Multimodal Semantic Spaces

Elia Bruni

KnowDive Seminar series - University of Trento

March 23, 2011

Aknowledgments

- This project is possible thanks to a Google Research Award

Outline

Outline

Traditional models

- A popular tradition of studying semantic representation has been driven by the assumption that word meaning can be learned from the linguistic environment
- Semantic space models, among which Latent Semantic Analysis (LSA, Landauer and Dumais 1997) is perhaps known best, operationalize this idea by capturing word meaning quantitatively in terms of simple co-occurrence statistics

Perceptual information

- Despite their popularity, these models offer a somewhat impoverished representation of word meaning based solely on information provided by the linguistic input; but we know that the cognitive system is also sensitive to perceptual information
- Studies in language acquisition suggest that word meaning arises not only from exposure to the linguistic environment but also from our interaction with the physical world
- For example, infants are from an early age able to form perceptually-based category representations (Quinn et al., 1993)

Perceptual Information

- Plausibly, words that refer to concrete entities and actions are among the first words being learned as these are directly observable in the environment (Bornstein et al., 2004)
- Children appear to respond to categories on the basis of visual features, e.g., they generalize object names to new objects often on the basis of similarity in shape (Landau et al., 1998) and texture (Jones et al., 1991)

The idea

- Today, through the Web, we have access to huge amounts of documents that contain both text and images
- The use of text to improve image labeling and retrieval is already an active and growing area of research (e.g, Feng and Lapata, 2008, Moringen, 2008, Wang et al., 2009)
- But in this project we want to go the other way around, and develop novel techniques to extract multimodal semantic spaces from texts and images, in order to improve the measurement of semantic similarity among words

A unified model

- A framework of word meaning that captures the mutual dependence between the linguistic and visual context
  - How to integrate linguistic and visual information in a single representation?
  - By converting the visual features from a continuous onto a discrete space, thereby rendering image features more like word units

Outline

The distributional hypothesis

- Words that occur in similar contexts tend to have similar meanings (Wittgenstein, 1953)
- The meaning of a word is the set of contexts in which it occurs in text
- It is a theory of meaning that can be easily operationlized into a procedure to "extract meaning" from text corpora on a large scale

The distributional hypothesis in everyday life

McDonald & Ramscar (2001)

- He filled the wampimuk with the substance, passed it around and we all drunk some
- We found a little, hairy wampimuk sleeping behind the tree

Semantic spaces

- The idea is to use corpus-based statistics to extract information about semantic properties of words
- Here we focus on models that:
  - Represent the meaning of words as *vectors* keeping track of the words' distributional history
  - Focus on the notion of *semantic similarity*, measured with geometrical methods in the space inhabited by the distributional vectors

Constructing the models

- Pre-process the source corpus
  - Tokenization, POS tagging
- Collect a co-occurrence matrix (with *distributional vectors* representing words as rows, and contextual elements of some kind as comlumns/dimensions)
- Transform the matrix: re-weighting raw frequencies
- Use resulting matrix to compute word-to-word similarity

Distributional vectors

- Count how many times each target word occurs in a certain context
- Build vectors out of (a function of) these context occurrence words
- Similar words will have similar vectors

Collecting context counts for target word dog

The dog barked in the park.
The owner of the dog put him
on the leash since he barked.

bark	++
park	+
owner	+
leash	+

The co-occurrence matrix

	leash	walk	run	owner	pet	bark
dog	3	5	2	5	3	2
cat	0	3	3	2	3	0
lion	0	3	2	0	1	0
light	0	0	0	0	0	0
bark	1	0	0	2	1	0
car	0	0	1	3	0	0

Contexts as vectors

	runs	legs
dog	1	4
cat	1	5
car	4	0

Computing the angle

Example

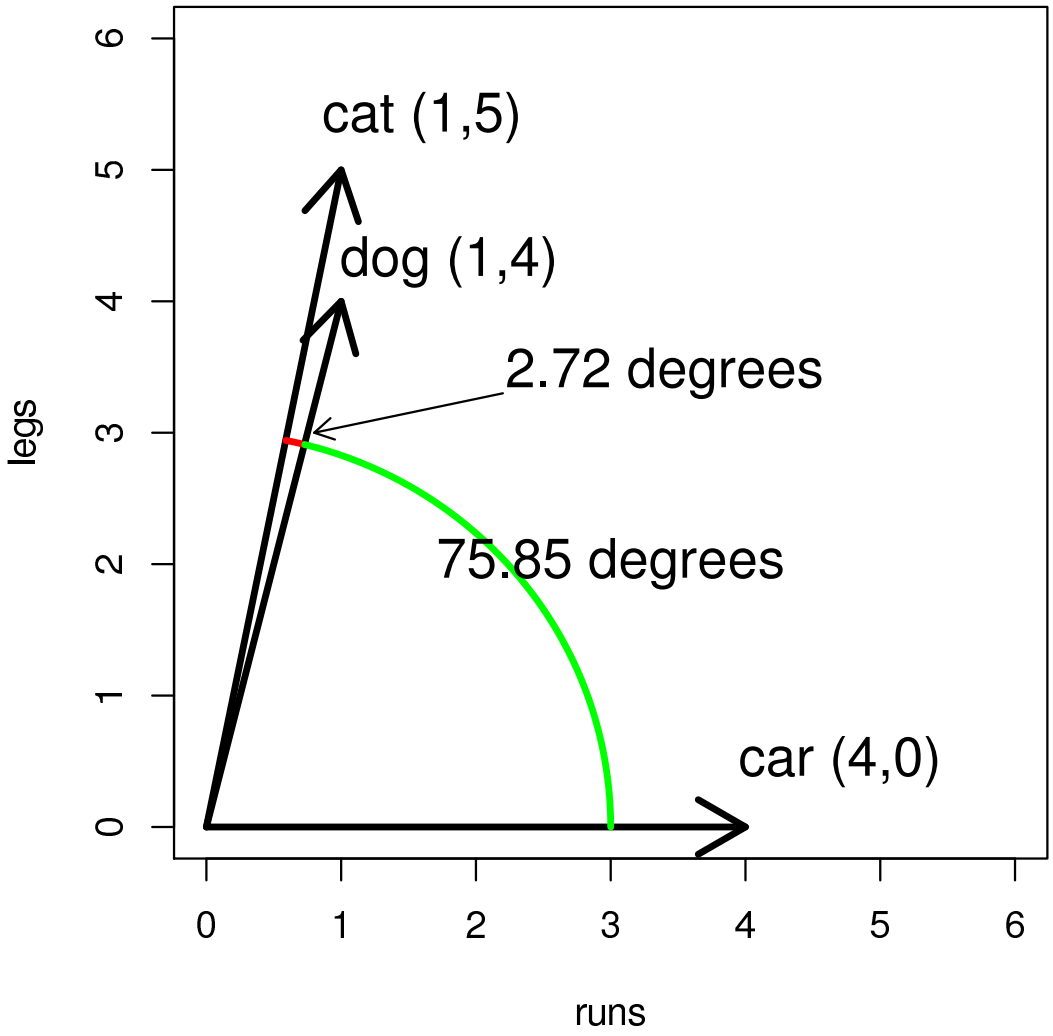


Figure: M. Baroni

Outline

Bag of visual words

- As bag-of-words approach employed in information retrieval, the “bag of visual codewords” is a similar technique used mainly for scene classification (Yang et al., 2007):
  1. To represent an image using BoW model, an image can be treated as a document
  2. However, "words" in images do not come off-the-shelf like in text documents
  3. To achieve bag-of-words representation of image document, pipeline in next slide is typically followed

The pipeline

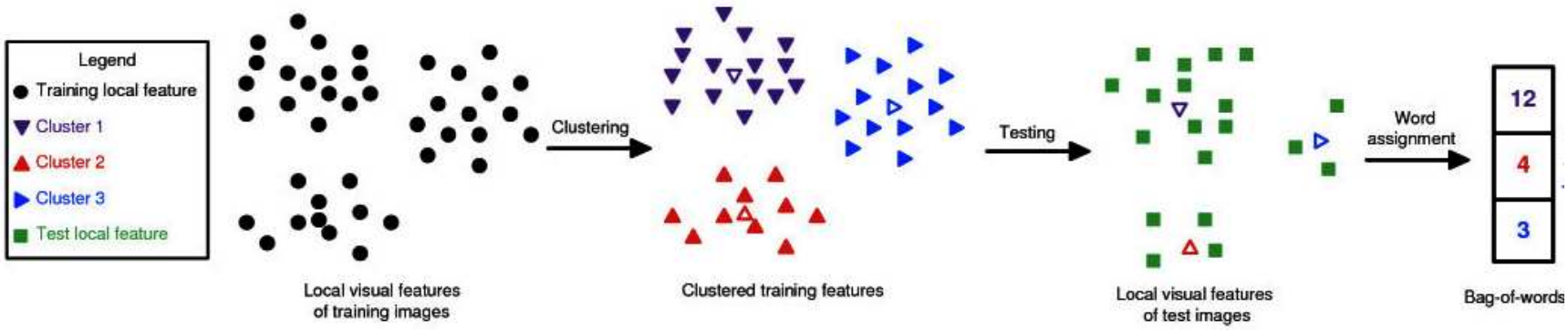


Figure: First, a visual codebook is constructed by applying a K-means to a subset of the local features from training images, and the center of each cluster is considered as a unique "visual word" in the codebook. Each local feature in a test image is then mapped to the closest visual word, and each test image is represented as a histogram of visual words (Ji et al, 2009)

Challenges in basic features extraction

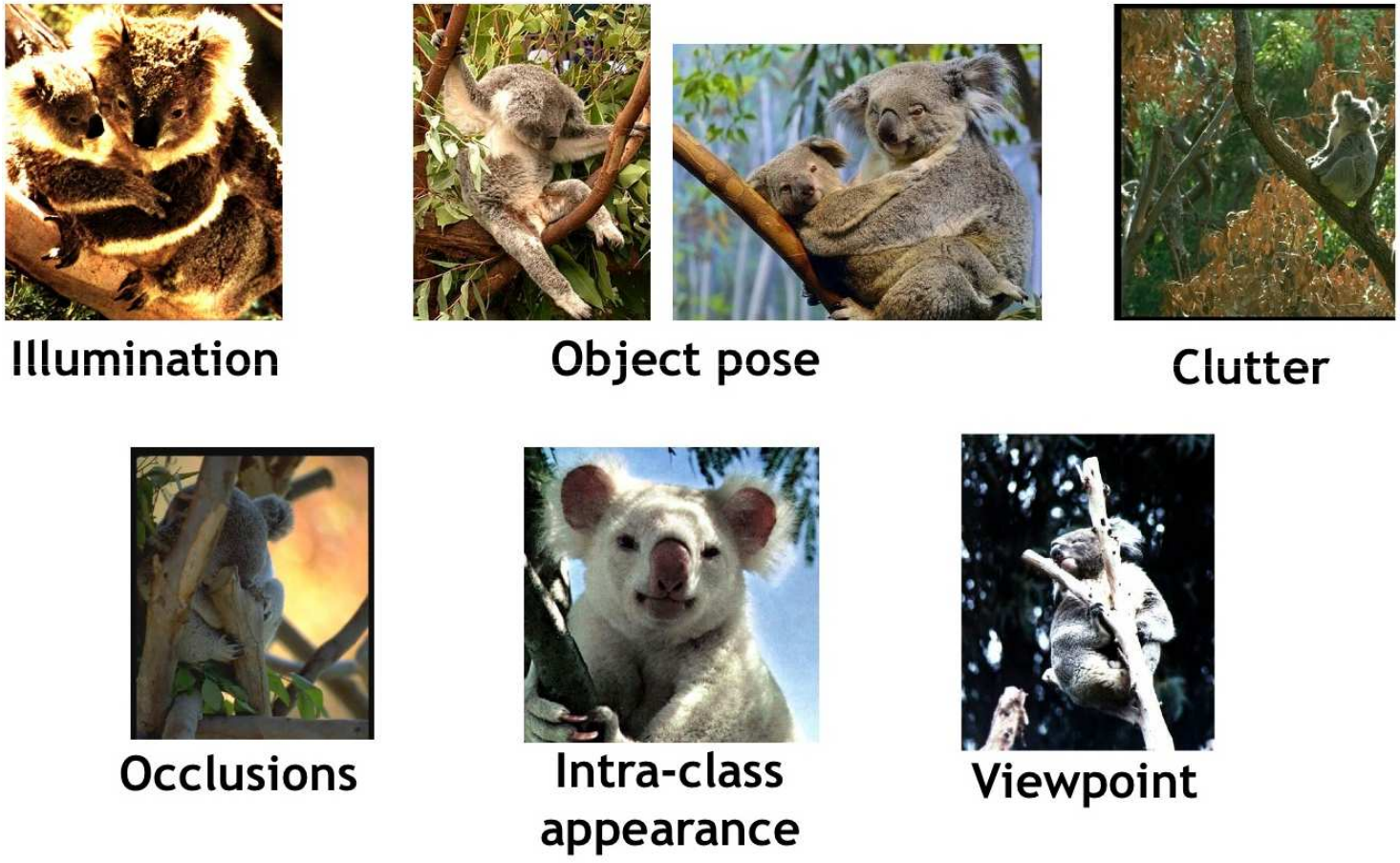


Figure: K. Grauman, B. Leibe

The approach

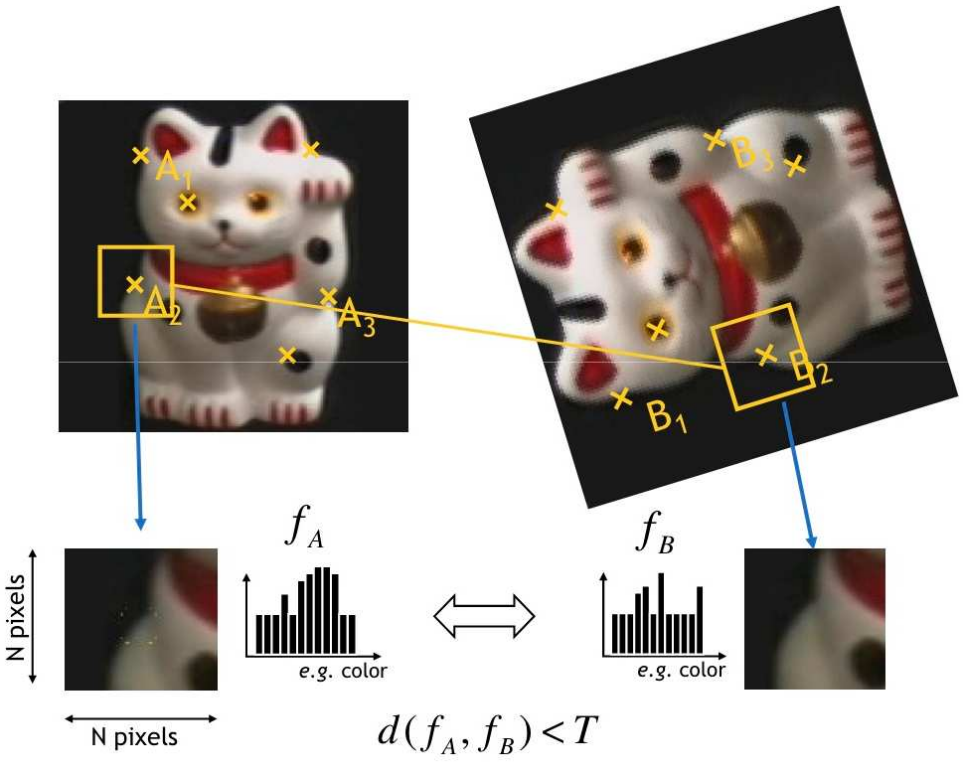


Figure: K. Grauman, B. Leibe

1. Find a set of distinctive keypoints
2. Define a region around each keypoint
3. Extract and normalize the region content
4. Compute a local descriptor from the normalized region
5. Match local descriptors

Indexing local features

- Each patch / region has a descriptor, which is a point in some high-dimensional feature space

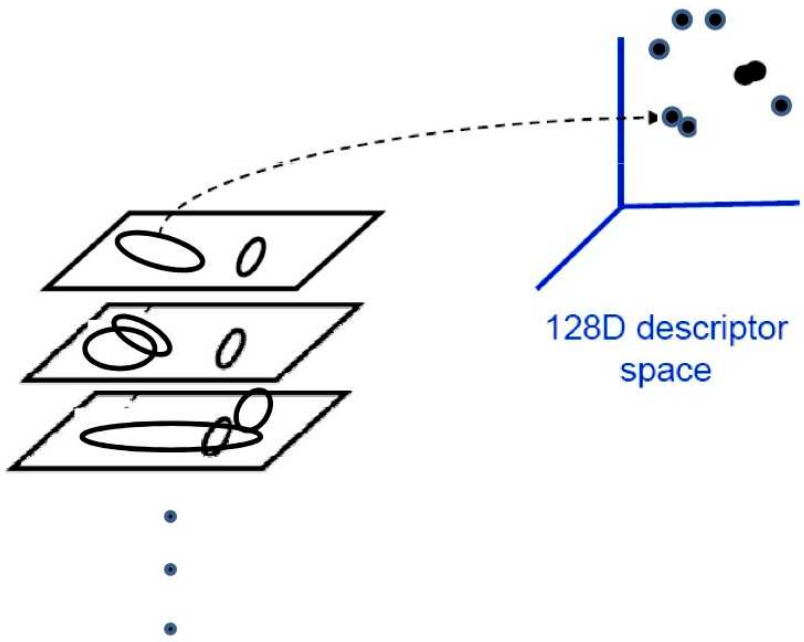


Figure: K. Grauman, B. Leibe

SIFT feature vector

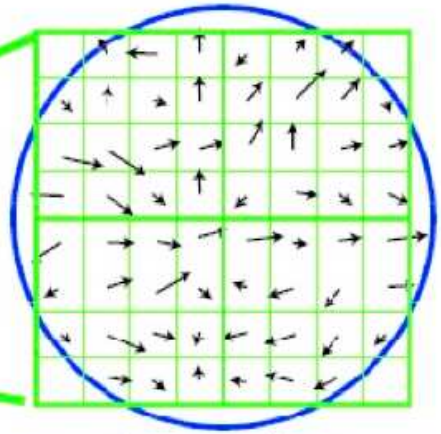
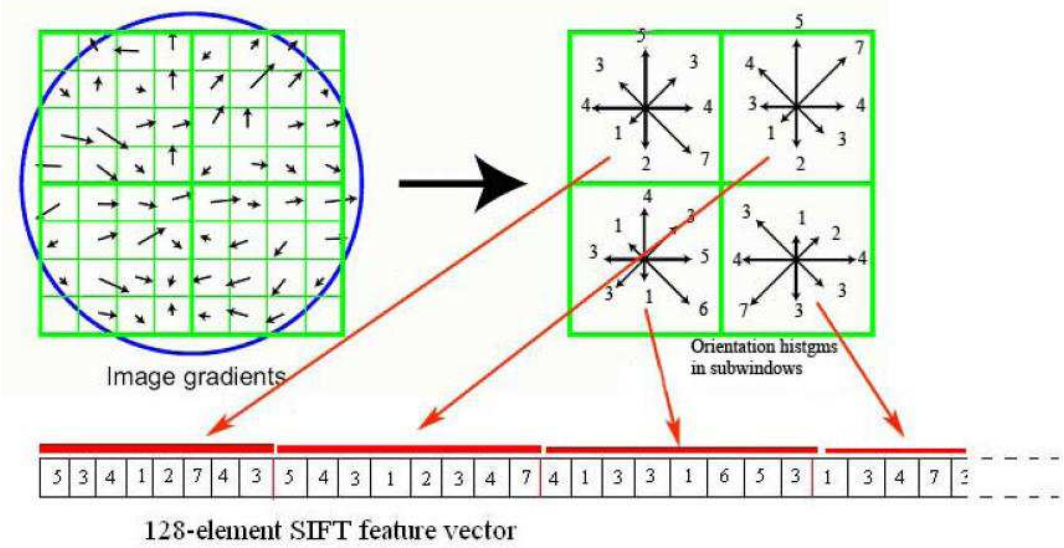


Image gradients



128-element SIFT feature vector

Clustering

- We want to perform a matching between descriptors of different images, so we need to extract types across images
- We use the vector quantization (VQ) technique which clusters the keypoint descriptors in their feature space into a large number of clusters using the K-means clustering algorithm
- Each cluster is a visual word that represents a specific local pattern shared by the keypoints in that cluster

Clustering

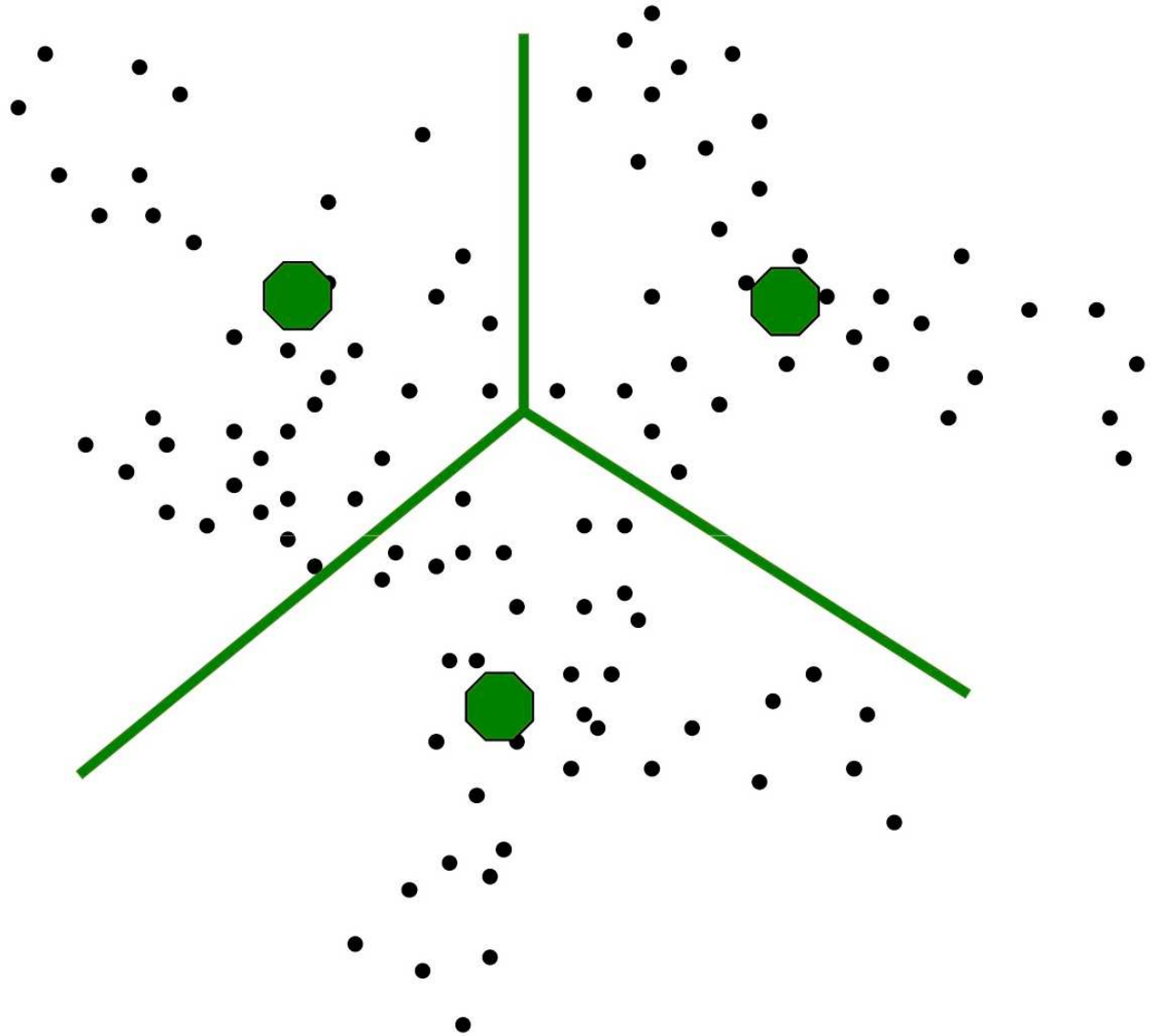


Figure: K. Grauman, B. Leibe

Visual words: main idea

- Map each high-dimensional descriptor (token) to its corresponding visual word (type) by quantizing the feature space
  - Quantize via clustering , let clusters centers be the prototype “words”

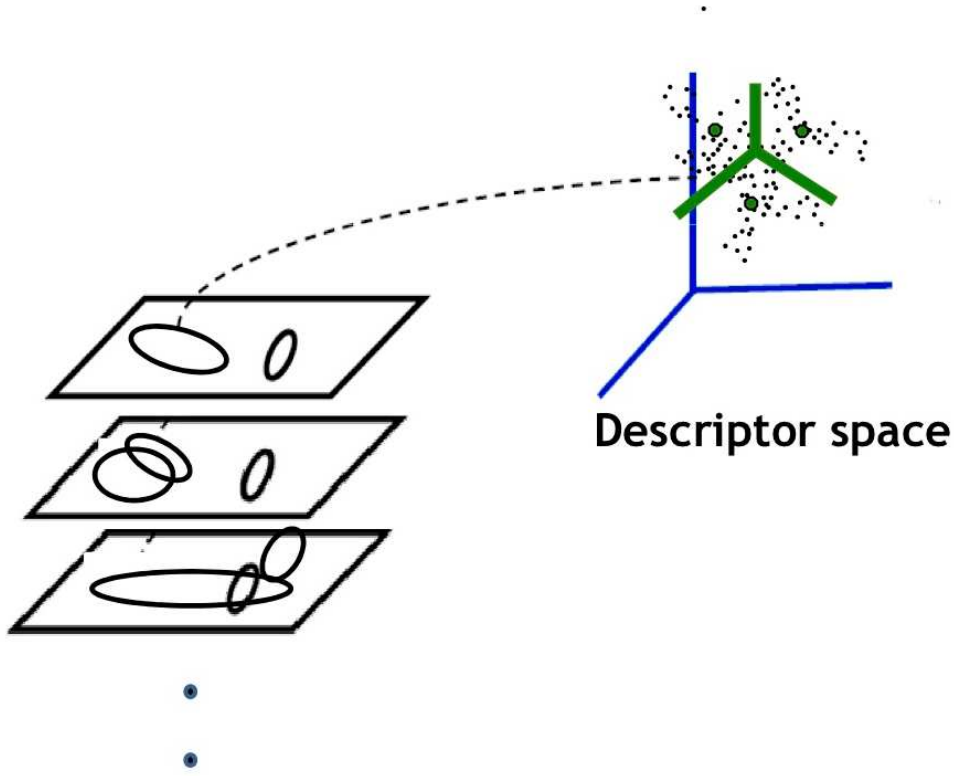


Figure: K. Grauman, B. Leibe

Visual words: main idea

- Map each high-dimensional descriptor (token) to its corresponding visual word (type) by quantizing the feature space
  - Determine which word to assign to each new image region by finding the closest cluster center

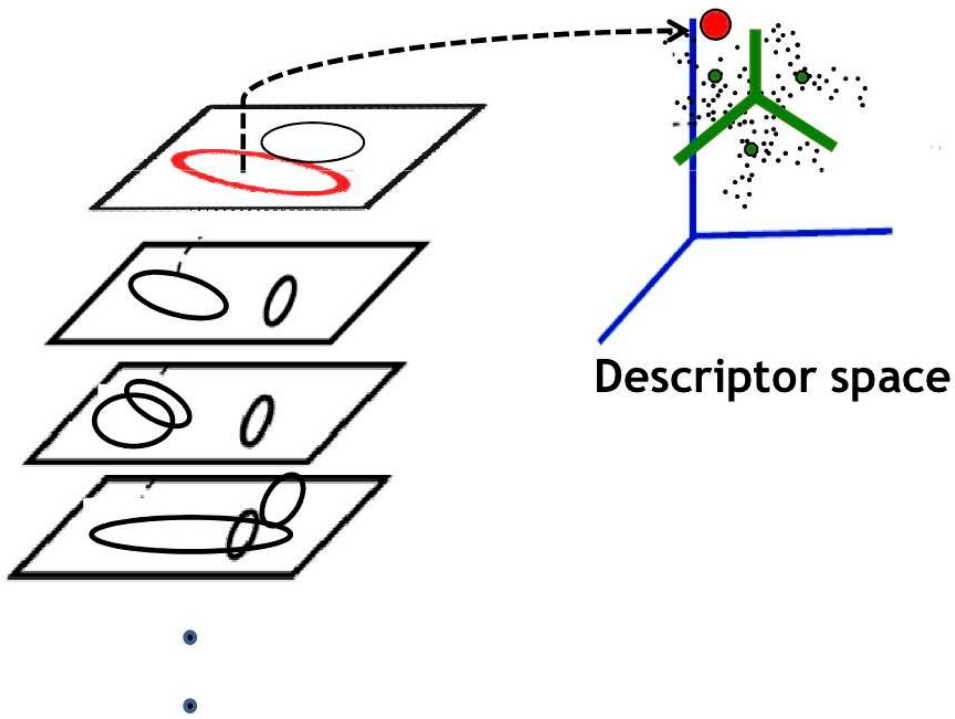


Figure: K. Grauman, B. Leibe

Recap

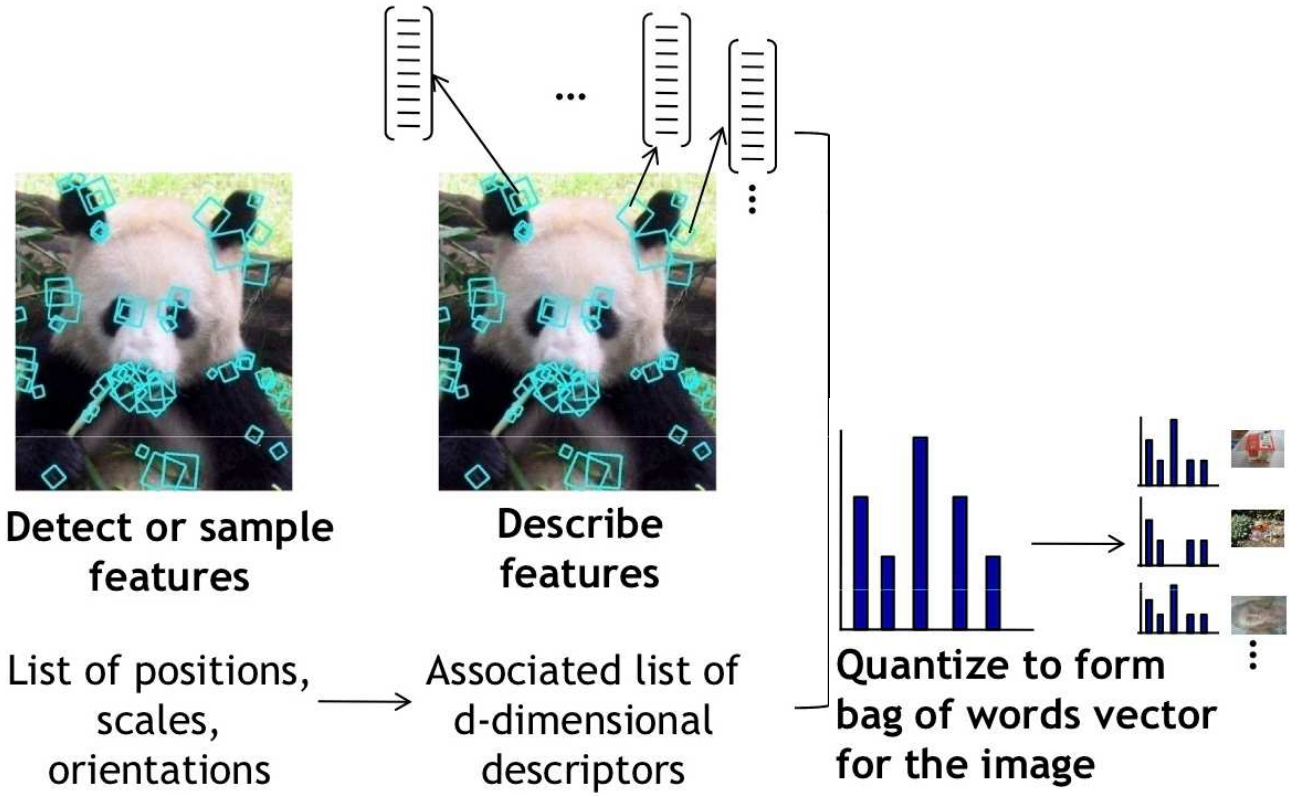


Figure: K. Grauman, B. Leibe

Outline

Bag of visual words

- Formally, each image I is expressed in a bag-of-words format vector, [v1, v2, . . . , vL], where vi = n only if I has n regions labeled with vi
- Once we have represented images as bag-of-visual-words vectors, we can say that a word appearing in proximity of an image is co-occurring with a set of visual words

Words and visual words concatenation

- Now
  - both images and documents in our corpus are now represented as bags-of-words
  - visual and textual modalities express the same content
- We can represent the document and its associated image as a mixture of textual and visual words

	leash	walk	run	owner	vw_1	vw_2	vw_3	vw_4
dog	3	5	2	5	7	3	0	4
cat	0	3	3	2	5	5	0	3
lion	0	3	2	0	5	5	3	6
light	0	0	0	0	0	0	4	0
bark	1	0	0	2	7	2	0	2
car	0	0	1	3	0	1	1	0

Outline

The Distributional Memory

- Our off-the-shelf textual distributional model
- Shown to be at the state of the art in many semantic tasks (Baroni and Lenci 2010)
- Available from: `http://clic.cimec.unitn.it/dm`
- Ask Marco for futher details!

ESP game

- Invented by L. von Ahn (2003)
- 50k labeled images
- Labeled through a game:
  - two people are partnered to label images
  - they both see the same image and the task is to agree on an appropriate word to label the image
  - once a word is entered by both partners, that word becomes a label for the image

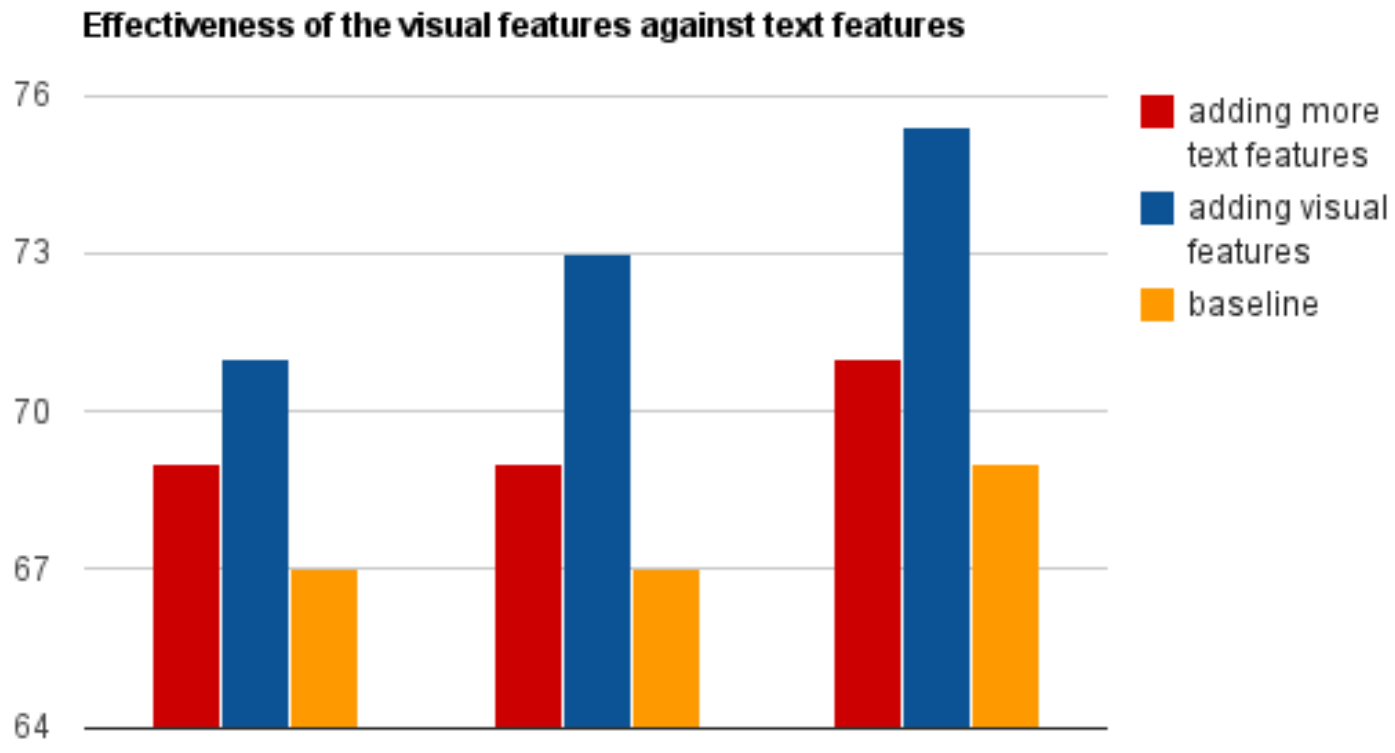
Wordsim353

- (Approximately) continuous similarity judgment
- 203 noun pairs rated by 13 subjects on a 0-10 similarity scale and averaged (our coverage 73%)
  - E.g.: *money-cash*, 9.08; *coast-hill*, 4.38; *stock-life*, 0.92
- (Spearman) correlation between cosine of angle between pair context vectors and the judgment averages

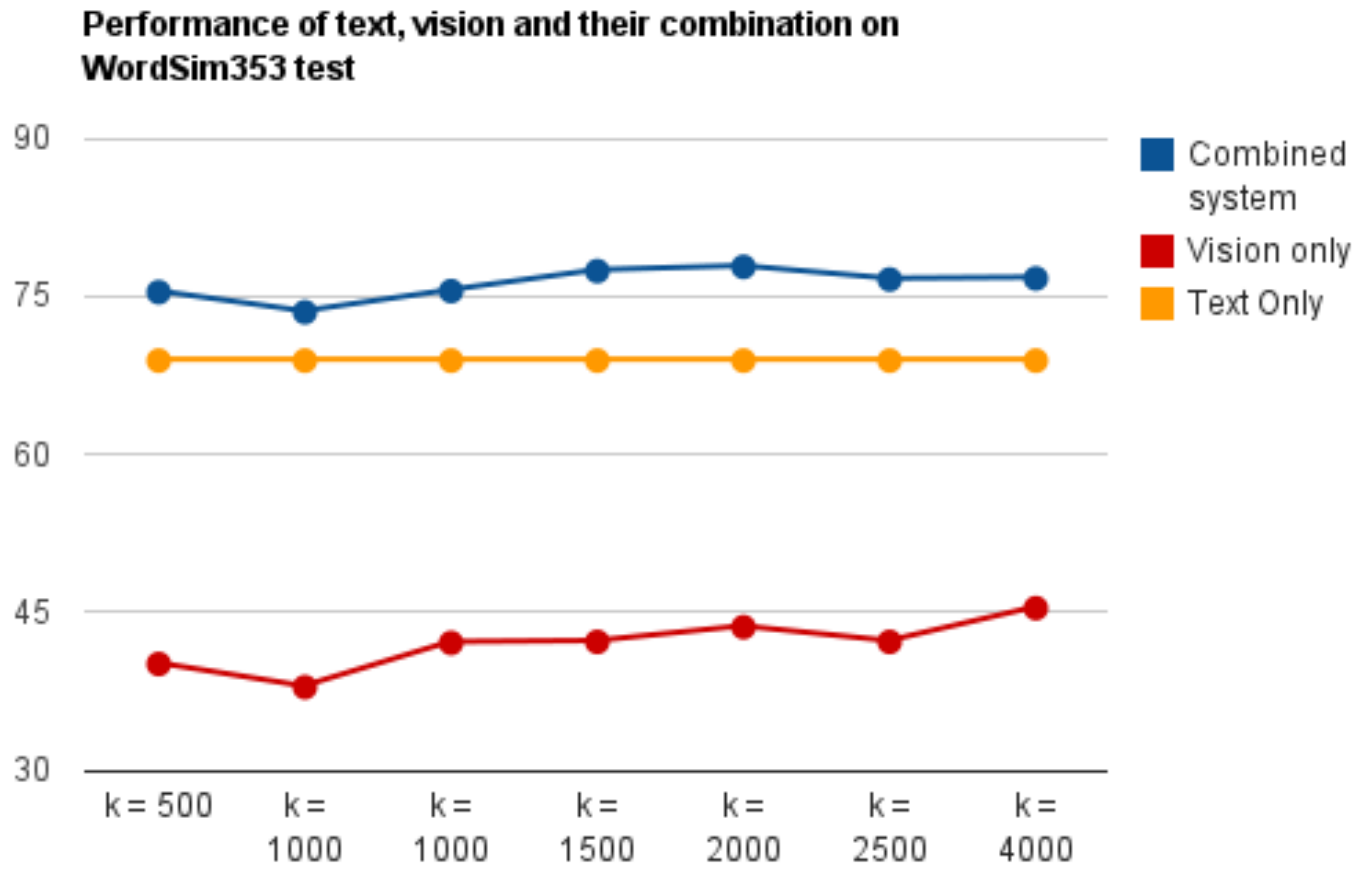
State of the art for Wordsim353

Method	Source	Spearman
(Strube and Ponzetto, 2006)	Wikipedia	0.19–0.48
(Jarmasz, 2003)	WordNet	0.33–0.35
(Jarmasz, 2003)	Roget's	0.55
(Hughes and Ramage, 2007)	WordNet	0.55
(Finkelstein et al., 2002)	Web corpus, WN	0.56
(Gabrilovich and Markovitch, 2007)	ODP	0.65
(Gabrilovich and Markovitch, 2007)	Wikipedia	0.75

Effectiveness



Performance



Coclusion

- Computer vision has moved towards discrete, visual word representations of images
- This allows seamless integration with text-corpus-based models of semantics
- What next?
  - test on more benchmarks
  - combination techinques
  - most interestingly: where and how do visual words help, exactly?

That's all!

Thank you!