

Content Based Image Retrieval

François Lauze

Department of Computer Science
University of Copenhagen



Plan for today (and perhaps Wednesday)

- Discuss how to implement a Database of Images, with Image-based query.
- For that, I will introduce the idea of visual words, codebook, bag of visual words.
- The two main ingredients for building visual words are image descriptors and data clustering.



Outline

- 1 Introduction
- 2 Bag of Visual Words
- 3 Similarity Measures
- 4 Conclusion
- 5 Assignment



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?



Another approach

- Replace simple textual description by content based analysis.
- Use for instance Intensity, color, texture, shape etc... in a global or local setting.
- Find efficient ways to answer the question:
What makes an image of an apple resemble an image of an apple?



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”...



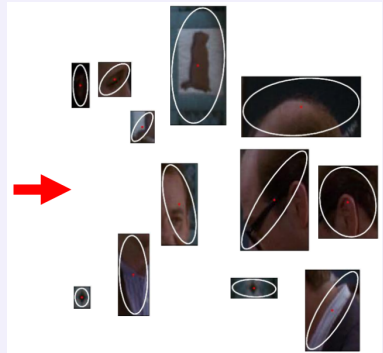
Outline

- 1 Introduction
- 2 Bag of Visual Words
- 3 Similarity Measures
- 4 Conclusion
- 5 Assignment



Visual Words

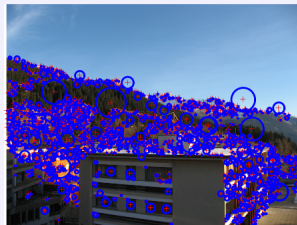
- Visual words: what to choose?



- Here patches centered around Harris Interest points.
- Other possibilities. Here we use SIFT descriptor (more than 16000 citations means it must be interesting)

Visual Words

- Choose output from SIFT as building blocks for visual words.
- Produces generally 100 – 10000 features per image.



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

Learning the Visual Words: Training

- Collect descriptor vectors from a training set of images.
- Group them 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

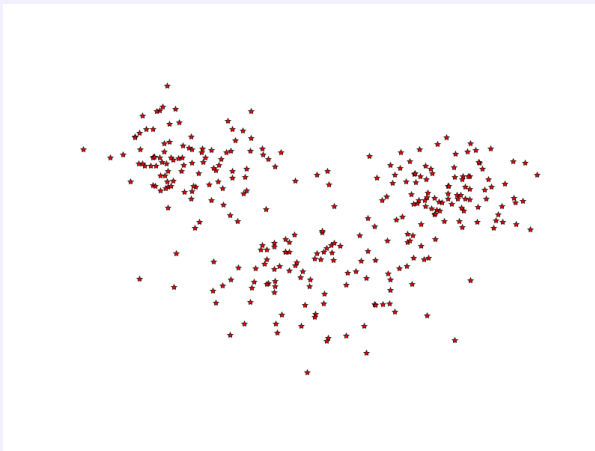
(from Wikipedia) Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.

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



A 2D Example

- We want to cluster the following data



- Visually 3 clusters.

K-means Clustering

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

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

is minimum, with

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

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



Standard Algorithm (Lloyd's Algorithm)

- Choose first K candidate cluster means m_1, \dots, m_K .
- 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.

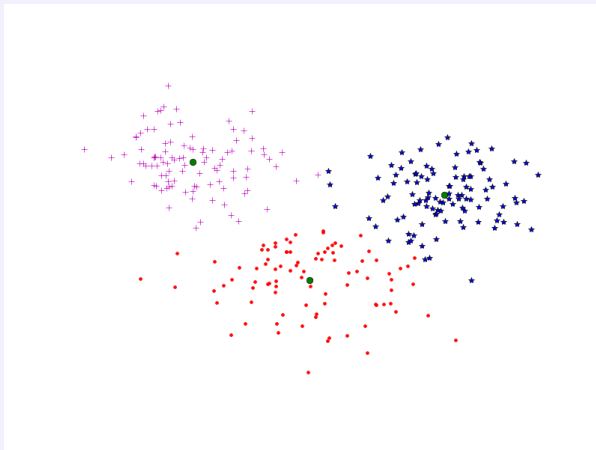


- The first K candidates can be specified, or chosen randomly.
- If a cluster becomes empty, it should be removed and K replaced by $K - 1$
- Several runs can be performed, retaining the one with lowest distortion.

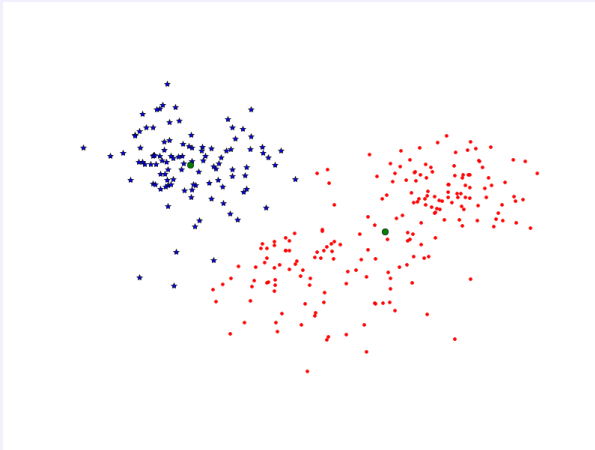


2D example, $K = 3$

- Run of K -means on the previous data.



- Run with $K = 2$.



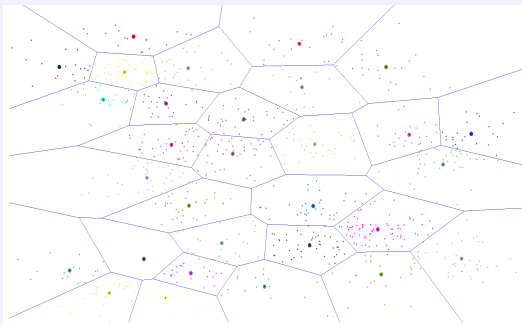
- K can have a deep impact in the clustering results.

Other Clustering Methods

- Hierarchical clustering: group data points by proximity, creates a binary tree-structure. Each non-leaf node contains average distance between it 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, Expectation - Maximization Algorithms. Can be very powerful.
- Non vector space methods. Complicated data that cannot easily be represented by vectors...



Clustering and Words



- Words are cluster centers
- Each observation is assigned to the center of 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** 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.



Outline

- 1 Introduction
- 2 Bag of Visual Words
- 3 Similarity Measures**
- 4 Conclusion
- 5 Assignment



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.
- Easy to implement within a standard relational database.
- Discard bin sizes in histograms.
- Other problems ?



Euclidean and Histogram Distances for BoVW

- Euclidean Distance between two vectors

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

- 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

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



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** (might be seen as a tautological statement!)
- Inverse Document Frequency weighting reduces their importance: For a given word w and a corpus D

$$\text{idf}_w = \frac{\#D}{\#\{d \in D, w \in d\}}$$

idf_w is a global weight.



- The **tf-idf** approach reweights the normalized histogram entries of document d (the term frequencies) by the *idf*'s.
- 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 BoVWs v_d and v_q .
- Define the similarity

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

- In practice, the arc cosine is not computed. $\text{sim}(d, q)$ is the range $[0, 1]$, maximum similarity is 1.



Outline

- 1 Introduction
- 2 Bag of Visual Words
- 3 Similarity Measures
- 4 Conclusion**
- 5 Assignment

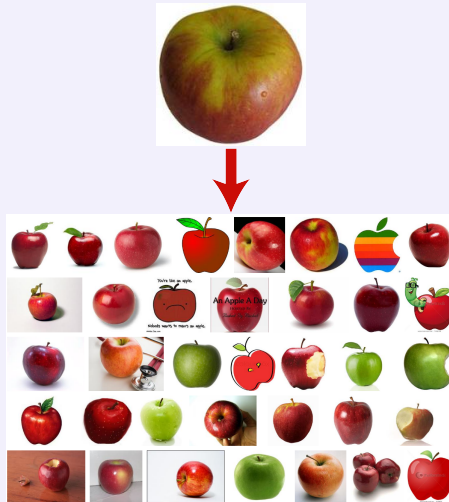


Conclusion

- We saw the major steps for a Content-Based Image Retrieval System based on the notion of learned vocabulary and bag of visual words.
- Step of training – learning vocabulary
- Step of Indexing – computing visual words
- Tools for searching / ranking – computing similarity measures.
- Bag of Words is also used for object classification and recognition.
- Main Limitation is that BoVW ignores spatial relationships among the words
- Incorporating them is a hot research topic.



Successful query?



A really smart CBIR system?



Outline

- 1 Introduction
- 2 Bag of Visual Words
- 3 Similarity Measures
- 4 Conclusion
- 5 Assignment**



Assignment

- The assignment consists in implementing a prototypical Content Based Image Retrieval System :-)
- We'll discuss some details Wednesday.
- Enjoy!

