UNIVERSITY OF CAMBRIDGE

Microsoft Research Cambridge

TextonBoost :

Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation

J. Shotton[*], **J. Winn**[†], **C. Rother**[†], and **A. Criminisi**[†]

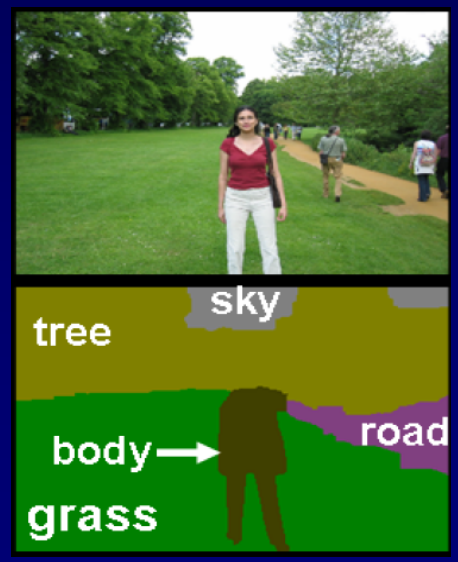
[*] University of Cambridge

[†] Microsoft Research Ltd, Cambridge, UK

Introduction

- Simultaneous recognition and segmentation
  - Explain every pixel (dense features)
  - Appearance + shape + context
  - Exploit class generalities + image specifics

- Contributions
  - New low-level features
  - New texture-based discriminative model
  - Efficiency and scalability



Example Results

Structure of Presentation

- The MSRC 21-Class Object Recognition Database
- New ‘Shape Filter’ Features
- Randomised boosting with Shared Features
- Adapting to the Pascal VOC Challenge

Image Databases

Object classes	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car
Bike	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat

- **MSRC 21-Class Object Recognition Database**
  - 591 hand-labelled images (45% train, 10% validation, 45% test)
- **Corel** (7-class) and **Sowerby** (7-class) [He *et al.* CVPR 04]

Sparse vs Dense Features

- Successes using sparse features, e.g.

  [Sivic *et al.* ICCV 2005], [Fergus *et al.* ICCV 2005], [Leibe *et al.* CVPR 2005]

- But…
  - do not explain whole image
  - cannot cope well with all object classes

- **We use *dense* features**
  - ‘shape filters’
  - local texture-based image descriptions

- Cope with
  - textured and untextured objects, occlusions, whilst retaining high efficiency

problem images for sparse features?

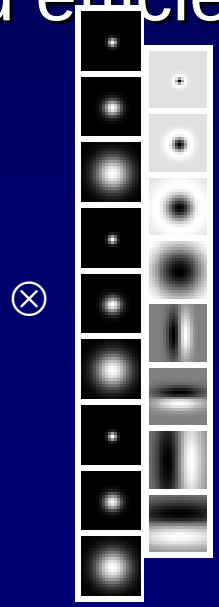
Textons

- Shape filters use *texton* maps [Varma & Zisserman IJCV 05]

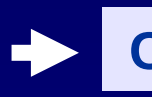
  [Leung & Malik IJCV 01]

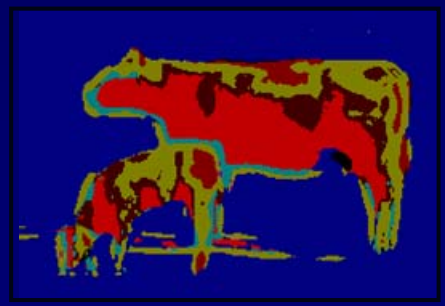
- Compact and efficient characterisation of local texture

Input image



Filter Bank

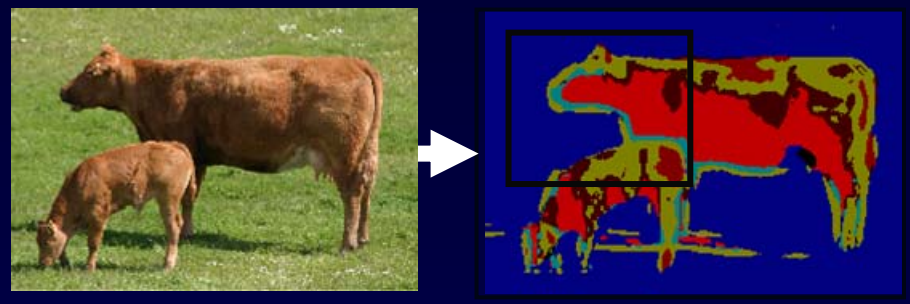


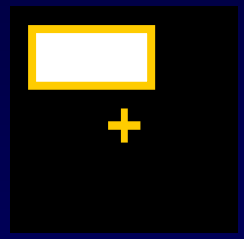
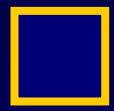


Texton map

Colours ⇔ Texton Indices

Shape Filters



- Pair: ( , )

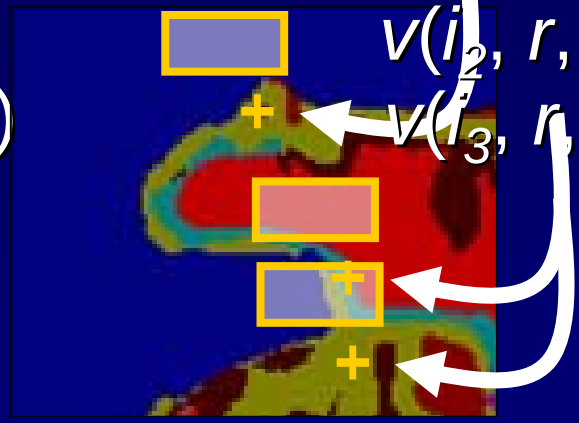
rectangle *r* texton *t*

- Feature responses $v(i, r, t)$

$v(i_1, r, t) = a$

$v(i_2, r, t) = 0$

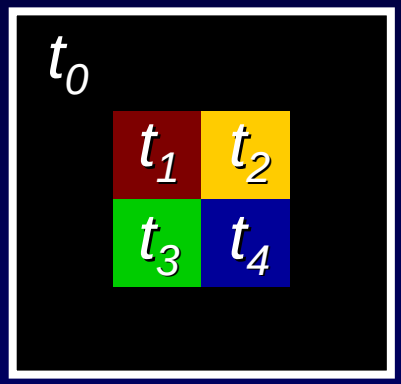
$v(i_3, r, t) = a/2$



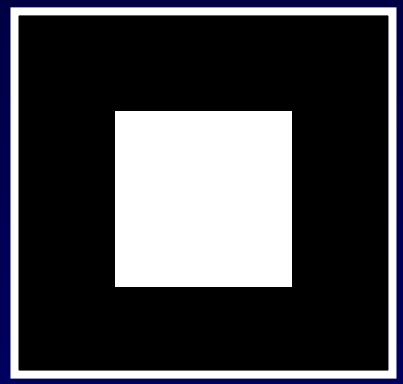
appearance context

- Integral images

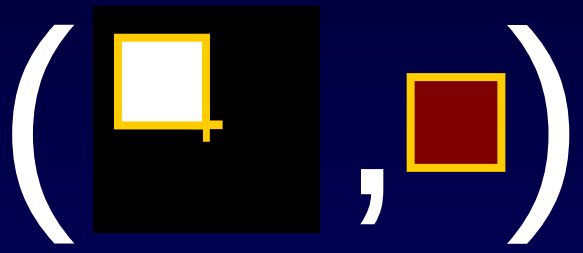
Shape and Appearance

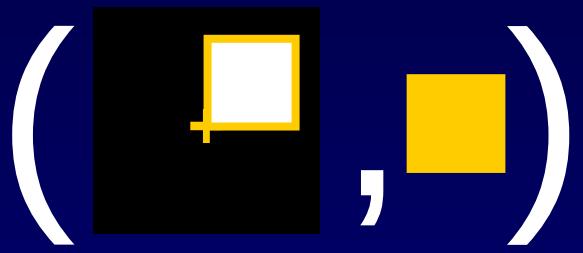


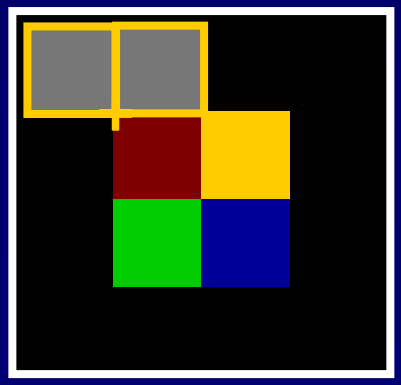
texton map



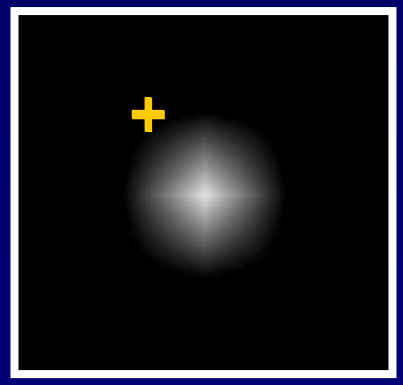
ground truth

$(r_1, t_1) =$ 

$(r_2, t_2) =$ 



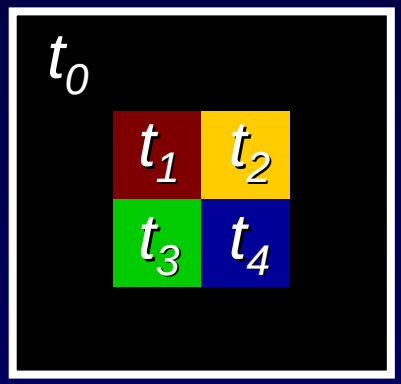
texton map



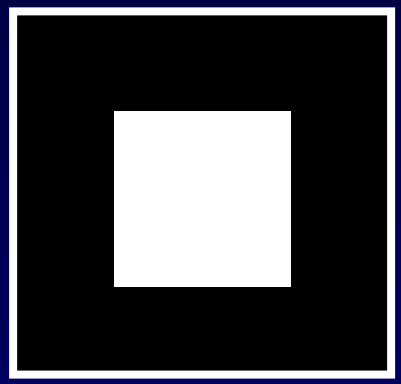
feature response image

$v(i, r_2, t_2)$

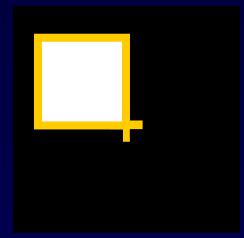
Shape and Appearance

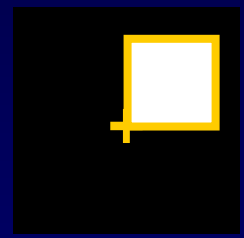


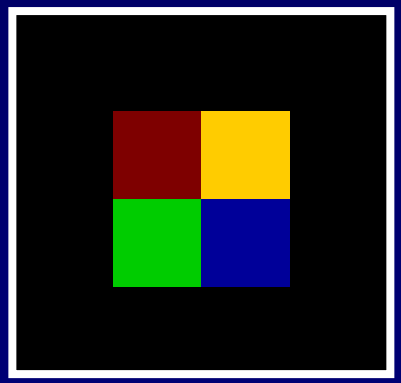
texton map



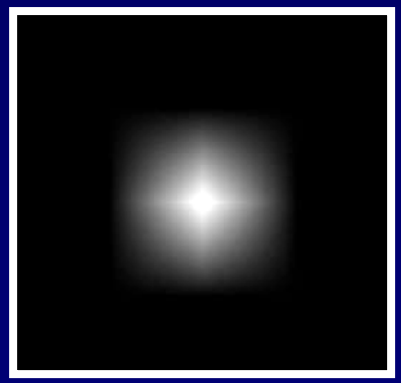
ground truth

$(r_1, t_1) =$ 

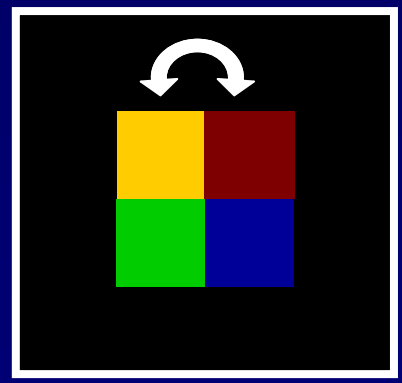
$(r_2, t_2) =$ 



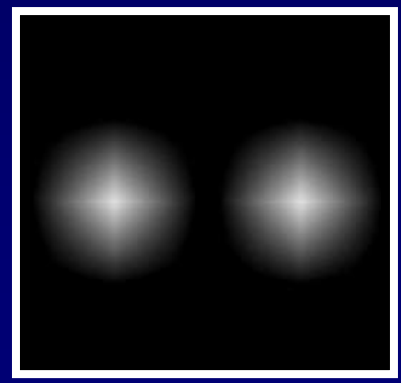
texton map



summed response images
$v(i, r_1, t_1) + v(i, r_2, t_2)$



texton map



summed response images
$v(i, r_1, t_1) + v(i, r_2, t_2)$

Shape-Texture Potentials

- Joint Boost algorithm [Torralba *et al.* CVPR 2004]
  - iteratively combines many shape filters
  - builds multi-class logistic classifier

- Resulting combination exploits:

Shape

Texture

Context (!)

- Shape-Texture potentials:

$$\underbrace{\psi_i(c_i, \mathbf{x}; \boldsymbol{\theta}_\psi)}_{\text{shape-texture potentials}} = \log \underbrace{\widetilde{P}_i(c_i|\mathbf{x})}_{\text{logistic classifier}}$$

Feature Selection by Boosting

30 rounds

1000 rounds

2000 rounds

input image

inferred segmentation
colour = most likely label

confidence
white = high entropy
black = low entropy

Object classes	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car
Bike	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat

Feature Selection by Boosting

30 rounds

1000 rounds

2000 rounds

input image

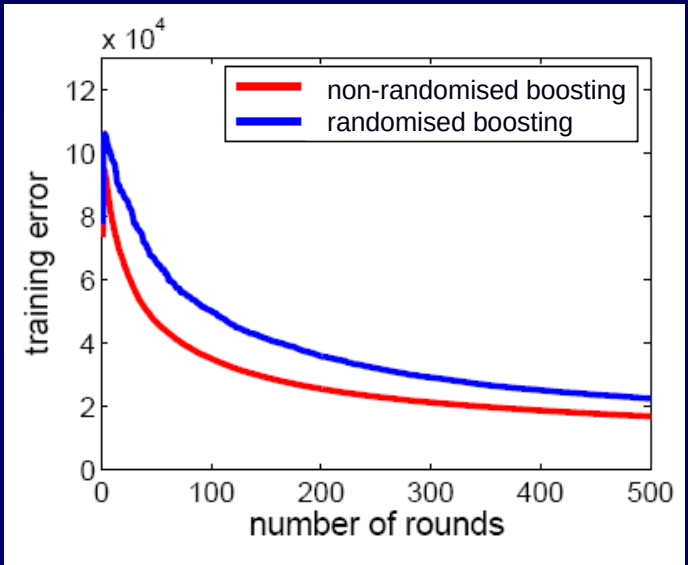
inferred segmentation
colour = most likely label

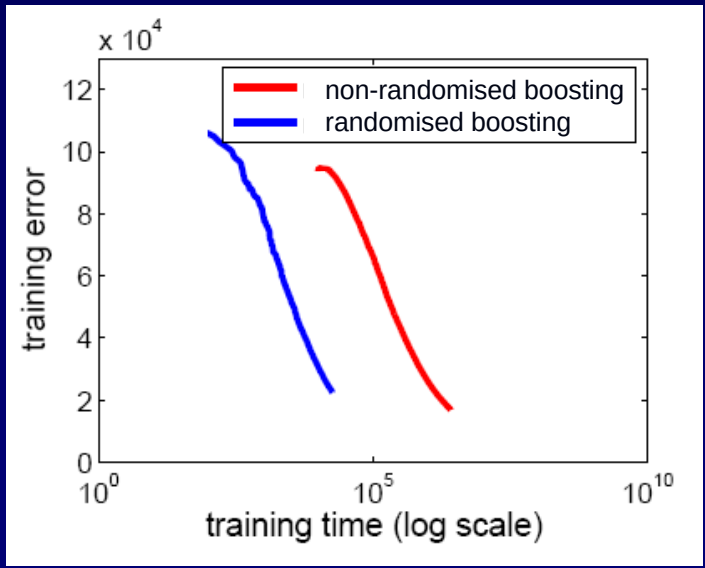
confidence
white = high entropy
black = low entropy

Object classes	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car
Bike	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat

Randomised Boosting

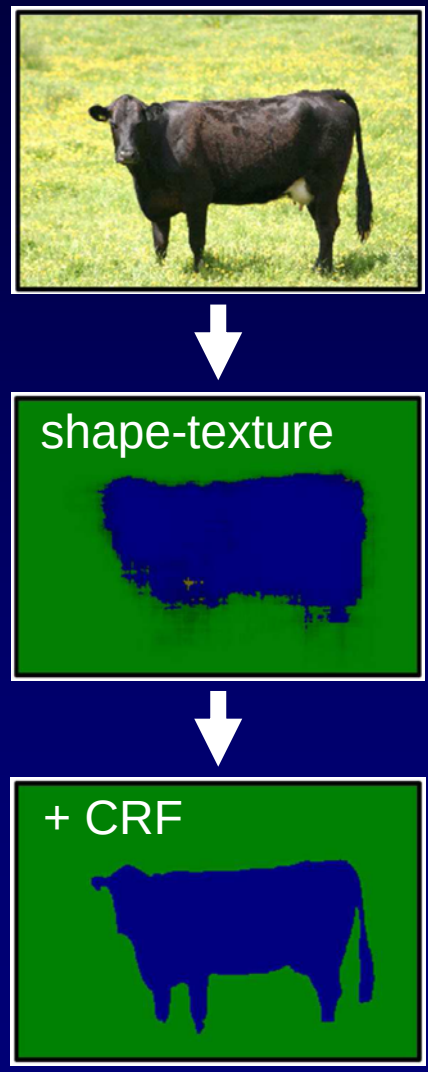
- Avoid expensive search over all features
  - only check random fraction (e.g. 0.3%) at each round
  - over several thousand rounds probably try all possible features





Accurate Segmentation?

- Shape-texture potentials alone
  - effectively recognise objects
  - but not sufficient for pixel-perfect segmentation
- Conditional Random Field (CRF) –
  see oral presentation tomorrow!



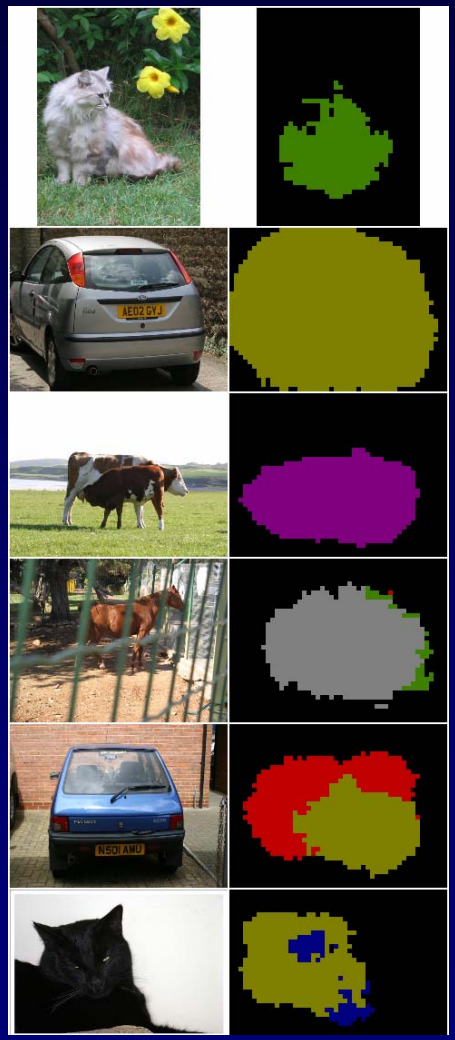
Adapting TextonBoost to the Pascal VOC Challenge

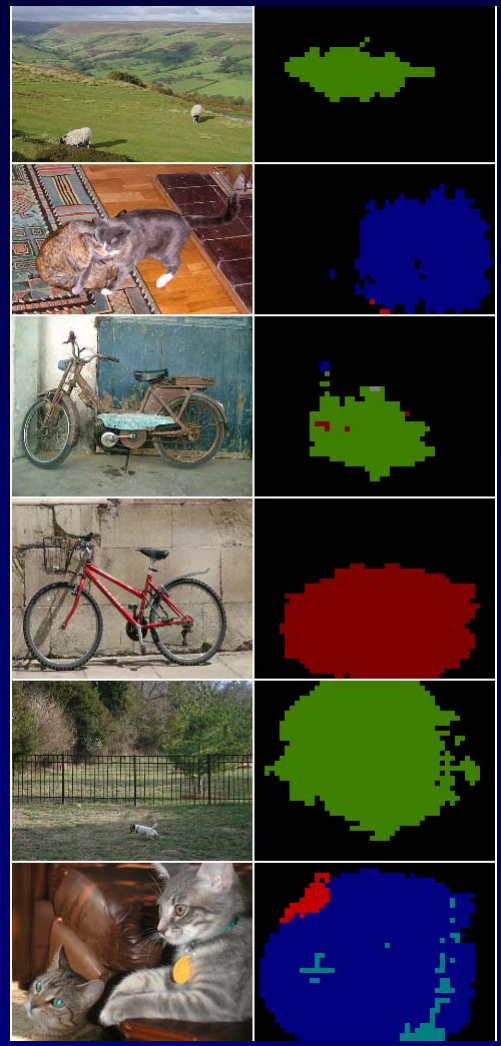
Training

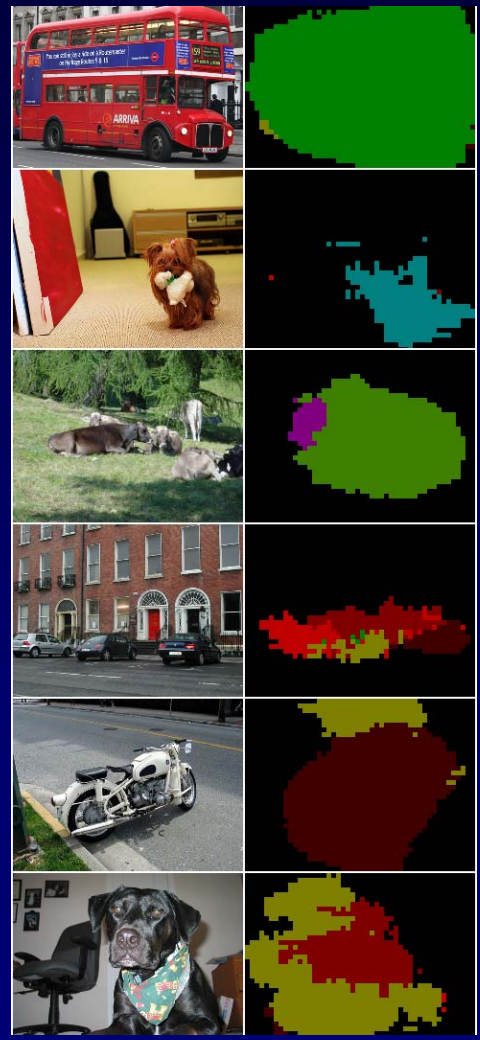
- Pascal training data is bounding boxes.
- Need pixelwise labelling – use GrabCut based on bounding box (noisy labelling!):

- Add ‘background’ label for non-object regions and train background class.
- ~1 day training time (for 10 classifiers on 1/3

Results








void
bicycle
bus
car
cat
cow
dog
horse
motorbike
person
sheep

Classification (competition 1)

- To give uncertainty measure, use only boosted softmax classifier and normalised sum of classifier over all image pixels.

Area under curve (AUC)

bicycle	bus	car	cat	cow	dog	horse	motorbike	person	sheep
0.873	0.864	0.887	0.822	0.850	0.768	0.754	0.844	0.715	0.866

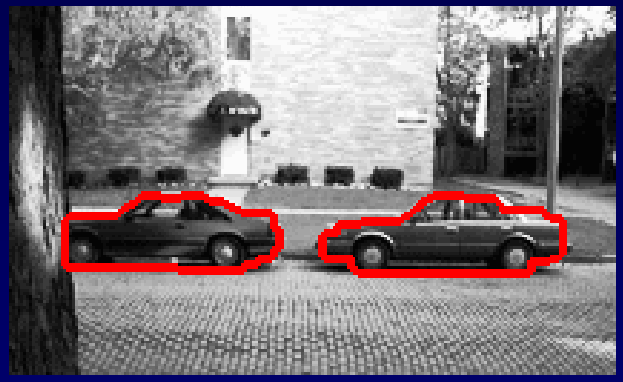
- Test time: 30sec image (three seconds per classifier)

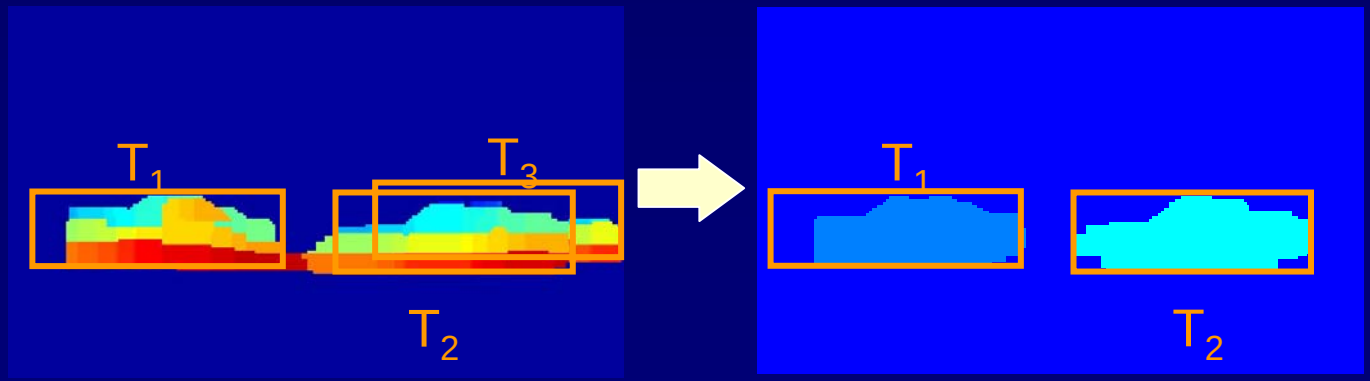
VOC experiments by Jamie Shotton

Detection (competition 3)

- **Work in progress:** scale/viewpoint invariant Layout Consisent Random Field

Input image





Layout-consistent regions

Instance labelling

Detection (competition 3)

- **Work in progress:** scale/viewpoint invariant Layout Consisent Random Field
- Instead, used connected-components of most probable labelling (ignoring if <1000 pixels) and then computed normalised sum (as before)



Average precision (AP)

bicycle	bus	car	cat	cow	dog	horse	motorbike	person	sheep
0.249	0.13	0.25	0.151	0.14	0.11	0.09	0.178	0.030	0.13
	8	4		0	8	1			1

Suggestions for Pascal VOC 2007

- Include other types of object classes:
  - unstructured classes (e.g. sky, grass)
  - semi-structured classes (e.g. building).
- Have small number of pixel-wise labelled images and include a segmentation competition.
- Keep it hard!!!

Thank you

TextonBoost code will be available shortly from
http://mi.eng.cam.ac.uk/~jdjs2/