Content Based Image Retrieval

Søren I. Olsen

Department of Computer Science
University of Copenhagen

Plan for today

- How to query Image Databases
- The idea of visual words, codebook, bag of visual words.
- The two main ingredients for building visual words are image descriptors and data clustering.

Querying Images

- Example from Google Images: Query using the word `apple`

- Maybe not what expected! Try yourself and observe difference when using `apple` and `an apple`

- Simple textual description may fail. Reasons are multiple:
  - Language ambiguity ...
  - Lack of proper image annotation ...
  - Can you think of others?

- Simple textual description may fail. Reasons are multiple:
  - Language ambiguity ...
  - Lack of proper image annotation ...
  - Can you think of others?

Photograph of Andreas Gursky

Another approach

- Replace simple textual description by content based analysis.
- **Exercise:** What content?
  ~~Discuss for 1 minutes with your neighbor what might be good~~ Think for 1 munute about what might be good image features for CBIR

Another approach

- Replace simple textual description by content based analysis.
- **Exercise:** What content?
  ~~Discuss for 1 minutes with your neighbor what might be good~~ Think for 1 munute about what might be good image features for CBIR
- Use for instance Intensity, color, texture, shape etc... in a global or local setting.

An Inspiration: Text Mining

The leading cause of death in the Western World is heart disease and consequently study of normal and pathological heart behavior has become the topic of rigorous research. In particular the study of the shape and motion of the heart is important because many heart diseases are thought to be strongly correlated to the shape and motion of the heart. Important examples of such heart diseases include ischemia and right ventricle (RV) hypertrophy.

An automated analysis must address the following tasks: 1) Extraction of 3-D information from the 2-D slices, 2) Computation of correspondence - the exact motion of the living tissue over time, 3) Generation of the anatomically correct model, 4) Provisions for normal variations with underlying geometric model, 5) Relation of the acquired geometric and motion data to specific diseases.

Our group has developed several methods over the past several years towards the automated analysis of the heart's motion. Due to the common presence of cluttered objects, complex backgrounds, high noise and intensity inhomogeneities in cardiac images, the segmentation problem remains a very difficult task.

"Tonight, in this election, you, the American people reminded us that, while our road has been hard, while our journey has been long, we have picked ourselves up, we have fought our way back," Obama said in his victory speech in Chicago. "We know in our hearts that for the United States of America, the best is yet to come."

Obama defeated Republican Mitt Romney, winning at least 303 electoral votes in yesterday's election with 270 needed for the victory. With one state -- Florida -- yet to be decided, Romney had 206 electoral votes.

The president faces a partisan divide in Congress, with Republicans retaining their House majority while Democrats kept control of the Senate, and a looming fiscal crisis of automatic spending cuts and tax increases set to begin next year unless a compromise is reached.

Romney Remarks

"This is a time for great challenges for America, and I pray that the president will be successful in guiding our nation," Romney said in a concession speech in Boston, where he had watched returns with family and friends. He called Obama to concede and offer congratulations shortly before his remarks.

Some statistics from these documents

Partial word count of the most frequent ones, eliminating very common “the, and, of, is . . . ”

Document 1

- 7 times **heart**
- 4 times **disease(s)**
- 3 times **motion**
- 2 times **geometric**
- 2 times **shape**
- 2 times **model**

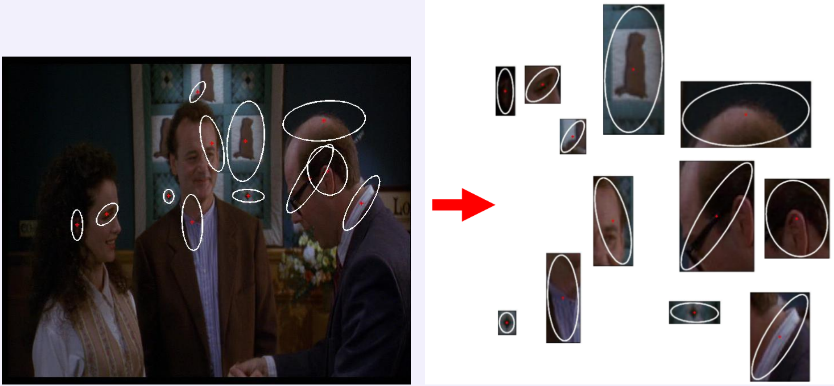
Document 2

- 3 times **elect(ion|oral)**
- 2 times **America(n)**
- 2 times **president**
- 2 times **speech**
- 2 times **state(s)**
- 1 time **heart**

- Even discarding grammatical structure, very different distributions.
- Some variation around a word were grouped “disease|diseases”, “election|electoral”, “America|American”...

Visual Words

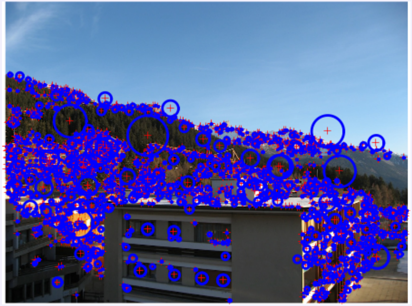
- Visual words: what to choose?



- Here patches centered around Harris Interest points.
- Other possibilities. Here we use SIFT descriptor (more than 64000 citations means it must be interesting)

Visual Words

- Choose output from SIFT as building blocks for visual words.
- Produces generally 1000 – 100.000 features per image.



- Some are similar in content (as for text, e.g., “Election / Electoral”)
- So group them by similarity.

What and how to sample

- Interest points or points sampled in a grid
- Points detected as Blobs, Corners, Centers of symmetri, ...
- Areas: Segments, Objects (what is that?)

Which descriptors are best (most discriminative, yet robust/general)?

- Histograms of color, textons (texture elements) ?
- Coded spatial layout (eg. SIFT-like) ?
- ...

MPEG 7

- Is a Multimedia Content Description Interface Standard
  ISO-standardised in 2002
- Defines a set of image and video descriptors including color descriptors, shape descriptors, motion descriptors etc.
- Includes language for specifying the relations between descriptors and their spatial (or other) relationships.
- Try Google MPEG 7 Standards.
- There is much more to the story than you will learn here.

Learning the Visual Words: Training

- Collect all descriptor vectors from a training set of images showing one class. They must have something in common.
- Group descriptors by similarity: clustering.
- Choose prototypical representatives of each cluster: the **visual words**.
- The set of visual words obtained is called **vocabulary** or **codebook**.

Clustering

Definition

(from Wikipedia) Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.

- I will describe one standard technique for vector clustering: K-means clustering.

Good clustering

Exercise:
Discuss 1 minutes with ~~your neighbor~~ yourself: What characterizes a good clustering

Good clustering

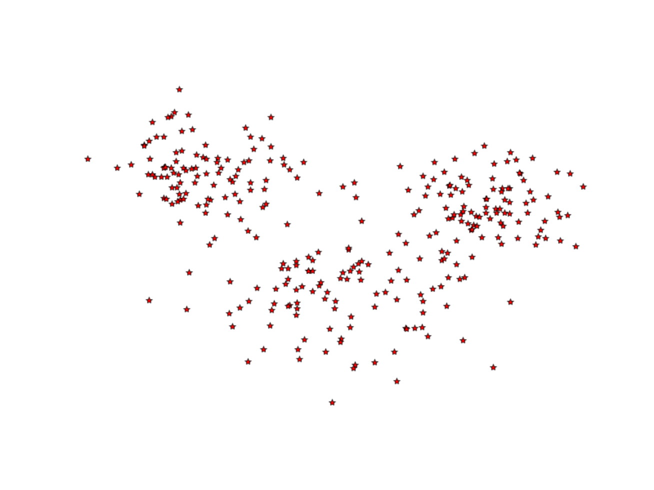
Exercise:
Discuss 1 minutes with ~~your neighbor~~ yourself: What characterizes a good clustering

- Small intra-cluster distances
- Large inter-cluster distances

Central for clustering is to have a good distance measure

A 2D Example

- We want to cluster the following data



- Visually 3 clusters.

K-means Clustering

- Given n data vectors $x_1, \ldots, x_n$ in $\mathbb{R}^d$, find a *partition* $\mathcal{S}$ of $\{1, \ldots, n\}$ into K subsets $S_1, \ldots, S_K$, such that the *Distortion* $\mathcal{D}$

  $$\mathcal{D}(S_1, \ldots, S_K) = \sum_{i=1}^{K} \sum_{j \in S_i} \|x_j - \mu_i\|^2$$

  is minimum, with

  $$\mu_i = \frac{1}{\# S_i} \sum_{j \in S_i} x_m = \text{mean of the } x_j, j \in S_i.$$

- The means μ_i become the prototypical representatives of the vectors, i.e., **the words**

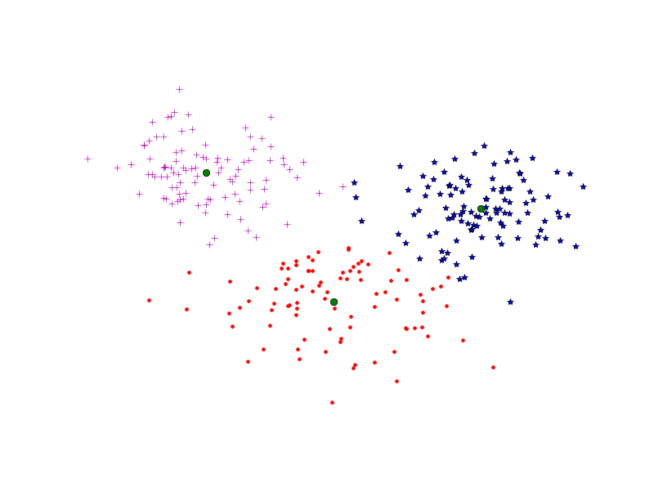
Standard Algorithm (Lloyd's Algorithm)

- Choose K candidate cluster means $m_1, \ldots, m_K$ (randomly).
- Then iterates the following two steps until no significant change occurs in the distortion $\mathcal{D}$
  1. Assignment step: assign each observation x_j to the cluster with closest mean m_i
  2. Update step: recompute the means of the clusters.
- Exercise: Discuss for one minute with ~~your neighbor~~ yourself: Does K-means always converge ?

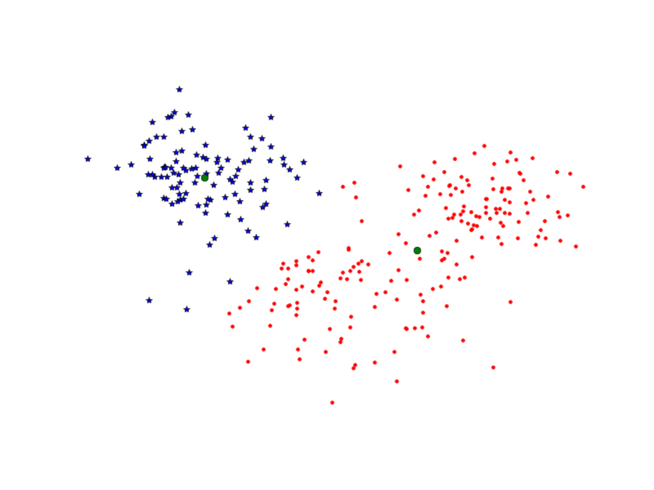
- K-means always converge, but not necessarily fast nor to the right clustering.
- To avoid local minima, run the *k-means* with different initial candidate clusters og choose the best, i.e. the one with minimal *Distortion* $\mathcal{D}$.
- The *Distortion* $\mathcal{D}$ is specified by the sum of distances from each point to the nearest cluster (but there are other approaches).
- You have to fix K. There is no easy way to choose the optimal value.

2D example, $K = 3$

- Run of K-means on the previous data.



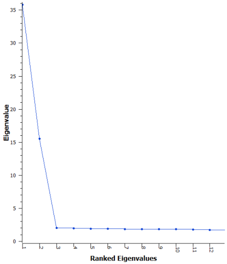
- Run with $K = 2$.



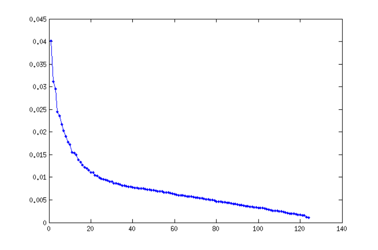
- K can have a deep impact in the clustering results.

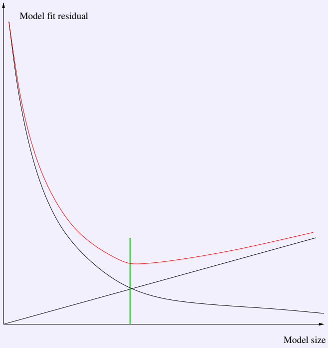
Which K ?

- Most algorithms for clustering assumes K known
- For very low-dimensional data visual exploration techniques (eg. plot3($\cdot$) in Matlab) may be used to guess K
- For high-dimensional data an optimal choice is difficult to guess
- K too small $\Longrightarrow$ Cannot model data
- K too large $\Longrightarrow$ Overfitting
- Problem: Often the average training error decreases smoothly with K (no sudden decrease)
- For CBIR, K should be counted in thousands, and an accurate optimal choice is neither needed nor easy to find.



Eigenvalue
Ranked Eigenvalues



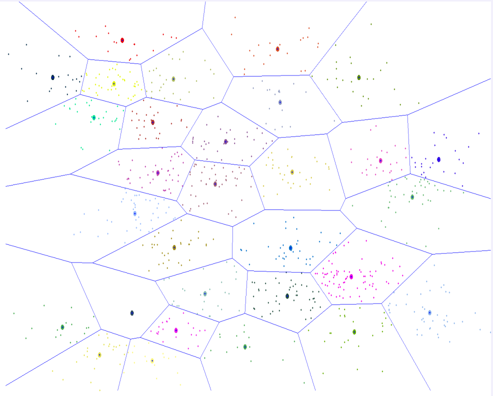


Model fit residual
Model size

Other Clustering Methods

- *Minimum Spanning tree*: Connect all points to its closest neighbor. Then iteratively remove the longest edge until left with K clusters.
- Hierarchical clustering: group data points by proximity, creates a binary tree-structure. Each non-leaf node contains average distance between its subtrees. Clustering is performed by distance threshold. Number of clusters is not predefined.
- Distribution model: observed data is produced by K distributions: clusters belong most likely to the same distribution, EM - *Expectation - Maximization* Algorithms. Can be very powerful.

Clustering and Words



- Words are cluster centers
- Each observation is assigned to the Voronoi cell it belongs to.

Bag of Visual Words Representation of Images

- Once the vocabulary is obtained, each image is “projected” to the vocabulary:
  1. Compute descriptors for the image,
  2. Assign each of them to its closest word
  3. Count the number of occurrences of each word in the image. i.e. compute the histogram of the words for this image.
- This histogram is the **Bag of words** (BOW) representation of the image.
- Spatial arrangement of the words is forgotten. Words are just “thrown in a bag”.
- Histograms can be normalized. Each entry becomes the **word frequency** (or **term frequency**) in the image.

When do BOW malfunction

Exercise:

Discuss 1 minutes with ~~your neighbor~~ yourself when a Bag of Words representation is insufficient or misleading

When do BOW malfunction

Exercise:

Discuss 1 minutes with ~~your neighbor~~ yourself when a Bag of Words representation is insufficient or misleading

Example: When spatial arrangement do matter,

When do BOW malfunction

Exercise:

Discuss 1 minutes with ~~your neighbor~~ yourself when a Bag of Words representation is insufficient or misleading

Example: When spatial arrangement do matter,

Ensemble of Common Words

- Simple similarity measure: count the amount of common words between images
- Can be used for query: return the images that have these “words” in common.
- A subset of words can be used.

Euclidean and Histogram Distances for BoW

- Euclidean Distance between two normalized vectors

$$v_1 = (v_{11}, \ldots, v_{2K}), v_2 = (v_{21}, \ldots, v_{2K}), d(v_1, v_2) = \sqrt{\sum_{i=1}^{K} (v_{1i} - v_{2i})^2}$$

- Bhattacharyya Distance for Normalized Histograms:

$$d(v_1, v_2) = \sum_{i=1}^{K} \sqrt{v_{1i}\, v_{2i}}$$

- Kullback-Leibler Divergence for Normalized Histograms

$$D_{KL}(v_1 || v_2) = \sum_{i=1}^{K} v_{1i} \ln \frac{v_{1i}}{v_{2i}}$$

Not a distance as not symmetric, but can be symmetrized

$$d_{KL}(v_1, v_2) = \frac{1}{2} \left(D_{KL}(v_1 || v_2) + D_{KL}(v_2 || v_1) \right)$$

- χ^2-distance measure $\sum \frac{|v_{1i} - v_{2i}|}{|v_{1i} + v_{2i}|}$
- *Earth movers (Wasserstein-) distance*

The Term Frequency – Inverse Document Frequency

- Some words (terms) are more common than other, not in one document but in a **corpus** (or data set).
- Such terms are in general **less informative**
- Inverse Document Frequency *idf* weighting reduces their importance: For a given word w and a corpus D

$$\mathrm{idf}_w = \frac{\text{Number of documents or images}}{\text{number of document in wich w appears}}$$

idf_w is a global weight for the word w.

- Sometimes the logarithm of the fraction above is used

$$\mathrm{idf}_w = \log\left(\frac{\text{Number of documents or images}}{\text{number of document in wich w appears}}\right)$$

- The **tf-idf** approach reweights the normalized histogram entries of document d (the term frequencies) by the *idf*'s.

  tf-idf = $\mathrm{idf}_w \times$ **tf**

- Alternatively the Euclidean Distance or angle used as similarity. For a document d in the data set and a query document q Compute the tf-idf reweighted BoW's v_d and v_q.
- Define the similarity

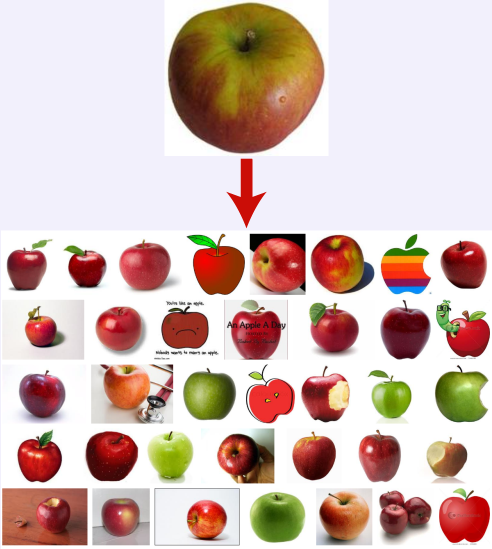
$$\mathrm{sim}(d, q) = \mathrm{angle}(v_d, v_q) = \arccos \frac{v_d \cdot v_q}{\|v_d\| \|v_q\|}$$

- In practice, only the argument to the arc cosine may be computed.

Conclusion

- We have been through the major steps for a Content-Based Image Retrieval System based on the notion of learned vocabulary and bag of visual words.
- Selection interest point detectors and descriptors
- Step of training – learning vocabulary using clustering
- Step of Indexing – computing visual words and histograms (BoW's) of visual words
- Tools for searching / ranking – computing similarity measures.
- Bag of Words is also used for object classification and recognition.
- Main limitation is that BoW ignores spatial relationships among the words
- Incorporating them is a research topic.

Successful query?



Assignment - details

- The assignment consists in implementing a prototypical Content Based Image Retrieval System
- Four parts:
    1. Gather descriptors in training image data set
    2. Construct codebook using k-means and Bag-Of-Words
    3. Project test images onto codebook, and generate BoW
    4. Retrieve according to similarity measure

What descriptors

- You are advised to use SIFT features
- Few versions of SIFT include color. You may miss a lot of information.
- Color histograms etc.
- Other descriptors: Texture
- Advise: Keep it simple

Training and testing

- Split data into two parts. Don't look at the test data before testing
- During development you may split the training data into a construction part and an evaluation part
- Never use the test set for tuning !
- Download Caltech-101 image database (131 MB) with 101 categories or the newer Caltech-256 base (1.2 GB)

What error ?

- Error on training data versus error on test data
- What if the training error is less than the test error?
- Overfitting is a serious problem. Often caused by using too few training data (or too complicated model)
- We want to generalize in order to retrieve/classify new data

Cross-validation

- Divide data into say 10 sets (10-fold CV)
- Train on all but one set that is used for testing
- Retrain and test on all possible combinations
- Compute the average test error

How to report

- Amount: 8 pages including everything: Try to be concise, structured etc. Keep text less than 2-3 pages
- Discussion: This is the most important part. Report WHY you did as you did.
- Tables: Remember to tell me what I should notice
- Figures: Remember axis labels etc. What should I see?
- Images: Don't show them too small. What should I see?
- Explanations: Don't expect me to guess what you mean or see yourself

QUESTIONS