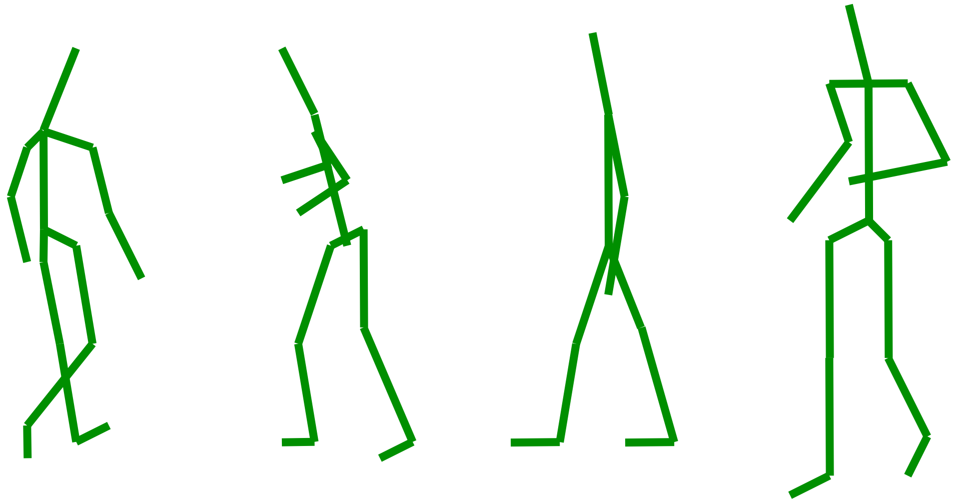
Pose inference: trivial using a NN approach

Simplest Regression-based Method

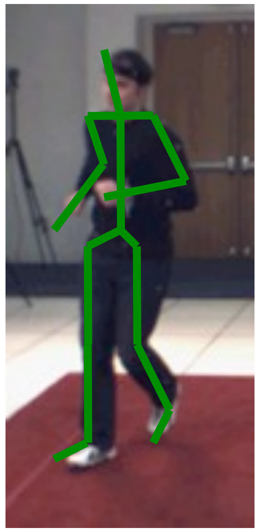
[Shakhnarovich, Viola, Darrell, ICCV'03]

Given: large database of image-pose pairs





Pose inference: trivial using a NN approach



Nearest Neighbor Regression

[Shakhnarovich, Viola, Darrell, ICCV'03]

- Speeding up the NN lookup using hashing functions

Nearest Neighbor Regression

[Shakhnarovich, Viola, Darrell, ICCV'03]

- Speeding up the NN lookup using hashing functions
- Better results are obtained by wighted average of k- Nearest Neighbors

Nearest Neighbor Regression

[Shakhnarovich, Viola, Darrell, ICCV'03]

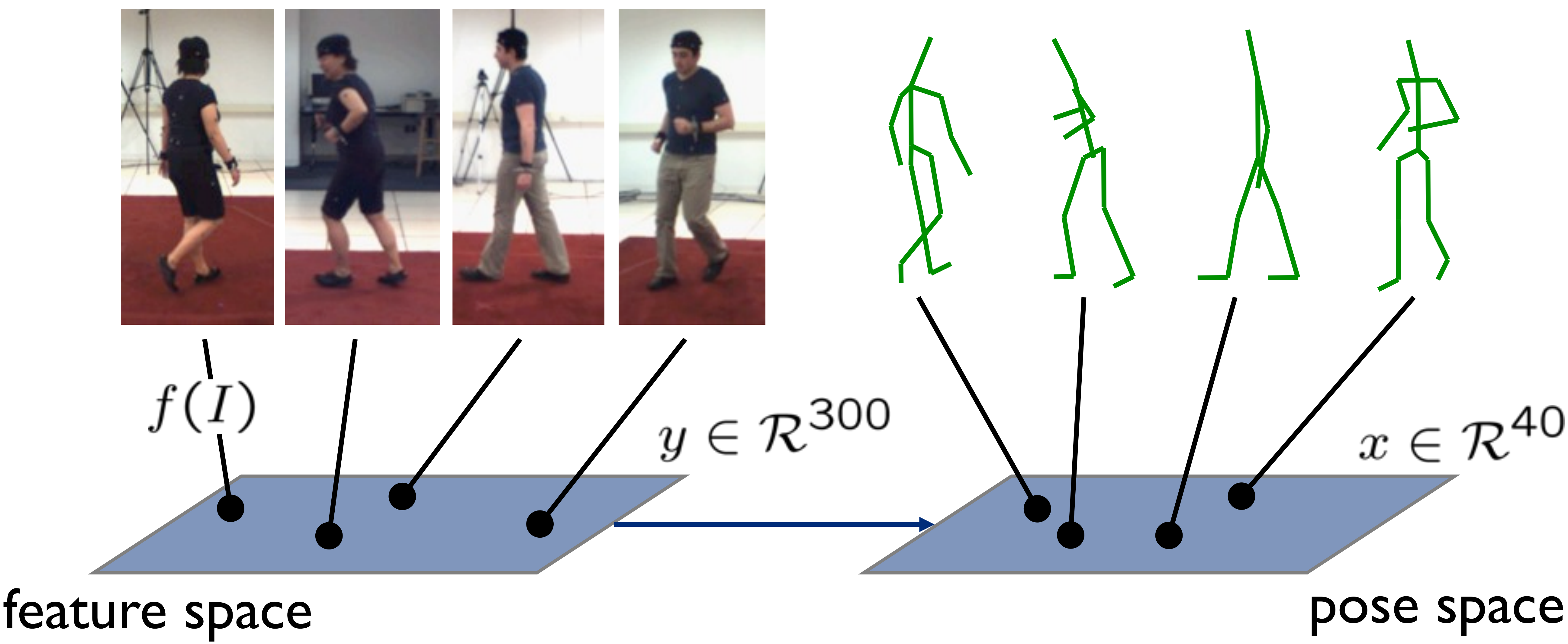
- Speeding up the NN lookup using hashing functions
- Better results are obtained by wighted average of k- Nearest Neighbors

Linear Regression

[Agarwal, Triggs, CVPR'04]

Learn a functional mapping from features to pose

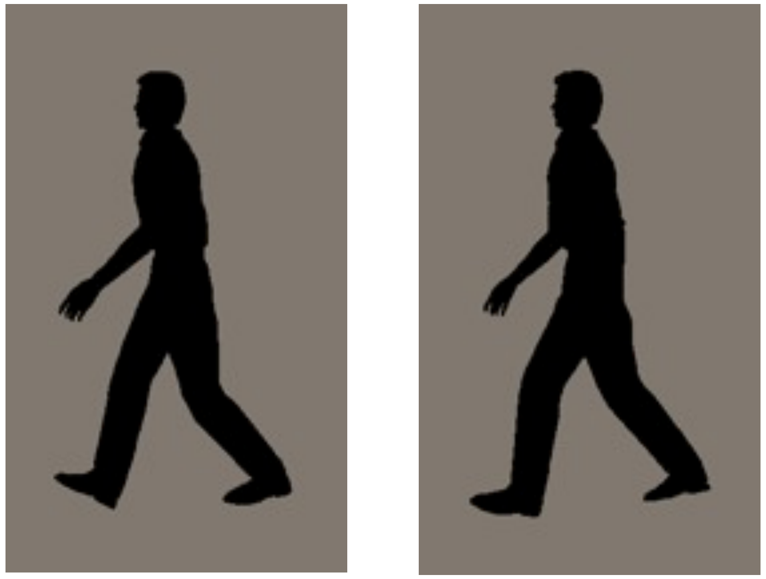
(e.g. Linear Regression: $x = g(y) = \mathbf{A}y + b$)



pose = g (features)

Imaging Ambiguities

Imaging Ambiguities



[Agarwal and Triggs, CVPR'05]

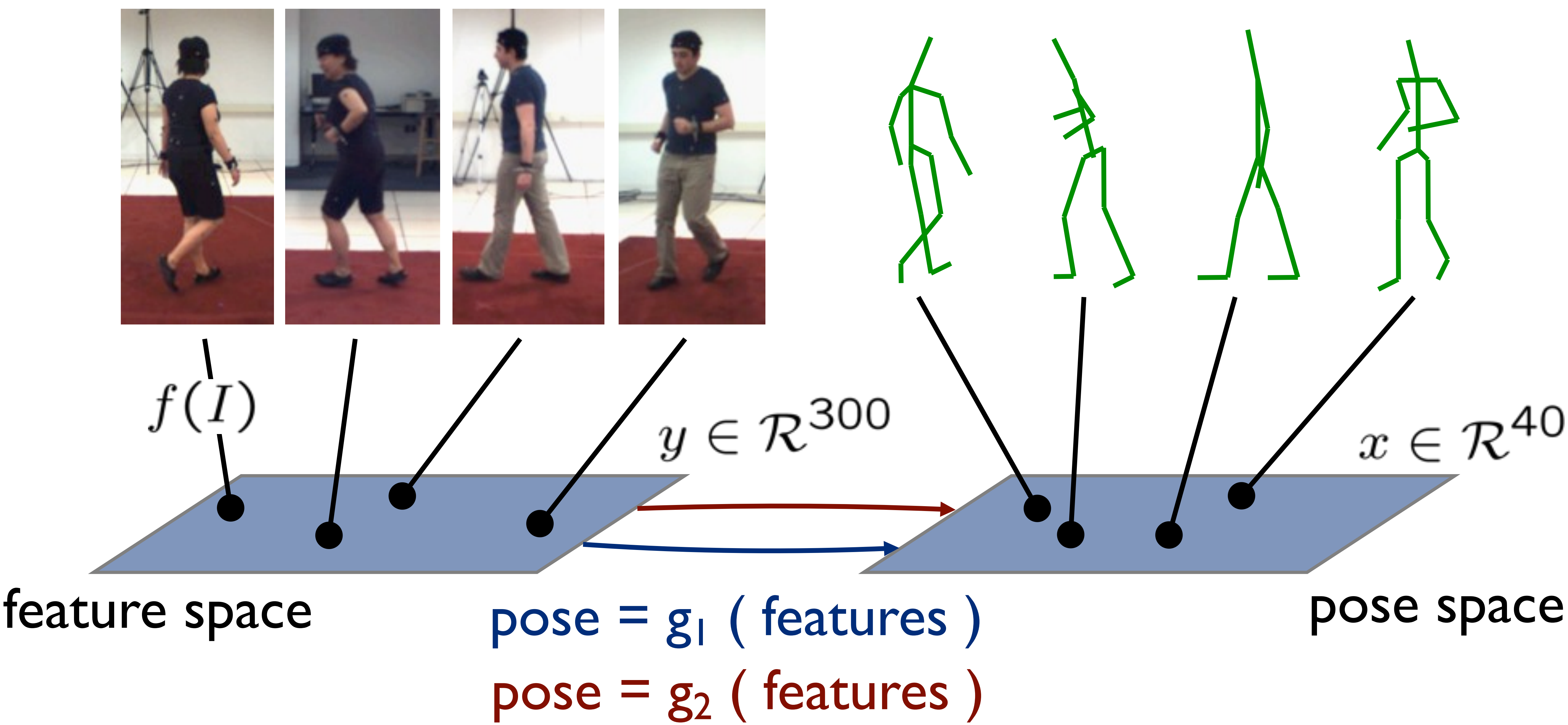
Imaging Ambiguities

[Agarwal and Triggs, CVPR'05]

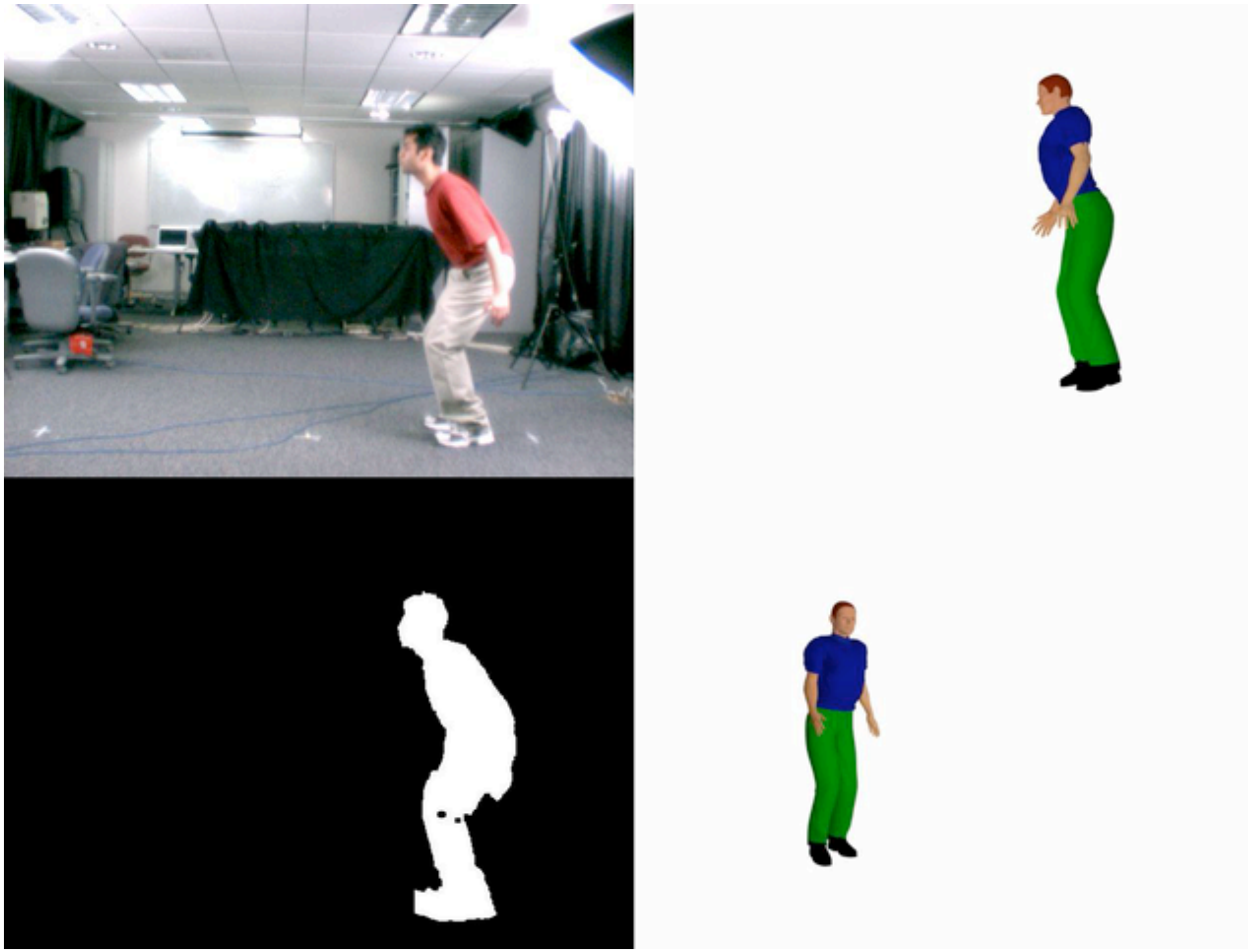
Mixture of Experts

[Sminchisescu et al PAMI'07, Bo et al CVPR'08]

Muti-modal probabilistic functions

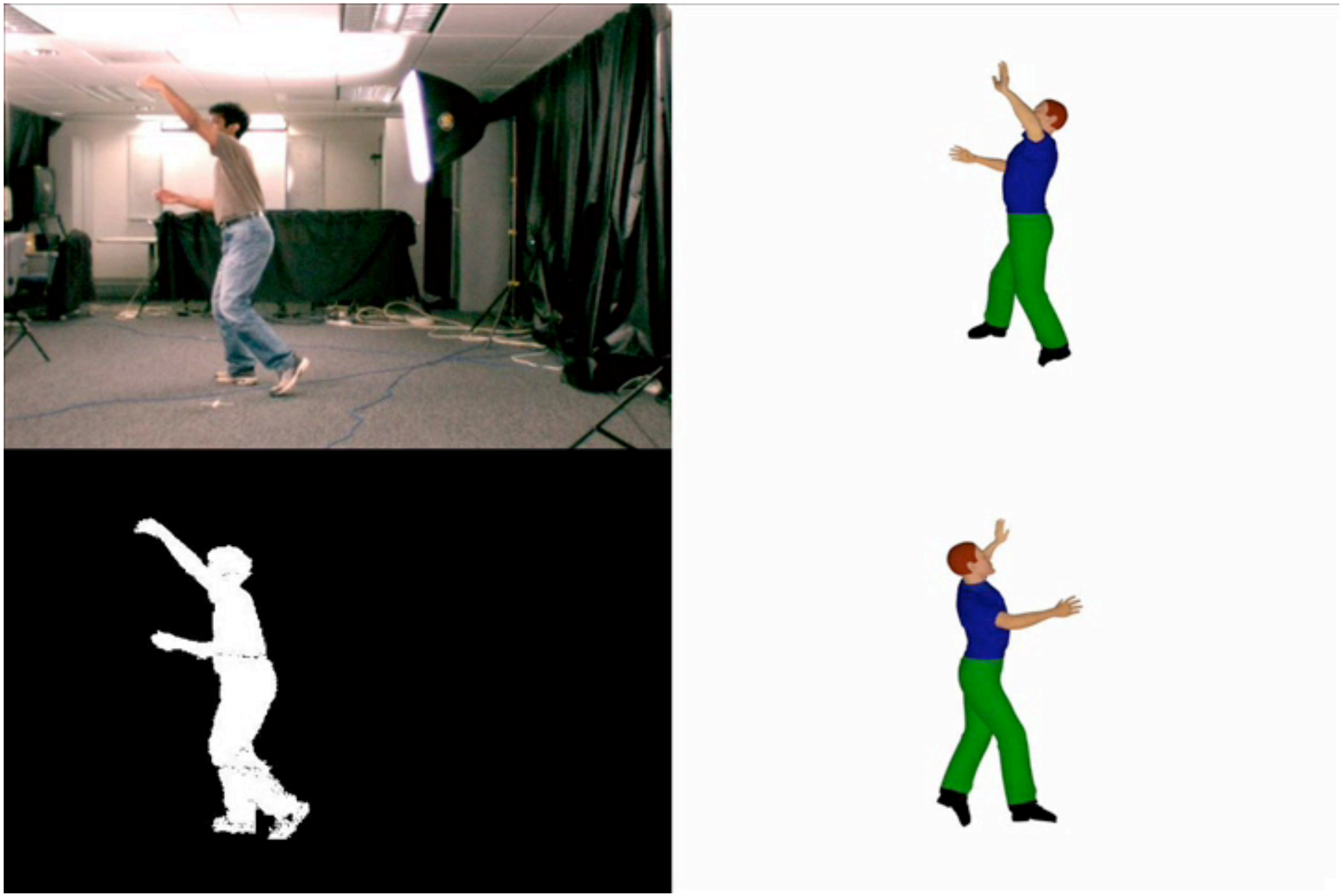


Mixture of Experts



[Sminchisescu et al PAMI'07, Bo et al CVPR'08]

Mixture of Experts



[Sminchisescu et al PAMI'07, Bo et al CVPR'08]

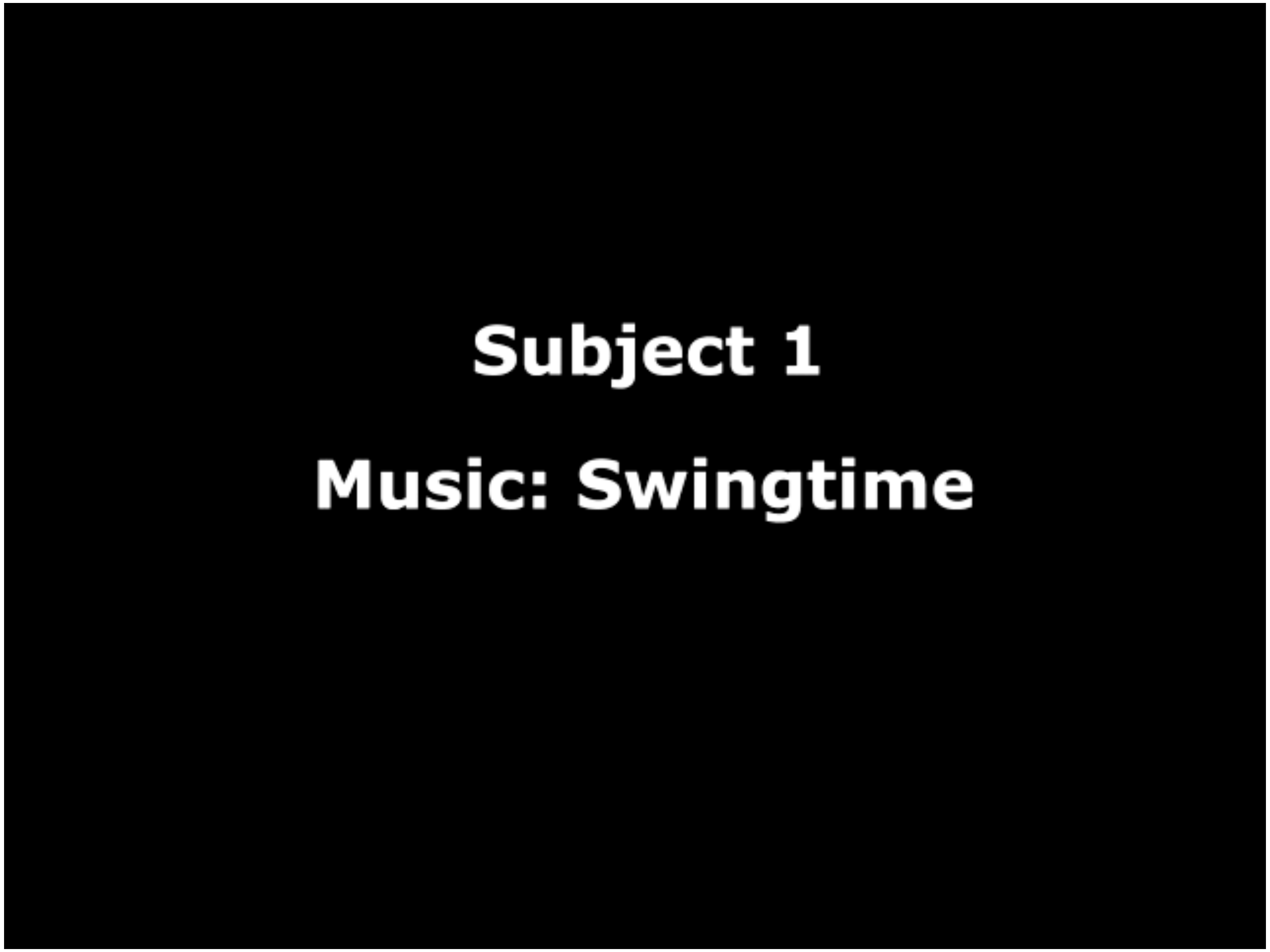
An Interesting Application

[Ren, Shakhnarovich, Hodgins, Pfister, Viola, ACM SIGGRAPH, 2004]

In this case multiple (3) cameras are used, but using similar regression-based (correspondence free) approach

An Interesting Application

[Ren, Shakhnarovich, Hodgins, Pfister, Viola, ACM SIGGRAPH, 2004]



In this case multiple (3) cameras are used, but using similar regression-based (correspondence free) approach

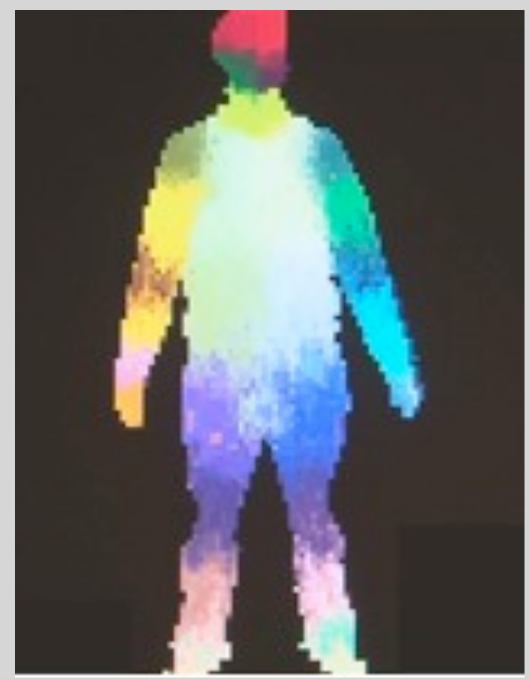
Kinect

Depth image:

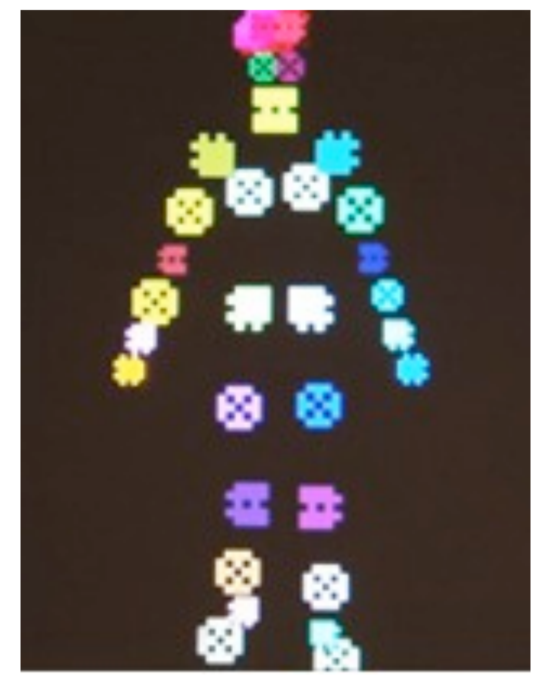
- Resolves ambiguities in pose
- Make it easy to segment person from background

Color Image

Depth Image

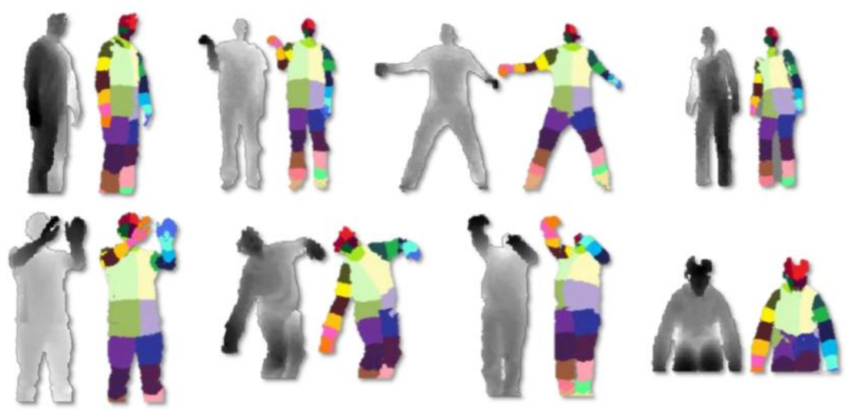


Body Part Segmentation



3D Joint Estimation

Step 1: Create a synthetic dataset



- 15 different body types
- About 100,000 poses
- Render depth image-pose pairs

[Shotton, Fitzgibbon, Cook, Sharp, Finocchio, Moore, Kipman, Blake, CVPR'11]

Step 2: Learn mapping to body parts

Train a randomized decision forests

[Slide after John MacCormick]

Step 2: Learn mapping to body parts

Train a randomized decision forests

It’s like a sophisticated game of 20 questions
(decision tree)

[Slide after John MacCormick]

Step 2: Learn mapping to body parts

Train a randomized decision forests

It’s like a sophisticated game of 20 questions
(decision tree)

Should you take an umbrella?

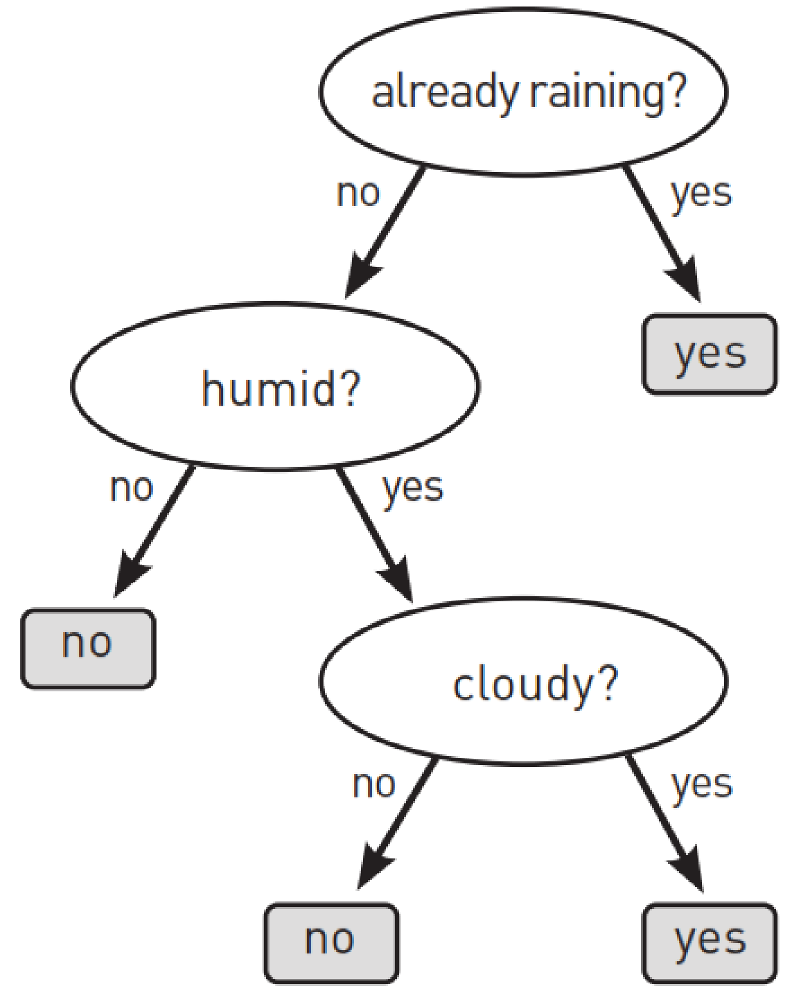
[Slide after John MacCormick]

Step 2: Learn mapping to body parts

Train a randomized decision forests

It's like a sophisticated game of 20 questions
(decision tree)

Should you take an umbrella?

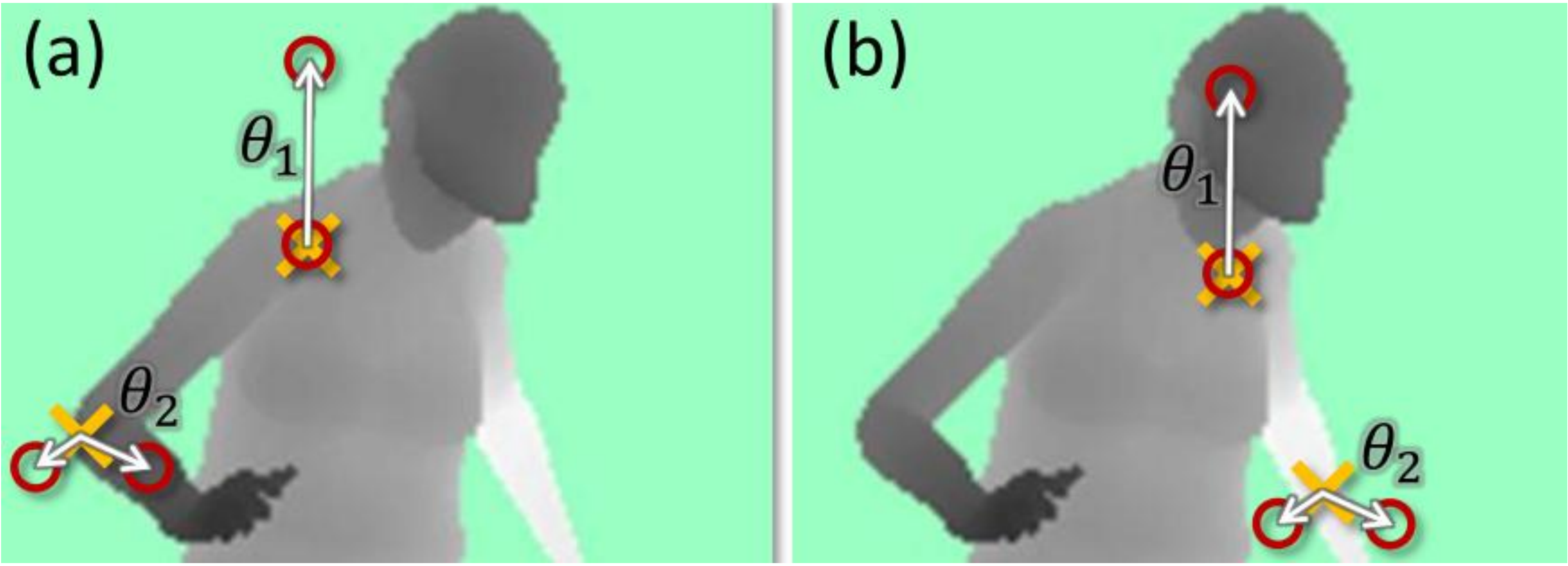


[Slide after John MacCormick]

Questions Kinect Asks

How does the (normalized) depth at the given pixel compares to the (normalized) depth at a pixel with a given offset

[Shotton, Fitzgibbon, Cook, Sharp, Finocchio, Moore, Kipman, Blake, CVPR'11]



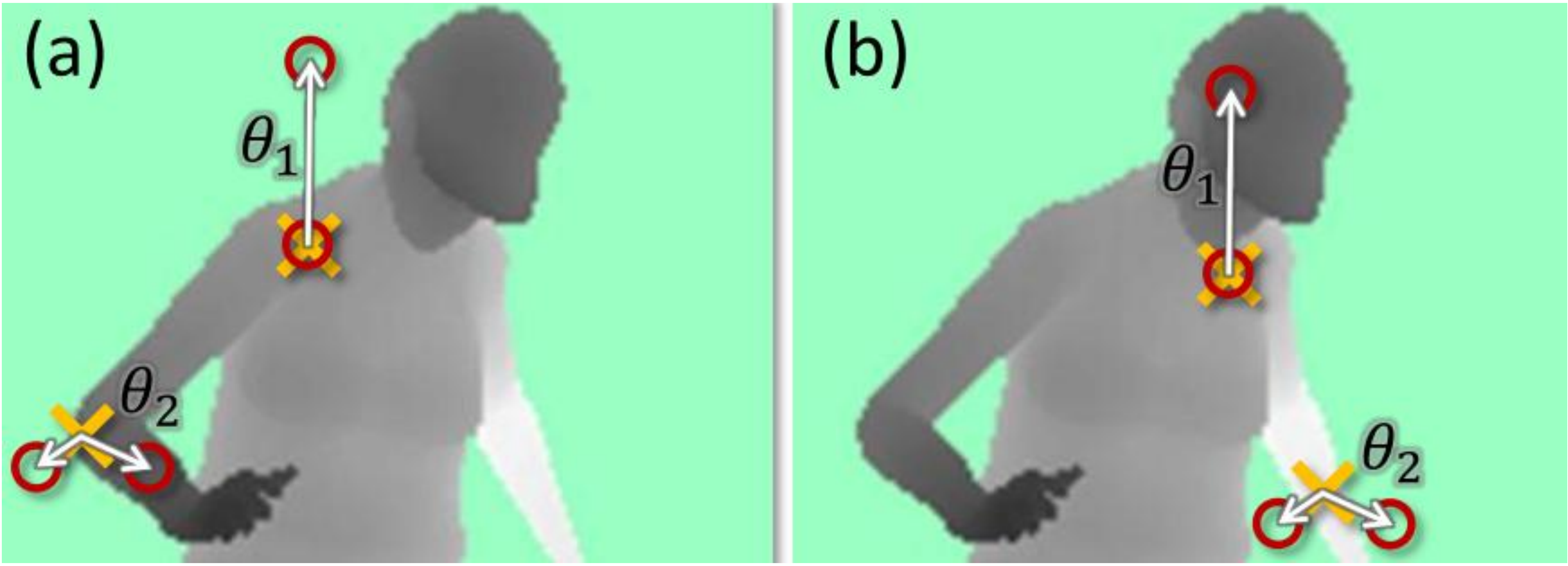
[Slide after John MacCormick]

Questions Kinect Asks

How does the (normalized) depth at the given pixel compares to the (normalized) depth at a pixel with a given offset

Note: this is only a form of the question, there are millions of these types of questions that can be asked depending on the parameters (e.g., offset, comparisons)

[Shotton, Fitzgibbon, Cook, Sharp, Finocchio, Moore, Kipman, Blake, CVPR'11]



[Slide after John MacCormick]

Learning a Decision Tree

- Need to choose a sequence of questions to ask
- Which question is most useful to ask next?

[Slide after John MacCormick]

Learning a Decision Tree

- Need to choose a sequence of questions to ask
- Which question is most useful to ask next?

e.g. for taking an umbrella is it more useful to ask "is it raining?" or "is it cloudy?"

[Slide after John MacCormick]

Learning a Decision Tree

- Need to choose a sequence of questions to ask
- Which question is most useful to ask next?

e.g. for taking an umbrella is it more useful to ask "is it raining?" or "is it cloudy?" Why?

[Slide after John MacCormick]

Learning a Decision Tree

- Need to choose a sequence of questions to ask
- Which question is most useful to ask next?

e.g. for taking an umbrella is it more useful to ask "is it raining?" or "is it cloudy?" Why?

- Mathematically this takes the form of information gain (which is derived from entropy)

(I am not going to go through the details)

[Slide after John MacCormick]

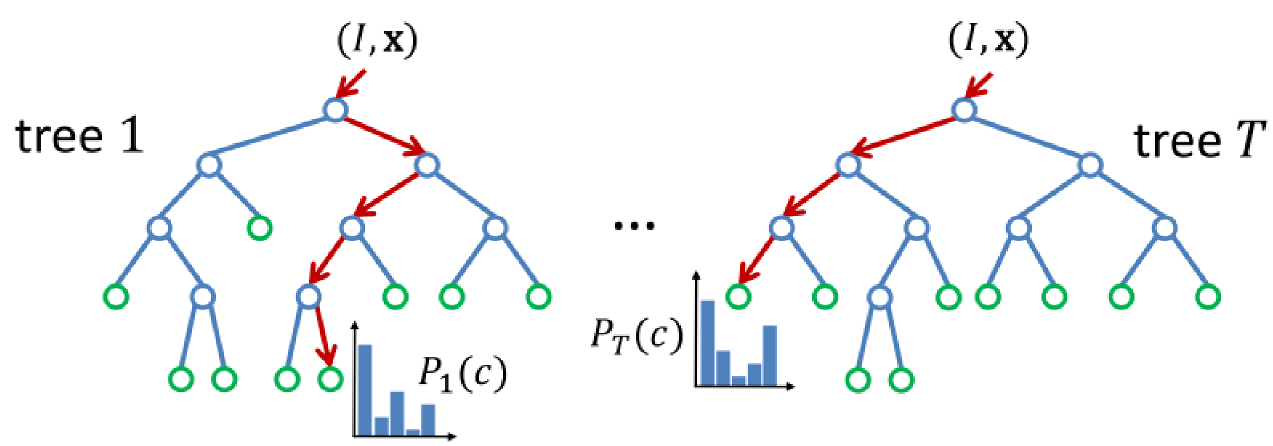
Randomized Decision Forest

Randomized

- Too many possible questions, so use a different random sub-set of 2,000 each time

Forest

- Instead of training one decision tree, train many
- Use results from all to make a decision



[Slide after John MacCormick]

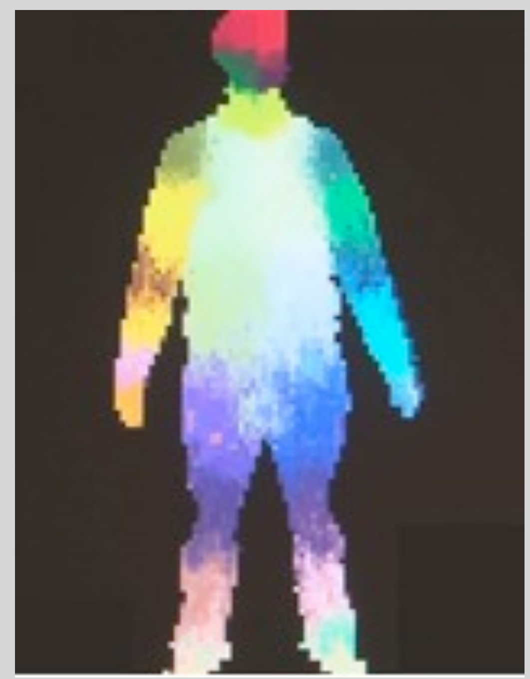
Kinect

Depth image:

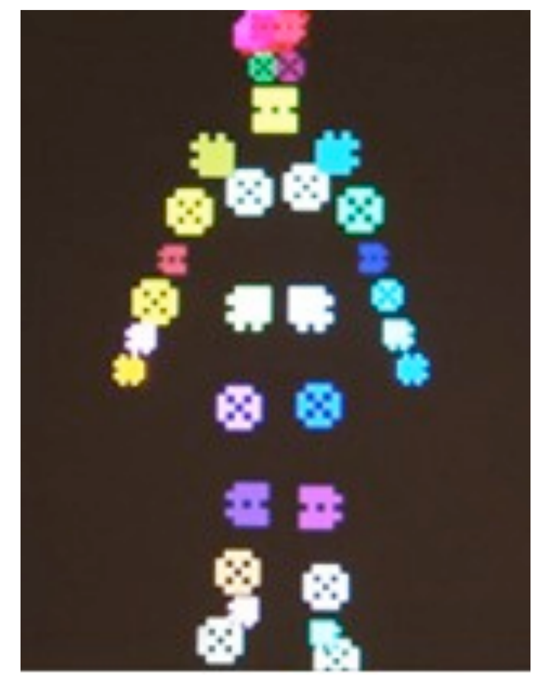
- Resolves ambiguities in pose
- Make it easy to segment person from background

Color Image

Depth Image



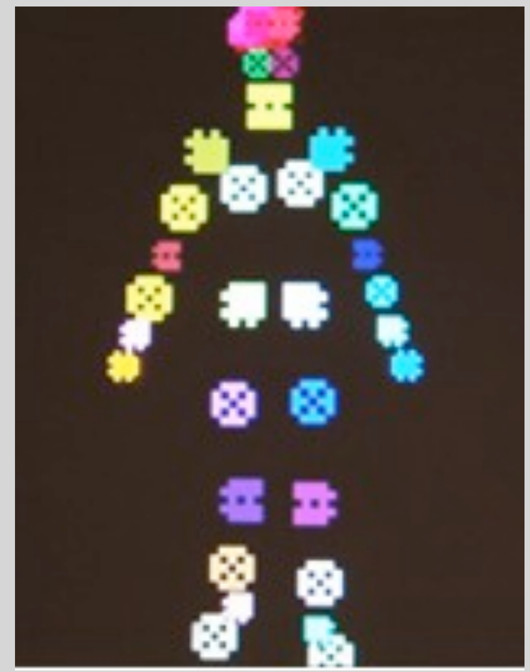
Body Part Segmentation



3D Joint Estimation

Body Parts to Skeleton

Body Part
Segmentation



3D Joint
Estimation

Find centroids of parts

Use robust (and fast) algorithm -- Mean Shift

Kinect

Is body segmentation really needed?

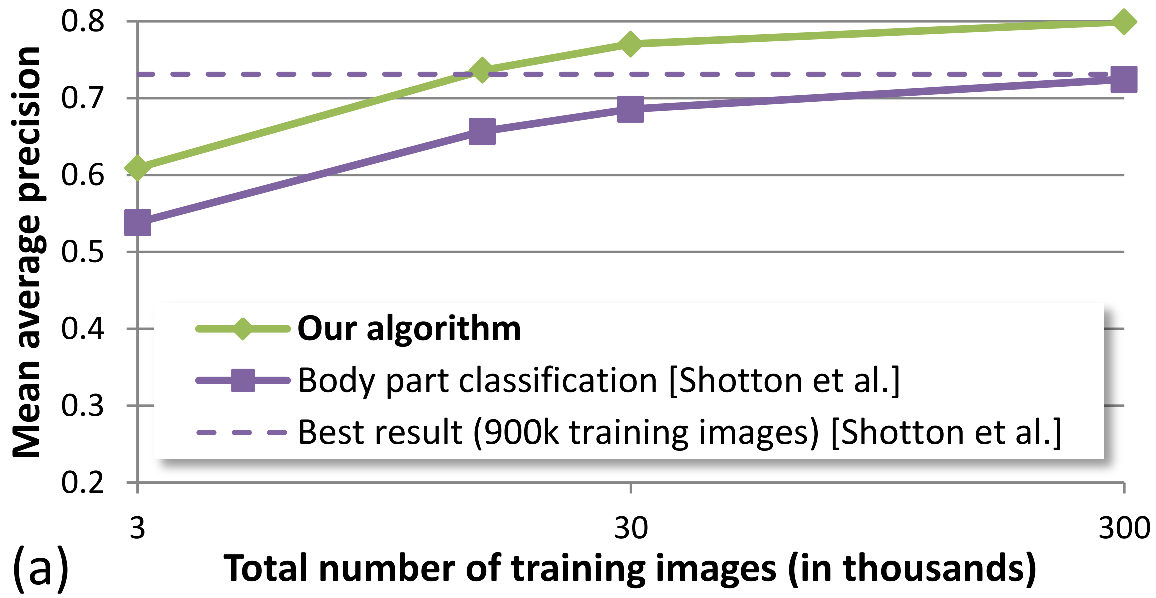
Color Image

Depth Image

Body Part Segmentation

3D Joint Estimation

Kinect



[Girshick, Shotton, Kohli, Criminisi, Fitzgibbon, ICCV, 2011]

You can do better regressing directly to 3D joints

Color Image

Depth Image

3D Joint Estimation

Kinect

Done using similar regression forest as before

You can do better regressing directly to 3D joints

Color Image

Depth Image

3D Joint
Estimation

Nearest Neighbor Regression

[Shakhnarovich, Viola, Darrell, ICCV'03]

In practice, similar to k Nearest Neighbor, except much faster

Nearest Neighbor Regression

[Shakhnarovich, Viola, Darrell, ICCV'03]

In practice, similar to k Nearest Neighbor, except much faster

Discussion

Closed Universe


All of these are input devices

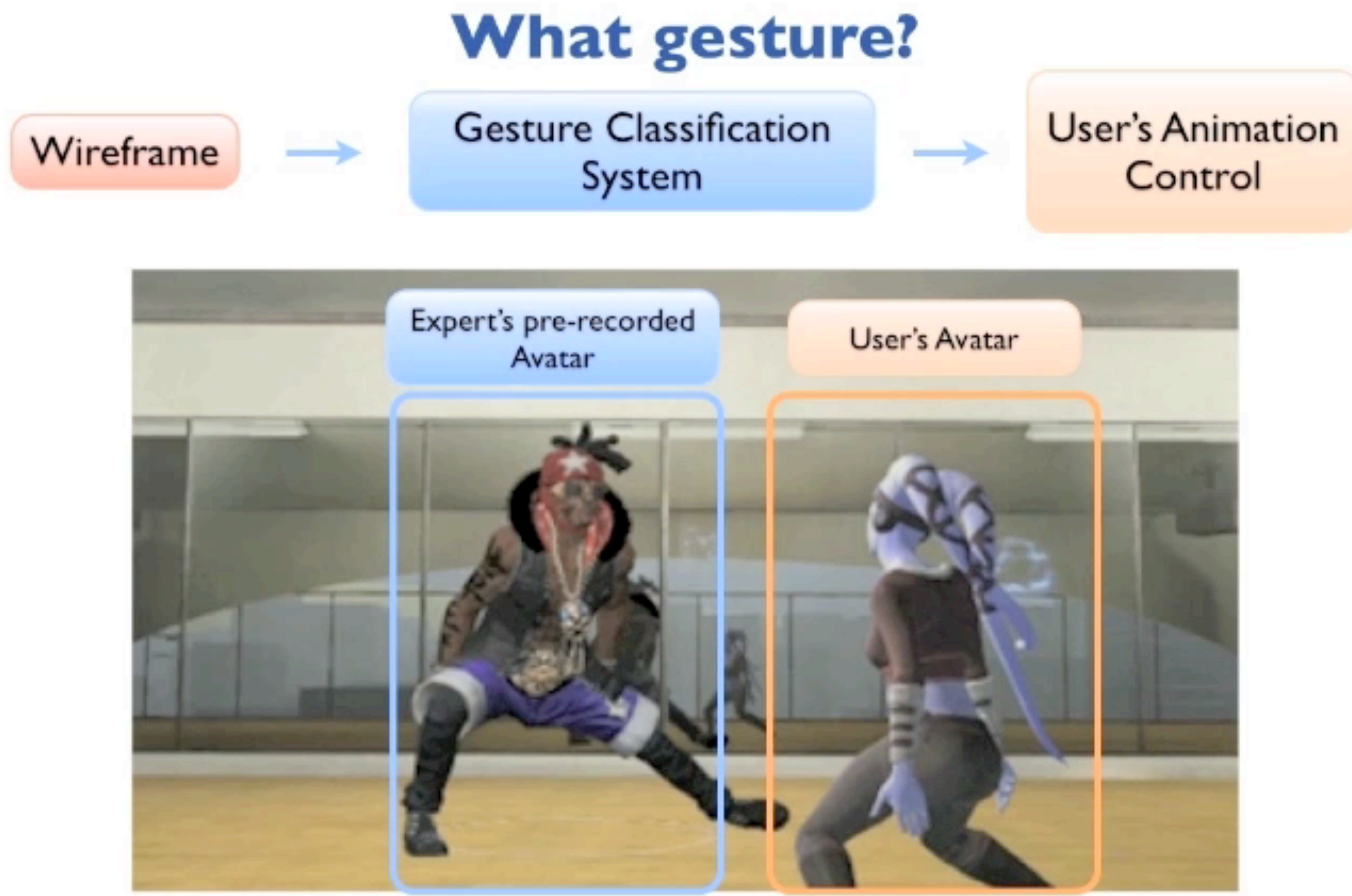
- Most games need few discrete controls
- Game designers are typically able to define controls that are SO different that any noise in location/skeleton will not really effect performance

Designing Around Limitations

[Raptis, Kirovski, Hoppes, SCA, 2011]

- Game designers are really good about designing around limitations of input devices

Designing Around Limitations



[Raptis, Kirovski, Hoppes, SCA, 2011]

- Game designers are really good about designing around limitations of input devices

Game-driven Response

- Even for very complex game interfaces, at any given point of the game only few gestures are possible
- Avatar can ask you to perform any motion, but once asked, the system only cares if you perform *that* motion

Leap Motion