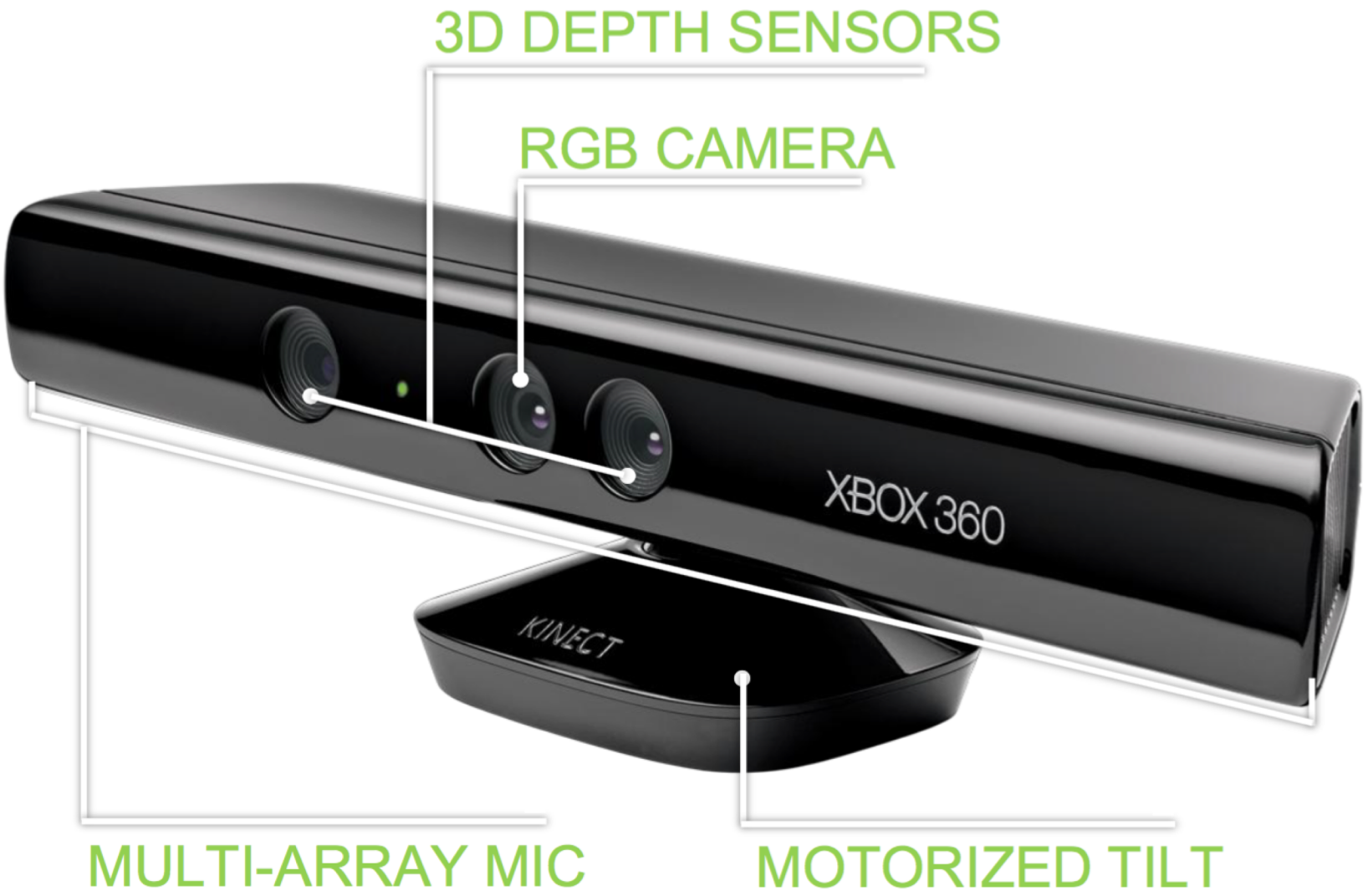
CSCM77
Structured Light & 3D Pose: Part 1
(a study on Kinect system)

Prof. Xianghua Xie

x.xie@swansea.ac.uk

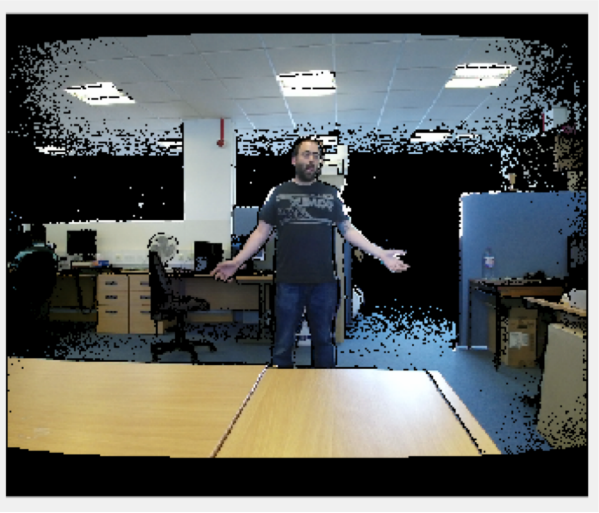
http://csvision.swan.ac.uk

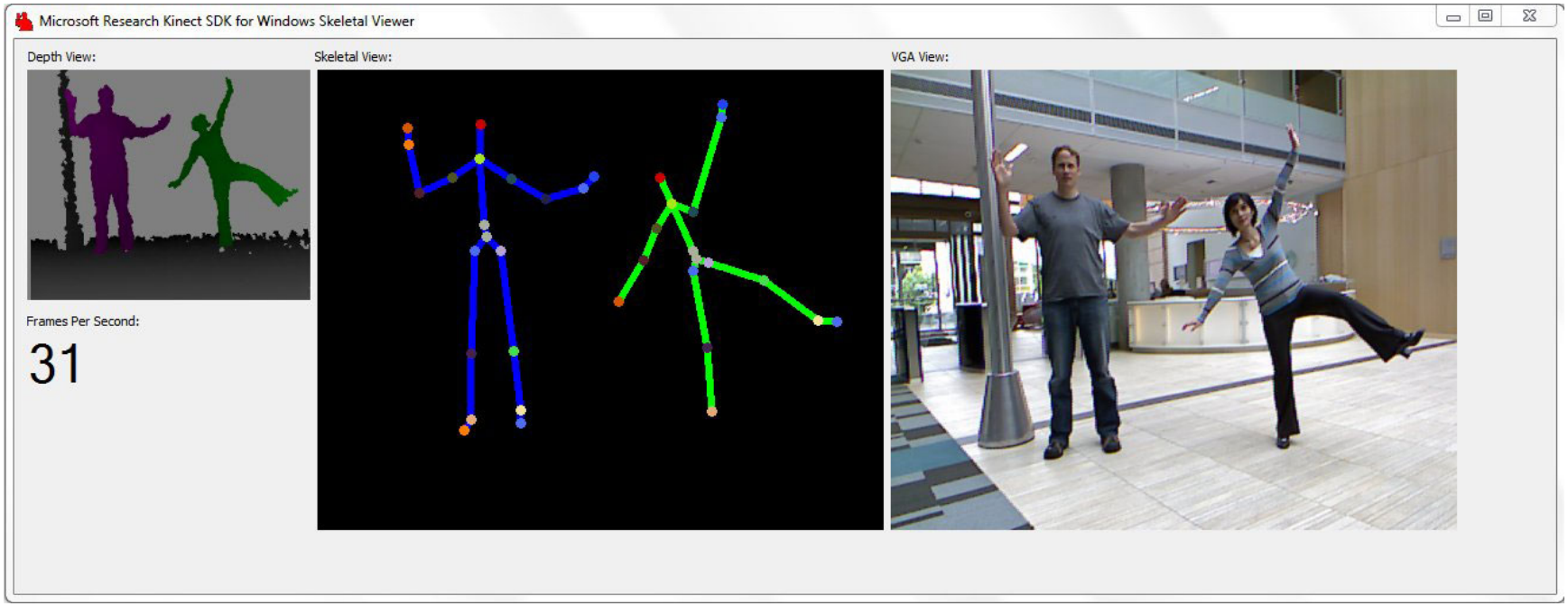
Windows Kinect System: hardware



Kinect Output

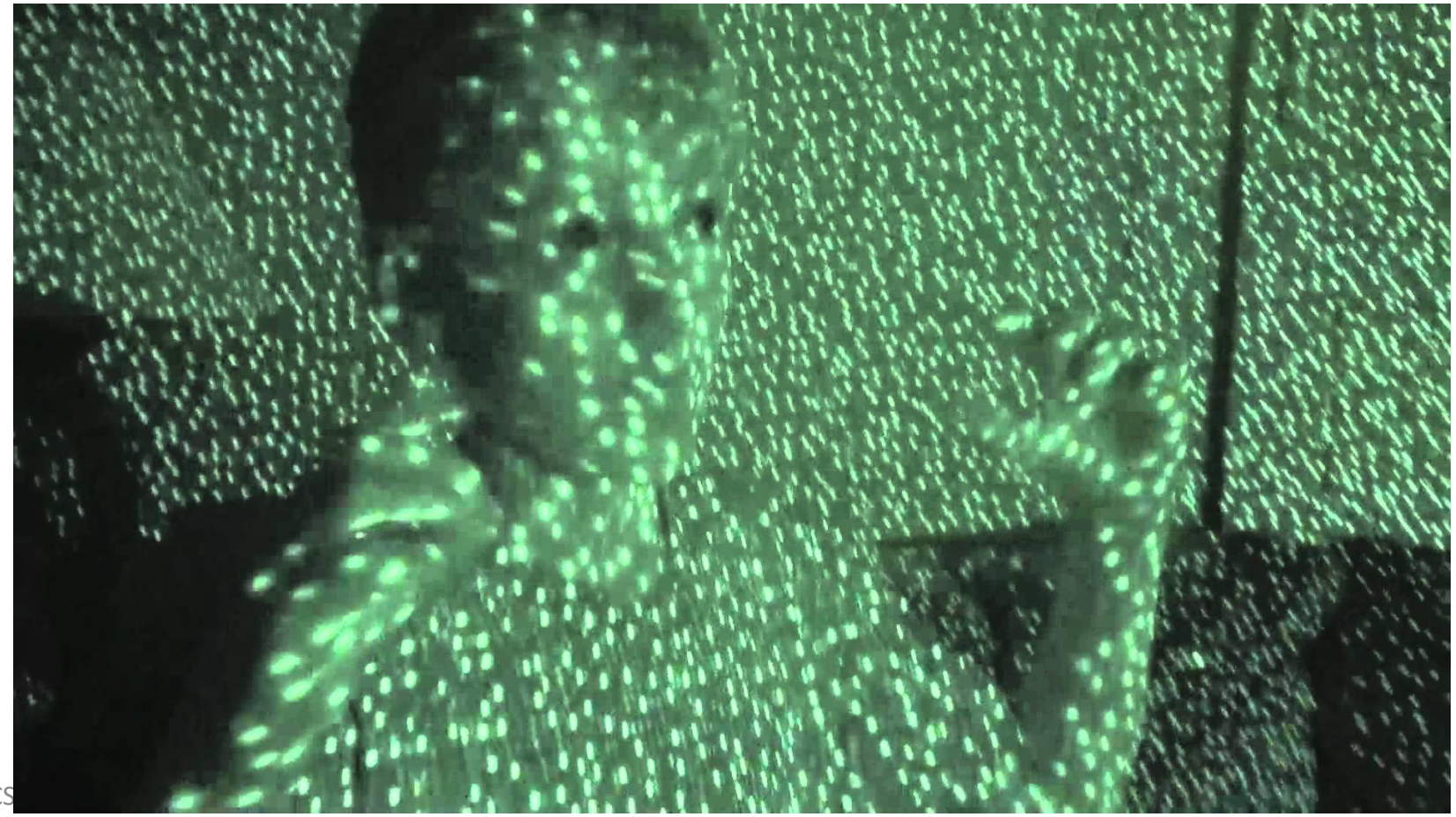
- Depth
- RGB texture
- Human pose (skeletonised)





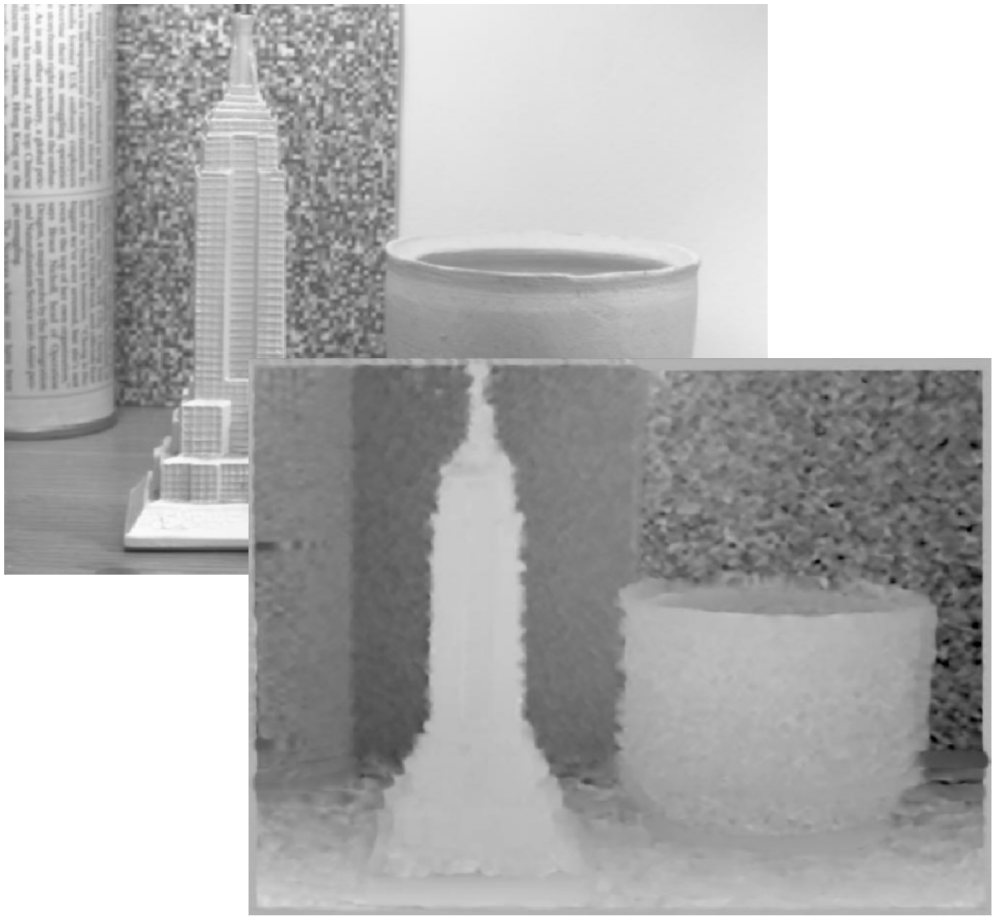
Depth Estimation using Structured Light

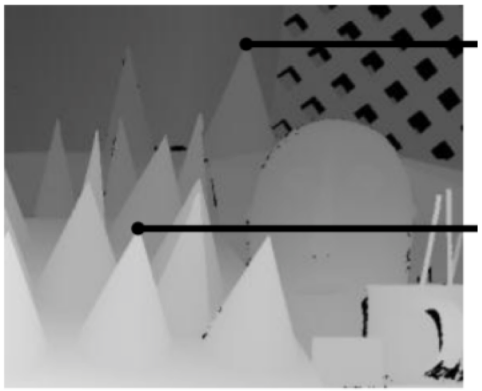
- A speckle pattern of infrared laser light
- Microsoft licensed the technology from a company called PrimeSense



Depth Estimation using Structured Light

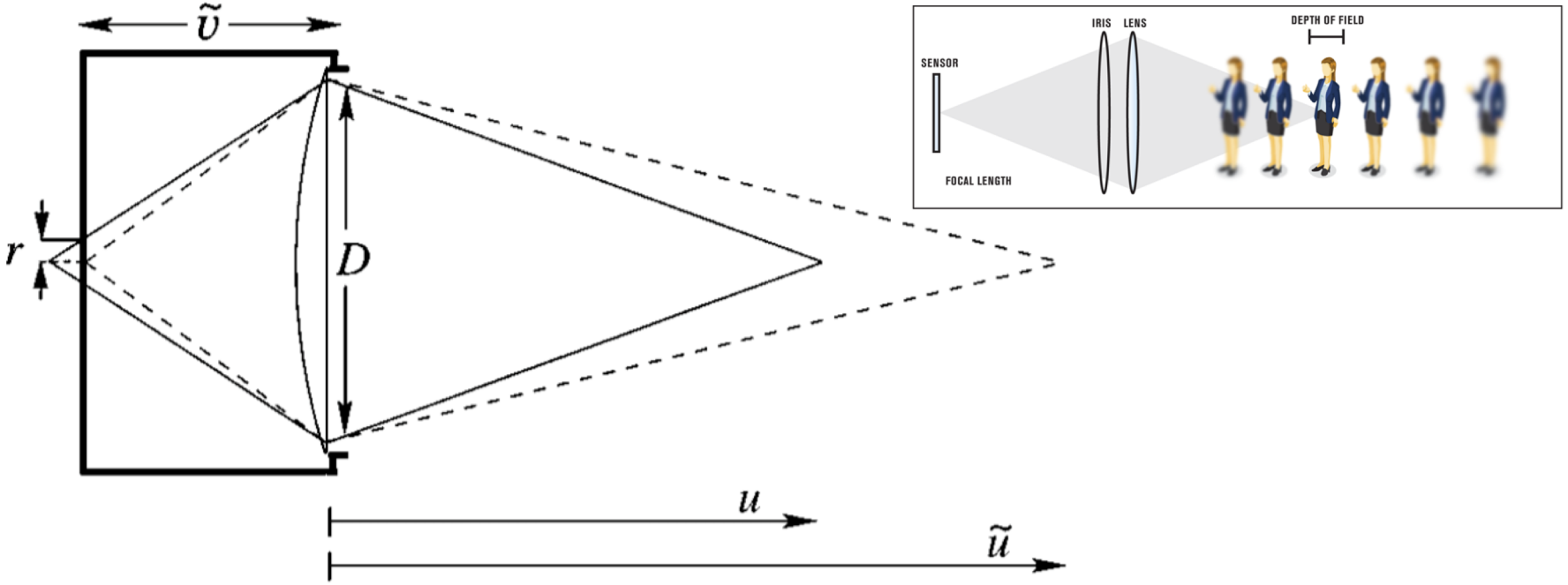
- Depth from focus
- Depth from stereo





Depth from focus

- Objects that in focus appear to be crispy/sharp
- Objects that deviate from focus appear to be blurred
- Depth from focus
  - Explore the correlation between blurriness and depth

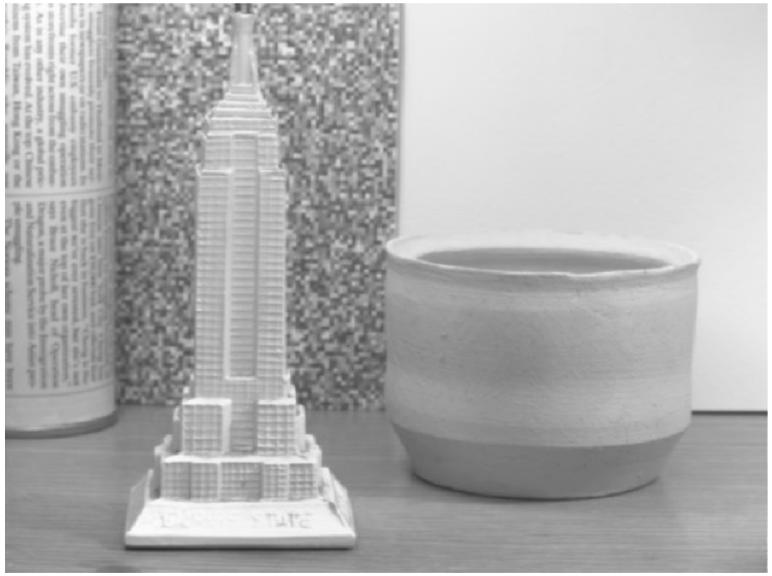


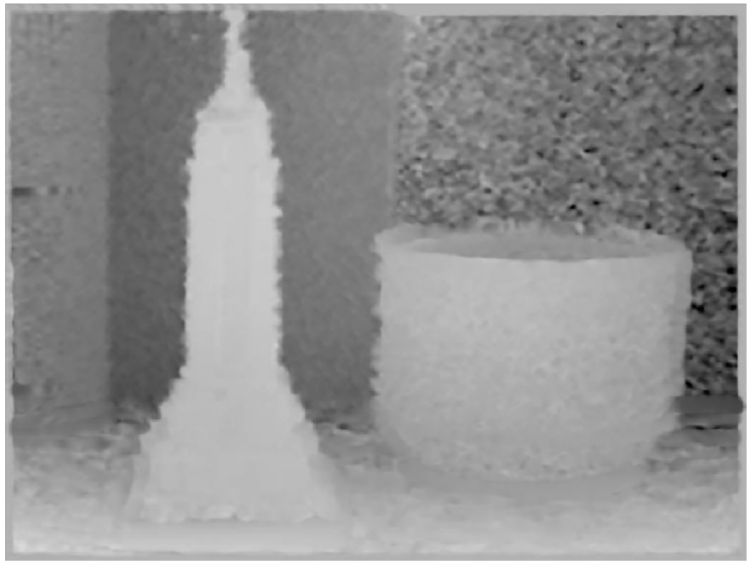
imaging system with an aperture D is tuned to view in focus object points at distance $\tilde{u}$.
image of an object point at distance u is a blur circle of radius r in the sensor plane

Depth from focus

Depth from focus

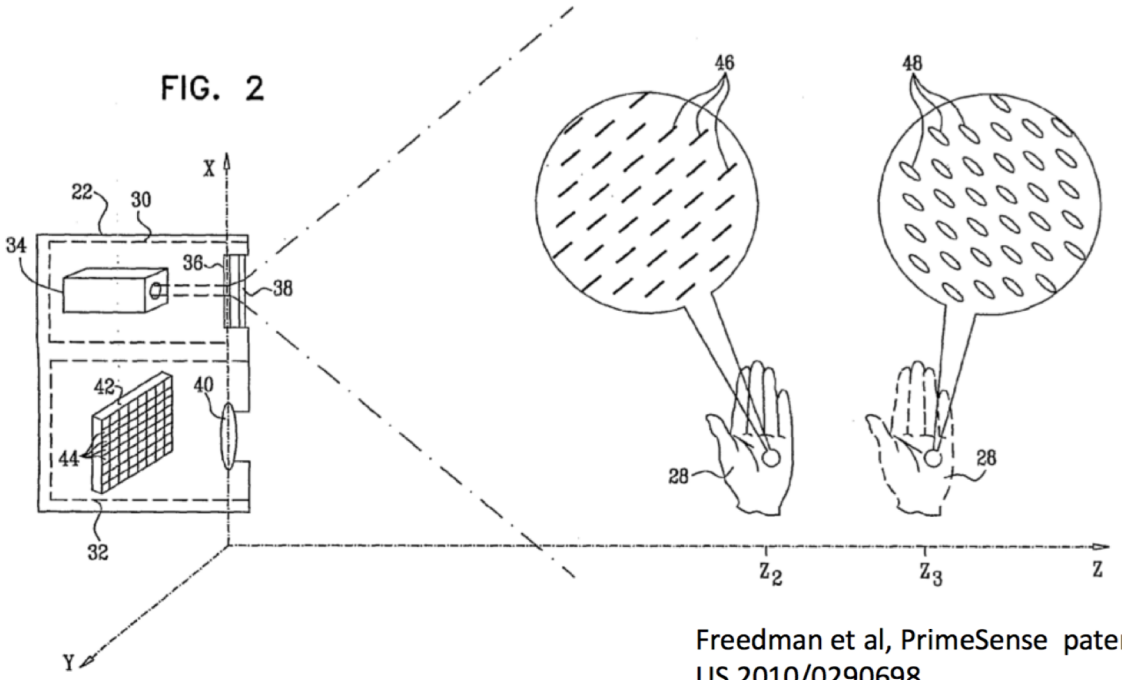
- Objects that in focus appear to be crispy/sharp
- Objects that deviate from focus appear to be blurred
- Depth from focus
  - Explore the correlation between blurriness and depth

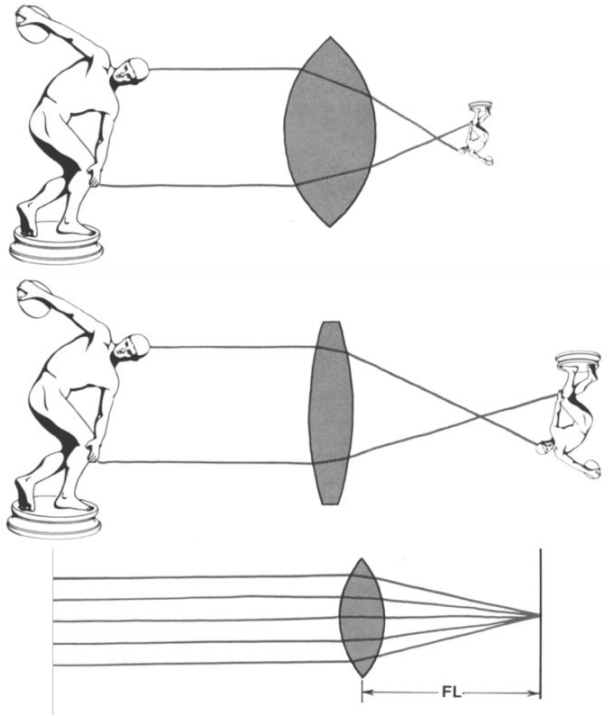




Depth from focus

- Furthermore, Kinect uses “astigmatic” lens with different focal length in x- (horizontal) and y- (vertical) directions.
- Projected circles then becomes an ellipse whose orientation depends on depth.



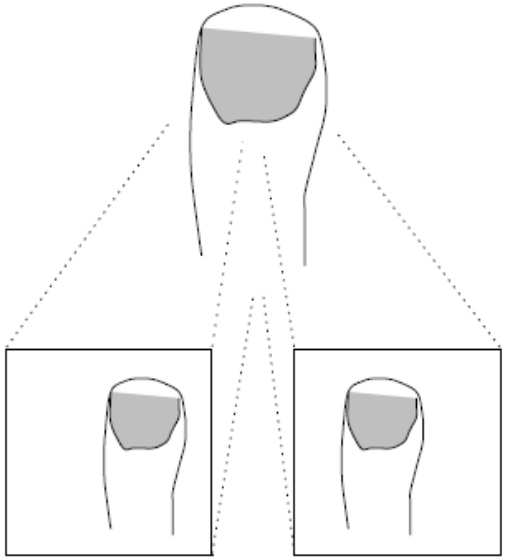


Freedman et al, PrimeSense patent application
US 2010/0290698

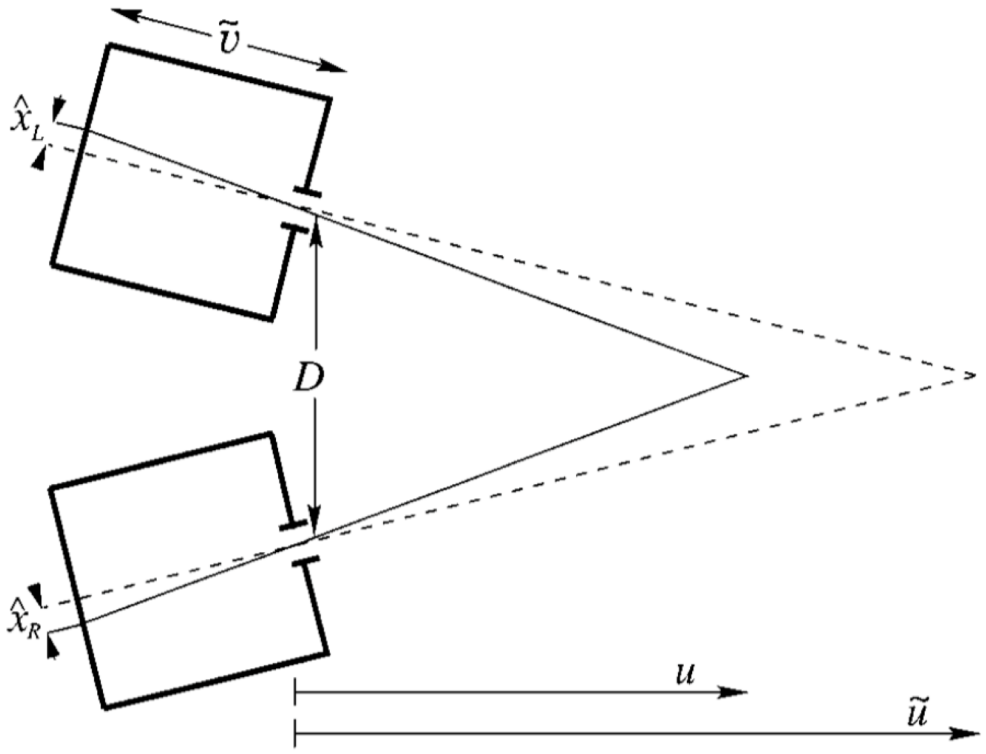
Depth from stereo

- Taking two images simultaneously from two different locations
- Objects appear in different position in each image depending on its depth in the scene (distance from the cameras)
- The position difference in the two images is known as **disparity**
  - there is a correlation between depth and disparity

- Stereo vision: inferring 3D structure from a pair of images taken from different viewpoints



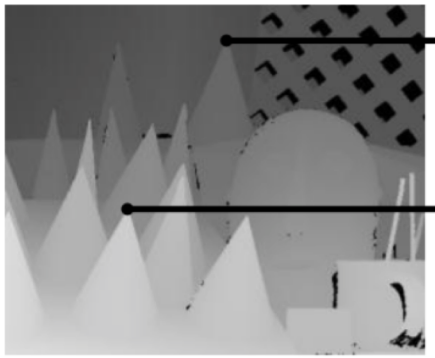
Depth from stereo



A stereo system with a baseline D. $\tilde{v}$ is the distance from the entrance pupil to the sensor. The vergence eliminates the disparity for the object point at distance $\tilde{u}$. The resulting disparity caused by the object point at u is then used to infer its depth.

Depth from stereo

Disparity Map



Disparity value: 25px

Disparity value: 50px

Left Image

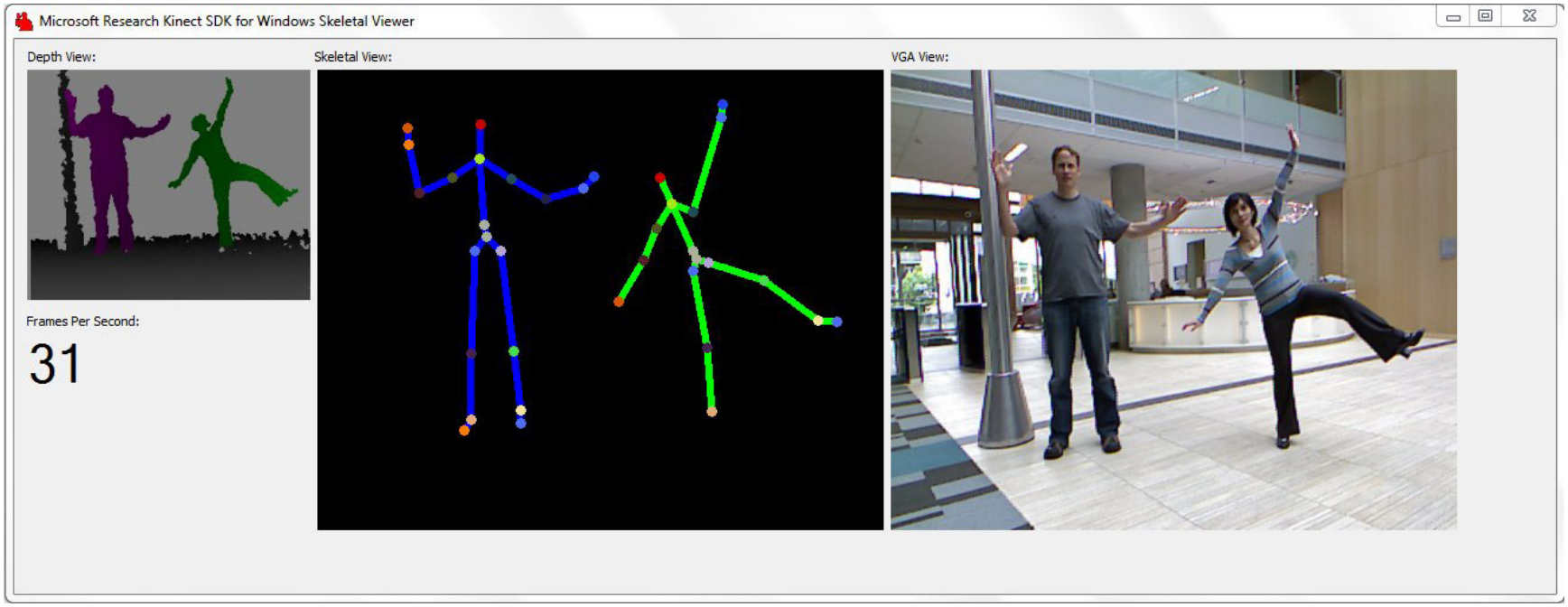
Right Image

Disparity 25px: far from camera

Disparity 50px: close to camera

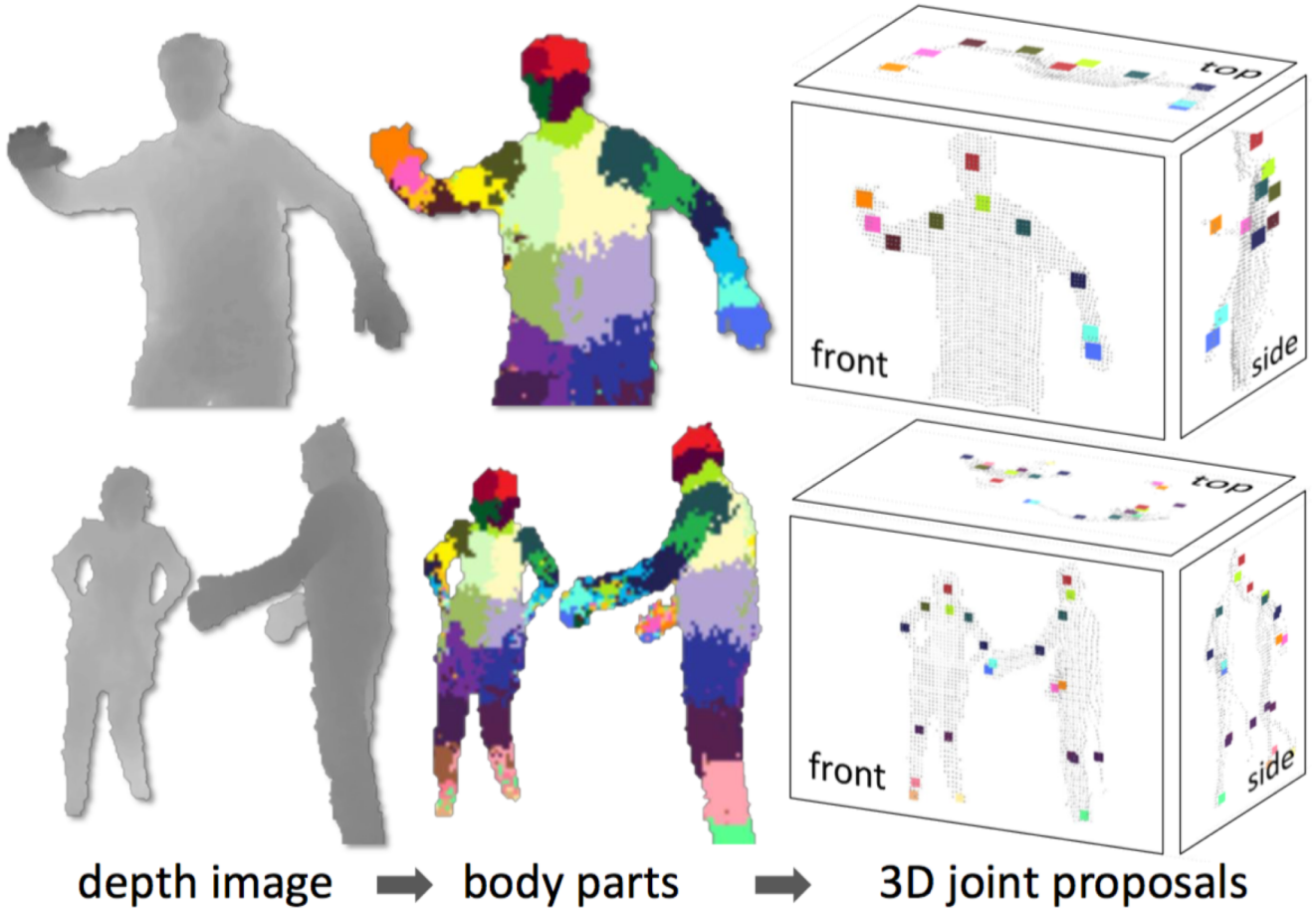
From depth to 3D joint positions

- How do we obtain 3D joint positions from depth maps?



From depth to 3D joint positions: 2 stages

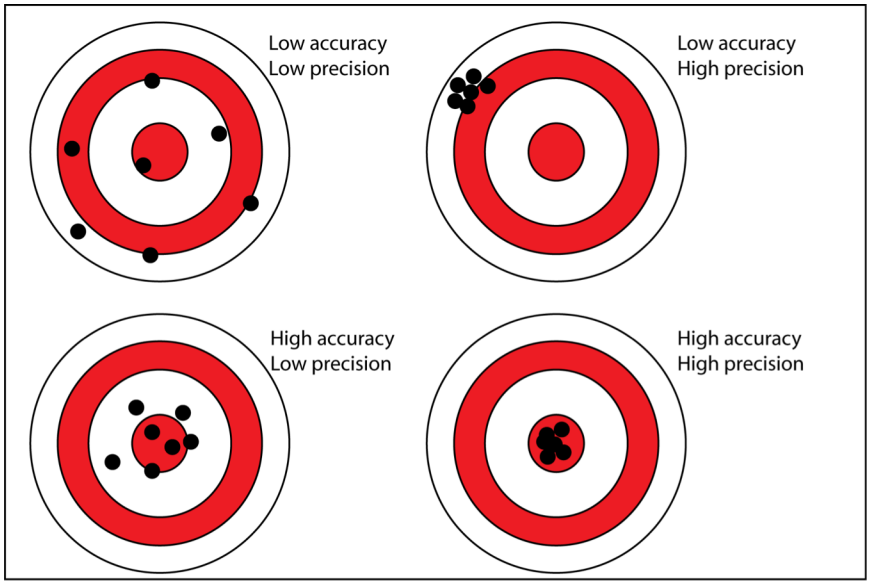
- 1. Learning and estimating “body parts”
- 2. Deducing joint positions from body parts



Body part estimation from depth maps

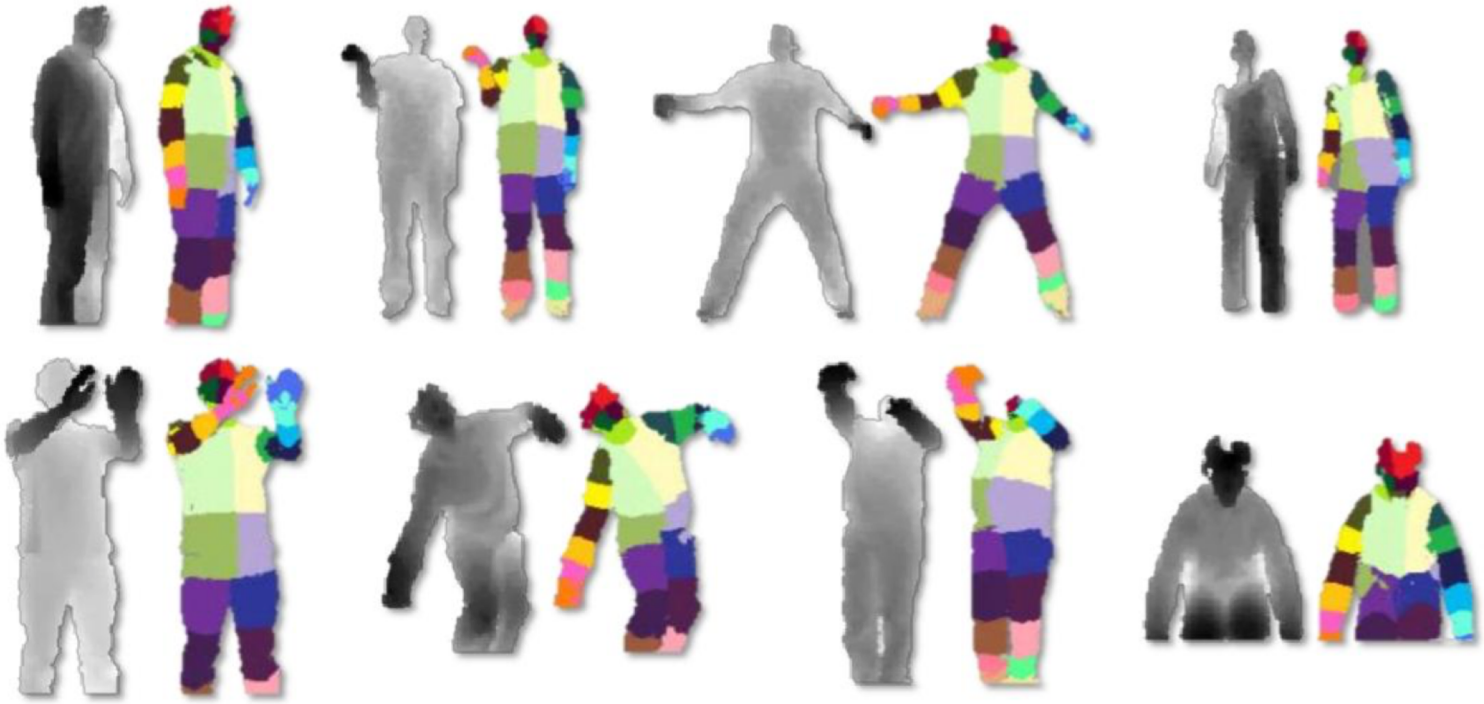
- Data
  - Training set
    - Data sets that are used to learn and discover predictive relationships.
  - Testing set
    - Testing data are used to evaluate the accuracy and precision of the learned predictive relationships.

Note, things that are precise are not necessarily accurate.



Body part estimation: Training Data

- Real world training data
  - 100,000 depth images with known skeletons (obtained from Motion Capture systems).
  - Body parts are colour labelled.



Body part estimation: Training Data

- Synthetic training data
  - For each real image, use computer graphics to render dozens more instances with varying pose and shape parameters.
  - Thus, obtain over a million training examples.

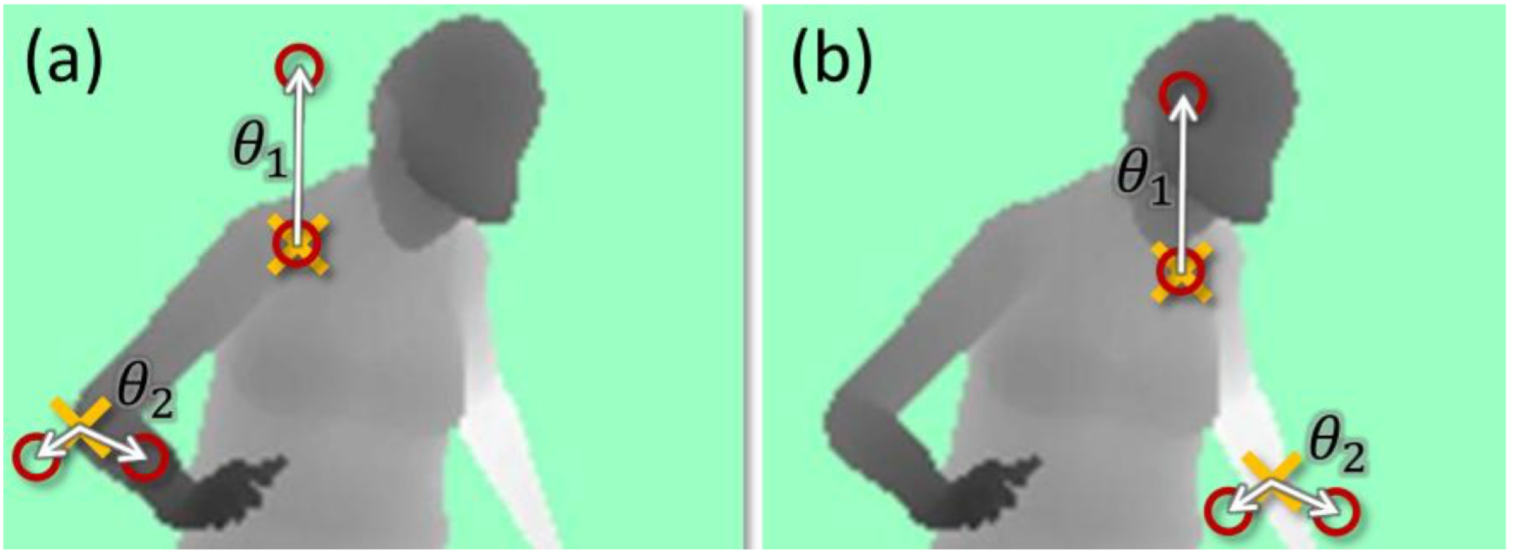
Body part estimation: depth features

- In order to pin-point each pixel in the depth map to individual body parts:
  - 1. What are the useful features to achieve this?
  - 2. What is a reliable decision-making method?

Feature extraction + Machine learning

Body part estimation: depth features

- Rather simplistic features: depth differences in local neighbourhoods
  - Individually these features provide only a weak signal about which part of the body the pixel belongs to, but in combination they are sufficient to accurately disambiguate all trained parts.
  - Very efficient (and can be implemented on GPU).



The yellow crosses indicates the pixel x being examined. The red circles indicate the offset pixels. In (a), the two example features give a large depth difference response. In (b), the same two features at new image locations give a much smaller response.

Body part estimation: classification

- Those depth feature, together with their body part labels, are fed into a Classifier
  - Random Forests classifier
    - For now, just treat this as a Black Box
    - Will be explained in Part Two

Training

Testing

3D joint localisation from estimated body parts

- Apply a clustering method (Mean Shift) to estimate the joint positions from classified body parts