



University of Cyprus – MSc Artificial Intelligence

MAI644 – COMPUTER VISION Lecture 12: Visual Bag of Words

Melinos Averkiou

CYENS Centre of Excellence University of Cyprus - Department of Computer Science m.averkiou@cyens.org.cy











Last time

- A simple Image Classification pipeline
 - Classification overview
- K-nearest neighbor algorithm
 - kNN: algorithm
 - kNN: analysis







Today's Agenda

- Visual bag of words (BoW)
 - Background
 - Algorithm
- Applications
 - Image search
- Spatial Pyramid Matching

[material based on Niebles-Krishna]







Today's Agenda

- Visual bag of words (BoW)
 - Background
 - Algorithm
- Applications
 - Image search
- Spatial Pyramid Matching























Origin 1: Texture Recognition



Example textures (from Wikipedia)











Origin 1: Texture Recognition

Texture is characterized by the repetition of basic elements or *textons*



Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003









Origin 1: Texture Recognition

Recognition based on identity of the textons, not their spatial arrangement (although that is very important too!)

Universal texton dictionary









Origin 1: Texture Recognition











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 Iraq** islam julie lebanon love madam marine math medicare moderation neighborhoods nuclear offensive palestinian payroll province pursuing **qaeda** radical regimes resolve retreat rieman sacrifices science sectarian senate

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

US Presidential Speeches Tag Cloud <u>http://chir.ag/phernalia/preztags/</u>









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



US Presidential Speeches Tag Cloud http://chir.ag/phernalia/preztags/









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

2007-0	1-23: St	ate of the Union Address George W. Bush (2001-)
abandon choices c deficit c	1962-	10-22: Soviet Missiles in Cuba John F. Kennedy (1961-63)
expand	abando build u	1941-12-08: Request for a Declaration of War Franklin D. Roosevelt (1933-45)
palestinia	declinec elimina	abandoning acknowledge aggression aggressors airplanes armaments armed army assault assembly authorizations bombing britain british cheerfully claiming constitution curtail december defeats defending delays democratic dictators disclose
septemt violenc	halt ha modern	economic empire endanger IACLS false forgotten fortunes france ITEECOOM fulfilled fullness fundamental gangsters german germany god guam harbor hawaii hemisphere hint hitler hostilities immune improving indies innumerable
	recessio surveill	invasion islands isolate Japanese labor metals midst midway navy nazis obligation offensive officially pacific 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 Wartime washington

US Presidential Speeches Tag Cloud http://chir.ag/phernalia/preztags/









Bags of features for object recognition



face, flowers, building

Works pretty well for image-level classification and for recognizing object *instances*

Csurka et al. (2004), Willamowski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)





This Master is run under the context of Action

under GA nr. INEA/CÉF/ICT/A2020/2267423

No 2020-EU-IA-0087, co-financed by the EU CEF Telecom



Bags of features for object recognition

class	bag of features	bag of features	Parts-and-shape model
Class	Zhang et al. (2005)	Willamowski et al. (2004)	Fergus et al. (2003)
airplanes	98.8	97.1	90.2
cars (rear)	98.3	98.6	90.3
cars (side)	95.0	87.3	88.5
faces	100	99.3	96.4
motorbikes	98.5	98.0	92.5
spotted cats	97.0		90.0

Today's Agenda

- Visual bag of words (BoW)
 - Background
 - Algorithm
- Applications
 - Image search
- Spatial Pyramid Matching

Bag of features

- First, take a bunch of images, extract features, and build up a "dictionary" or "visual vocabulary" – a list of common features
- Given a new image, extract features and build a histogram for each feature, find the closest visual word in the dictionary

Bag of features: outline

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
- 4. Represent images by frequencies of "visual words"

1. Extract features

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

1. Feature extraction

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

1. Feature extraction

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

1. Feature extraction

• 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)

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

Image patch examples of visual words

Sivic et al 2005

2. Learn the visual vocabulary

2. Learn the visual vocabulary

Slide credit: Josef Sivic

2. Learn the visual vocabulary Visual vocabulary $[\equiv] [\equiv] [\equiv] ...$ Clustering

Slide credit: Josef Sivic

Example visual vocabulary

Fei-Fei et al. 2005

Visual vocabularies: issues

- How to choose vocabulary size?
 - Too small: visual words not representative of all patches
 - Too large: quantization artifacts, overfitting
- Computational efficiency
 - Solution: Vocabulary trees (Nister & Stewenius, 2006)

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

3. 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 codevector
 - 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 codevector in a codebook
 - Codebook = visual vocabulary
 - Codevector = visual word

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

Today's Agenda

- Visual bag of words (BoW)
 - Background
 - Algorithm
- Applications
 - Image search
- Spatial Pyramid Matching

Image classification

• Given the bag-of-features representations of images from different classes, how do we learn a model for distinguishing them?

Uses of BoW representation

- Treat as feature vector for standard classifier
 - e.g k-nearest neighbors, support vector machine

- Cluster BoW vectors over image collection
 - Discover visual themes

Large-scale image search

Bag-of-words models have been useful in matching an image to a large database of object *instances*

How do I find this image in the database?

Large-scale image search

Build the database:

- Extract features from the database images
- Learn a vocabulary using k-means (typical k: 100,000)
- Compute *weights* for each word
- Create an inverted file mapping words → images

Weighting the words

• Just as with text, some visual words are more discriminative than others

the, and, or vs. cow, AT&T, Cher

- The bigger fraction of the documents a word appears in, the less useful it is for matching
 - e.g., a word that appears in *all* documents is not helping us

TF-IDF weighting

- Instead of computing a regular histogram distance, we'll weight each word by its *inverse document frequency*
- Inverse Document Frequency (IDF) of word *j* =

log <u>number of documents</u> number of documents in which *j* appears

TF-IDF weighting

- Term Frequency (TF) of word *j* is the number of times it appears in the 'document', i.e., the image
- To compute the value of bin *j* in image *I*, compute TF-IDF: Term frequency of *j* in *I* \mathbf{X} Inverse Document Frequency of *j*

Inverted file

- Each image has ~1,000 features
- We have ~100,000 visual words

 \rightarrow each histogram is extremely sparse (mostly zeros)

- Inverted file
 - mapping from words to 'documents', i.e., images

"a":	{2}		
"banana":	{2}		
"is":	{O,	1,	2}
"it":	{O,	1,	2}
"what":	{O,	1}	

Inverted file

- Can quickly use the inverted file to compute similarity between a new image and all the images in the database
 - Only consider database images whose bins overlap the query image

Large-scale image search

top 6 results

- Cons:
 - performance degrades as the database grows

Large-scale image search

- Cons:
 - performance degrades as the database grows

Large-scale image search

- Pros:
 - Works well for CD covers, movie posters
 - Real-time performance possible

Real-time retrieval from a database of 40,000 CD covers Nister & Stewenius, Scalable Recognition with a Vocabulary Tree

Example bag-of-words matches

This Master is run under the context of Action

50

Example bag-of-words matches

Today's Agenda

- Visual bag of words (BoW)
 - Background
 - Algorithm
- Applications
 - Image search
 - Action recognition
- Spatial Pyramid Matching

What about spatial info?

Pyramids

- Very useful for representing images.
- Pyramid is built by using multiple copies of image.
- Each level in the pyramid is 1/4 of the size of previous level.
- The lowest level is of the highest resolution.
- The highest level is of the lowest resolution.

Bag of words + pyramids

Locally orderless representation at several levels of spatial resolution

Bag of words + pyramids

Locally orderless representation at several levels of spatial resolution

Bag of words + pyramids

Results: Scene category dataset

Multi-class classification results (100 training images per class)

	Weak features		Strong features	
	(vocabulary size: 16)		(vocabulary	size: 200)
Level	Single-level	Pyramid	Single-level	Pyramid
$0(1 \times 1)$	45.3 ± 0.5		72.2 ± 0.6	
$1 (2 \times 2)$	53.6 ± 0.3	56.2 ± 0.6	77.9 ± 0.6	79.0 ± 0.5
$2(4 \times 4)$	61.7 ± 0.6	64.7 ± 0.7	79.4 ± 0.3	81.1 ±0.3
$3 (8 \times 8)$	63.3 ± 0.8	66.8 ±0.6	77.2 ± 0.4	80.7 ± 0.3

Lazebnik, Schmid & Ponce (CVPR 2006)

Slide credit: Svetlana Lazebnik

Results: Caltech101 dataset

http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html

Multi-class classification results (30 training images per class)

	Weak features (16)		Strong feat	ures (200)
Level	Single-level	Pyramid	Single-level	Pyramid
0	15.5 ± 0.9		41.2 ± 1.2	
1	31.4 ± 1.2	32.8 ± 1.3	55.9 ± 0.9	57.0 ± 0.8
2	47.2 ± 1.1	49.3 ± 1.4	63.6 ± 0.9	$\textbf{64.6} \pm 0.8$
3	52.2 ± 0.8	54.0 ± 1.1	60.3 ± 0.9	$64.6\pm\!0.7$

Lazebnik, Schmid & Ponce (CVPR 2006)

Slide credit: Svetlana Lazebnik

Thank you.

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

