Motion and Tracking

COSC450

4 – Simultaneous Localisation and Mapping

Simultaneous Localisation and Mapping

Localisation:

- Input: A 3D model and 2D image
- Output: Pose of the camera

Challenges:

- 3D-2D feature matching
- Robust pose estimation
- Dealing with ambiguity

Mapping:

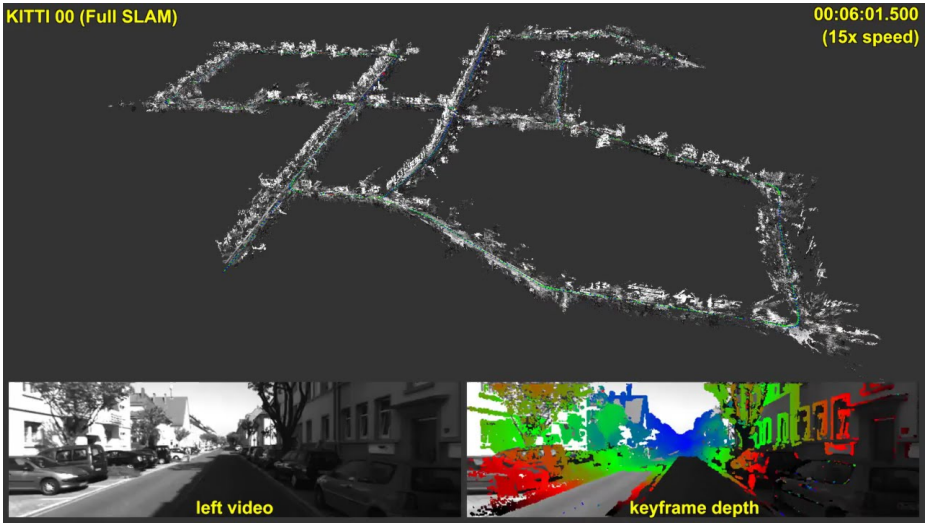
- Input: 2D images (and camera poses?)
- Output: A 3D model of the world

Challenges:

- Determining camera pose
- Dealing with large areas
- Drift / loop closure

Doing two hard things at once – is it twice as hard?

SLAM Example – Large Scale Direct SLAM



https://www.youtube.com/watch?v=oJt3Ln8H03s, Engel, Stückler & Cremers, IROS 2015

SLAM Problem Formulation

- We have a sequence of observations, $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t$
- We want to estimate a map, $\mathbf{m}_1, \mathbf{m}_2, \ldots \mathbf{m}_t$, and location, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$, over time
- This is often represented in a probabilistic framework:

$$P(\mathbf{m}_t, \mathbf{x}_t | \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_n)$$

- Expectation-Maximisation algorithms split this into two parts:
  1. Assuming we know the map, what is our pose? Estimate $P(\mathbf{x}_t | \mathbf{m}_t, \mathbf{x}_{t-1}, \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t)$
  2. Assuming we know the pose, what is the map? Estimate $P(\mathbf{m}_t | \mathbf{x}_1, \mathbf{m}_{t-1}, \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t)$

  What do we expect our pose to be? Maximise the map probability based on this pose.
- Often we can just update based on the latest measurements:

$$P(\mathbf{x}_t | \mathbf{m}_t, \mathbf{x}_{t-1}, \mathbf{o}_t[, \mathbf{o}_{t-1}]) \qquad P(\mathbf{m}_t | \mathbf{x}_t, \mathbf{m}_{t-1}, \mathbf{o}_t[, \mathbf{o}_{t-1}])$$

FastSLAM

Fast Slam (Montemerlo *et al.*, 2002) uses particle and Kalman filters

- Simple case – assume correpsonding 2D features between frames as observations
- The map consists of K 3D landmark points, $m_1, m_2, \ldots, m_k$
- Assuming the landmarks' measurements are independent lets us factor the problem as

$$P(\mathbf{m}_t, \mathbf{x}_t | \mathbf{o}_t) = P(\mathbf{x}_t | \mathbf{o}_t) \prod_{j=1}^{k} P(m_j | \mathbf{x}_t, \mathbf{o}_t)$$

- A particle filter is used to predict the new pose
- A Kalman filter is then used to estimate the map probability

FastSLAM – Particle Filter

We want to estimate $P(\mathbf{x}_t|\mathbf{o}_t)$

- We use a particle filter from $\mathbf{x}_{t-1}$
- This gives us an *a priori* estimate of $P(\mathbf{x}_t)$
- We need to update this based on the measurements, $\mathbf{o}_t$
- To do this we need to predict the observations...
- ...and this depends on the map

FastSLAM – Extended Kalman Filter

The map is first estimated for *each particle*

- We want to estimate $P(m_j|\mathbf{x}_t, \mathbf{o}_t)$
- We know (recursively) the probabilities from the previous frame
- We have a model of how the measurements depend on state and map
- This is generally non-linear – an *extended* Kalman filter
- One Kalman filter per 3D measurement (m_j) per particle
- Lots of EFKs, but all just 2D (measurement)
- This gives a lot of estimates for each 3D point (1 per particle)
- All Gaussians, so easy to combine, so we have a map estimate

Drift and Loop Closure

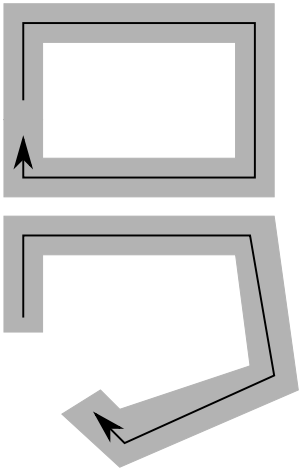
SLAM updates map/location frame-by-frame

- There are small errors each frame
- These accumulate over time
- Eventually the map becomes inaccurate

Loop closure provides corrections

- Detect when you return to a location
- Can then determine the error...
- ...and fix your path around the loop

How to detect when you return to a place?



Bag-of-Words in Document Search

Inspiration from information retrieval

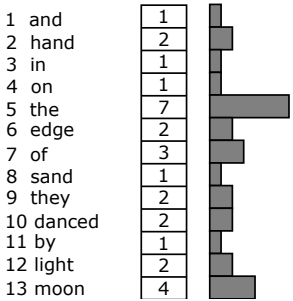
- Document classification and search
- Analyse documents for common words

Bag-of-Words (BoW) models are common

- Each word is given an index
- Documents represented as an array
- Array records frequency of each word
- Related words are grouped together
  compute computer computing

And hand in hand on the edge of the sand
They danced by the light of the moon,
the moon,
the moon,
They danced by the light of the moon.

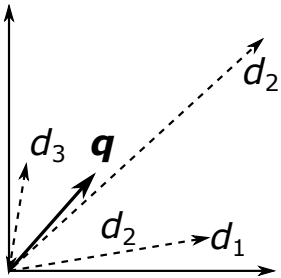
Edward Lear, *The Owl and the Pussycat*



Comparing Bags-of-Words

Distance between histograms

- Can view BoW as a vector
- Euclidean distance?
  - What happens if you compare a document to itself repeated twice?
- Can normalise lengths
- Can just have binary vectors
  - Presence/absence of words, not count
- Can use angle between vectors
  - Easier to compute the cosine



$$\cos\theta(\mathbf{q}, \mathbf{d}_i) = \frac{\mathbf{q} \cdot \mathbf{d}_i}{\|\mathbf{q}\|\|\mathbf{d}_i\|}$$

Document Classification

Given an email is it Spam?

- We have a set of labelled emails
- This has N_S spam, and N_N not spam
- Given a document, D, find $P(S|D)$
- Apply Bayes' rule:

$$P(S|D) = \frac{P(D|S)P(S)}{P(D)} \propto P(D|S)P(S)$$

- $P(S) = \frac{N_S}{N_N}$

What is $P(D|S)$?

- We have a vocabulary of words, $V = \{w_1, w_2, \ldots, w_k\}$
- We know $P(w_i|S)$ from training
- We make an observation $\mathbf{c} = [c_1, c_2, \ldots, c_k]$
- Here c_i is number of times the ith word appears in the email

$$\begin{aligned} P(D|S) &= P(\mathbf{c}|S) \\ &= \prod_{i=1}^{k} (P(w_i|S)^{c_i}) \end{aligned}$$

Log-Likelihood

Minimising products is hard

- Can convert to a sum with logarithms
- We started wanting $P(S|D)$
- We end up computing $P(D|S)$
- This is a likelihood,

$$\mathcal{L}(S|D) = P(D|S)$$

- We end up with a *log-likelihood*
- This is a linear function of $\mathbf{c}$

$$\begin{aligned} P(S|D) &\propto P(D|S)P(S) \\ &= P(S)\prod_{i=1}^{k}(P(w_i|S)^{c_i}) \\ \log P(S|D) &= \log P(S) + \sum_{i=1}^{k}(\log P(w_i|S)^{c_i}) \\ &= \log P(S) + \sum_{i=1}^{k}(c_i \log P(w_i|S)) \\ &= a + \mathbf{c}\cdot\mathbf{b} \end{aligned}$$

Document Clustering

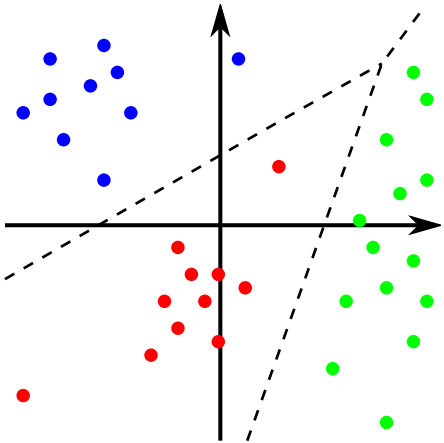
What if we have unlabelled data?

- Can analyse the BoWs
- Look for natural groupings
- Can use k-means clustering
- This will give k classes

Now we can classify new documents

- Find nearest cluster centre
- Or Bayesian approach

$$\arg\max_i P(D|C_i),$$



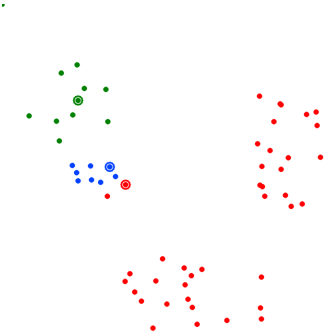
k-Means Clustering

Suppose we have:

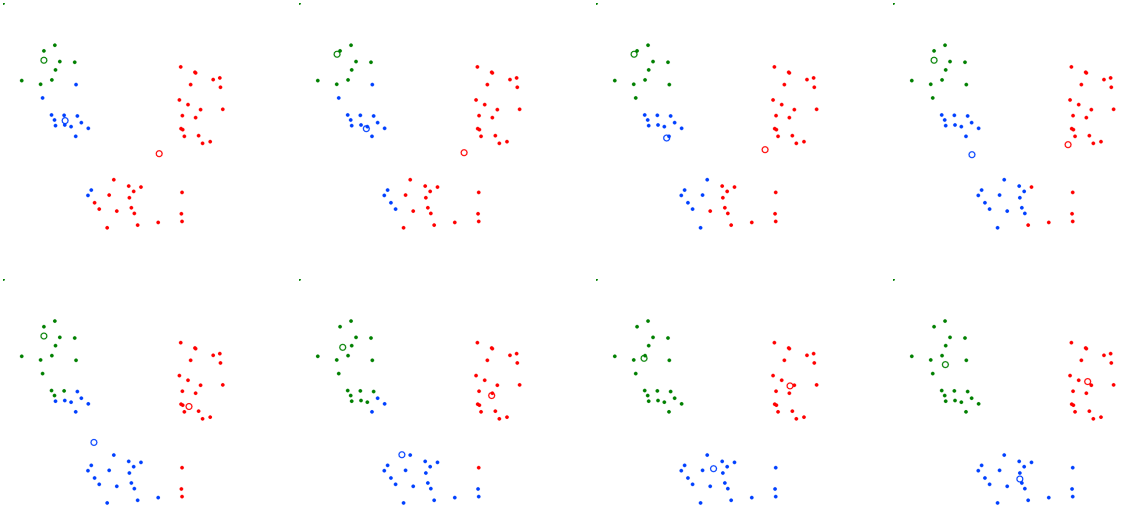
- A set of n data points, $\{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n\}$
- A desired number of clusters, k

1. Pick k points as cluster centres
2. Assign each point to the nearest centre
3. Update each cluster centre to be the average of all points in the cluster
4. If anything has changed goto 2

- How do we choose initial centres?
- How do we detect convergence?



k-Means Clustering



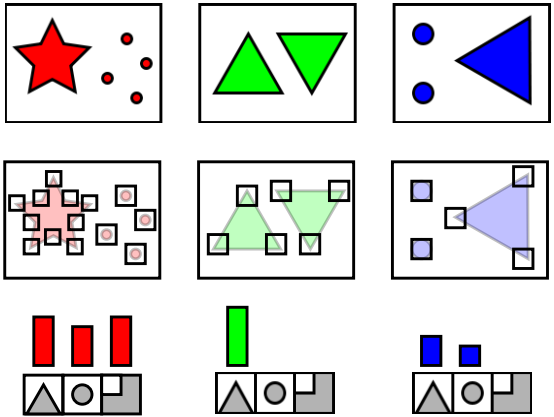
Visual Bag-of-Words

How does this relate to images?

- Documents naturally divide into words
- Feature descriptors can be used
- But features are rarely identical
- So cluster them to make a dictionary

Visual Bag-of-Words

- Given a large collection of descriptors
- Apply k-means for some large k
- This gives a vocabulary of k words
- Images can now be described by a BoW



Using Visual Bags-of-Words

Image Search

- Find similar images to a query image
- Compute BoW for each image in the database that we want to search
- Organise these into a *k*d-Tree
- Given a query, compute its BoW
- Return approximate nearest neighbours using the *k*d-Tree

Object Classification

- We have a set of labelled images
- We want to label a new image
- Compute BoW for each labelled image
- Find average BoW for each label
- Compute query image BoW
- Assign it the nearest label

Evaluating Search Results

How do you measure search 'goodness'

- Searching a database of n_d images
- The database has r_d relevant items
- The search returns n_s images
- And r_s of these are relevant

We define precision and recall

- How many returned results are relevant?
- How many relevant results are returned?

$$\text{Precision} = \frac{\text{retrieved items} \cap \text{relevant items}}{\text{retrieved items}}$$

$$P = \frac{r_s}{n_s}$$

$$\text{Recall} = \frac{\text{retrieved items} \cap \text{relevant items}}{\text{relevant items}}$$

$$R = \frac{r_s}{r_d}$$

Evaluating Classification

Precision and recall are also used here

- The definitions differ slightly
- Consider finding items of class C
- True positive (T_P) – item is in C, and is classified as C
- False positive (F_P) – item is not in C, but is classified as C
- True negative (T_N) – item is not in C, and is not classified as C
- False negative (F_N) – item is in C, but is not classified as C

$$\text{Precision} \quad P = \frac{T_P}{T_P + F_P}$$

$$\text{Recall} \quad R = \frac{T_P}{T_P + F_N}$$

$$\text{Accuracy} \quad A = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

Using 'Word' Frequency

Not all words are equally important

- 'A', 'and', 'the' in English text
- These should be given less weight
- Given a set of documents,

$$D = \{d_1, d_2, \ldots, d_n\},$$

which contain words or terms,

$$T = t_1, t_2, \ldots, t_k$$

- Term frequency, $\text{tf}(t_i, d_j)$, is number of times t_i appears in d_j

Some words are more common

- Inverse document frequency

$$\text{idf}(t_i, D) = \log \frac{n}{|\{d \in D : t_i \in d\}|}$$

- Term frequency – Inverse document frequency

$$\text{tf–idf}(t_i, d_j, D) = \text{tf}(t_i, d_j) \times \text{idf}(t_i, D)$$

- Can then use tf–idf instead of frequency