

# FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance

Presented by  
Zifei (David) Zhong

11/13/2023

authored by Mark Cummins and Paul Newman;  
appeared in the International Journal of Robotics Research, 2008

# Introduction

## Fast Appearance-Based Mapping (FAB-MAP)

- Problem: Recognizing locations based on their appearance.
- Solution: Learning a generative model for the appearance to achieve:
  1. Compute the similarity of two observations.
  2. Compute the probability that the two observations originate from the same location.
- Superiorities:
  1. Address perceptual aliasing (less false positives)
  2. Improve inference reasoning (less false negatives)
  3. Accommodate new locations
  4. Linear-time complexity

# Chow Liu Tree

## Approximating High Dimensional Discrete Distributions

- The Chow Liu algorithm approximates a discrete distribution  $P(Z)$  by the closest tree-structured Bayesian network  $Q(Z)_{opt}$ , in the sense of minimizing the Kullback-Leibler divergence.
- For a distribution over  $n$  variables, a mutual information graph  $\mathcal{G}$  is the complete graph with  $n$  nodes and  $\frac{n(n-1)}{2}$  edges, where each edge  $(z_i, z_j)$  has weight equal to the mutual information  $I(z_i, z_j)$  between variable  $i$  and  $j$ :

$$I(z_i, z_j) = \sum_{z_i \in \Omega, z_j \in \Omega} p(z_i, z_j) \log \frac{p(z_i, z_j)}{p(z_i)p(z_j)}$$

# Chow Liu Tree

## Advantages & Illustration

- The maximum-weight spanning tree of the mutual information graph  $\mathcal{G}$  will have the same structure as  $Q(Z)_{opt}$ .
- The Chow Liu algorithm guarantees the optimal approximation, and requires only first order conditional probabilities.

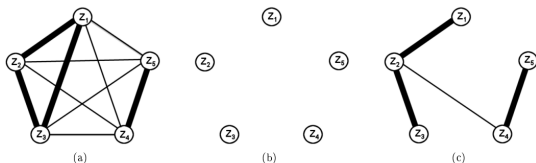


Figure 1: (a) Graphical model of the underlying distribution  $P(Z)$ . Mutual information between variables is shown by the thickness of the edge. (b) Naive Bayes approximation. (c) Chow Liu tree.

# Probabilistic Navigation using Appearance

## Overview

- The world is modeled as a set of discrete locations, each location being described by a distribution over “appearance words”.
- Incoming sensory data is converted into a bag-of-words representation.
- For each location  $L_i$ , we ask how likely it is that the observation comes from  $L_i$ 's distribution.
- Find the probability that the observation comes from a location not in the map, and update the map if a location found.

# Probabilistic Navigation using Appearance

## Representing Appearance

- “bag-of-words” representation for raw sensor data
  1. A scene is represented as a collection of attributes (words) chosen from a set (vocabulary) of size  $|V|$ .
- Observation  $Z_k$  of local scene appearance at time  $k$ :
  1.  $Z_k = \{z_1, \dots, z_{|V|}\}$
  2.  $z_i$  is a binary variable indicating the presence/absence of the  $i$ th word of the vocabulary.

# Probabilistic Navigation using Appearance

## Representing Locations

- Map of environment at time  $k$  is a collection of  $n_k$  discrete and disjoint locations  $\mathcal{L}^k = \{L_1, \dots, L_{n_k}\}$ .
- Hidden variable  $e_i$ : an event that an object which generates observations of type  $z_i$  exists.
- Location  $L_i$ 's model: a set  $\{p(e_1 = 1|L_i), \dots, p(e_{|V|} = 1|L_i)\}$ , where each  $e_i$  is generated independently by the location.
- Detector  $\mathcal{D}$ :

$$\mathcal{D} : \begin{cases} p(z_i = 1|e_i = 0), \text{ false positive probability} \\ p(z_i = 0|e_i = 1), \text{ false negative probability} \end{cases}$$

# Probabilistic Navigation using Appearance

## Representing Locations: illustration

- Factoring  $p(Z|L_i)$  into two parts:
  1. (Learn online) A simple model that  $e_i$  only depends on  $L_i$ .
  2. (Learn offline) A complex model that captures the correlations between detections of appearance words  $p(z_i|Z_k)$ .

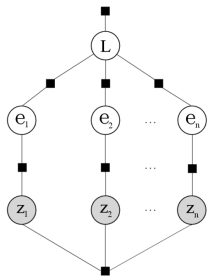


Figure 2: Factor graph of our generative model. Observed variables are shaded, latent variables unshaded. The environment generates locations  $L_i$ . Locations independently generate words  $e_j$ . Words generate observations  $z_k$ , which are interdependent.



# Probabilistic Navigation using Appearance

## Estimating Location via Recursive Bayes

- Calculating  $p(L_i|\mathcal{Z}^k)$ :

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

1.  $\mathcal{Z}^k$ : the set of all observations up to time  $k$
2.  $p(L_i|\mathcal{Z}^{k-1})$ : prior belief about our location
3.  $p(Z_k|L_i, \mathcal{Z}^{k-1})$ : observation likelihood, equal to  $p(Z_k|L_i)$
4.  $p(Z_k|\mathcal{Z}^{k-1})$ : normalization term

# Probabilistic Navigation using Appearance

## Observation Likelihood

- Naive Bayes assumption:

$$p(Z_k|L_i) \approx p(z_n|L_i) \cdots p(z_2|L_i)p(z_1|L_i),$$

and

$$\begin{aligned} p(z_j|L_i) &= \sum_{s \in \{0,1\}} p(z_j|e_j = s, L_i)p(e_j = s|L_i) \\ &= \sum_{s \in \{0,1\}} p(z_j|e_j = s)p(e_j = s|L_i) \end{aligned}$$

# Probabilistic Navigation using Appearance

## Observation Likelihood

- Chow Liu assumption:

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

where  $z_r$  is the root of the tree, and  $z_{p_q}$  is the parent of  $z_q$  in the tree. With further expansion,

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

$$p(z_q|e_q, z_{p_q}) \approx \left(1 + \frac{\alpha}{\beta}\right)^{-1},$$

# Probabilistic Navigation using Appearance

## Observation Likelihood

where

$$\alpha = p(z_q = s_{z_q})p(z_q = \bar{s}_{z_q} | e_q = s_{e_q})p(z_q = \bar{s}_{z_q} | z_p = s_{z_p}),$$

$$\beta = p(z_q = \bar{s}_{z_q})p(z_q = s_{z_q} | e_q = s_{e_q})p(z_q = s_{z_q} | z_p = s_{z_p}),$$

Now,  $\alpha$  and  $\beta$  are expressed entirely in terms of quantities that can be estimated from training data.

# Probabilistic Navigation using Appearance

## Discovery of New Places

- To deal with the possibility that a new observation comes from a previously unknown location, calculation of  $p(Z_k|\mathcal{Z}^{k-1})$  is required.
- Dividing the world into two sets: mapped places  $M$ , and unmapped places  $\bar{M}$ , then

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

where  $n_s$  is the number of samples, and  $p(L_{new}|\mathcal{Z}^{k-1})$  is the prior probability of being at a new place.

# Probabilistic Navigation using Appearance

## Location Prior & Smoothing

- With sequentially collected observations, if the robot is at place  $i$  at time  $t$ , it has equal probability of being at one of the places  $\{i - 1, i, i + 1\}$  at time  $t + 1$ ; otherwise assume uniform prior.
- Smoothing the likelihood estimation:

$$p(Z_k | L_i) \rightarrow \sigma p(Z_k | L_i) + \frac{1 - \sigma}{n_k},$$

where  $n_k$  is the number of places in the map,  $\sigma$  is the smoothing parameter (0.99 in experiments).

# Probabilistic Navigation using Appearance

## Updating Place Models

- When a new place is created, its appearance model is initialized so that all words exist with marginal probability  $p(e_i = 1)$  derived from the training data.
- Given an observation that relates to the new place, each component of the appearance model can be updated by

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

# Probabilistic Navigation using Appearance

## Input Parameters

- Detector model,  $p(z_i = 1|e_i = 0)$  and  $p(z_i = 0|e_i = 1)$ .
- Smoothing parameter  $\sigma$ .
- Prior probability that a topological link with an unknown endpoint leads to a new place.



# Evaluation

## Building the Vocabulary Model

- Use the SURF detector/description to extract region of interest from images, and map the 128D descriptors visual words.
- Construct Chow Liu tree.
  1. Each node in the graph corresponds to a visual word.
  2. Compute the maximum-weight spanning tree of the graph.
  3. 2800 images from 28km of urban streets environment.
  4. 11k visual words in the constructed Chow Liu tree.

# Sample Vocabulary



Figure 3: A sample word in the vocabulary, showing typical image patches and an example of the interest points in context. Interest points quantized to this word typically correspond to the top-left corner of windows. The most correlated word in the vocabulary is shown in Figure 4.



Figure 4: A sample word in the vocabulary, showing typical image patches and an example of the interest points in context. Interest points quantized to this word typically correspond to the cross-piece of windows. The most correlated word in the vocabulary is shown in Figure 3.

# Sample Chow Liu Tree

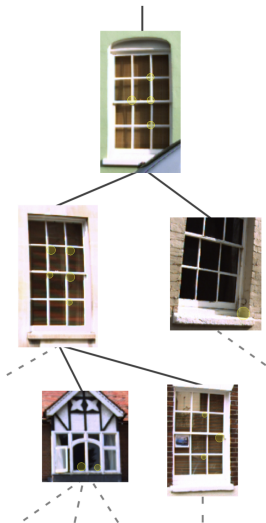


Figure 5: Visualization of a section the Chow Liu tree computed for our urban vocabulary. Each word in the tree is represented by a typical example. Clockwise from top, the words correspond to the cross-pieces of window panes, right corners of window sills, top-right corners of window panes, bottom-right corners of window panes and top-left corners of window panes. Under the Chow Liu model the joint probability of observing these words together is 4,778 times higher than under the a Naive Bayes assumption.

# Appear-based Matching (City Centre dataset)

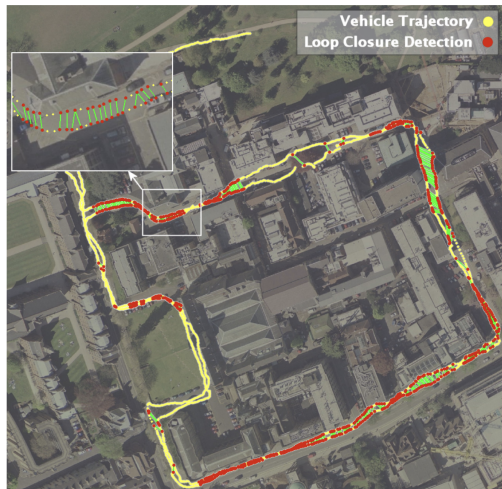


Figure 6: Appearance-based matching results for the City Centre dataset overlaid on an aerial photograph. The robot travels twice around a loop with total path length 2km, collecting 2,474 images. Positions (from hand-corrected GPS) at which the robot collected an image are marked with a yellow dot. Two images that were assigned a probability  $p \geq 0.99$  of having come from the same location (on the basis of appearance alone) are marked in red and joined with a line. There are no incorrect matches that meet this probability threshold. This result is best viewed as a video (Extension 2).

# Appear-based Matching (New College dataset)



Figure 7: Appearance-based matching results for the New College dataset overlaid on an aerial photograph. The robot traverses a complex trajectory of 1.9km with multiple loop closures. 2,146 images were collected. Positions (from hand-corrected GPS) at which the robot collected an image are marked with a yellow dot. Two images that were assigned a probability  $p \geq 0.99$  of having come from the same location (on the basis of appearance alone) are marked in red and joined with a line. There are no incorrect matches that meet this probability threshold. This result is best viewed as a video (Extension 1).

# Precision Recall

- At 100% precision, the system achieves 48% recall on the New College dataset, and 37% on the City Center dataset.
- Typically 37% recall rate is sufficient to detect almost all loop closure.

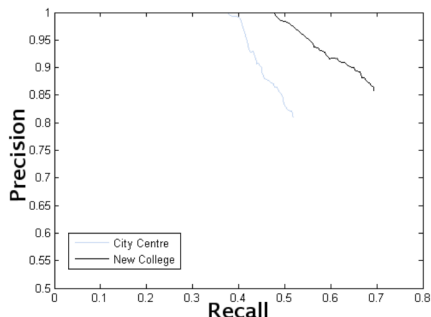


Figure 8: Precision-Recall curves for the City Centre and New College datasets. Notice the scale.

# Samples from Results: similar scenes, different locations

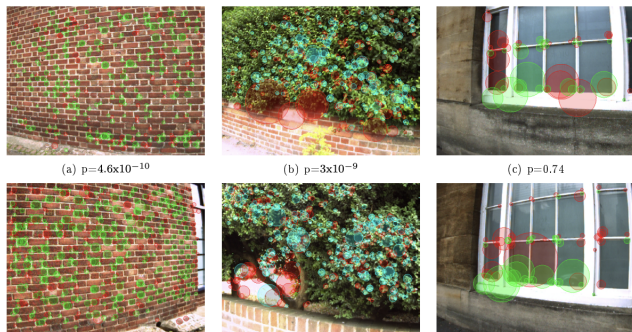


Figure 9: Some examples of remarkably similar-looking images from different parts of the workspace that were correctly assigned low probability of having come from the same place. The result is possible because most of the probability mass is captured by locations in the sampling set – effectively the system has learned that images like these are common. Of course, had these examples been genuine loop closures they might also have received low probability. We would argue that this is correct behaviour, modulo the fact that the probabilities in (a) and (b) are too low. The very low probabilities in (a) and (b) are due to the fact that good matches for the query images are found in the sampling set, capturing almost all the probability mass. This is less likely in the case of a true but ambiguous loop closure. Words common to both images are shown in green, others in red. (Common words are shown in blue in (b) for better contrast). The probability that the two images come from the same place is indicated between the pairs.

## Samples from Results: different scenes, same location

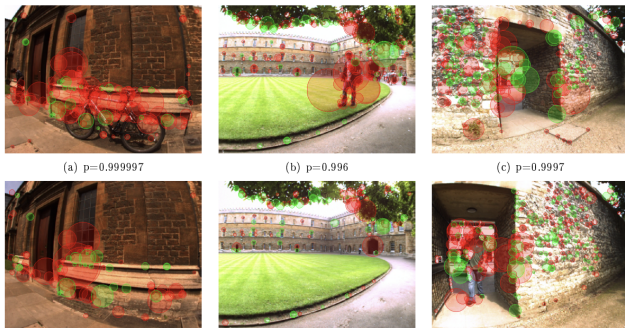


Figure 10: Some examples of images that were assigned high probability of having come from the same place, despite scene change. Words common to both images are shown in green, others in red. The probability that the two images come from the same place is indicated between the pairs.



# Comparison of Different Approximations

Algorithm	Recall - New College	Recall - City Centre	Run Time
Mean Field, Naive Bayes	34%	16%	0.6ms/place
Mean Field, Chow Liu	35%	31%	1.1ms/place
Monte Carlo, Naive Bayes	40%	31%	0.6ms/place + 1.71 secs sampling
Monte Carlo, Chow Liu	48%	37%	1.1ms/place + 3.15 secs sampling

Table 1: Comparison of the four different approximations. The recall rates quoted are at 100% precision. The time to process a new observation is given as a function of the number of places already in the map, plus a fixed cost to perform the sampling. Timing results are for a 3GHZ Pentium IV.

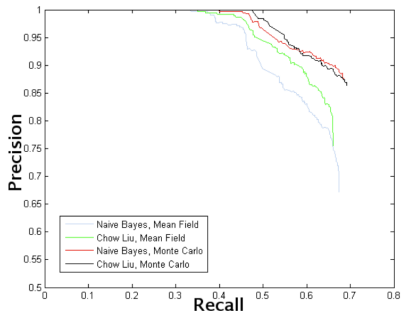


Figure 11: Precision-Recall curves for the four variant algorithms on the New College datasets. Notice the scale. Relative performance on the City Centre dataset is comparable.

# Chow Liu Approximation vs. Naive Bayes

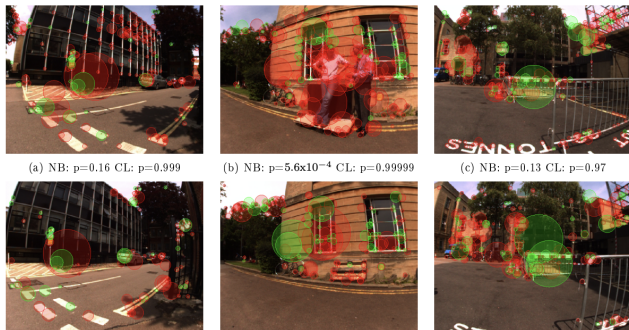


Figure 12: Some examples of situations where the Chow Liu approximation outperforms Naive Bayes. In (a), a change in lighting means that the feature detector does not fire on the windows of the building. In (b), the people are no longer present. In (c), the foreground text and the scaffolding in the top right are not present in the second image. In each of these cases, the missing features are known to be correlated by the Chow Liu approximation, hence the more accurate probability. Words common to both images are shown in green, others in red. The probability that the two images come from the same place (according to both the Chow Liu and Naive Bayes models) is indicated between the pairs.

# False Positives Possibility & Discussion

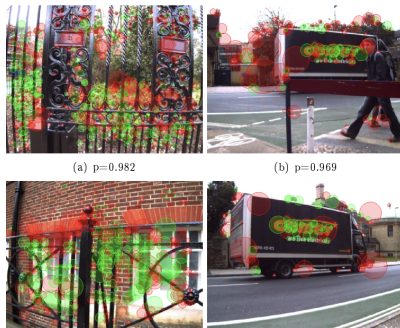




Figure 13: Some images from different locations incorrectly assigned high probability of having come from the same place. In (a), the training set contains no examples of railings, so the matched features are not known to be correlated. In (b), we encounter the same truck again in a different part of the workspace. Errors of this type are particularly challenging. Notice that while both images are assigned high probability of a match, a typical true loop closure is assigned much higher probability. Neither of these image pairs met our  $p = 0.99$  data association threshold.

# References

-  Probabilistic Appearance Based Navigation and Loop Closing.  
Mark Cummins, Paul Newman.  
*Proc. of IEEE International Conference on Robotics and Automation*,  
2007.
-  Video Google: A Text Retrieval Approach to Object Matching in  
Videos.  
Josef Sivic, Andrew Zisserman  
*Proc. of IEEE International Conference on Computer Vision*, 2003.