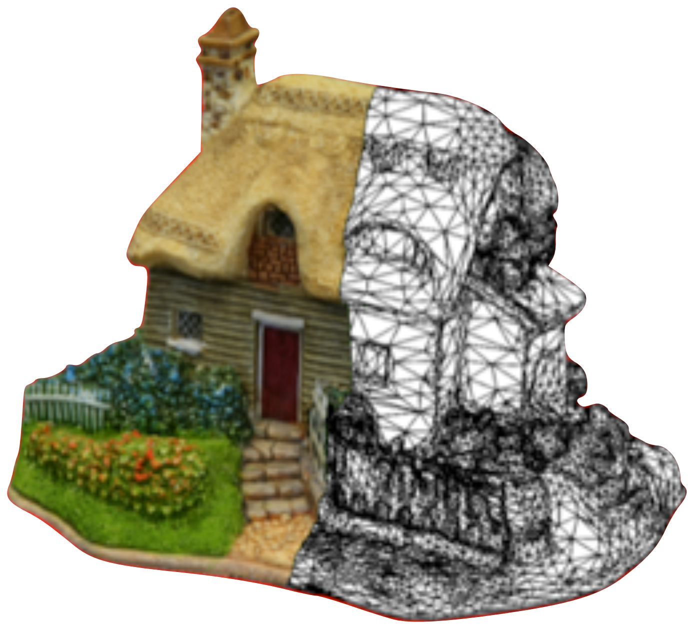


Adventures in 3d Computer Vision

George Vogiatzis

TOSHIBA Leading Innovation

Aston University Birmingham

UNIVERSITY OF CAMBRIDGE

Outline

- 3d vision for capturing 3d shape
  - Applications, mature technologies and their limitations
- Video-based multi-view stereo
- Automatic calibrated multi-segmentation
- Face capture with multi-spectral photometric stereo

Aston University
Birmingham

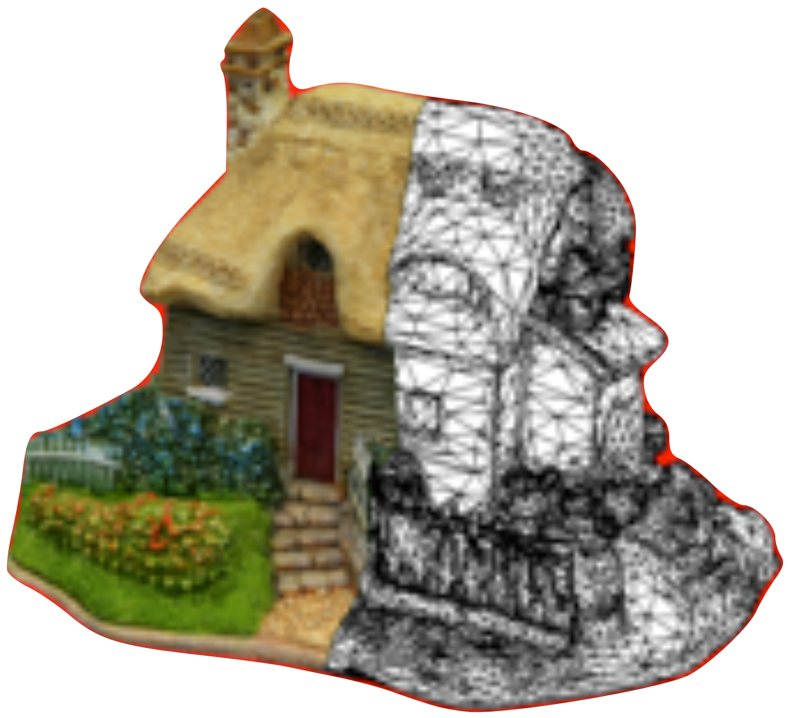
Outline

- 3d vision for capturing 3d shape
    - Applications, mature technologies and their limitations
- Video-based multi-view stereo
- Automatic calibrated multi-segmentation
- Face capture with multi-spectral photometric stereo

Aston University
Birmingham

Models of 3d shape

- There is a ever growing need for photorealistic 3d models
- 3d model = “digital copy” of real object
- Allows us to
  - inspect details
  - measure properties
  - reproduce in different material



Aston University
Birmingham

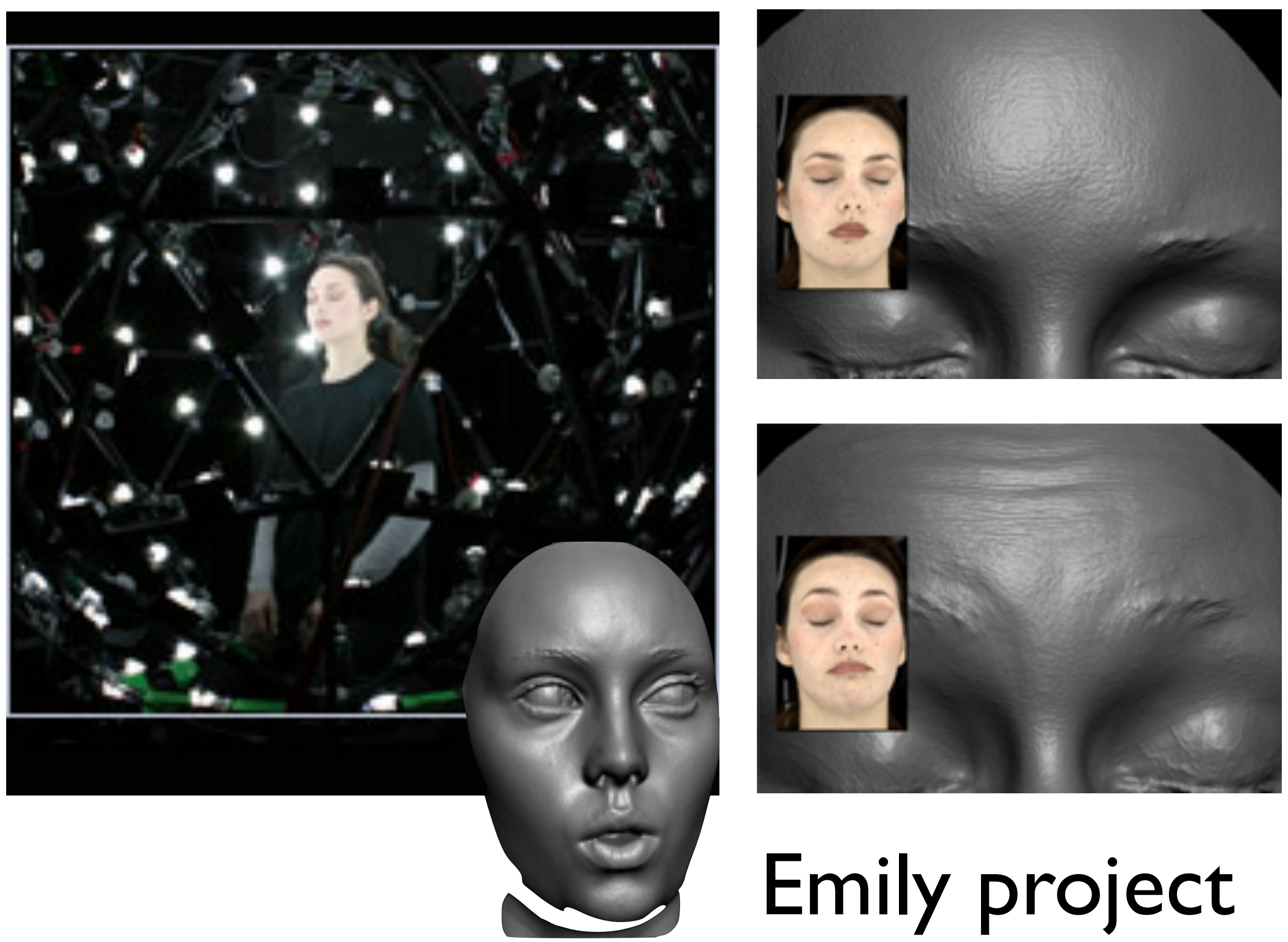
Applications

- Cultural heritage preservation

Aston University
Birmingham

Applications

- Computer games and Film



Emily project

Developing "assets"

Aston University
Birmingham

Applications

- City modelling

Nokia Maps 3D WebGL

http://maps3d.svc.nokia.com/webgl/index.html

Aston University
Birmingham

Applications

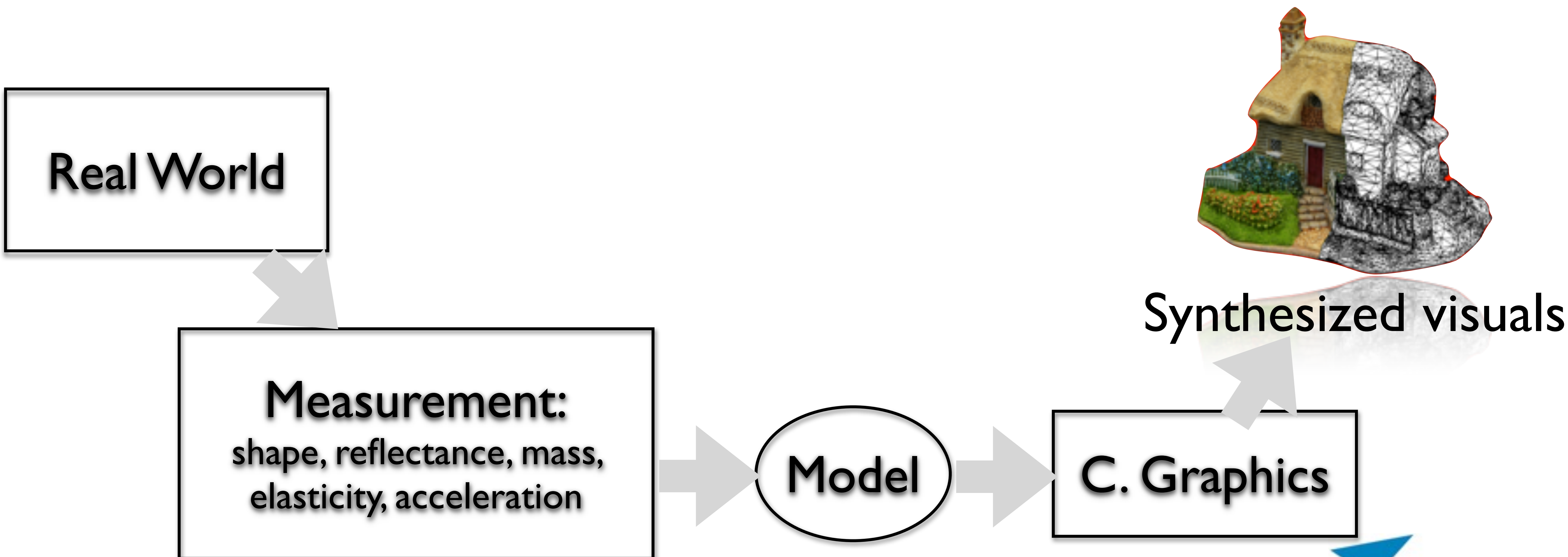
- E-commerce
  - www.metail.co.uk

UNIVERSITY OF CAMBRIDGE

Aston University Birmingham

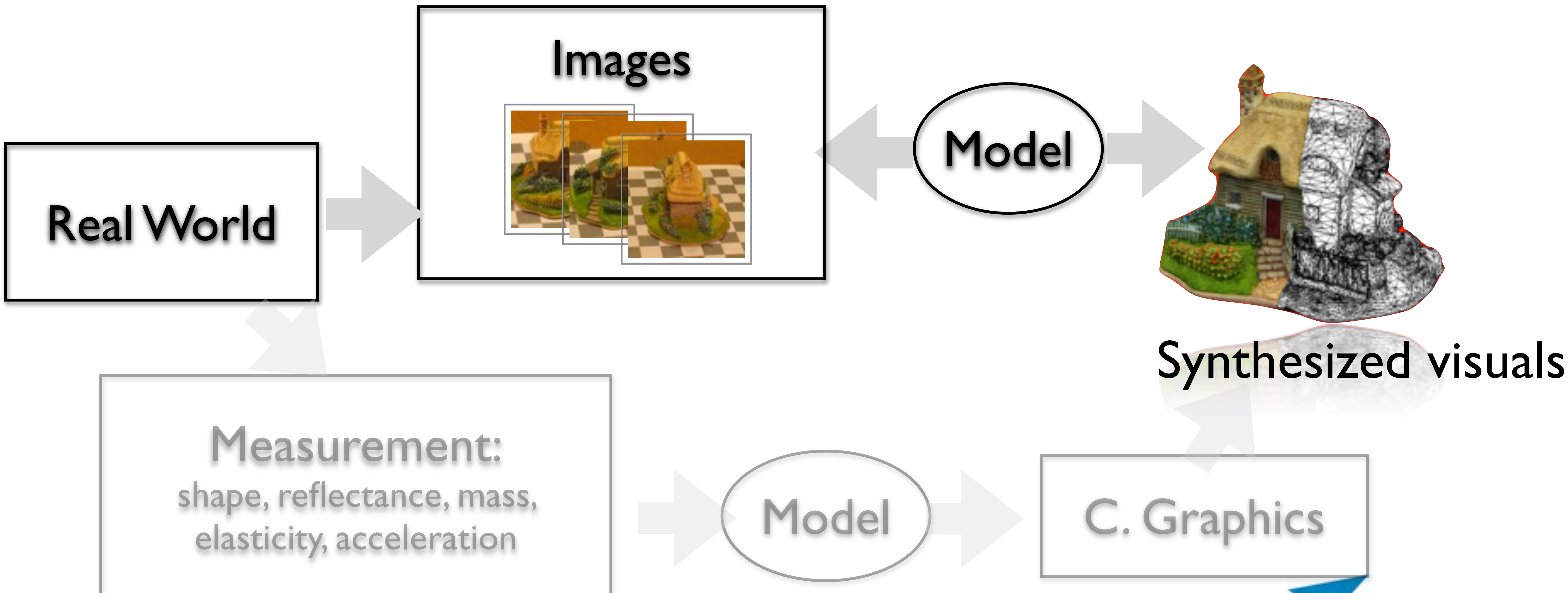
3d vision for capturing shape

- Best example of "**comp. vision in the real world**"
- Why is it successful?



3d vision for capturing shape

- Best example of "**comp. vision in the real world**"
- Why is it successful?



Aston University
Birmingham

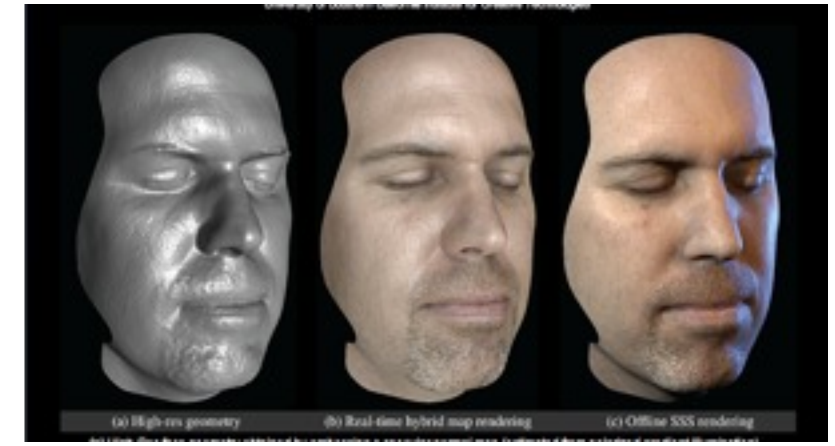
3d vision technologies

- Shape from X, where X=
  - Shading,
  - Photometric stereo,
  - Silhouettes,
  - Vanishing points,
  - Optic flow,
  - Polarization,
  - Texture,
  - Defocus,
  - Refraction paterns,
  - Atmospheric perpective,
  - Learning photo-popups,
  - ……

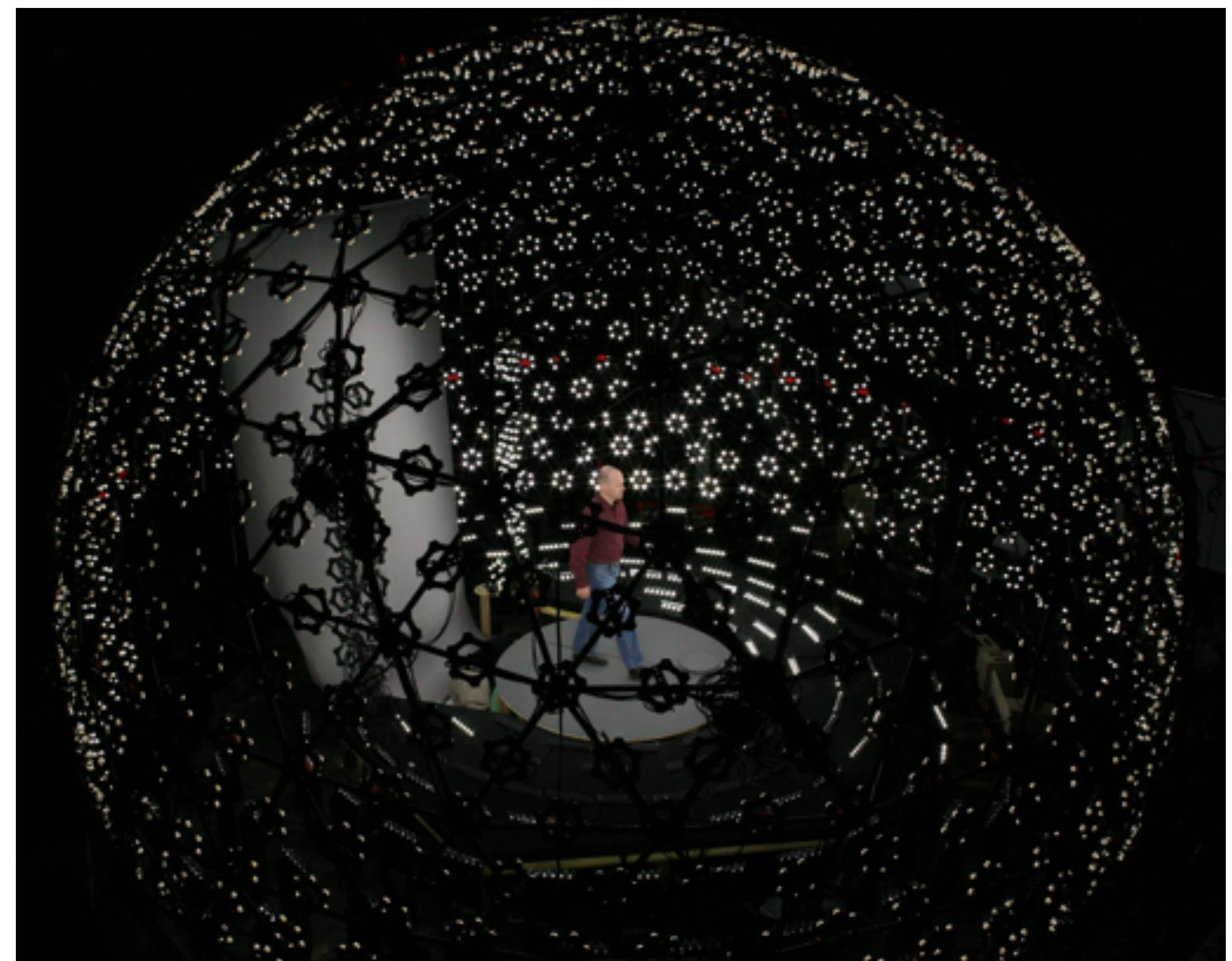
- Some of these are already maturing into commercial-grade solutions

Aston University
Birmingham

Mature technologies



- Photometric/Polarimetric surface capture
  - Image metrics, Light Stage 1-6
  - high-end Film Industry applications



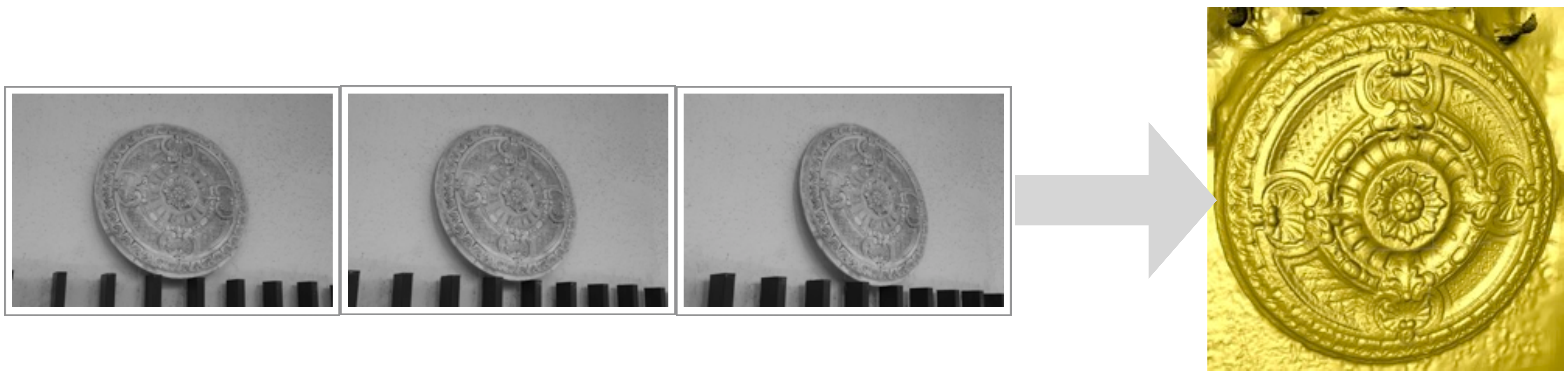
+ high-quality results
- complex setup
- expensive

Mature technologies

- Multi-View Stereo
  - Capture 10-100 high-res stills (>12 Mpx)

  + Very cheap, lightweight method

  + Easy to deploy outdoors

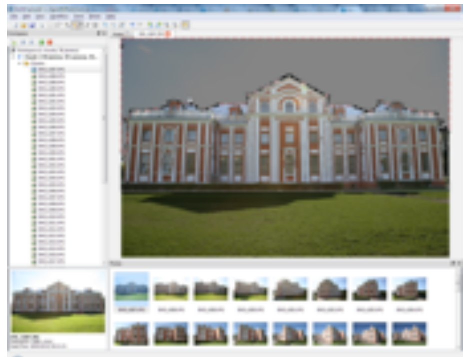


Aston University
Birmingham

MVS systems

- Accurate, dense, and robust multiview stereopsis (PMVS) *[ENS, Furukawa & Ponce '07]*
  - Binaries available, widely downloaded and used
- Using Multiple Hypotheses to Improve Depth-Maps for MVS *[Cambridge, Campbell et al '08]*
  - Several commercial implementations
- Towards high-resolution large-scale multi-view stereo *[IMAGINE, Vu et al '09]*
  - Licensed to Autodesk "123D Catch" Free to use

- Agisoft's Photoscan (basic version $179)



Aston University
Birmingham

Multi-view stereo

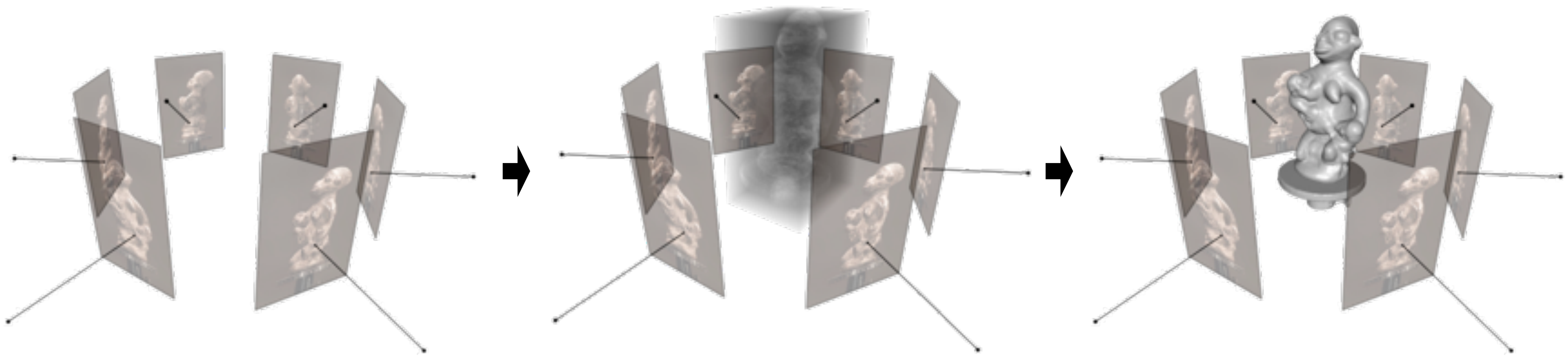


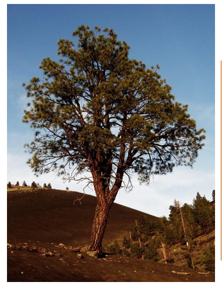
Image acquisition, camera calibration

photo-consistency 3d volume from images

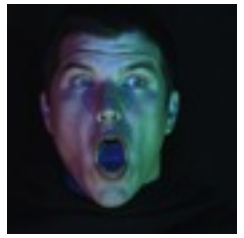
3d surface from photo-consistency

- Key Limitations:
  - Sensors constantly evolve. High-res stills not the final answer. What about Video? RGB-D?

  - Types of objects: world does not consist of well textured, granite-like objects.



  - What about deformations?



Aston University
Birmingham

Outline

- 3d vision for capturing 3d shape
  - Applications, mature technologies and their limitations
- Video-based multi-view stereo
- Automatic calibrated multi-segmentation
- Face capture with multi-spectral photometric stereo

Aston University
Birmingham

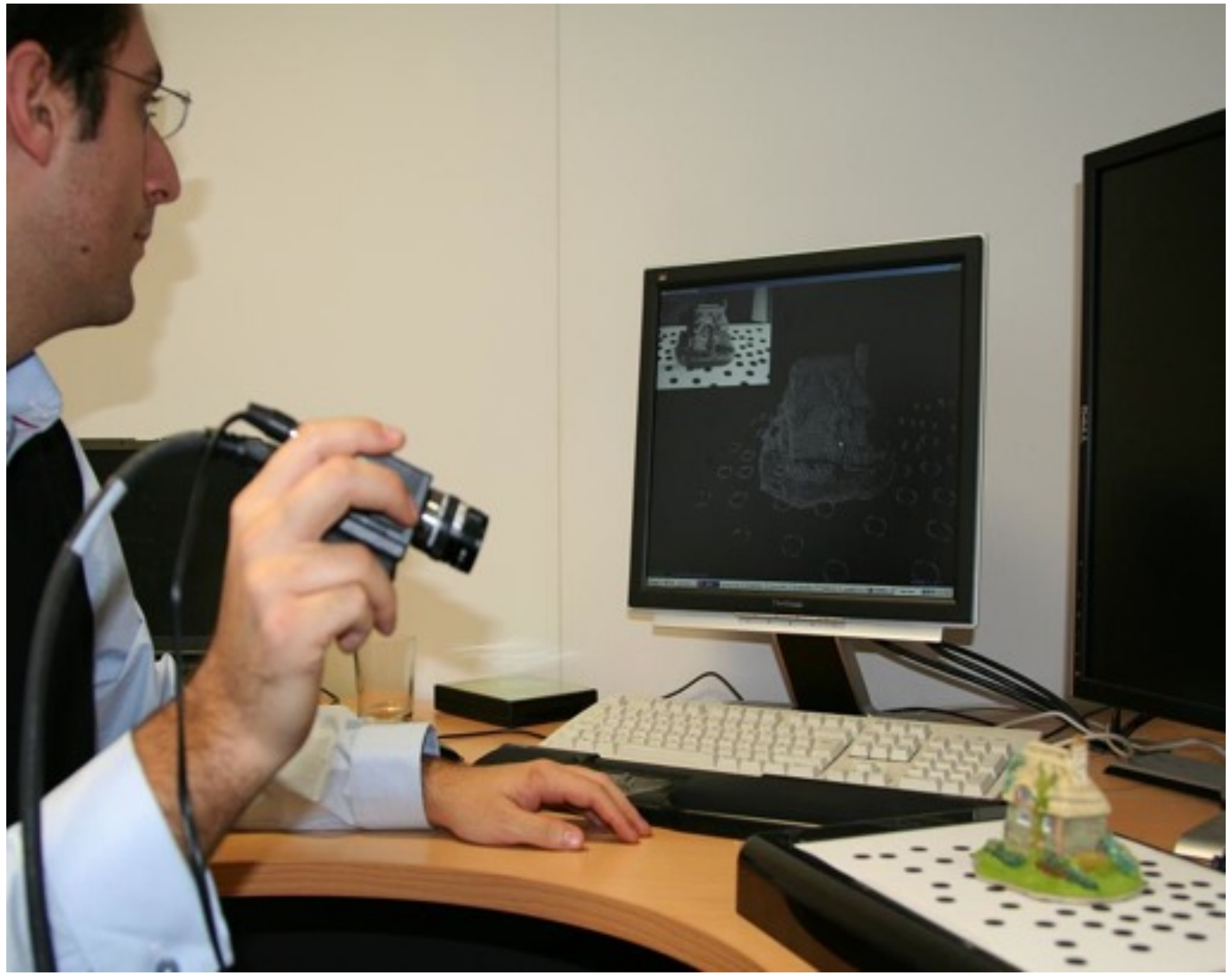
Video-based Multi-View Stereo?

Immediate feedback

- Interactive reconstruction
- Feedback leads to better models
- Still passive & cheap

Requirements:

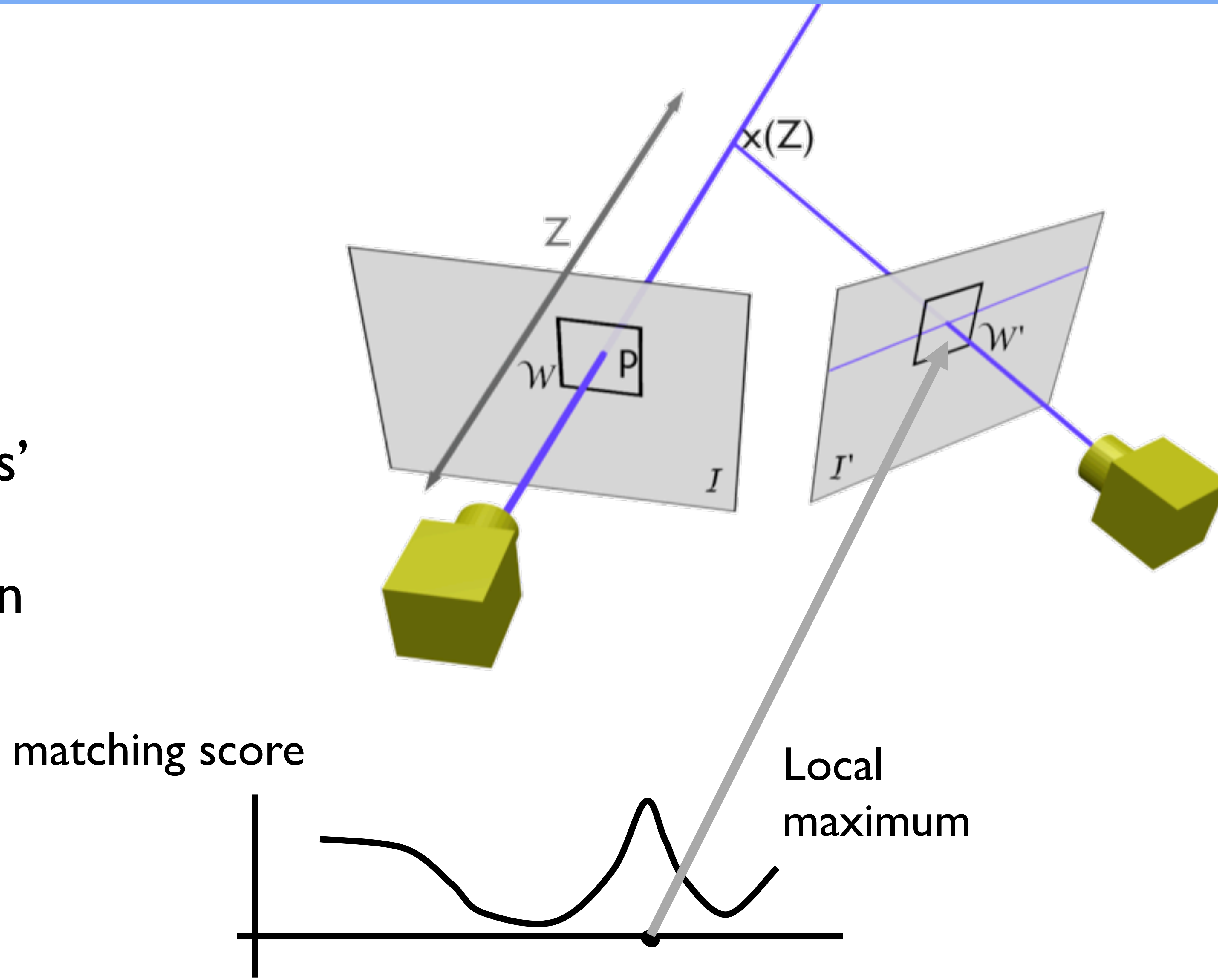
- online camera pose estimation (visual SLAM)
- real-time
- interactive
- lots of data



Aston University
Birmingham

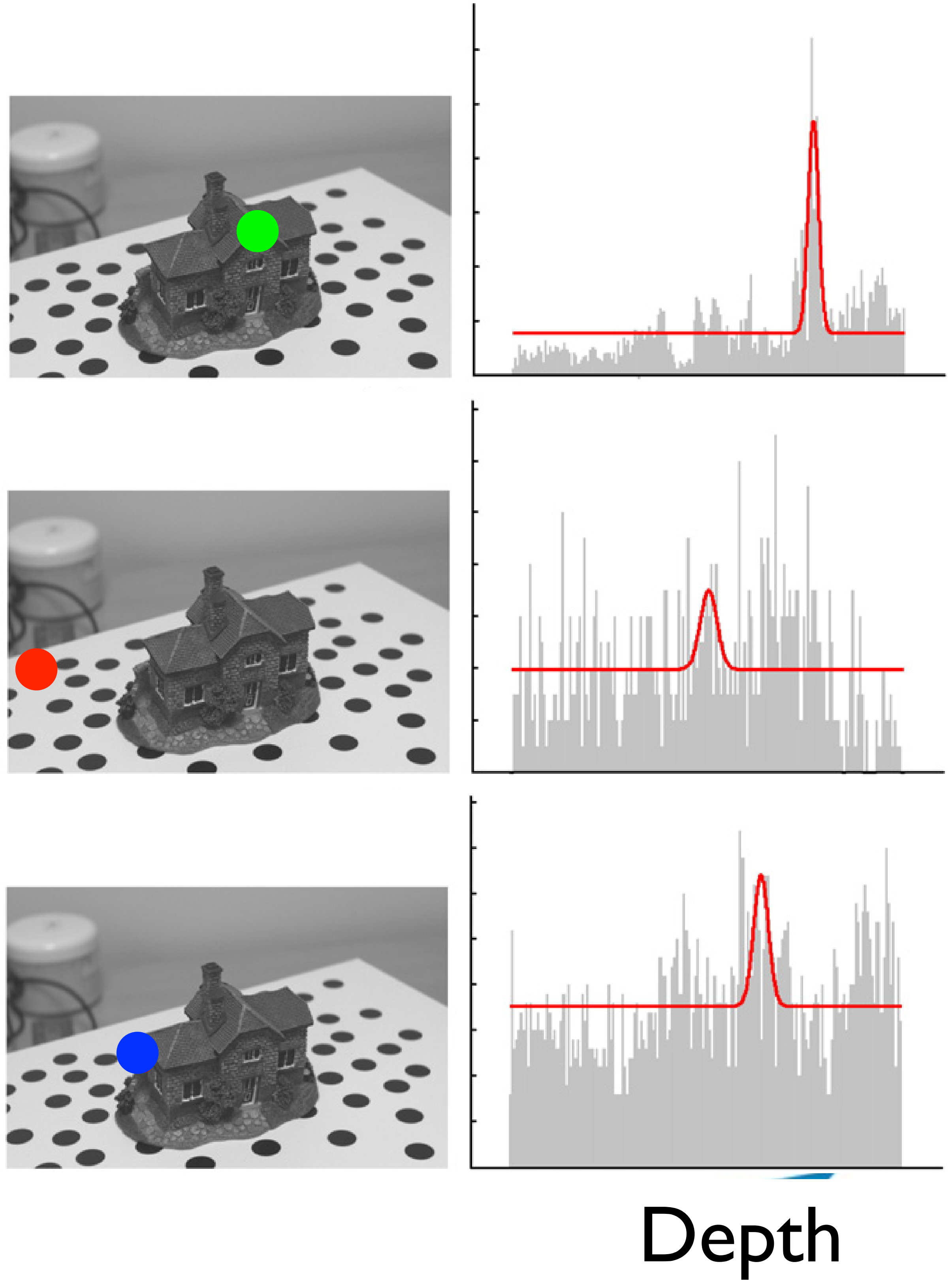
Pixel = depth sensor

- Reference pixel fixed during depth inference (we store the patch)
- NCC search along each incoming video frame
- Peaks in NCC score correspond to 'measurements' in depth.
- Our aim: to *infer* the unknown depth behind the reference pixel sensor



Measurement model

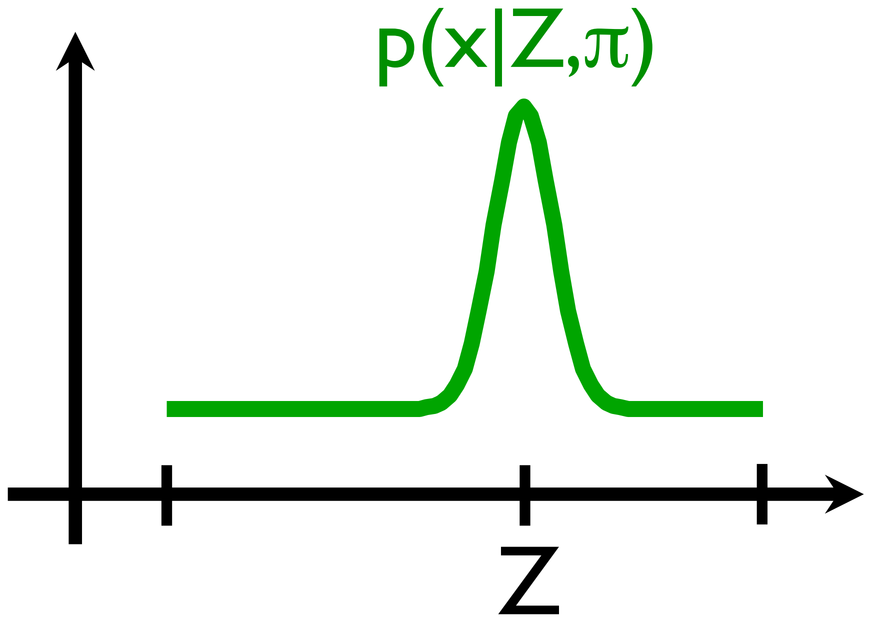
- Well textured pixel
- Untextured point
- Occluded point



Strategy

- Model sensor probabilistically as a **Gaussian+Uniform** mixture

$$\boxed{p(x|Z,\pi)=\pi\, N(x|Z,\tau^2) + (1-\pi)\, U(x)}$$

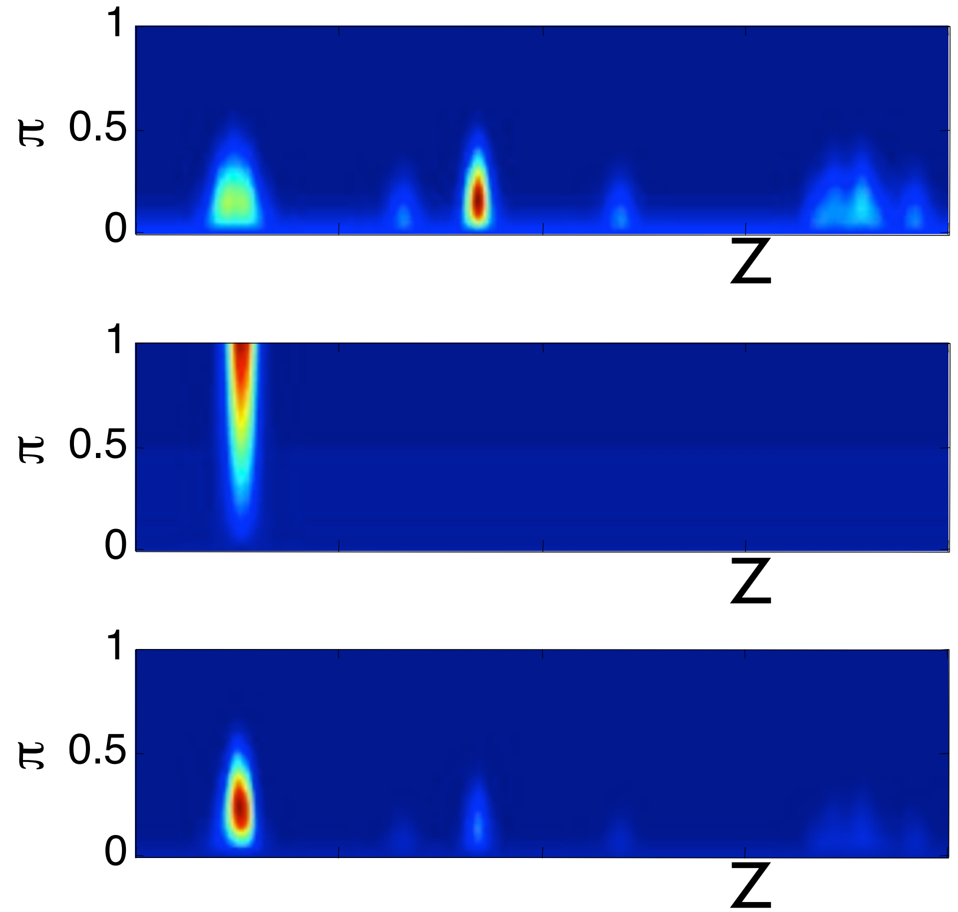


- –Z is the actual depth we are looking for
- –π is the inlier ratio, also unknown
- –x is the measurement (data)

- Can fit using EM but not in one pass!

Aston University
Birmingham

Sequential inference

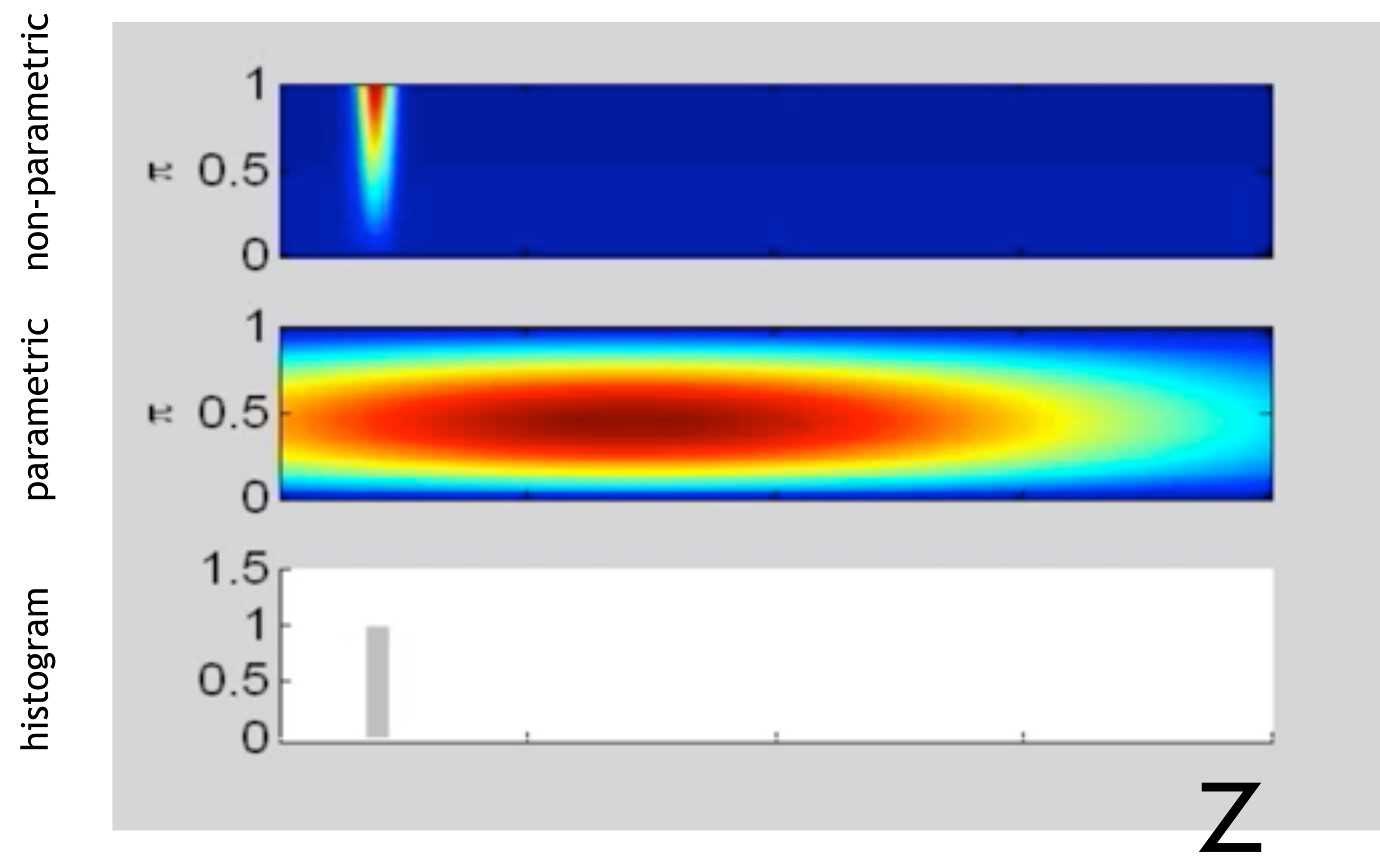


- Posterior at time t,
  $p(Z, \pi | x_1, \ldots, x_t)$
- Likelihood of measurement at t+1,
  $p(x_{t+1} | Z, \pi)$
- Posterior at time t+1,
  - $p(Z, \pi | x_1, \ldots, x_t) \propto p(x_{t+1} | Z, \pi) \times p(Z, \pi | x_1, \ldots, x_t)$
- What form can $p(Z, \pi | x_1, \ldots, x_t)$ take?
  - Closed form is intractable, Non-parametric 2d histogram is too memory intensive
  - Approximate with a parametric N(Z)xBeta(π) form
    - Variational argument (minimises KL divergence)
    - Needs 4 numbers per pixel to represent posterior
  - Can't do full variational approx. in one pass
    - moment matching

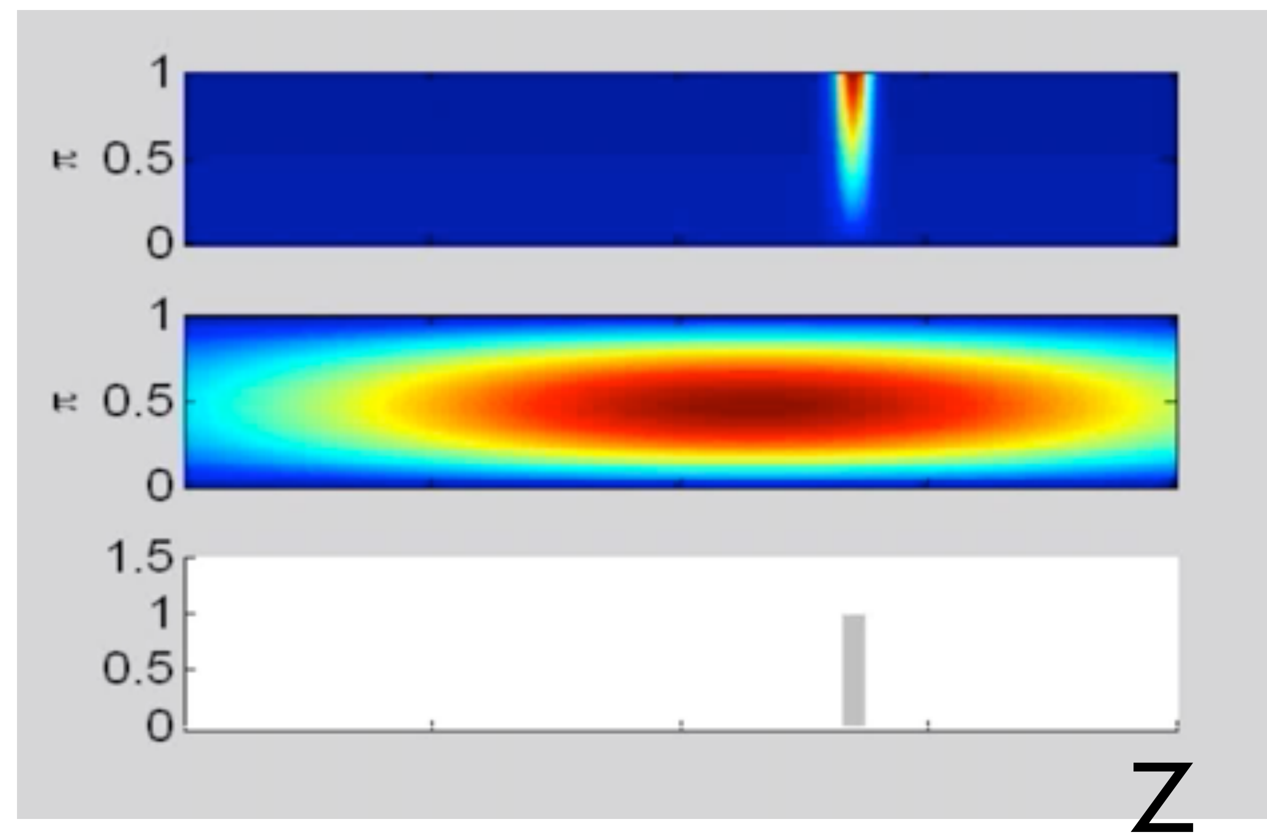
Aston University
Birmingham

How well does it work?

Successful case



Failure case



Still ok because inferred inlier ratio is low

Aston University
Birmingham

Outline of algorithm

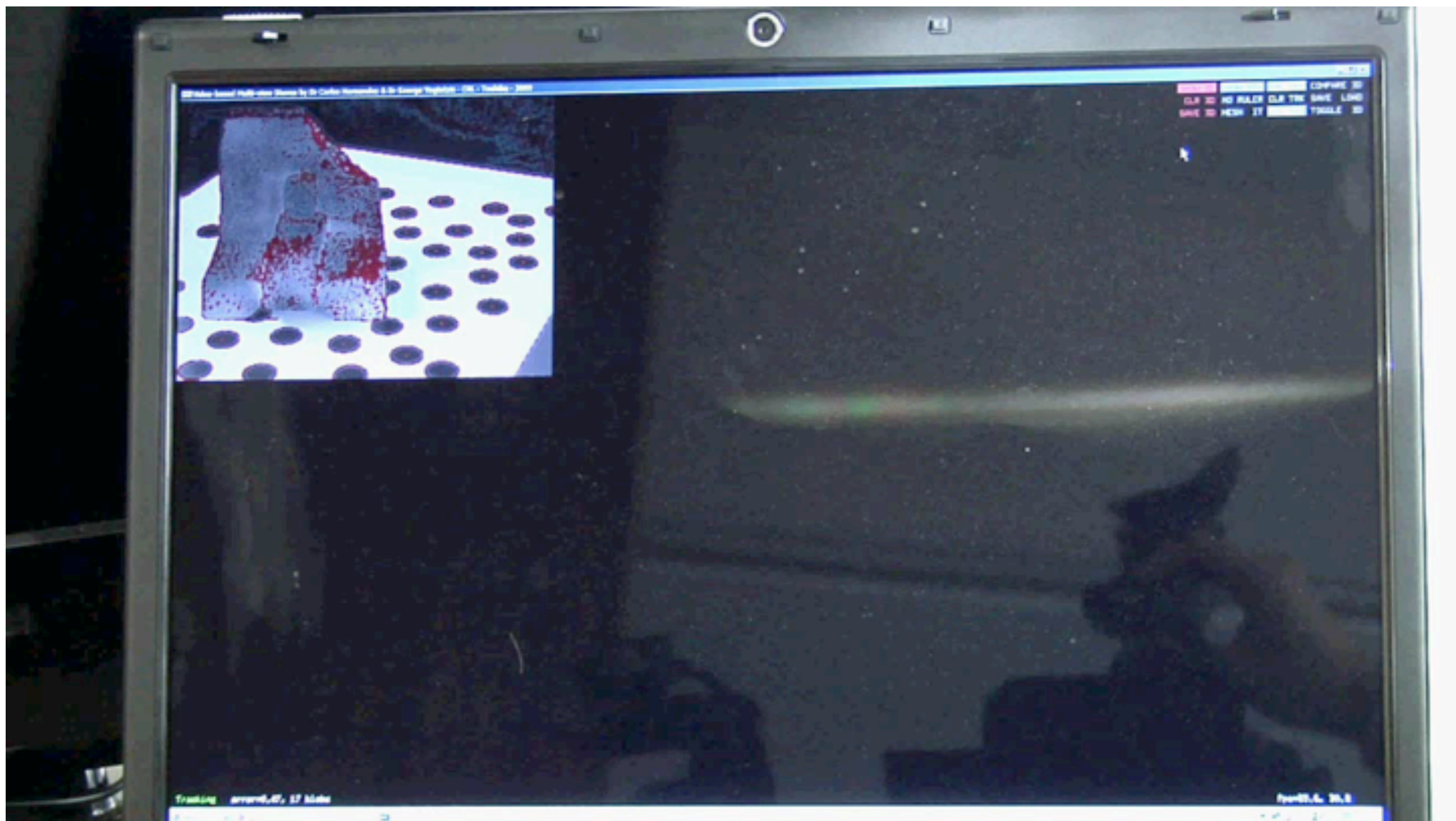
- Initialise a number of pixel depth sensors in first frame
- For every new incoming frame
  1. Measure pixel depth for each sensor
  2. Update (Z, π) posteriors using measurements
  3. Remove sensors whose expected inlier ratio drops below a threshold
  4. Convert into 3d points sensors whose posterior depth variance drops below a threshold
  5. Replace removed or converted sensors by new ones on current frame

Aston University
Birmingham

Interactive Multi-view stereo

Benefits:

- Feedback leads to better models
- Still passive & cheap



Aston University
Birmingham

Evaluation

Compared against [Merrel'08]

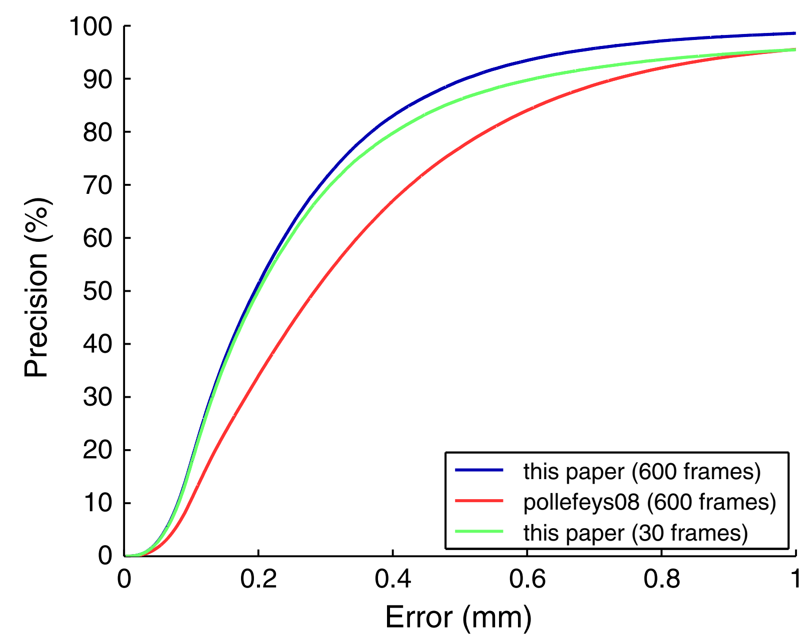
Pending a more thorough evaluation with [Newcombe '10] and others

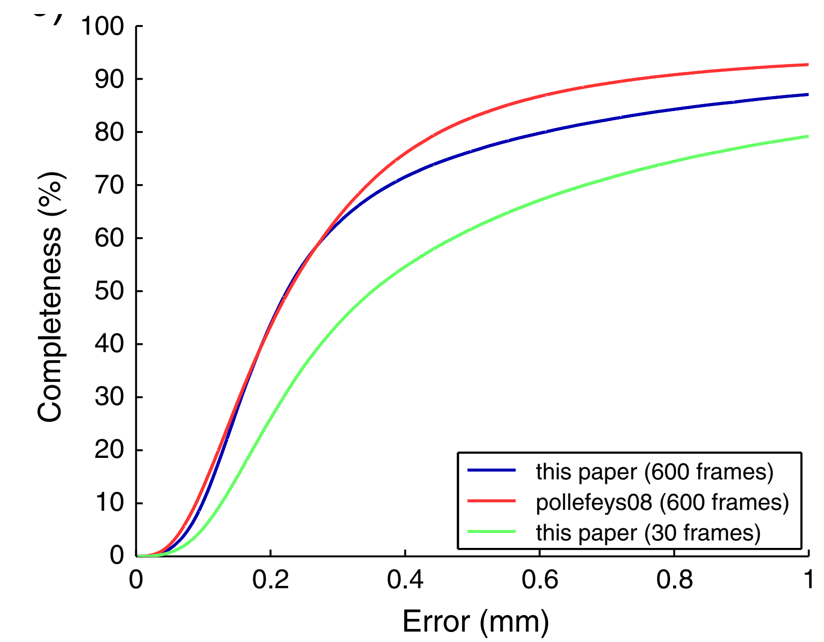
Merrel '08

Our system with 30 video frames

Our system with 600 video frames

Ground-truth





- Less complete than independent depth-maps [Merrel '08] but
- More accurate

Aston University
Birmingham

Video based MVS

Video-based, real-time multi-view stereo
Vogiatzis and Hernández, Image and Vision Computing, 29 (7), p.434-441, Jun 2011

TOSHIBA
Leading Innovation >>>

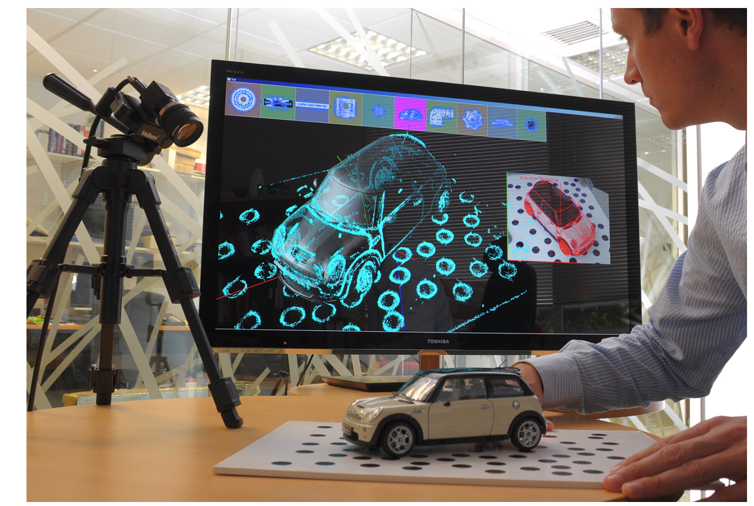


Figure 1. **Our system** in use. The user rotates the object on the pattern, and it is reconstructed on screen (cyan dots). After object inference the recognized objects are overlaid on the point cloud and on the video output.

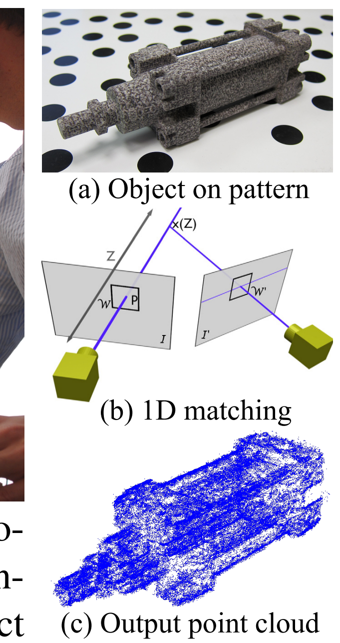


Figure 2. **Geometry capture** key steps.

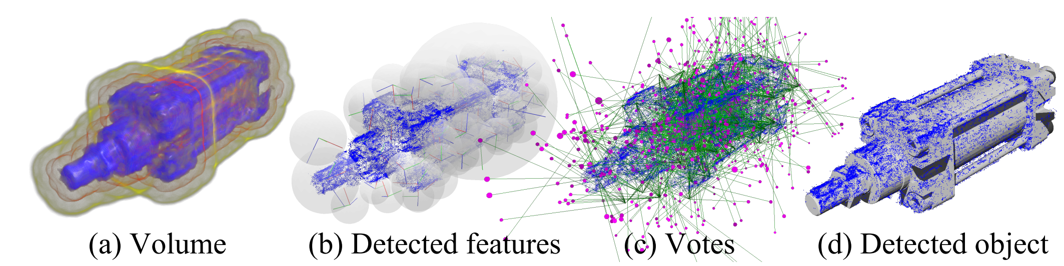


Figure 3. **Object inference** key steps.

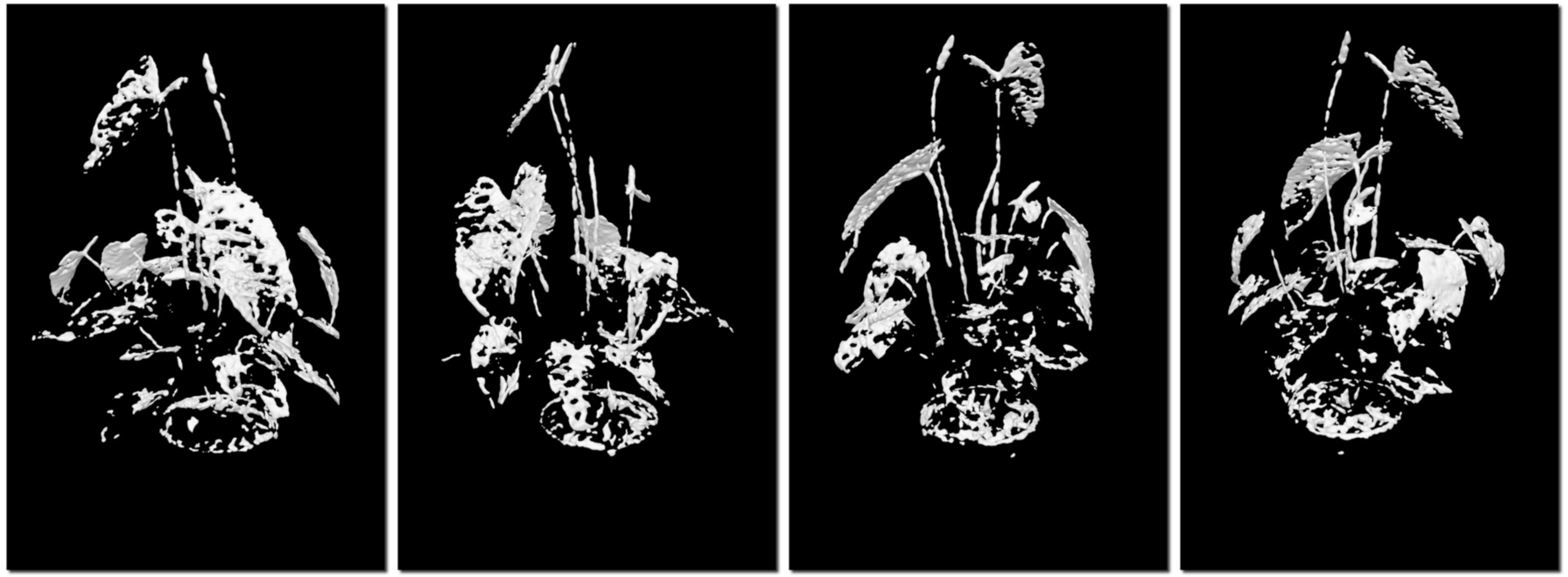
Aston University
Birmingham

Outline

- 3d vision for capturing 3d shape
  - Applications, mature technologies and their limitations Video-based multi-view stereo
- Automatic calibrated multi-segmentation
- Face capture with multi-spectral photometric stereo

Aston University
Birmingham

Textureless objects



Furukawa ’07

Silhouettes

- Shape-from-silhouettes can
  - handle lack of texture
  - improve MVS results
    - Outer bound [Vogiatzis '05]
    - Occlusion reasoning [Kolev '11, Hernandez '04]

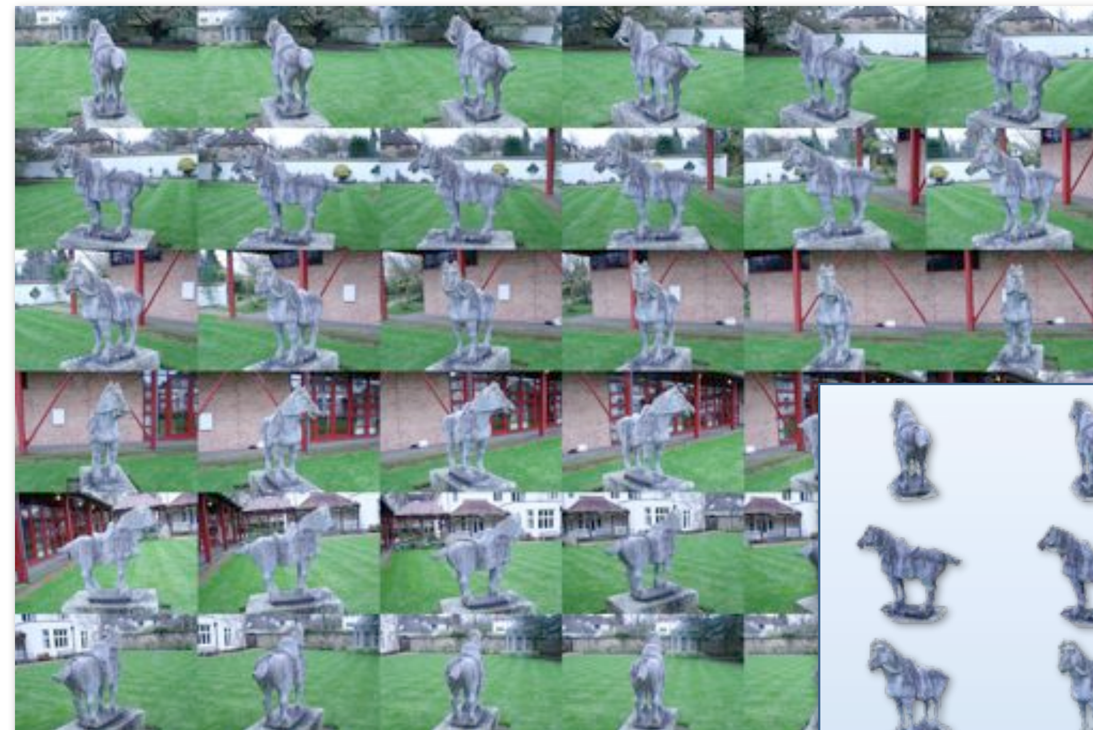
- Images in the 'real world'
- Perform segmentation automatically

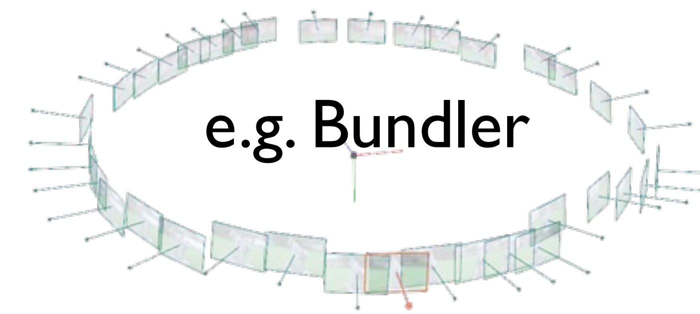
- Large number of images
- Avoid per-image interaction
  - Bounding boxes/brush strokes

Aston University
Birmingham

Our task

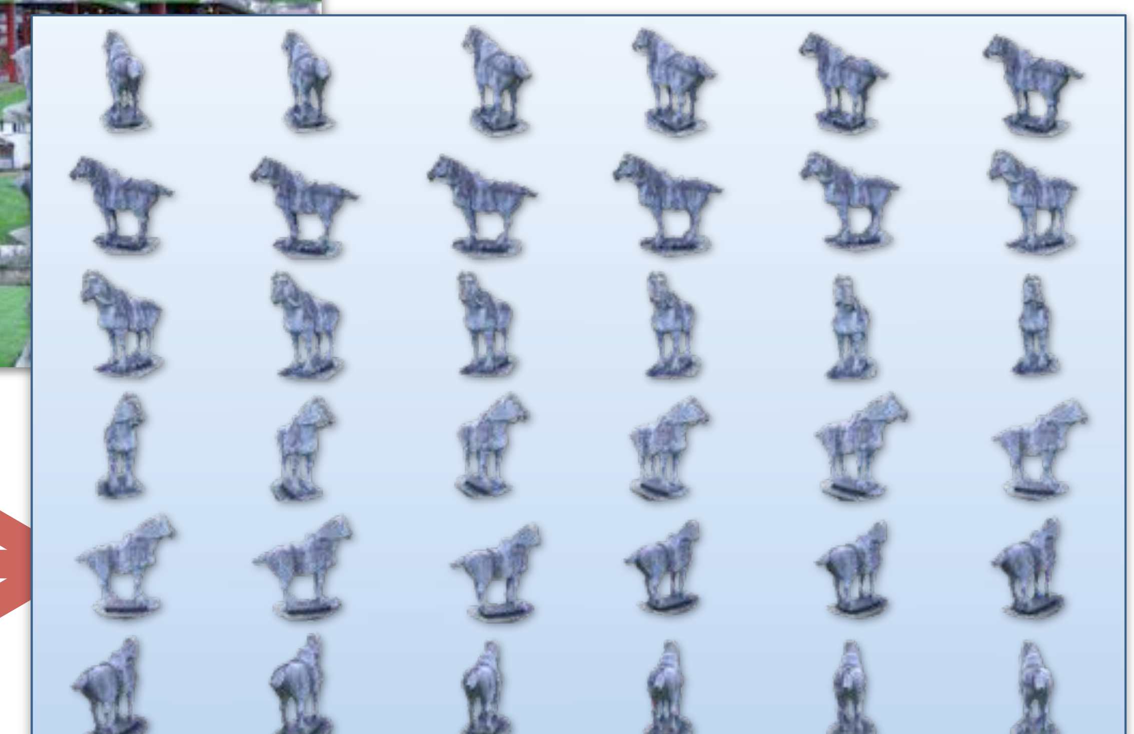
- From a set of pose-calibrated images
- automatically obtain silhouettes of a rigid object





Have Camera Calibration

Segment Automatically

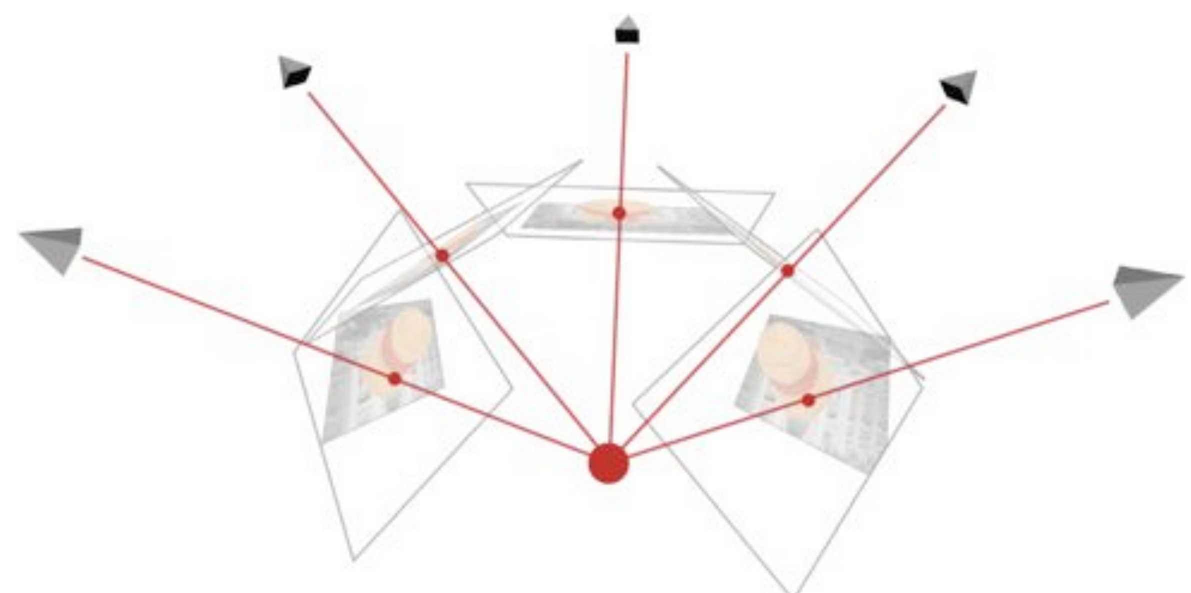


Aston University
Birmingham

Segmentation Constraints

Campbell '07, Lee '07

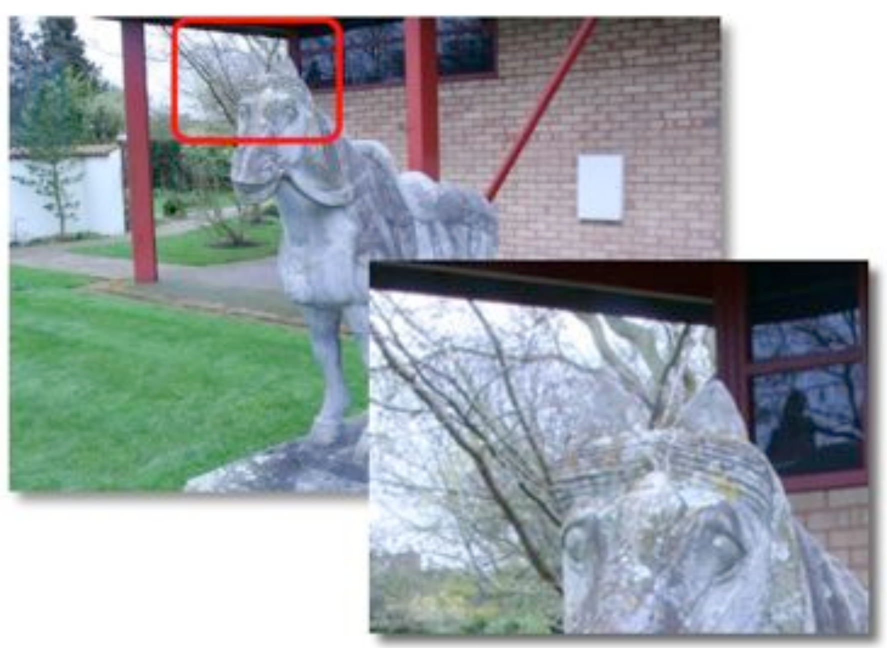
- Silhouette coherence
  - visual hull projection must maximally fill the silhouettes



- Fixation constraint
  - Object of interest is at centre of images
- Appearance consistency
  - FG and BG have their own colour model [Grabcuts]

Segmentation Constraints

- Limitations [Campbell ’07, Lee ’07]
  - Generative appearance model
    - Gaussian Mixture model in colour space
  - FG/BG not separable in the space
  - Silhouette coherency alone not sufficient



Input Images (6 of 36)

Result of **[Campbell *et al.* 2007]**

Aston University
Birmingham

‘Weak’ stereo for multi-Segmentation

Campbell et al, CVMP 2012

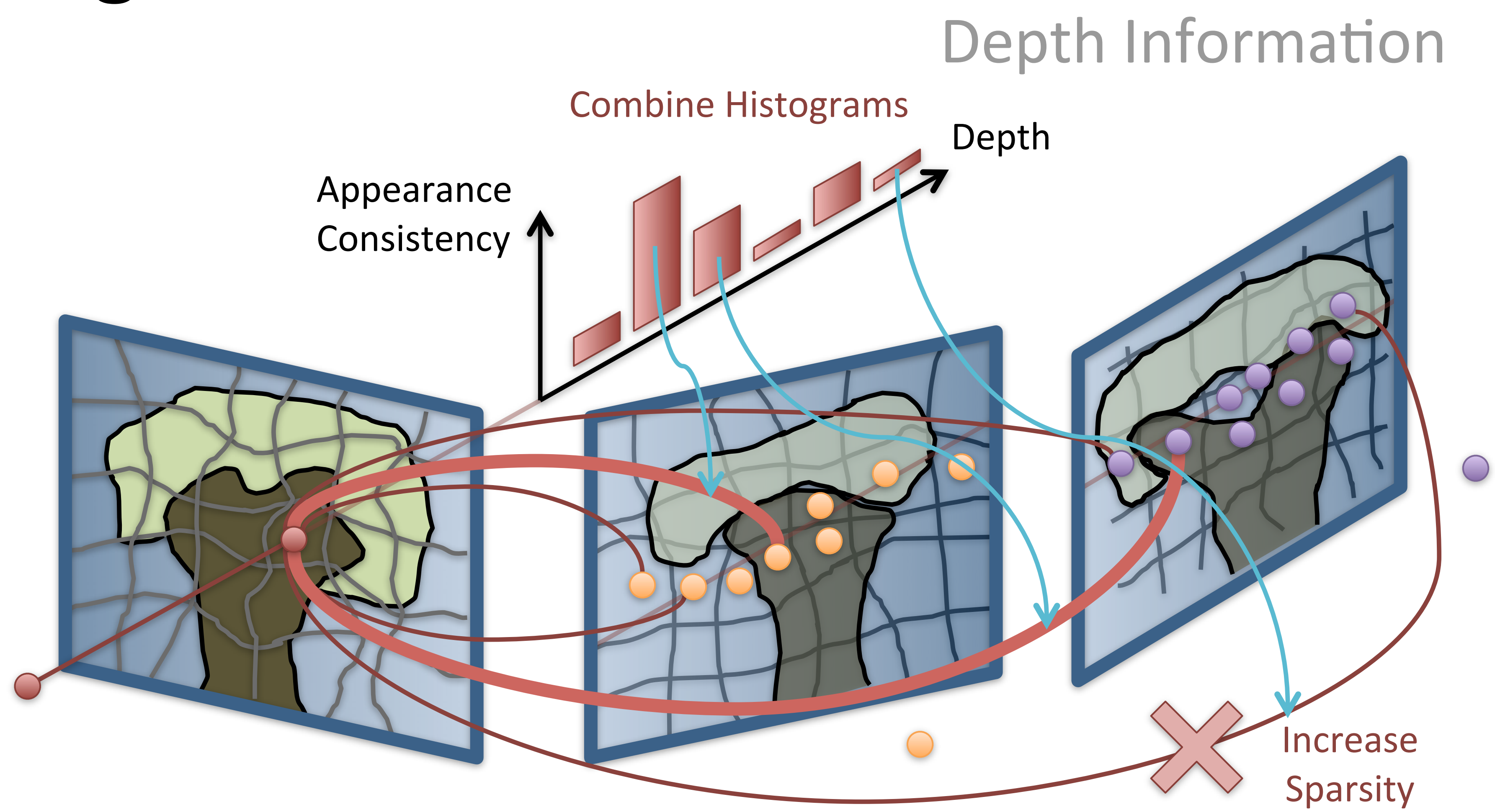
- Quantize images into super-pixels (Turbopixels [Levinshtein ’09])
- Label each super-pixel as FG/BG using Maxflow/Mincut
  - *Unary term:* colour model
  - *Pairwise term:* pixels encouraged to have same label if
    - they have similar colour
    - they obey epipolar constraint
    - other similar superpixels vote for same depth

    } ‘Weak’ stereo
- Iterate while
  - enforcing silhouette coherence,
  - refining colour models

Aston University
Birmingham

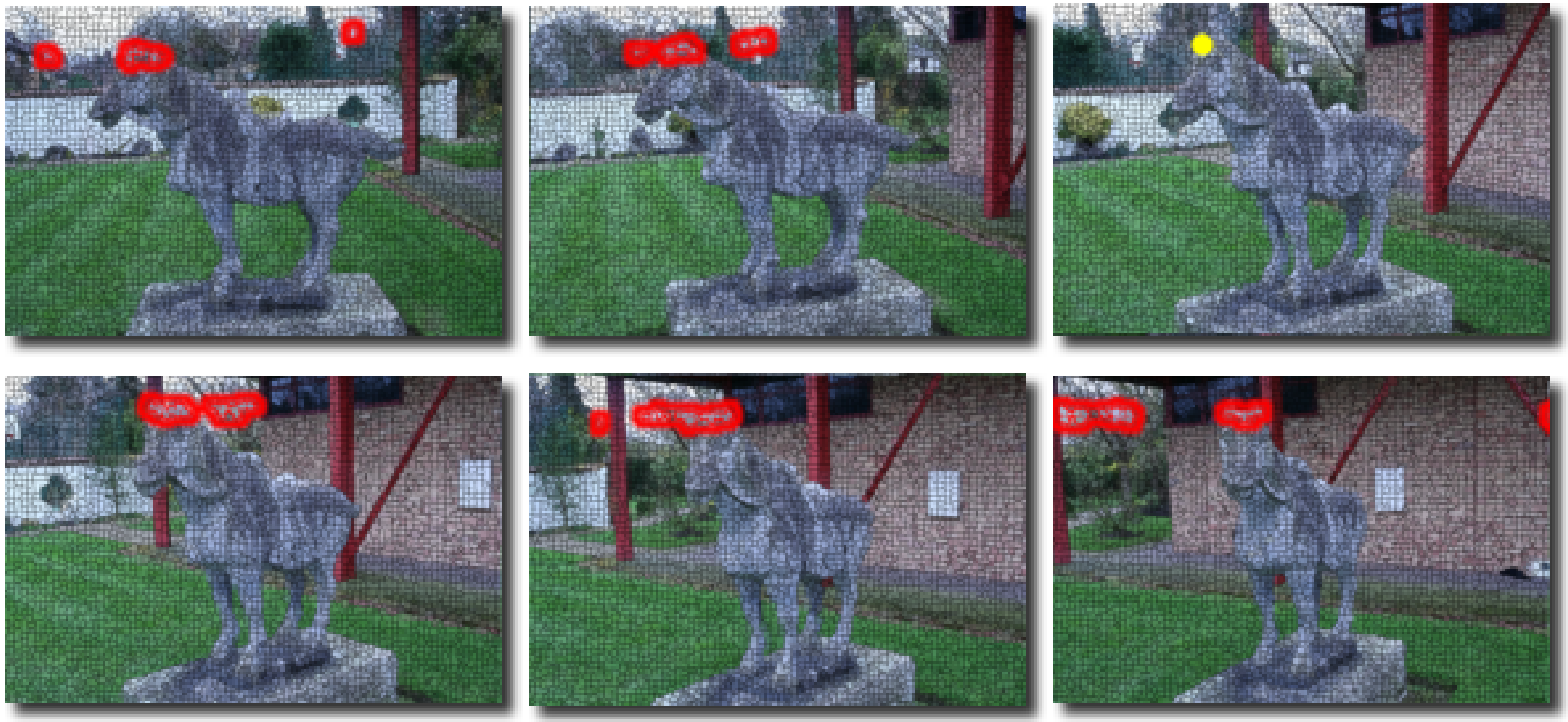
Creating the graph

Algorithm



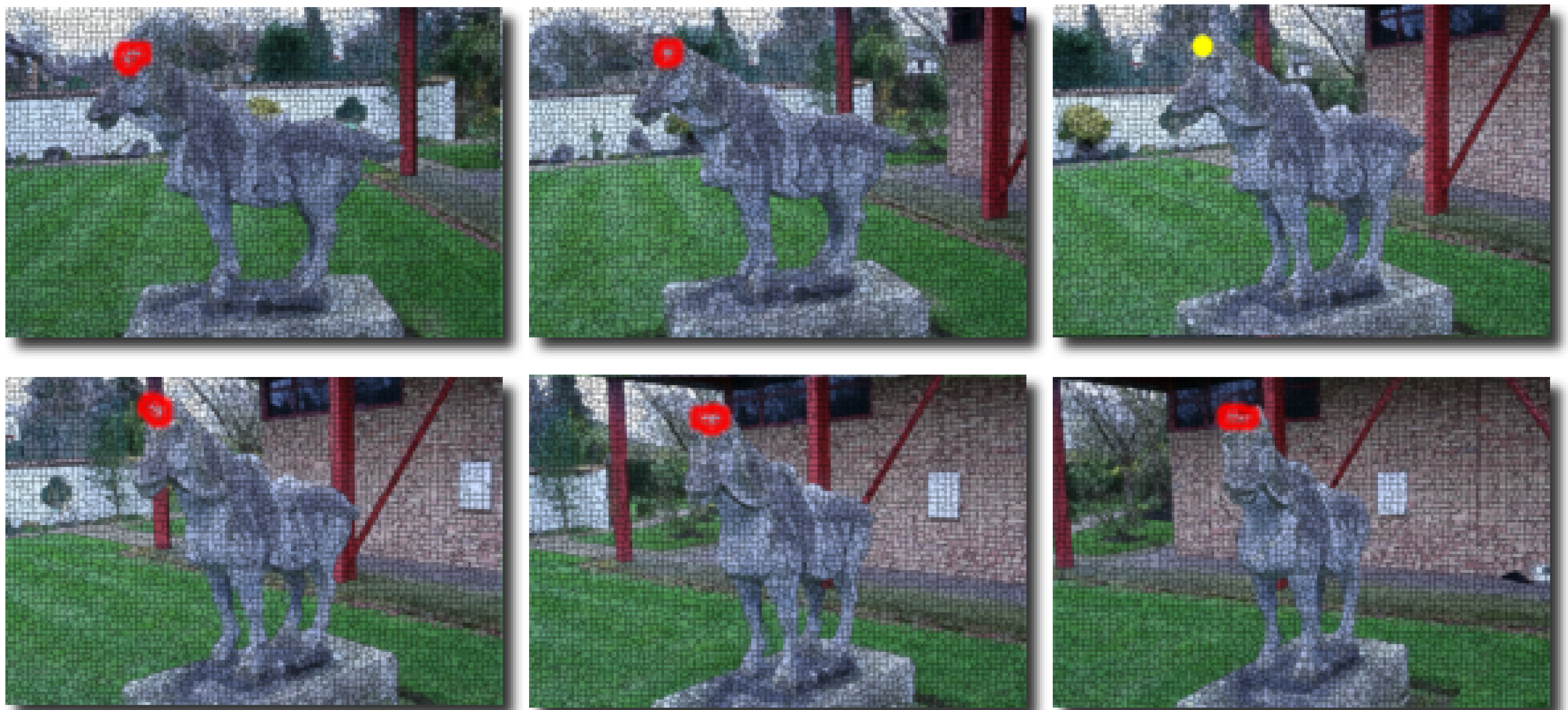
- 'Soft' depth information from weak superpixel stereo
- Build histogram (with outlier model)

Edge connections



Without depth (appearance only)

Edge connections



Depth and appearance

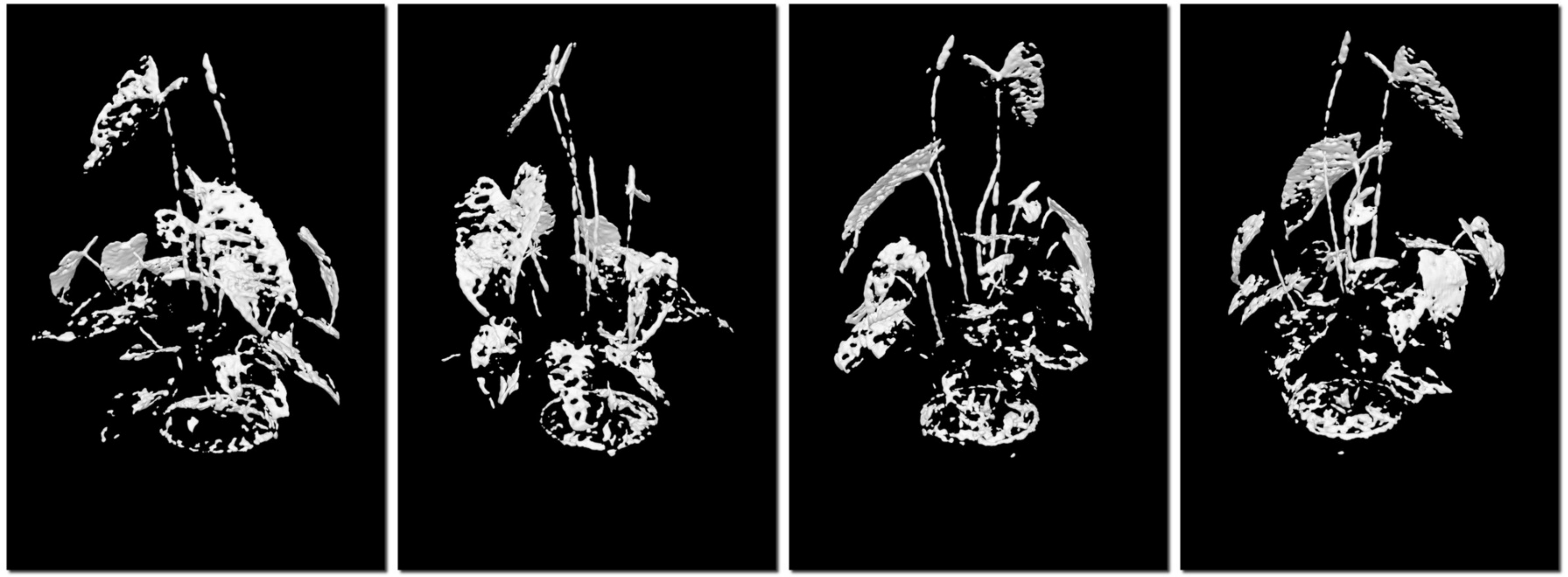
Aston University
Birmingham

Results-horse

Our Result

Head and tail recovered

Results-plant



Furukawa ’07

Aston University
Birmingham

Results-plant

Campbell '11

Aston University
Birmingham

Automatic Calibrated Multi-Segmentation

- Successful automatic Multi-View segmentation
- Iterative algorithm
  - Segments all images simultaneously
- Improve over existing methods
  - Addition of Depth Information
  - Enforce additional constraints

Computation Time (Matlab):

- Super-pixels: 60s / image
- W matrix: 120s
- Graph-cut iteration: 7s

Aston University
Birmingham

Outline

- 3d vision for capturing 3d shape
  - Applications, mature technologies and their limitations
  - Video-based multi-view stereo
- Automatic calibrated multi-segmentation
- Face capture with multi-spectral photometric stereo

Aston University
Birmingham

Colour photometric stereo

- Original idea proposed in 80s [Petrov 87] & [Woodham94]
- In [Hernandez 07] we applied it on moving objects of constant albedo
- Leads to very simple / low cost setup

If a white object is illuminated by a red, a green and a blue light source, the color reflected by a point on the surface is in 1-1 correspondence with the local orientation.

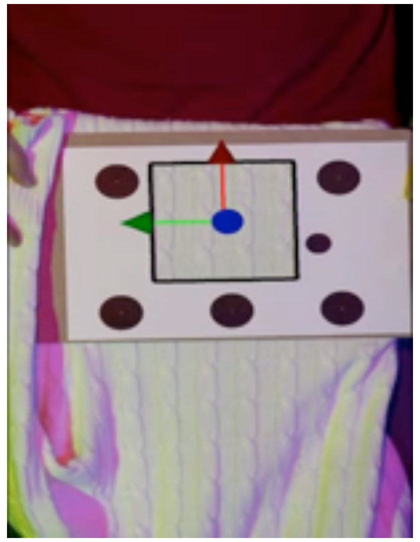
A. Petrov. Light, color and shape. Cognitive Processes and their Simulation (in Russian), pages 350-358, 1987

Calibration of photometric stereo

- Must estimate light-directions and intensities
  - Can be seen as mapping between (surface orientation, albedo) and pixel intensity profile in all images
- Can estimate light directions with complex mirror setup

- Can also fit mapping to known data points [Hertzman04]

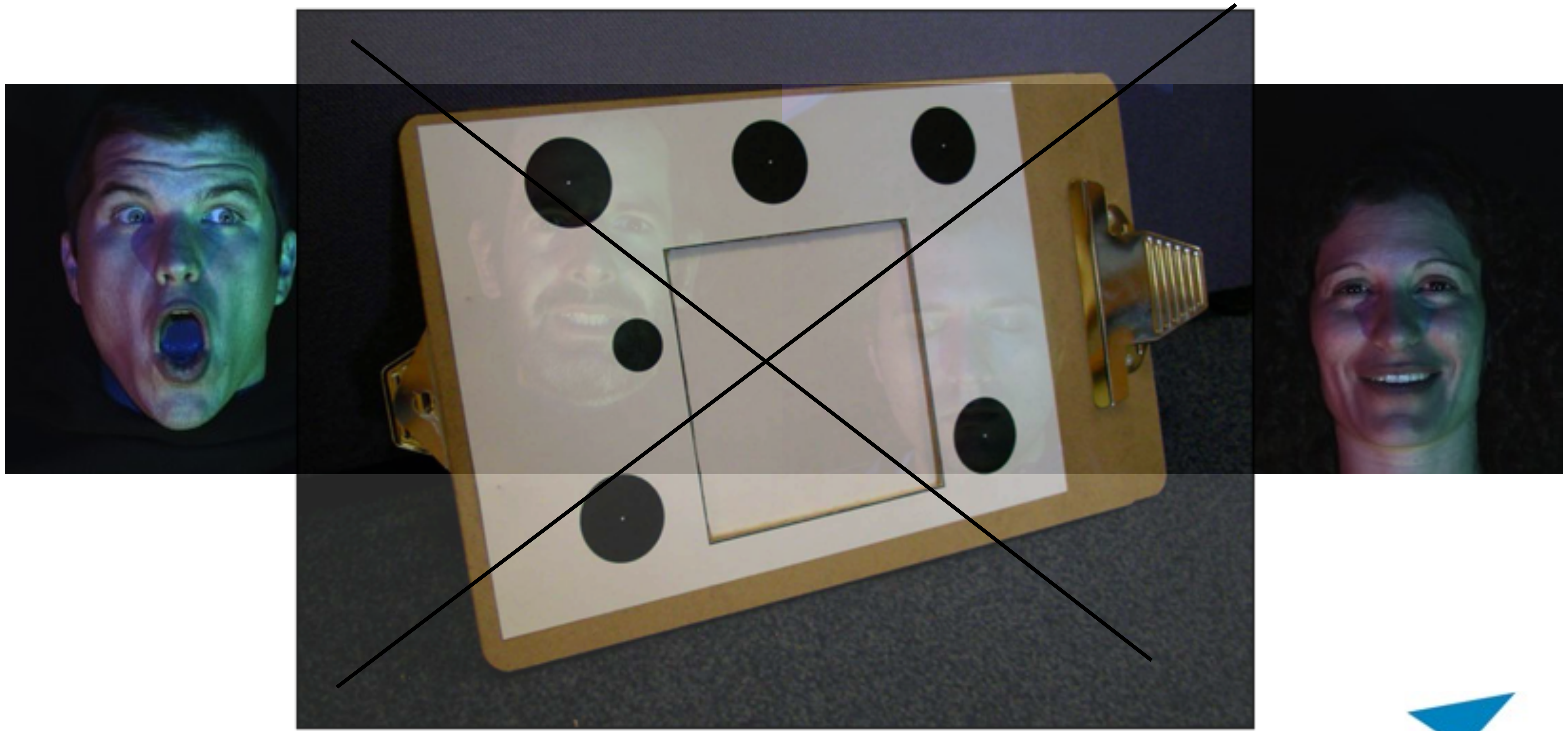
- Colour Photometric Stereo
  - Estimate mapping: RGB space → normal space
    [Patterson05] [Hernandez07]
    Material Dependent!



Aston University
Birmingham

Colour photometric stereo for faces

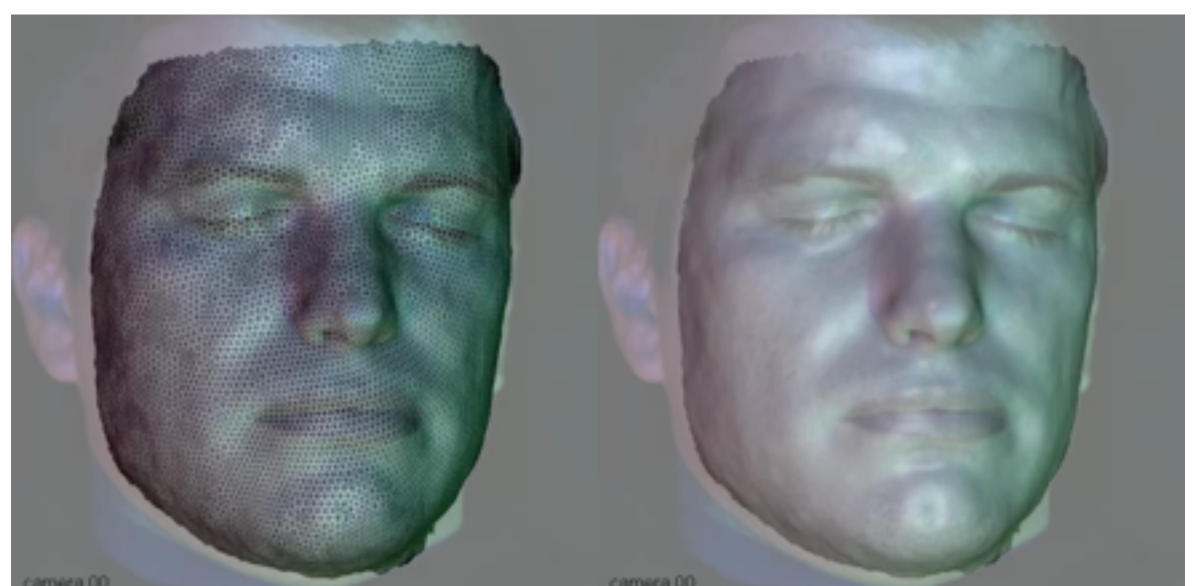
- Examples of faces captured using the colour photometric stereo setup



Aston University
Birmingham

Colour photometric stereo for faces

- ...but we can construct a partial & noisy example object using Multi View Stereo



Aston University
Birmingham

Mono-chromaticity

- Face consists of multiple shades of same colour.
- This leads to

$$\mathbf{c} = \mathbf{V} \cdot \mathbf{L}\rho\mathbf{n}.$$

where c is the RGB triplet, L is the matrix of light directions and V is the colour 'mixture' matrix

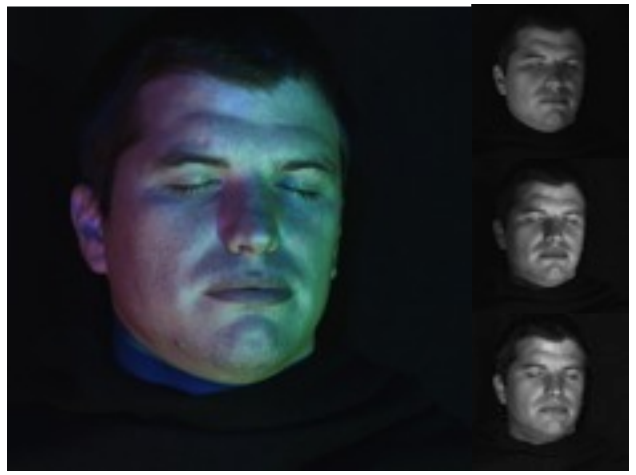
- we wish to estimate this mapping but do not know which data points we can use!

Aston University
Birmingham

Robust model fitting

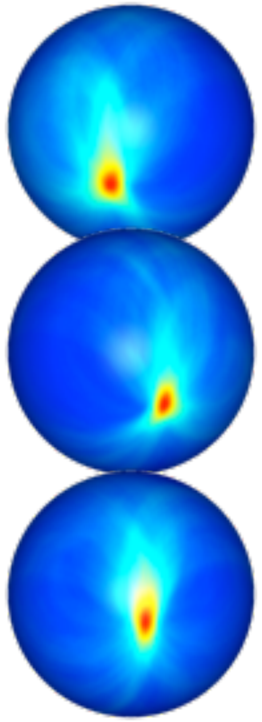
$$\mathbf{c} = \mathbf{V} \cdot \mathbf{L}\rho\mathbf{n}.$$

- Since we don't really care about monochromatic albedo ρ
- treat ρ**n** and **c** as vectors only defined up to a scaling factor.
- L*V maps from a 2d projective space to a 2d projective space
- This is just a 2d Homography!
  - use your favourite RANSAC + nonlinear fit!



Sample input image

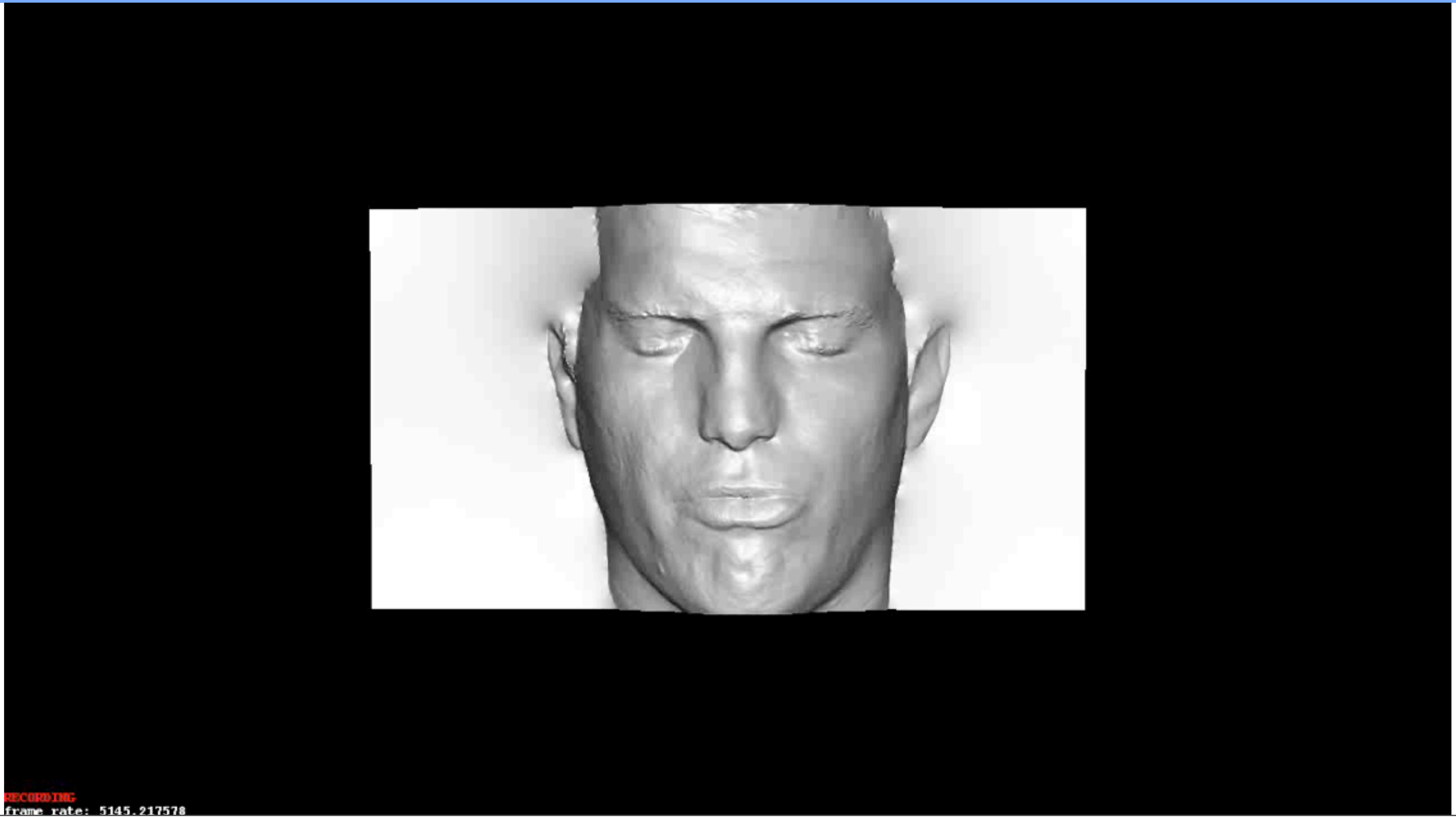
Inliers



Consensus w.r.t light direction

Aston University Birmingham

Facial expressions



Conclusion

- Colour photometric stereo for faces *[Vogiatzis & Hernandez IJCV 2011]*
  - Photometric stereo gives lifelike detail, but low frequency shape is not as good
  - Combine with depth sensor (see *[Anderson et al 2011]*)
  - Some faces are remarkably Lambertian, others are not
  - The single albedo chromaticity assumption works well in practice
  - **Deformable surface registration** must be part of mocap solution. Some solutions exist but all with weaknesses

- Calibrated Multi-Segmentation *[Campbell et al CVMP 2012]*
  - **Can we extend the "weak shape-from-X" idea** to other algorithms?

- Video based MVS *[Vogiatzis & Hernandez IVC]*
  - **Building higher level models**: is important for many users

- Thank you

- more in http://george-vogiatzis.org

Collaborators:

Gabe Brostow
Neill Campbell
Roberto Cipolla
Carlos Hernandez
Bjorn Stenger
Oliver Woodford

TOSHIBA Leading Innovation >>>
UNIVERSITY OF CAMBRIDGE
Aston University Birmingham