What, Where & How Many?
Combining Object Detectors and CRFs

Ľubor Ladický, Paul Sturgess, Karteek Alahari,
Christopher Russell, Philip H.S. Torr

Oxford Brookes
Vision Group

http://cms.brookes.ac.uk/research/visiongroup/

Complete Scene Understanding

Involves

- Localization of all instances of foreground objects (“things”)
- Localization of all background classes (“stuff”)
- Pixel-wise segmentation
- 3D reconstruction
- Pose detection
- Action recognition
- Event recognition

Complete Scene Understanding

Oxford Brookes Vision Group

Involves

- Localization of all instances of foreground objects ("things")
- Localization of all background classes ("stuff")
- Pixel-wise segmentation

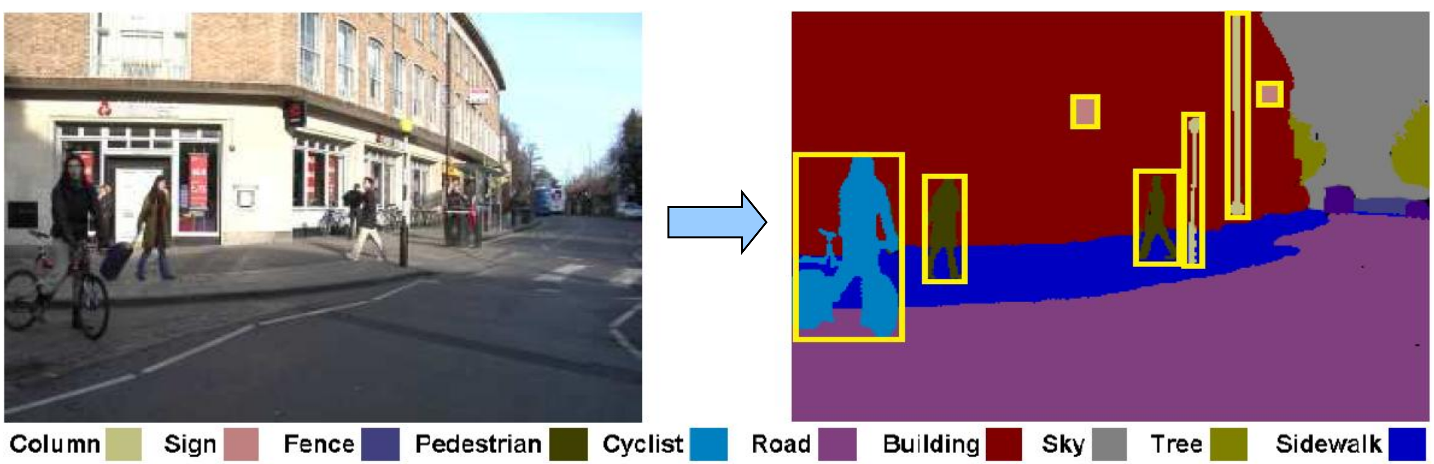
Semantic Segmentation

- 3D reconstruction
- Pose detection
- Action recognition
- Event recognition

Complete Scene Understanding

Involves

- Localization of all instances of foreground objects (“things”)
- Localization of all background classes (“stuff”)
- Pixel-wise segmentation



Semantic Scene Understanding

We're interested in whole scene understanding
Given an image, detect every *thing* in it.

Thing : An object with a specific size and shape.

Adelson, Forsyth et al. 96

We're interested in whole scene understanding
Given an image, detect every *thing* in it.

Thing : An object with a specific size and shape.

Adelson, Forsyth et al. 96

Semantic Scene Understanding

We're interested in whole scene understanding
Given an image, detect every *thing* in it.

Thing : An object with a specific size and shape.

Adelson, Forsyth et al. 96

Semantic Scene Understanding

Oxford Brookes Vision Group

We're interested in whole scene understanding
Given an image, detect every *thing* in it.

Thing : An object with a specific size and shape.

Adelson, Forsyth et al. 96

Semantic Scene Understanding

We're interested in whole scene understanding
Given an image, detect every *thing* in it.

Thing : An object with a specific size and shape.

Adelson, Forsyth et al. 96

We're interested in whole scene understanding

Given an image, label all the *stuff*

Stuff : Material defined by a homogeneous or repetitive pattern, with no specific spatial extent / shape.

Adelson, Forsyth et al. 96

Semantic Scene Understanding

We're interested in whole scene understanding
Given an image, label all the *stuff*

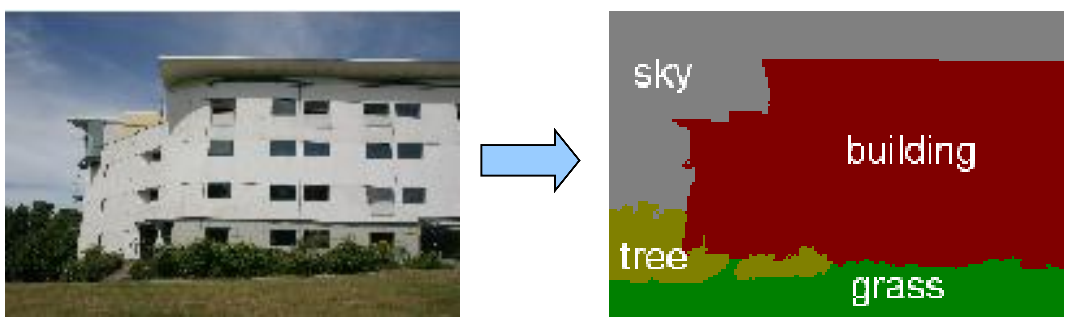
Stuff : Material defined by a homogeneous or repetitive pattern, with no specific spatial extent / shape.

Adelson, Forsyth et al. 96

Our plan is to combine

- State of the art sliding window object detection

- State of the art segmentation techniques



Algorithms for Object Localization

Sliding window detectors

- HOG descriptor (Dalal & Triggs CVPR05)
- Based on histograms of features (Vedaldi et al. ICCV09)
- Part-based models (Felzenszwalb et al. CVPR09)

Algorithms for Object Localization

Oxford Brookes Vision Group

Sliding window detectors

- HOG descriptor (Dalal & Triggs CVPR05)
- Based on histograms of features (Vedaldi et al. ICCV09)
- Part-based models (Felzenszwalb et al. CVPR09)

Algorithms for Object Localization

Oxford Brookes Vision Group

Sliding window detectors

- HOG descriptor (Dalal & Triggs CVPR05)
- Based on histograms of features (Vedaldi et al. ICCV09)
- Part-based models (Felzenszwalb et al. CVPR09)

Algorithms for Object Localization

Non-maxima suppression

- Greedily
- Using field of indicator variables (Ramanan et al ICCV09)
  - One binary variable $y_i \in \{ 0, 1 \}$ per location/scale

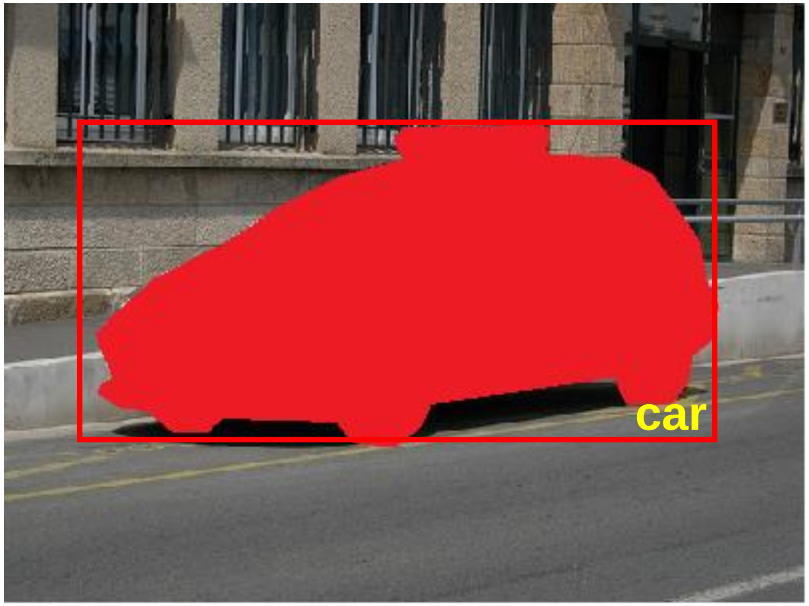
Algorithms for Object Localization

Oxford Brookes Vision Group

Non-maxima suppression

- Greedily
- Using field of indicator variables (Ramanan et al ICCV09)
  - One binary variable $y_i \in \{ 0, 1 \}$ per location/scale

Sliding window detectors



- Sliding window + Segmentation
  - OBJCUT (Kumar et al. 05)
  - Updating colour model (GrabCut - Rother et al. 04)

Sliding window detectors not good for "stuff"

Try to detect "sky" !

Sliding window detectors not good for “stuff”

Sky is irregular shape not suited
to the sliding window approach

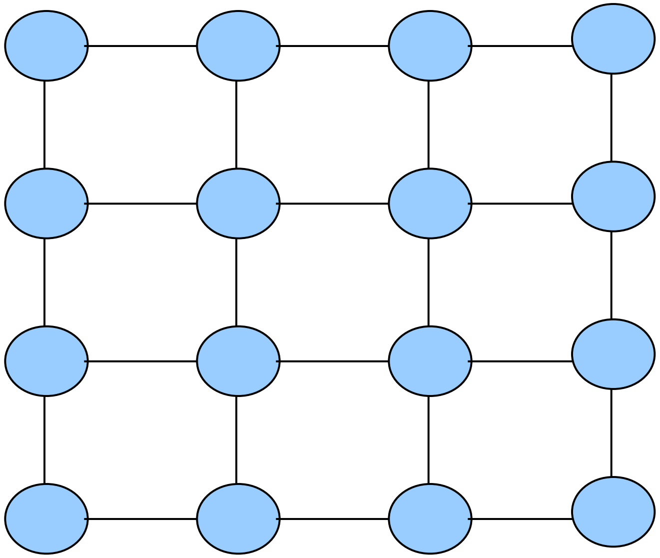
Sliding window detectors

- State-of-the-art for object detection (“things”)
- Do not work for background classes (“stuff”)
  - No distinct shape
  - Cannot be enclosed in a box
- Cannot recover from incorrect detections

Pairwise CRF over pixels

Input image

CRF construction



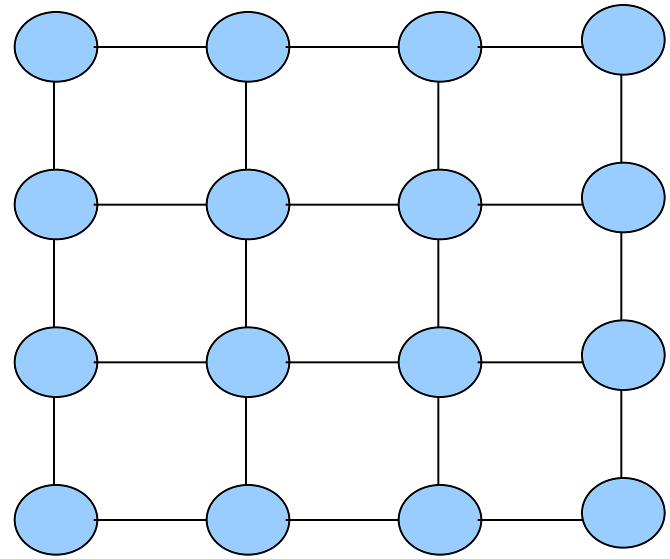
Final segmentation

Shotton et al. ECCV06

Pairwise CRF over pixels

Input image

CRF construction



Training of Potentials

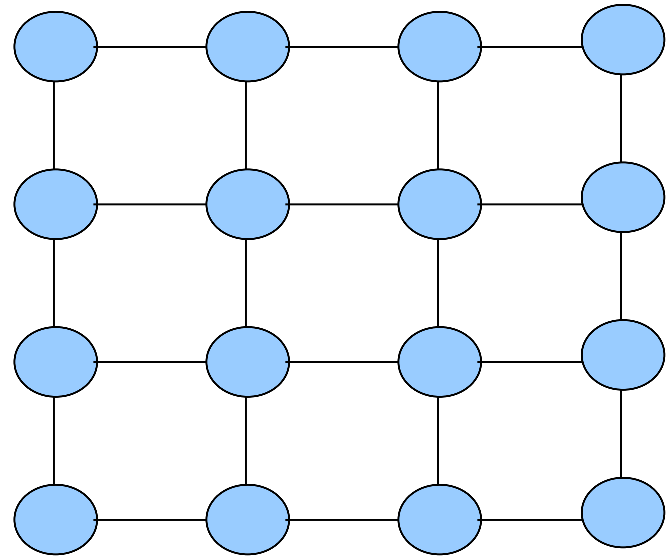
$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j)$$

Shotton et al. ECCV06

Pairwise CRF over pixels

Input image

CRF construction



Training of Potentials

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j)$$

MAP

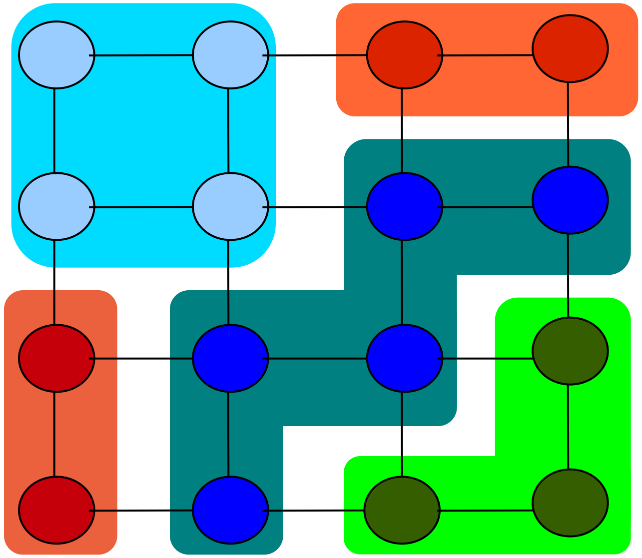
Final segmentation

Shotton et al. ECCV06

Pairwise CRF over Super-pixels / Segments

Input image

Unsupervised segmentation

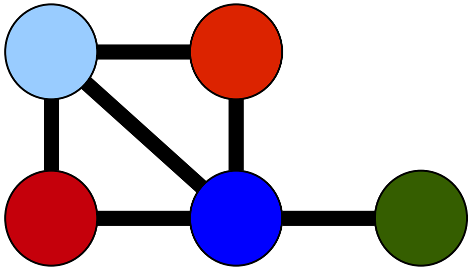


Shi, Malik PAMI2000, Comaniciu, Meer PAMI2002, Felzenszwalb, Huttenlocher, IJCV2004

Pairwise CRF over Super-pixels / Segments

Input image

Unsupervised segmentation

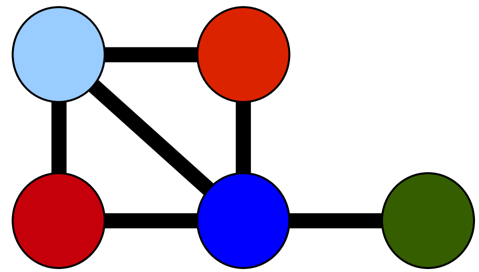


Shi, Malik PAMI2000, Comaniciu, Meer PAMI2002, Felzenszwalb, Huttenlocher, IJCV2004

Pairwise CRF over Super-pixels / Segments

Input image

Unsupervised segmentation



Training of potentials

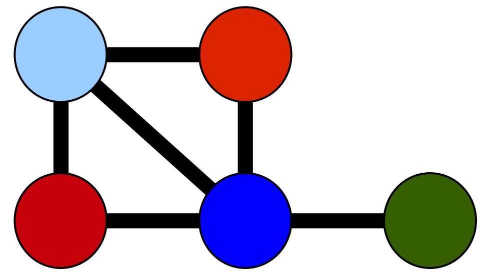
$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j)$$

Batra et al. CVPR08, Yang et al. CVPR07, Zitnick et al. CVPR08, Rabinovich et al. ICCV07, Boix et al. CVPR10

Pairwise CRF over Super-pixels / Segments

Input image

Unsupervised segmentation



Training of potentials

MAP

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j)$$

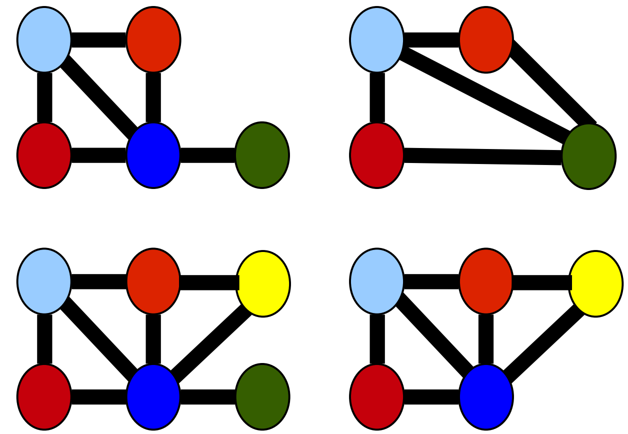
Final segmentation

Batra et al. CVPR08, Yang et al. CVPR07, Zitnick et al. CVPR08, Rabinovich et al. ICCV07, Boix et al. CVPR10

Associative Hierarchical CRF

Input image

Multiple segmentations or hierarchies

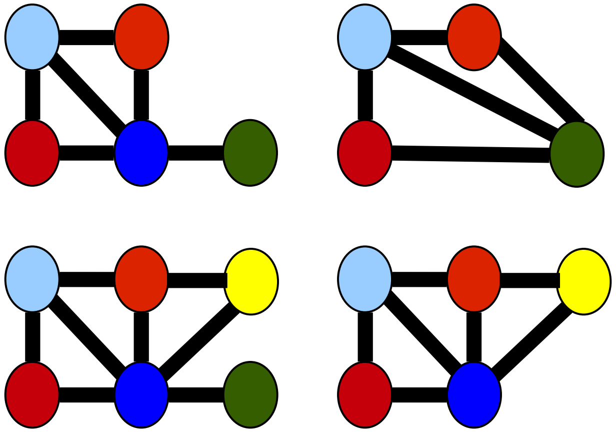


Ladický et al. ICCV09

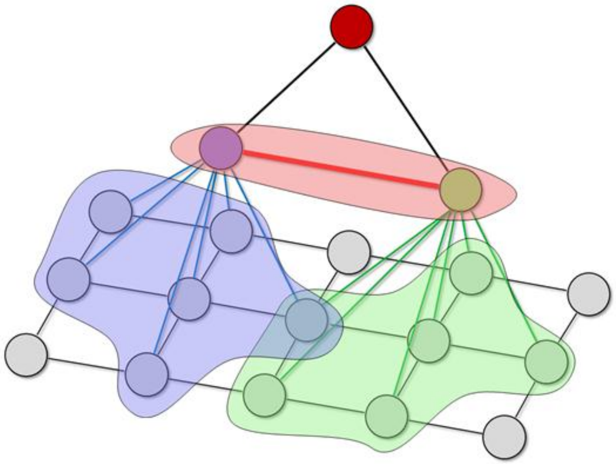
Associative Hierarchical CRF

Input image

Multiple segmentations or hierarchies



CRF construction

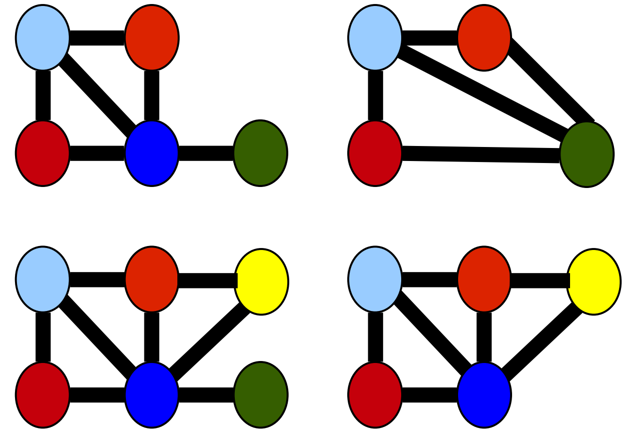


Ladický et al. ICCV09

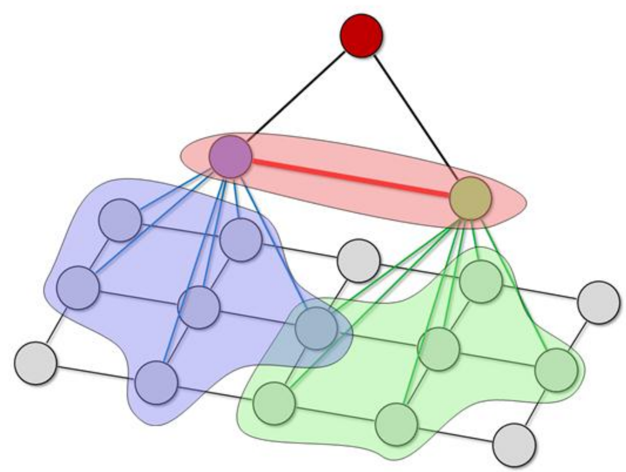
Associative Hierarchical CRF

Input image

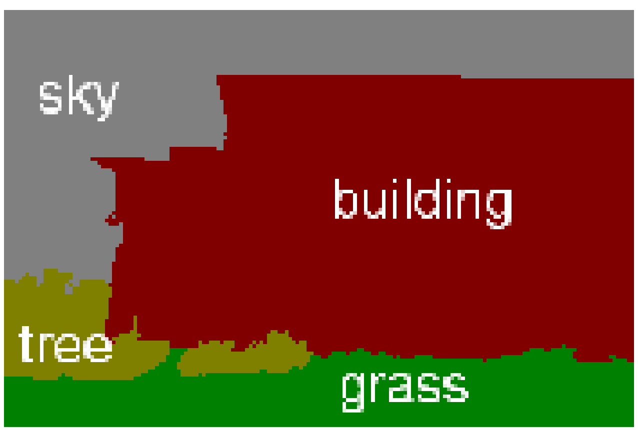
Multiple segmentations or hierarchies



CRF construction



MAP



Final segmentation

Ladický et al. ICCV09, Russell et al. UAI10

Associative Hierarchical CRF

- State-of-the-art performance on MSRC & CamVid
- Merges information at multiple scales
- Can recover from incorrect segmentations

Ladický et al. ICCV09, Sturgess et al., BMVC09

Associative Hierarchical CRF

- State-of-the-art performance on MSRC & CamVid
- Merges information at multiple scales
- Can recover from incorrect segmentations

However,

- No concept of "things"
- Cannot distinguish between multiple instances

Ladický et al. ICCV09, Sturgess et al., BMVC09

CRF Formulation with Detectors

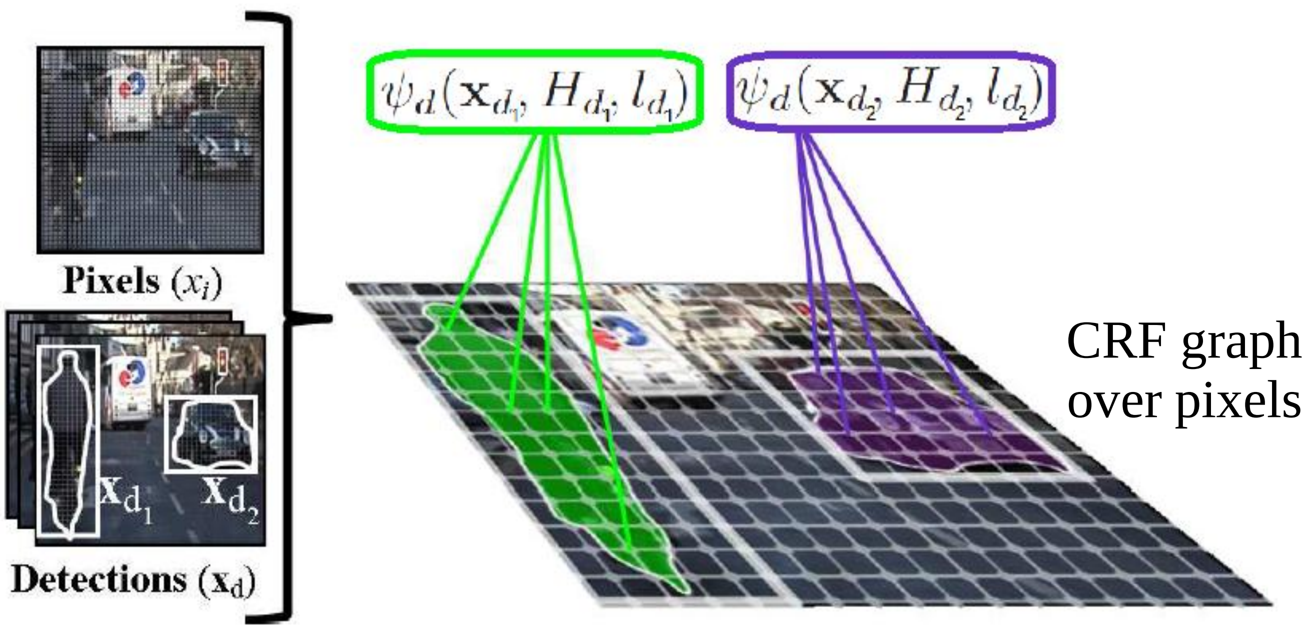
- CRF formulation altered with a potential for each detection

CRF Formulation with Detectors

- CRF formulation altered with a potential for each detection

$$E(\mathbf{x}) = E_{pix}(\mathbf{x}) + \sum_{d \in \mathcal{D}} \psi_d(\mathbf{x}_d, H_d, l_d)$$

AH-CRF energy
without detectors



CRF Formulation with Detectors

- CRF formulation altered with a potential for each detection

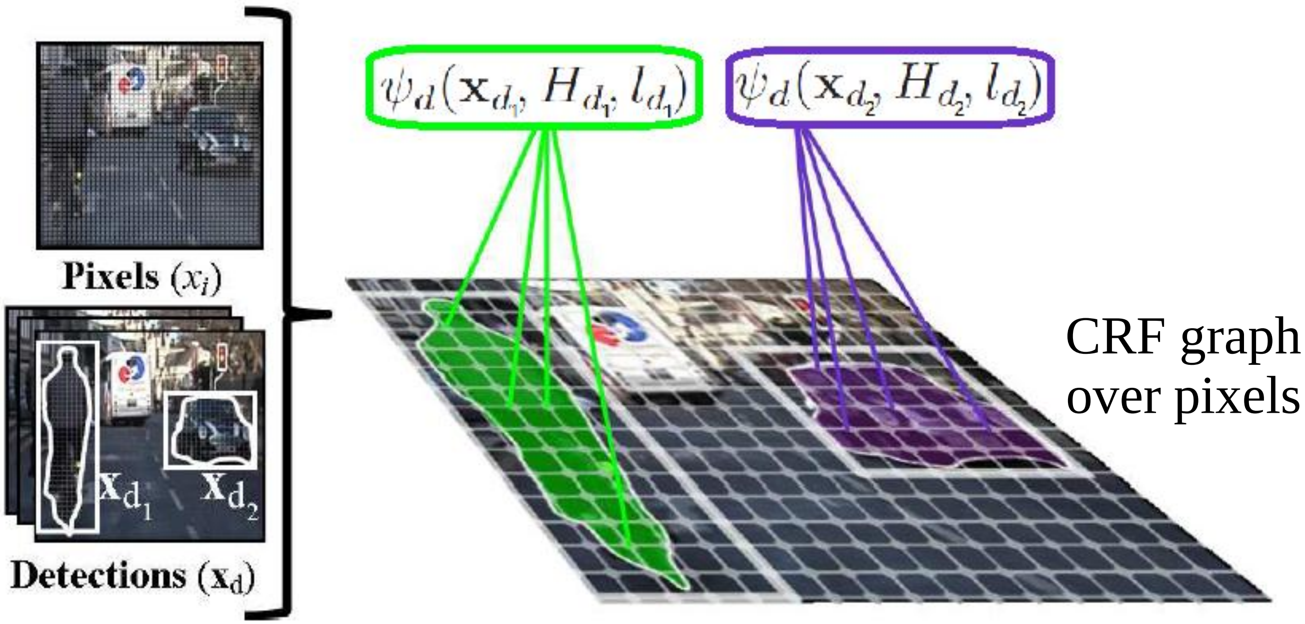
$$E(\mathbf{x}) = E_{pix}(\mathbf{x}) + \sum_{d \in \mathcal{D}} \psi_d(\mathbf{x}_d, H_d, l_d)$$

AH-CRF energy without detectors

Set of pixels of d-th detection

Classifier response

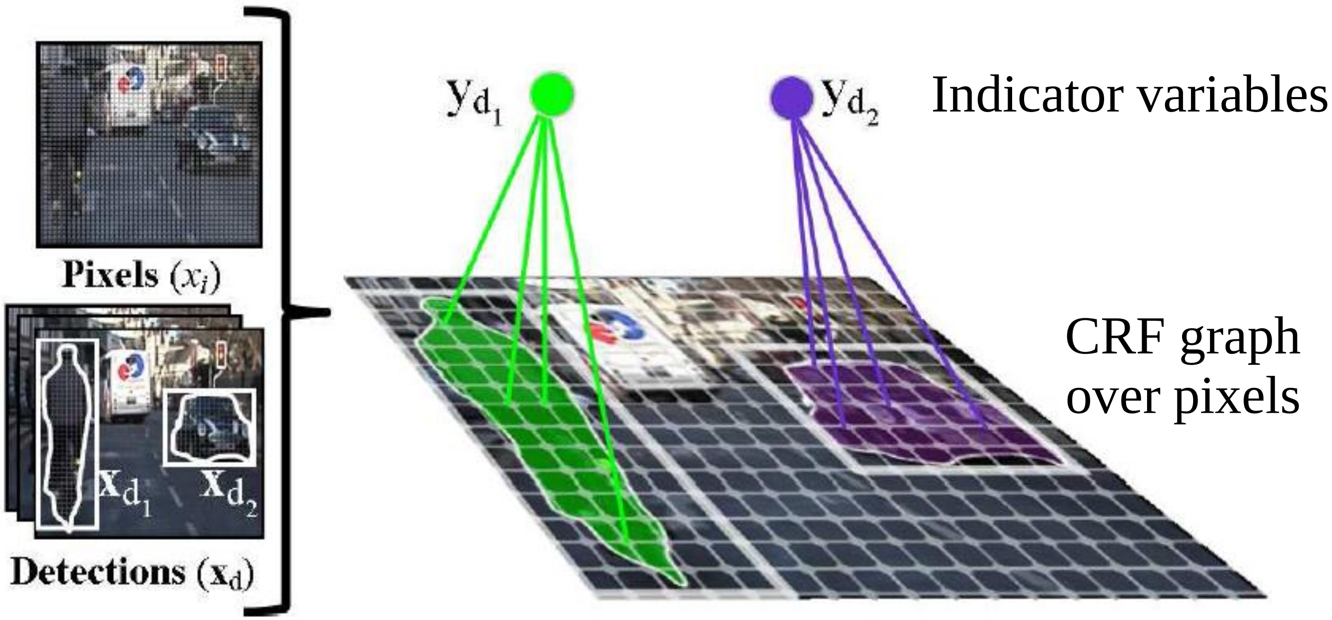
Detected label



CRF Formulation with Detectors

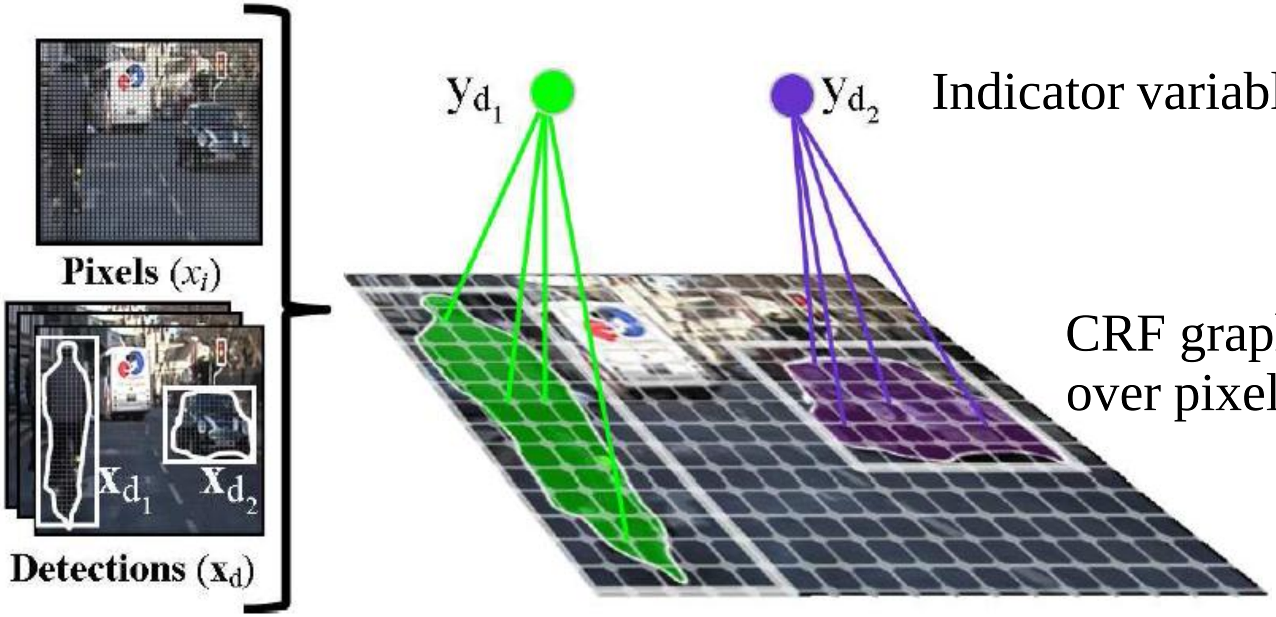
- Joint CRF formulation should contain
  - Possibility to reject detection hypothesis
  - Recover the status of the detection (*0 / 1*)
- Thus, potential is a minimum over indicator variable $y_d \in \{0, 1\}$

$$\psi_d(\mathbf{x}_d, H_d, l_d) = \min_{y_d} \phi_d(y_d, \mathbf{x}_d, H_d, l_d)$$



CRF Formulation with Detectors

- Detection potential should decrease the energy of labelling agreeing with the detection hypothesis
- Partial disagreement penalized but not directly rejects the hypothesis



CRF Formulation with Detectors

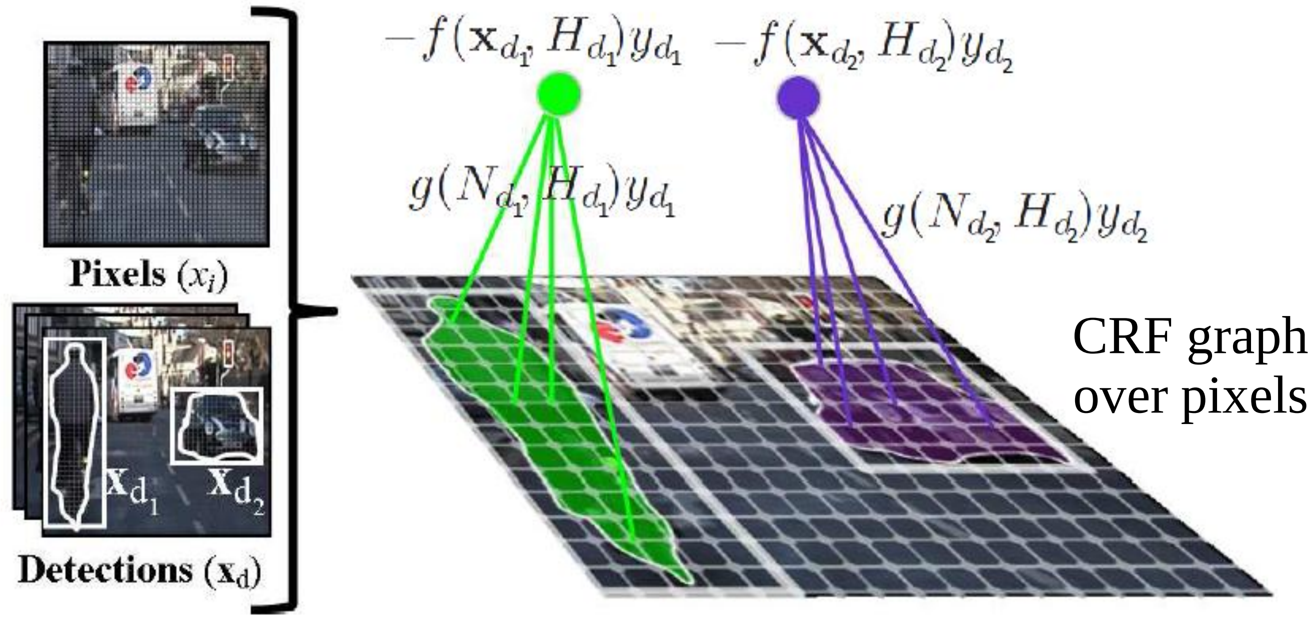
- Detection potential should decrease the energy of labelling agreeing with the detection hypothesis
- Partial disagreement penalized but not directly rejects the hypothesis

$$\psi_d(\mathbf{x}_d, H_d, l_d) = \min_{y_d}(-f(\mathbf{x}_d, H_d)y_d + g(N_d, H_d)y_d)$$

Detection hypothesis status

Detection strength

Partial inconsistency cost



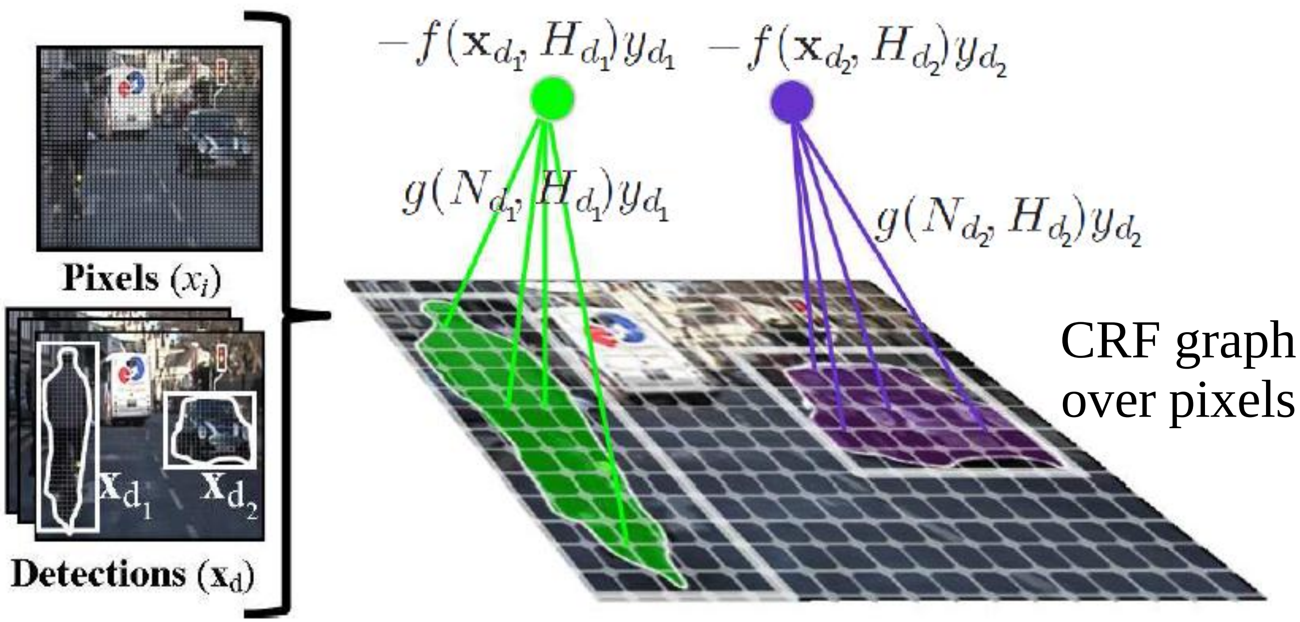
CRF Formulation with Detectors

- Detection potential should decrease the energy of labelling agreeing with the detection hypothesis
- Partial disagreement penalized but not directly rejects the hypothesis

$$\psi_d(\mathbf{x}_d, H_d, l_d) \;= \min_{y_d}(-f(\mathbf{x}_d, H_d)y_d + g(N_d, H_d)y_d)$$

Linear thresholded

Linear



Inference over CRF with Detectors

- This higher order potential can be transformed to

$$\psi_d(\mathbf{x}_d, H_d) = \min_{y_d \in \{0,1\}} (-f(\mathbf{x}_d, H_d) y_d + k_d N_d y_d)$$
$$= -f(\mathbf{x}_d, H_d) + \min(f(\mathbf{x}_d, H_d), k_d \sum_{j \in \mathbf{x}_d} \delta(x_j \neq l_d))$$

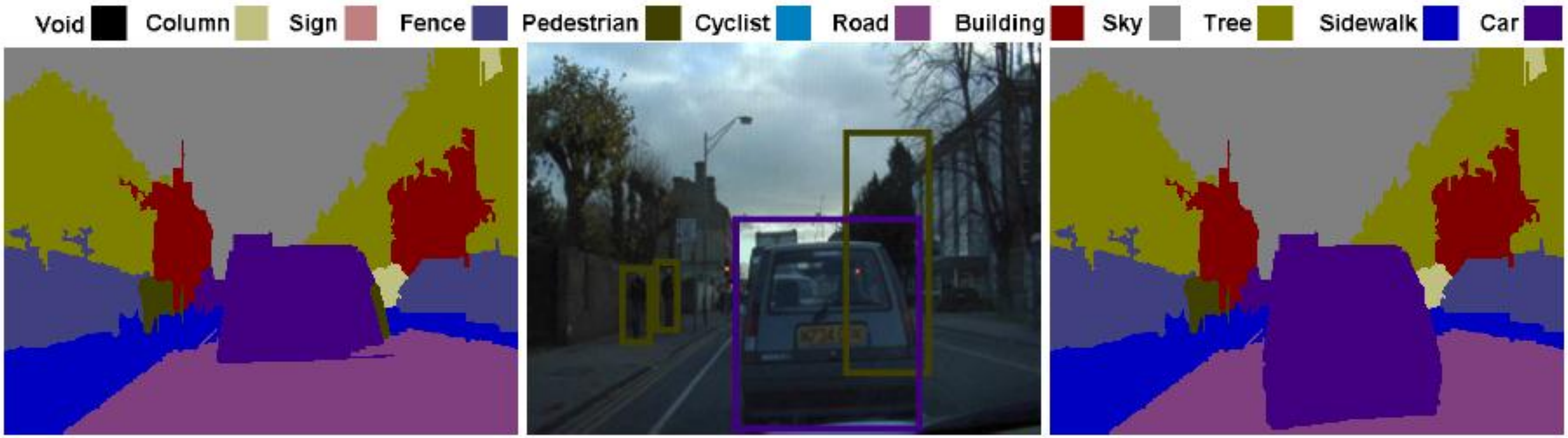
which take the form of Robust P^N (Kohli et al. CVPR08)

$$\psi_h(\mathbf{x}) = \min(\gamma_{max}, \min_l(\gamma_l + k_l \sum_{i \in \mathbf{x}} \delta(x_i \neq l)))$$

- Solvable with all graph cut-based methods

Results on CamVid dataset

Oxford Brookes Vision Group



Result without detections — Set of detections — Final Result

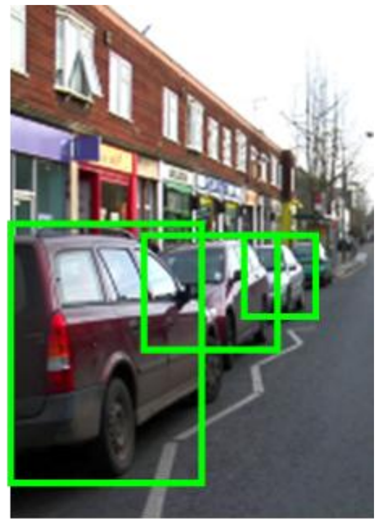
	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Sidewalk	Bicyclist	**Global**	**Average**
Brostow et al.	46.2	61.9	89.7	68.6	**42.9**	89.5	**53.6**	46.6	0.7	60.5	22.5	69.1	53.0
Sturgess et al.	**84.5**	72.6	**97.5**	72.7	34.1	**95.3**	34.2	45.7	8.1	77.6	28.5	**83.8**	59.2
Our method	81.5	**76.6**	96.2	**78.7**	40.2	93.9	43.0	**47.6**	**14.3**	**81.5**	**33.9**	**83.8**	**62.5**

Brostow et al. ECCV08, Sturgess et al. BMVC09

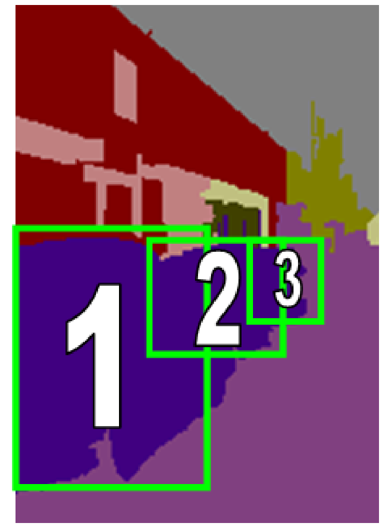
Results on CamVid dataset

Oxford Brookes Vision Group

Result without detections



Set of detections



Final Result

Also provides number of object instances (using y_d's)

Results on VOC2009 dataset

Input image	CRF without detectors	CRF with detectors	Input image	CRF without detectors	CRF with detectors

	Background	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dining table	Dog	Horse	Motor bike	Person	Potted plant	Sheep	Sofa	Train	TV/monitor	Average
BONN_SVM-SEGM	**83.9**	64.3	21.8	21.7	**32.0**	**40.2**	**57.3**	49.4	**38.8**	5.2	**28.5**	22.0	19.6	33.6	45.5	33.6	27.3	**40.4**	18.1	33.6	**46.1**	**36.3**
CVC_HOCRF	80.2	**67.1**	**26.6**	**30.3**	31.6	30.0	44.5	41.6	25.2	5.9	27.8	11.0	23.1	**40.5**	**53.2**	32.0	22.2	37.4	**23.6**	40.3	30.2	34.5
UOCTTI_LSVM-MDPM	78.9	35.3	22.5	19.1	23.5	36.2	41.2	50.1	11.7	8.9	**28.5**	1.4	5.9	24.0	35.3	33.4	**35.1**	27.7	14.2	34.1	41.8	29.0
NECUIUC_CLS-DTCT	81.8	41.9	23.1	22.4	22.0	27.8	43.2	**51.8**	25.9	4.5	18.5	18.0	**23.5**	26.9	36.6	**34.8**	8.8	28.3	14.0	35.5	34.7	29.7
LEAR_SEGDET	79.1	44.6	15.5	20.5	13.3	28.8	29.3	35.8	25.4	4.4	20.3	1.3	16.4	28.2	30.0	24.5	12.2	31.5	18.3	28.8	31.9	25.7
BROOKESMSRC_AHCRF	79.6	48.3	6.7	19.1	10.0	16.6	32.7	38.1	25.3	5.5	9.4	25.1	13.3	12.3	35.5	20.7	13.4	17.1	18.4	37.5	36.4	24.8
Our method	81.2	46.1	15.4	24.6	20.9	36.9	50.0	43.9	28.4	**11.5**	18.2	**25.4**	14.7	25.1	37.7	34.1	27.7	29.6	18.4	**43.8**	40.8	32.1

Summary

Oxford Brookes Vision Group

- Novel formulation of CRF with detectors
- Can recover from incorrect detections
- Possible to obtain instances of objects
- Efficiently solvable

Future and ongoing work

- Automatic non-maximum suppression
- Part-based models in CRF
- New things & stuff dataset available soon!
- Questions?