Advanced Techniques for Mobile Robotics

Bag-of-Words Models & Appearance-Based Mapping

Wolfram Burgard, Cyrill Stachniss,
Kai Arras, Maren Bennewitz
Motivation: Analogy to Documents

Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach the brain from our eyes. For a long time it was thought that the retinal image was transmitted point by point to visual centers in the brain; the cerebral cortex was a movie screen, so to speak, upon which the image in the eye was projected. Through the discoveries of Hubel and Wiesel we now know that behind the origin of the visual perception in the brain there is a considerably more complicated course of events. By following the visual impulses along their path to the various cell layers of the optical cortex, Hubel and Wiesel have been able to demonstrate that the message about the image falling on the retina undergoes a step-wise analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.

image source: L. Fei-Fei

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% jump in exports to $750bn, compared with a 18% rise in imports to $660bn. The figures are likely to further annoy the US, which has long argued that China's exports are unfairly helped by a deliberately undervalued yuan. Beijing agrees the surplus is too high, but says the yuan is only one factor. Bank of China governor Zhou Xiaochuan said the country also needed to do more to boost domestic demand so more goods stayed within the country. China increased the value of the yuan against the dollar by 2.1% in July and permitted it to trade within a narrow band, but the US wants the yuan be allowed to trade freely. However, Beijing has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.
Object Classification / Scene Recognition

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

image source: L. Fei-Fei
Bag of Visual Words

- Visual words = independent features
Bag of Visual Words

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

slide adapted from: L. Fei-Fei example patch
Feature Detection and Representation

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

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

example patch
Learning the Dictionary

clustering, e.g., k-means

cluster center = code words

slide adapted from: L. Fei-Fei
Example Codewords Dictionary
Example Image Representation

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

![Histogram](image.png)

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
  
  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:
Learning the Visual Vocabulary

feature extraction

SURF
Clustering in Feature Space
Bag-of-Words Representation

feature detection

compute descriptor vectors

quantize

Word 753
Inference in FAB-MAP

map:

new place

current observation:

\[ Z = [0 \ 1 \ 0 \ 1 \ 1 \ldots ] \]

\[ Z_k = \{ z_1, \ldots, z_{|v|} \} \]

observation at time \( k \),

\(|v| = \text{number of words in dictionary}\)
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
  \[ \mathcal{D} : \begin{cases} 
  p(z_i = 1|e_i = 0), & \text{false positive probability.} \\
  p(z_i = 0|e_i = 1), & \text{false negative probability.} 
  \end{cases} \]
Graphical Model

detector model

\[ e_1 \to \tilde{z}_1 \]
\[ e_2 \to \tilde{z}_2 \]
\[ e_3 \to \tilde{z}_3 \]
\[ \ldots \]
\[ e_{|L|} \to \tilde{z}_{|L|} \]

word existence

word observation
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)
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

Detector model

Chow Liu tree

word existence

word observation
Inference in FAB-MAP

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

- **all observations up to k**
- **location i**
- **observation likelihood**
- **prior**
- **normalizing term**
Observation Likelihood

- Chow Liu tree for the joint distribution

\[
p(Z_k|L_i) = p(z_r|L_i) \prod_{q=2}^{|v|} p(z_q|z_{pq}, L_i)
\]
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

\[
p(Z_k | L_i) = p(z_r | L_i) \prod_{q=2}^{v} p(z_q | z_{pq}, L_i)
\]

\[
p(z_q | z_{pq}, L_i) = \sum_{s_{eq} \in \{0,1\}} p(z_q | e_q = s_{eq}, z_{pq}) p(e_q = s_{eq} | L_i)
\]

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

\[ p(L_i \mid Z^k) = \frac{p(Z_k \mid L_i)p(L_i \mid Z^{k-1})}{p(Z_k \mid Z^{k-1})} \]

- We need to evaluate the normalizing term since the current observation might come from a location not yet contained in the map.
\[ p(Z_k | \mathcal{Z}^{k-1}) = \sum_{\text{all } L} p(Z_k | L)p(L | \mathcal{Z}^{k-1}) \]

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

mapped places

unmapped places
\[ p(Z_k | Z^{k-1}) = \sum_{all \ L} p(Z_k | L)p(L | Z^{k-1}) \]

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

mapped places \hspace{2cm} \text{unmapped places}

approximate by sampling:

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

prior probability \hspace{2cm} \text{sampled place models}

of being at a new location
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

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

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, ....
Lighting Change
Dynamic Objects
Perceptual Aliasing Correctly Rejected
Highest Confidence False Positives
Loop Closure (70 km Data Set)

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