Advanced Techniques for Mobile Robotics

Bag-of-Words Models & Appearance-Based Mapping

Wolfram Burgard, Cyrill Stachniss,

Kai Arras, Maren Bennewitz

Motivation: Analogy to Documents

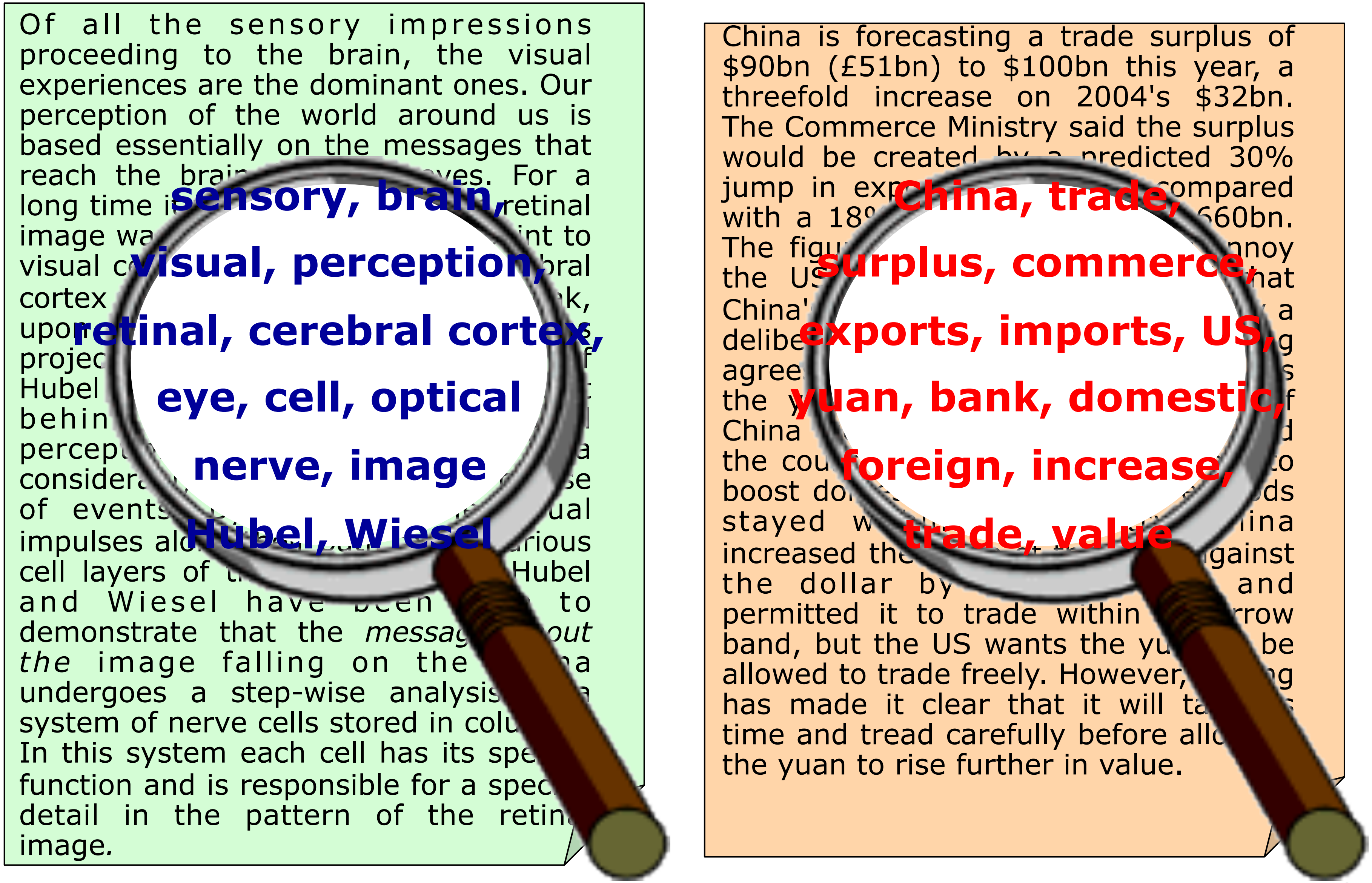


image source: L. Fei-Fei

Object Classification / Scene Recognition

- Analogy to documents: The content can be inferred from the frequency of words

object

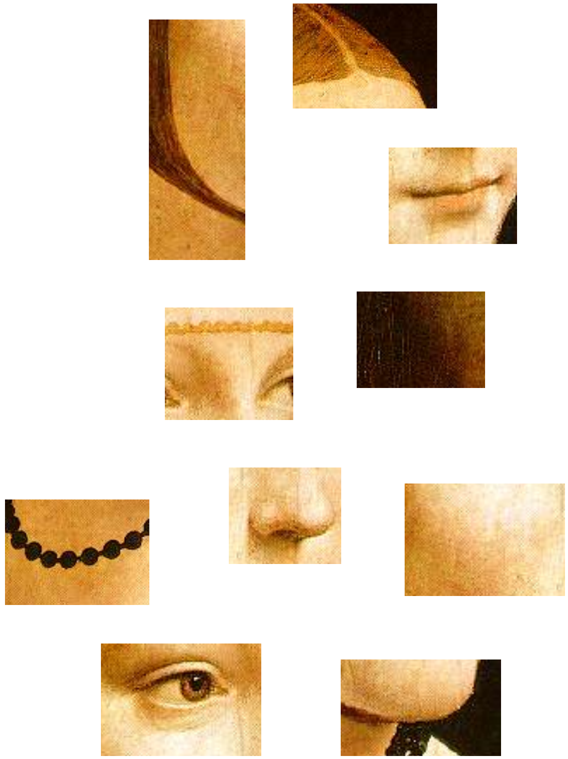
bag of
“visual words”

image source: L. Fei-Fei

Bag of Visual Words

- Visual words = independent features

face



features

image source: L. Fei-Fei

Bag of Visual Words

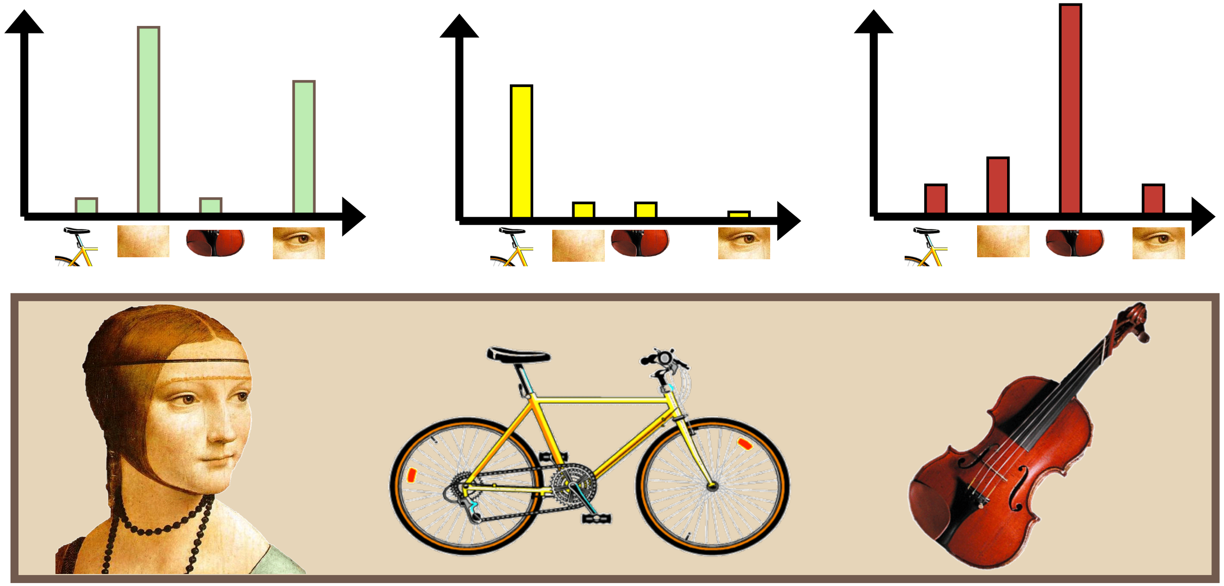
- Visual words = independent features
- Construct a dictionary of representative words

codewords dictionary

image source: L. Fei-Fei

Bag of Visual Words

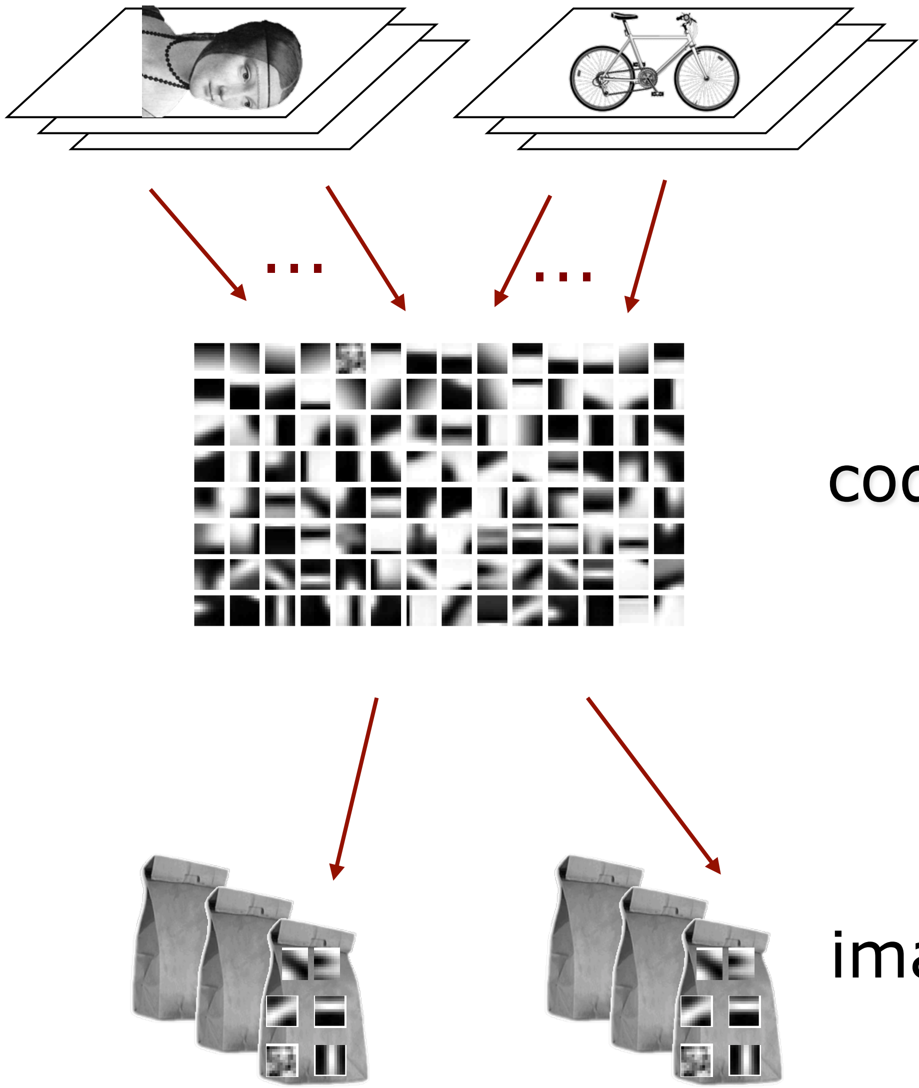
- Visual words = independent features
- Construct a dictionary of representative words
- Represent the images based on a histogram of word occurrences (bag)



Each detected feature is assigned to the closest entry in the codebook

image source: L. Fei-Fei

Overview



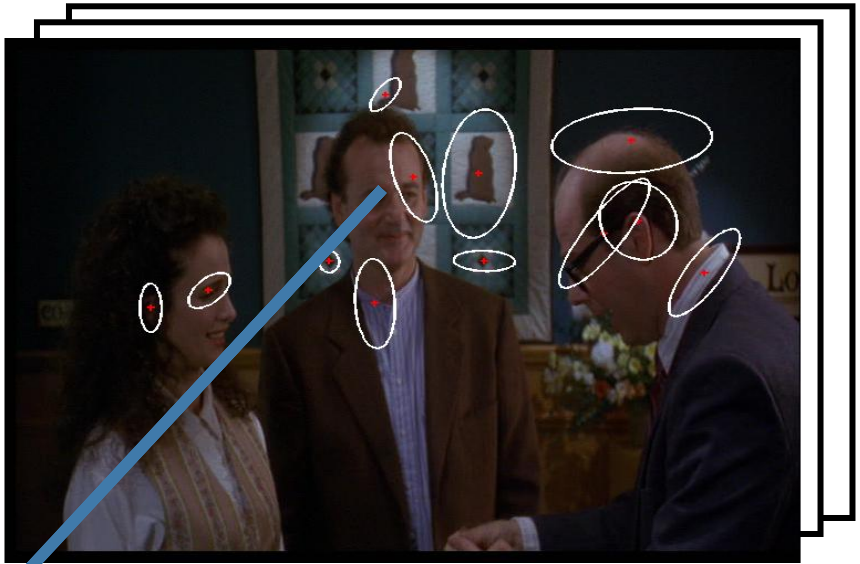
feature detection
&
representation

codewords dictionary

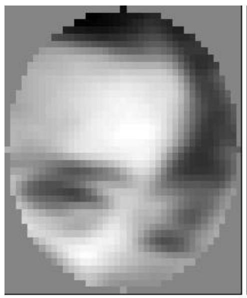
image representation

slide adapted from: L. Fei-Fei

Feature Detection and Representation



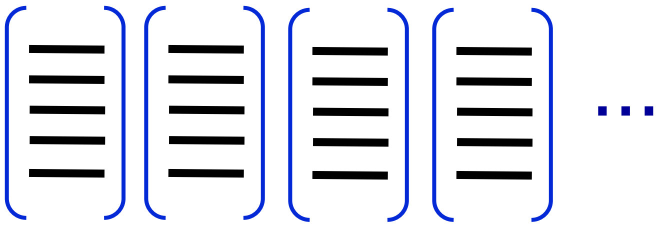
detected features in a set of training images (intensity changes)



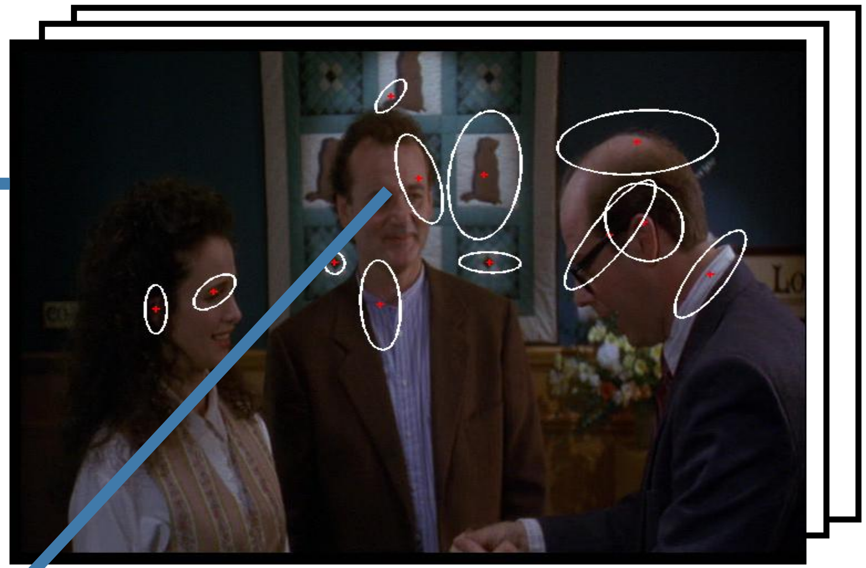
example patch

slide adapted from: L. Fei-Fei

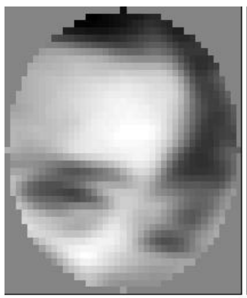
Feature Detection and Representation



descriptor vectors (e.g., SIFT/SURF, consider local orientations of gradients)



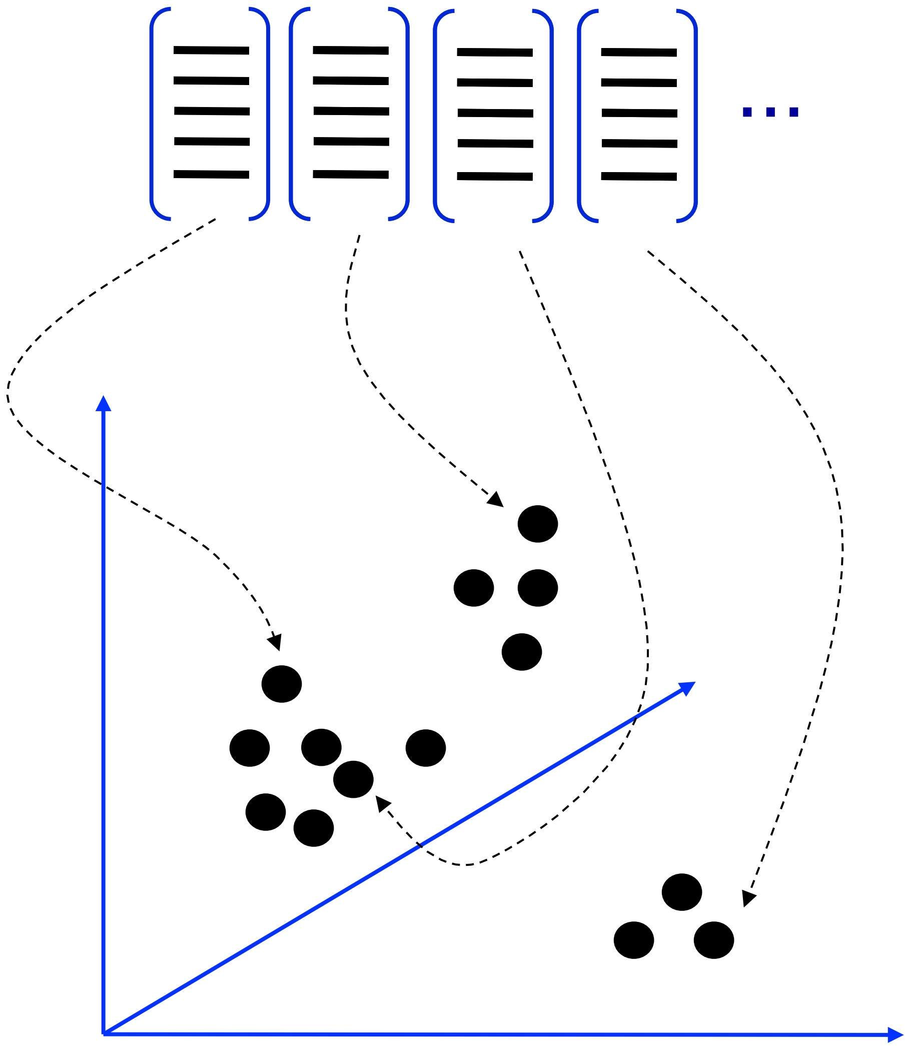
detected features in a set of training images (intensity changes)



example patch

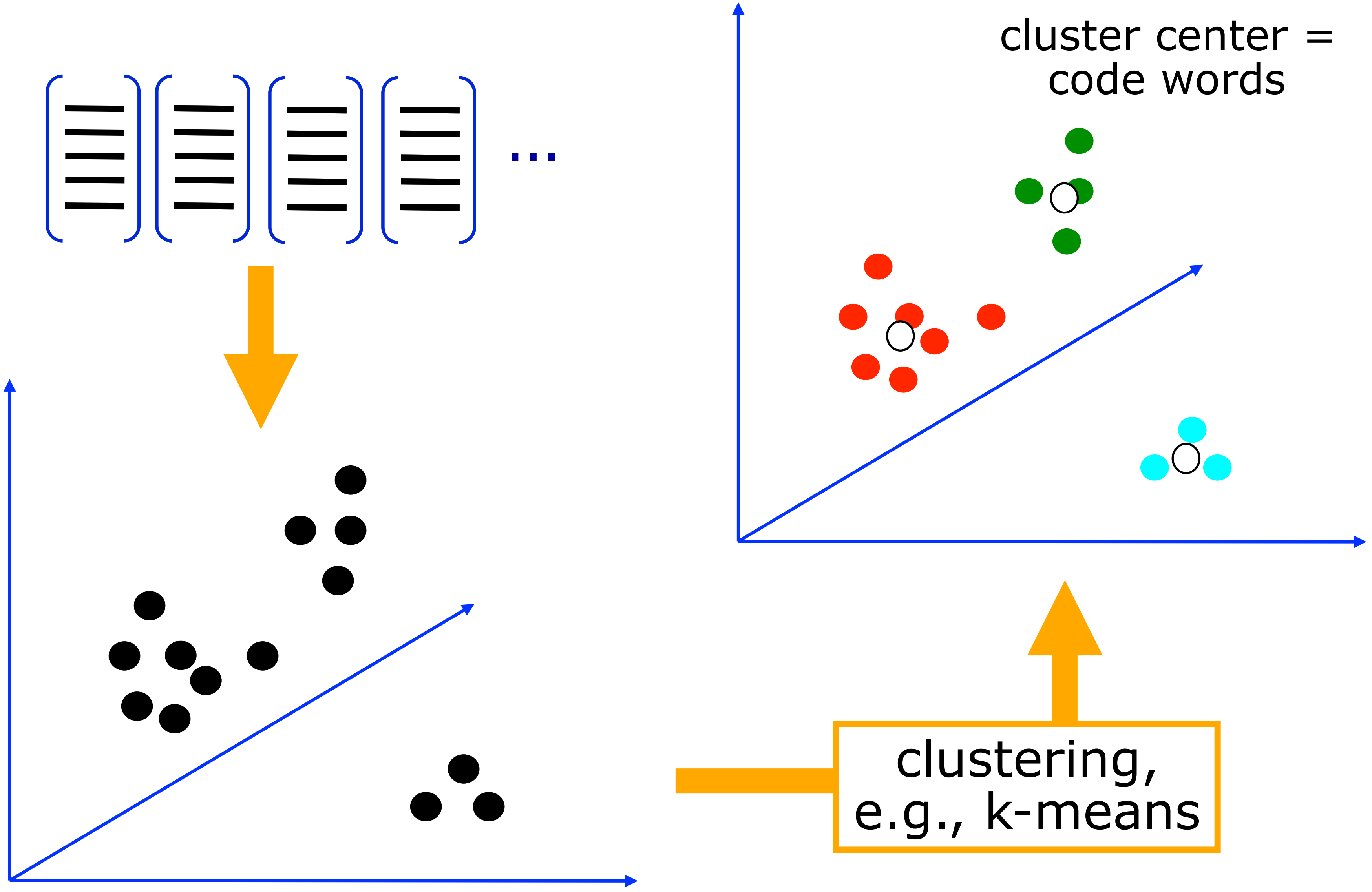
slide adapted from: L. Fei-Fei

Learning the Dictionary



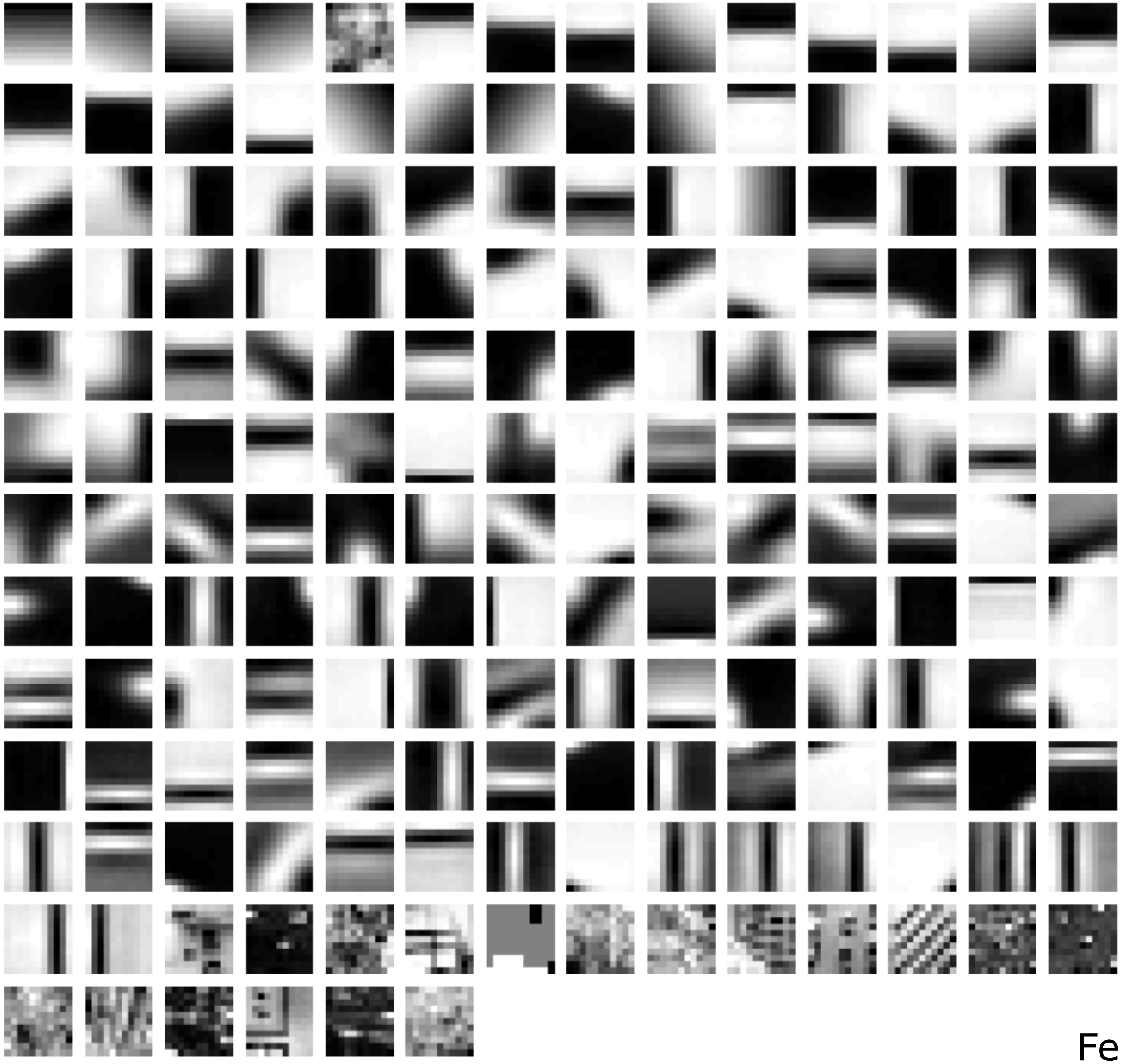
slide adapted from: L. Fei-Fei

Learning the Dictionary



slide adapted from: L. Fei-Fei

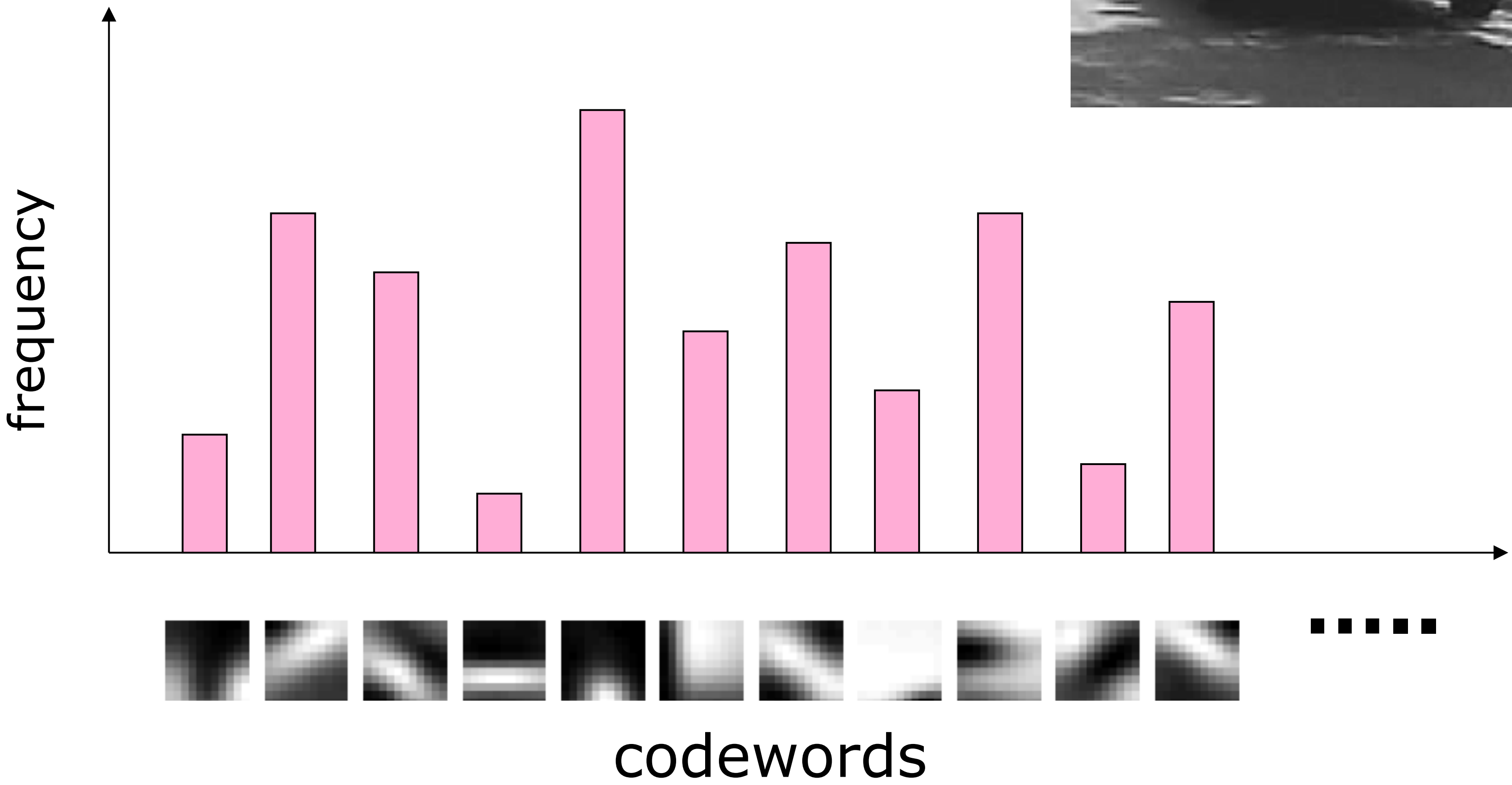
Example Codewords Dictionary



Fei-Fei et al. 2005

Example Image Representation

- Build the histogram by assigning each detected feature to the closest entry in the codebook



slide adapted from: L. Fei-Fei

Properties Bag-of-Words

- Compact summary of content
- Flexible to viewpoint, deformations
- Can be used for object / image classification by comparing the histograms (and applying some discriminative method)
- Ignores geometry
- Unclear how to choose optimal vocabulary
  - Too small: Words not representative of all patches
  - Too large: Artifacts, over-fitting

Appearance-Based Mapping with a Bag-of-Words Approach

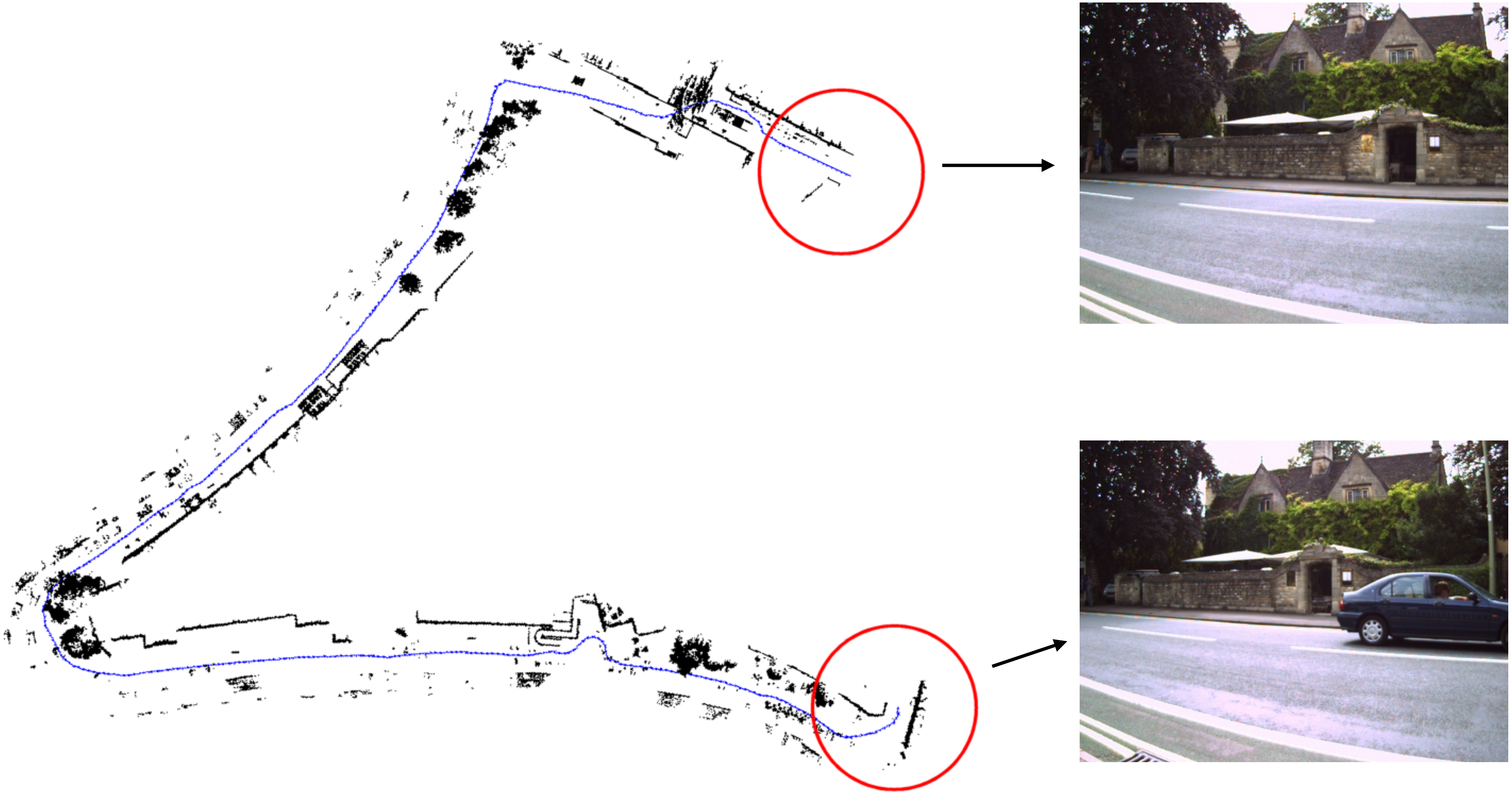
- Based on M. Cummins & P. Newman

  FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance
  Int. Journal of Robotics Research, 2008

  Appearance-only SLAM at Large Scale with FAB-MAP 2.0
  Int. Journal of Robotics Research, 2010

- Slides based on a presentation of Mark Cummins at R:SS 2009

Motivation: Failure of Metric SLAM



Appearance information can help to recover the pose estimate where metric approaches may fail

Appearance-Based Mapping (1)

- Recognize places based on visual appearance, even under difficult conditions
- Decide whether observations result from places already in the map, or from new, unseen places
- Difficult problem since different places may have similar visual appearance (and vice versa)
- Apply a bag-of-words approach
- Extension: Take into account that certain combinations of words co-occur

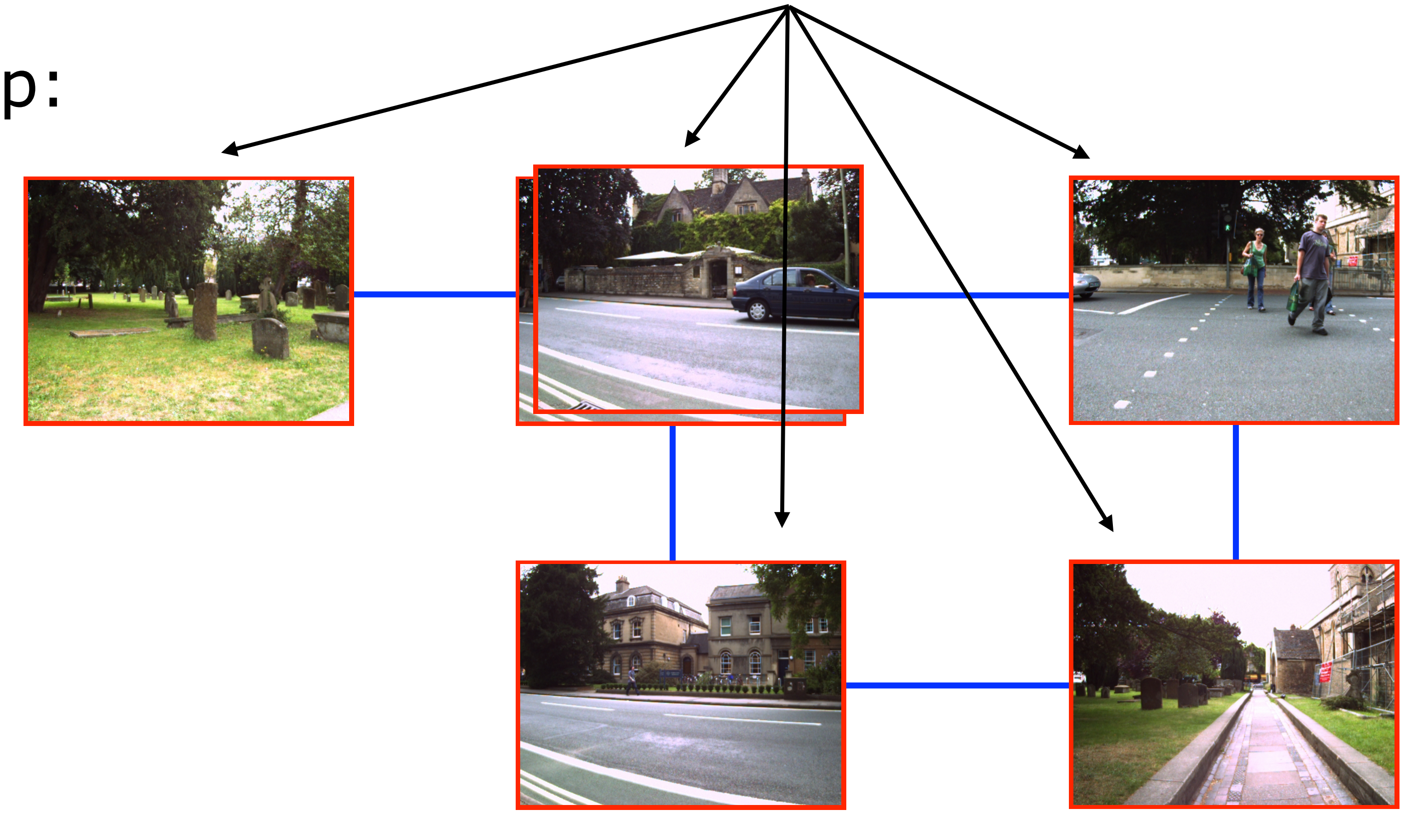
Appearance-Based Mapping (2)

- Parameterize the world as a set of discrete locations
- Estimate their positions in an appearance space
- Distinctive places can be recognized even after unknown motion (loop-closure)

Example

current
observation:

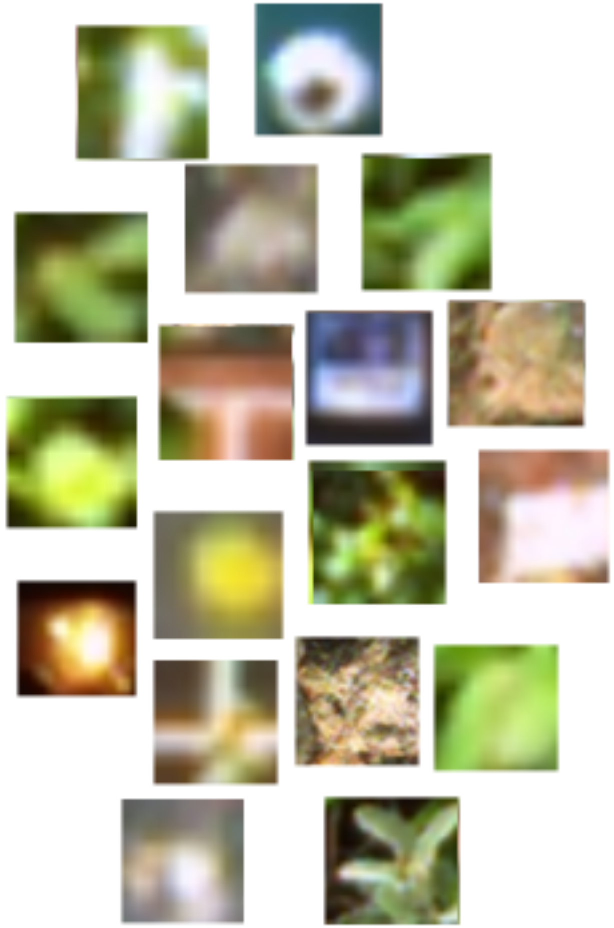
map:



Data Collection Platform

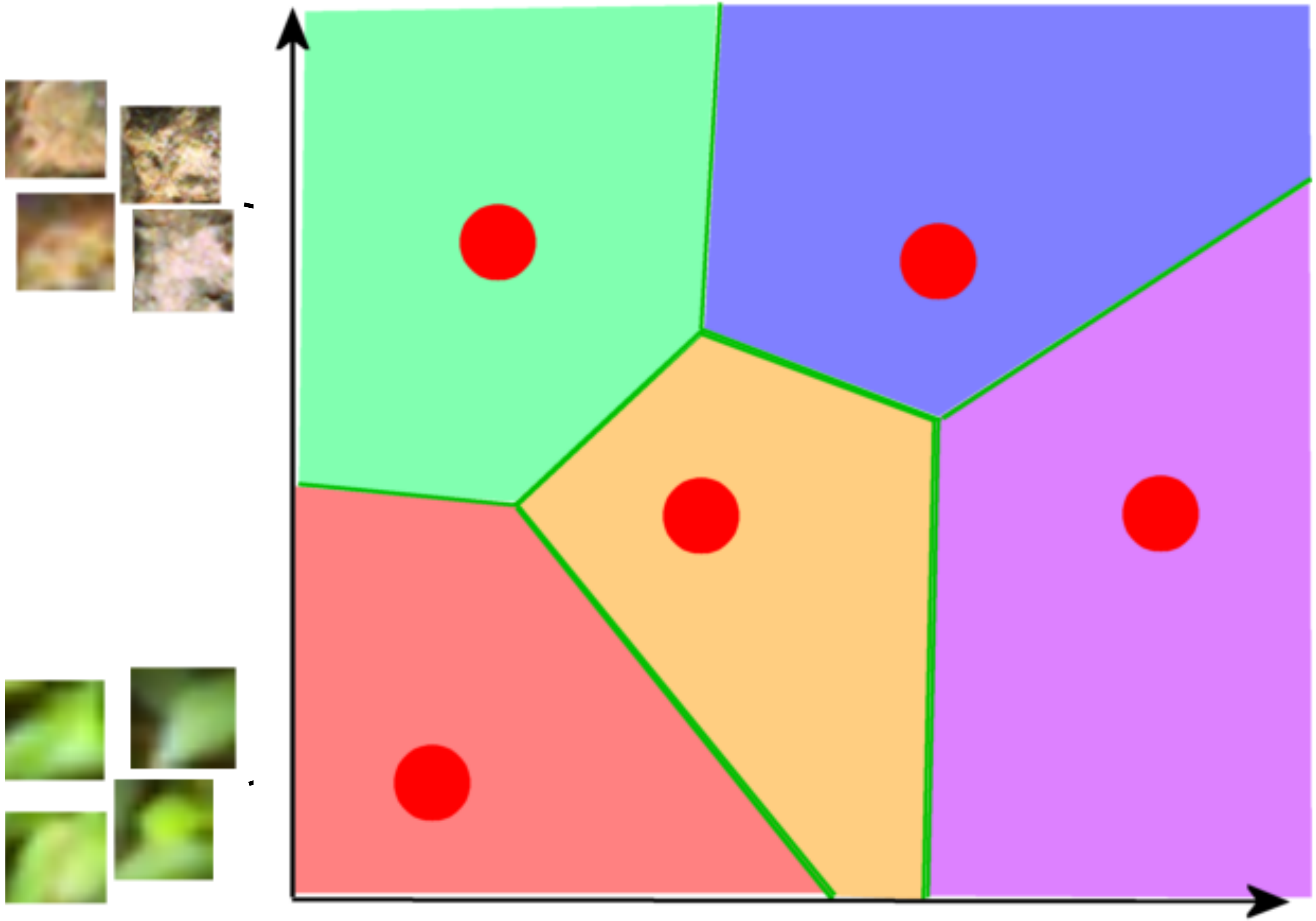
Learning the Visual Vocabulary

feature
extraction

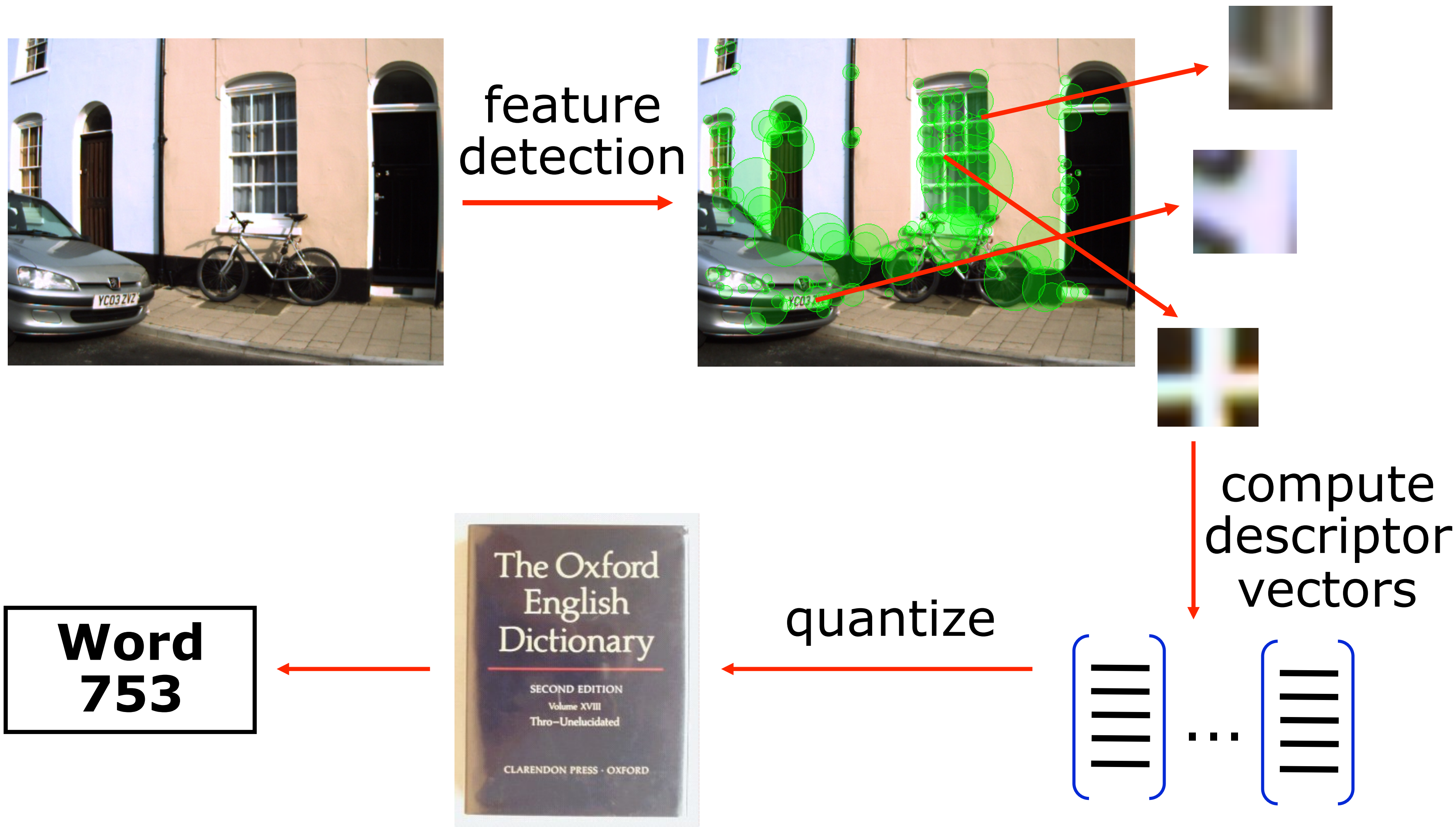


SURF

Clustering in Feature Space

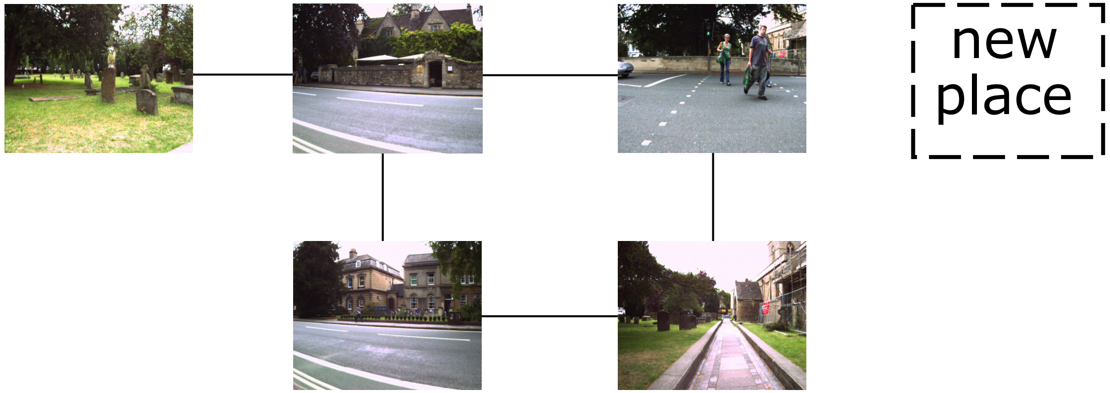


Bag-of-Words Representation



Inference in FAB-MAP

map:



current observation:

Word 1 Word 2 Word 3 Word 4 Word 5

Z = [0 1 0 1 1 …]

$$Z_k = \{z_1, \ldots, z_{|v|}\}$$

observation at time *k*,
|*v*| = number of words in dictionary

Environment Representation

- Collection of a set of discrete and disjoint locations at time *k*:

$$\mathcal{L}^k = \{L_1, \ldots, L_{n_k}\}$$

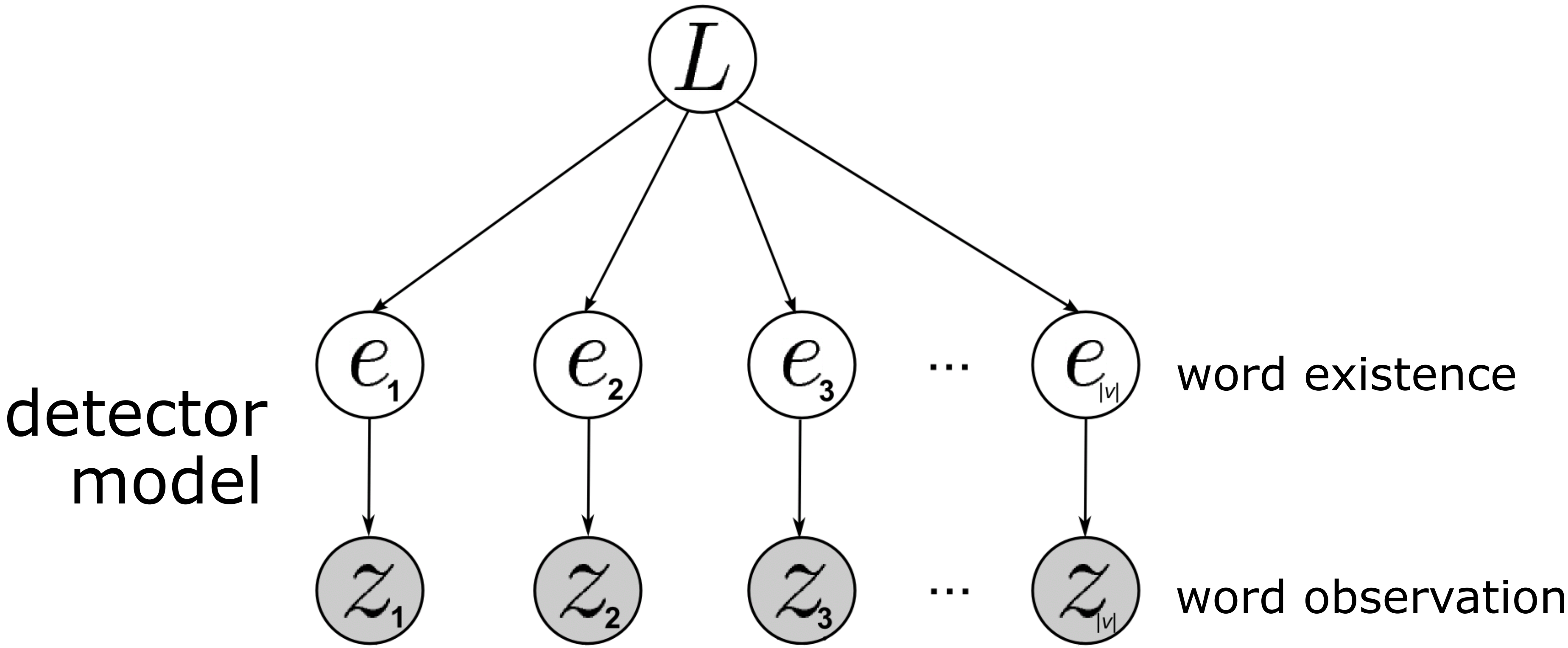
- Place appearance model: belief about the existence of scene elements (words)

$$\{p(e_1 = 1|L_i), \ldots, p(e_{|v|} = 1|L_i)\}$$

- Detector model relates feature existence and feature detection

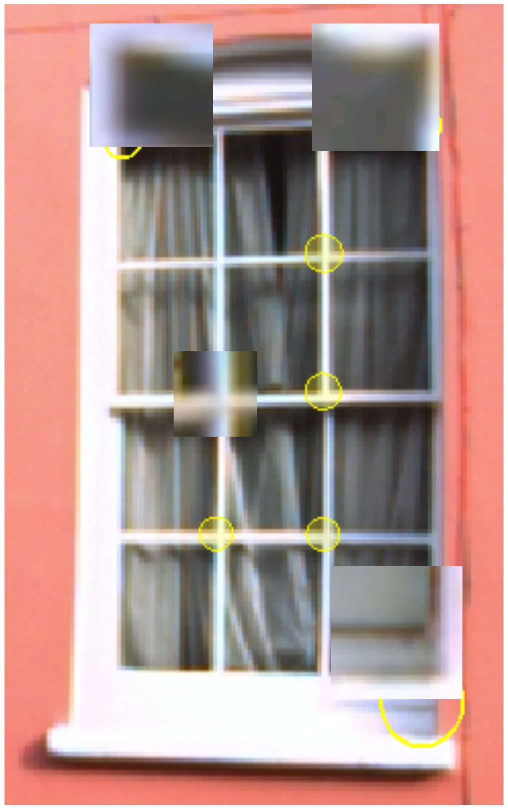
$$\mathcal{D}: \quad \begin{cases} p(\underbrace{z_i}_{\text{observation}} = 1|e_i = 0), & \text{false positive probability.} \\ p(z_i = 0|\underbrace{e_i}_{\text{existence}} = 1), & \text{false negative probability.} \end{cases}$$

Graphical Model



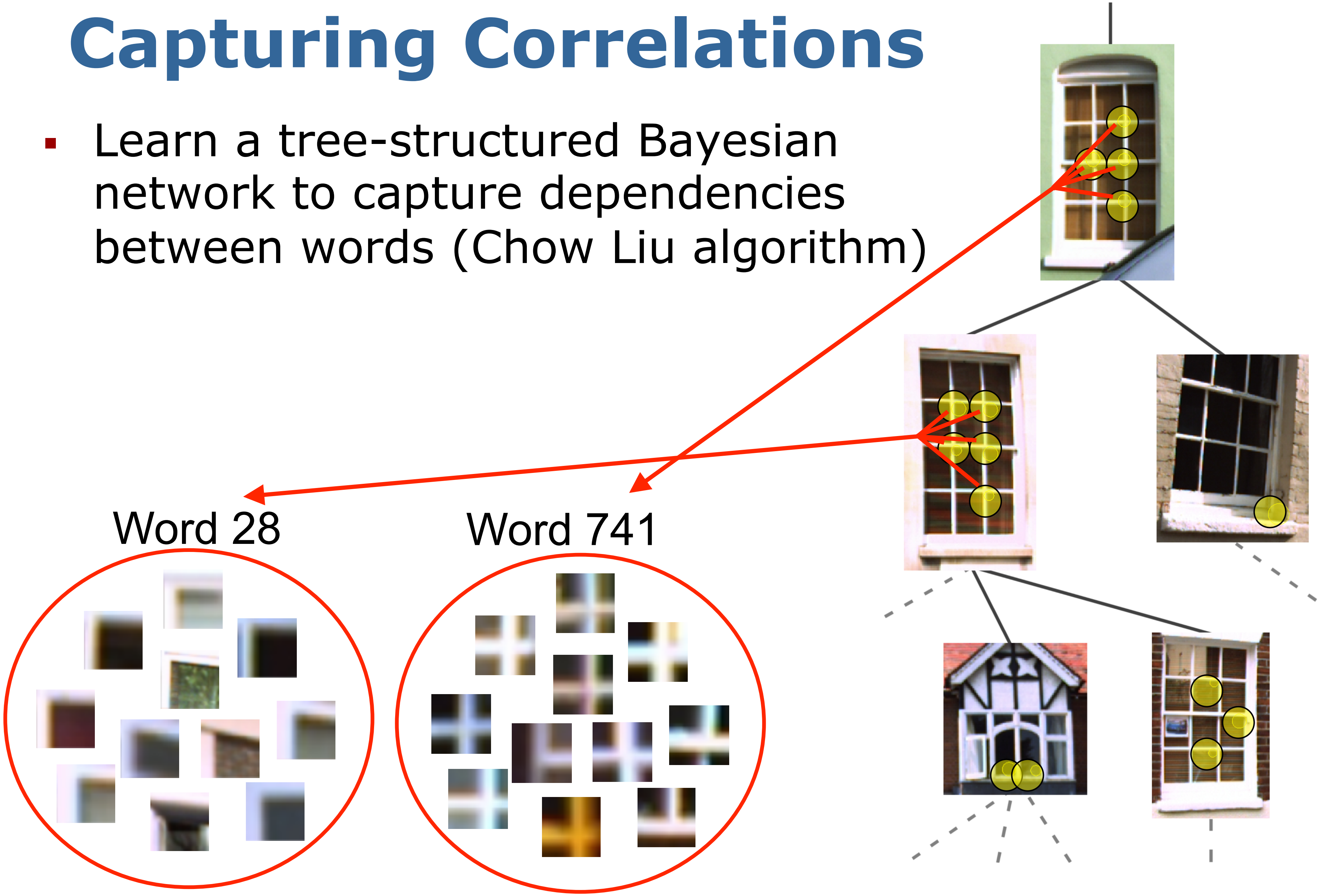
Correlations of Word Occurrence

- Visual words are not independent, instead they tend to co-occur



Capturing Correlations

- Learn a tree-structured Bayesian network to capture dependencies between words (Chow Liu algorithm)



Capturing Correlations

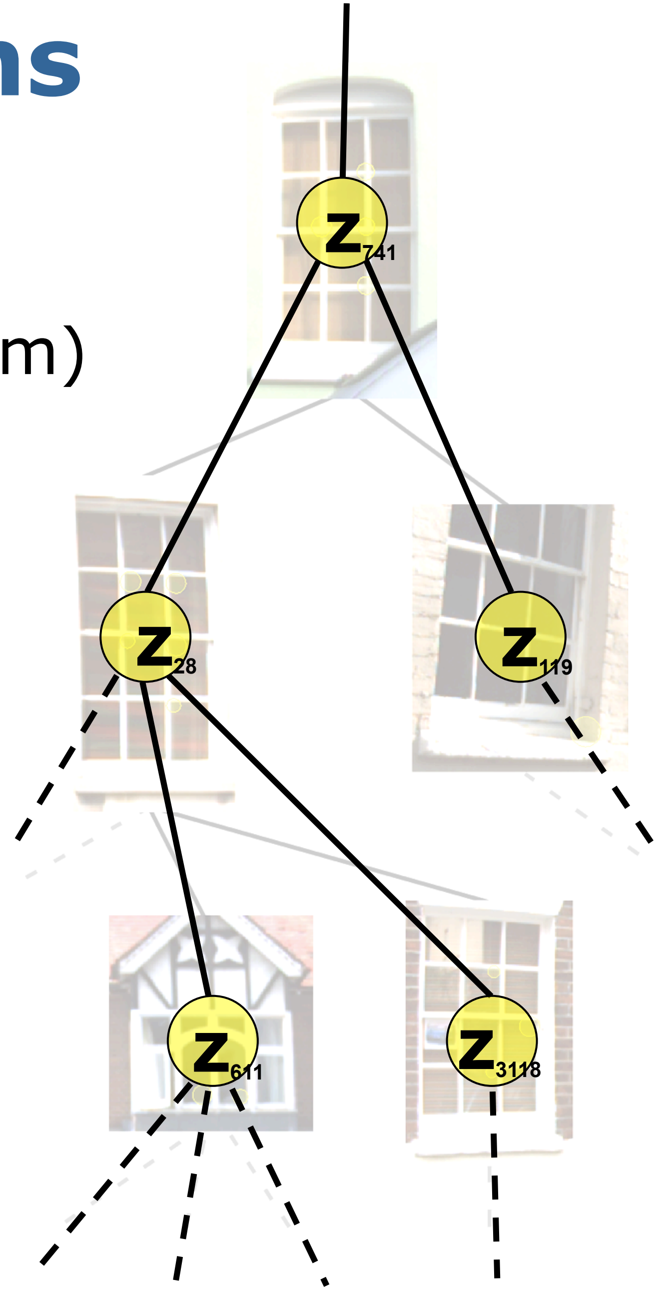
- Learn a tree-structured Bayesian network to capture dependencies between words (Chow Liu algorithm)

$$Z = \{z_1, \ldots, z_N\}$$

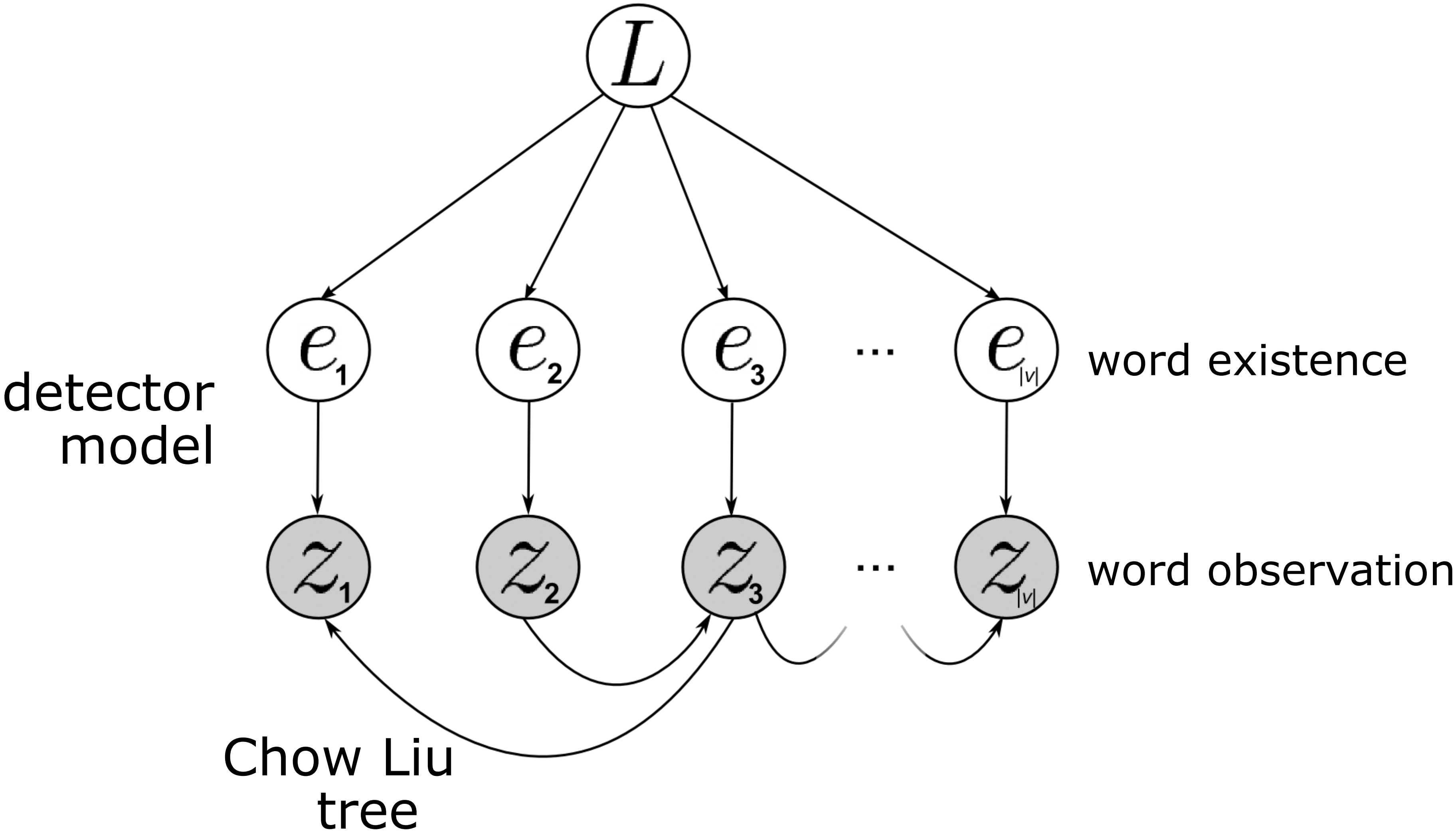
$$p(Z) = p(z_r) \prod_{i=1}^{N} p(z_i | z_{p_i})$$

root

parent of z_i



Graphical Model



Inference in FAB-MAP

all observations up to k

observation likelihood

prior

$$p(L_i|\mathcal{Z}^k) = \frac{p(Z_k|L_i)p(L_i|\mathcal{Z}^{k-1})}{p(Z_k|\mathcal{Z}^{k-1})}$$

location i

normalizing term

Observation Likelihood

- Chow Liu tree for the joint distribution

$$p(Z_k|L_i) = p(z_r|L_i) \prod_{q=2}^{|v|} p(z_q|z_{p_q}, L_i)$$

Observation Likelihood

- Chow Liu tree for the joint distribution

$$p(Z_k|L_i) = p(z_r|L_i) \prod_{q=2}^{|v|} p(z_q|z_{p_q}, L_i)$$

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

Observation Likelihood

- Chow Liu tree for the joint distribution

$$p(Z_k|L_i) = p(z_r|L_i) \prod_{q=2}^{|v|} p(z_q|z_{p_q}, L_i)$$

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

can be further expanded and estimated from training data

appearance model (updated online)

Location Prior

- Use a simple motion model to compute
$$p(L_i|\mathcal{Z}^{k-1})$$
- If the vehicle is at location *i* at time *k*-1, it is likely to be at one of the topologically adjacent locations at time *t*
- In case of unknown neighbors, part of the probability mass is assigned to a “new place” node (no odometry is used)

Normalization

all observations up to k

observation likelihood

prior

$$p(L_i|\mathcal{Z}^k) = \frac{p(Z_k|L_i)p(L_i|\mathcal{Z}^{k-1})}{p(Z_k|\mathcal{Z}^{k-1})}$$

location i

normalizing term

- We need to evaluate the normalizing term since the current observation might come from a location not yet contained in the map

$$p(Z_k|\mathcal{Z}^{k-1}) = \sum_{all\ L} p(Z_k|L)p(L|\mathcal{Z}^{k-1})$$

$$p(Z_k|\mathcal{Z}^{k-1}) \quad = \quad \sum_{m \in M} p(Z_k|L_m)p(L_m|\mathcal{Z}^{k-1}) + \sum_{n \in \overline{M}} p(Z_k|L_n)p(L_n|\mathcal{Z}^{k-1})$$

mapped places

unmapped places

$$p(Z_k|\mathcal{Z}^{k-1}) = \sum_{all\ L} p(Z_k|L)p(L|\mathcal{Z}^{k-1})$$

$$p(Z_k|\mathcal{Z}^{k-1}) \quad = \quad \sum_{m \in M} p(Z_k|L_m)p(L_m|\mathcal{Z}^{k-1}) + \sum_{n \in \overline{M}} p(Z_k|L_n)p(L_n|\mathcal{Z}^{k-1})$$

mapped places

unmapped places

approximate by sampling:

$$p(Z_k|\mathcal{Z}^{k-1}) \approx \sum_{m \in M} p(Z_k|L_m)p(L_m|\mathcal{Z}^{k-1}) + p(L_{new}|\mathcal{Z}^{k-1}) \sum_{u=1}^{n_s} \frac{p(Z_k|L_u)}{n_s}$$

prior probability
of being at a new location

sampled place models

Updating Place Models

- Maximum likelihood data association after each observation
- Update the relevant place appearance model

$$\{p(e_1 = 1|L_i), \ldots, p(e_{|v|} = 1|L_i)\}$$

- Each component is updated according to

prior

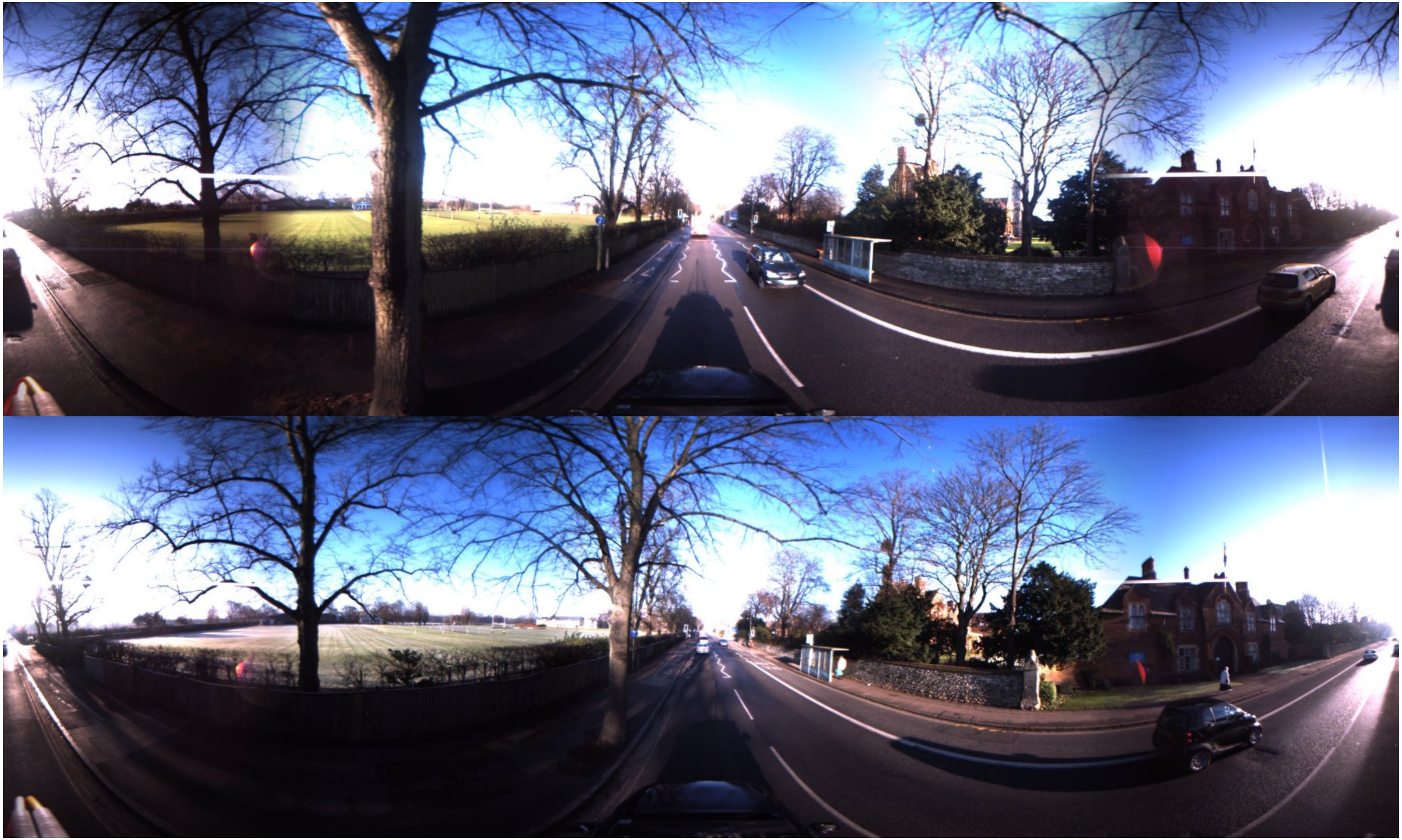
$$p(e_j = 1|L_j, \mathcal{Z}^k) = \frac{p(Z_k|e_i{=}1)p(e_i{=}1|L_j, \mathcal{Z}^{k-1})}{p(Z_k|L_j)}$$

Bayes' rule + two assumption:
observations independent given place
detection errors independent of location

Experimental Results

- 2k images, collected 30m apart, for training (vocabulary + Chow Liu tree)
- Vocabulary: 100k words
- 1000 km test data set: 103k images, ~8m apart, with 50k loop closures, 21h driving
- Robust matching even when place appearance changes
- Correct loop closures under perspective changes, rotation, lighting changes, dynamic objects, ...

Perspective Change



Rotation

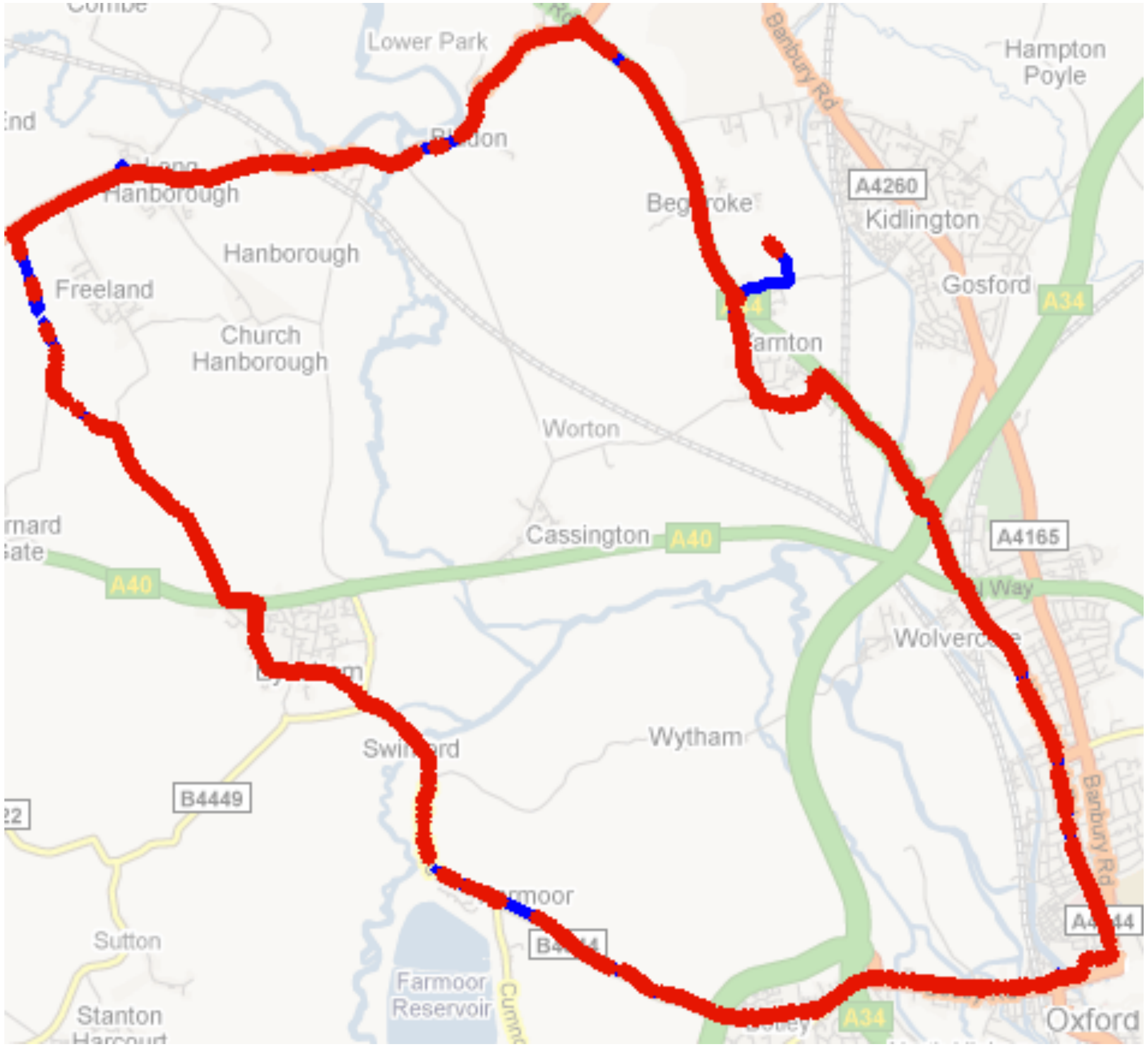
Lighting Change

Dynamic Objects

Perceptual Aliasing Correctly Rejected

Highest Confidence False Positives

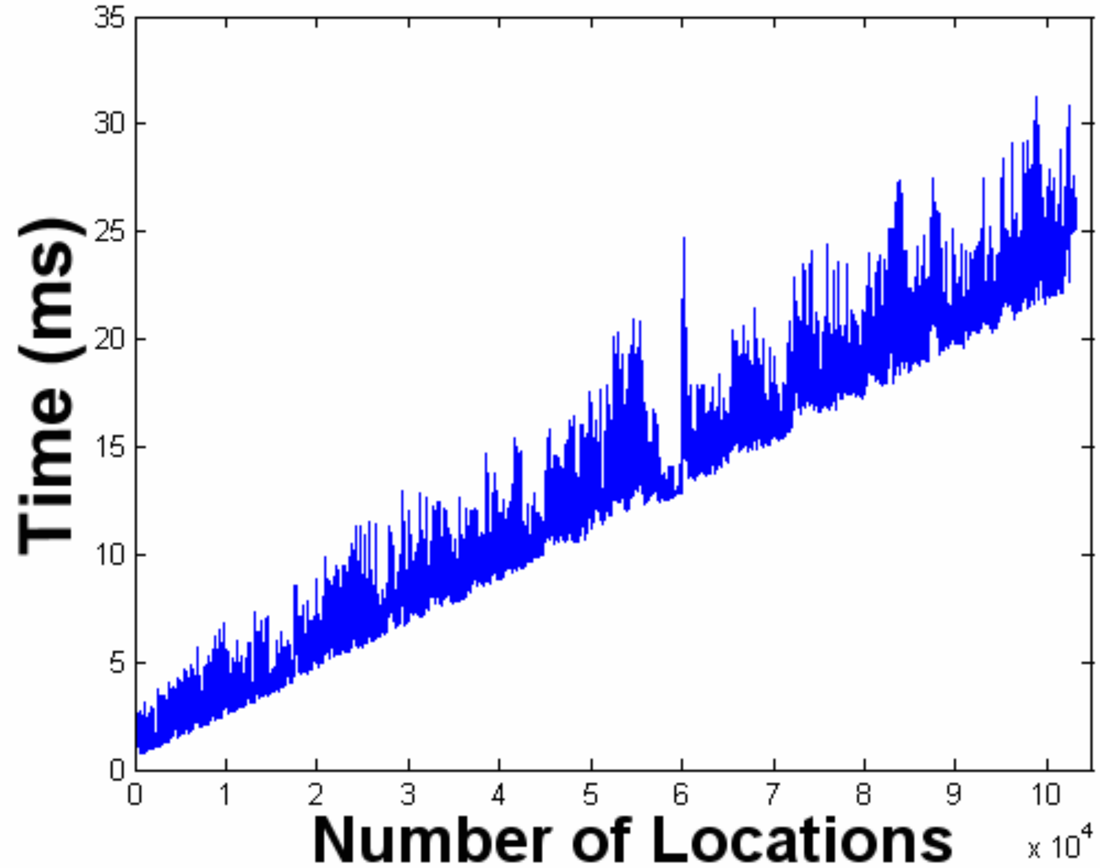
Loop Closure (70 km Data Set)



Trajectory from GPS data


Video
LS07 YJT
EA58 ZGF
Data Collection Platform

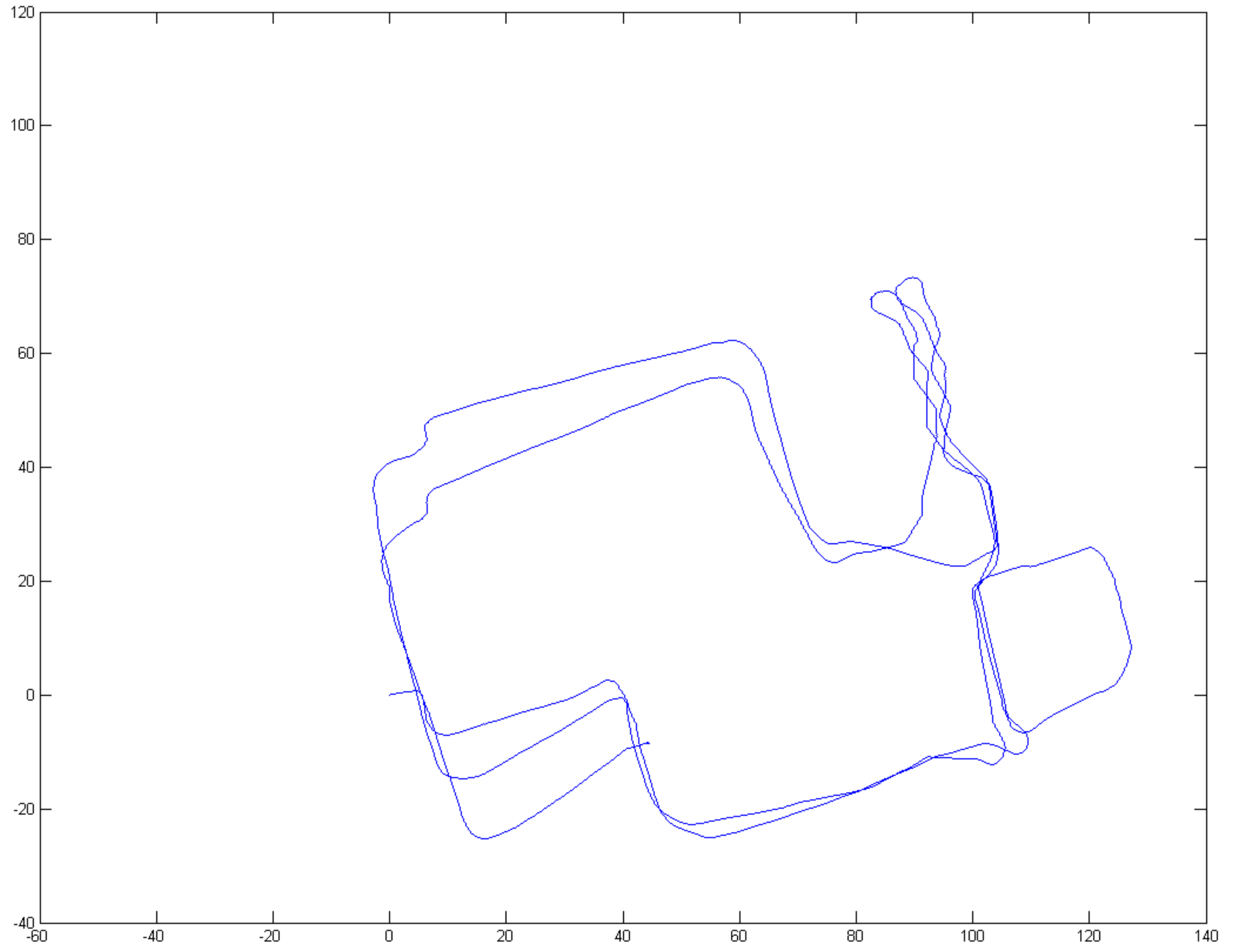
Timing Performance



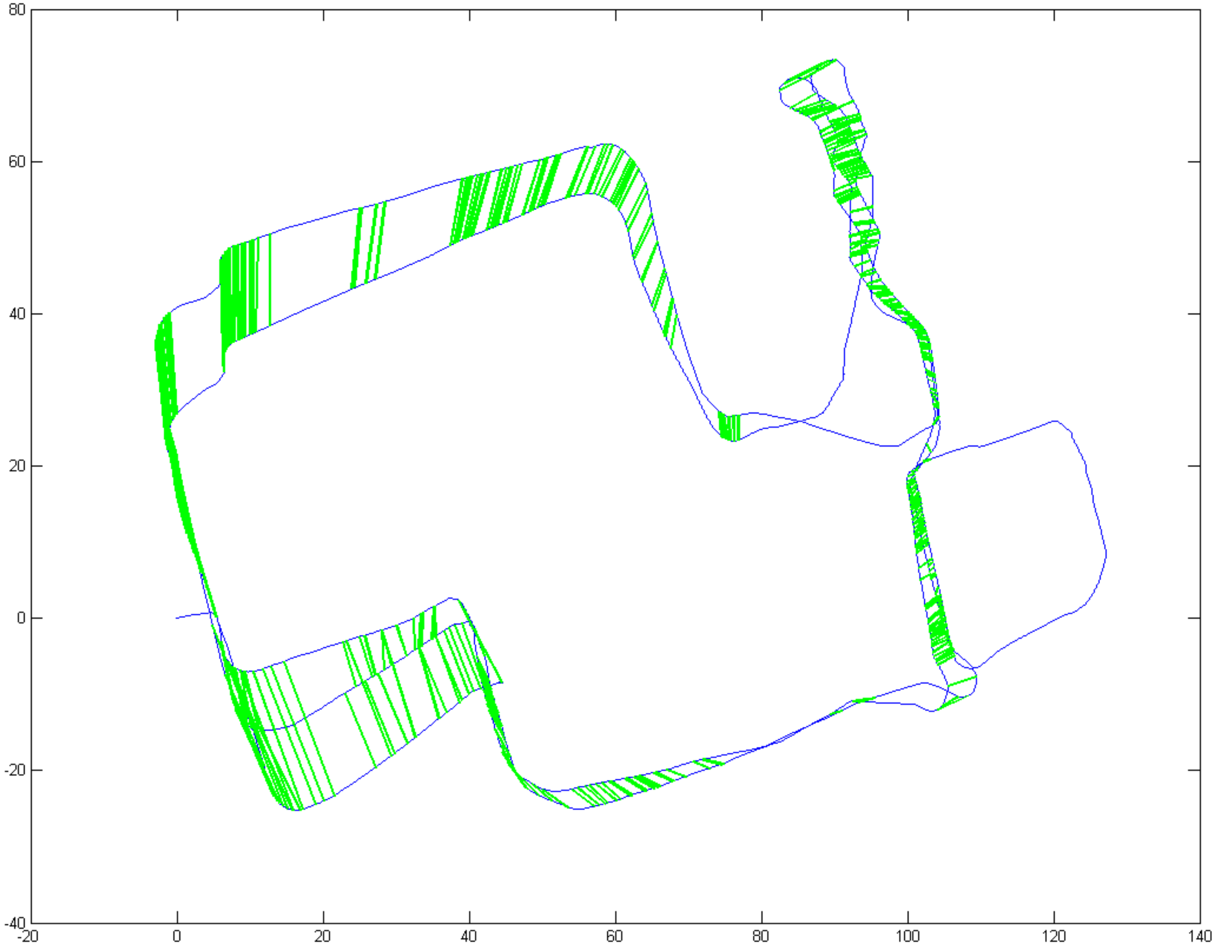
Mean computation times:

- Inference: 25 ms for 100k locations
- SURF detection + quantization: 483 ms

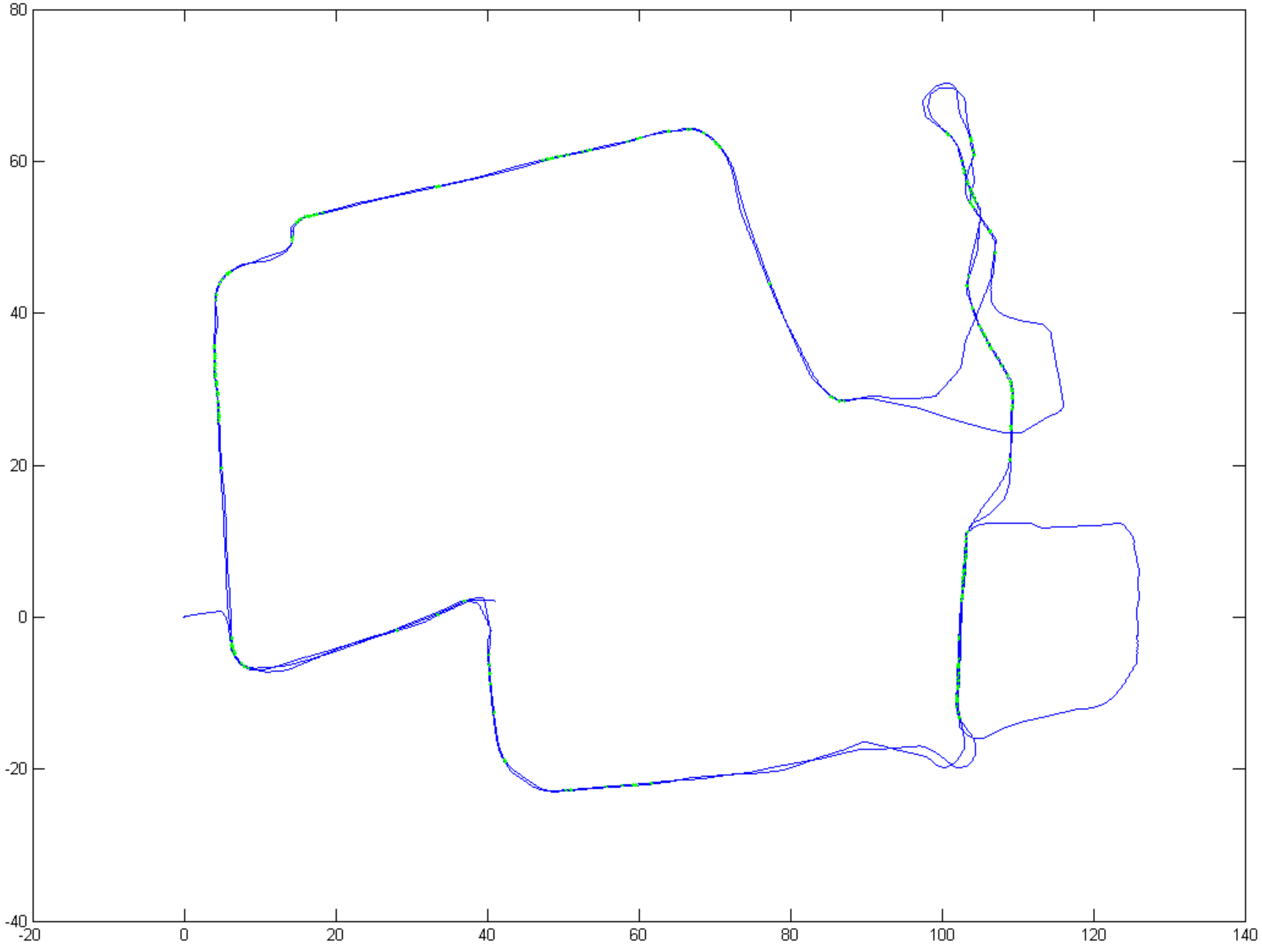
Visual Odometry



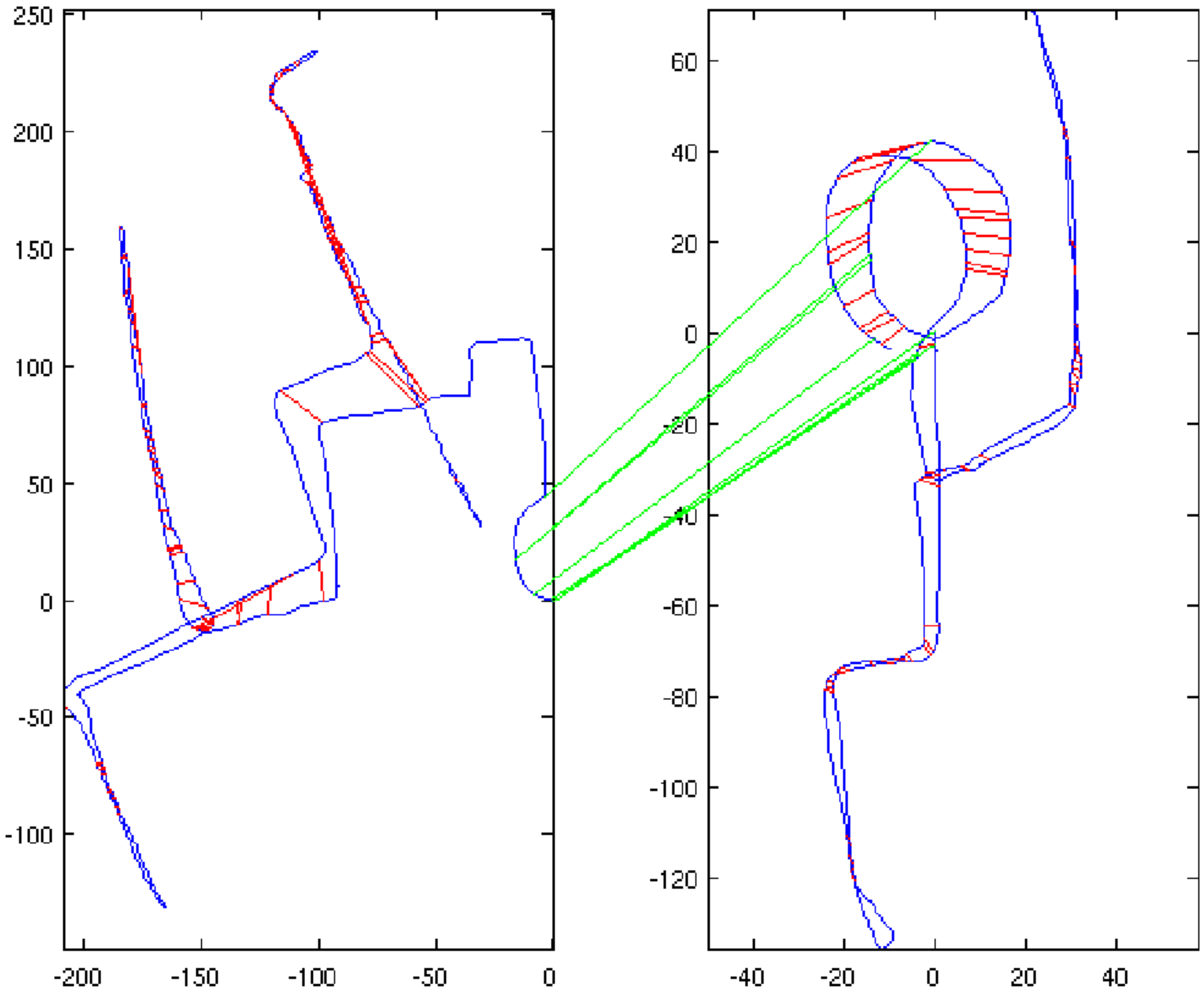
Visual Odometry with Loop Closure Constraints



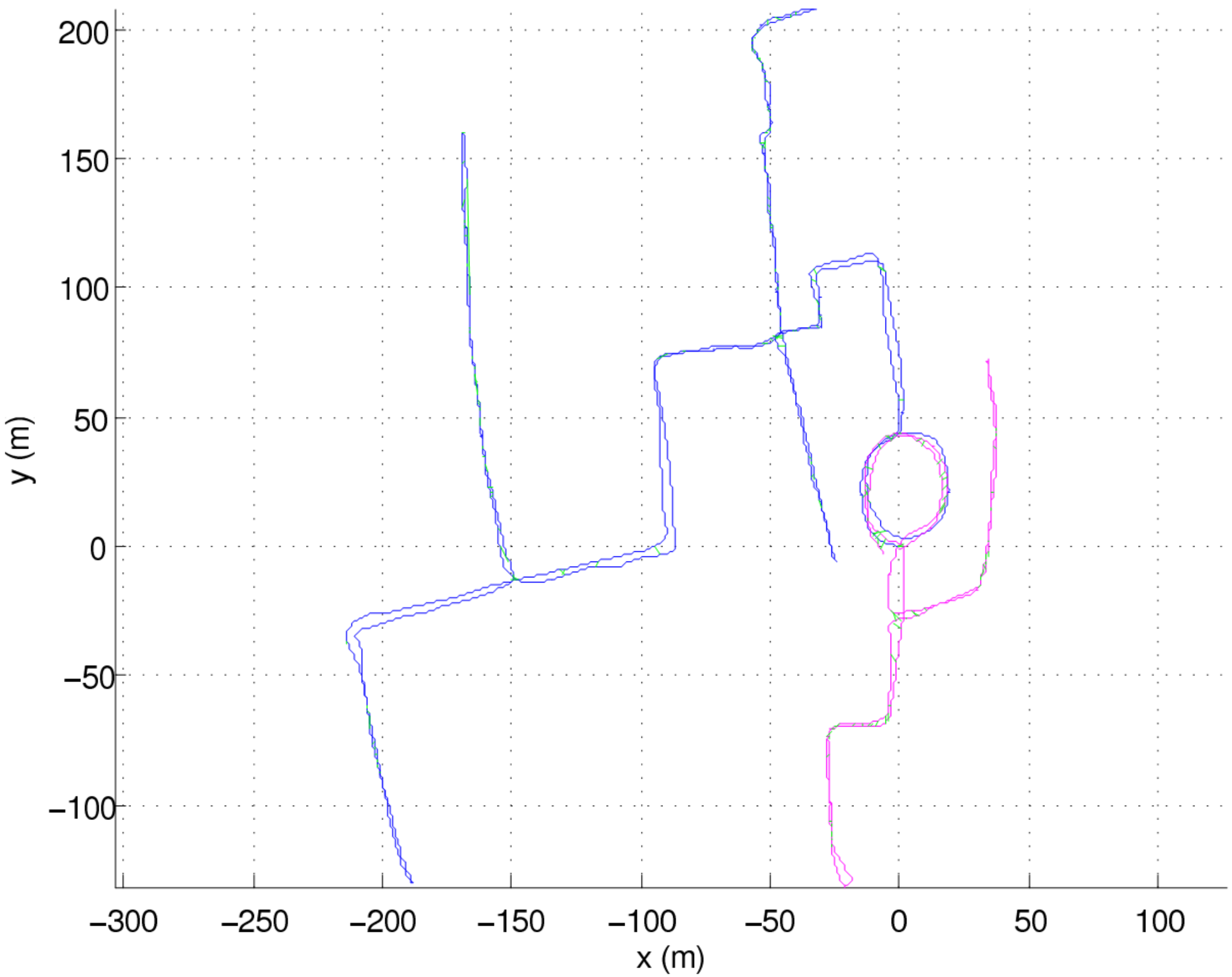
Combined Result



Multi-Session Mapping



Multi-Session Mapping



Summary

- Appearance-only navigation
- Bag-of-words approach to recognize places
- Chow Liu tree to capture dependencies
- Probabilistic framework can deal with perceptual aliasing and new place detection
- Successfully detects loops in challenging outdoor environments
- Fast enough for online loop closure detection
- Can be used to complement metric SLAM

Further Reading

- M. Cummins & P. Newman

  FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance
  Int. Journal of Robotics Research, 2008

  Appearance-only SLAM at Large Scale with FAB-MAP 2.0
  Int. Journal of Robotics Research, 2010