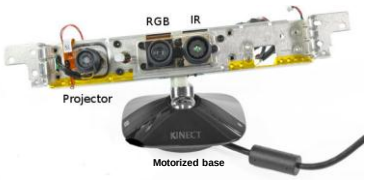
# Kinect

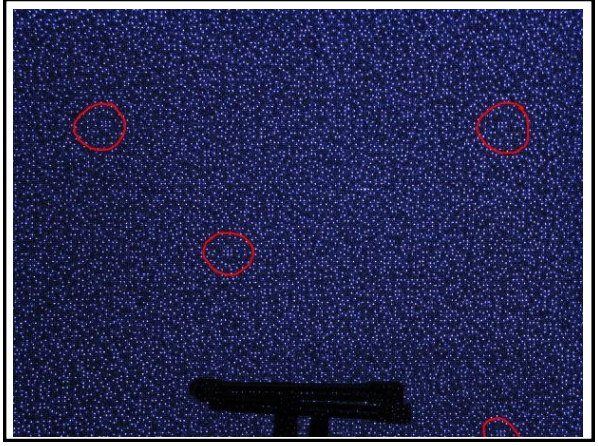
Case Study

CSE P 576

Larry Zitnick (larryz@microsoft.com)

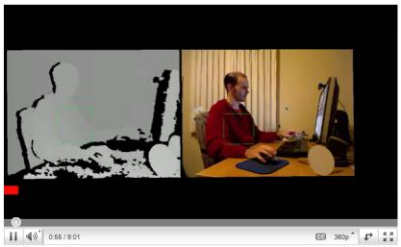

http://www.youtube.com/watch?v=dTKINGSH9Po&feature=related



## Depth

http://www.youtube.com/watch?v=inim0xWiR0o

http://www.youtube.com/watch?v=7TGF30-5KuQ&feature=related

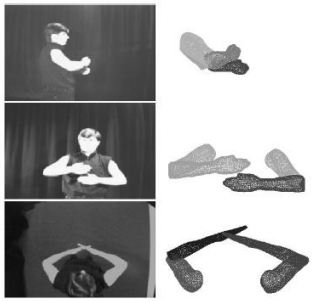


## Questions

- Why a dot pattern?
- Why a laser?
- Why only one IR camera?
- Is the dot pattern random?
- Why is heat a problem?
- How is it calibrated?
- Why isn't depth computed everywhere?
- Would it work outside?

## Pose recognition

- Research in pose recognition has been on going for 20+ years.
- Many assumptions: multiple cameras, manual initialization, controlled/simple backgrounds



Model-Based Estimation of 3D Human Motion, Ioannis Kakadiaris and Dimitris Metaxas, PAMI 2000


Fig. 22. Our automatic tracker on commercial sports footage with fast and extreme motions. On the **top**, we show results from a 300 frame sequence of a baseball pitch from the 2002 World Series. On the **bottom**, we show results from the complete medal-winning performance of Michelle Kwan from the 1998 Winter Olympics. We label frame numbers from the 7,800-frame sequence. For each sequence, our system first runs a walking pose finder on each frame, and then uses the single frame with the best score (shown in the **left insets**) to train the discriminative appearance models. In the baseball sequence, our system is able to track through frames with excessive motion blur and interlacing effects (the **center insert**). In the skating sequence, our system is able to track through extreme poses for thousands of frames. This means that our system can track long sequences with extreme poses, complex backgrounds, and fast movement *without manual intervention*.

Tracking People by Learning Their Appearance, Deva Ramanan, David A. Forsyth, and Andrew Zisserman, PAMI 2007

## Kinect

- Why does depth help?

## Algorithm design

Shotton et al. proposed two main steps:

1. Find body parts
2. Compute joint positions.

Real-Time Human Pose Recognition in Parts from Single Depth Images
Jamie Shotton Andrew Fitzgibbon Mat Cook Toby Sharp Mark Finocchio Richard Moore Alex Kipman Andrew Blake, CVPR 2011

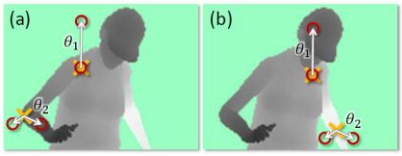
## Finding body parts

- What should we use for a feature?
- What should we use for a classifier?

## Finding body parts

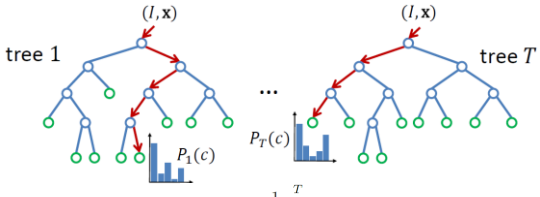
- What should we use for a feature?
  - Difference in depth
- What should we use for a classifier?
  - Random Decision Forests

## Features




$$f_\theta(I,\mathbf{x}) = d_I\left(\mathbf{x} + \frac{\mathbf{u}}{d_I(\mathbf{x})}\right) - d_I\left(\mathbf{x} + \frac{\mathbf{v}}{d_I(\mathbf{x})}\right)$$

## Classification




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

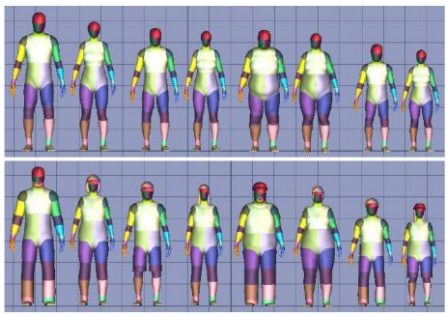
Learning:

1. Randomly choose a set of thresholds and features for splits.
2. Pick the threshold and feature that provide the largest information gain.
3. Recurse until a certain accuracy is reached or depth is obtained.

## Implementation details

- 3 trees (depth 20) (why so few?)
- 300k unique training images per tree.
- 2000 candidate features, and 50 thresholds
- One day on 1000 core cluster.
- Why RDF and not AdaBoost, SVMs, etc.?

## Synthetic data



## Synthetic training/testing

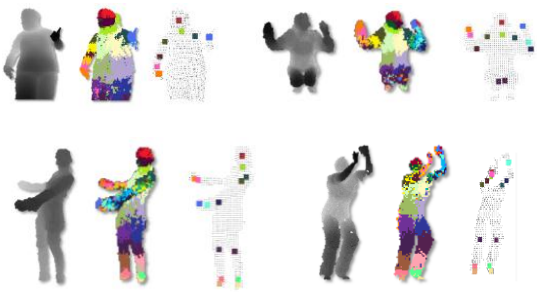

## Real test

## Results

## Joint estimation

- Apply mean-shift clustering to the labeled pixels. (why mean shift?)
- "Push back" each mode to lie at the center of the part.

## Results



## Failures

- Why would the system fail?

## Video

- http://research.microsoft.com/pubs/145347/CVPR%202011%20-%20Final%20Video.mp4

Story about the making of Kinect:

http://www.wired.co.uk/magazine/archive/2010/11/features/the-game-changer