Computer vision at Cambridge

Roberto Cipolla

Department of Engineering

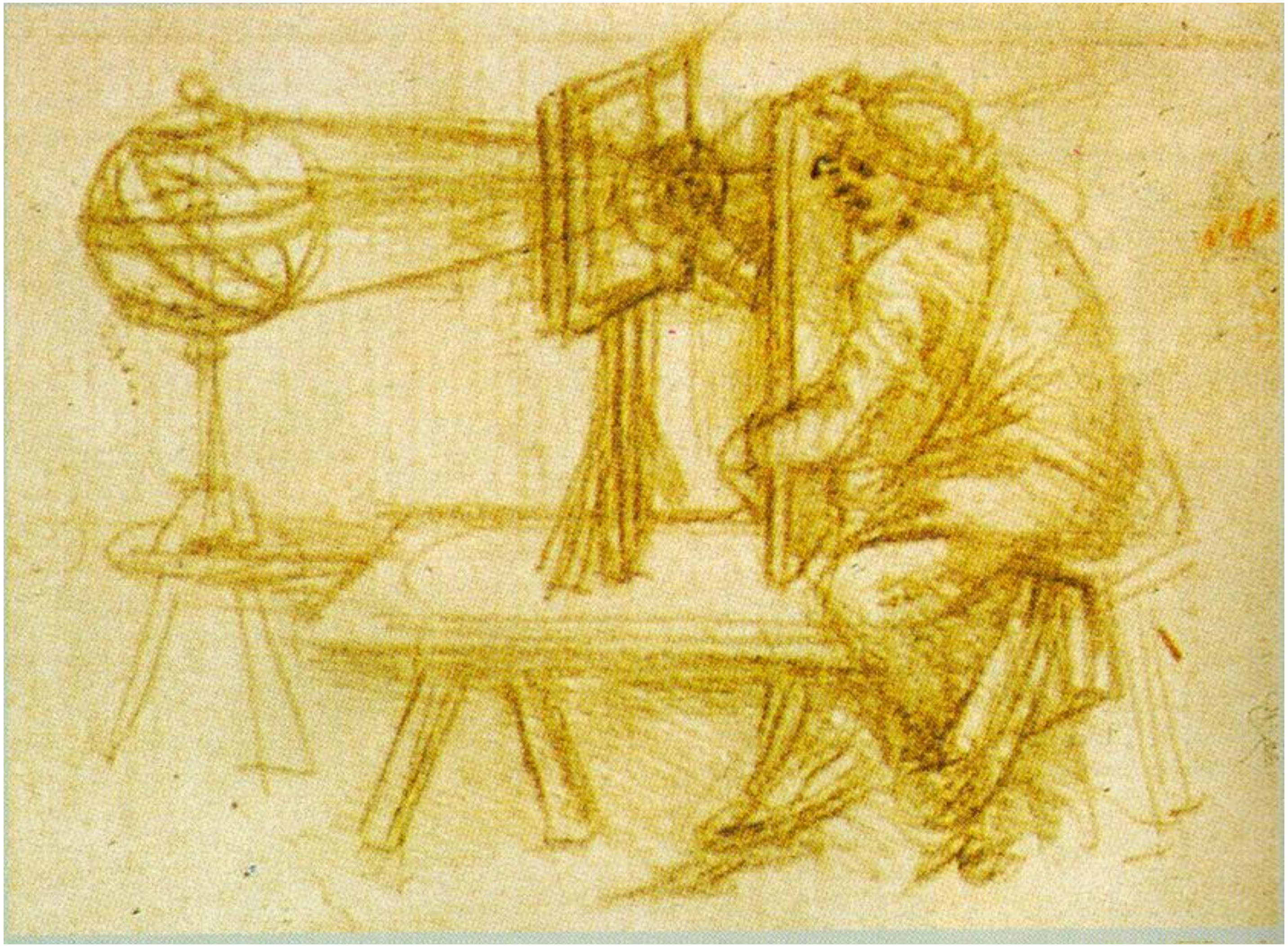
Research team

http://www.eng.cam.ac.uk/~cipolla/people.html

Part II:

Matching images and simple object recognition

Perspective



Overview – Part II

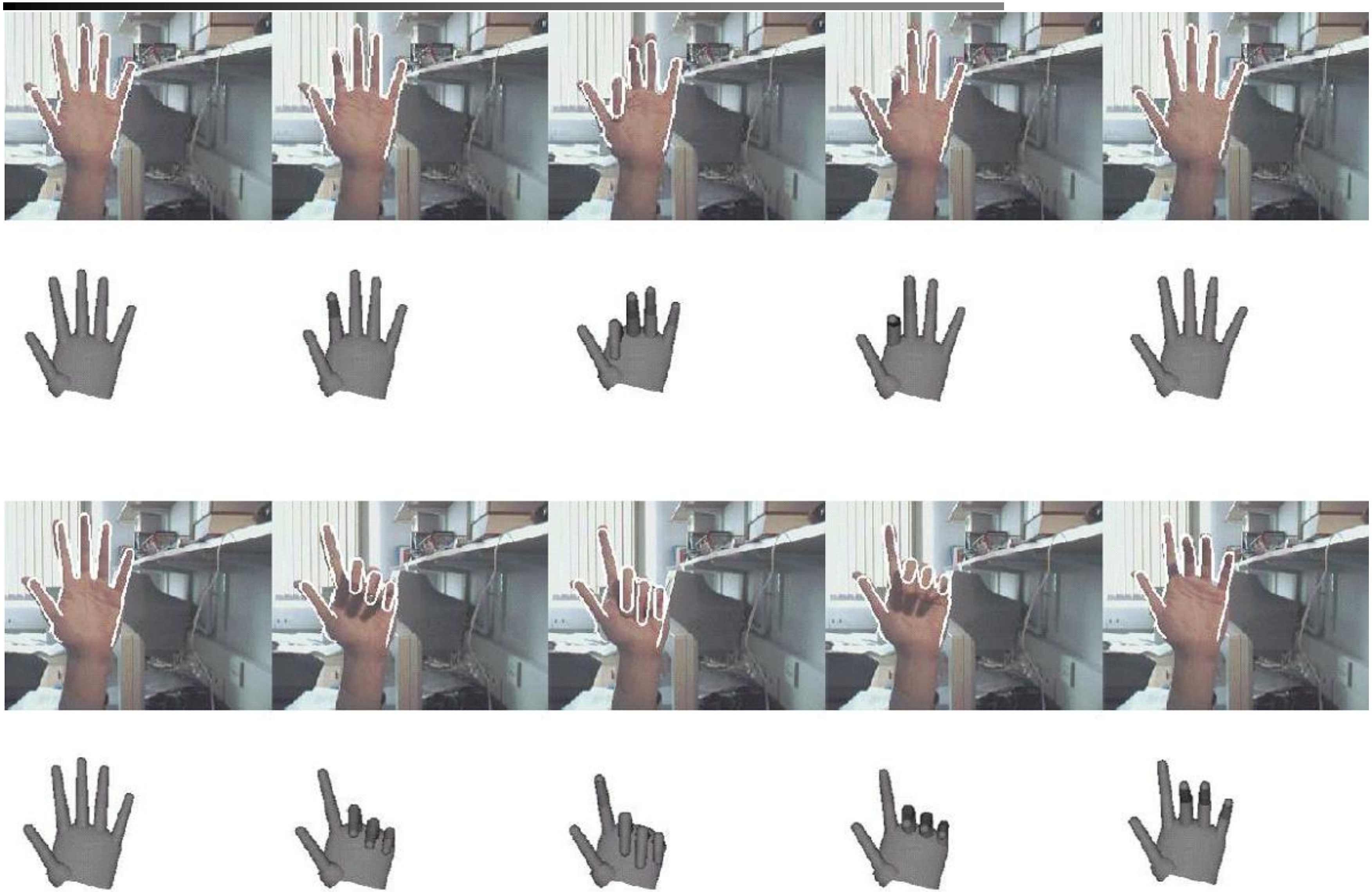
1. Image matching:

   Image matching and image-based localisation from a single photo.

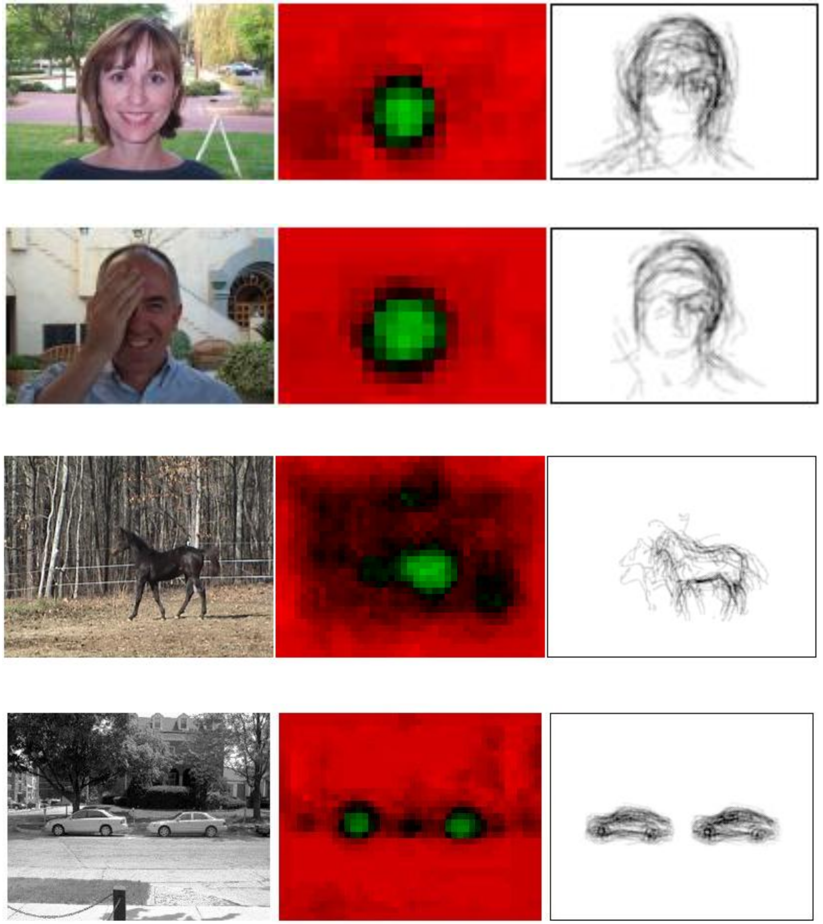
2. Object detection

Recognition of pictures

Hand detection



Object Detection



input | classification map | contours

1. Image matching

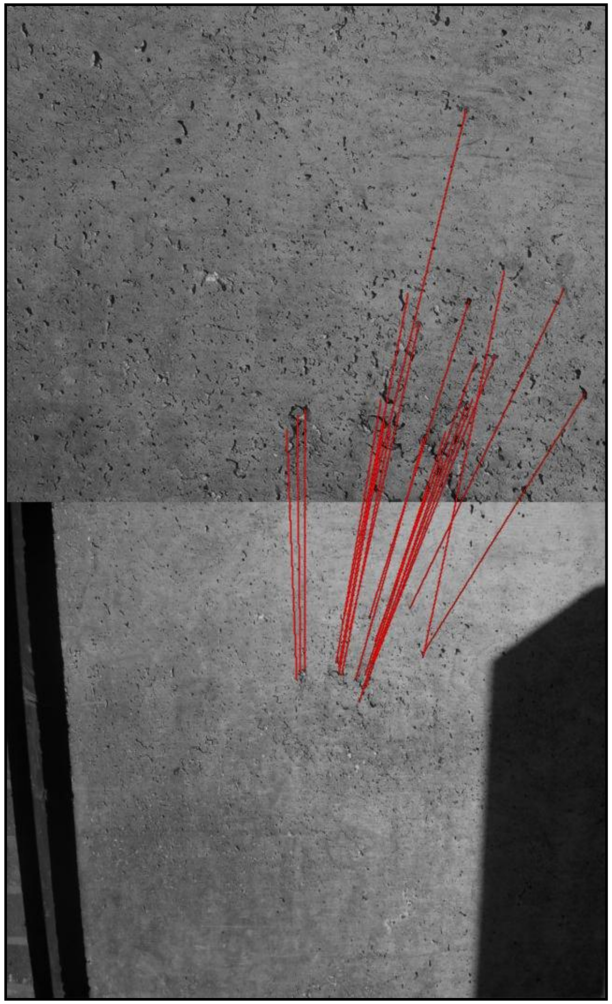
1. Image matching

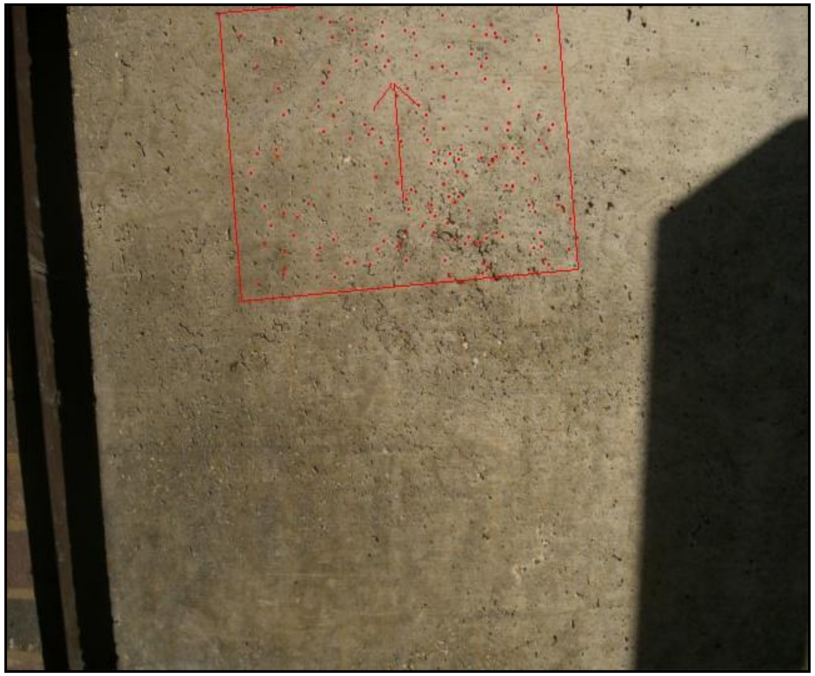
Demo – visual inspection

UNIVERSITY OF CAMBRIDGE

Demo – visual inspection

Matching concrete imags





Wide baseline matching

UNIVERSITY OF CAMBRIDGE

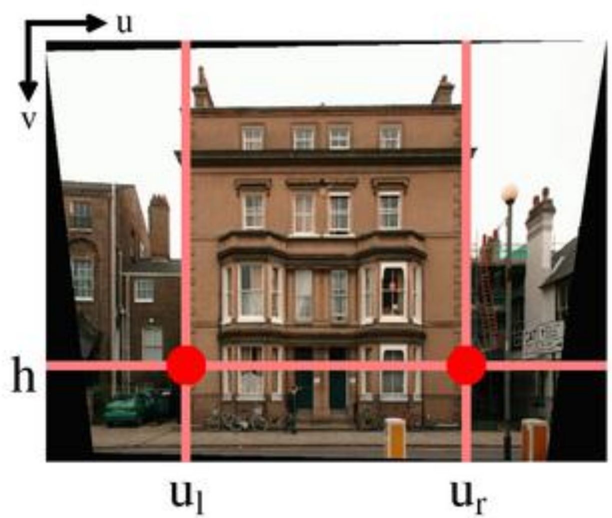
Image-based localisation

Where I am?

Determine pose from single image by matching

Register database view

First align database view to map



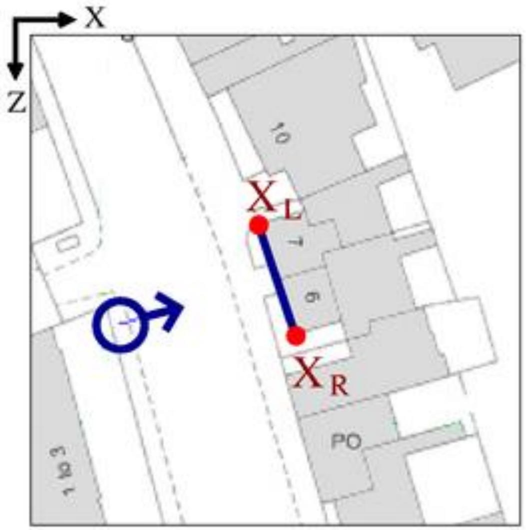


Image-based localisation

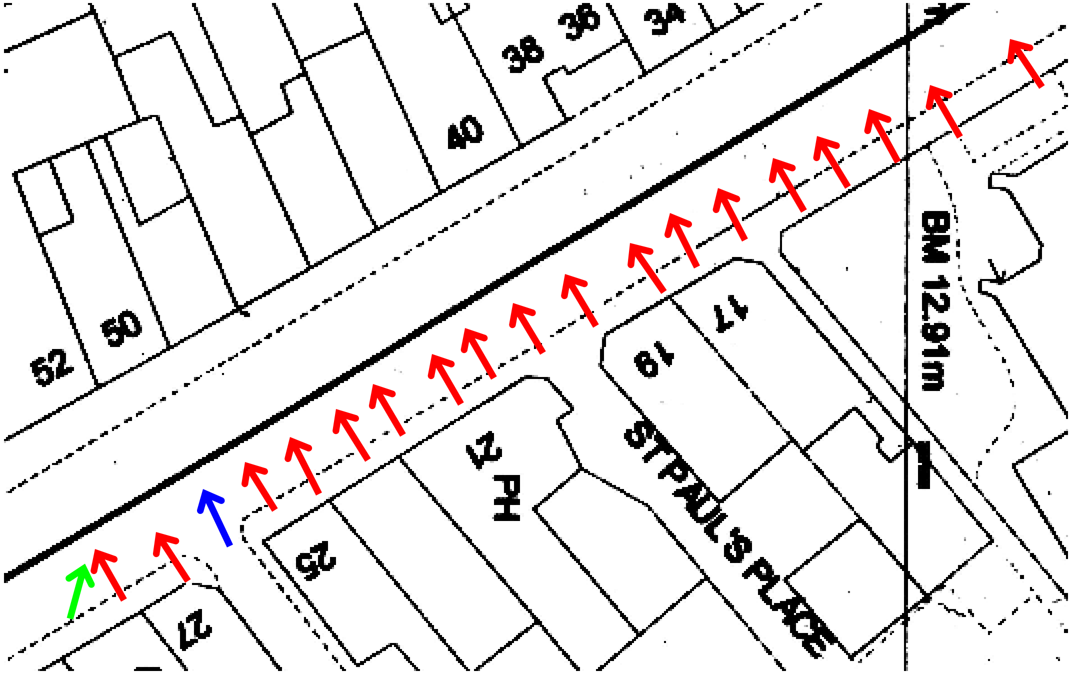
- Determine pose from single image
- Match to database
- Triangulate position

Image-based localisation

- Determine pose from single image
- Match to database
- Triangulate position

Localisation

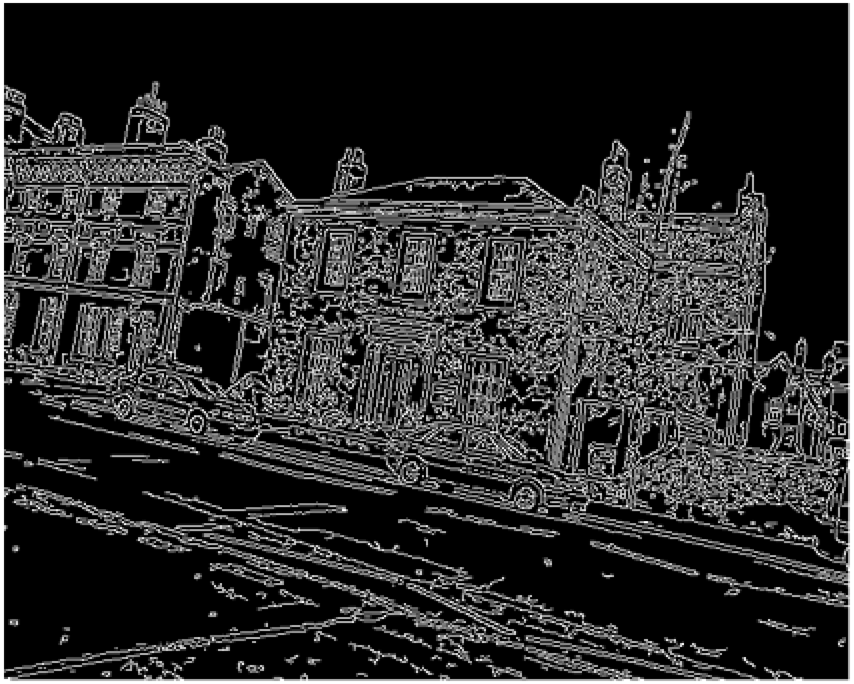
- Determine pose from single image
- Match to database
- Triangulate position

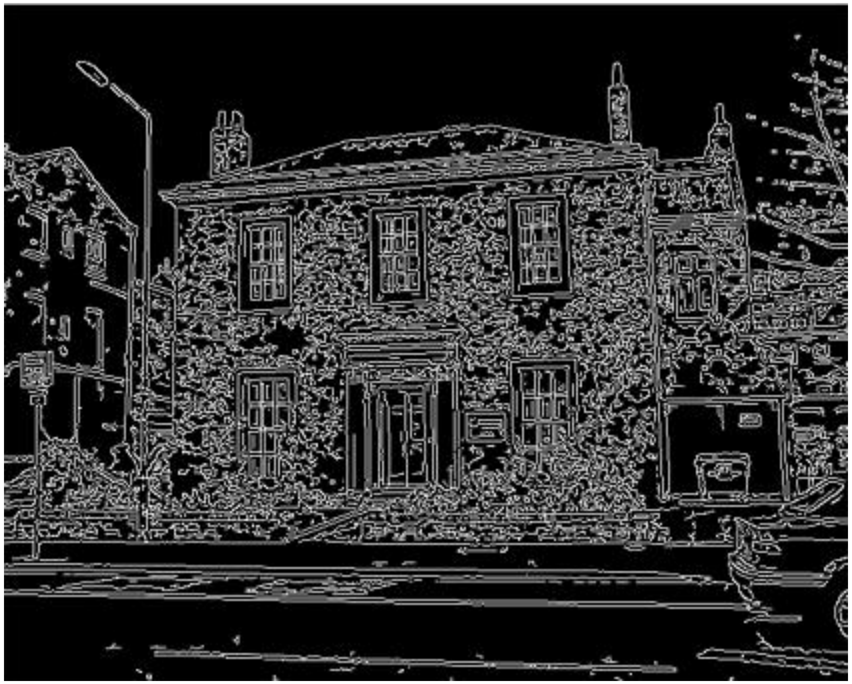


Constrained matching

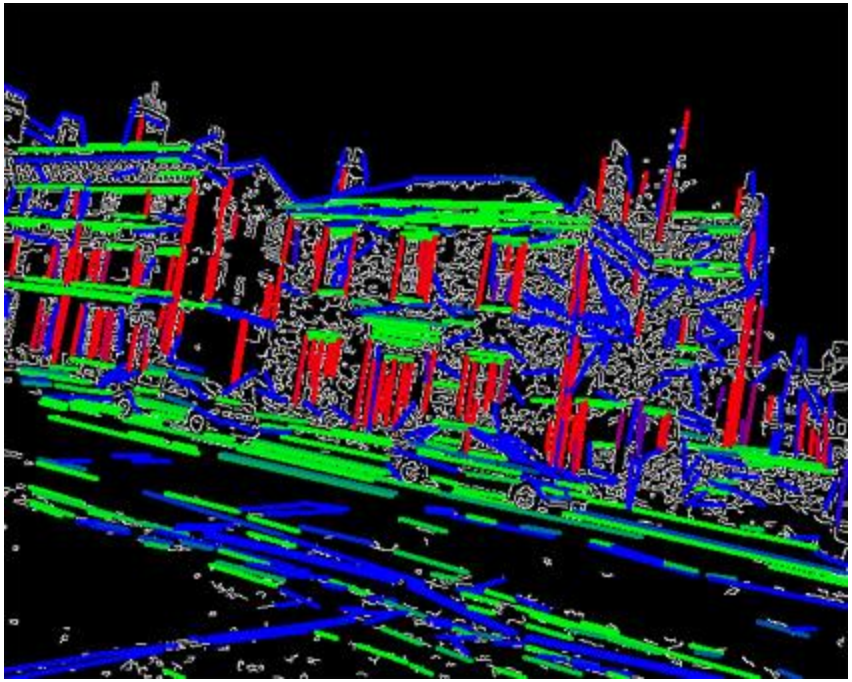
UNIVERSITY OF CAMBRIDGE

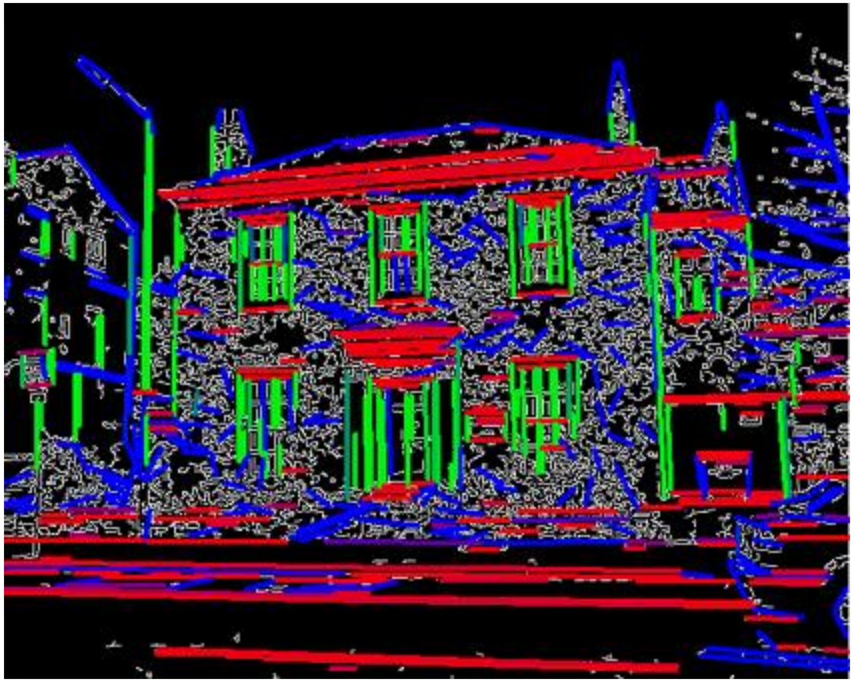
Constrained matching





Constrained matching





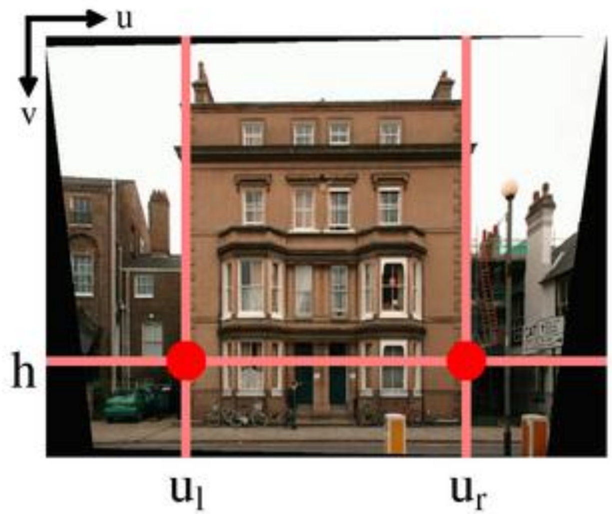
Constrained matching

Constrained matching

UNIVERSITY OF CAMBRIDGE

Register database view

First align database view to map



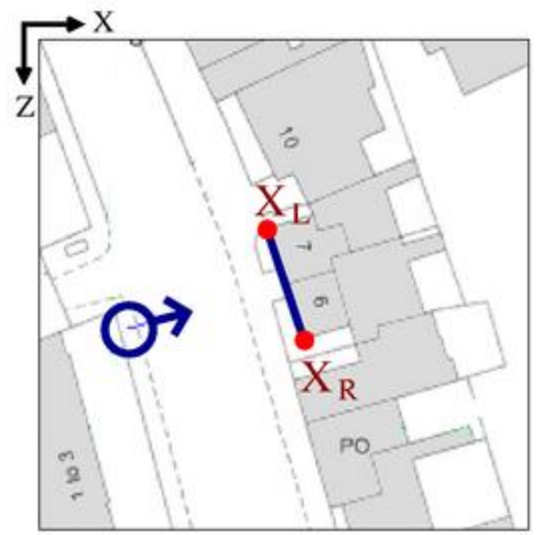


Image-based localisation

. . .

. . .

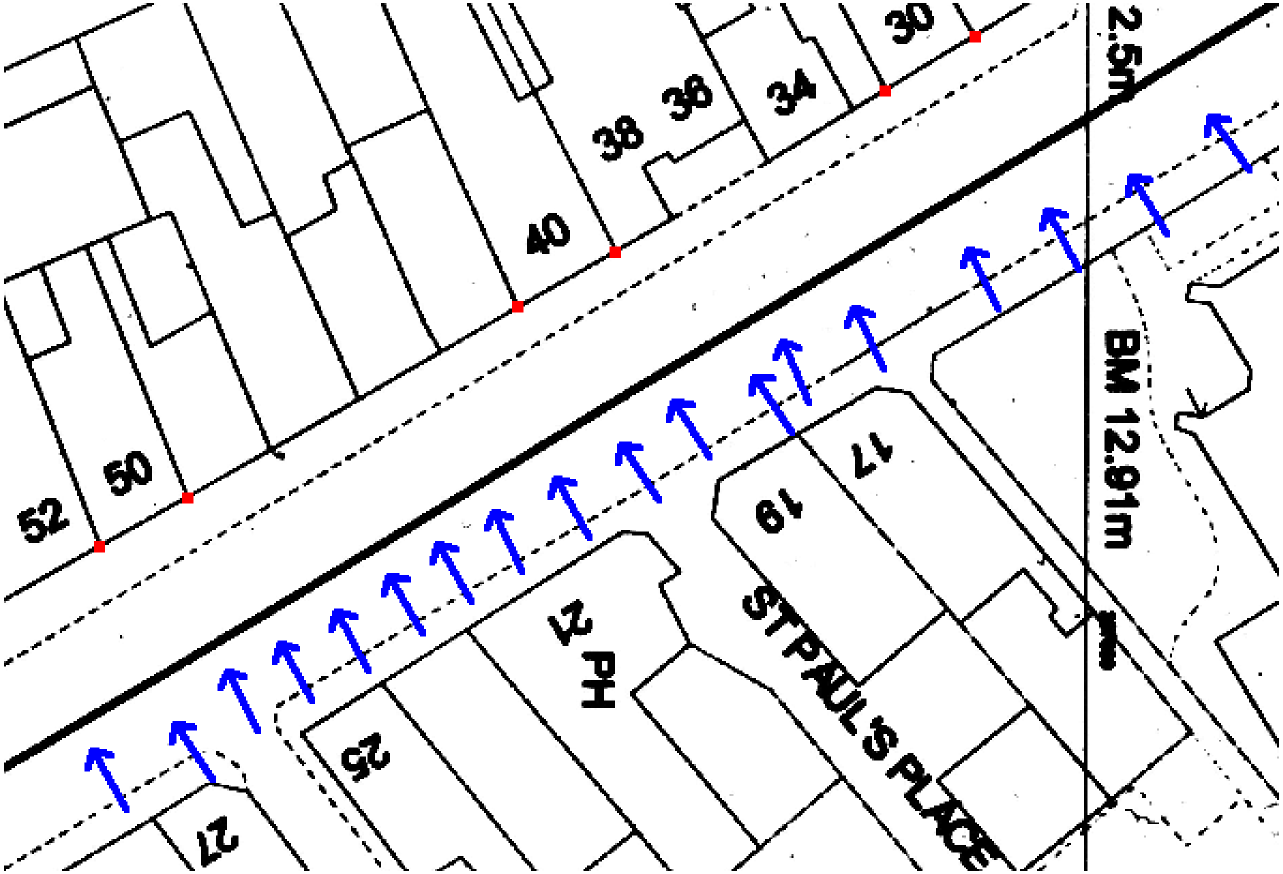


Image-based localisation

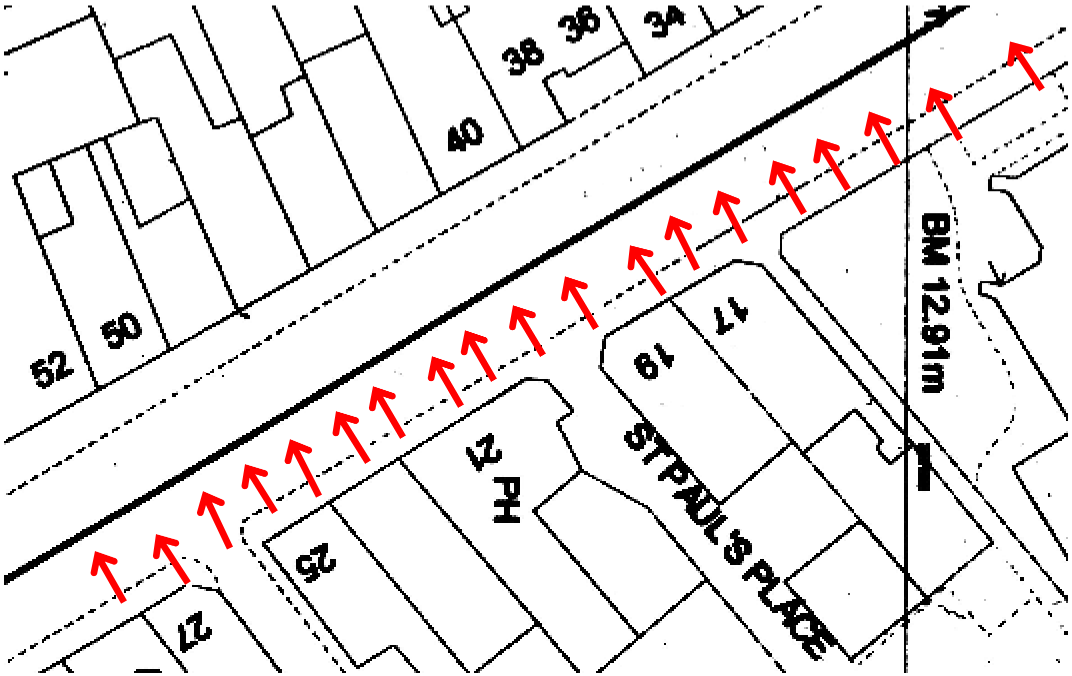
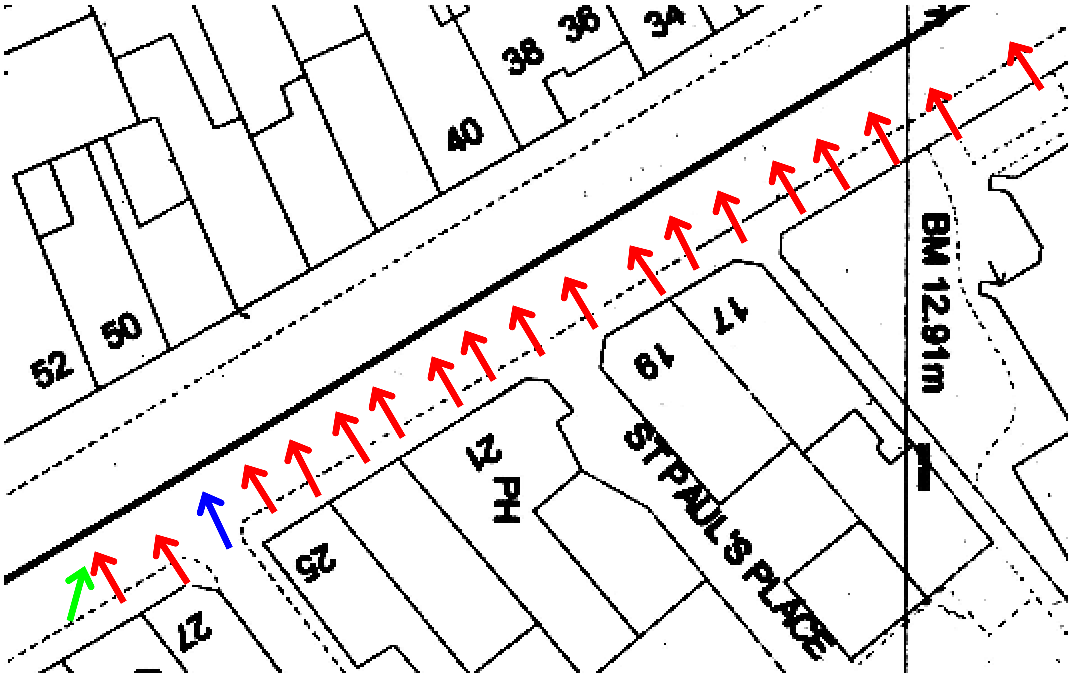


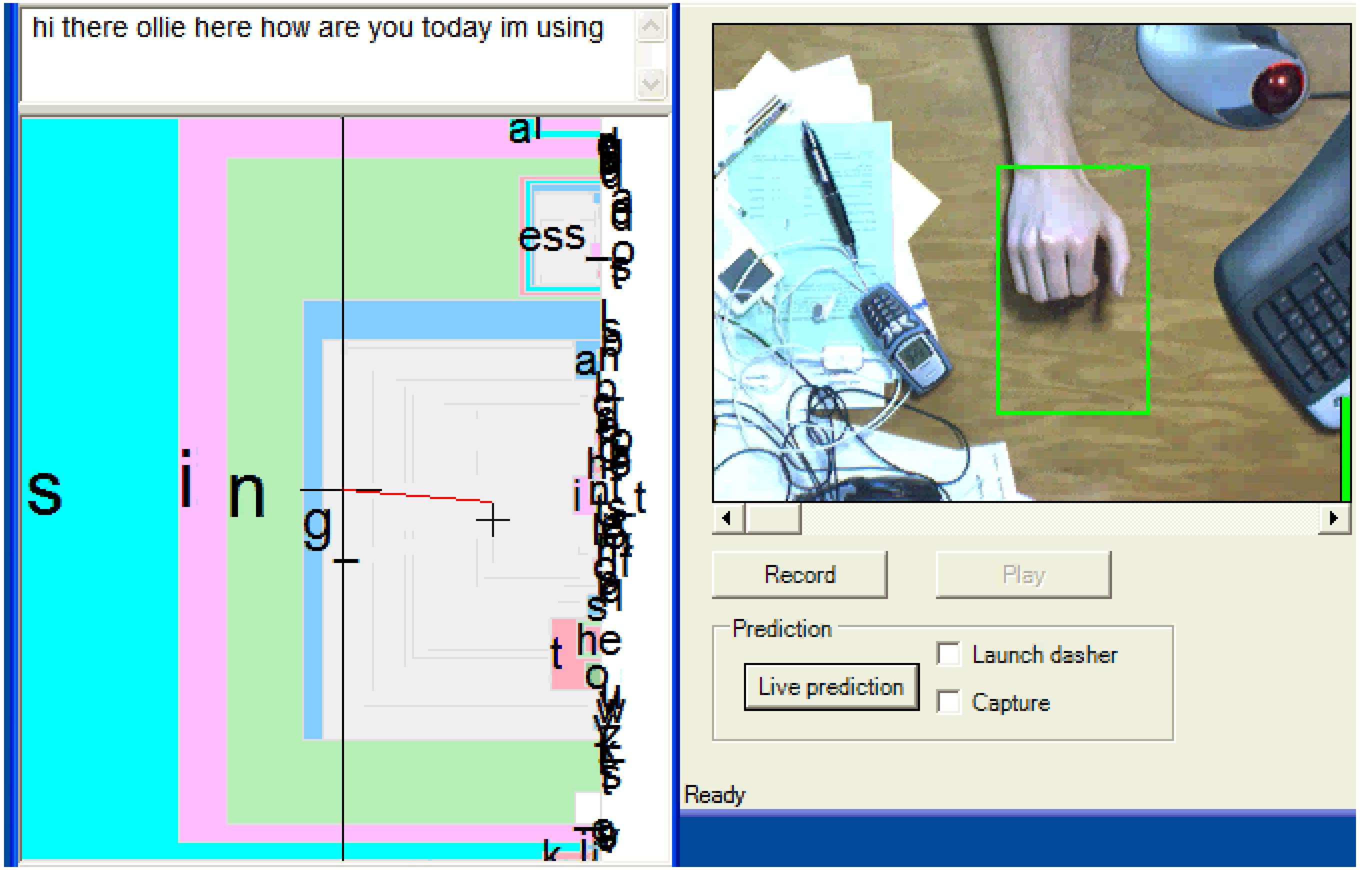
Image-based localisation



Part II:
3D object detection

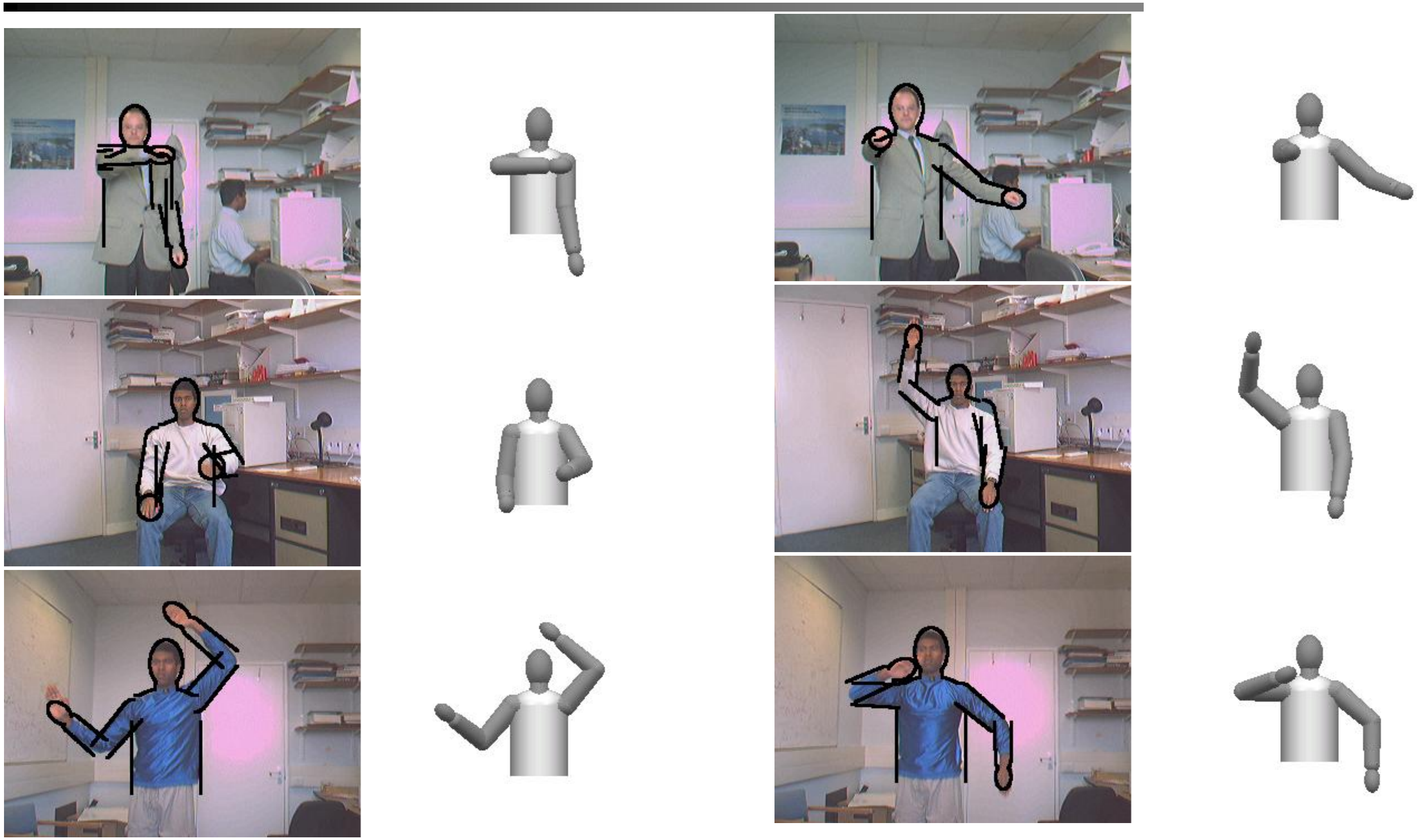
2. Object detection and tracking

Real-time visual controller for Dasher

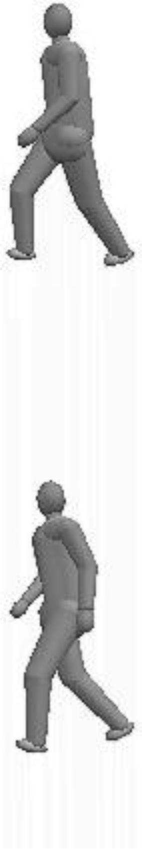


Hand detection system

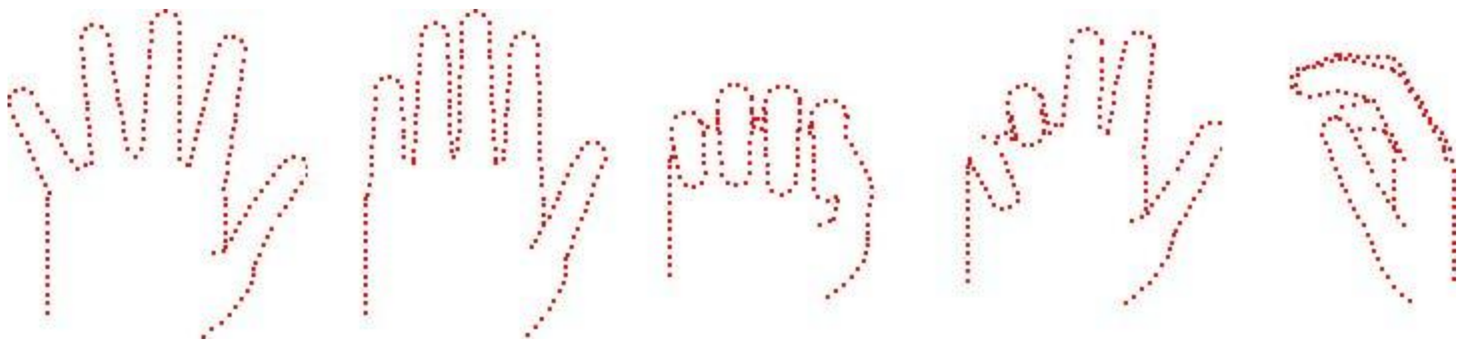
People and pose detection



People and pose detection



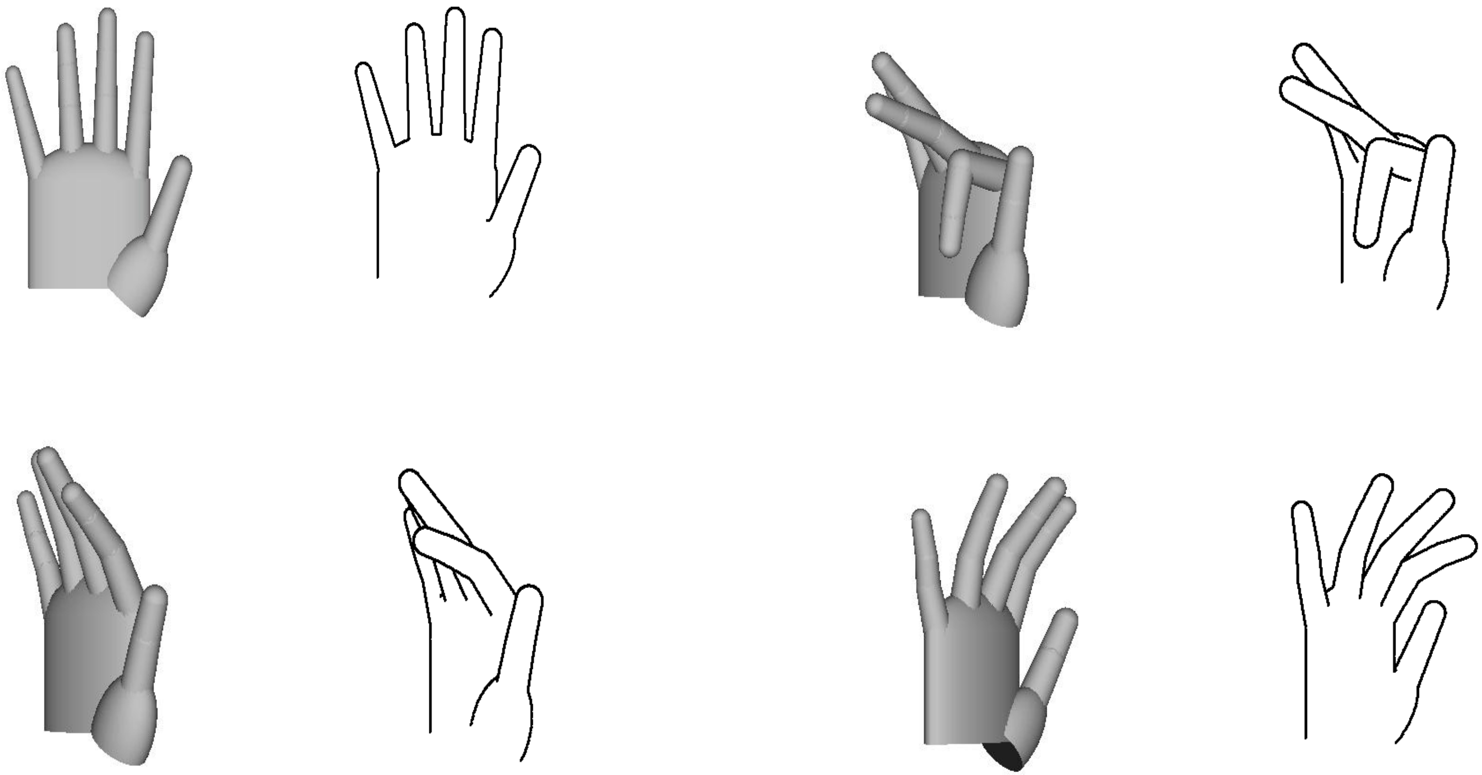
Template-based Detection



- Large number of templates are generated off-line to handle global motion and finger articulation.
- Need for
  - Inexpensive template-matching function
    - Distance Transform and Chamfer Matching
  - Efficient search structure
    - Bayesian Tree structure

3D Hand Model

- Used as generative model
- Constructed from 35 truncated quadrics (ellipsoids, cones)
- Efficient contour projection
- 27 degrees of freedom



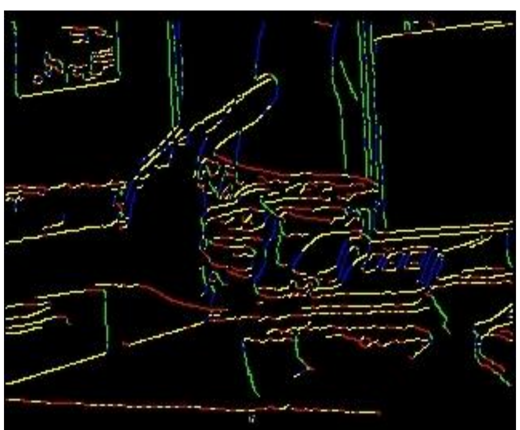
Likelihood : Edges

Input Image

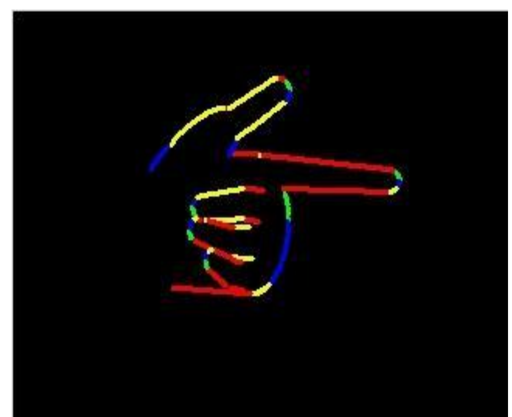
3D Model

Edge Detection

Projected Contours



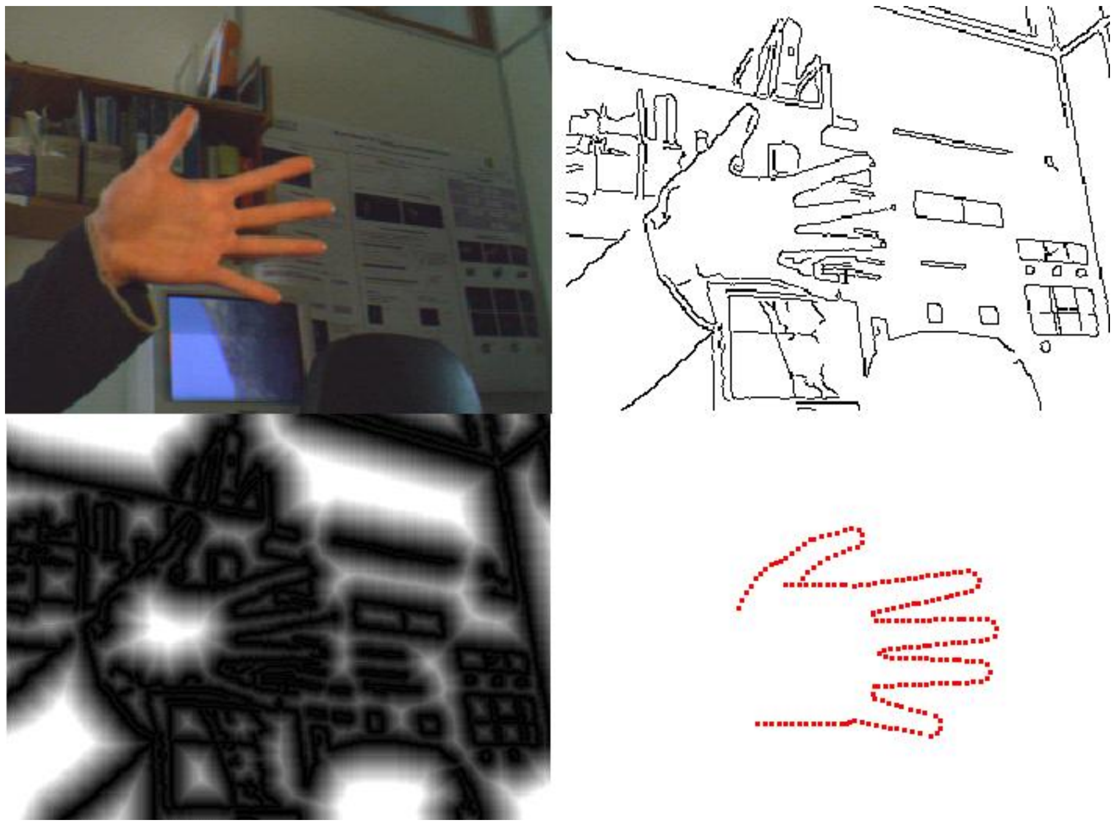
Robust Edge Matching



Chamfer Matching

Input image

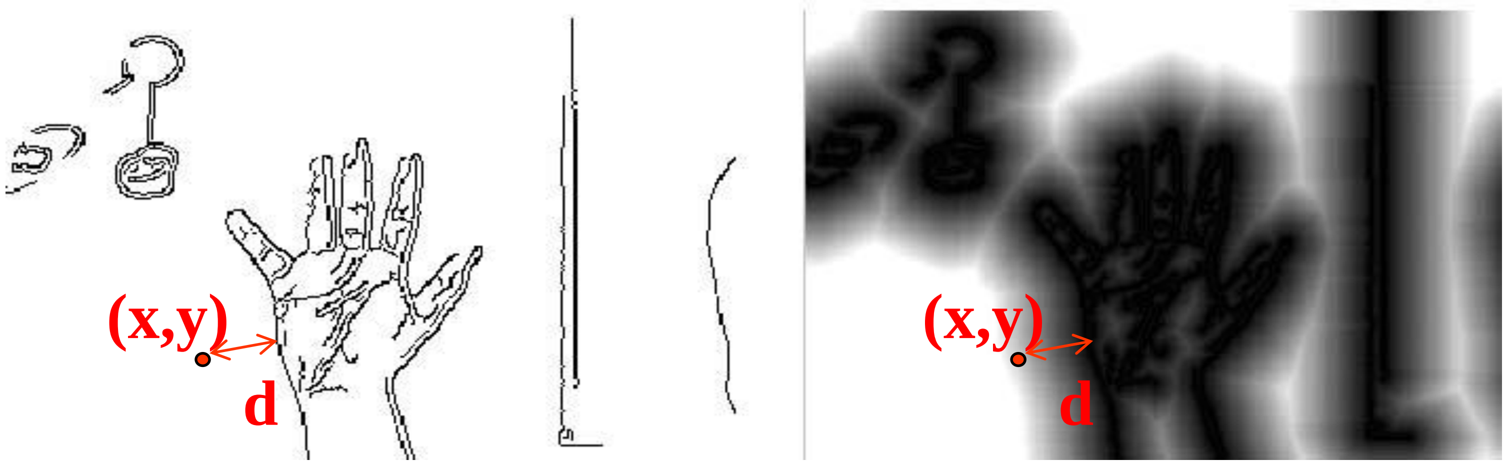
Canny edges



Distance transform

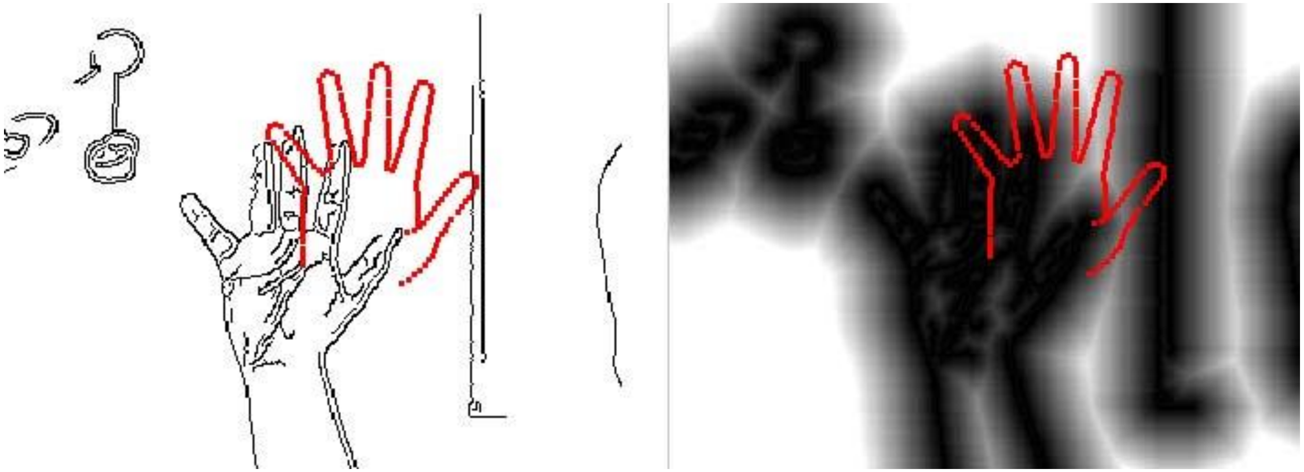
Projected Contours

Distance Transform



- Distance image contains distance to nearest edge
- Calculated only once for each frame

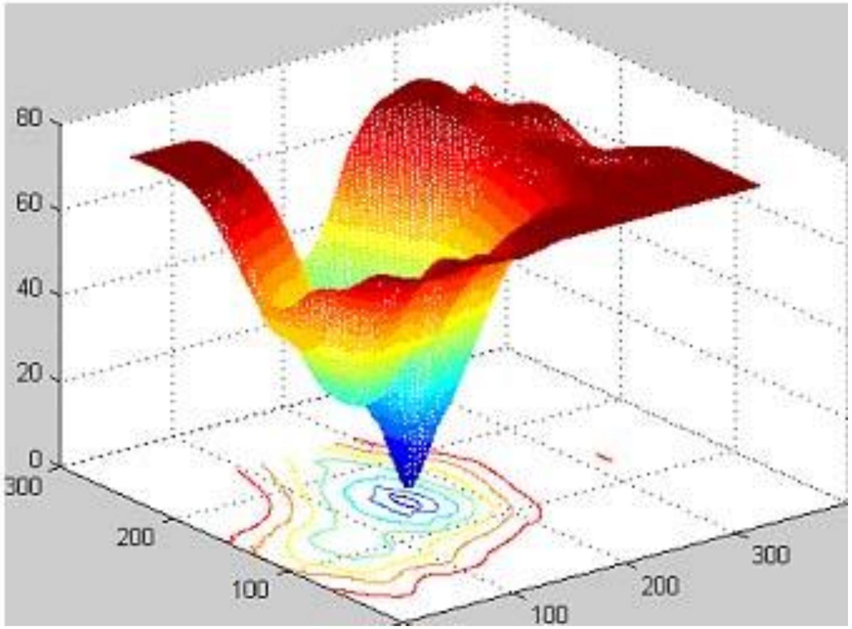
Chamfer Matching



$$d_{cham}(\mathcal{U},\mathcal{V}) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \min_{v\in\mathcal{V}} ||u-v||^2$$

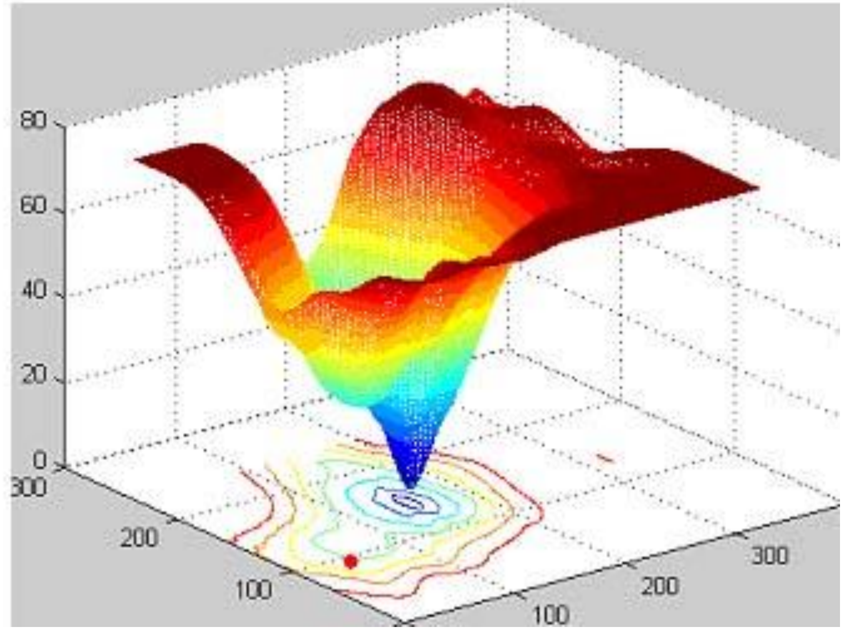
- Chamfer distance: average distance from template points to nearest image point
- Nearest distances obtained from distance image
- Efficient

Chamfer Matching

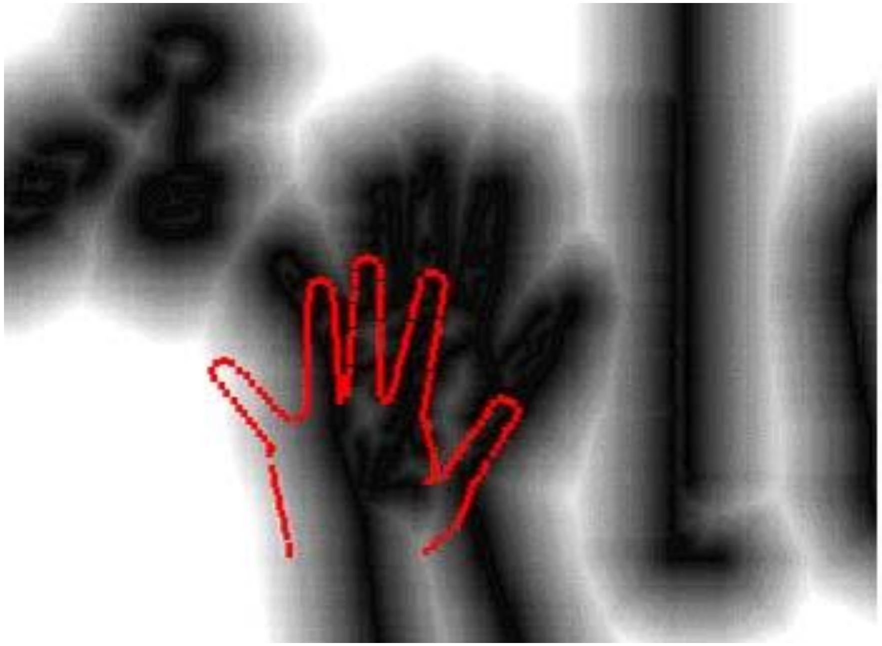


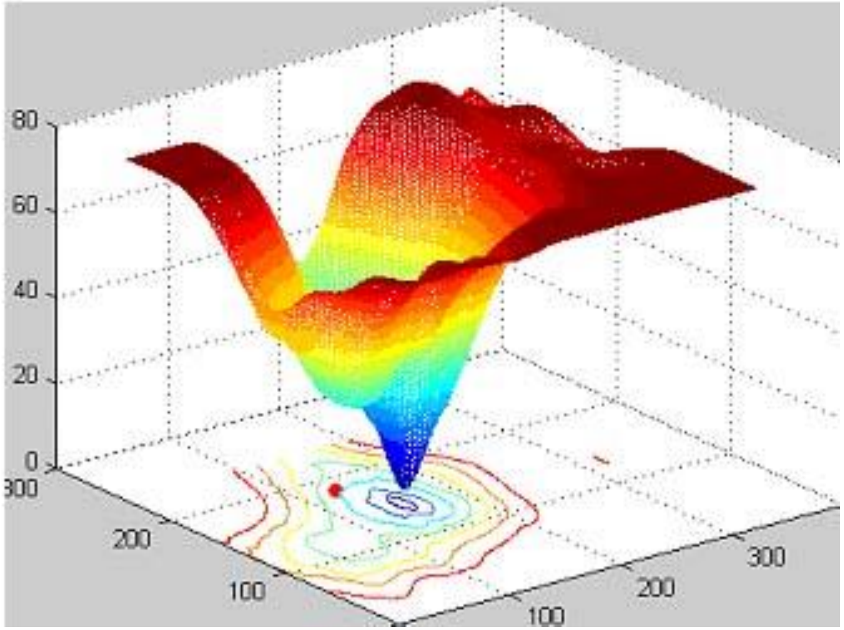
- Distance image provides a smoothed cost function
- Efficient searching techniques can be used to find correct template

Chamfer Matching

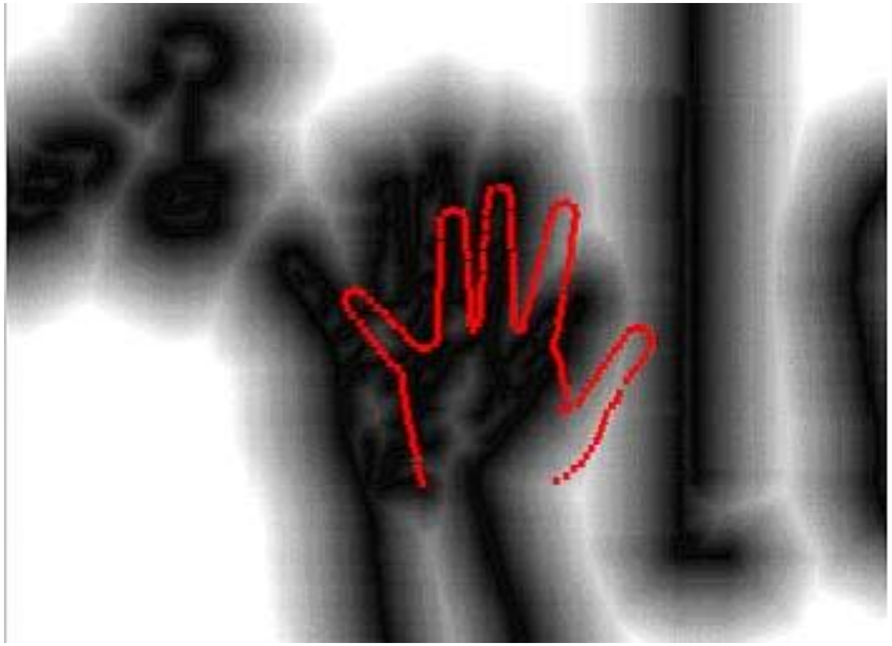


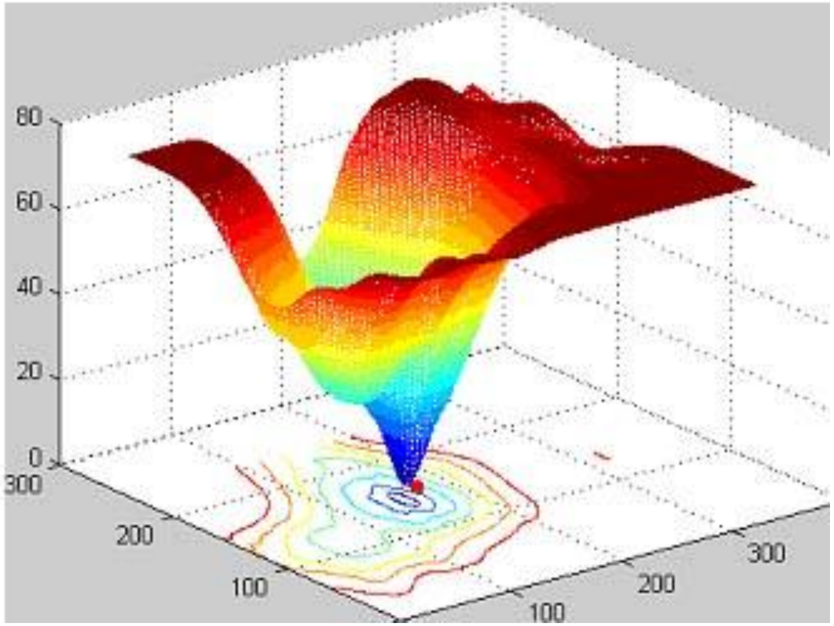
Chamfer Matching



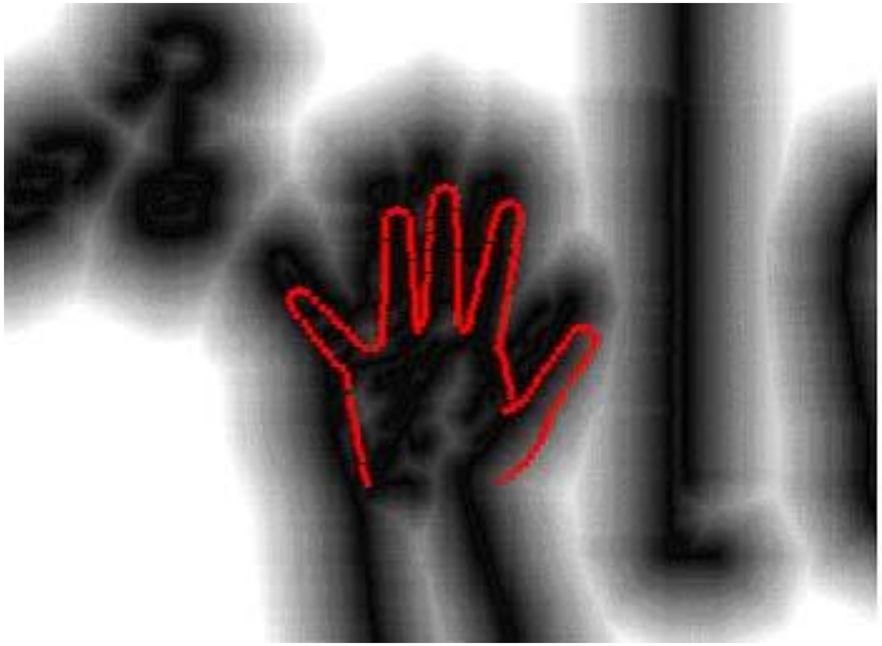


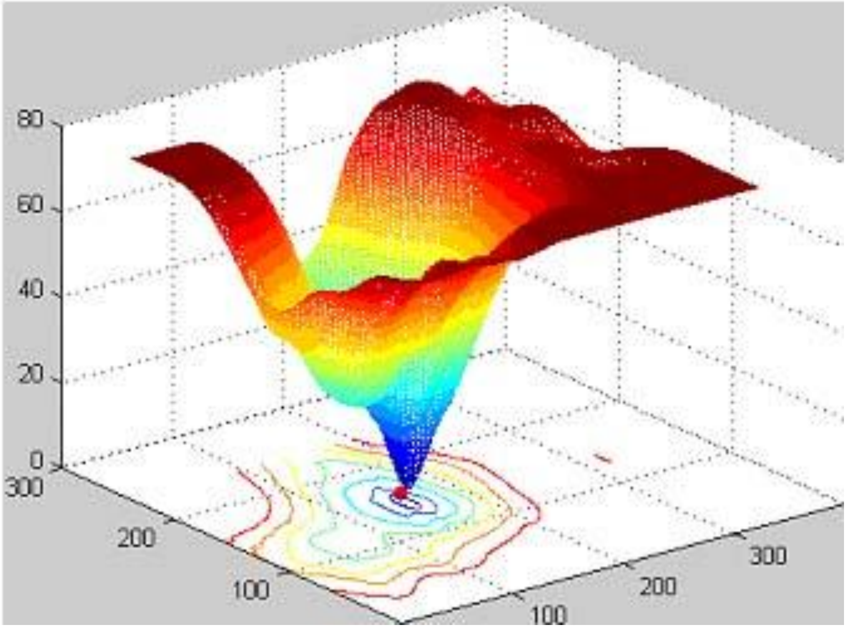
Chamfer Matching



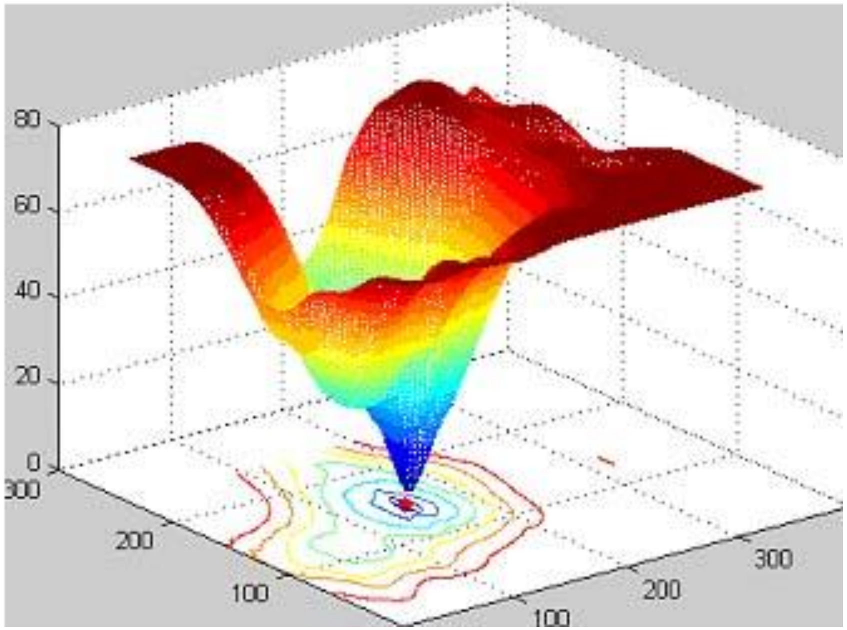


Chamfer Matching



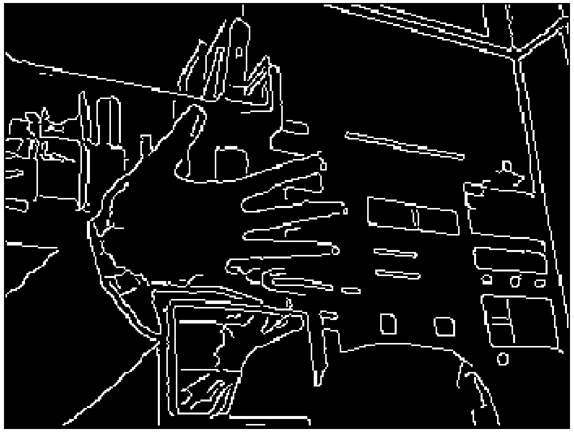


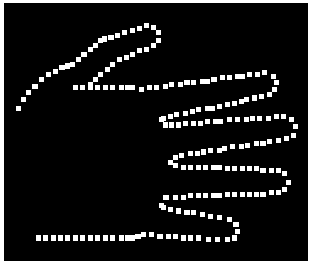
Chamfer Matching

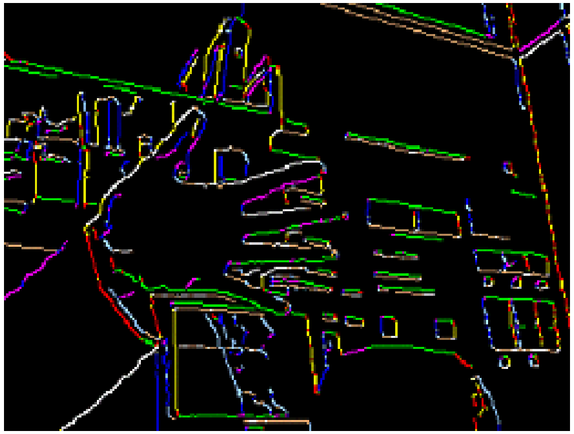


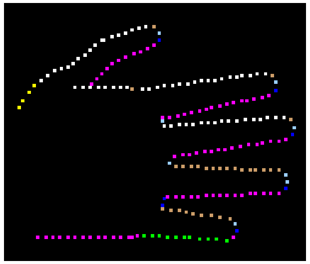
Multiple Edge Orientations

- Edge pixels are divided into 8 groups based on orientation
- Distance transforms are calculated separately for each group
- Total cost is obtained by adding individual chamfer distance









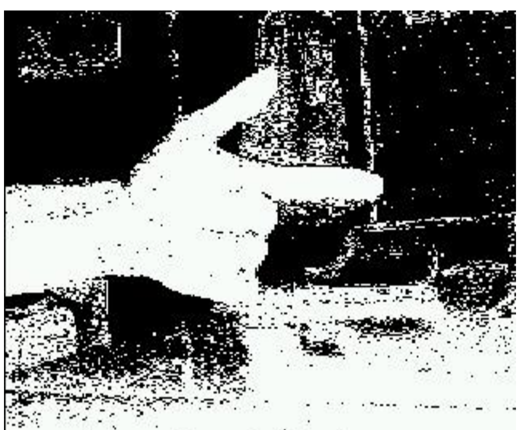
Likelihood : Colour

Input Image

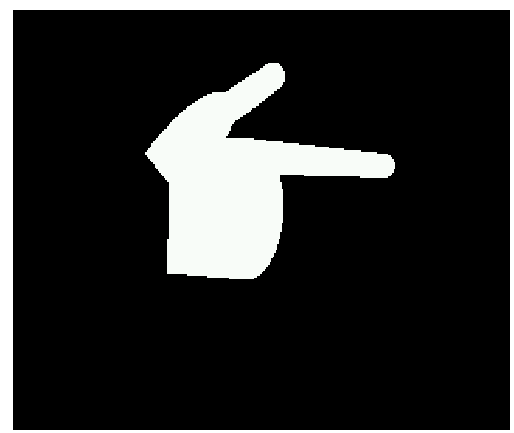
3D Model

Skin Colour Model

Projected Silhouette

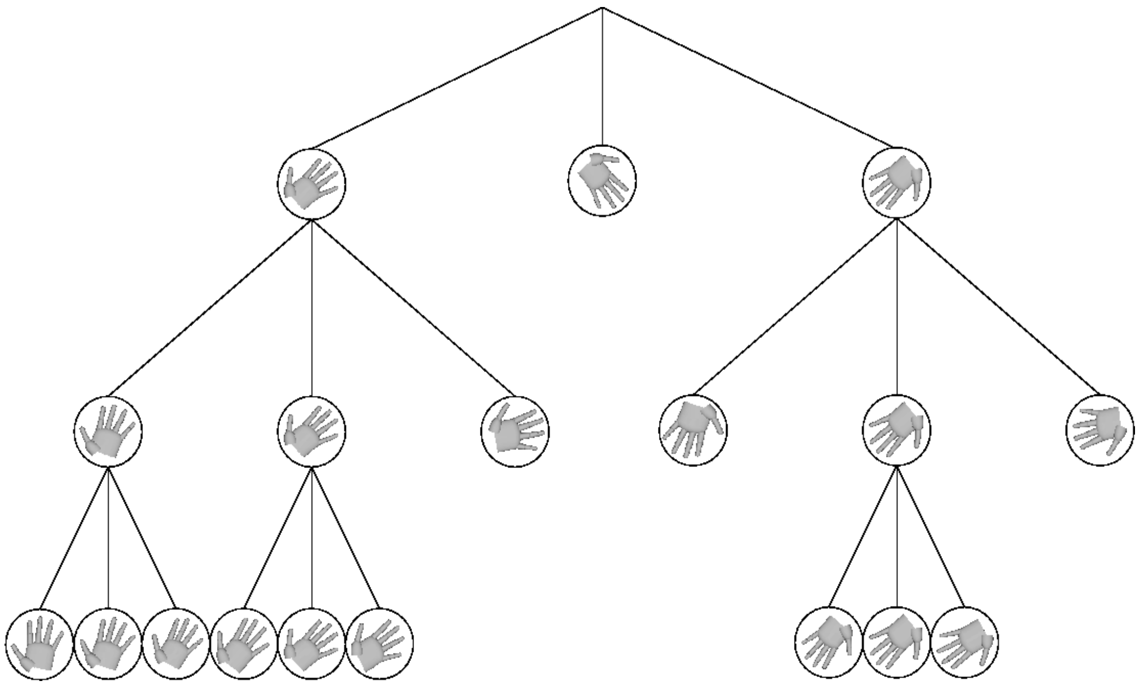


Template Matching

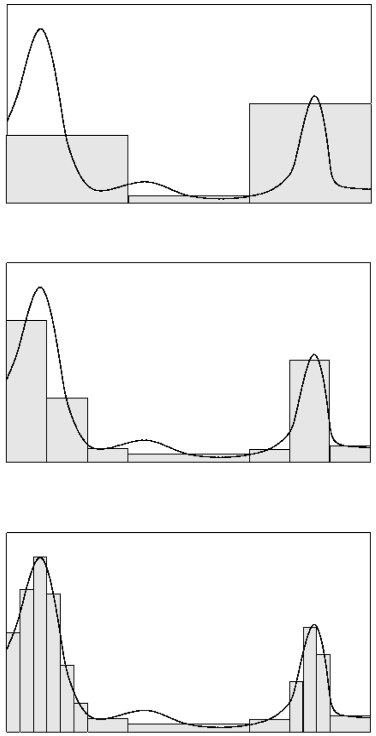


Bayesian-Tree

State space partitioning

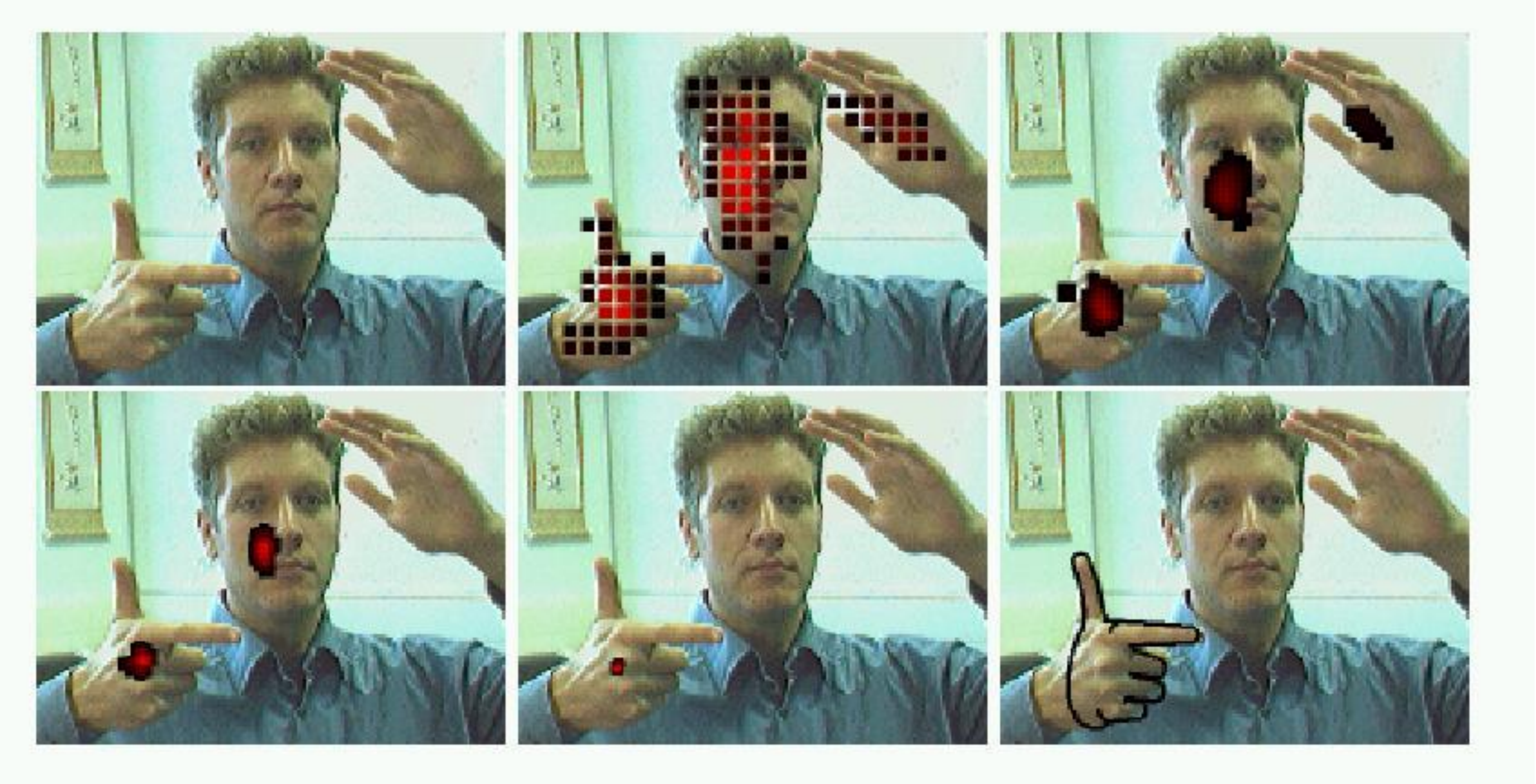


Estimation of posterior pdf



- The search-tree is brought into a Bayesian framework by adding the prior knowledge from previous frame.
- The Bayesian-Tree can be thought as approximating the posterior probability at different resolutions.

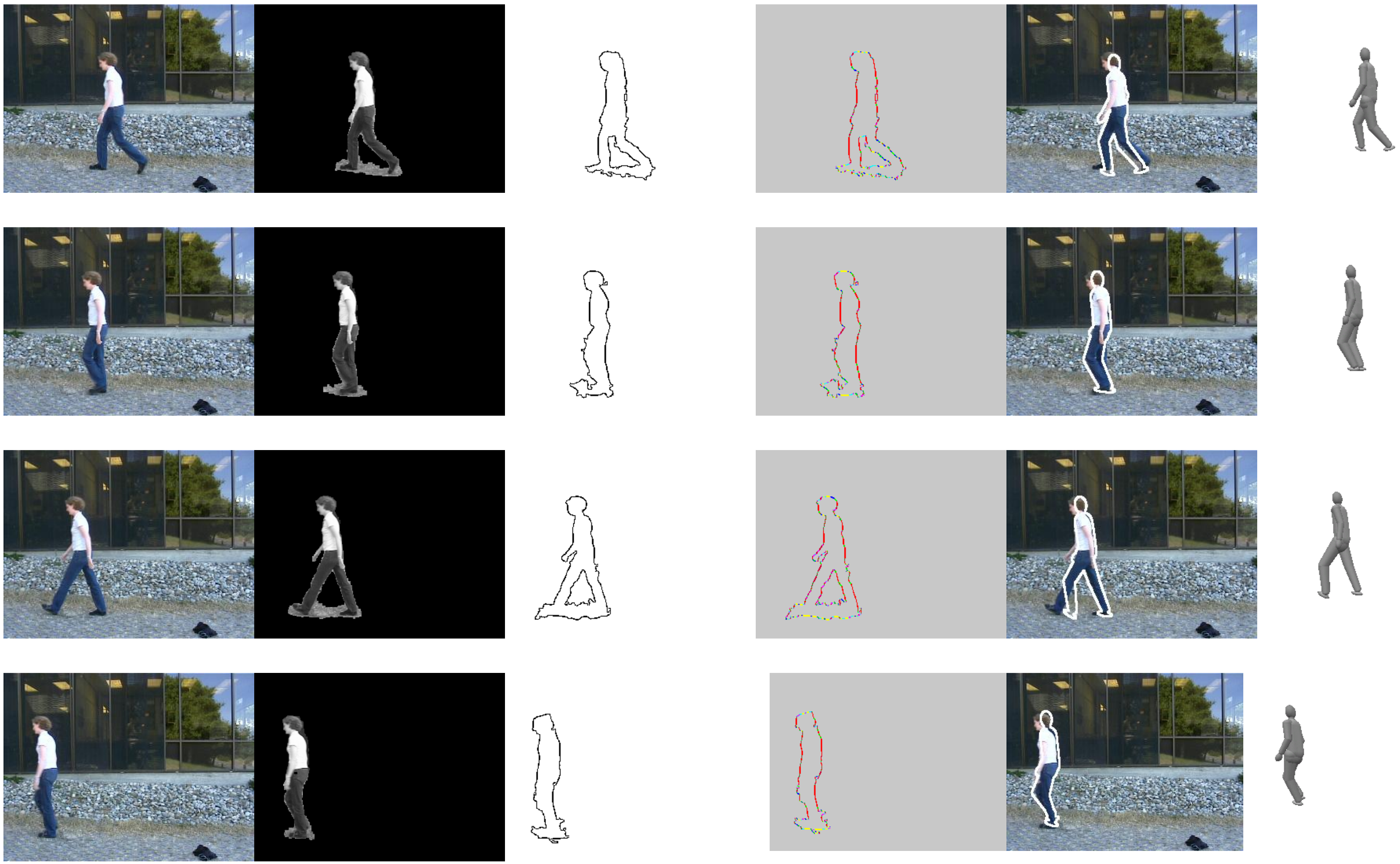
Evaluation at Multiple Resolutions



Tracking - 3D mouse

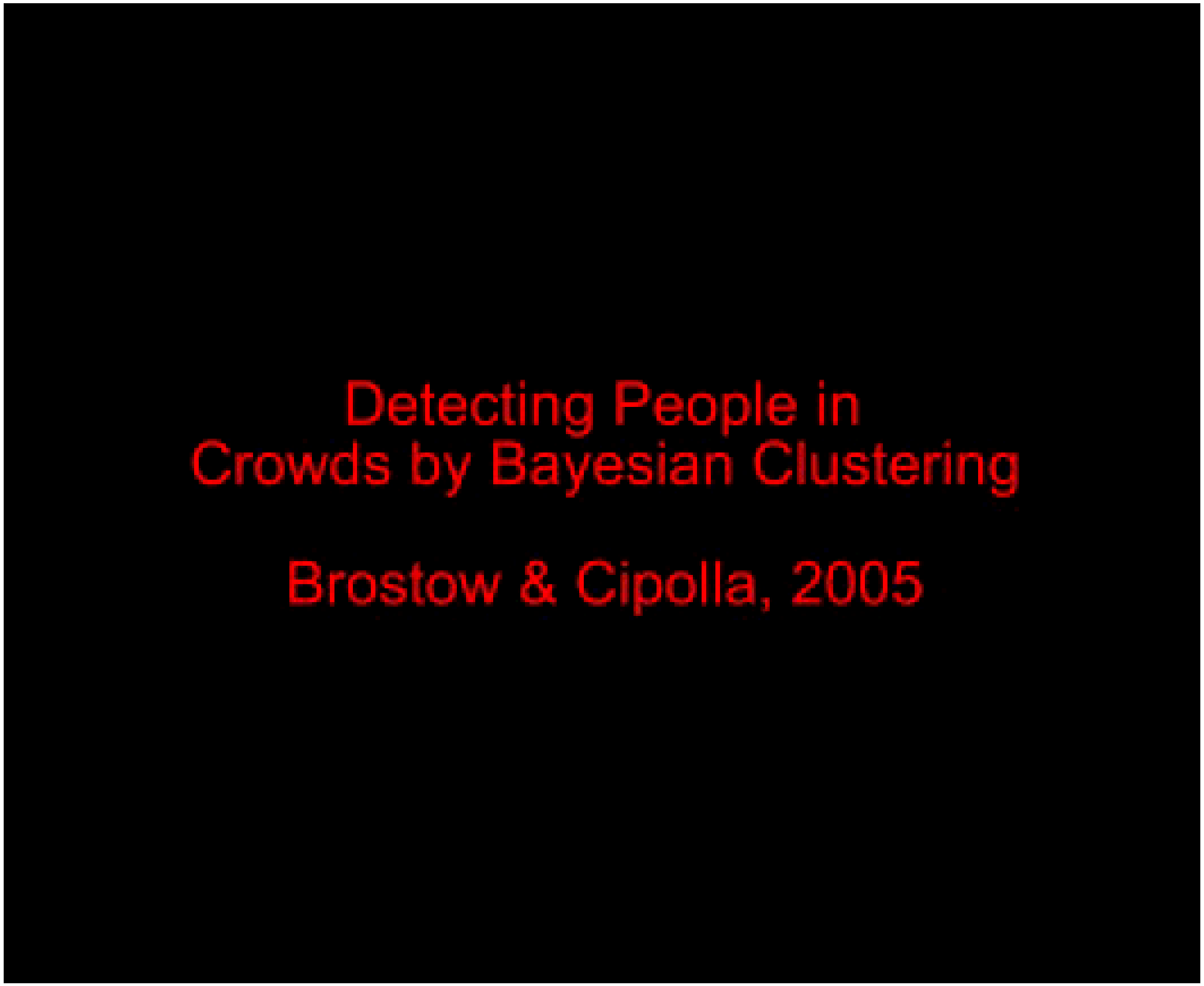
Detection of people

Input | Foreground | Silhouette | Oriented Edges | Image & Template | Solid Model



Detecting and tracking people in crowds

Tracking people in crowds

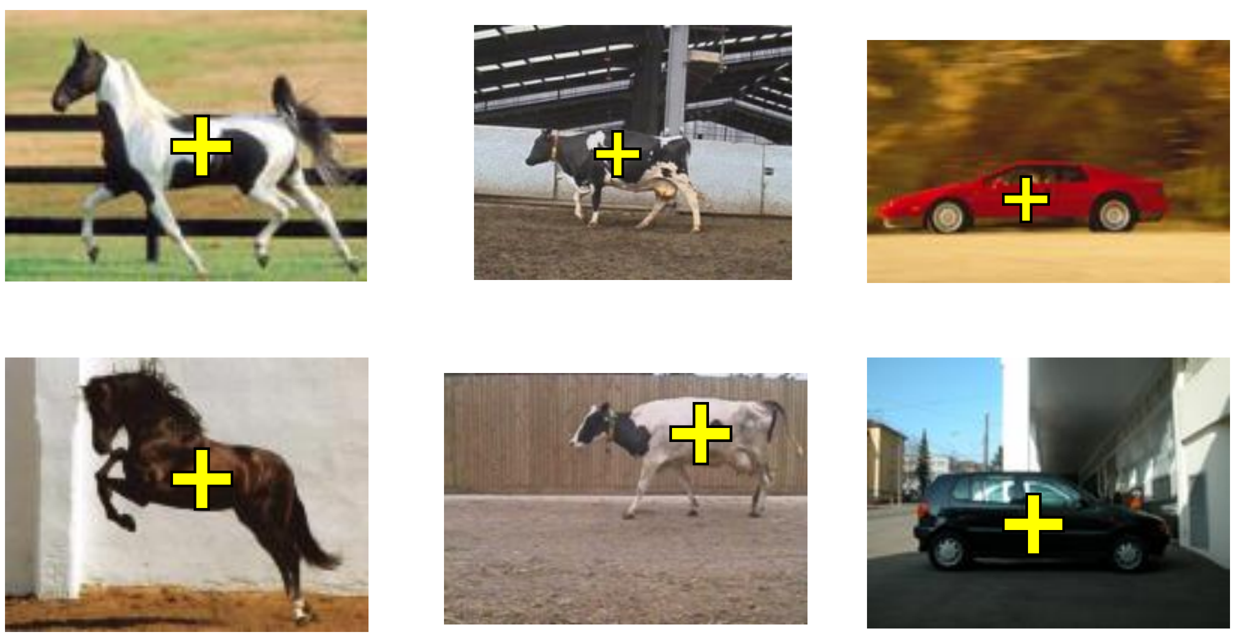


4. Learning to detect object categories

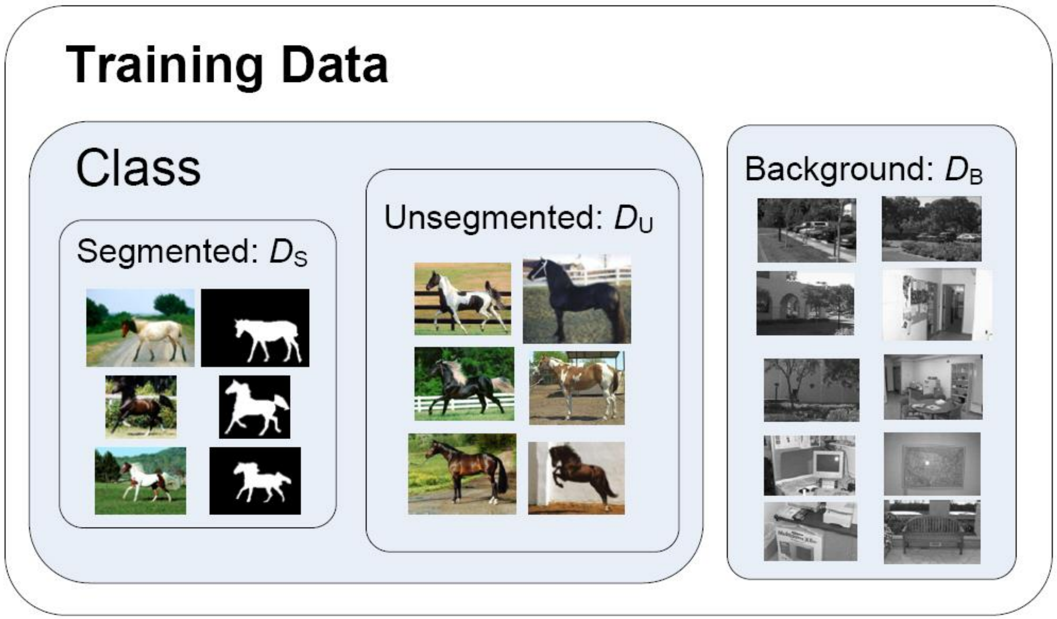
Learning and adaptability

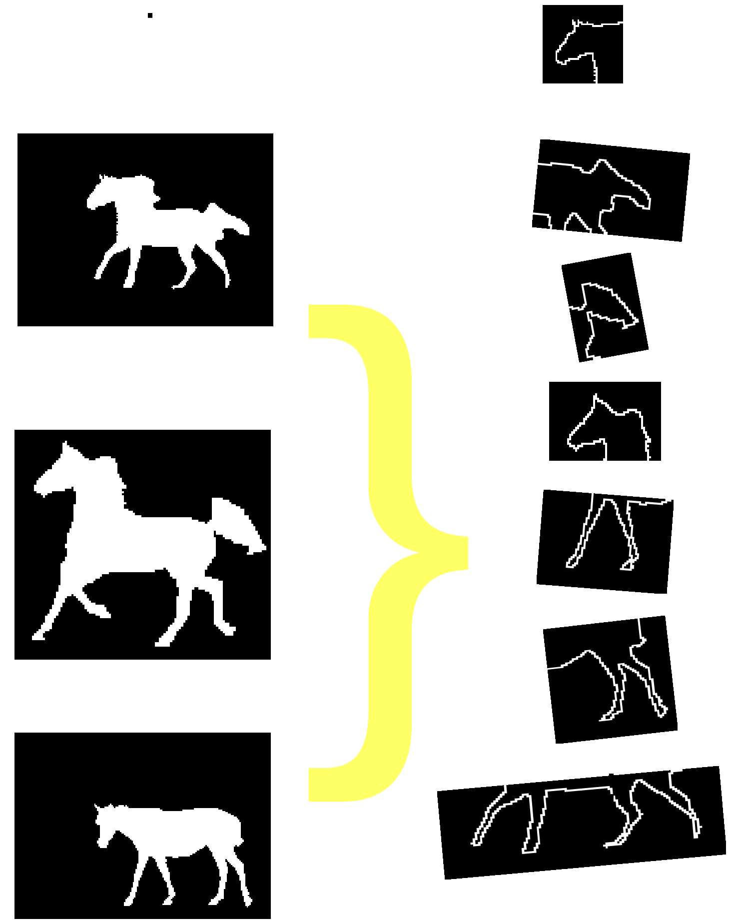
- Learn to recognise images of a particular class, localised in space and scale
- i.e. find the horse/cow/car etc!

Desired Results

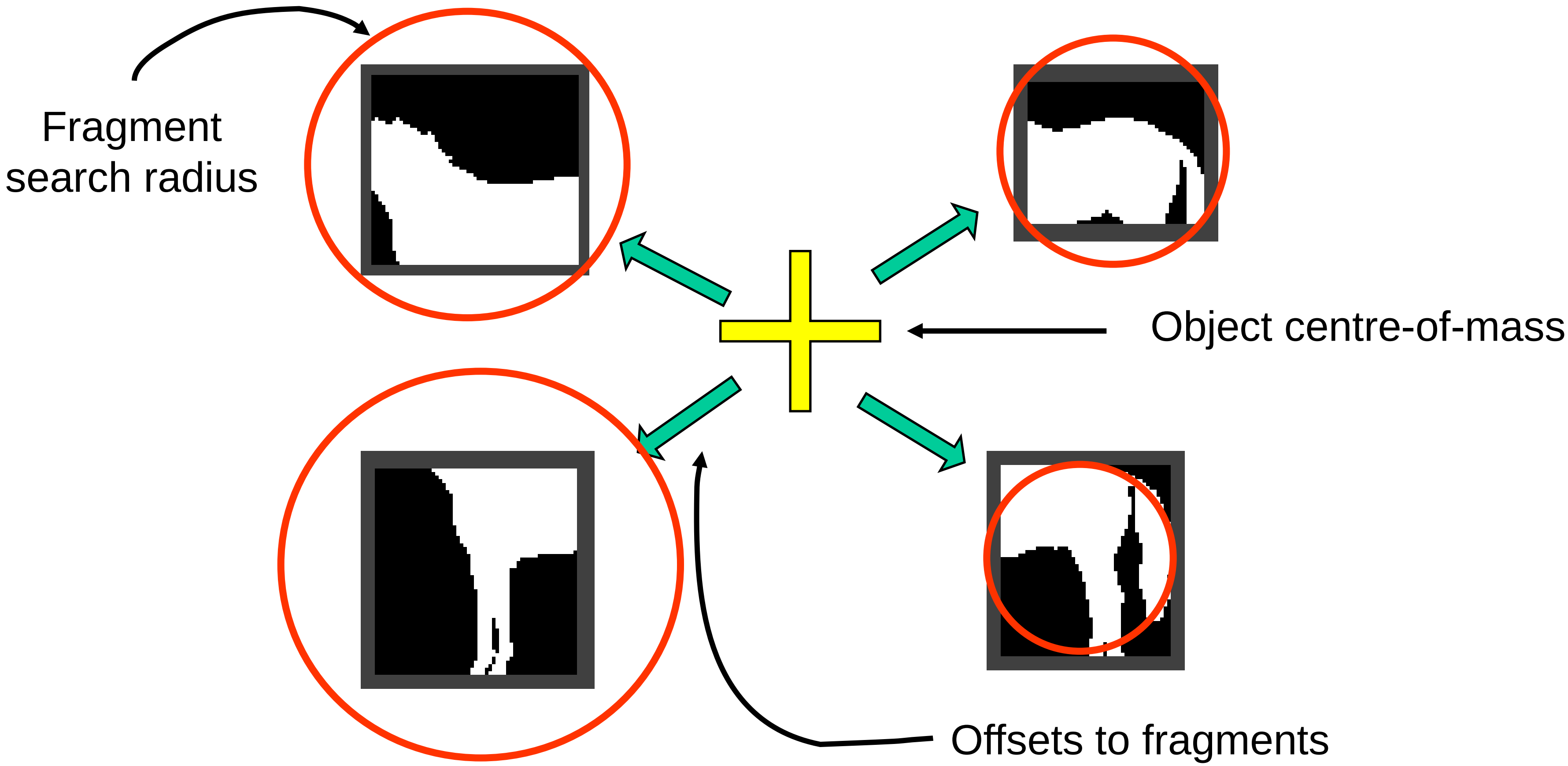


Learning and Adaptability





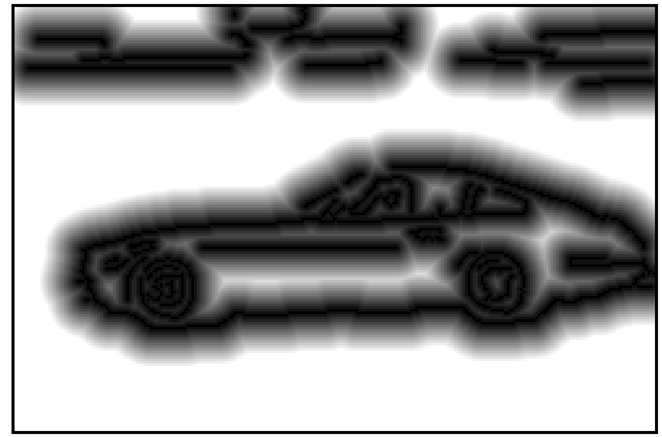
Learn Object Model



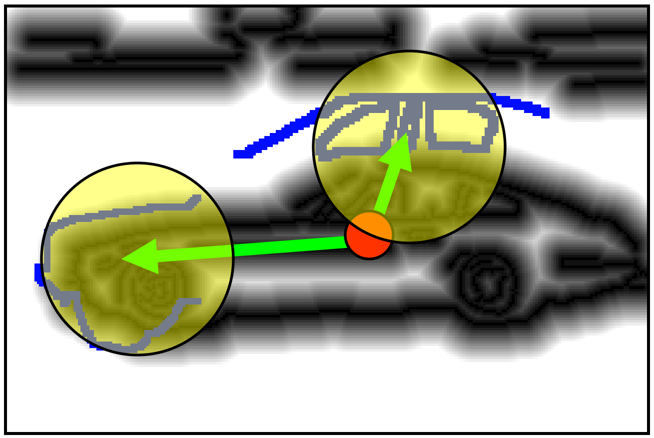
Matching Each Part

Oriented Distance Transform

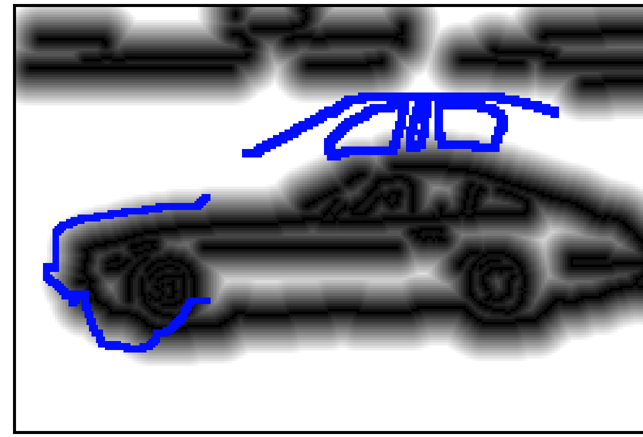
Oriented Canny Edge Detector



Matching Each Part



Oriented Chamfer Matching



$$d^{(T,E)}_{\text{cham}_\tau}(\mathbf{x}) = \frac{1}{N_T}\sum_{\mathbf{t}\in T}\min(\mathrm{DT}_E(\mathbf{t}+\mathbf{x}),\tau)$$

$$d^{(T,E)}_{\text{orient}}(\mathbf{x}) = \frac{1}{N_T}\sum_{\mathbf{t}\in T}|o(\mathbf{t}) - o(\mathrm{ADT}_E(\mathbf{t}+\mathbf{x}))|$$

Oriented Chamfer Score: $d^{(T,E,\lambda)}(\mathbf{x}) = d^{(T,E)}_{\text{cham}_\tau}(\mathbf{x}) + \lambda d^{(T,E)}_{\text{orient}}(\mathbf{x})$

$$\mathbf{r}(F,E|\mathbf{c}) = \arg\min_{\mathbf{x}\in R_{\hat{\mathbf{x}},\sigma}}\left(d^{(T,E,\lambda)}(\mathbf{x}) + W(\mathbf{x}|\hat{\mathbf{x}},\sigma)\right)$$

“Feature Responses”: $v(F,E|\mathbf{c}) = d^{(T,E,\lambda)}(\mathbf{r}(F,E|\mathbf{c}))$

Object Detection (1)

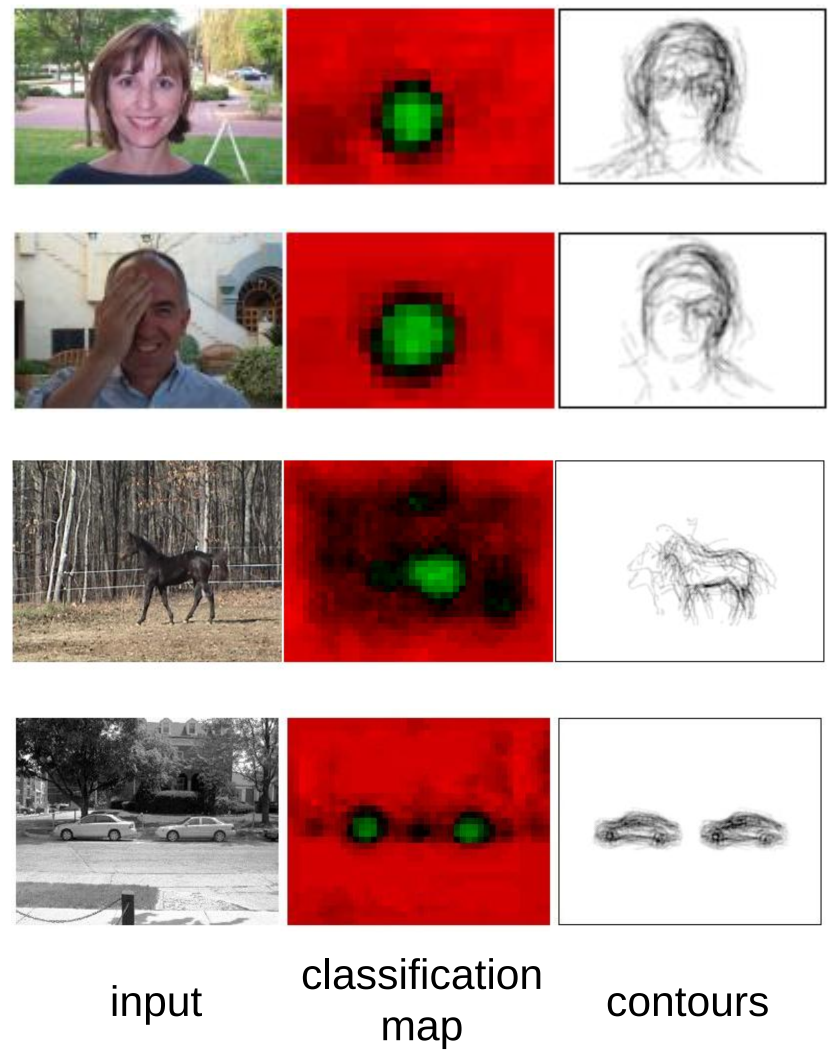
- Given a learned model, detection uses a classification function *K*(**c**)
  - boosted additive model combines feature responses *v*(*F*,*E*|**c**) for each part:

[redacted]

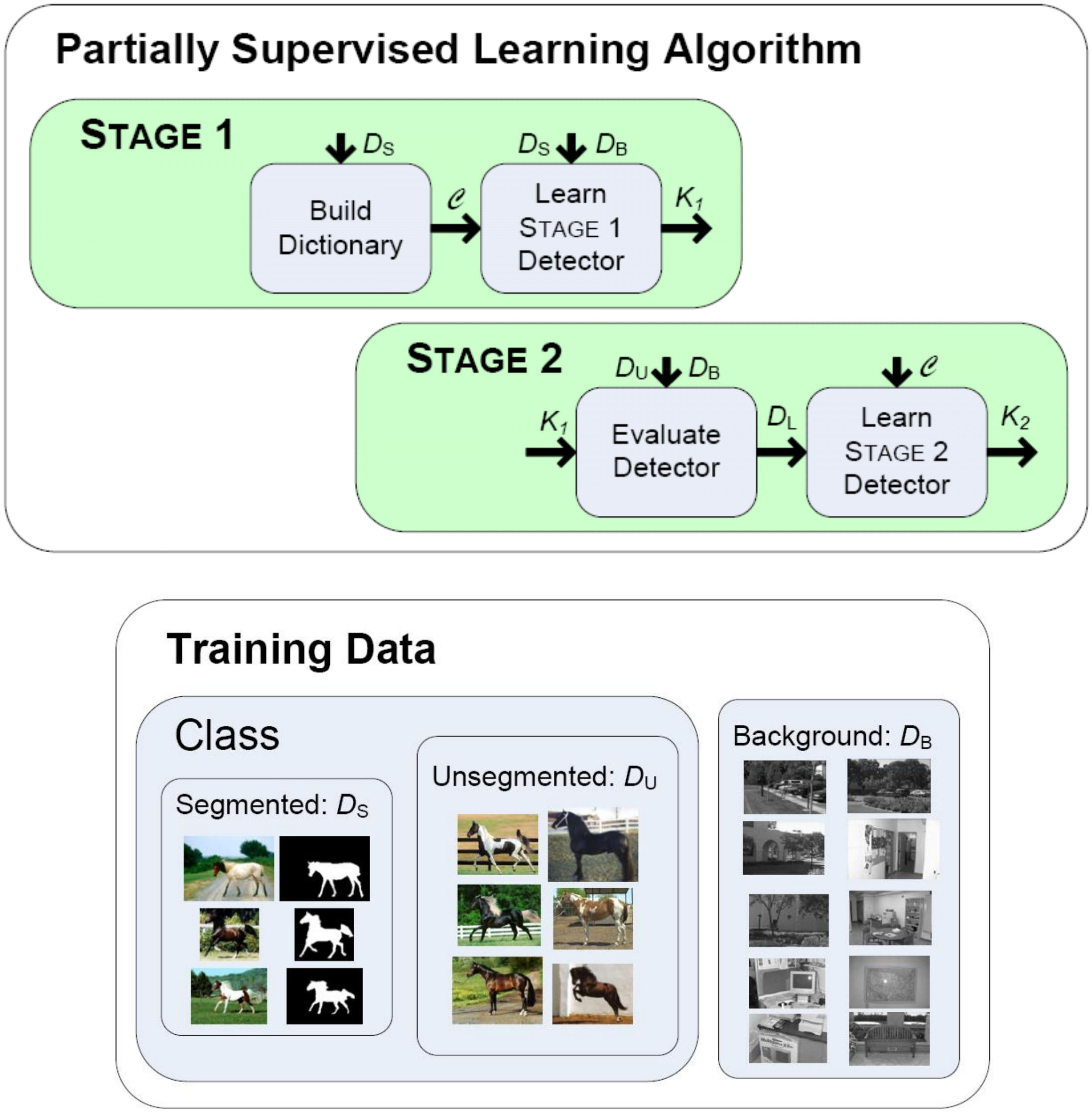
$$K(\mathbf{c}) = \sum_{m=1}^{M} a_m \delta(\overbrace{v(F_m, E|\mathbf{c})} > \theta_m) + b_m$$

Object Detection (2)

- Evaluate $K(\mathbf{c})$ for all centroids in test image gives classification map
  - confidence value as function of position
    - +ve (green) => object present
    - -ve (red) => no object
- Globally thresholded local maxima give detections

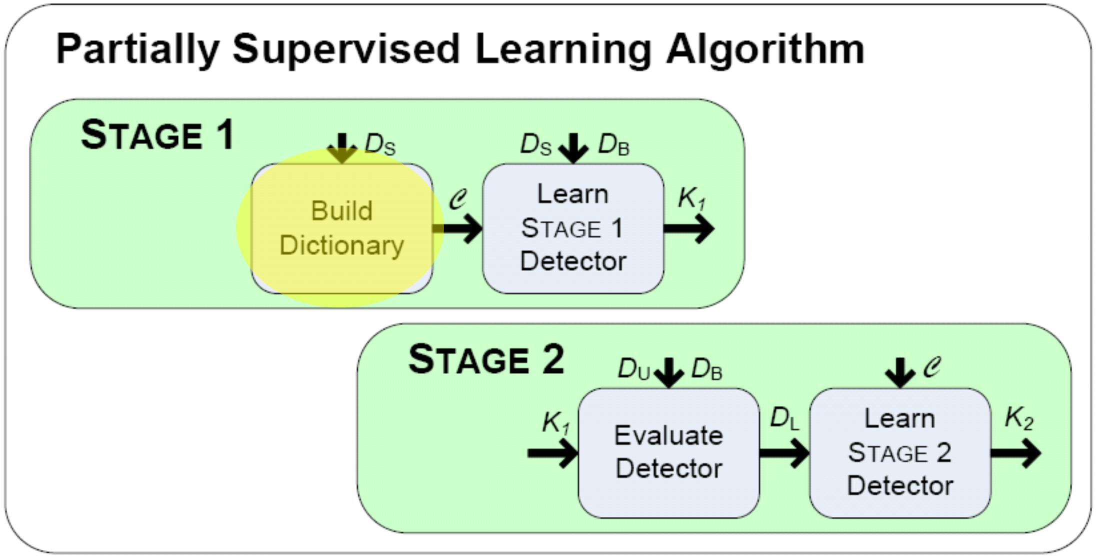


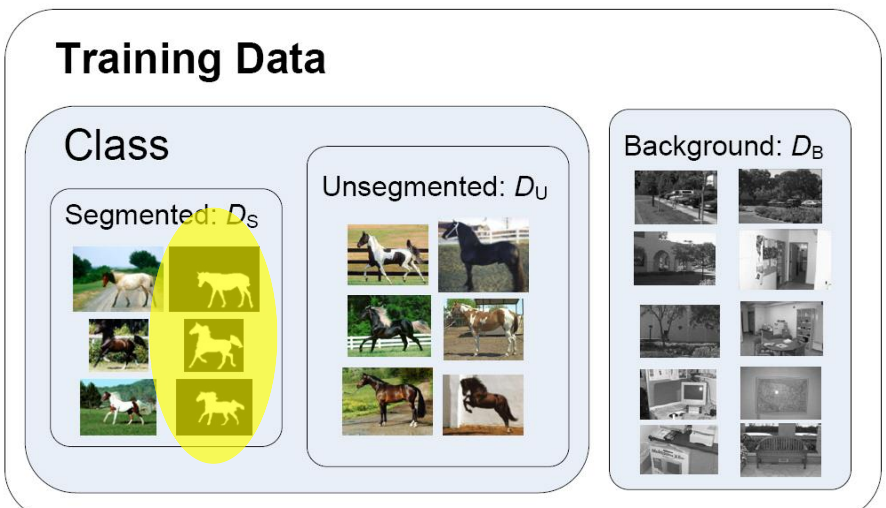
Learning Paradigm

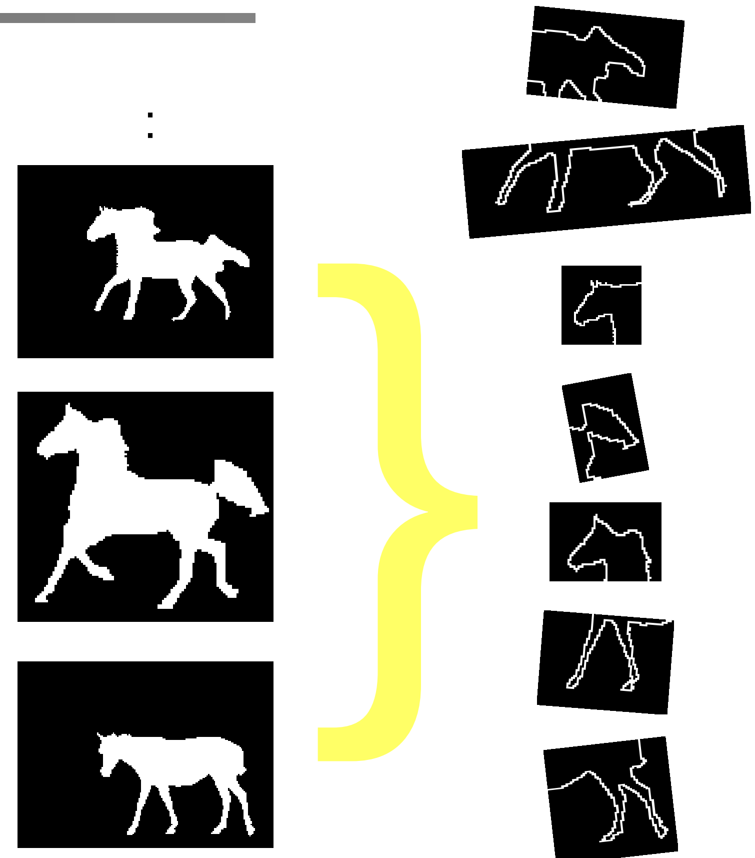


- Stage 1
  - Fully supervised
  - Uses small (~10 images) database of segmented images
- Stage 2
  - Leverages a second, larger, set of *unsegmented* images to improve detector performance

Learning Paradigm





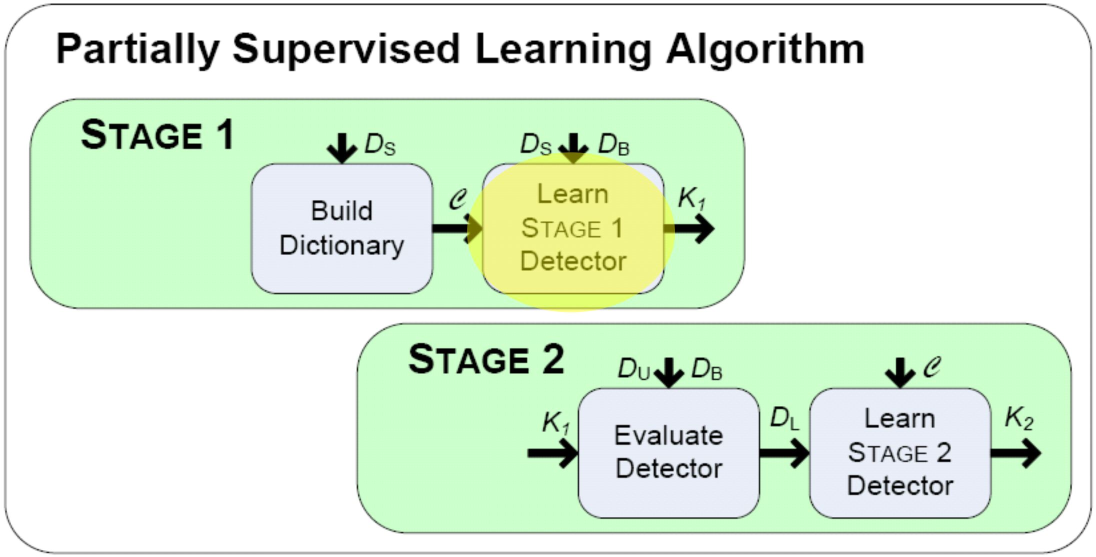


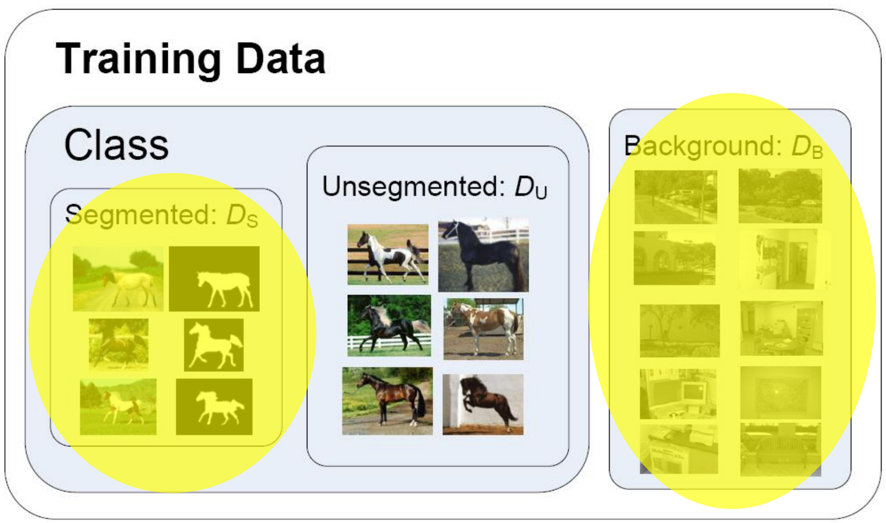
Masks (~10)

Contour Fragments (~1000)

NB Slight random transformation to aid generalisation ability

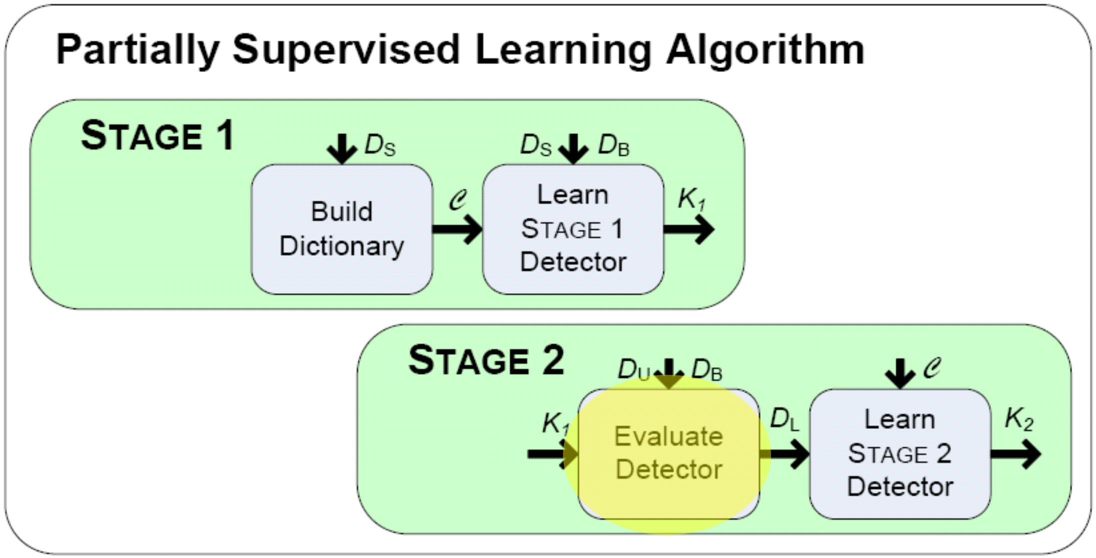
Learning Paradigm

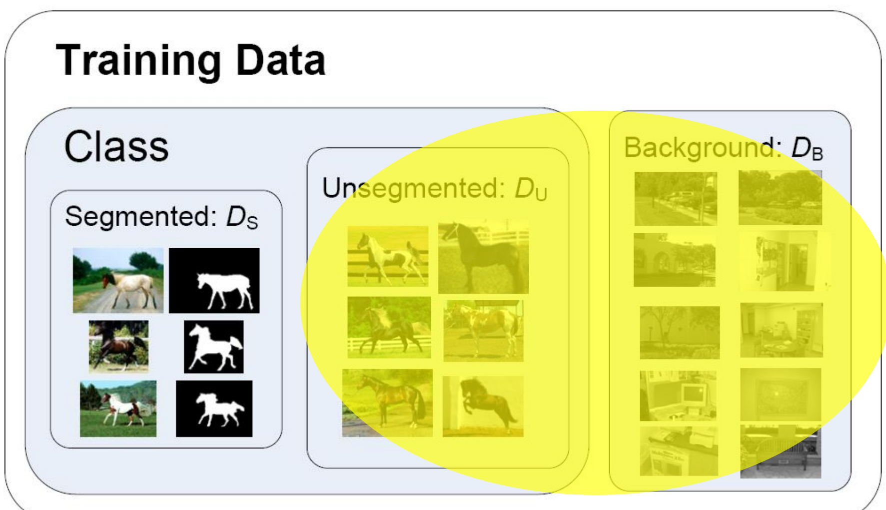




- Stage 1 Detector
  - Learned from small segmented database and full background database
  - Gives $K_1(\mathbf{c})$, a rudimentary detector

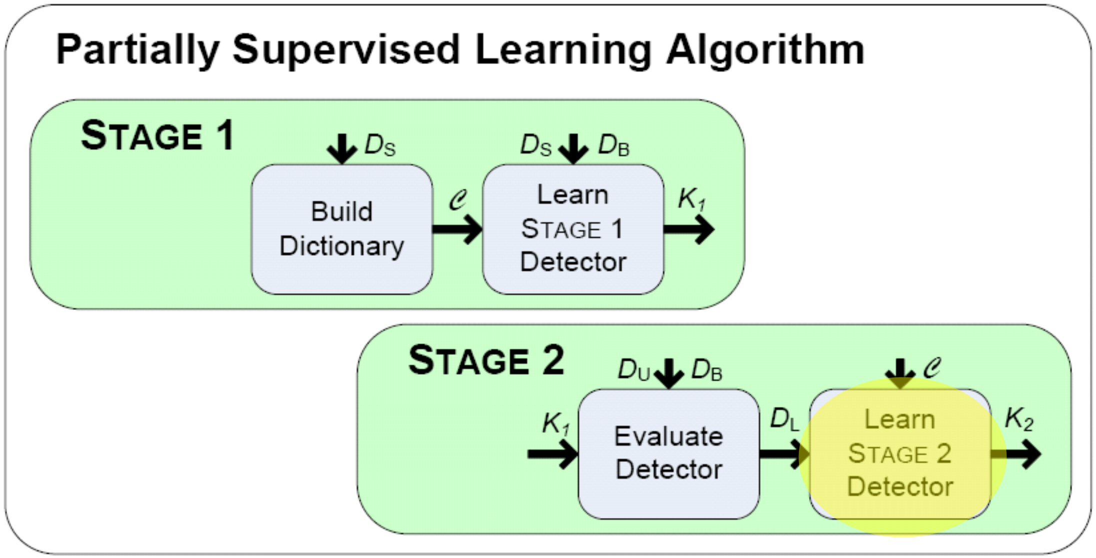
Learning Paradigm

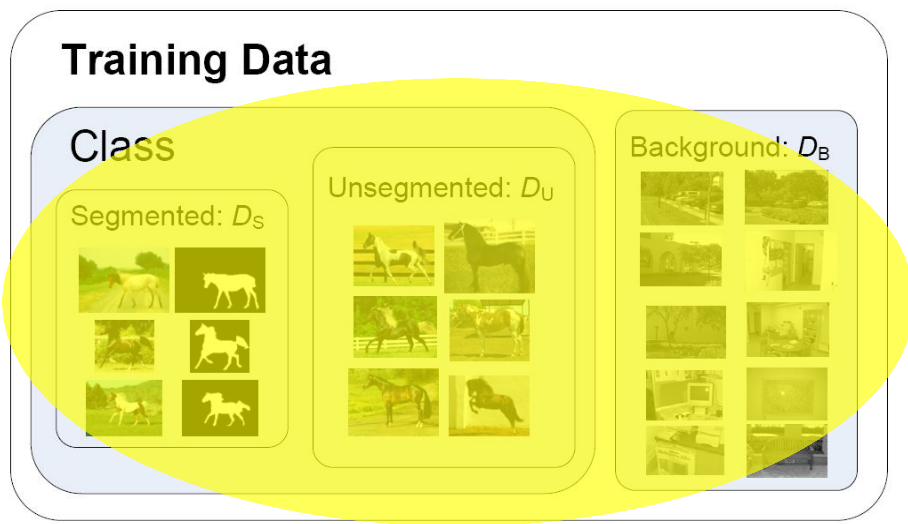




- Stage 2
  - Evaluate rudimentary detector $K_1(\mathbf{c})$
  - On unsegmented class images
    - locates approximate object centroid positions
  - On background images
    - locates problem areas (*clutter*)

Learning Paradigm

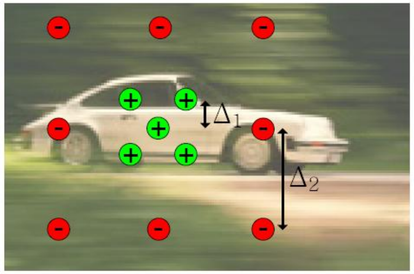




- Stage 2 Detector
  - Learned from complete dataset
  - Gives $K_2(\mathbf{c})$, a better detector than $K_1(\mathbf{c})$

Training examples

- Boosting algorithm requires for a given centroid
  - a vector of feature responses
  - a ‘target’ classification
    - +1 (green) => object present at this centroid
    - -1 (red) => object not present at this centroid
- Each training image can therefore generate multiple *training examples* to aid the localisation of the resulting classification maps
  - around each true centroid in positive training images, several examples are taken in a given pattern
  - examples are taken in background training images at points of clutter (in Stage 2)



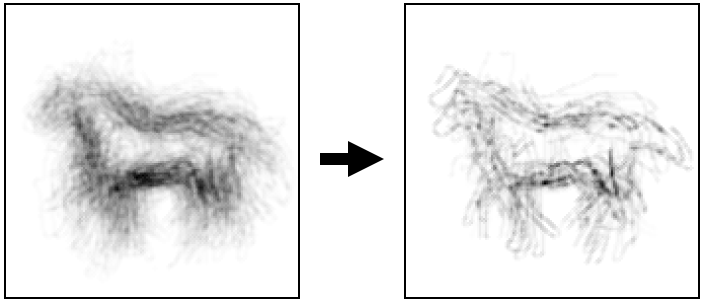
positive training image

background training image

Learning a Detector

- Evaluate chamfer scores for each random contour fragment at each training example
  - gives feature vector
  - target value known since training data labelled
- Boosting algorithm greedily selects a discriminative subset of fragments, performing simultaneously:

1. Feature Selection



1000 random fragments

50 discriminative fragments

(contours drawn at estimated positions)

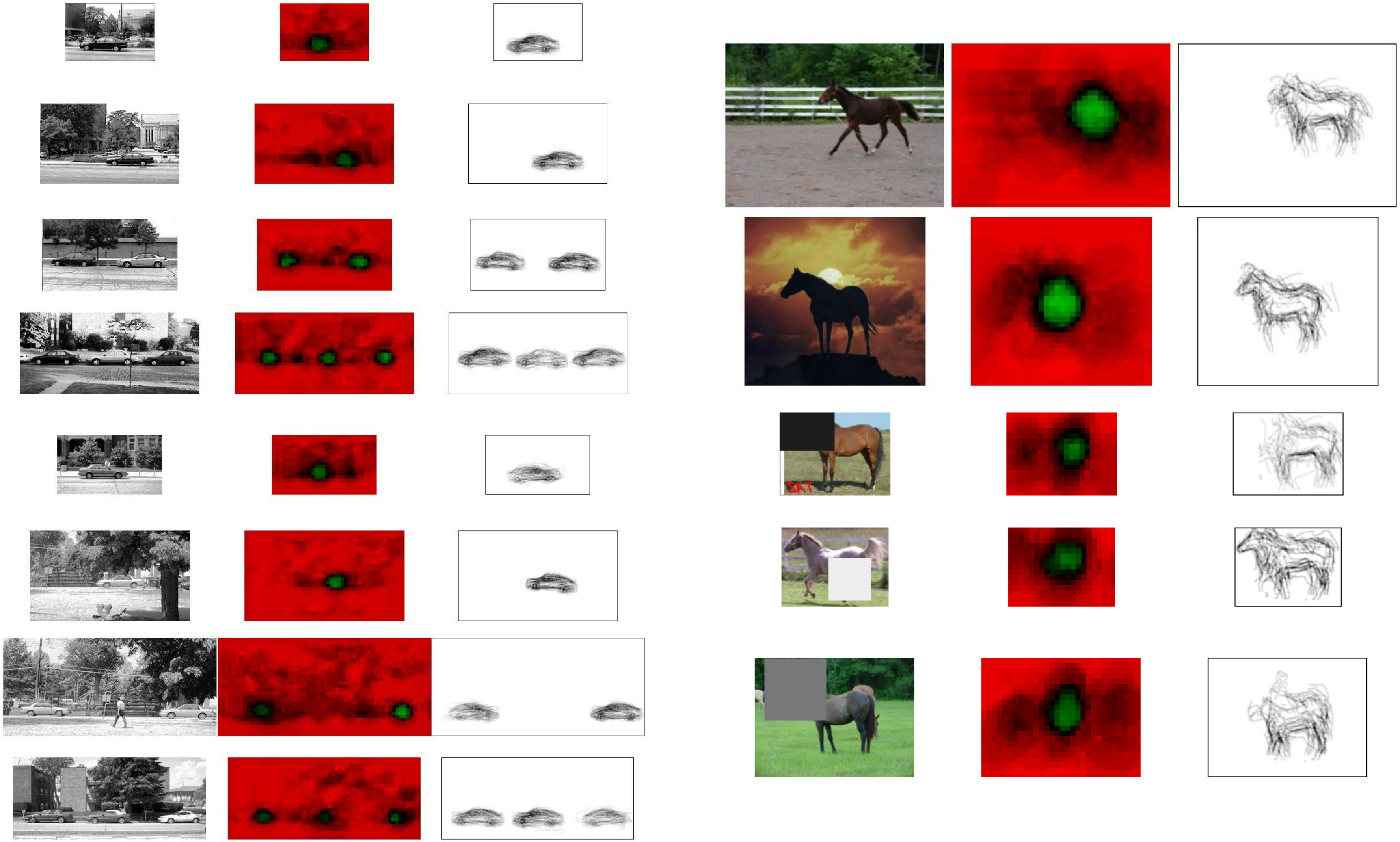
2. Model Parameters Estimation

Select σ, λ for each part

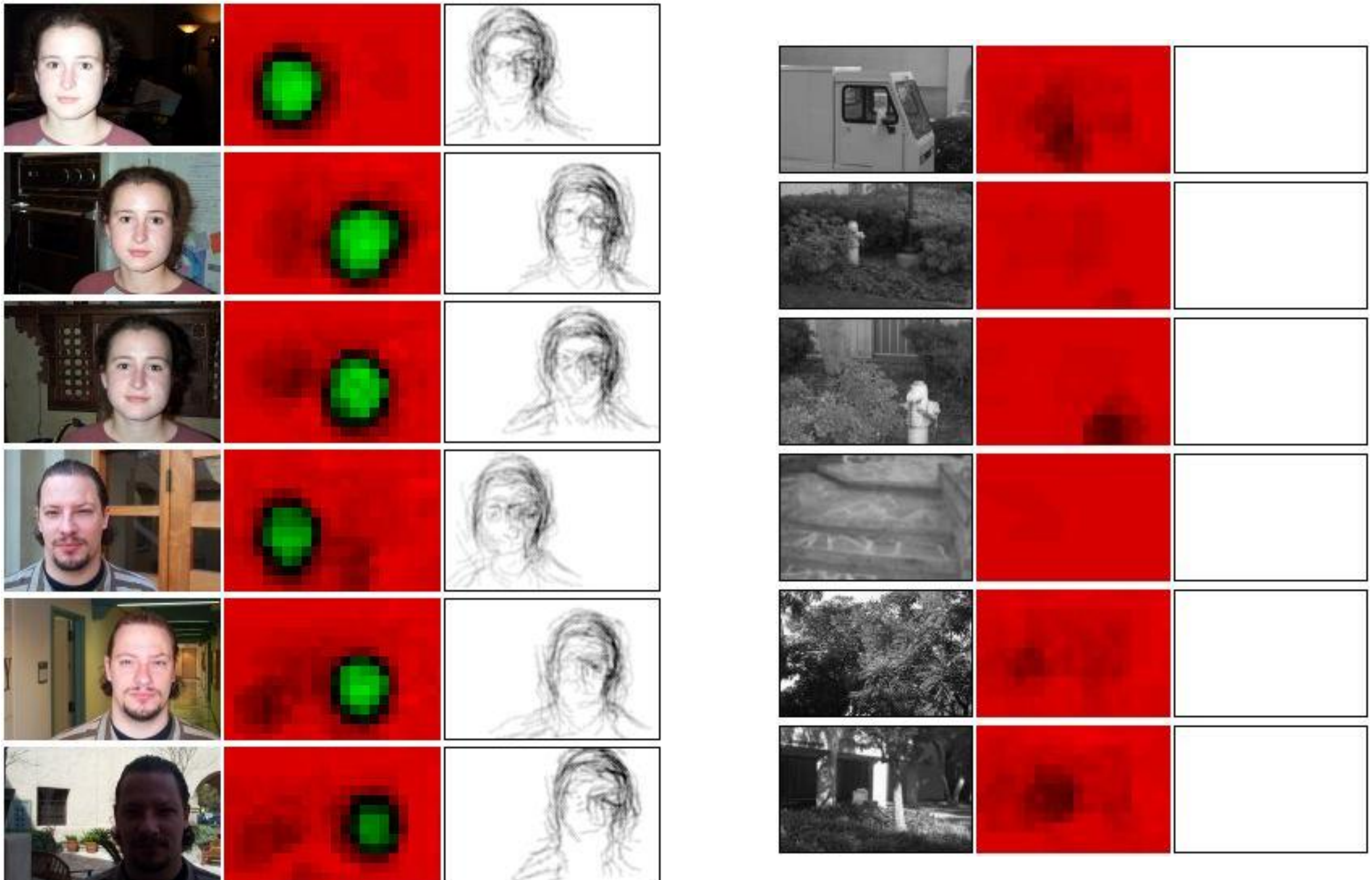
3. Weak-Learner Estimation

Select θ, a, b for each part

Results



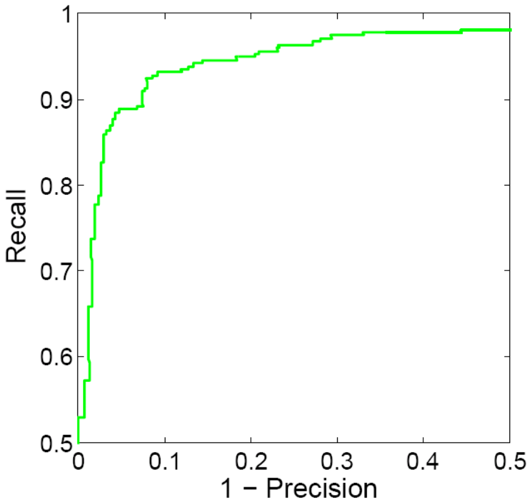
Results



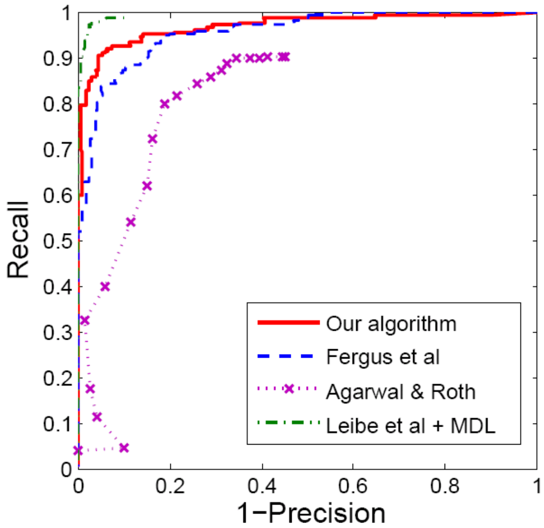
Results

- Quantification with *recall-precision* curves
  - shows trade-off between
    - correct detection rate
    - proportion of all detections that are correct
  - as a global detection threshold is changed

- Equal error rates (only 50 positive training images):
  - Weizmann Horses: 92.1%
  - UIUC Cars: 92.8%
  - Caltech Faces: 94.0%

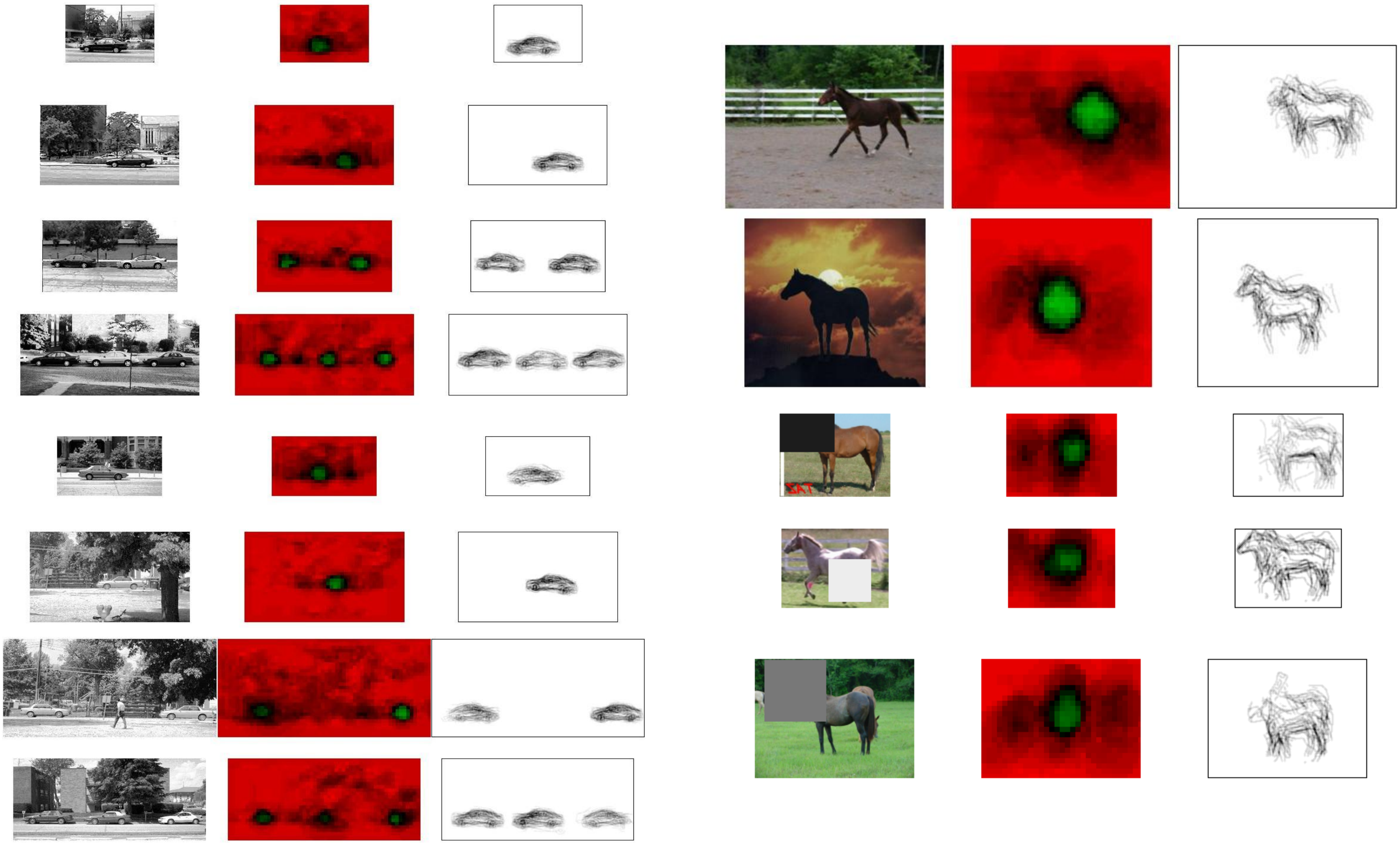


Horses

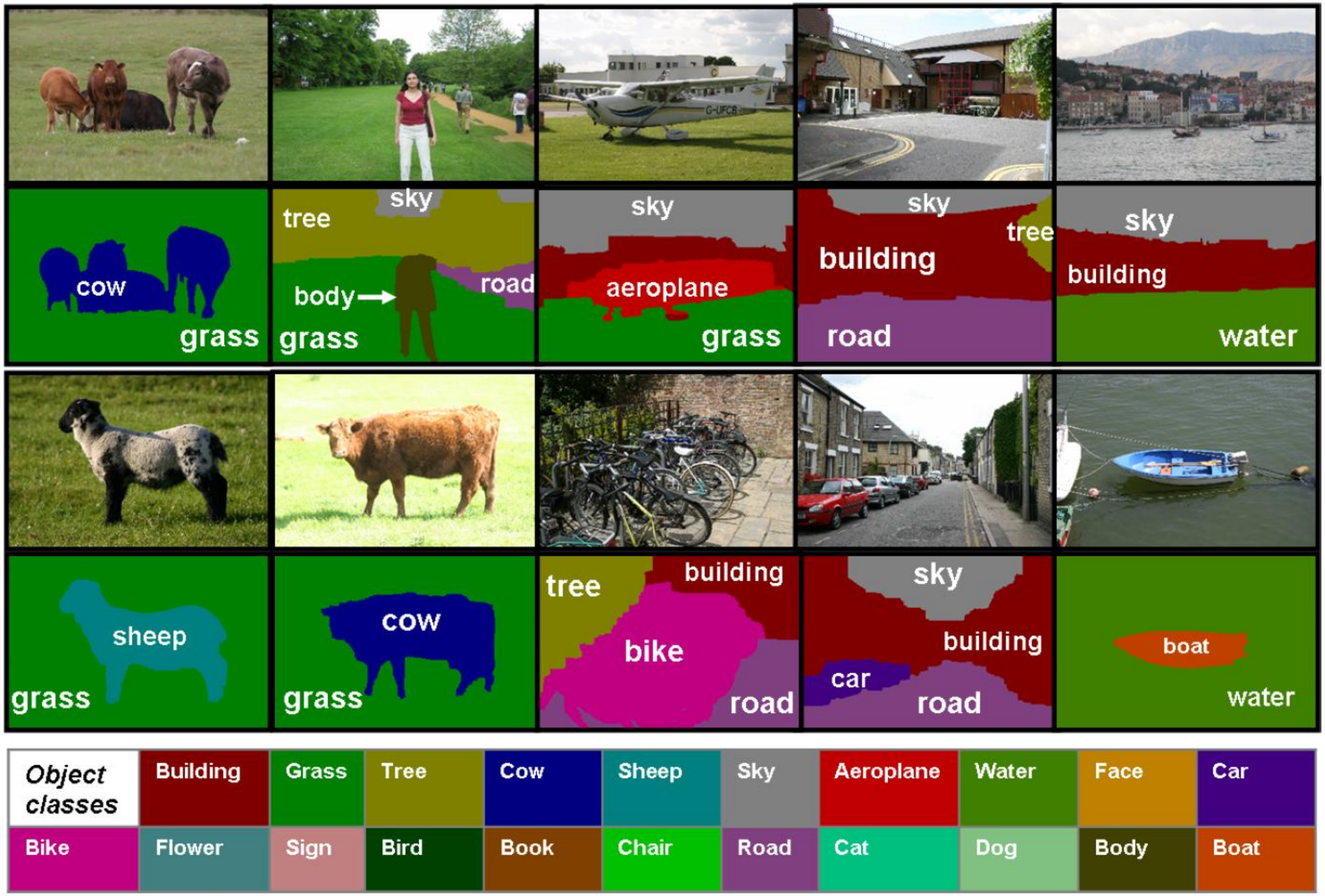


Cars

Results



Texture-based segmentation



Demos: Realtime mosaicing and editing

Summary

- 3D shape: making digital copies of sculpture from photographs from multiple viewpoints
- Recognition of a painting/picture from a single photo using a mobile (camera) phone
- Detection of objects: hands, faces and people
- Learning