

*Mobile Robotics
Group*



SEAS DTC
*Systems Engineering for Autonomous
Systems*

An Approach to Spatio-Temporally Consistent Scene Classification in Urban Workspaces

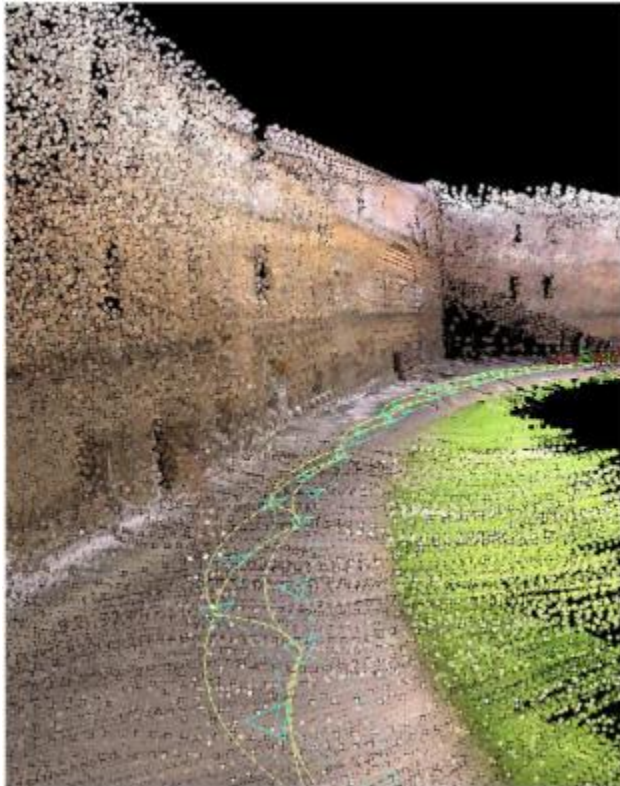
Ingmar Posner, Mark Cummins and Paul Newman

Mobile Robotics Group

Oxford University

From Numbers to Meaning...

- *We can easily build coloured point clouds, but we want more.*
- *We want scene labels...*



... to aid with

- Navigation and planning.
- Action selection.
- Human-machine interaction.



Appearance can *augment* metric/topological approaches.

Ideally, the classification framework should be:

- Principled.
 - Probabilistic.
- Introspective.
 - Able to handle context.
 - Know what measurements to trust (detector model).
- Flexible.
 - Adapt (learn) class models online and in real-time.
- Fast.
 - Classify in real-time.

Goal: Find a framework which delivers on these.

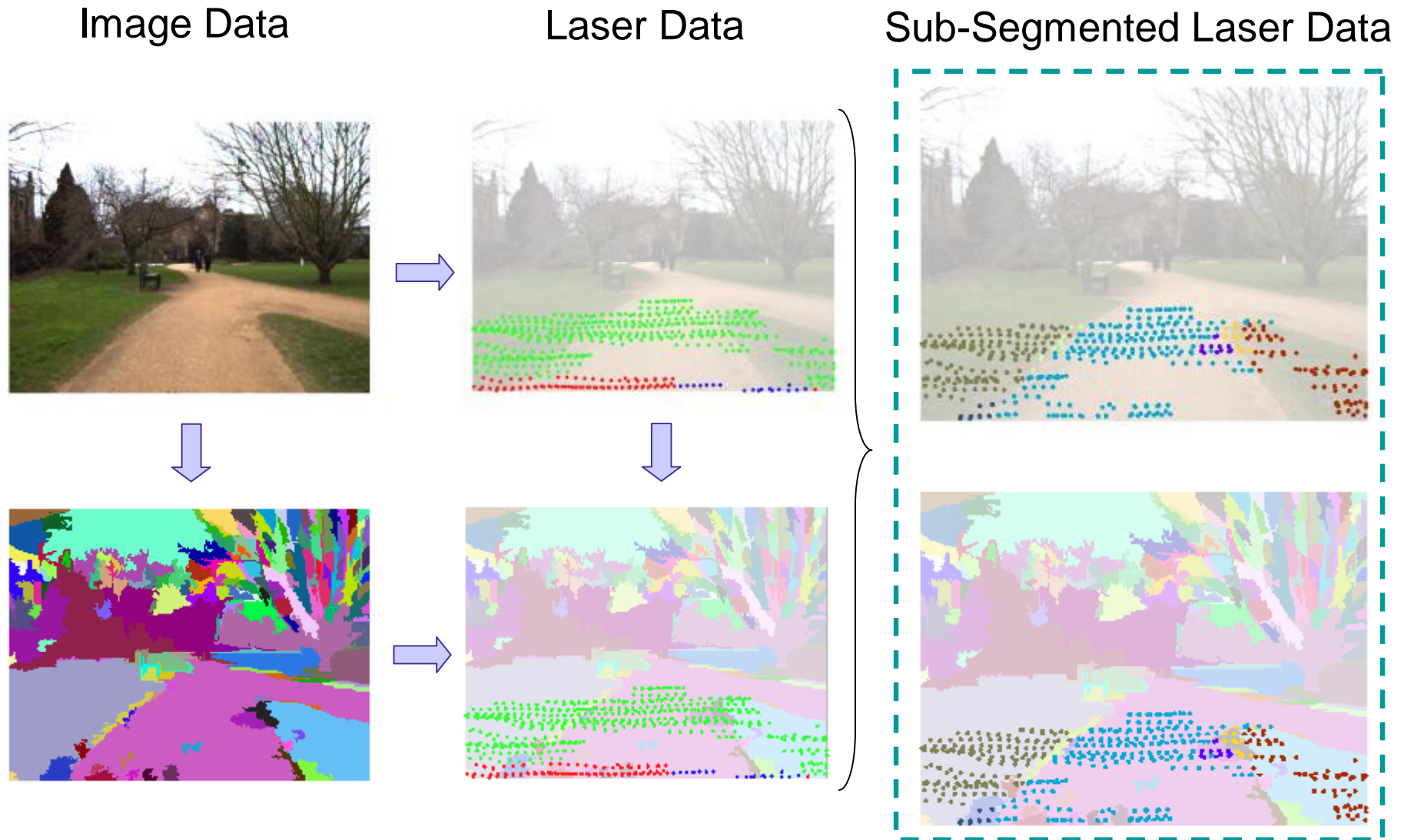
What's to come:

- Pipeline
 - Combining 3D laser and image data.
 - Representation.
 - Features.
- Classification Framework
 - Patch Classification.
 - Scene Classification.
- Results
- Conclusions

Pipeline

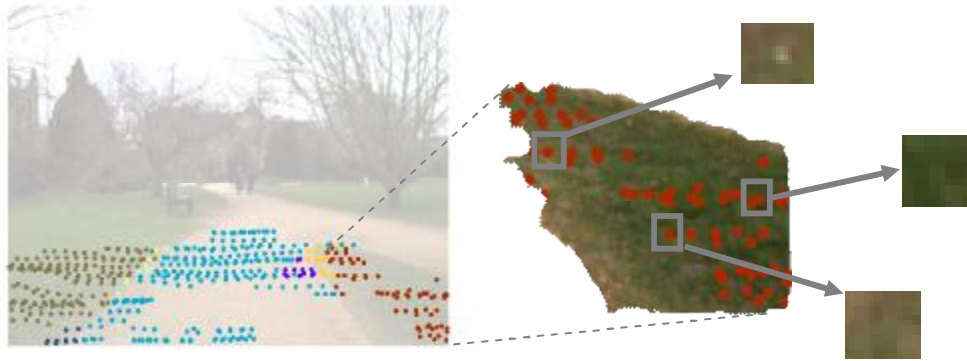


Similarity in Appearance Space

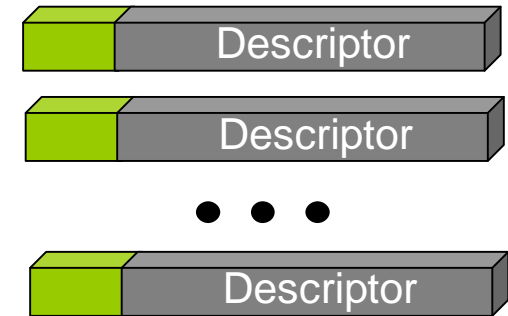


Operate on *image patches* with associated 3D geometry.

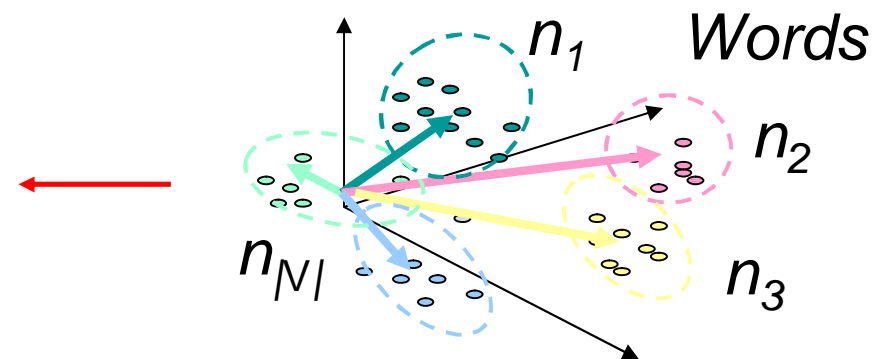
Bag-of-Words Representation



Compute Descriptors



Quantize



Cluster Centres are "Words"

Observation



$z =$

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \begin{array}{l} \text{Word 1} \\ \text{Word 2} \\ \text{Word 3} \\ \\ \text{Word } |V| \end{array}$$

Features

3D Geometry

- Surface Normal

2D Geometry

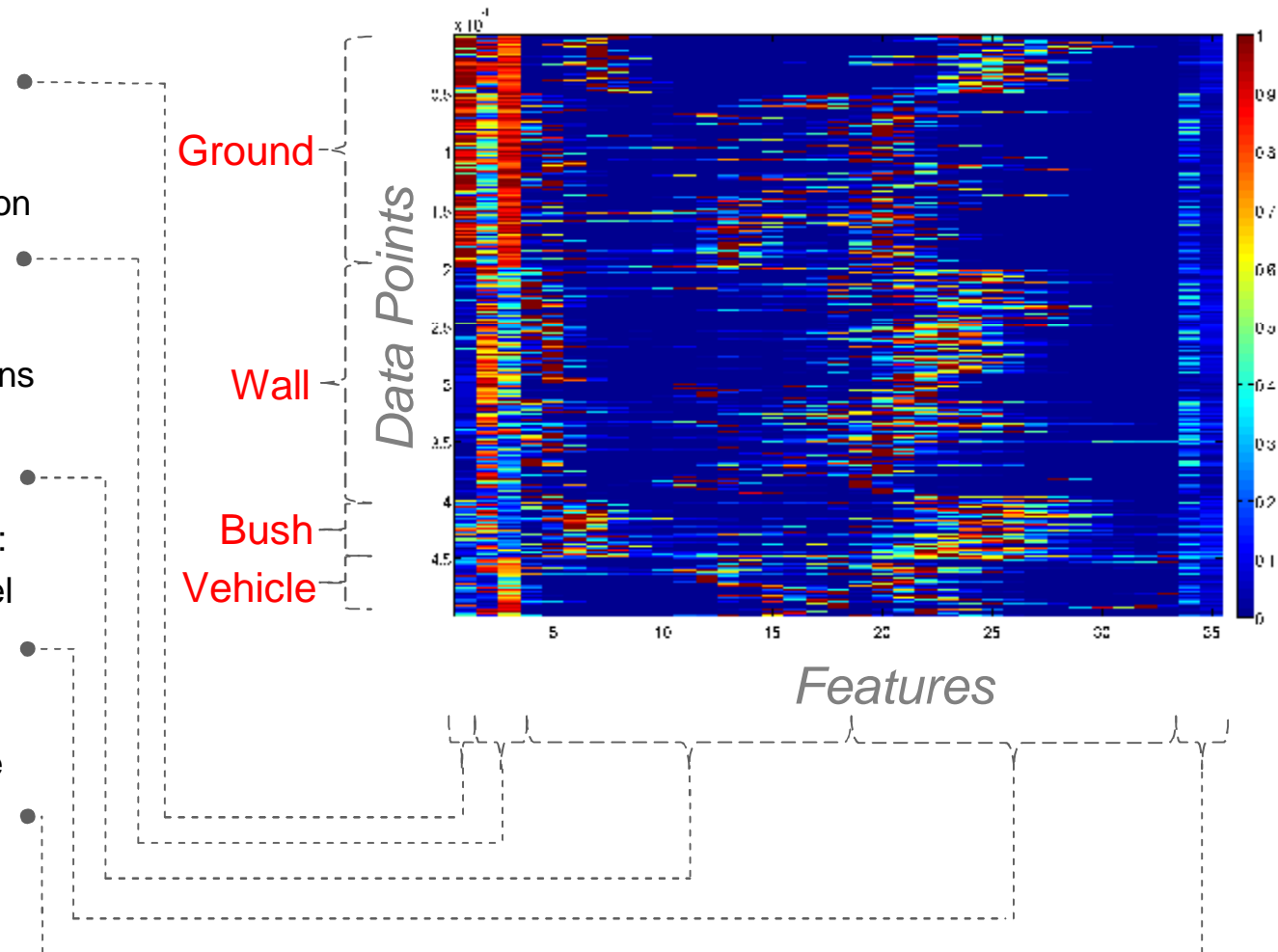
- Normalised x- and y-position in image

Colour

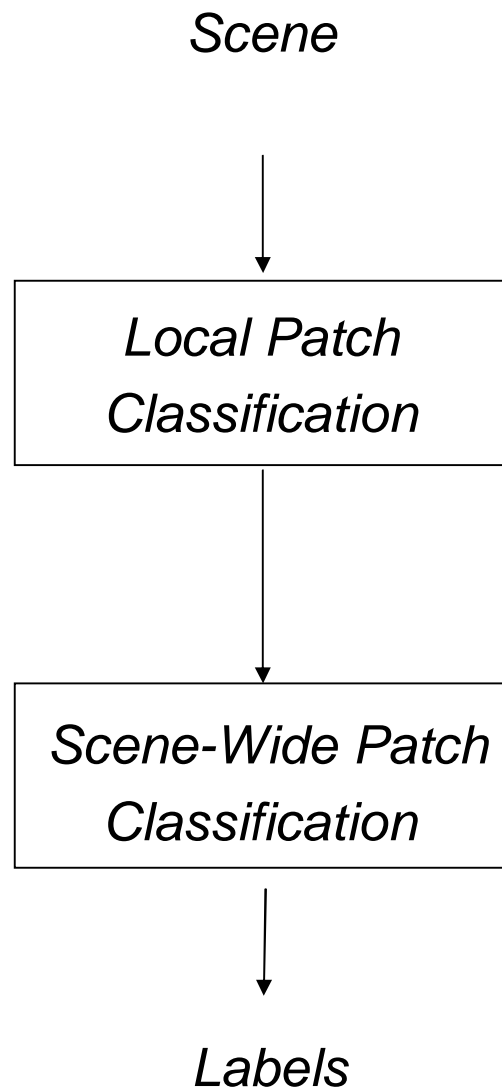
- Local hue histogram: 15 bins over a 15 x 15 pixel neighbourhood
- Local saturation histogram: 15 bins over a 15 x 15 pixel neighbourhood

Texture (via colour variation)

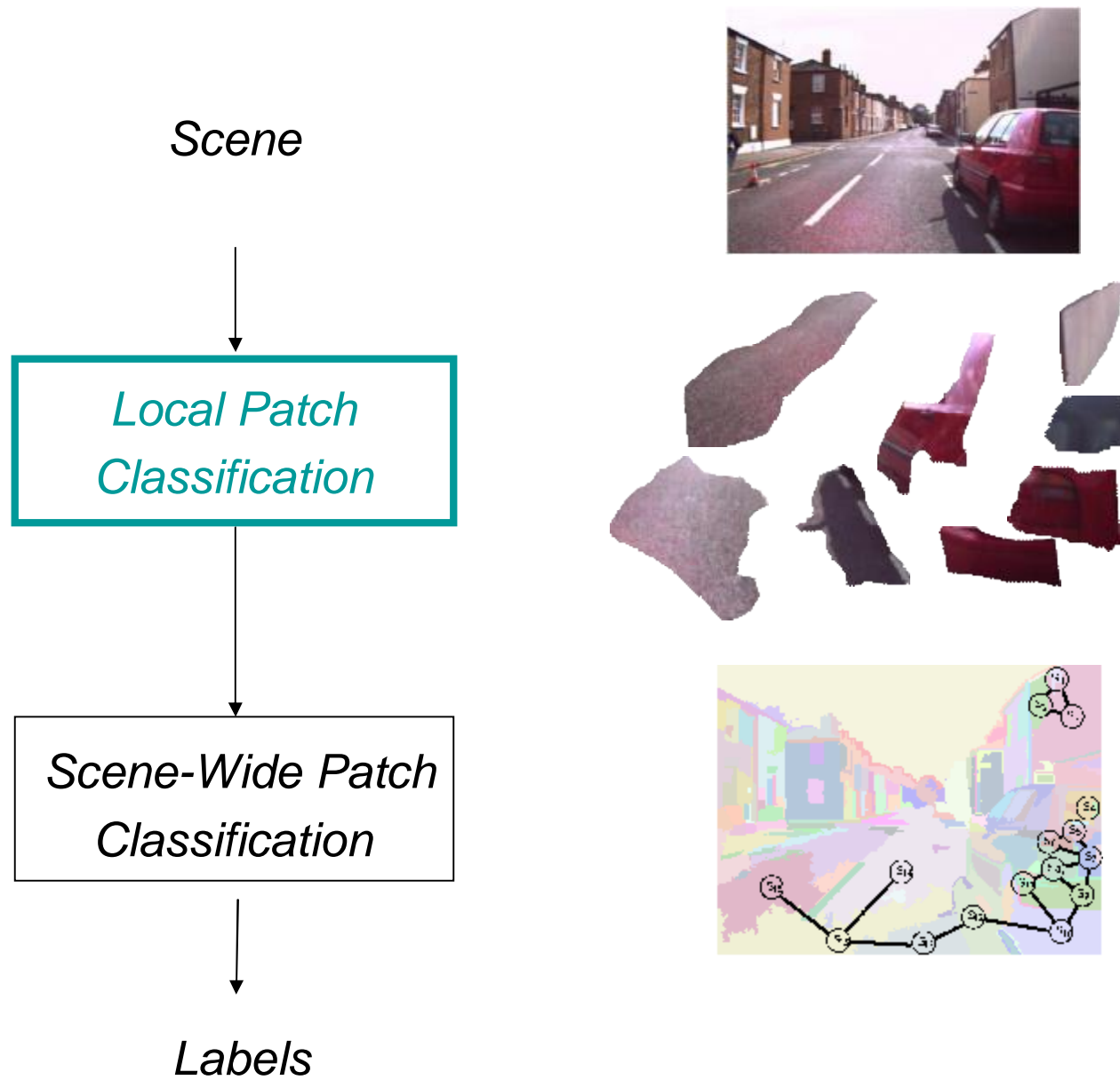
- Standard deviations of hue and saturation histograms.































Classification



Classification: Stage I



Stage I: Representing Classes

Exemplar:	1	2	3	...	n_k
Grass				• • •	
Tarmac / Paved				• • •	
Dirt Path				• • •	
Textured Wall				• • •	
Smooth Wall				• • •	
Bush / Foliage				• • •	
Vehicle				• • •	

$$\mathcal{C}^k = \{C_1^k, \dots, C_{n_k}^k\}$$

Stage I: Exemplars and the Sensor Model

Representing Exemplars:



$$C_i^k \triangleq \{p(e_1|C_i^k), \dots, p(e_{|v|}|C_i^k)\}$$

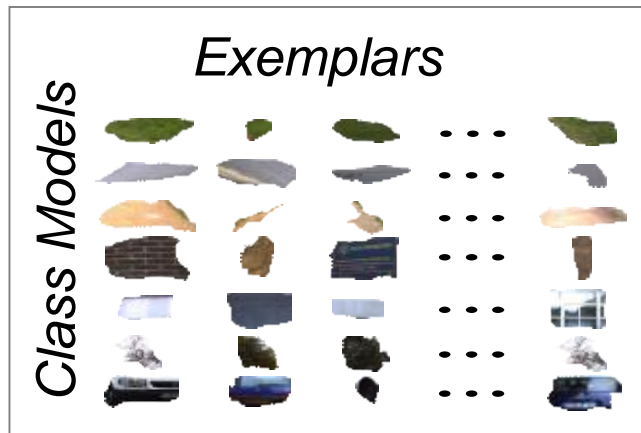
e_j - event that a *generator* for word j exists

The Detector Model:

$$\mathcal{D} : \begin{cases} p(z_j = 1|e_j = 0), & \text{false positive probability.} \\ p(z_j = 0|e_j = 1), & \text{false negative probability.} \end{cases}$$

Detector Model: we do not assume a perfect sensor.

Stage I: Patch Classification



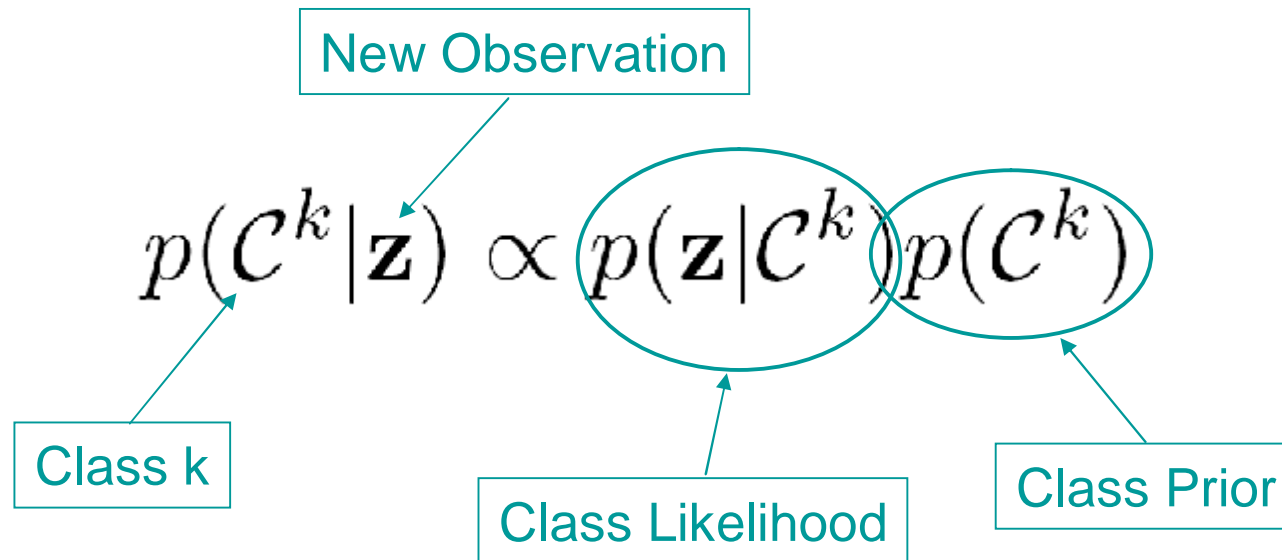
Which class k ?



New Observation



$$\mathbf{z}^T = [1, 1, 0, \dots, 1]$$



Stage I: The Class Likelihood

Assume:

- 1) *None of the training data are mislabeled.*
- 2) *All exemplars within a class are equally likely.*

$$p(\mathbf{z}|\mathcal{C}^k) = \sum_{i=1}^{n_k} p(\mathbf{z}|C_i^k, \mathcal{C}^k) p(C_i^k | \mathcal{C}^k)$$

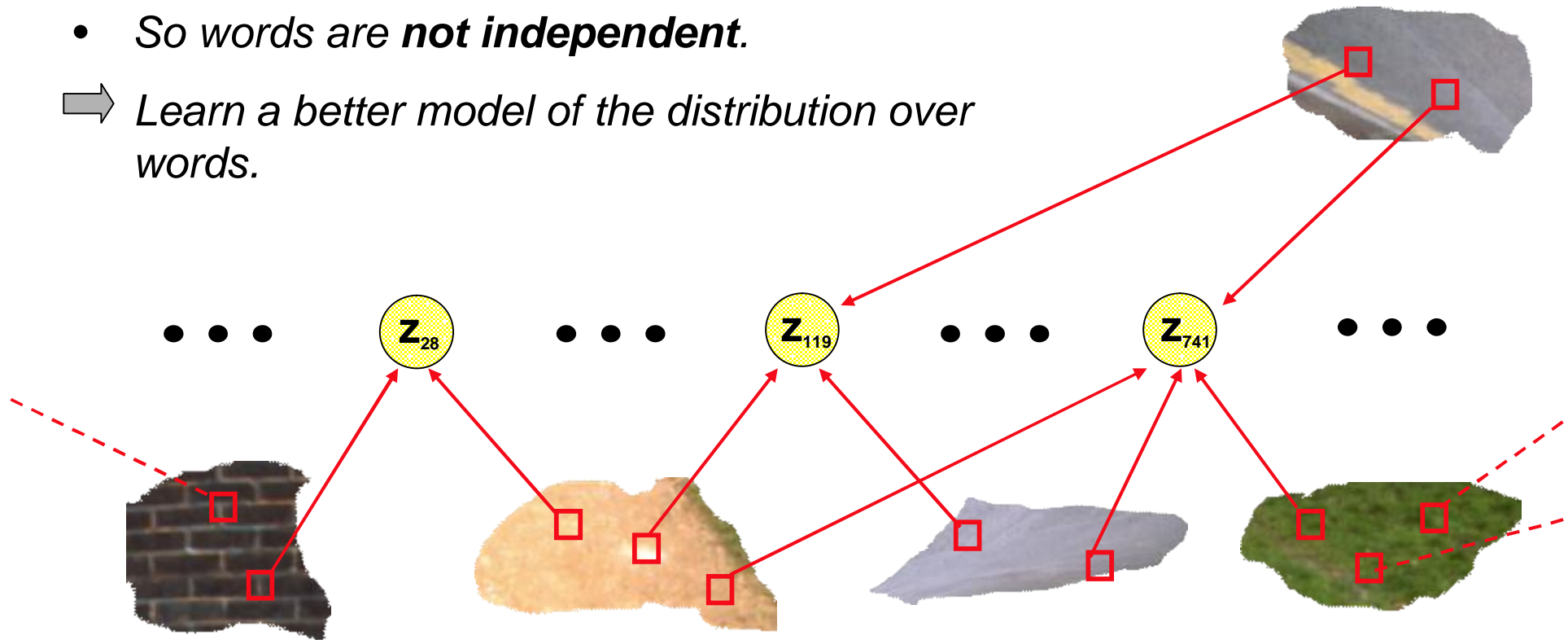
$$= \frac{1}{n_k} \sum_{i=1}^{n_k} p(\mathbf{z}|C_i^k)$$

Stage I: Estimation the Class Likelihood

Generative Model for Bag-of-Words Data

- Certain words **tend to co-occur**, because they are generated by the same object in the world.
- So words are **not independent**.

→ Learn a better model of the distribution over words.



Stage I: Estimation the Class Likelihood

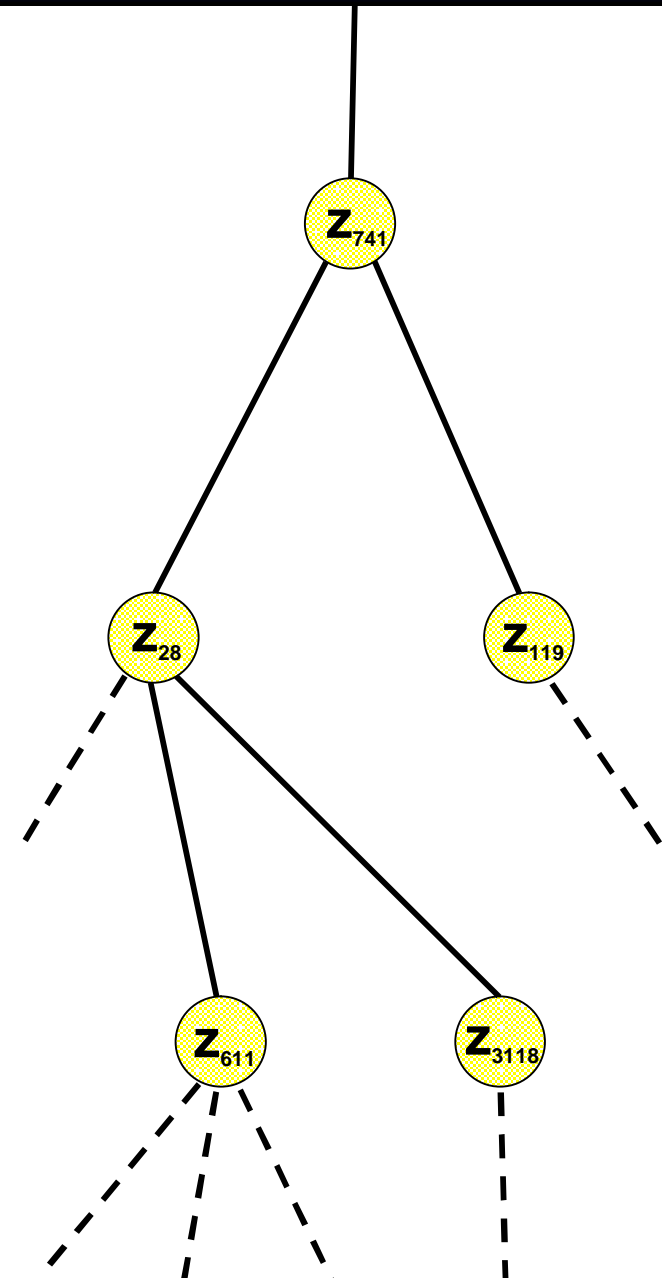
Generative Model for Bag-of-Words Data

- *Instead we learn a tree-structured Bayesian network to capture a first-order approximation to the PDF over word occurrence, using the **Chow Liu algorithm**:*

$$\begin{aligned} p(\mathbf{z}) &= p(z_1, z_2, \dots, z_{|v|}) \\ &\approx p(z_r) \prod_{q=1}^{|v|} p(z_q | z_{p_q}) \end{aligned}$$

- *And so:*

$$p(\mathbf{z} | C_i^k) \approx p(z_r | C_i^k) \prod_{q=1}^{|v|} p(z_q | z_{p_q}, C_i^k)$$

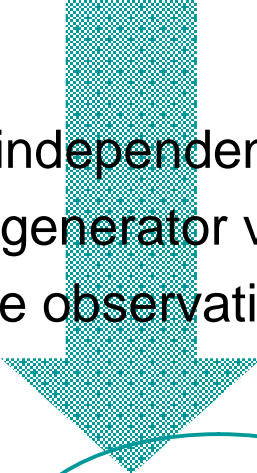


Stage I: Patch Classification

$$p(z_q | z_{p_q}, C_i^k) = \sum_{s_{e_q} \in \{0,1\}} p(z_q | e_q = s_{e_q}, z_{p_q}, C_i^k) p(e_q = s_{e_q} | z_{p_q}, C_i^k)$$

Assumptions:

- 1) Detector errors are independent of class (exemplar).
- 2) The probability of a generator variable for word i existing is independent of the observations of *all other words*.


$$p(z_q | z_{p_q}, C_i^k) = \sum_{s_{e_q} \in \{0,1\}} p(z_q | e_q = s_{e_q}, z_{p_q}) p(e_q = s_{e_q} | C_i^k)$$

We can compute this using the Exemplar Model

Stage I: Learning a Class Model

Remember - a class model consists of exemplars:

$$\mathcal{C}^k = \{C_1^k, \dots, C_{n_k}^k\}$$

Learn class models using *labeled training data*. So for each new exemplar:

$$p(e_q = 1 | C_i^k, \mathbf{z}) = \frac{p(\mathbf{z} | e_q = 1, C_i^k) p(e_q = 1 | C_i^k)}{p(\mathbf{z} | C_i^k)}$$

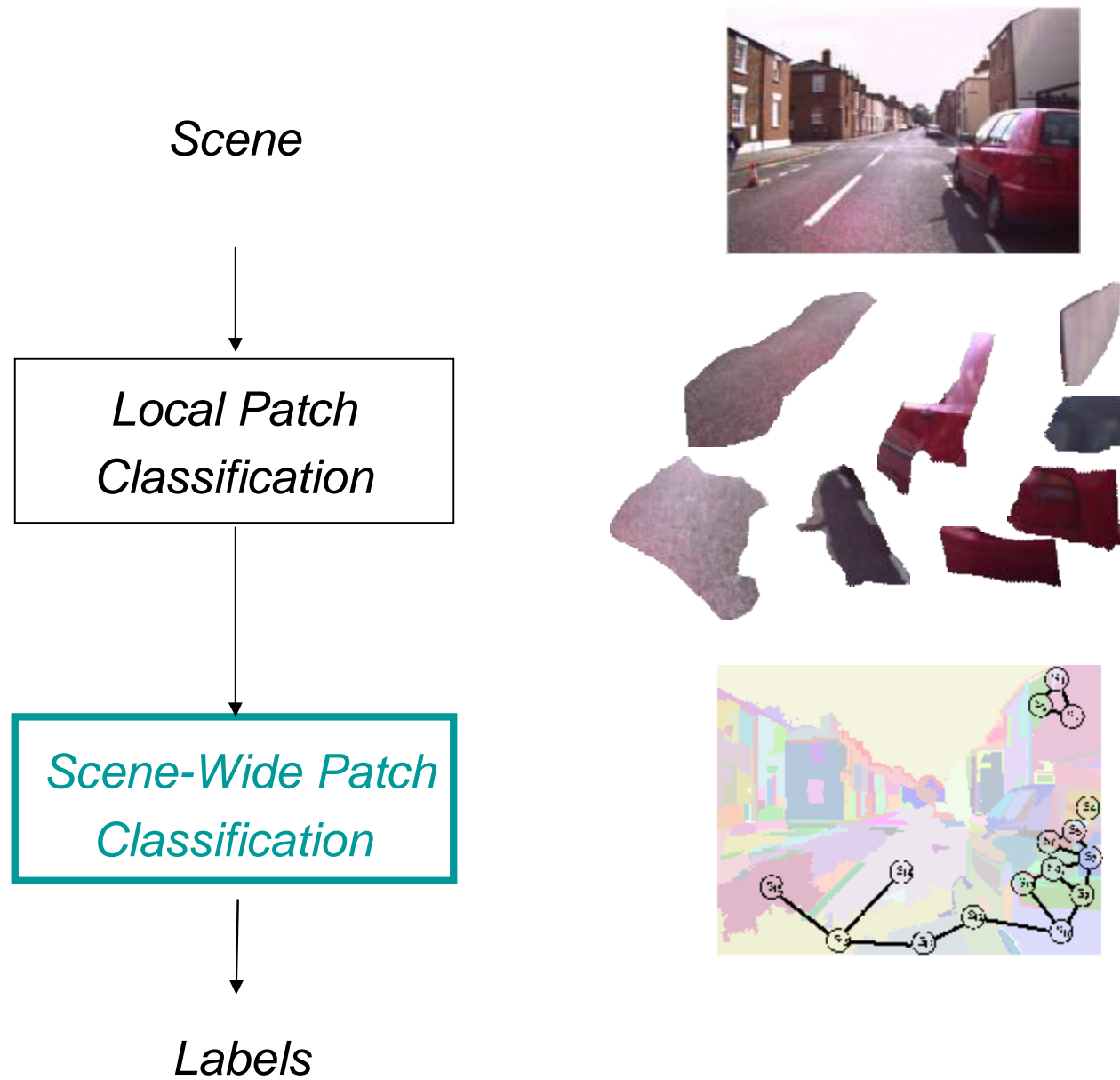
Likelihood term given word existence

Likelihood Term

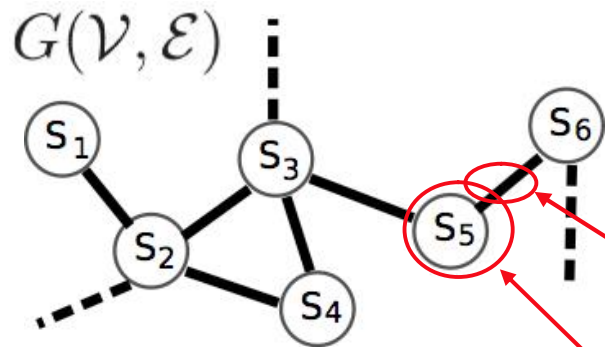
Word Existence Prior

Readily learned *online!*

Classification: Stage II



Stage II: MRFs



Markov Random Fields (MRF)

- Family of undirected graphical models.
- Model joint distribution over (hidden) states and the available data.

Consider: N_n - # of nodes
 N_c - # of classes

$\mathbf{x} \in \mathbb{Z}^{N_n}$ - label configuration
 $x_s \in \{1, \dots, N_c\}$

Perform MAP inference by minimizing an energy function:

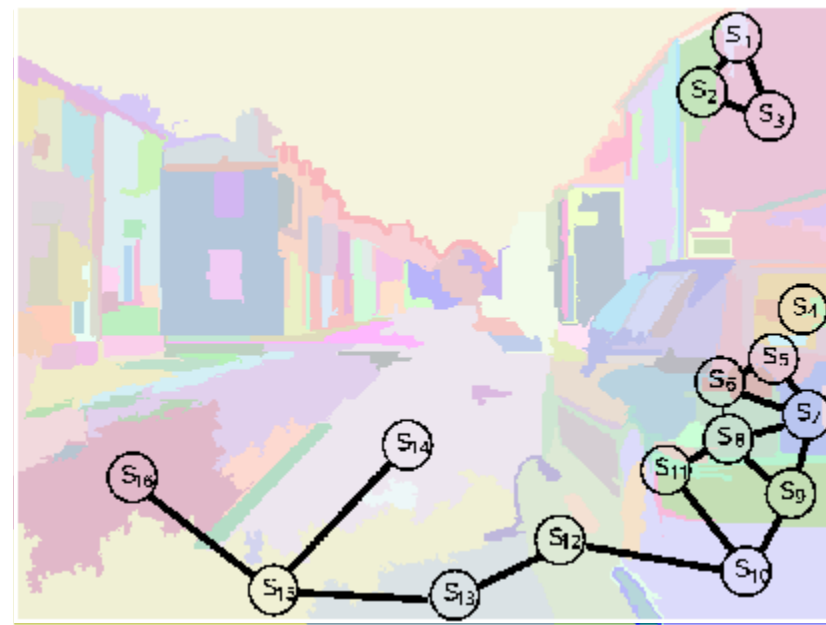
$$E(\mathbf{x}|\theta) = \sum_{s \in \mathcal{V}} \theta_s(x_s) + \sum_{(s,t) \in \mathcal{E}} \theta_{st}(x_s, x_t)$$

Well known problem in CV & ML. Can be tackled using

- Graph Cut Variants.
- (Max-Product) Loopy BP.
- Sequential Tree-reweighted Message Passing (TRW-S).

Need to determine model *structure* and *parameters*.

Stage II: Model Structure



- Operate on *superpixels*.
- Extract structure from neighbourhood relations in 2D.
- Capture adjacency dependencies.

Fast and intuitive extraction of *sparse* graphs.

Stage II: Model Formulation

Want to minimize

$$E(\mathbf{x}|\theta) = \sum_{s \in \mathcal{V}} \theta_s(x_s) + \sum_{(s,t) \in \mathcal{E}} \theta_{st}(x_s, x_t)$$

Unary potentials:

$$\theta_s(x_{sk}) = 1 - p(\mathcal{C}^k | \mathbf{z}^s) \quad \mathbf{z}^s \text{ - observation for node } s$$

Stage I

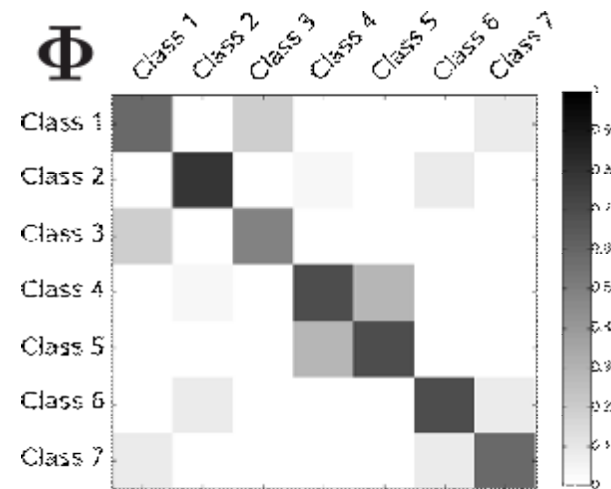
- Assuming our classes fully partition the world (i.e. $p(\mathcal{C}^k | \mathbf{z}^s)$ is a distribution)

Cost of transition between labels i and j .

Binary potentials:

$$\theta_{st}(x_i, x_j) = 1 - \phi_{i,j}$$

- Constitute our environmental prior.
- Learned from training data.

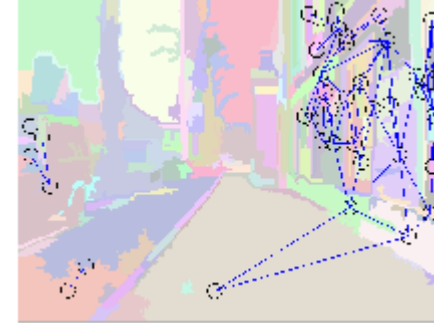


Stage II: MRF Smoothing in Action

Scene



Model



Pre MRF



Post MRF



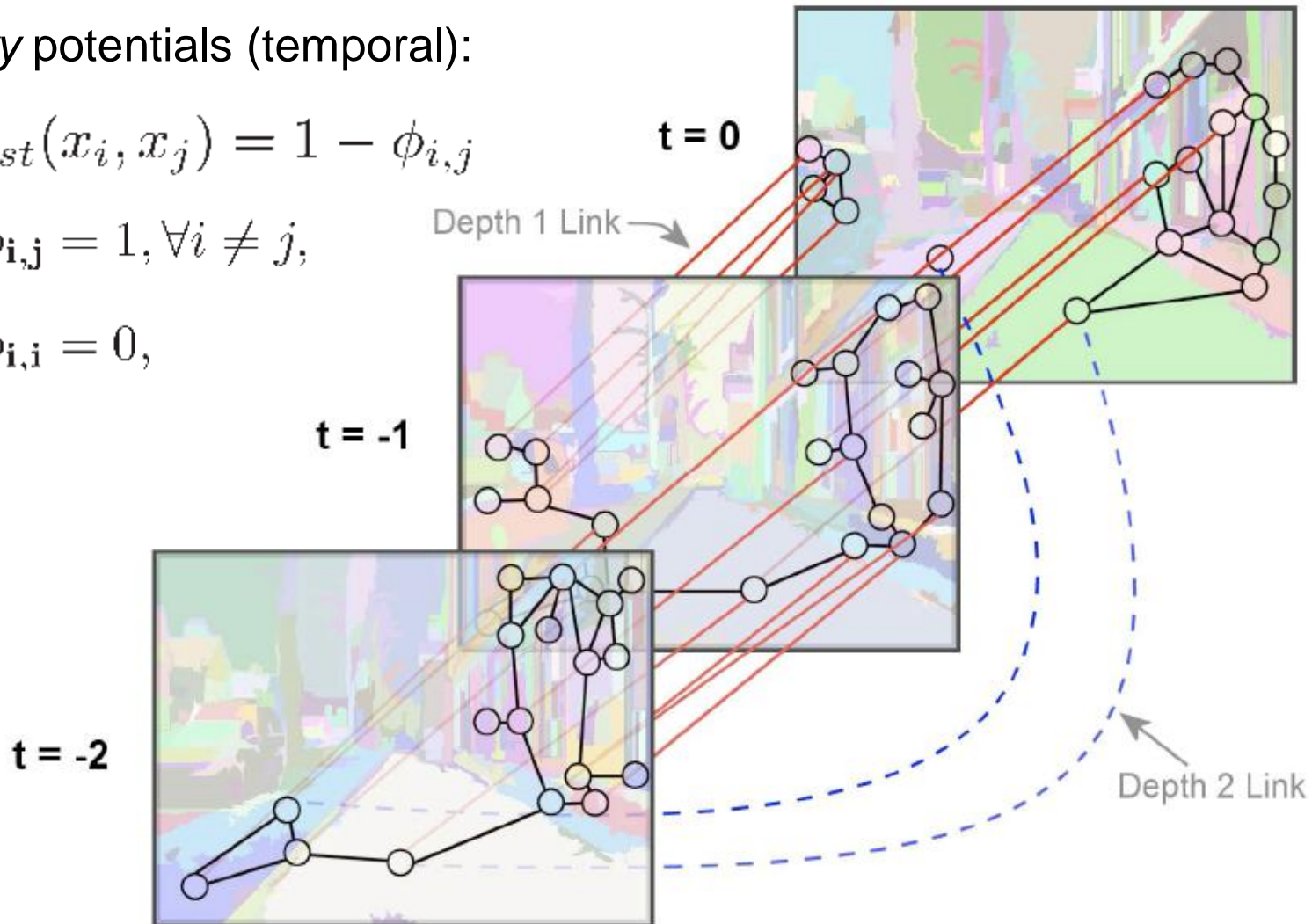
Temporal Smoothing

Binary potentials (temporal):

$$\theta_{st}(x_i, x_j) = 1 - \phi_{i,j}$$

$$\phi_{i,j} = 1, \forall i \neq j,$$

$$\phi_{i,i} = 0,$$



Training - Stage I



- Jericho / Oxford
 - 13.2 km track
 - 16,538 images
 - ca. 174×10^6 laser points

Test



- Oxford Science Park
 - 3.3 km track
 - 8,536 images
 - ca. 74×10^6 laser points

Results

Original Image (Oxford Science Park, No. 522)

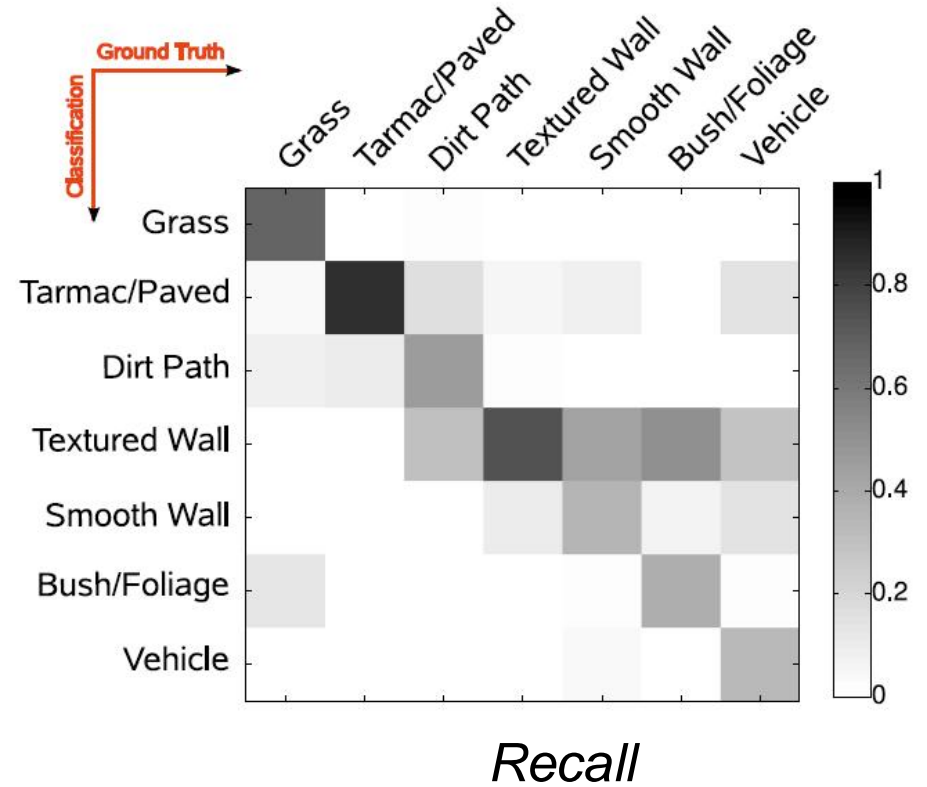
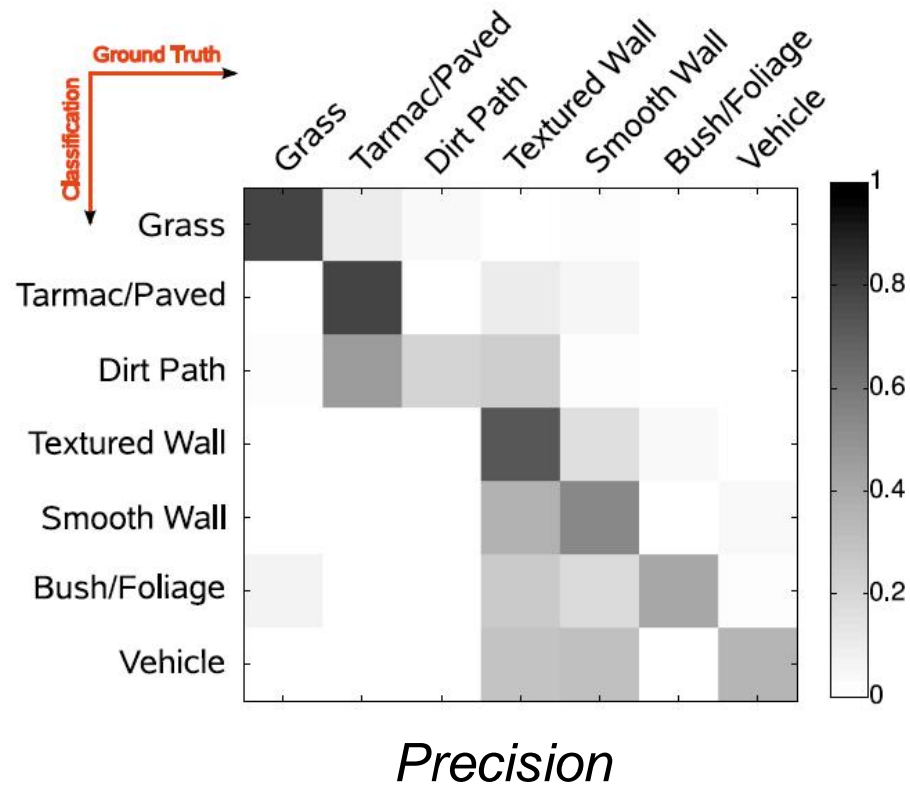


Classification

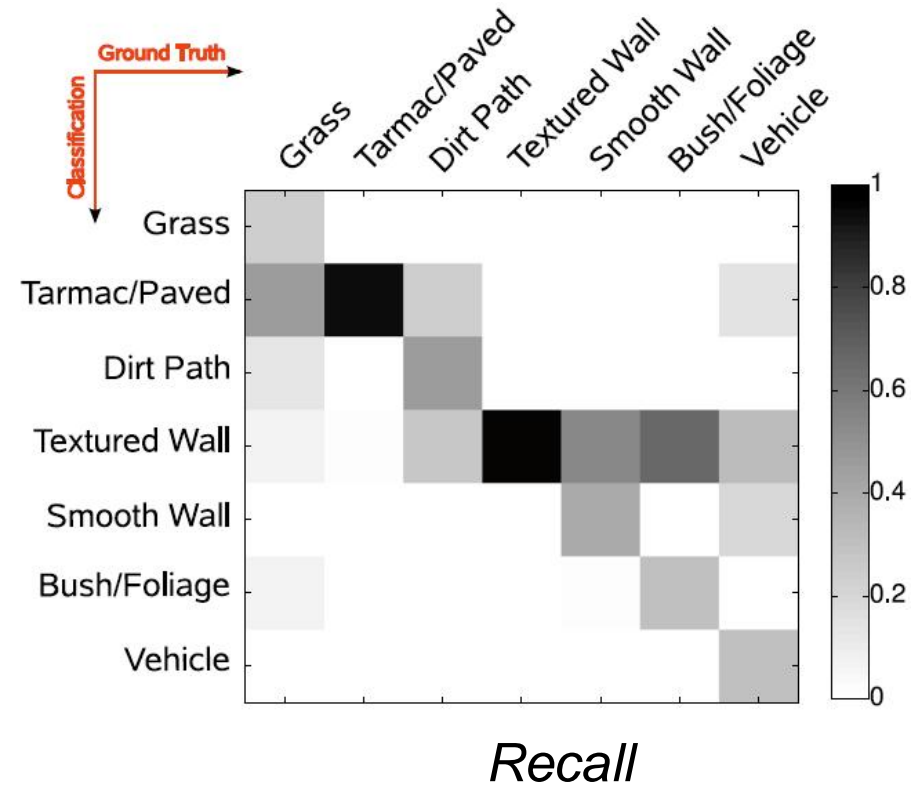
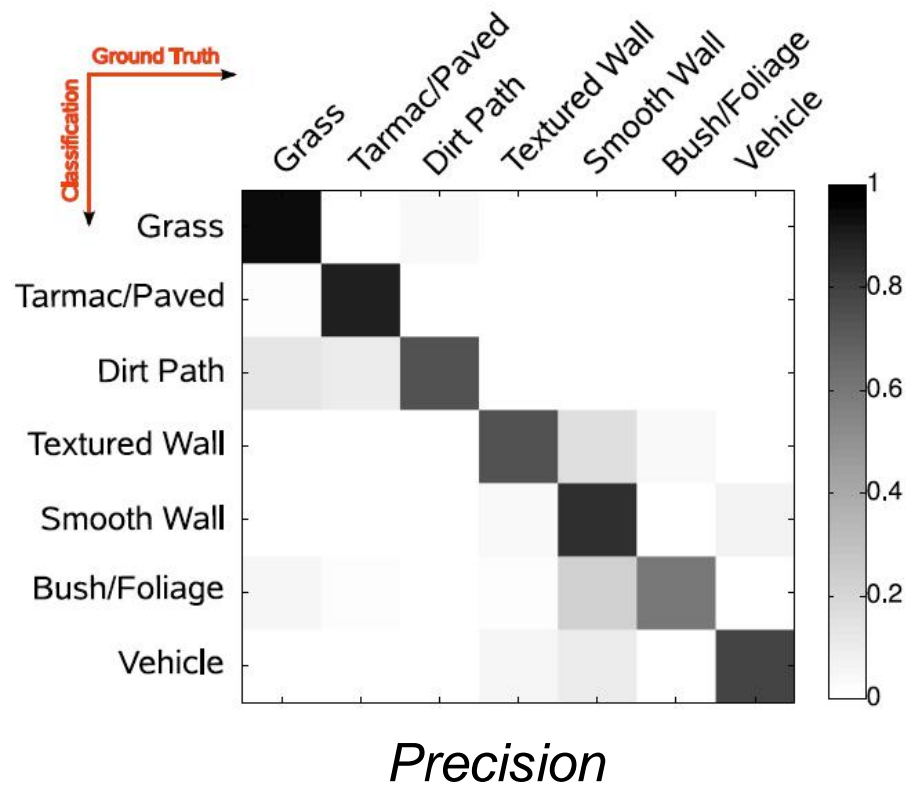


- Grass
- Tarmac
- Dirt Track
- Textured Wall
- Smooth Wall
- Bush/Foliage
- Vehicle

Results: Pre-MRF



Results: Post-MRF

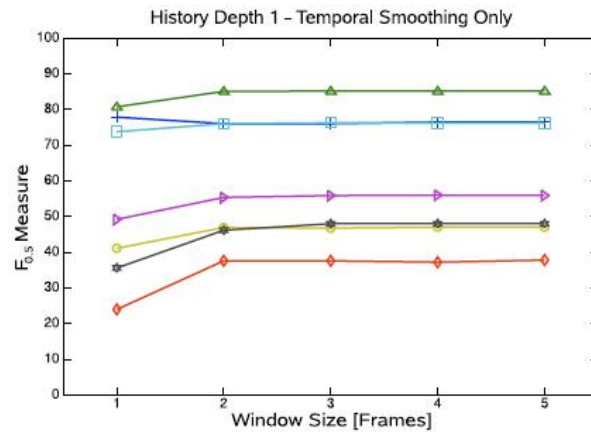


Results: Numbers

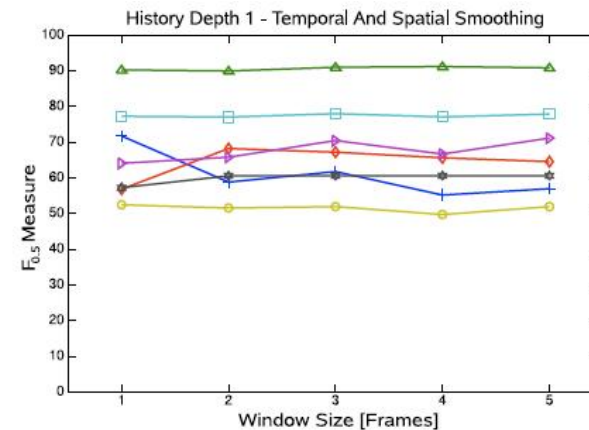
Class Details		Pre MRF			Spatial Context			Spatio-Temporal Context		
Name	# Patches	Precision [%]	Recall [%]	$F_{0.5}$	Precision [%]	Recall [%]	$F_{0.5}$	Precision [%]	Recall [%]	$F_{0.5}$
Gr	82	80.3	69.5	77.9	89.2	40.2	71.7	95.5	25.6	61.8
Ta	1286	79.5	86.1	80.8	89.2	94.9	90.3	89.9	95.7	91.0
Di	127	21.4	47.2	24.0	60.2	46.5	56.8	75.6	46.5	67.2
Te	2199	73.3	75.5	73.8	74.0	93.8	77.3	74.3	97.5	78.0
Sm	898	54.1	36.2	49.2	76.4	39.0	64.1	86.3	40.7	70.5
Bu	175	41.7	38.9	41.1	59.6	35.4	52.5	61.5	32.0	52.0
Ve	165	35.9	34.6	35.6	69.1	33.9	57.3	79.7	30.9	60.6

Legend for class shortcuts: **G**rass, **T**armac/Paved, **D**irt Path, **T**extured Wall, **S**mooth Wall, **B**ush/Foliage, **V**ehicle

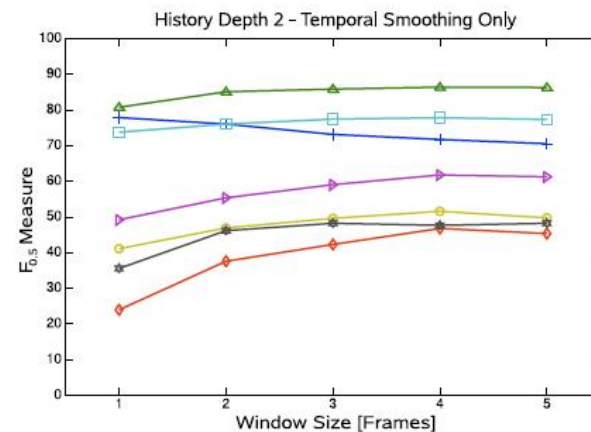
Spatial vs. Temporal Smoothing



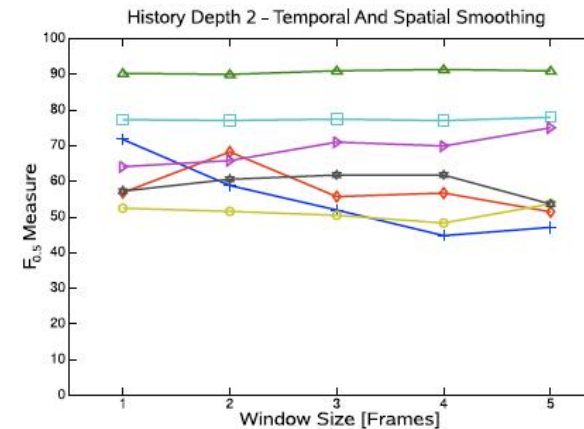
(a)



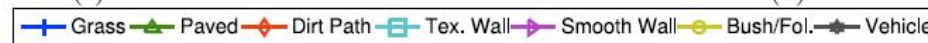
(b)



(c)



(d)



- (a) and (c) present results obtained using temporal smoothing only.
- (b) and (d) present results with both temporal and spatial edge information included.

Results: Timing

Process	Mean (ms)	Max (ms)
Plane Segmentation	2000	2800
Feature Extraction	89	125
Feature Quantization	4	90
Image Segmentation	960	1130
Patch Classification	850	3480
MRF Construction	63.5	453.9
MRF Inference	6.0	22.0
<i>Overall</i>	4.0 seconds	8.1 seconds

Our real-time constraint: ~ 3 s

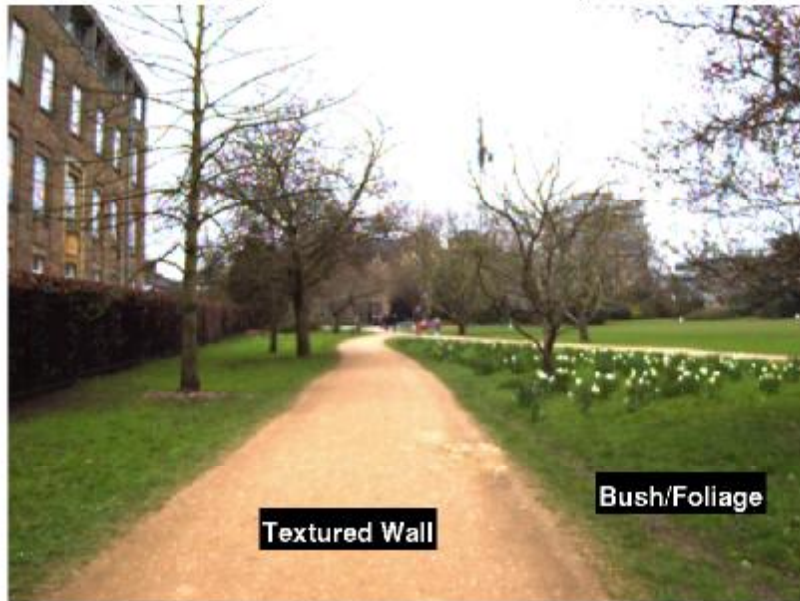
So, let's see:

- Principled.
 - Probabilistic.
- Introspective.
 - Able to handle context.
 - Know what measurements to trust (detector model).
- Flexible.
 - Adapt (learn) class models online and in real-time.
- Fast.
 - Classify in real-time.

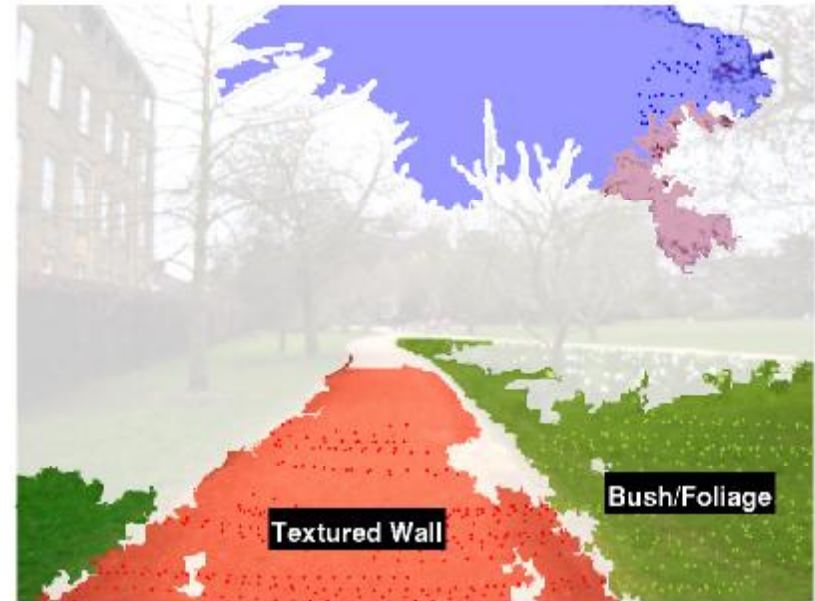
Failure Cases

The things we didn't tell you ...

Original Image (Oxford Science Park, No. 1)



Classification



Grass Tarmac Dirt Track Textured Wall Smooth Wall Bush/Foliage Vehicle

... there are currently some obvious shortcomings.

Comparative Analysis

Class Details	Voted SVM			Pre MRF			Spatio-Temporal Context		
Name	Precision [%]	Recall [%]	$F_{0.5}$	Precision [%]	Recall [%]	$F_{0.5}$	Precision [%]	Recall [%]	$F_{0.5}$
Gr	96.6	98.1	96.9	80.3	69.5	77.9	95.5	25.6	61.8
Ta	97.7	89.0	95.8	79.5	86.1	80.8	89.9	95.7	91.0
Di	46.4	84.8	51.0	21.4	47.2	24.0	75.6	46.5	67.2
Te	82.7	73.5	80.7	73.3	75.5	73.8	74.3	97.5	78.0
Sm	56.9	64.4	58.3	54.1	36.2	49.2	86.3	40.7	70.5
Bu	60.6	62.8	61.0	41.7	38.9	41.1	61.5	32.0	52.0
Ve	43.7	80.1	48.1	35.9	34.6	35.6	79.7	30.9	60.6

Legend for class shortcuts: **G**rass, **T**armac/Paved, **D**irt Path, **T**extured Wall, **S**mooth Wall, **B**ush/Foliage, **V**ehicle

We propose a classification framework which

- Is fully probabilistic.
- Takes into account local and global context.
- Provides for a degree of introspection.
- Provides for online adaptation/learning of class models.
- Is Fast.

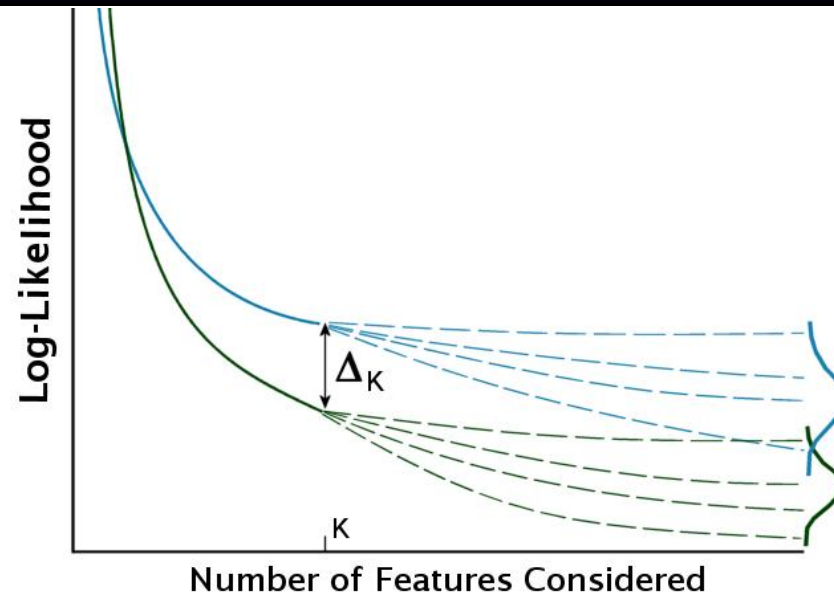
Room for improvement:

- First stage classification performance.
- Better features may provide a route to improvement.
- But any soft assignment classifier can be used.
- More elaborate cues in the MRF: relative location, containment relations.
- Relative weighting between unary and binary potentials.

Thank you...

Let's talk...

The Distribution over Final Log-Likelihood



X_i

The relative change in log-likelihood between the two hypotheses due to the i^{th} feature.

$$S_k = \sum_{i=k+1}^N X_i$$

Discard weaker hypothesis?

$$p(S_k > \Delta_k) < \epsilon$$

Bennett's Inequality

Define M:

$$p(|X_i| < M) = 1, \forall i$$

Maximum value
of any X_i

and v:

$$\sum_{i=n+1}^N E[X_i^2] < v$$

Sum of variances
of all X_i

Bennett's Inequality:

$$p(S > \Delta) < \exp\left(\frac{v}{M^2} \cosh(\sinh^{-1}\left(\frac{\Delta M}{v}\right)) - 1 - \frac{\Delta M}{v} \sinh^{-1}\left(\frac{\Delta M}{v}\right)\right)$$