### **Bag-of-Words Representation**

Many slides are by Tamara Berg, Svetlana Lazebnik, Fei-Fei Li, Rob Fergus, Josef Sivic, and Antonio Torralba, Minh Haoi

#### **Document Vectors**

• Represent a document as a bag of words

The paper bag is a remarkable contrivance. It serves us constantly and inconspicuously. It folds flat, yet opens into a structure that can stand open upon the table while we eat our sandwiches from it and chat with friends.

If we take the bag apart, we find it's made from a single paper cylinder. One end of the cylinder has been folded into a complex 3-dimensional pattern and finished off with a bit of paste. It would be, and once was, costly to make, because each fragile cylinder had to be folded manually into that hardy sack.



• Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)

• Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)

#### 2007-01-23: State of the Union Address

George W. Bush (2001-)

abandon accountable affordable afghanistan africa aided ally anbar armed army **baghdad** bless **challenges** chamber chaos choices civilians coalition commanders **commitment** confident confront congressman constitution corps debates deduction deficit deliver **democratic** deploy dikembe diplomacy disruptions **earmarks economy einstein elections** eliminates expand **extremists** failing faithful families **freedom** fuel **funding** god haven ideology immigration impose insurgents **iran iran use islam julie** lebanon love madam marine math medicare moderation neighborhoods nuclear offensive palestinian payroll province pursuing **gaeda** radical regimes resolve retreat rieman sacrifices science sectarian senate

september **shia** stays strength students succeed sunni tax territories **terrorists** threats uphold victory violence violent War washington weapons wesley

• Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)



• Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)

2007-01-23: State of the Union Address George W. Bush (2001-)		
abandon choices c	1962-	10-22: Soviet Missiles in Cuba John F. Kennedy (1961-63)
expand () insurgen palestinia	abando <b>buildu</b>	1941-12-08: Request for a Declaration of War Franklin D. Roosevelt (1933-45)
	declinec elimina	abandoning acknowledge aggression aggressors airplanes armaments <b>armed army</b> assault assembly authorizations bombing britain british cheerfully claiming constitution curtail december defeats defending delays <b>democratic dictators</b> disclose
septemb violenc	halt ha moderni	economic empire endanger TACUS false forgotten fortunes france <b>Treedom</b> fulfilled fullness fundamental gangsters german germany god guam harbor hawafi hemisphere hint hitler hostilities immune improving indies innumerable
	recessio surveill	invasion <b>islands</b> isolate <b>Japanese</b> labor metals midst midway navy nazis obligation offensive officially <b>pacific</b> partisanship patriotism pearl peril perpetrated perpetual philippine preservation privilege reject repaired resisting retain revealing rumors seas soldiers speaks speedy stamina strength sunday sunk supremacy tanks taxes
		treachery true tyranny undertaken victory War wartime washington

## **Bag-of-features models**





Many slides adapted from Fei-Fei Li, Rob Fergus, and Antonio Torralba

#### 1. Extract features







- 1. Extract features
- 2. Learn "visual vocabulary"



- 1. Extract features
- 2. Learn "visual vocabulary"
- 3. Quantize features using visual vocabulary

- 1. Extract features
- 2. Learn "visual vocabulary"
- 3. Quantize features using visual vocabulary
- Represent images by frequencies of "visual words"



- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005



- Regular grid
  - Vogel & Schiele, 2003
  - Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic et al. 2005



#### • Regular grid

- Vogel & Schiele, 2003
- Fei-Fei & Perona, 2005
- Interest point detector
  - Csurka et al. 2004
  - Fei-Fei & Perona, 2005
  - Sivic et al. 2005
- Other methods
  - Random sampling (Vidal-Naquet & Ullman, 2002)
  - Segmentation-based patches (Barnard et al. 2003)



#### Compute SIFT descriptor

[Lowe'99]



Normalize patch



#### **Detect patches**

[Mikojaczyk and Schmid '02] [Mata, Chum, Urban & Pajdla, '02] [Sivic & Zisserman, '03]

Slide credit: Josef Sivic







Slide credit: Josef Sivic



Slide credit: Josef Sivic

# Clustering

- The assignment of objects into groups (called clusters) so that objects from the same cluster are more similar to each other than objects from different clusters
- Often similarity is assessed according to a distance measure
- Clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics

## Some Data



- Ask user how many clusters they'd like. (e.g. k=5)
- 2. Randomly guess k cluster Center locations



- Ask user how many clusters they'd like. (e.g. k=5)
- 2. Randomly guess k cluster Center locations



- Ask user how many clusters they'd like. (e.g. k=5)
- Randomly guess k cluster Center locations
- Each datapoint finds out which Center it's closest to.



- Ask user how many clusters they'd like. (e.g. k=5)
- Randomly guess k cluster Center locations
- Each datapoint finds out which Center it's closest to.
- 4. Each Center finds the centroid of the points it owns...



- Ask user how many clusters they'd like. (e.g. k=5)
- Randomly guess k cluster Center locations
- Each datapoint finds out which Center it's closest to.
- 4. Each Center finds the centroid of the points it owns...
- 5. ...and jumps there



- Ask user how many clusters they'd like. (e.g. k=5)
- Randomly guess k cluster Center locations
- Each datapoint finds out which Center it's closest to.
- Each Center finds the centroid of the points it owns...
- 5. ...and jumps there

Repeat until terminated



- Randomly initialize k centers
  - $\mu^{(0)} = \mu_1^{(0)}, ..., \mu_k^{(0)}$
- Classify: Assign each point j∈{1,...N} to nearest center:

$$C^{(t)}(j) \leftarrow \arg\min_i ||\mu_i - x_j||^2$$

• Recenter: μ<sub>i</sub> becomes centroid of its point:

$$\mu_i^{(t+1)} \leftarrow \arg\min_{\mu} \sum_{j:C(j)=i} ||\mu - x_j||^2$$

- Equivalent to  $\mu_i \leftarrow$  average of its points!

## What is K-means optimizing?

 Potential function F(μ,C) of centers μ and point allocations C:

$$F(\mu, C) = \sum_{j=1}^{N} ||\mu_{C(j)} - x_j||^2$$

• Optimal K-means:  $\min_{\mu} \min_{C} F(\mu, C)$ 

#### Does K-means converge??? Part 1

• Optimize potential function:

$$\min_{\mu} \min_{C} F(\mu, C) = \min_{\mu} \min_{C} \sum_{i=1}^{k} \sum_{j:C(j)=i} ||\mu_i - x_j||^2$$

• Fix μ, optimize C

#### Does K-means converge??? Part 2

• Optimize potential function:

$$\min_{\mu} \min_{C} F(\mu, C) = \min_{\mu} \min_{C} \sum_{i=1}^{k} \sum_{j:C(j)=i} ||\mu_i - x_j||^2$$

• Fix C, optimize  $\mu$ 

## **Coordinate descent algorithms**

$$\min_{\mu} \min_{C} F(\mu, C) = \min_{\mu} \min_{C} \sum_{i=1}^{k} \sum_{j:C(j)=i} ||\mu_i - x_j||^2$$

- Want: min<sub>a</sub> min<sub>b</sub> F(a,b)
- Coordinate descent:
  - fix a, minimize b
  - fix b, minimize a
  - repeat
- Converges!!!
  - if F is bounded
  - to a (often good) local optimum
  - (For LASSO it converged to the global optimum, because of convexity)

• K-means is a coordinate descent algorithm!





Slide credit: Josef Sivic



Slide credit: Josef Sivic

#### From clustering to vector quantization

- Clustering is a common method for learning a visual vocabulary or codebook
  - Unsupervised learning process
  - Each cluster center produced by k-means becomes a codebook entry
  - Codebook can be learned on separate training set
  - Provided the training set is sufficiently representative, the codebook will be "universal"
- The codebook is used for quantizing features
  - A vector quantizer takes a feature vector and maps it to the index of the nearest entry in the codebook
  - Codebook = visual vocabulary
  - Codebook entry = visual word

#### **Example Visual Vocabulary**



Fei-Fei et al. 2005

## Image patch examples of visual words



## Visual vocabularies: Issues

- How to choose vocabulary size?
  - Too small: visual words not representative of all patches
  - Too large:
    - quantization artifacts, overfitting
    - computational expensive



## 3&4. Image Representation



Summarize entire image based on its distribution (histogram) of word occurrences.

#### 3&4. Image Representation



# Comparing bags of words

- Use some distance metric,
  - e.g., normalized scalar product between their (possibly weighted) occurrence counts
- Use some classifier/ranker, e.g., nearest neighbor, SVM



$$sim(d_j, q) = \frac{\langle d_j, q \rangle}{\|d_j\| \|q\|}$$
$$\sum_{i=1}^V d_j(i) * q(i)$$

$$\sqrt{\sum_{i=1}^{V} d_j(i)^2} * \sqrt{\sum_{i=1}^{V} q(i)^2}$$

for vocabulary of V words

Kristen Grauman

# tf-idf weighting

- Term frequency inverse document frequency
- Describe frame by frequency of each word within it, downweight words that appear often in the database
- (Standard weighting for text retrieval)



## Bags of words for content-based image retrieval

#### Visually defined query

"Find this clock"



#### "Groundhog Day" [Rammis, 1993]



Slide from Andrew Zisserman Sivic & Zisserman, ICCV 2003

## Example



#### Slide from Andrew Zisserman Sivic & Zisserman, ICCV 2003

#### retrieved shots







Start frame 52907

Key frame 53026

End frame 53028







Start frame 54342

Key frame 54376

End frame 54644







Start frame 51770

Key frame 52251

End frame 52348







End frame 54201



Start frame 54079



Start frame 38909

Key frame 39126

Key frame 54201

End frame 39300







End frame 41049



Start frame 40760





Start frame 39301

Key frame 39676

End frame 39730

# Video Google System

- 1. Collect all words within query region
- 2. Inverted file index to find relevant frames
- 3. Compare word counts
- 4. Spatial verification

Sivic & Zisserman, ICCV 2003 Demo online at: http://www.robots.ox.ac.uk/~vgg/research/vgoo gle/index.html



# Query region



K. Grauman, B. Leibe





Lazebnik, Schmid & Ponce (CVPR 2006)



Lazebnik, Schmid & Ponce (CVPR 2006)



Lazebnik, Schmid & Ponce (CVPR 2006)

- Divide an image into multiple level grids, normally 3:
  - Level 0: full image, Level 1: 2x2 grid, Level 2: 4x4 grid
- Compute a BoW feature vector for each region
- Weight the feature vectors based on level
  - Level 0: 1/2<sup>L</sup>
  - Other levels: 1/2<sup>L-l+1</sup>
- Concatenate the weighted feature vectors of multiple regions
- Do L<sub>2</sub> normalization

#### Bags of features for action recognition

#### Space-time interest points



Juan Carlos Niebles, Hongcheng Wang and Li Fei-Fei, <u>Unsupervised Learning of Human Action</u> <u>Categories Using Spatial-Temporal Words</u>, IJCV 2008.

#### Bags of features for action recognition



Juan Carlos Niebles, Hongcheng Wang and Li Fei-Fei, <u>Unsupervised Learning of Human Action</u> <u>Categories Using Spatial-Temporal Words</u>, IJCV 2008.

## Bags of words: pros and cons

- + flexible to geometry / deformations / viewpoint / scale
- + compact summary of image content
- + provides vector representation for sets
- + very good results in practice
- basic model ignores geometry must verify afterwards, or encode via features
- background and foreground mixed when bag covers whole image
- optimal vocabulary formation remains unclear

## What you need to know

- BoW representation
- Vocabulary construction
- K-means clustering
- Spatial Pyramid