FAB-MAP 3D: Topological Mapping with Spatial and Visual Appearance - “RGBD”

Rohan Paul and Paul Newman

Mobile Robotics Group

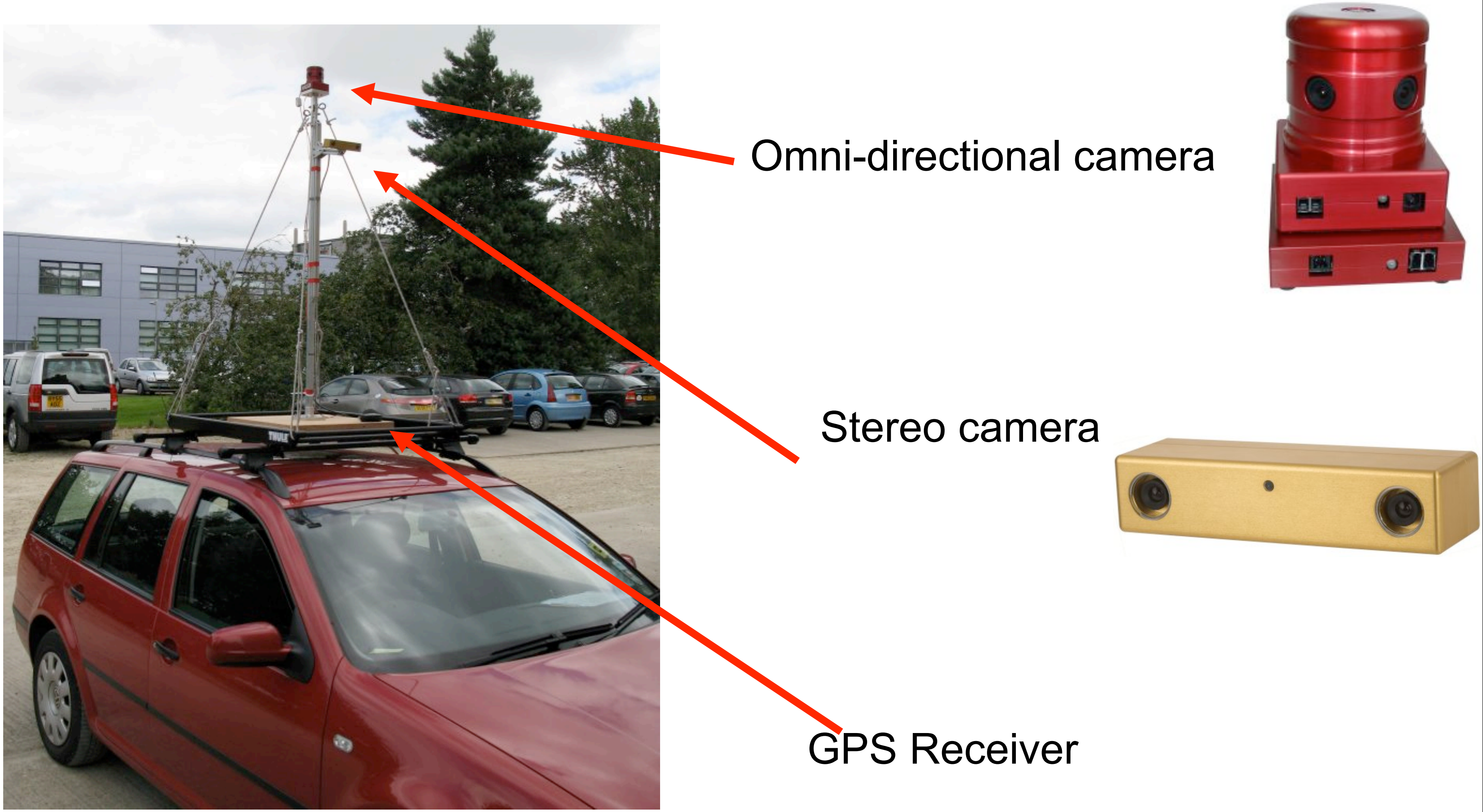
University of Oxford

We want Robots to Perceive, Understand and Manipulate the physical world.

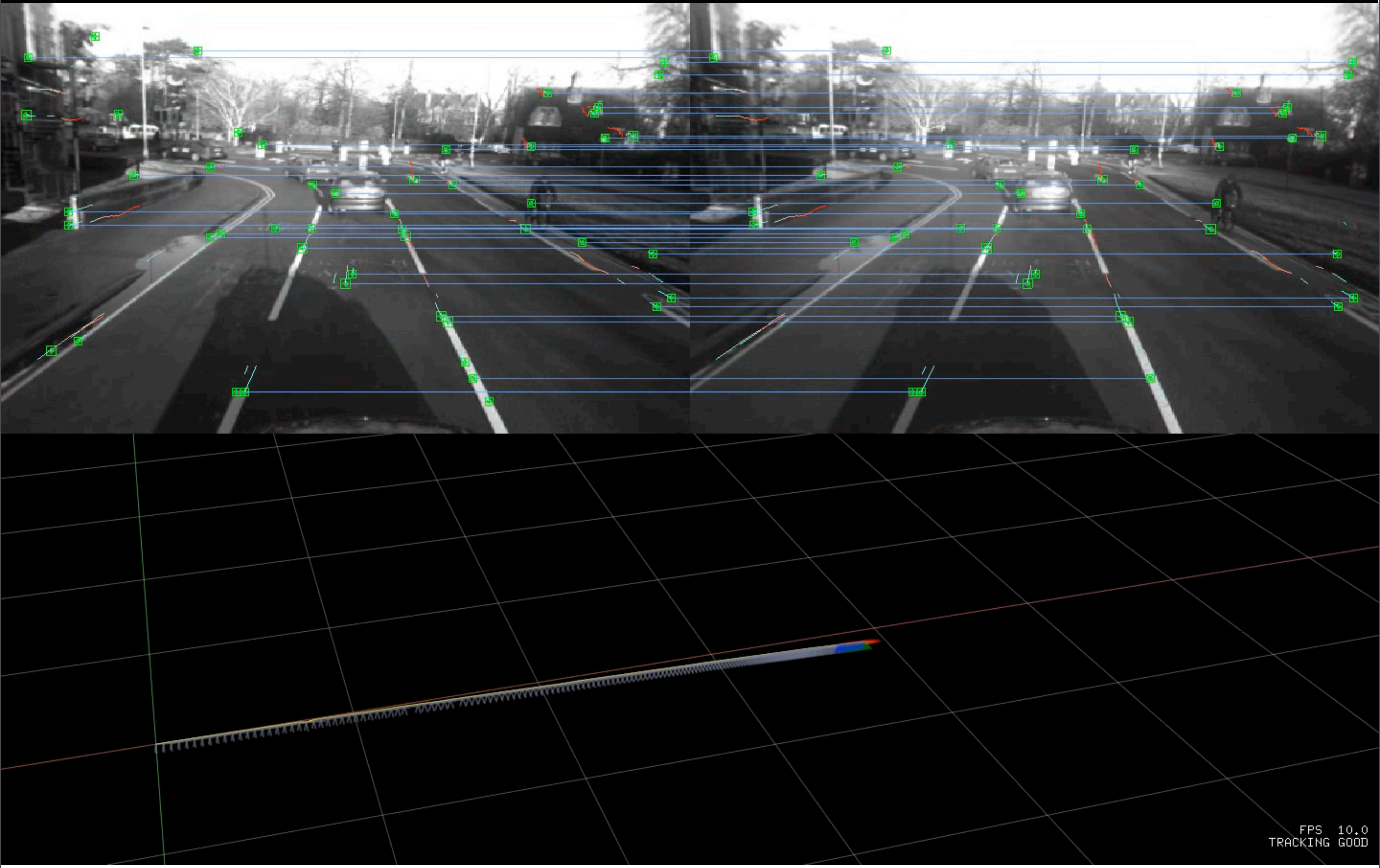
An intelligent robot must answer:

What does the world look like? and Where am I located?

Vehicle

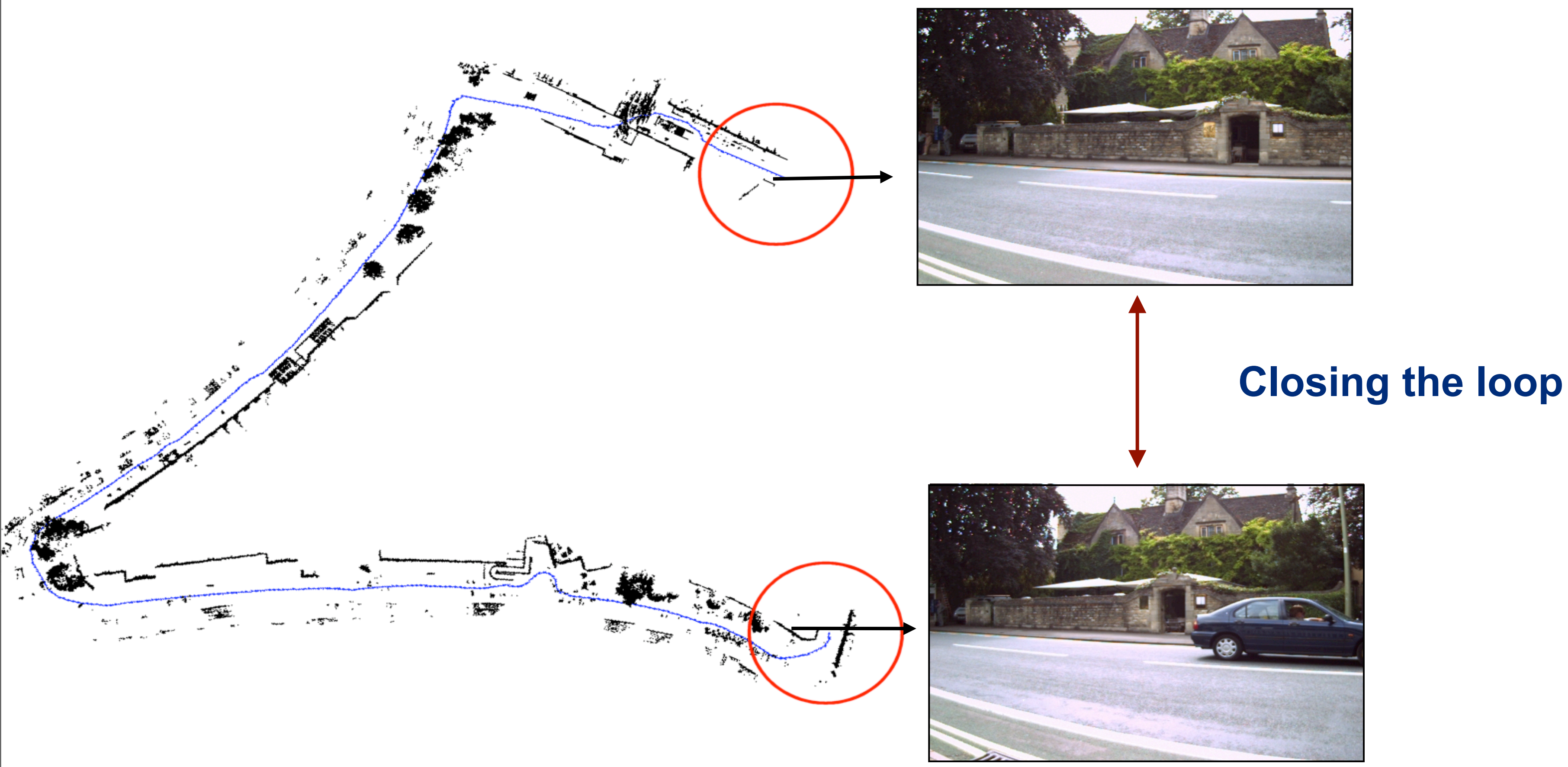


Creating a Map with Stereo Vision



[Sibley Mei Reid Newman RSS2009]

Loop Closure Detection from Appearance Alone



Scene Change

Place appearance changes between visits.

Real environment is highly **dynamic**

Perceptual Aliasing

Different places can appear identical.

Looking same does **not** mean its the same place

Place Recognition in Action: FAB-MAP

Problems: Missed Loop Closures

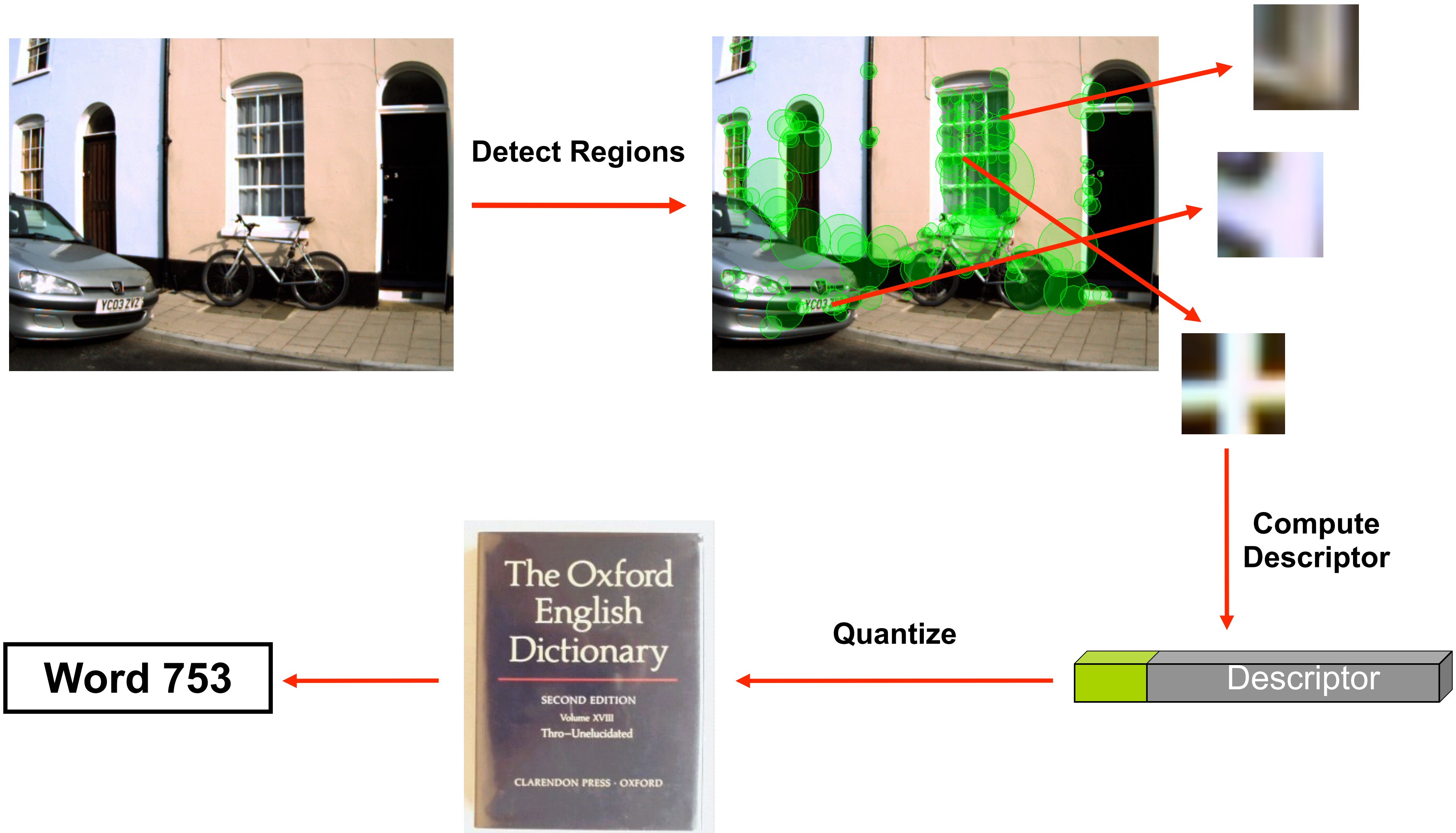
Only picks about 40% loop closures.

Problems: Wrong Loop Closures

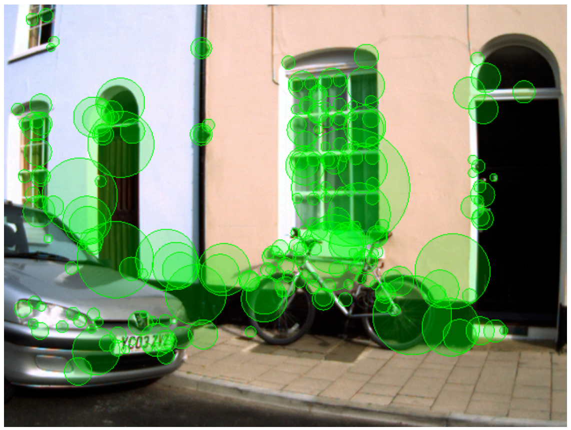
Its not the same place!

FAB-MAP Image Representation

Bag-of-Words Representation



FAB-MAP Limitations



Robot View

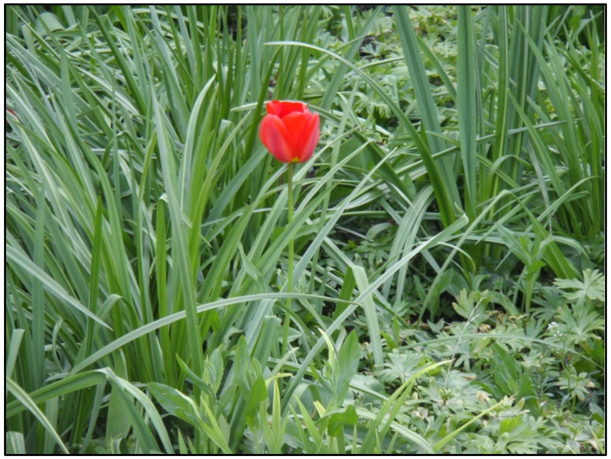
$$Z = \{0, 1, 0, 1, 1, \ldots\}$$

Presence or absence of a feature

FAB-MAP only considers presence or absence of a feature.

Spatial information is lost

Why is Spatial Information Important?

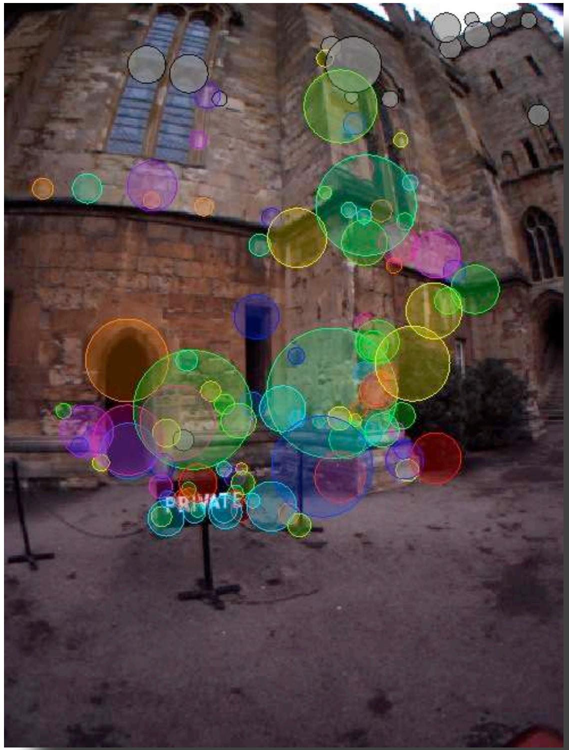


Set of Words = Same

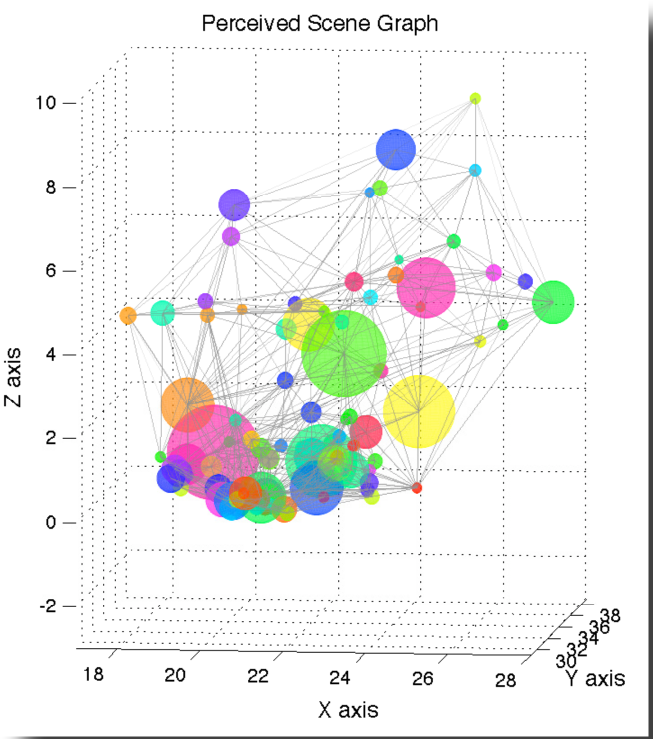
Set of Words + Spatial Configuration = Different

Location as a Constellation of Visual Features

Places are defined by their content AND their spatial configuration



Robot View

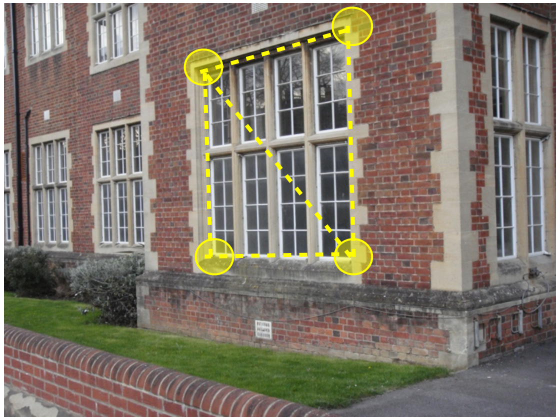


Perceived 3D Scene Graph

Model Locations as a Non-planar Random Graph

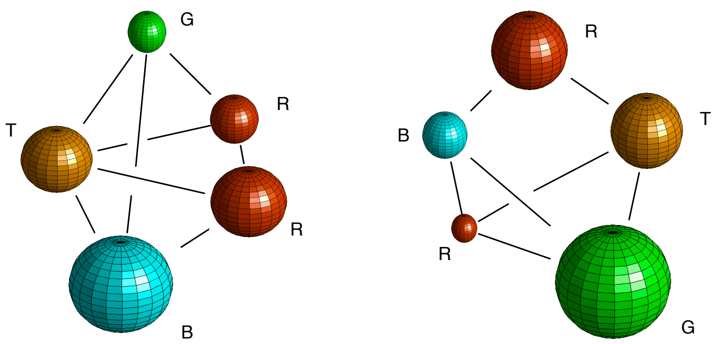
3D distances from Lidar, Depth cameras, Stereo or Structure from motion

Clear advantage : Invariance



Graph Structure is invariant under rigid transformations

- A probabilistic model of locations as a **random graph**.
- Capture presence of features and their **spatial configuration**.
- Use a **Detector model** to explain the noisy way in which the sensors perceive the world.
- Learn **correlations** between observed features and complex multi-modal **distributions** over distances.



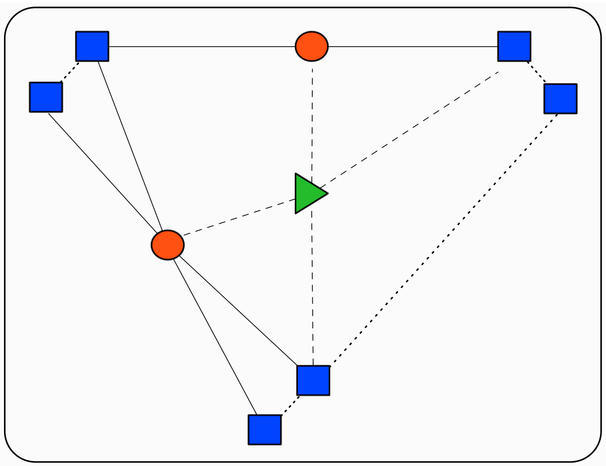
Key Point: By explicitly modeling spatial appearance we shall massively increase recall-precision coverage

Random Graph Location Model

Observation

Vocabulary

Perceived Graph



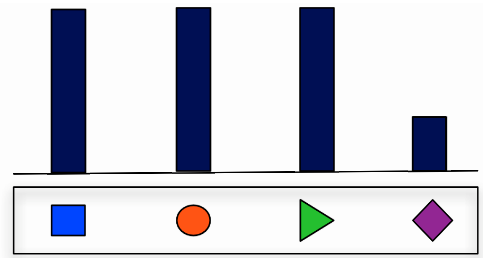
Word Detection

1 1 1 0

Visual Model

Likelihood of word existence

p = 1.0 p = 1.0 p = 1.0 p = 0.3

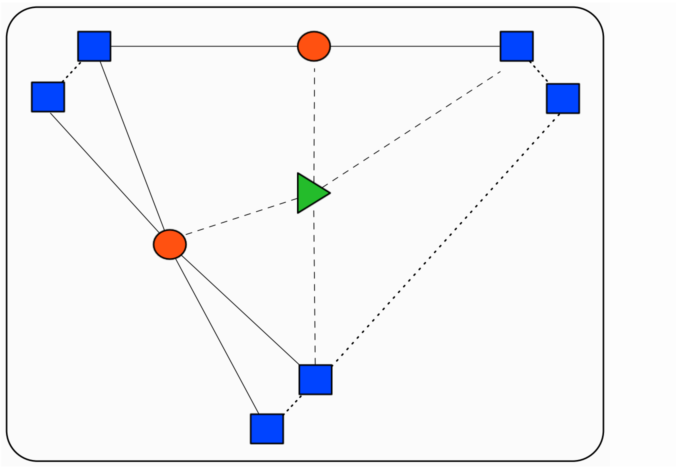


Random Graph Location Model

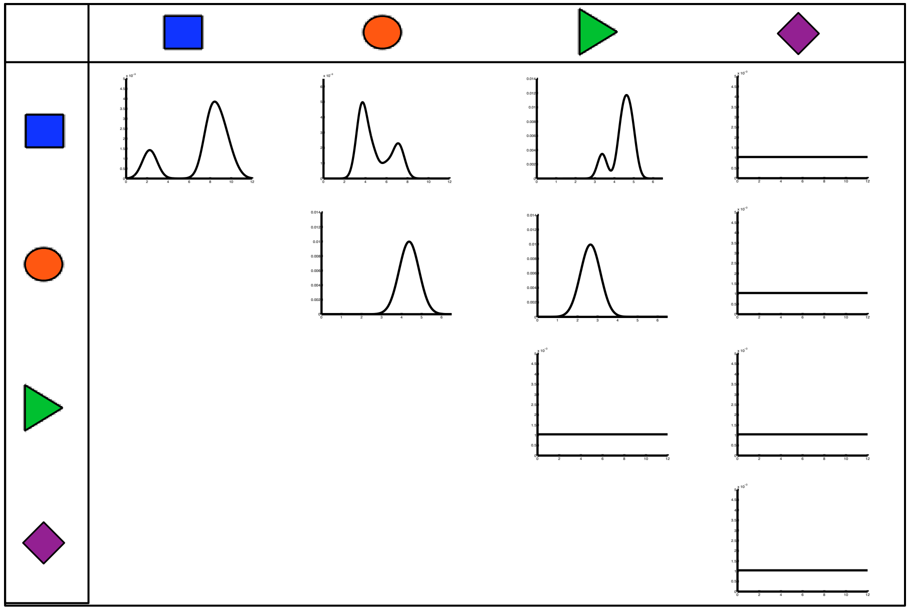
Observation

Spatial Model

Perceived Graph



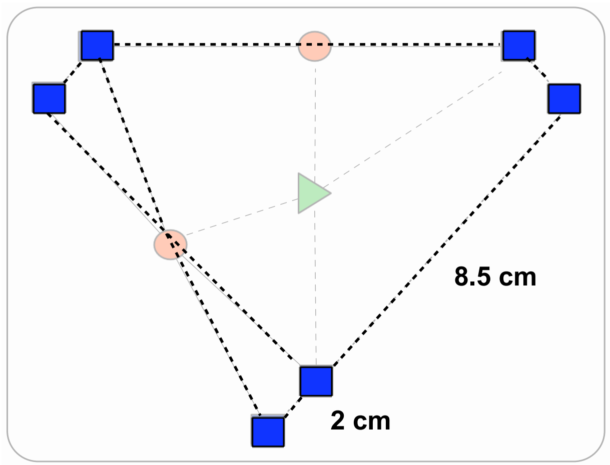
Distributions over inter-word distances



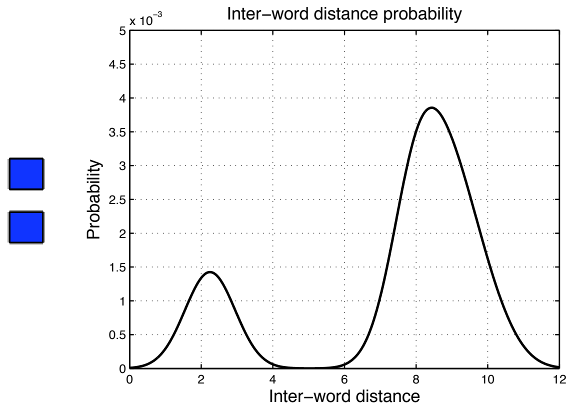
Random Graph Location Model

Observation

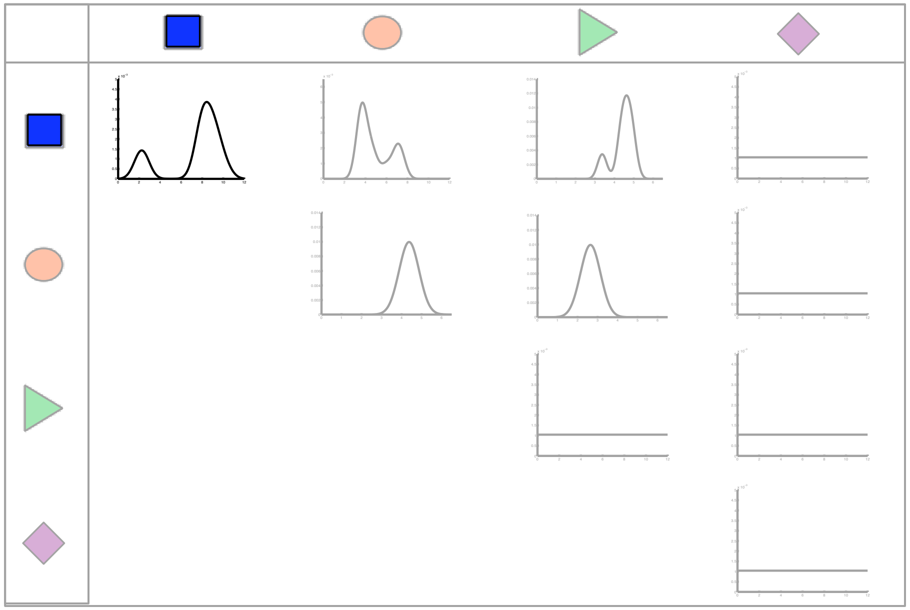
Perceived Graph



Location Model: Spatial



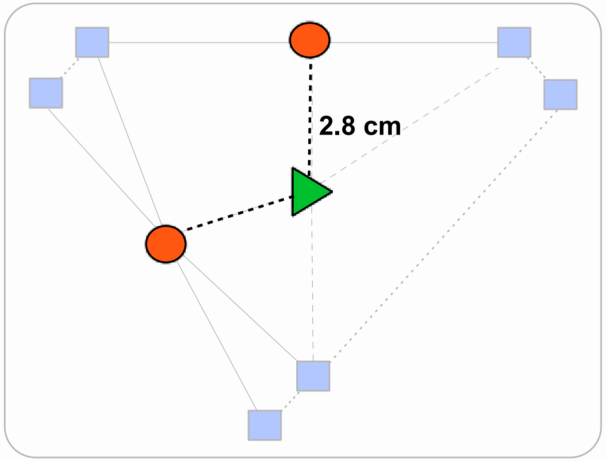
Distributions over inter-word distances



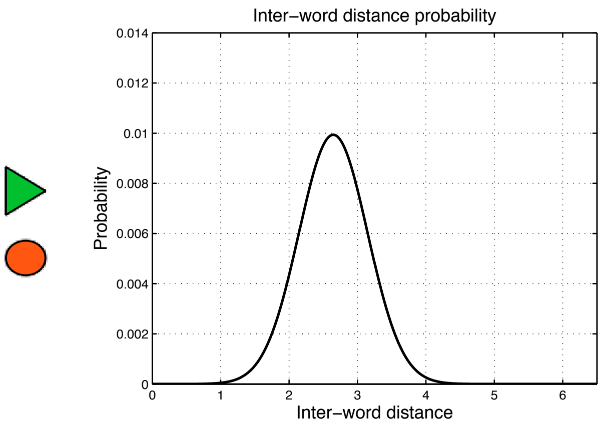
Random Graph Location Model

Observation

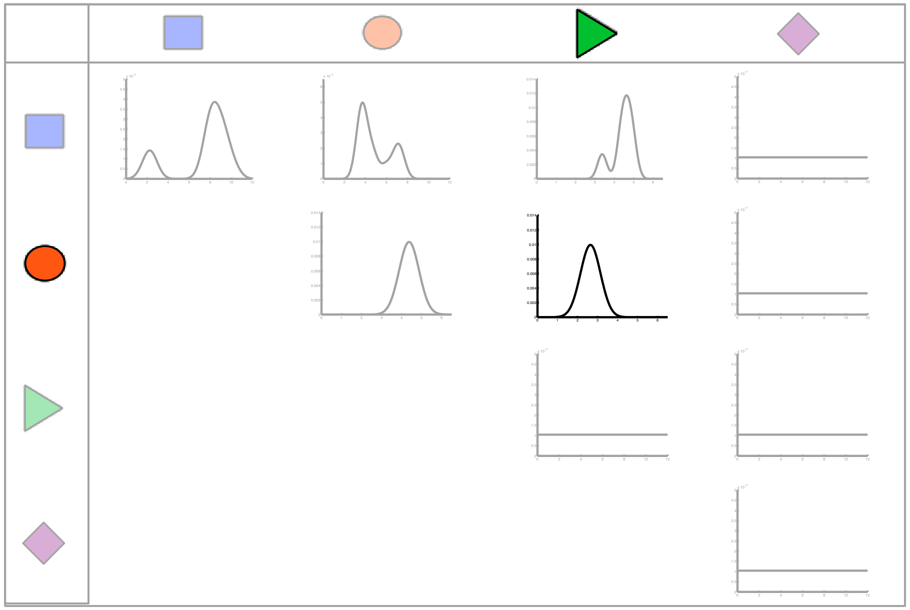
Perceived Graph



Location Model: Spatial

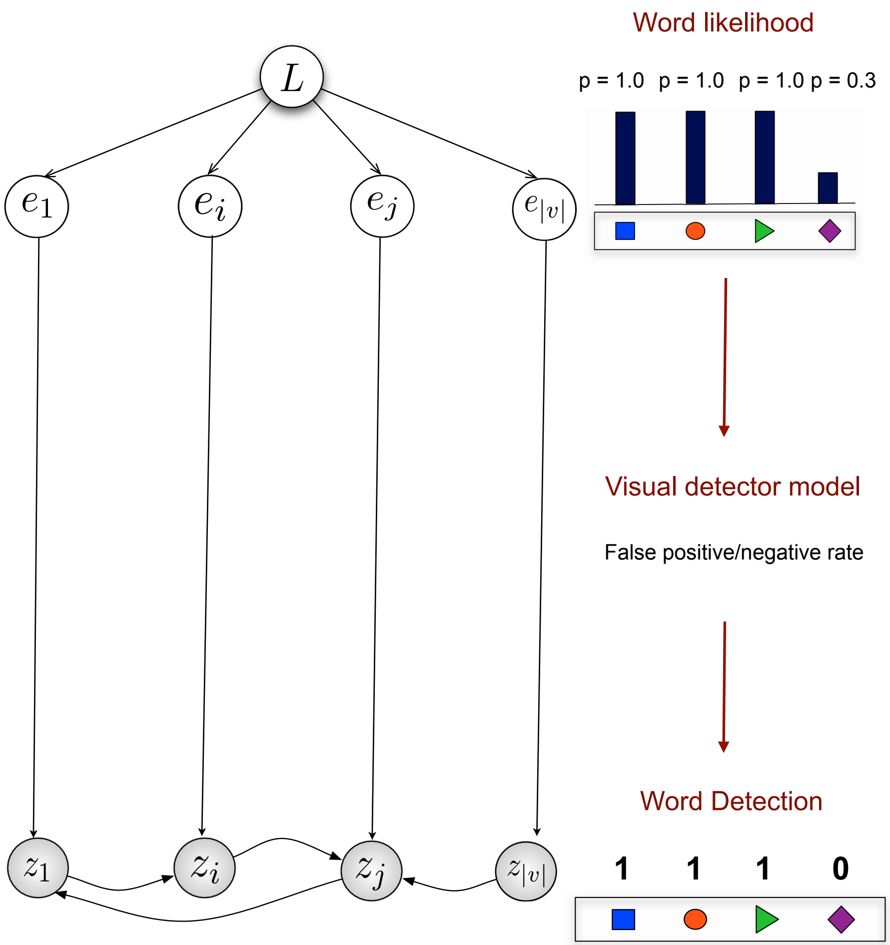


Distributions over inter-word distances

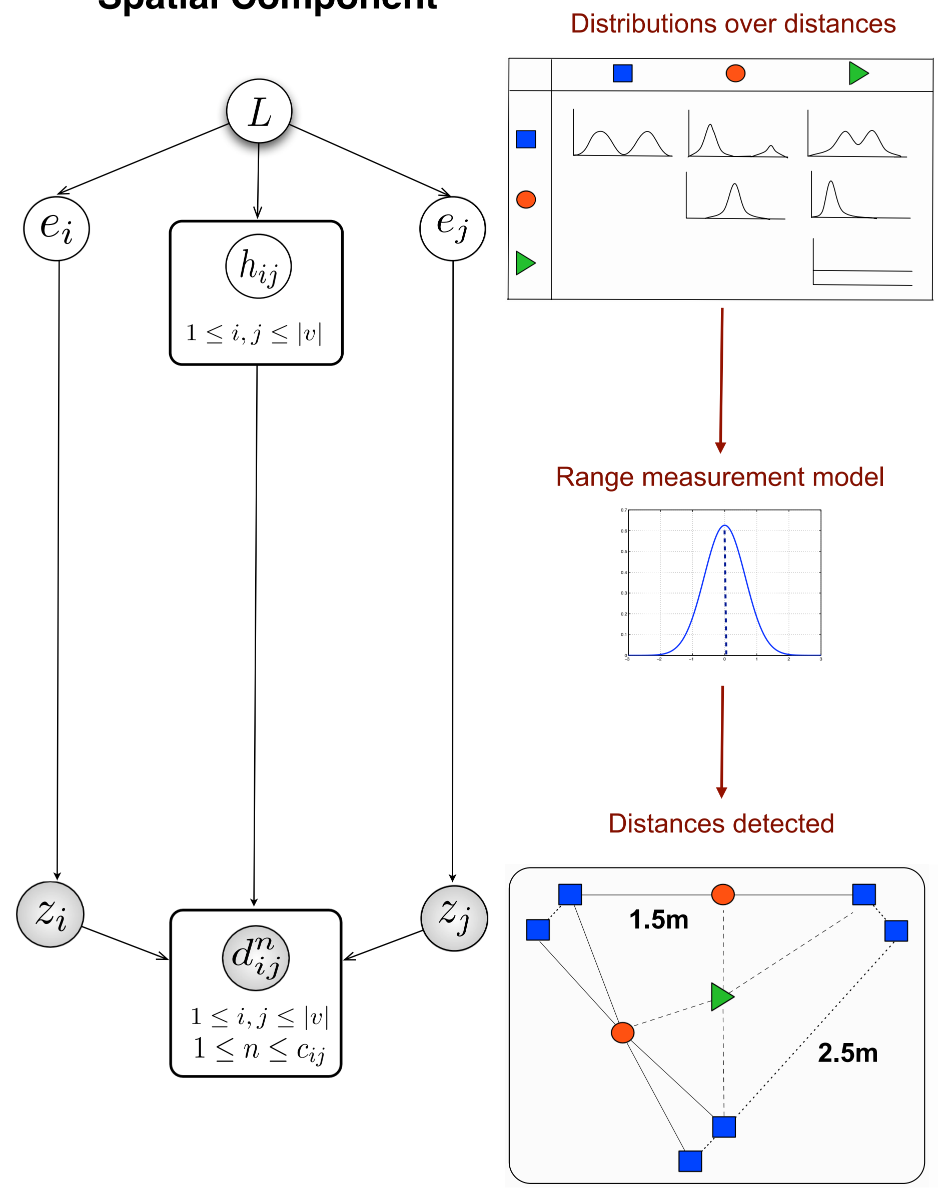


A Generative Model for Locations

Visual Component



Spatial Component



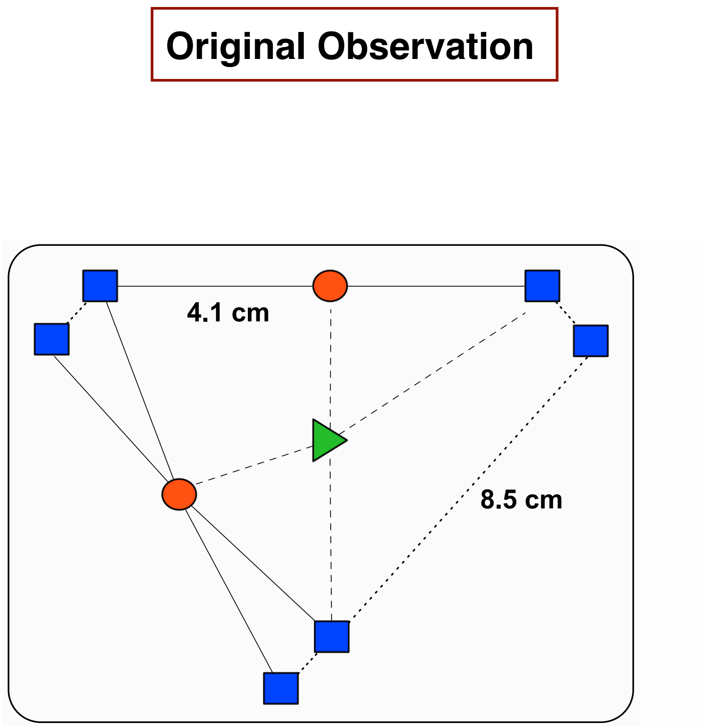
Understanding re-observation

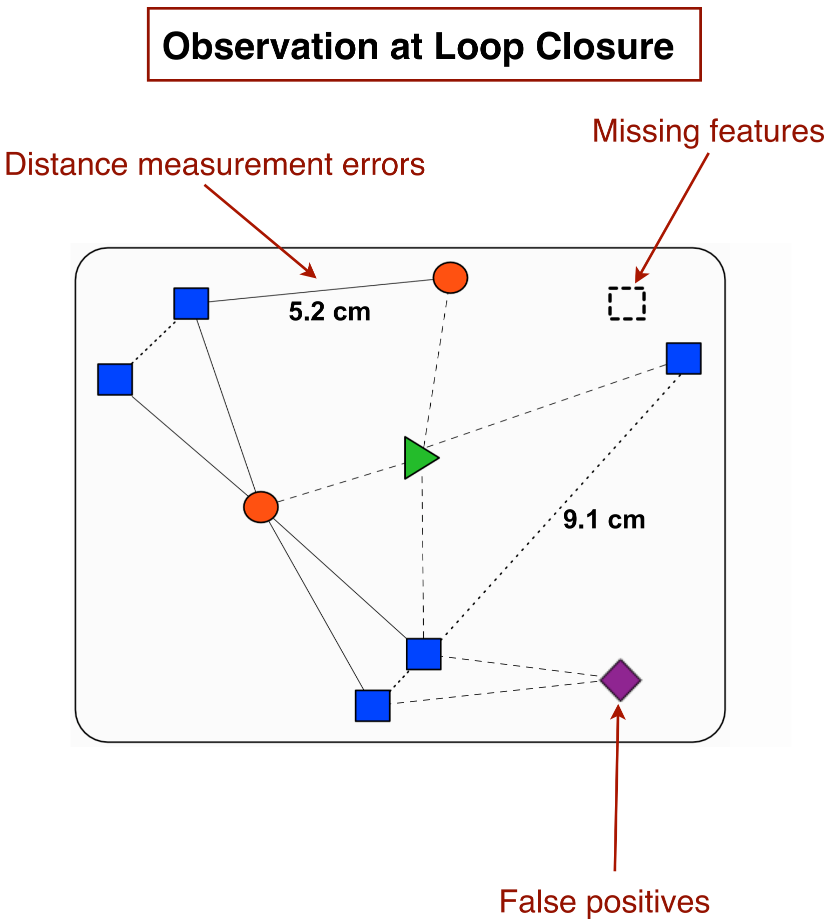
Original Observation

Observation at Loop Closure

- Scene change can occur due to illumination and viewpoint changes or dynamic objects.
- Noisy perception by sensors.

Understanding re-observation

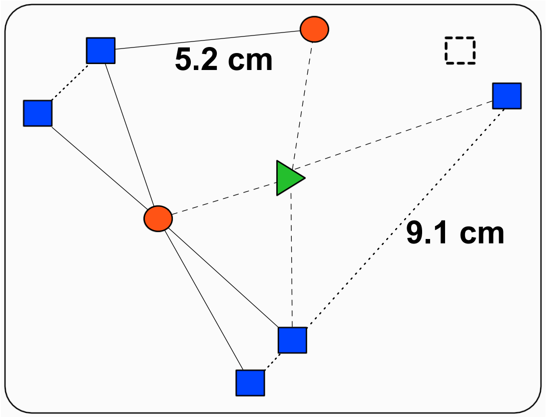




- Scene change can occur due to illumination and viewpoint changes or dynamic objects.
- Noisy perception by sensors.

Recognising a place for the second time....

Observation 2



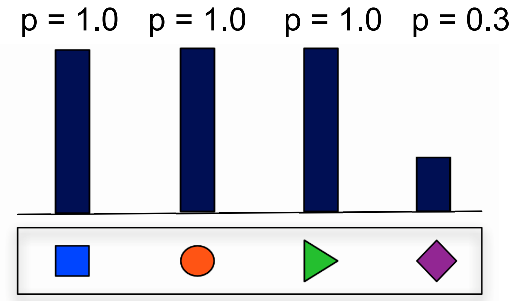
How likely are these features ?

Given these features,

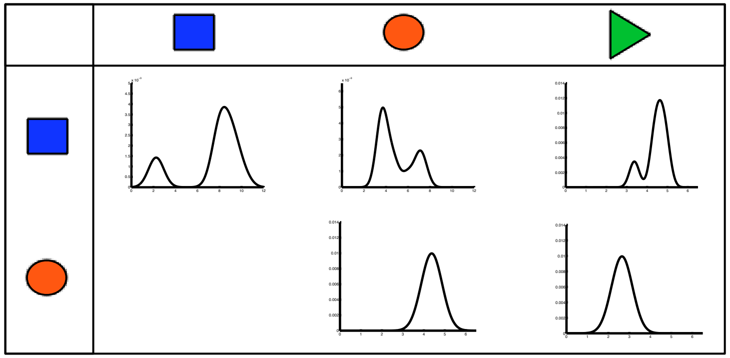
How likely is this configuration ?

Location Model

Likelihood of word existence



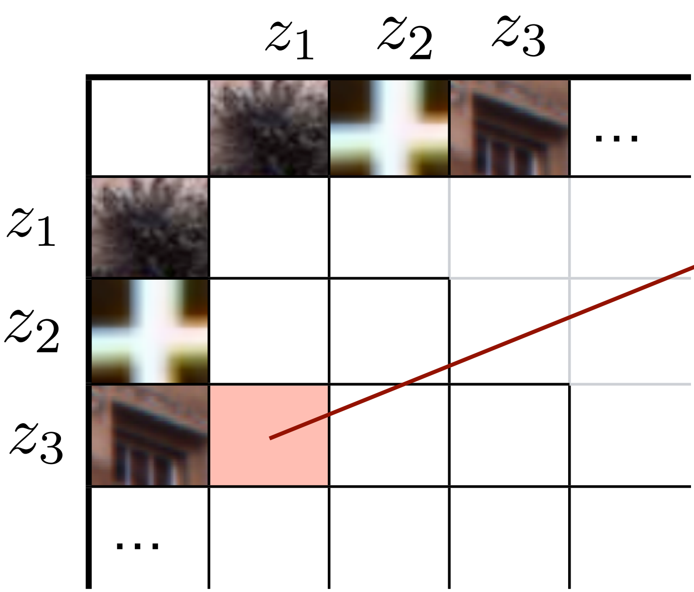
Distributions over inter-word distances



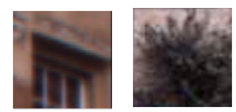
Learning Distributions over Feature Distances

Training Data

Word-pair Distances

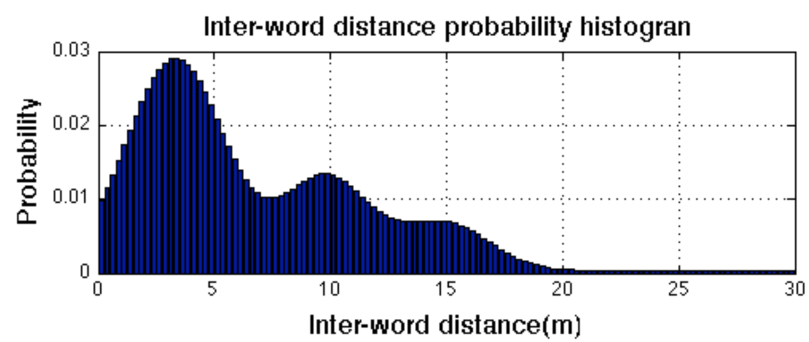


Observed distances between words



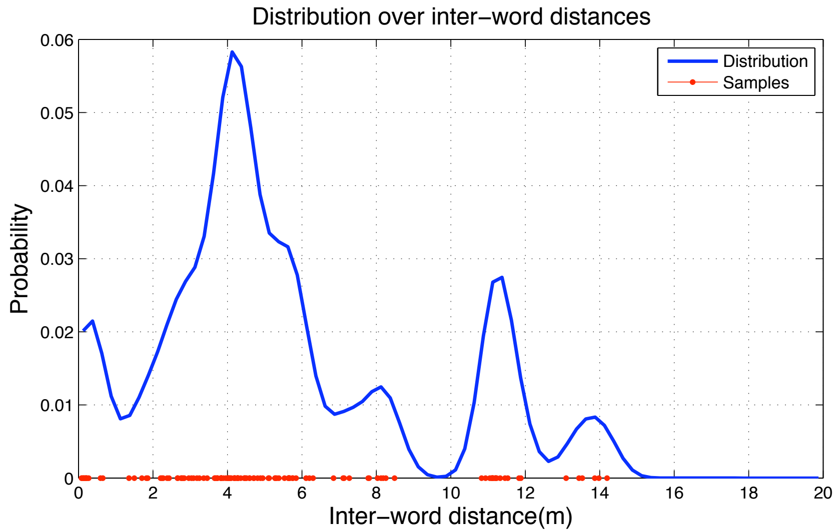
$$\{2.1, 2.3, 6.5, \ldots\}$$

Estimate likelihood over distances



Rohan Paul and Paul Newman

Kernel Density Estimation



Observed Distances

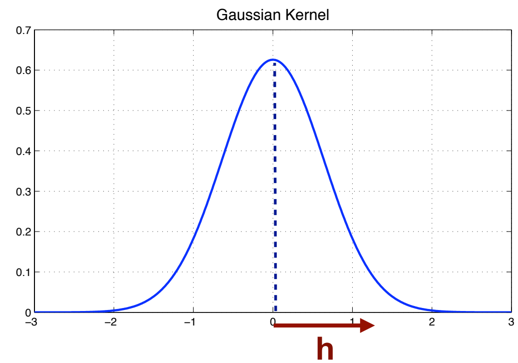
$$Distances = \{6.1, 3.1, 4.4, 7.8, ...\}$$

Non-parametric Kernel Density Estimation

$$\hat{p}(x) = \frac{1}{N\sqrt{2\pi}h}\sum_{i=1}^{N}\exp-(\frac{(x-x_i)^2}{2h^2})$$

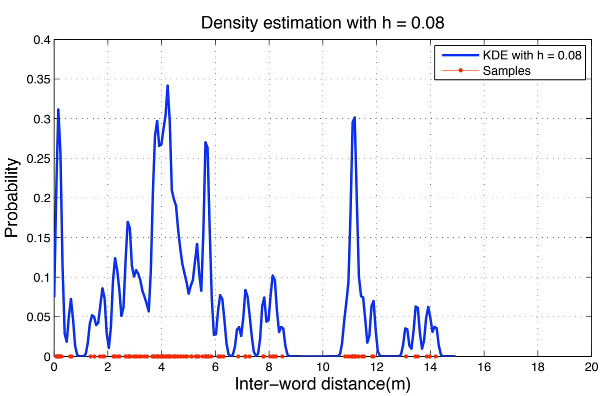
Kernel Bandwidth Selection

What kernel bandwidth to select ?

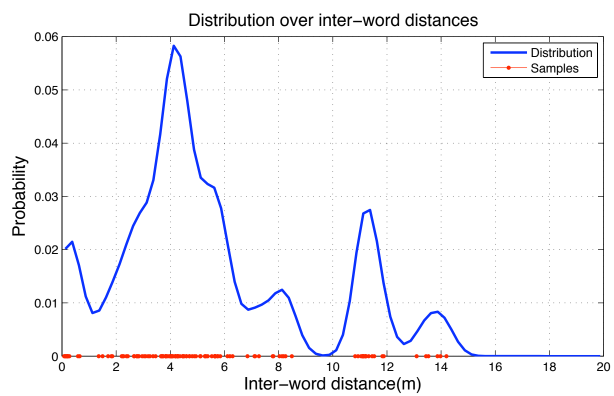


h = ?

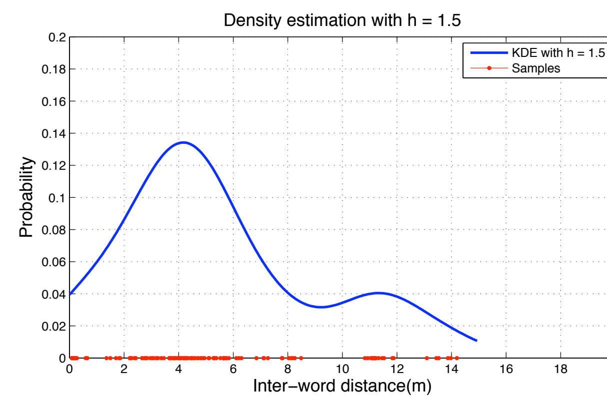
Too small: over-fitting	Optimal	Too large: under-fitting



h = 0.08



h = 0.4



h = 1.5

Optimal Bandwidth Selection

An optimal bandwidth for KDE

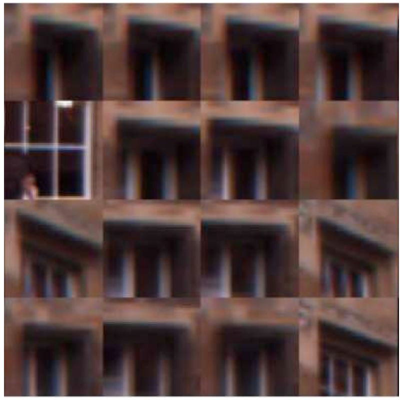
Error metric

- Mean Integrated Square Error $$\mathrm{MISE}(h) = E\int(\hat{f}_h - f)^2$$
- **Asymptotic** Mean Integrated Square Error $$\mathrm{AMISE}(h) = n^{-1}h^{-1}R(K) + h^4 R(f'')\left(\int x^2 K/2\right)^2$$
- Optimal h minimizing AMISE [Jones et al. 1996] $$h_{\mathrm{AMISE}} = \left[\frac{R(K)}{nR(f'')\left(\int x^2 K\right)^2}\right]^{1/5}$$

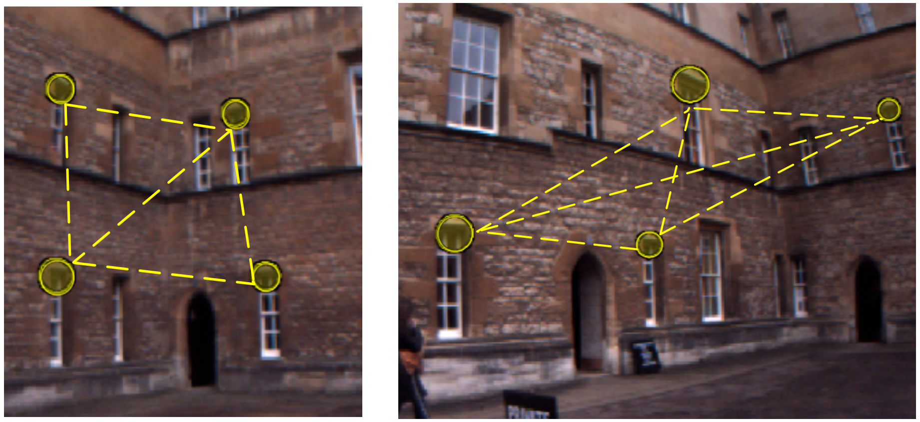
Can bandwidth be estimated fast? Yes.

Linear in Training Points

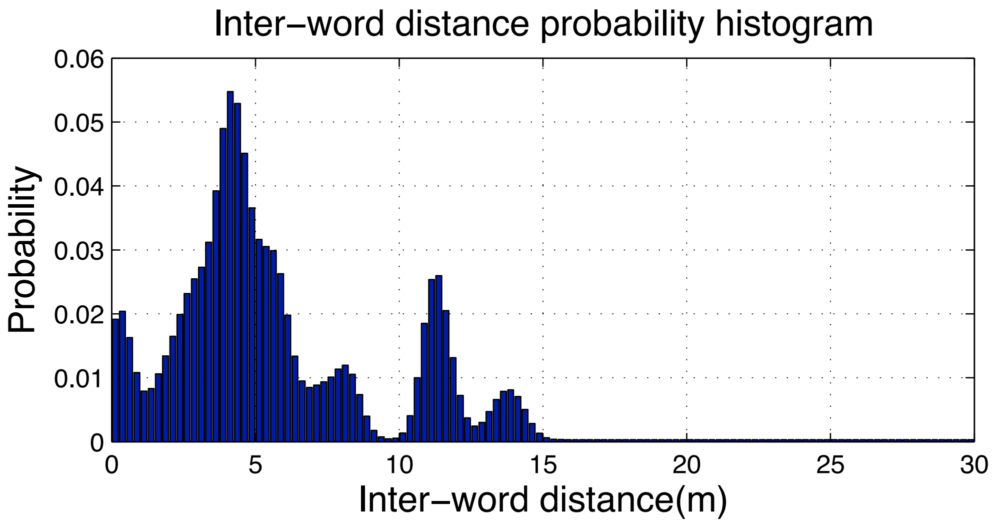
- Inverse Fast Gaussian Transform (IFGT) [Yang et al. 2003] — **Very fast. Lose error bound.**
- Computational Geometry: Dual Tree algorithm [Gray et al. 2003] — **Variety of kernels. Tight error bound.**



Visual features that appears on top of windows

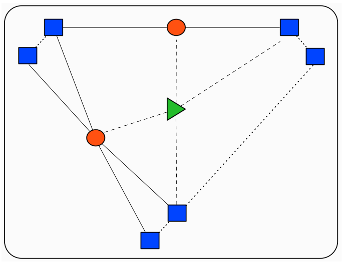


Visual feature observed in outdoor scenes



Multi-modal distribution learnt through kernel density estimation with optimal bandwidth selection

Observation

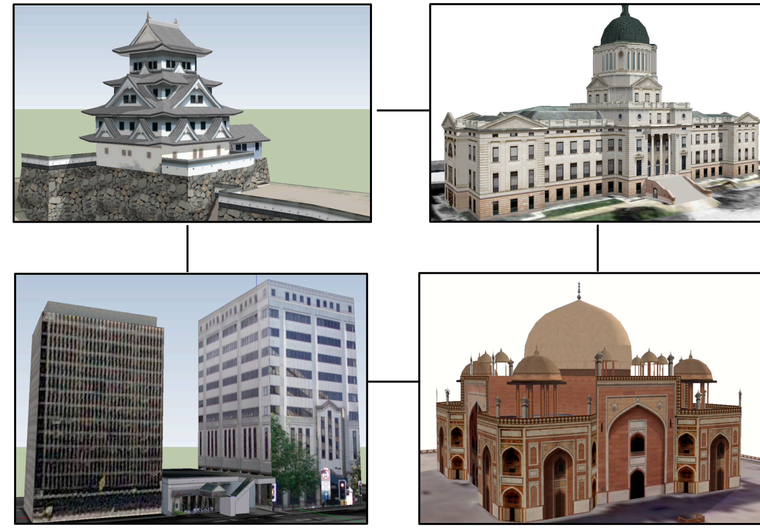


$G_k = \{Z_k, D_k\}$

Is this a new place?

New Place

Is this a place in the map?



How likely is this location given all places?

Observation Likelihood

Prior

$$p(L_n|\mathcal{G}^k) = \frac{p(G_k|L_n, \mathcal{G}^{k-1})p(L_n|\mathcal{G}^{k-1})}{p(G_k|\mathcal{G}^{k-1})}$$

Partition Function

A Promising Factorisation....

What's the probability that *observation,* G_k came from *location,* L_n

$$
\begin{aligned}
p(G_k|L_n, \mathcal{G}^{k-1}) &= p(G_k|L_n) \\
&= p(\{Z_k, D_k\}|L_n) \\
&= p(D_k|Z_k, L_n)p(Z_k,|L_n)
\end{aligned}
$$

How likely are the *graph distances,* D_k given these features and location, L_n

How likely are these features, Z_k at this location, L_n

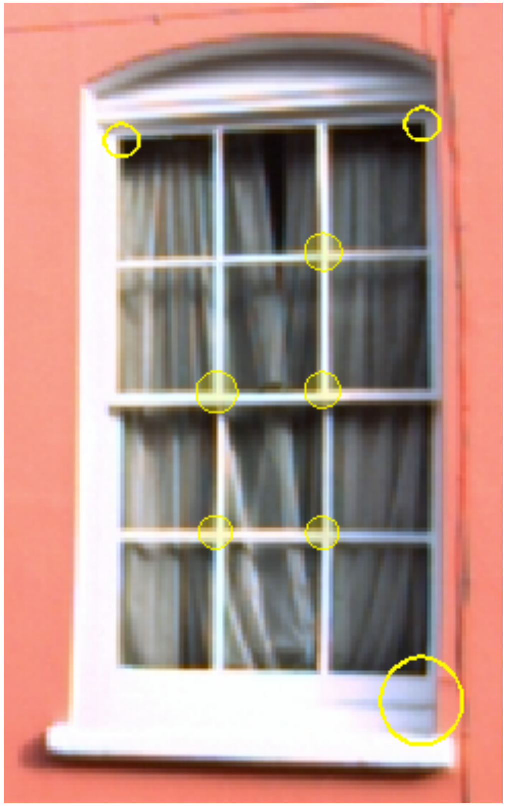
Spatial component

Visual component (this is Vanilla FABMAP!)

Visual Appearance

$$p(Z_k|L_n)$$

- Visual Words are **Not Independent**.
- Presence of some words is **correlated** as they are generated from the same underlying **objects**.
- Learn correlations via **mutual information** between features from training data.



[Cummins and Newman IJRR08]

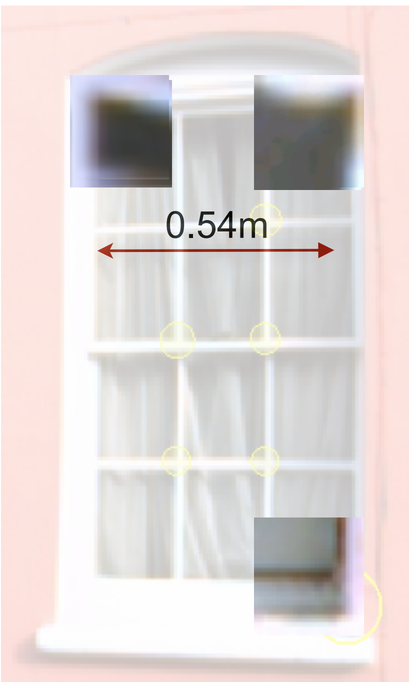
Evaluating Spatial Likelihood

Spatial appearance term conditioned on visual appearance

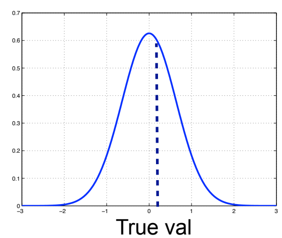
$$p(D_k|Z_k,L_n) = \prod_{i,j=1}^{|v|}\prod_{n=1}^{C_{ij}}\sum_{r=1}^{R}\underbrace{p(d^n_{ij}|h_{ij}=b_r)}_{Det_{range}}\underbrace{p(h_{ij}=b_r|L_n)}_{histogram}$$

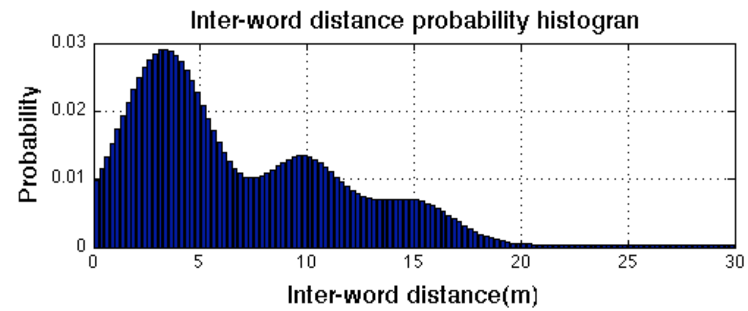
Likelihood of observing 0.54m given a noisy sensor

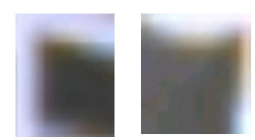
What is our prior belief over distances?



Range measurement model



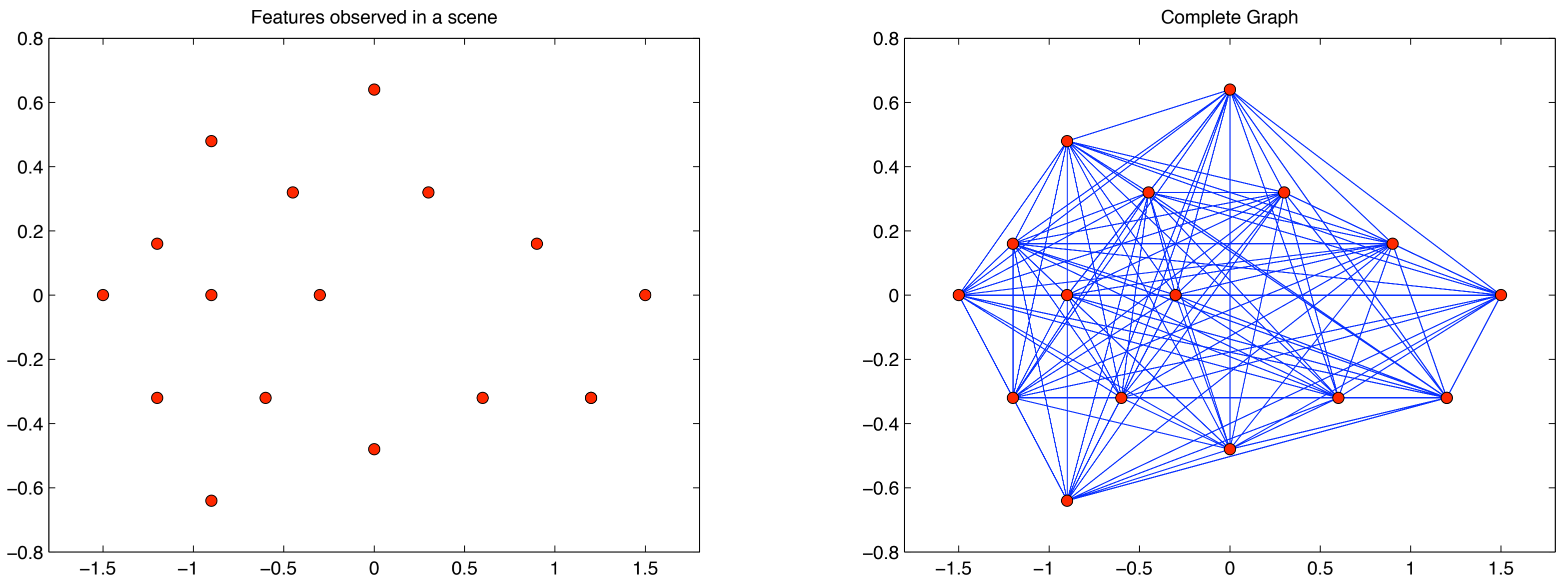




Account for all possible bins, across all measured range occurrences, over all observed pairs and incorporate a range sensor model

How many graph edges to keep?

If N_f features are detected in scene how many pairwise distances should be checked ?



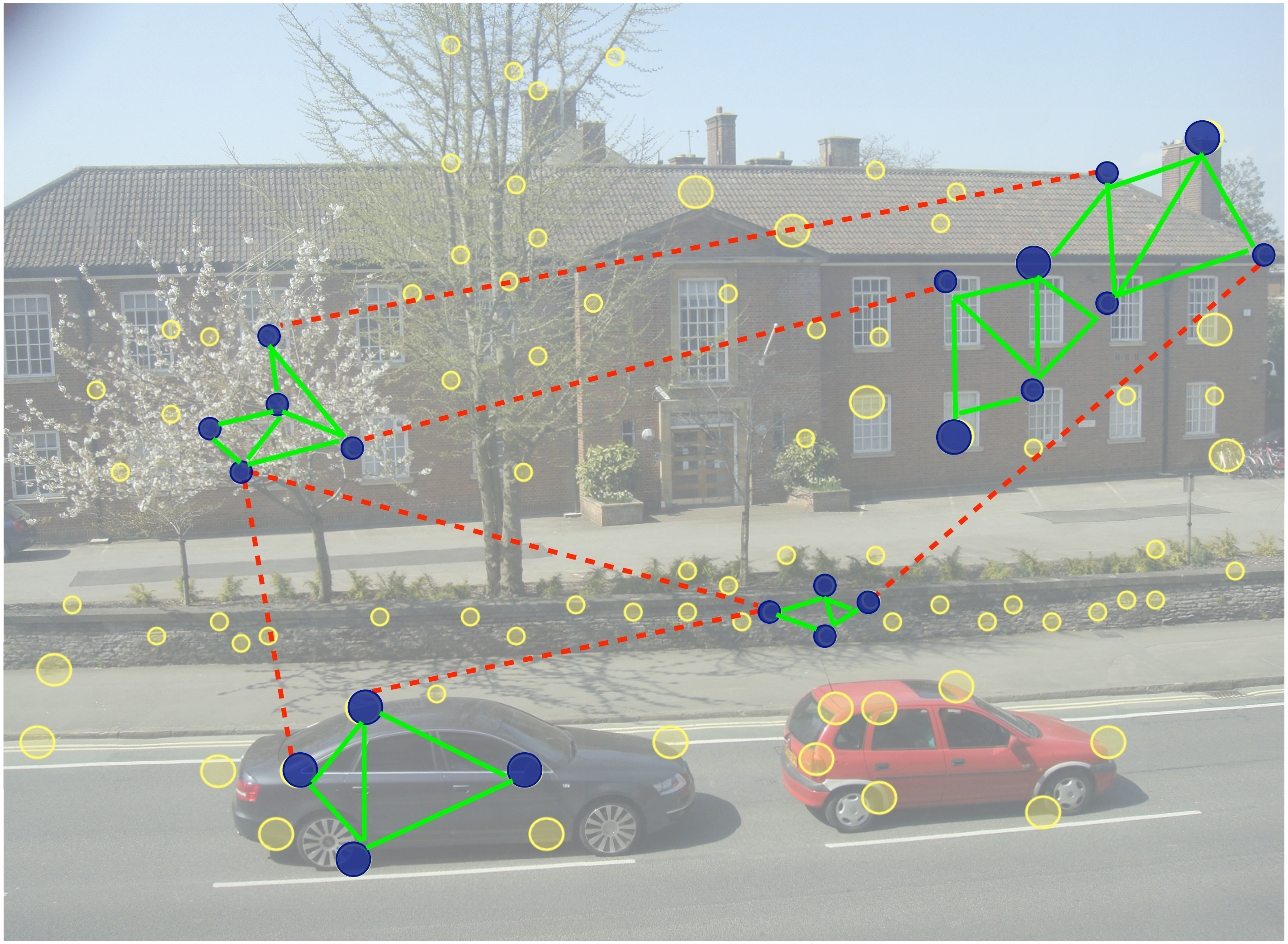
Checking all pairwise distances causes $O(N_f^2)$ histogram updates - not pleasant

Which graph edges to keep?

Insight: Features usually originate from objects possessing high local spatial correlation.

Consider distances to *neighboring* points.

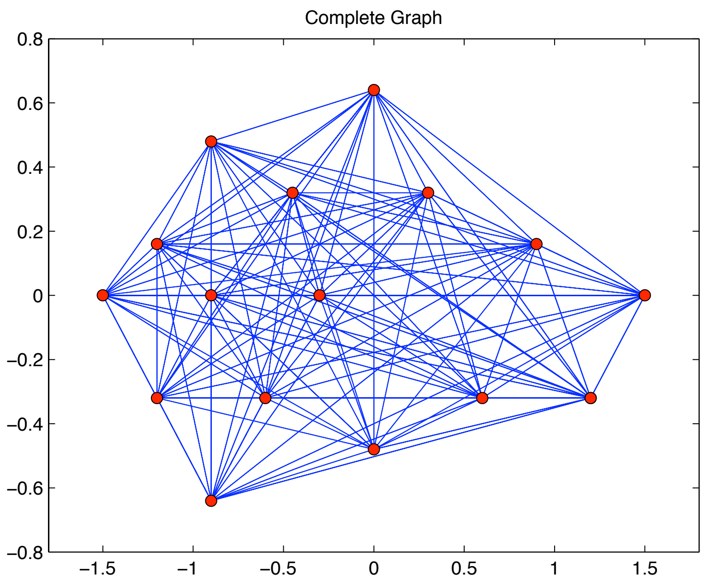
Local Spatial Correlations are Common



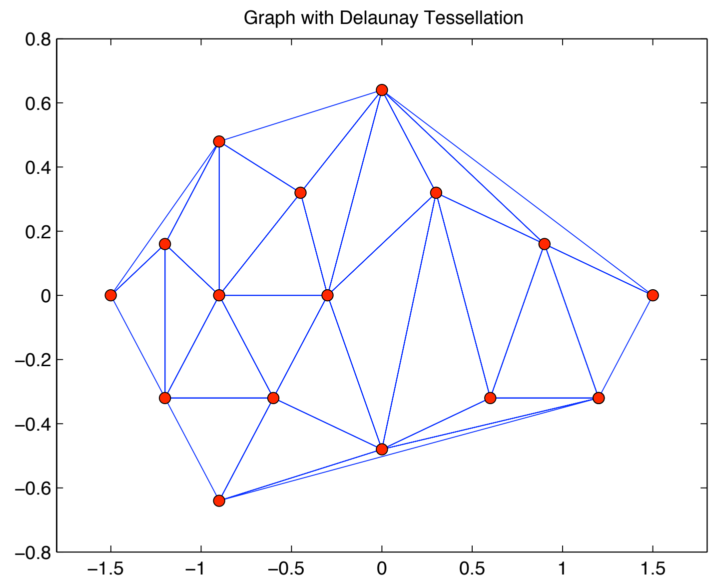
Insight: Preserve local spatial correlations by choosing only neighboring points.

Delaunay Tessellation

Delaunay Tessellation is a triangulation such that no point is inside the circumcircle of any triangle.



Complete Graph

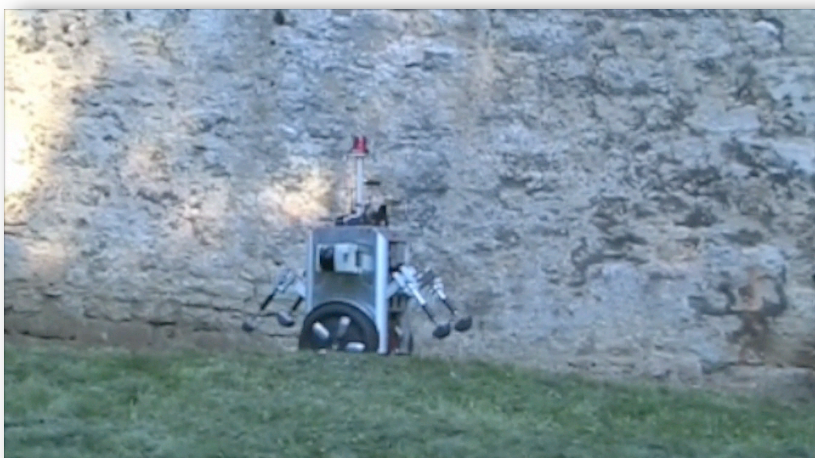


After Delaunay Tessellation

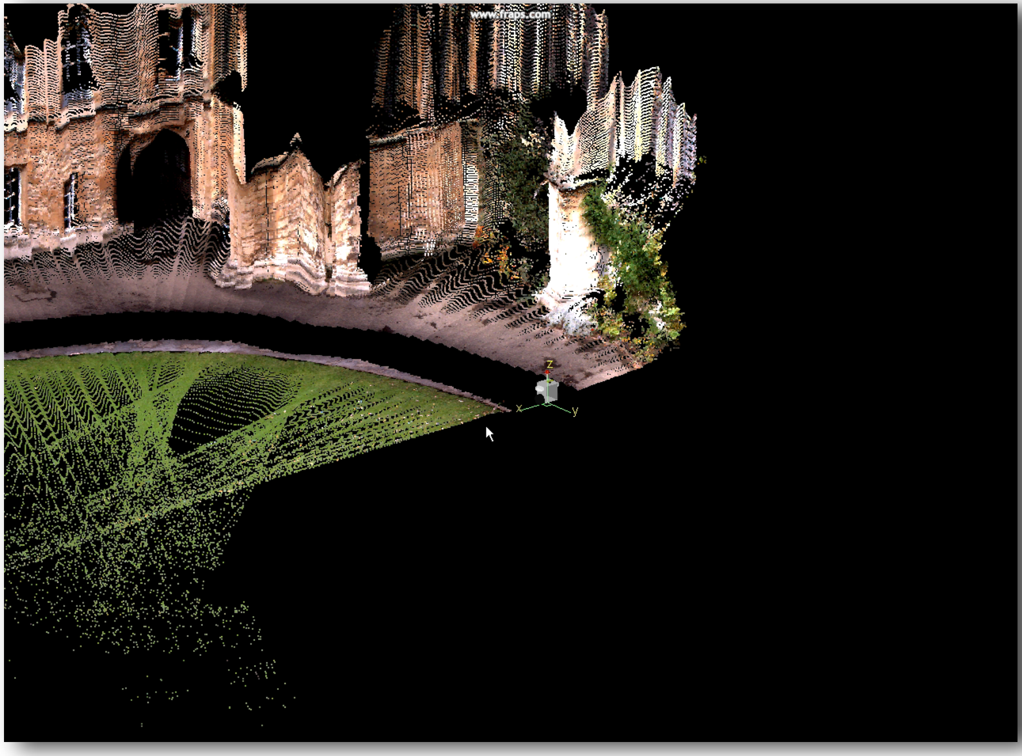
After tessellation $O(N_f)$ pairwise histograms are updated.

Tessellation algorithm has $O(N_f log N_f)$ complexity.

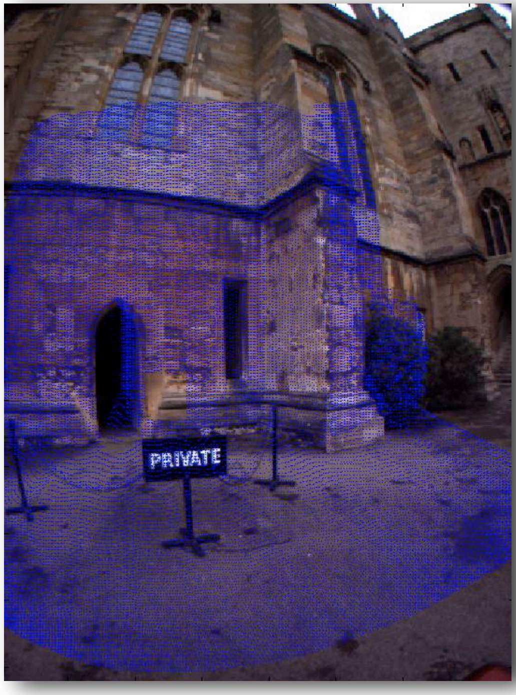
Evaluation - New College Data Set (Smith IJRR09)



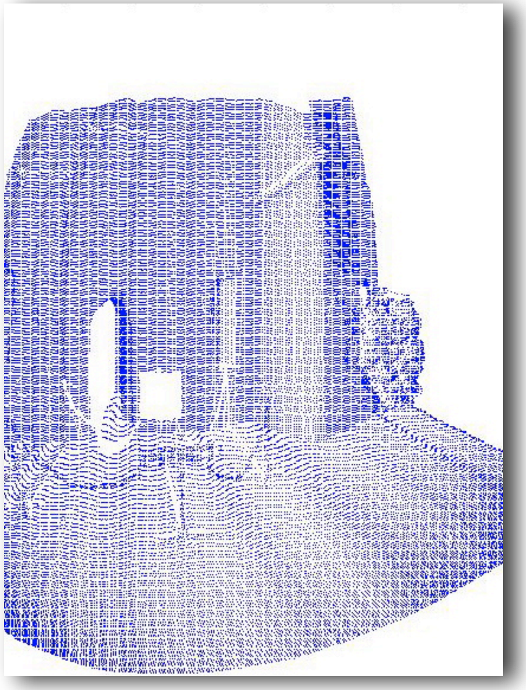
Combining Vision and Laser



Obtaining 3D Coordinates for Visual Features



Laser points projected into the camera frame



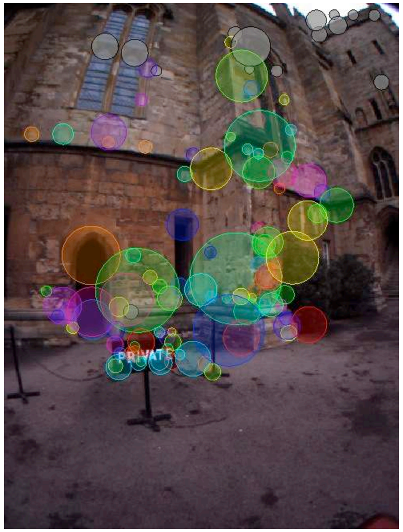
Laser point cloud

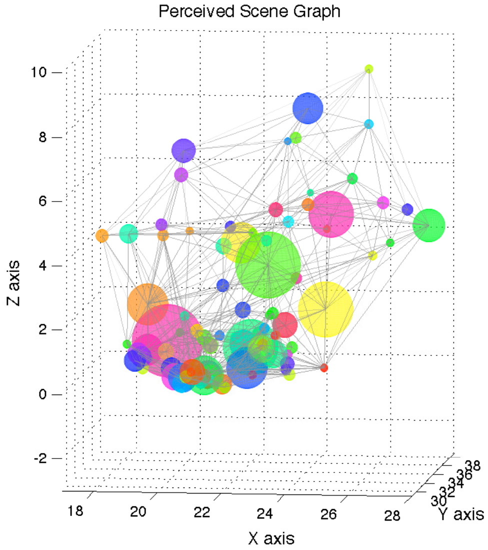
Obtaining 3D coordinates for visual words by projecting laser points into camera frame after cross-calibration.

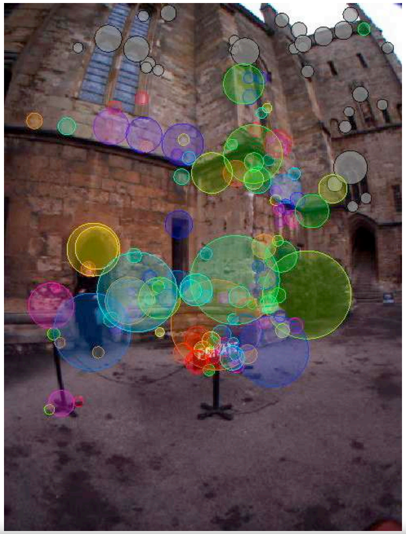
Results: Example Picking up Missed Loop Closures

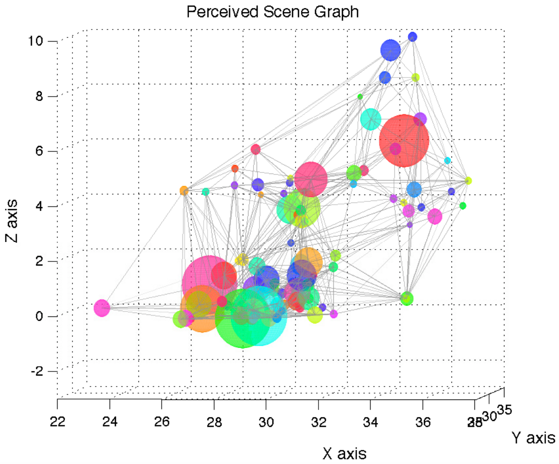
Scene

Perceived Graph



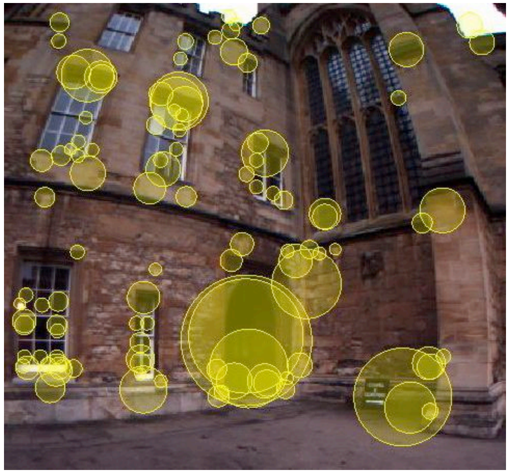
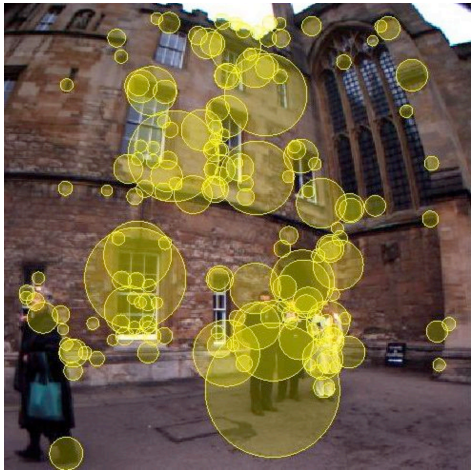




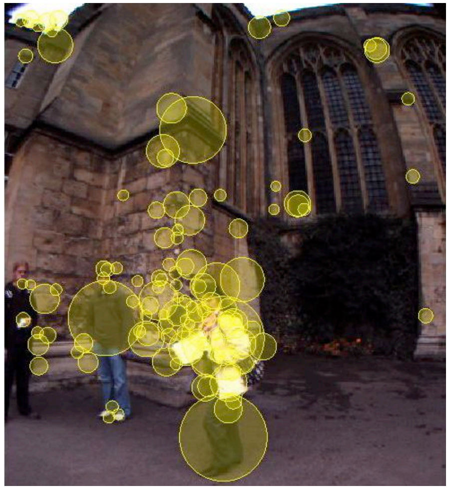


Distinctive spatial similarity in graphs used by FAB-MAP 3D to infer a loop closure.

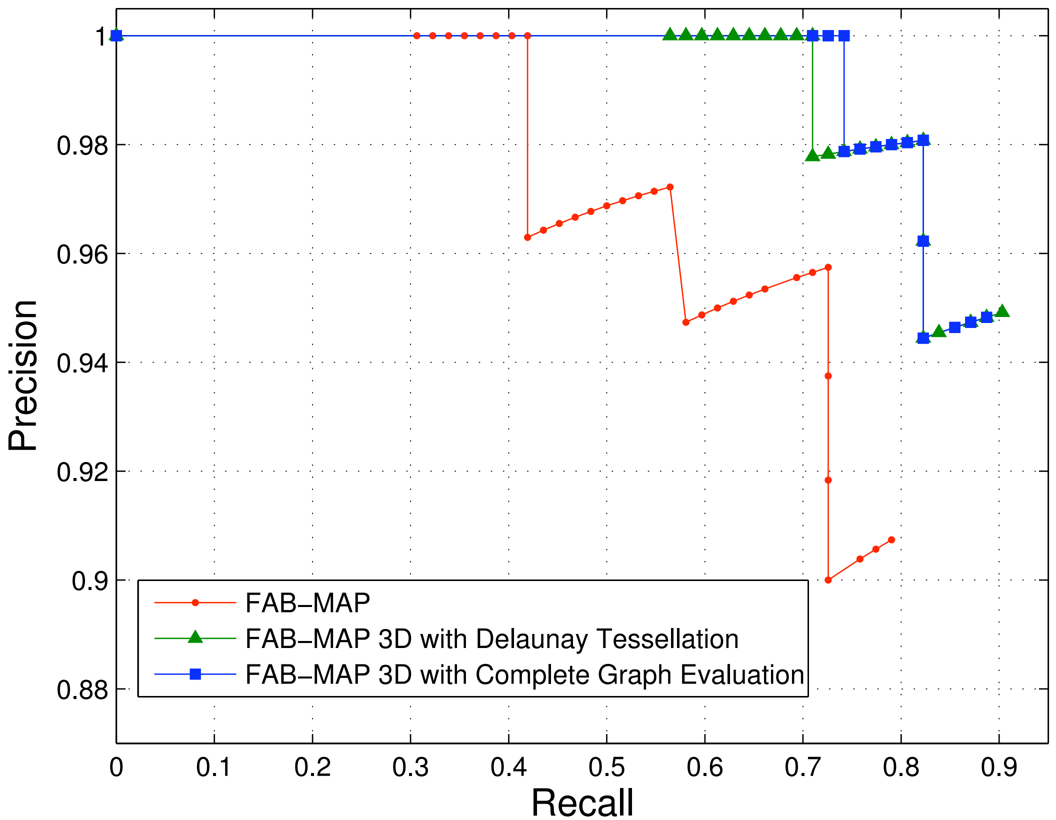
Results: Example Picking up Missed Loop Closures

Loop closures declared by FABMAP 3D using spatial similarity, missed by FABMAP

Precision-Recall Curves - The Central Result

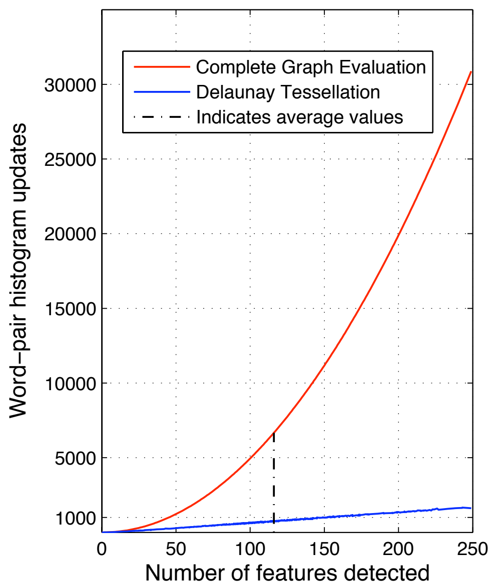


Algorithm	**Recall at 100% Precision**	**Order (in Scene Complexity)**	**Order in Num Scenes**
FAB-MAP	42%	Linear	Linear
FAB-MAP 3D with Complete Graph Evaluation	74%	Quadratic	Linear
FAB-MAP 3D with Delaunay Tessellation	71%	Log-Linear	Linear

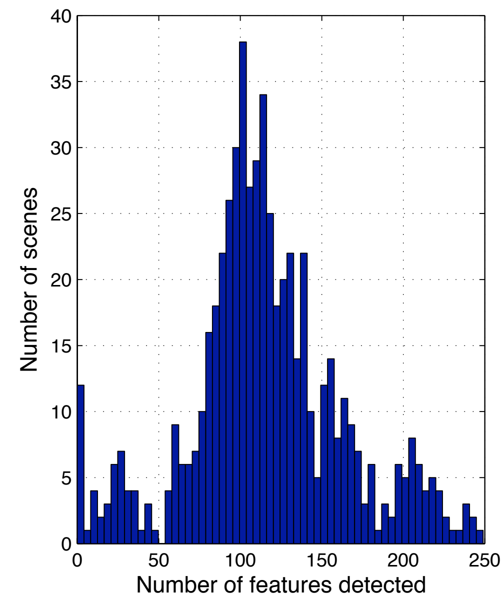
Computational Overhead of FAB-MAP 3D

- Online: avg. 314ms inference time/place
- Offline: 4.5 hrs for one off density estimation
- MATLAB implementation, 2.66GHz Intel Core 2 Duo machine.

Speed-up with Delaunay Tessellation

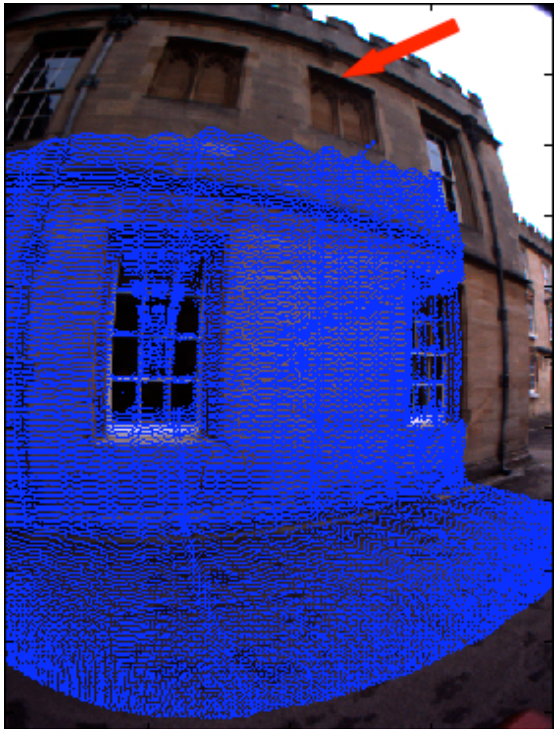


Histogram: Number of Features per Scene

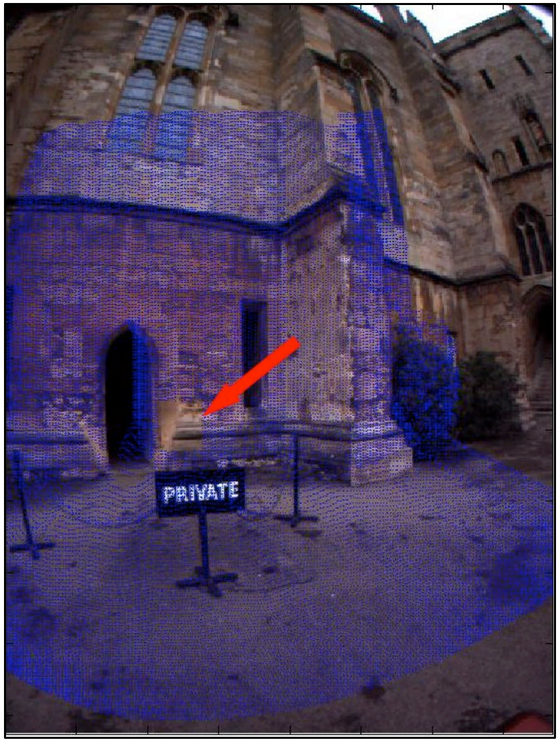


Mean 116, Median 112, Std. dev. 48

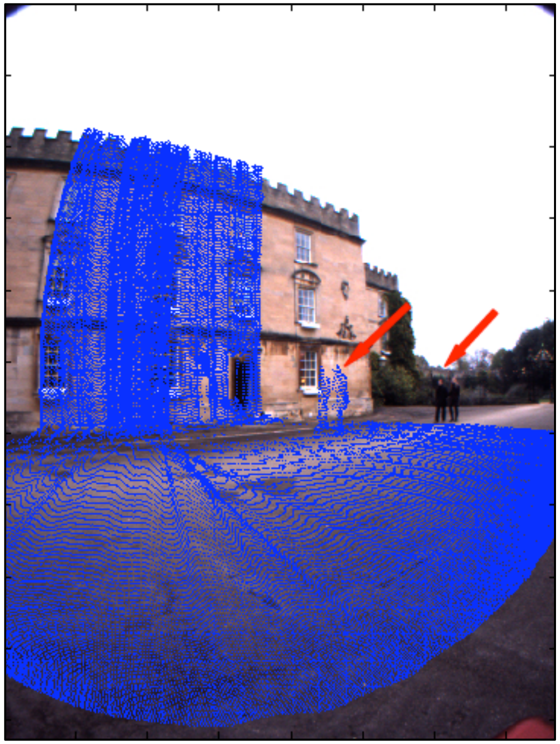
Limitations



Incomplete laser coverage



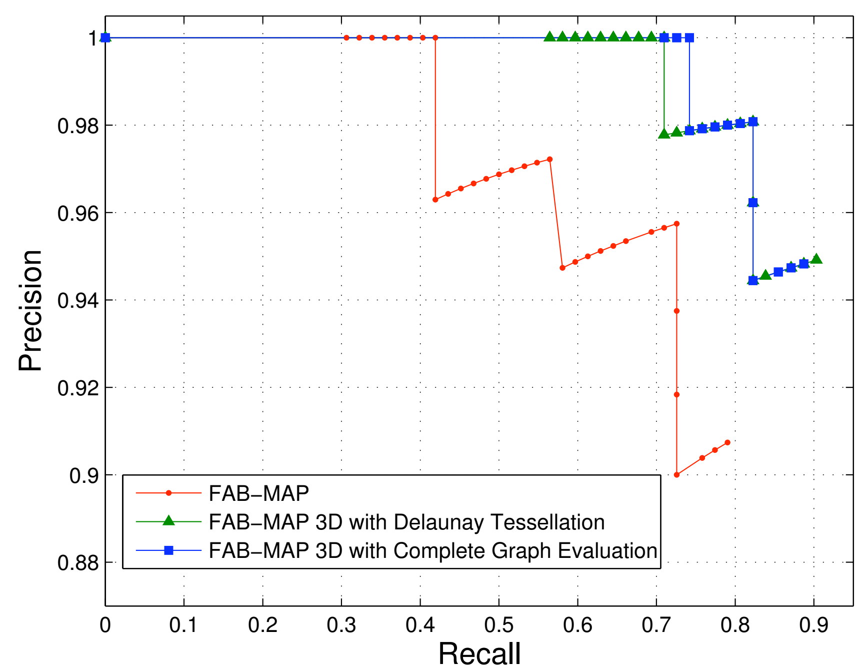
Shadows



Dynamic Objects

Conclusions - What have we done?

- Made use of easy to obtain local metric information (SFM, Stereo, Laser).
- We learn at run time a probabilistic generative place model which captures visual appearance (feature existence) *and* relative geometry.
- This makes a marked difference to precision recall - greatly increased recall at 100% precision
- Algorithm is linear in number of places



FABMAP-3D fully exploits scene structure and constitutes a new way to undertake robotic mapping with vision and laser.

Thank You

Rohan Paul and Paul Newman

Mobile Robotics Group

University of Oxford