Visual Features for Linguists

Basic image analysis techniques for multimodally-curious NLPers

Elia Bruni Marco Baroni

Center for Mind/Brain Sciences
University of Trento

ACL Tutorial 2013

Acknowledgments

- Partial financial support from a **Google** Research Award

- Thanks to the authors of all the online materials we liberally sampled from (see slide credits)

- Special thanks to **Jasper Uijlings**, our in-house image analysis tutor

Outline of the tutorial

1 Why image analysis?

2 Annotated image datasets

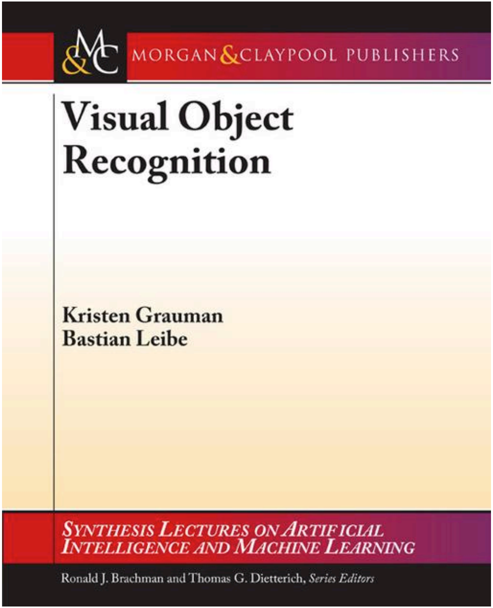
3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

Suggested reading



Outline of the tutorial

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

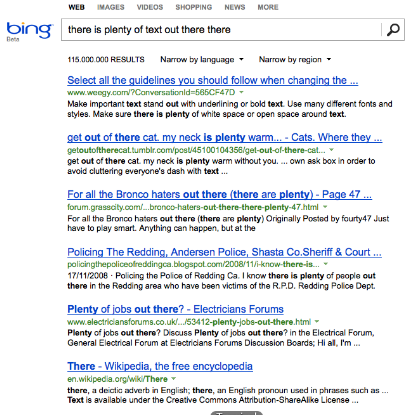
4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

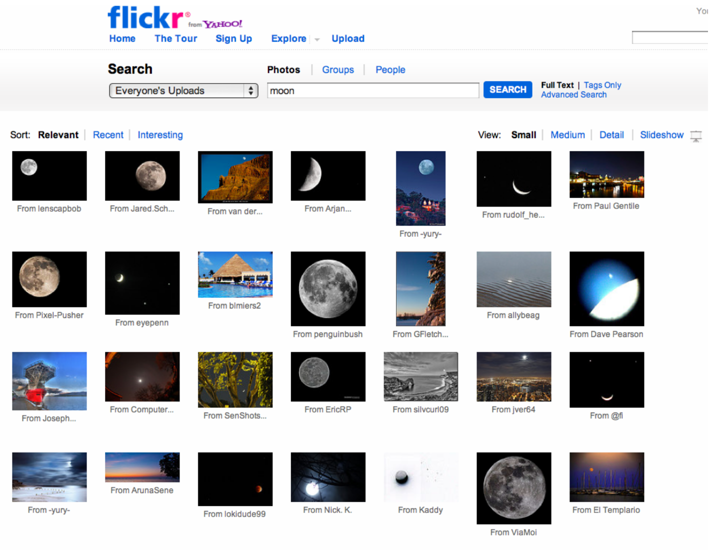
There is plenty of text out there

and we have the tools to process it




```
(ROOT
  (S
    (NP (EX There))
    (VP (VBZ is)
      (NP
        (NP (RB plenty))
        (PP (IN of)
          (NP (NN text))))
      (ADVP (IN out) (RB there)))
    (. .)))
```

There are also plenty of (labeled!) images out there!



And increasingly accurate tools to process them

. . . enabling sophisticated applications such as image stitching

And increasingly accurate tools to process them

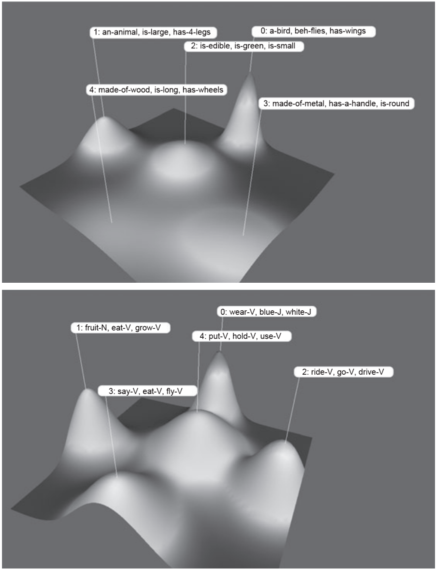
ImageNet Large Scale Visual Recognition Challenge

2012 winner performed 1K-object classification with 16% 5-guess error

Why should WE care?

- Besides all the new applications we can pursue by bringing together language and vision…
- image analysis might help us to deal with the very concrete problem of lack of grounding of linguistic symbols

Lack of grounding



Riordan and Jones, TopiCS 2011, Fig. 2

Lack of grounding

A sheep…

- According to the subject descriptions in the McRae et al.'s 2005 norms: **is white**, **has wool**, **has 4 legs**, **bahs**, …
- According to the text-generated descriptions of Baroni et al. 2010: needs a shepherd, might suffer of scrapie, grazes, in a farm, …

- Kelly et al. 2010, using large corpora, weak supervision, lexico-syntactic patterns, achieve max 24% precision, 48% recall at guessing McRae-subject-generated properties

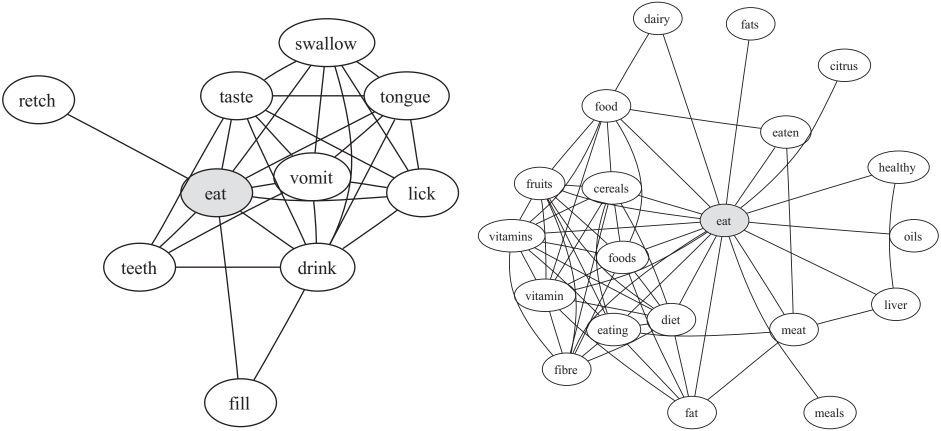
Lack of grounding

The psychedelic world of corpus-determined color

- **clover** is blue
- **coffee** is green
- **crows** are white
- **flour** is black
- **fog** is green
- **gold** is purple
- **mud** is red
- the **sky** is green
- **violins** are blue

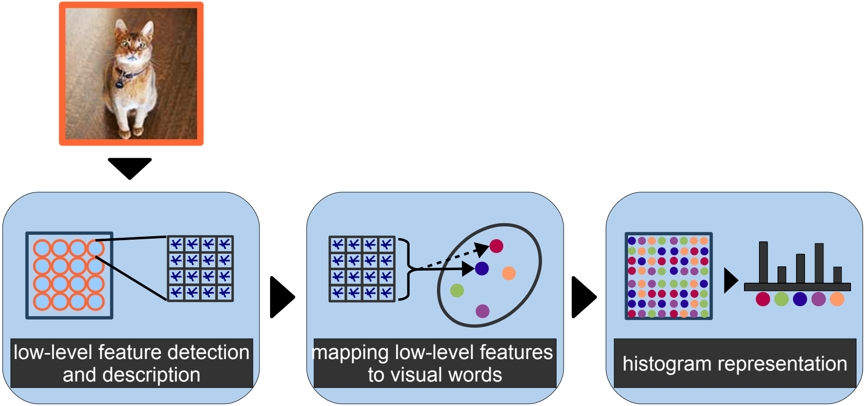
Bruni et al. ACL 2012

Lack of grounding



Andrews et al., Psych. Review 2009, Fig. 4

The image analysis pipeline



Outline of the tutorial

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

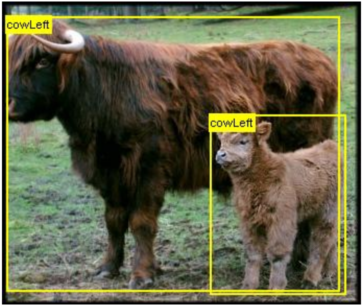
4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

The PASCAL VOC dataset

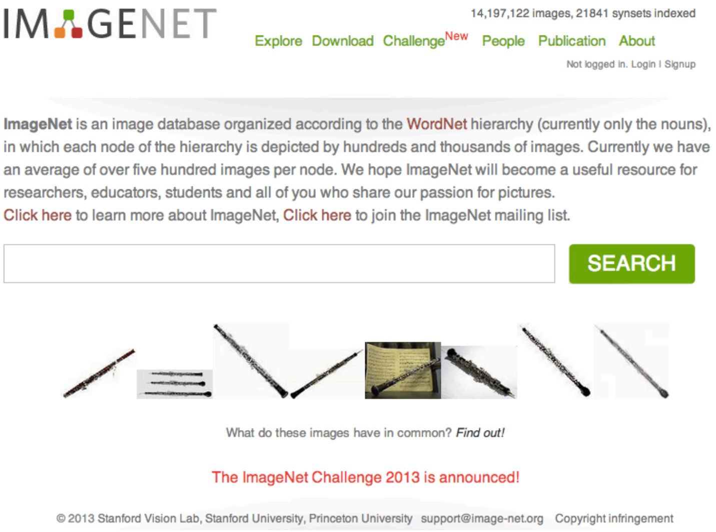
- Around 10,000 images, with around 25,000 target objects
  - Objects from 20 categories (**person**, **car**, **bicycle**, **cow**, **table**...)
  - Objects are annotated with labeled bounding boxes



Slide credit: Pedro Felzenswalb

Slide credit: Pedro Felzenswalb

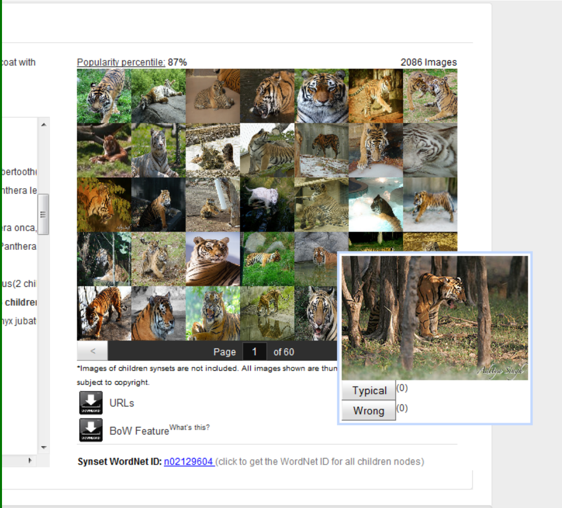
ImageNet at a glance



Slide credit: Fei-Fei Li

ImageNet at a glance

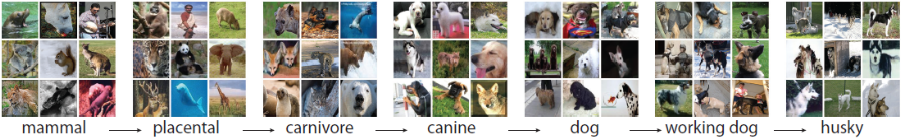
- Animals
  - Birds
  - Fish
  - Mammal
  - Invertebrate
- Scenes
  - Indoor
  - Geological formations
- Sport activities
- Materials and fabric
- Instrumentation
  - Tools
  - Appliances
  - …
- Plants
  - …



Slide credit: Fei-Fei Li

ImageNet is a knowledge database

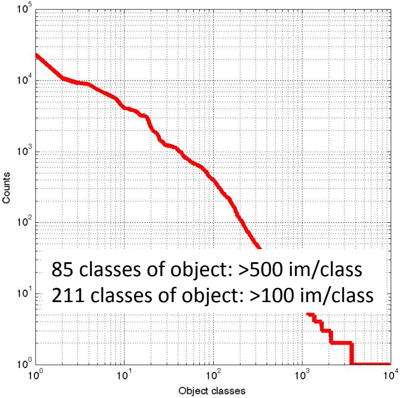
- Taxonomy on top of WordNet

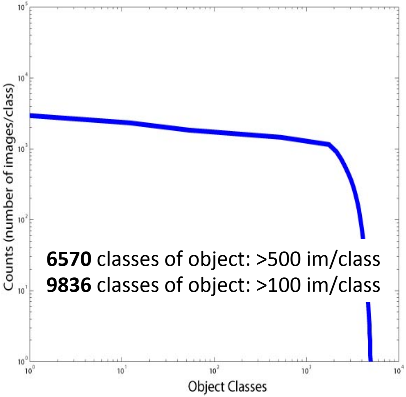


- S: (n) Eskimo dog, **husky** (breed of heavy-coated Arctic sled dog)
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) working dog (any of several breeds of usually large powerful dogs bred to work as draft animals and guard and guide dogs)
      - S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) *"the dog barked all night"*
        - S: (n) canine, canid (any of various fissiped mammals with nonretractile claws and typically long muzzles)
          - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) *"terrestrial carnivores have four or five clawed digits on each limb"*
            - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
              - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
                - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
                  - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
                    - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
                      - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
                        - S: (n) living thing, animate thing (a living (or once living) entity)
                          - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
                            - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
                              - S: (n) physical entity (an entity that has physical existence)
                                - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Slide credit: Fei-Fei Li

ImageNet is large scale





LabelMe

Russell et al. 2005;
statistics obtained in 2009

IMAGENET

Deng et al. 2009
statistics obtained in 2009

Slide credit: Fei-Fei Li

The ESP Game dataset

- Invented by Luis von Ahn (2003)
- 100K labeled images
- Labeled through a game:
  - two people are partnered together
  - both see the same image and have to agree on an appropriate word label
  - a word entered by both participants becomes a label for the image
- `http://www.cs.cmu.edu/~biglou/resources/`

Slide credit: Luis von Ahn

Grass

White

Sheep

+10

Sheep's

Sheep

Slide credit: Luis von Ahn

New image tag: **Sheep**

Slide credit: Luis von Ahn

mirror, mud, white, person, stuck, car, jeep, door, tire, wheel

triangle, pink, building, tower, square, towers

band, sing, hair, arm, singer, man, guitar, mic, microphone

desert, soldier, army, man

coin, round, money, face, gold, old, man

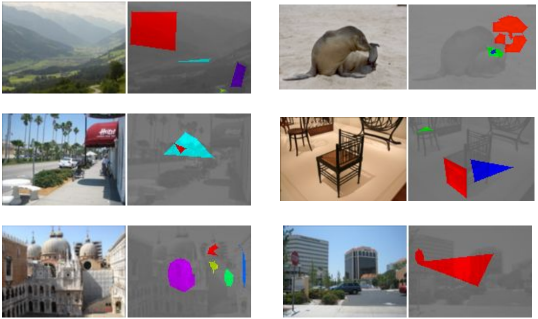
imagine, in-depth, depth, uro, in, reports, more, euro

Slide credit: Antonio Torralba

Polygon quality

Slide credit: Antonio Torralba

Not all data is reliable

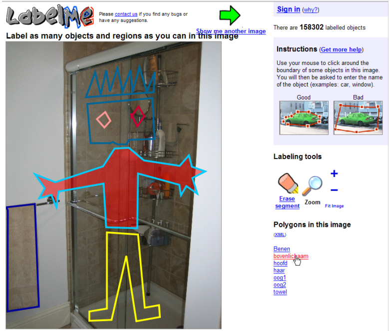


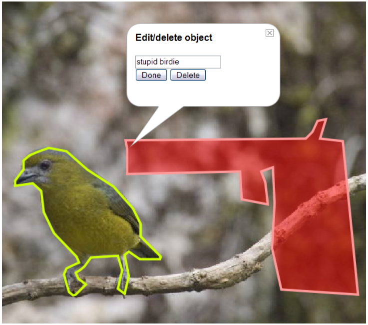
Most common labels:

test
adksdsa
woiieiie
...

Slide credit: Antonio Torralba

Online hooligans





Slide credit: Antonio Torralba

More on annotated image datasets

- L. von Ahn and L. Dabbish. 2004. Labeling Images with a Computer Game. *Proceedings of CHI*.
- B.C. Russell, A. Torralba, K.P. Murphy and W.T. Freeman. 2008. LabelMe: A Database and Web-Based Tool for Image Annotation. *IJCV*.
- J. Deng, W. Dong, R. Socher, L.J. Li and L. Fei Fei. 2009. Imagenet: A large-scale hierarchical image database. *Proceedings of CVPR*.
- M. Everingham, L. van Gool, C.K.I. Williams, J. Winn and A. Zisserman. 2010. The PASCAL Visual Object Classes (VOC) Challenge. *IJCV*.

Outline of the tutorial

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

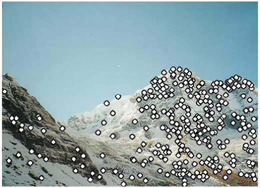
4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

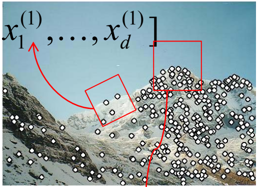
Local features: main components

- Detection: Identify the interest points



- Description: Extract feature descriptor surrounding each interest point

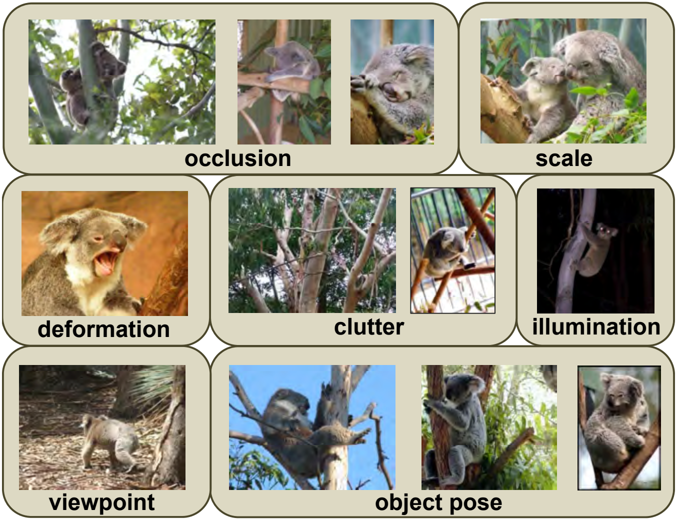
$$\mathbf{x}_1 = [x_1^{(1)}, \ldots, x_d^{(1)}]$$



$$\mathbf{x}_2 = [x_1^{(2)}, \ldots, x_d^{(2)}]$$

Slide credit: Kristen Grauman

Local features: challenges



Slide credit: Grauman and Leibe

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images
Feature detectors
Feature descriptors

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

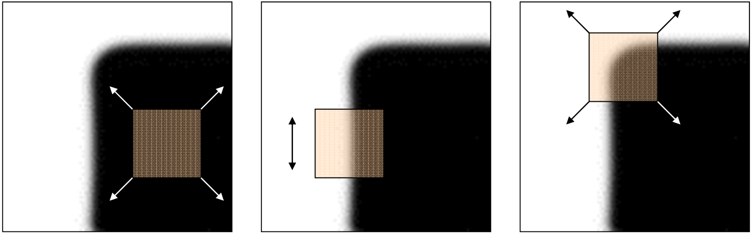
Local detectors

- Goals: Repeatable detection, precise localization, interesting content

Slide credit: Kristen Grauman

Corners as distinctive interest points

- Design criteria
  - We should easily recognize the point by looking through a small window (*locality*)
  - Shifting the window in *any direction* should give *a large change* in intensity (*good localization*)



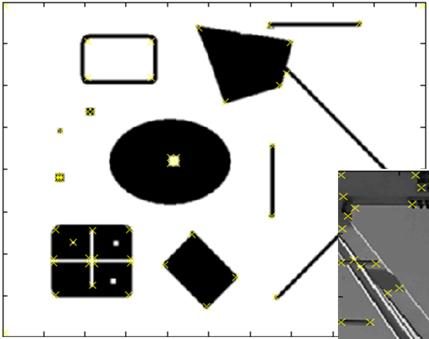
"flat" region:
no change in all directions

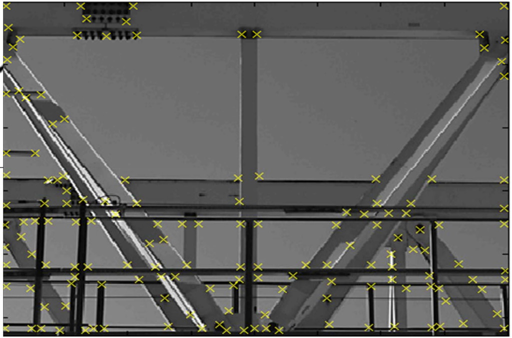
"edge":
no change along the edge direction

"corner":
significant change in all directions

Slide credit: Alyosha Efros

Harris Detector: Responses

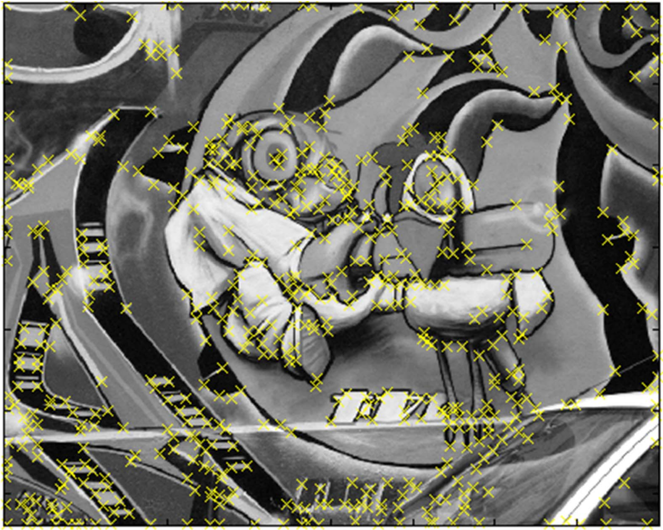




Effect: **A very precise corner detector.**

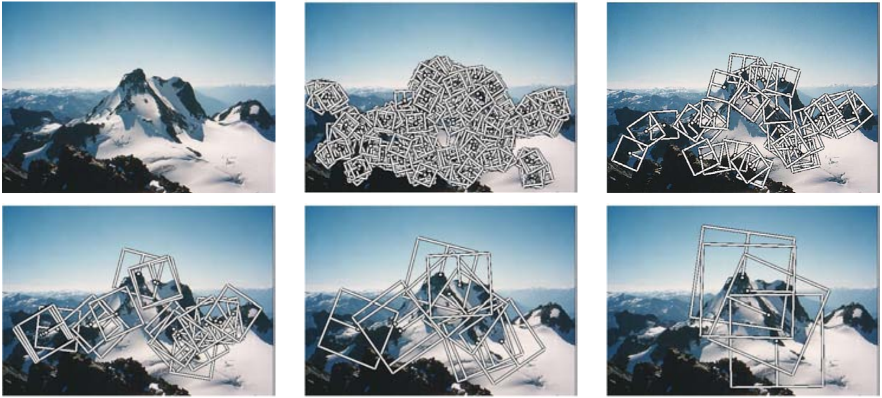
Slide credit: Krystian Mikolajczyk

Harris Detector: Responses



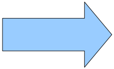
Slide credit: Krystian Mikolajczyk

From keypoints to regions



Slide credit: Rick Szeliski

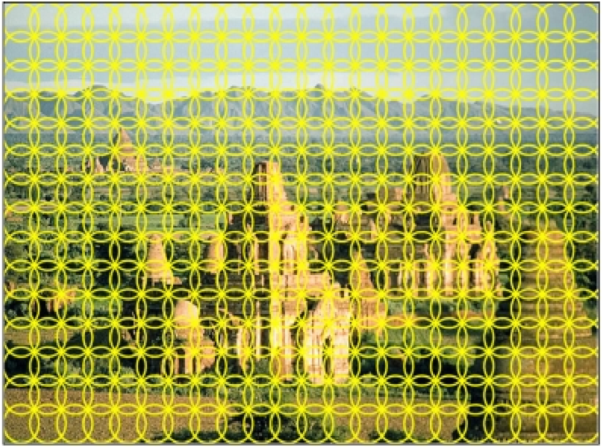
From keypoints (Harris) to regions (DoG)



Slide credit: Bastian Leibe

Recent advances

Dense feature extraction [Nowak et al. 2006]



1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

Feature detectors

Feature descriptors

4 Visual words for higher-level image representation

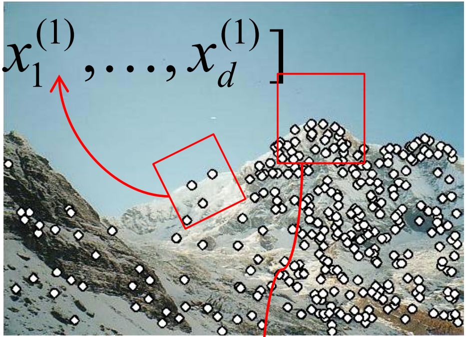
5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

Local descriptors

- We know how to detect points
- Next question: How to describe them?

$$\mathbf{x}_1 = [x_1^{(1)}, \ldots, x_d^{(1)}]$$

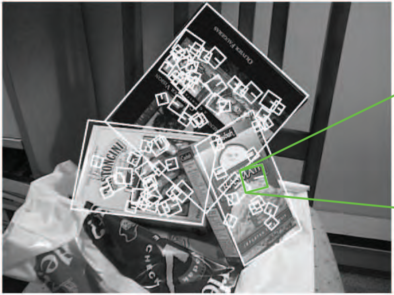


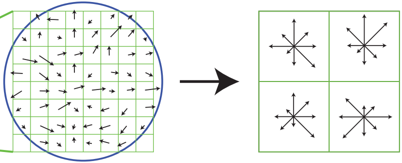
$$\mathbf{x}_2 = [x_1^{(2)}, \ldots, x_d^{(2)}]$$

Slide credit: Kristen Grauman

SIFT descriptor [Lowe 1999]

- Scale Invariant Feature Ttransform
- Descriptor computation:
  - Divide regions into 4x4 cells
  - Compute histogram of gradient orientations (8 reference angles) for all pixels inside each sub-patch - Gradient = directional change in the intensity or color in an image
  - Resulting descriptor: 4x4x8 = 128 dimensions



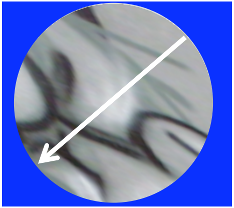


Slide credit: David Lowe

SIFT descriptor - rotation invariance

- Estimation of the dominant direction
  - Extract gradient orientation
  - Histogram over gradient orientation
  - Peak in this histogram

- Rotate patch in dominant direction



Slide credit: Cordelia Schmid

Many more descriptors than just SIFT

- SIFT[Lowe ’99]
  - Color SIFT [Sande et al. 2010]
  - Normalizing SIFT with square root transformation [Arandjelovic et Zisserman 2012]
- Textons [Leung and Malik ’01]
- HoG [Dalal and Triggs ’05]
- SURF [Bay et al. ’08]
- DAISY [Tolaetal. ’08, Windleretal ’09]
- Bag of (e.g., LAB) colors [Farhadi et al. ’09]

More on low-level features

- C. Harris and M. Stephens. 1988. A combined corner and edge detector. *Proceedings of Alvey Vision Conference*.
- D. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *IJCV*.
- K. Mikolajczyk and C. Schmid. 2005. A performance evaluation of local descriptors. *PAMI*.
- E. Nowak,F. Jurie and B. Triggs. 2006. Sampling strategies for bag-of-features image classification. *Proceedings of ECCV*.
- K.E.A. van de Sande, T. Gevers and C.G.M. Snoek. 2010. Evaluating color descriptors for object and scene recognition. *PAMI*.
- R. Arandjelovic and A. Zisserman. 2012. Three things everyone should know to improve object retrieval. *Proceedings of CVPR*.

Outline of the tutorial

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation
- Constructing a vocabulary of visual words
- Classic Bags-of-visual-words representation
- Recent advances

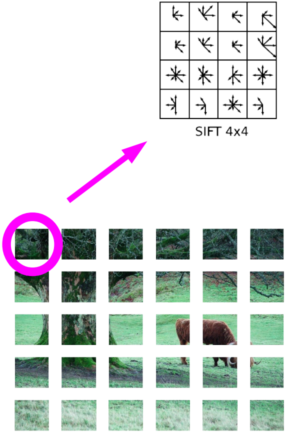
5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

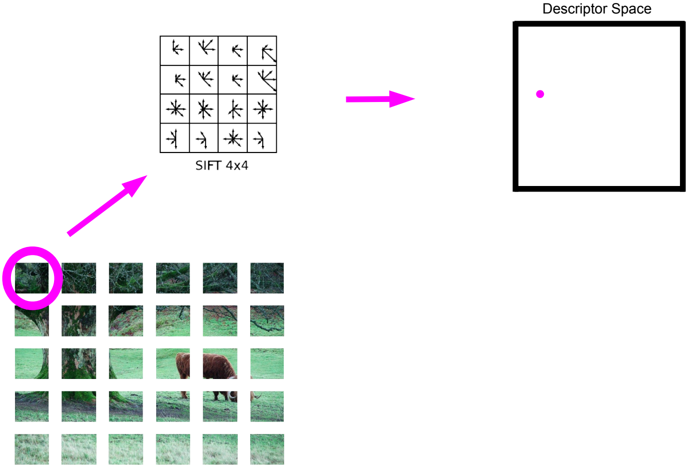
BoVW: Feature extraction

BoVW: Feature extraction

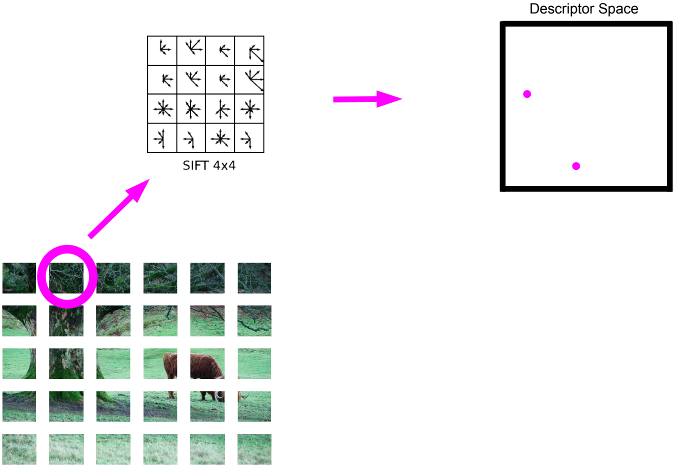
BoVW: Feature extraction



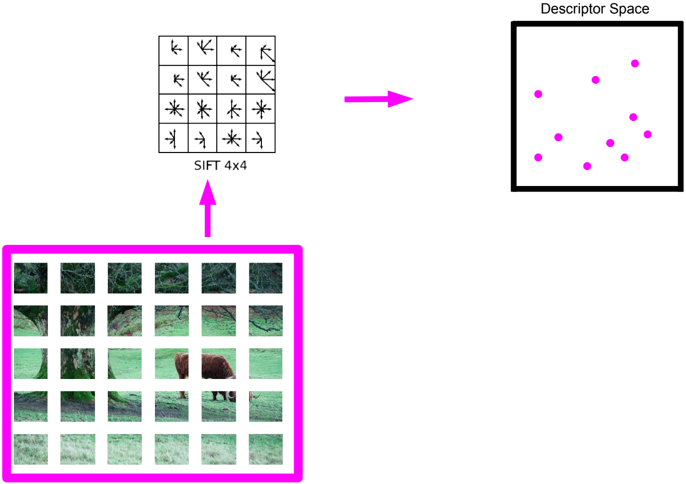
BoVW: Feature extraction



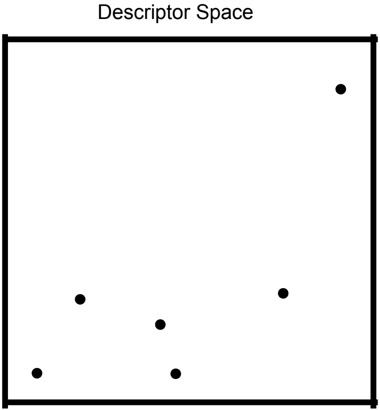
BoVW: Feature extraction



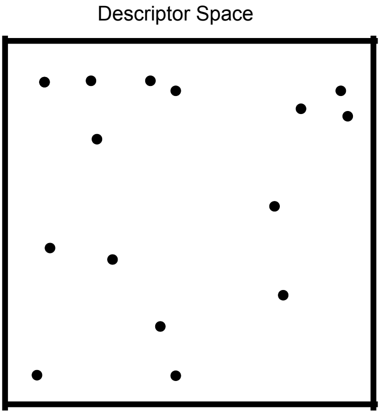
BoVW: Feature extraction



BoVW: Feature extraction

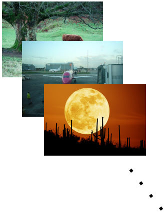


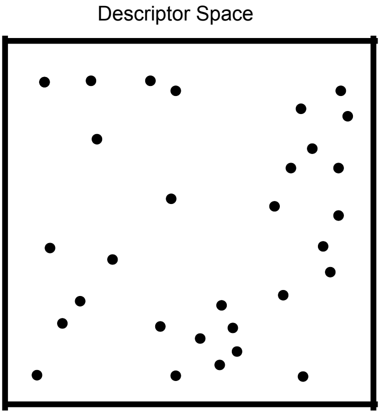
BoVW: Visual dictionary



The clusters partition the descriptor space. Each cluster is called a “Visual Word”.

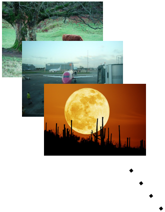
BoVW: Visual dictionary

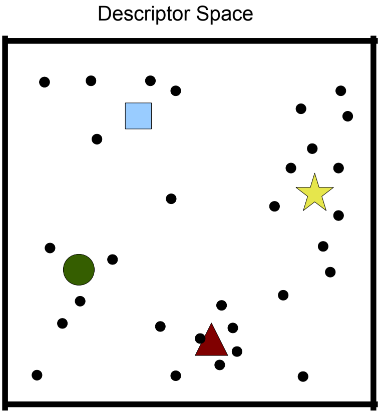




The clusters partition the descriptor space. Each cluster is called a “Visual Word”.

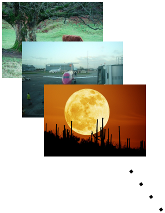
BoVW: Visual dictionary

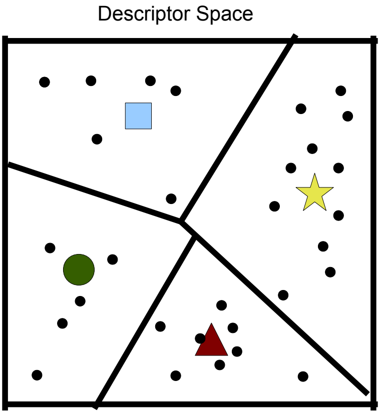




The clusters partition the descriptor space. Each cluster is called a “Visual Word”.

BoVW: Visual dictionary





The clusters partition the descriptor space. Each cluster is called a “Visual Word”.

1 Why image analysis?

2 Annotated image datasets

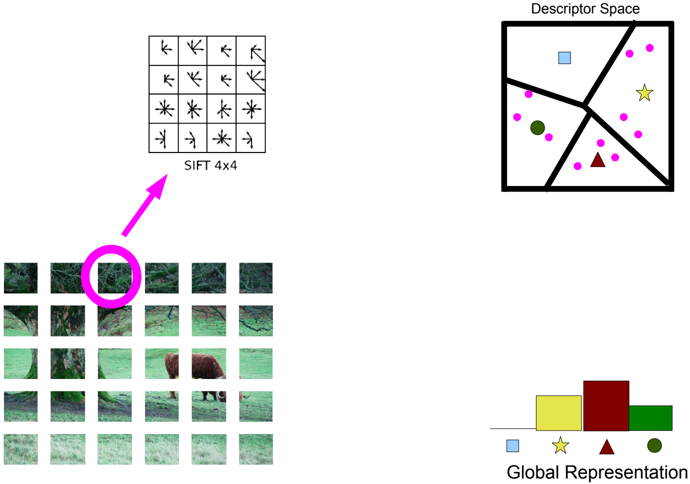
3 Extraction of low-level features from images

4 Visual words for higher-level image representation
- Constructing a vocabulary of visual words
- Classic Bags-of-visual-words representation
- Recent advances

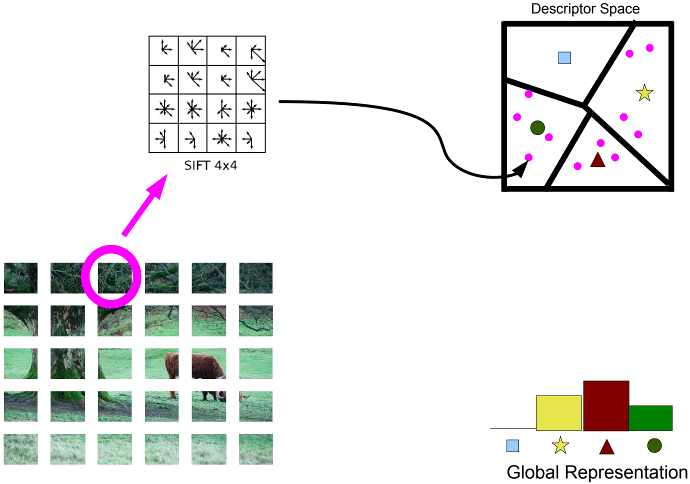
5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

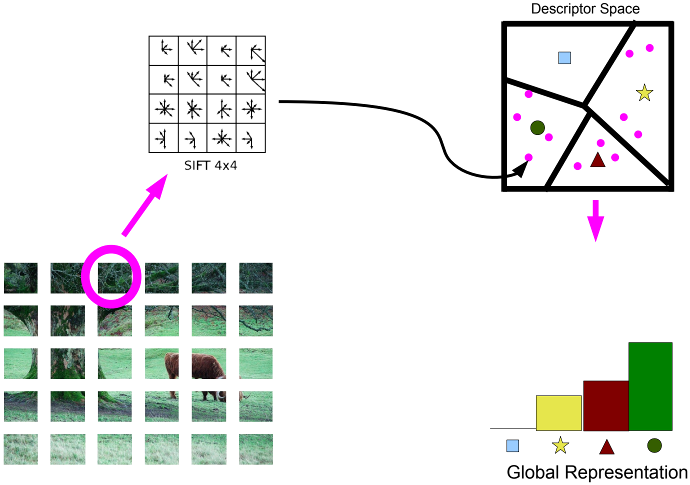
BoVW: Image representation



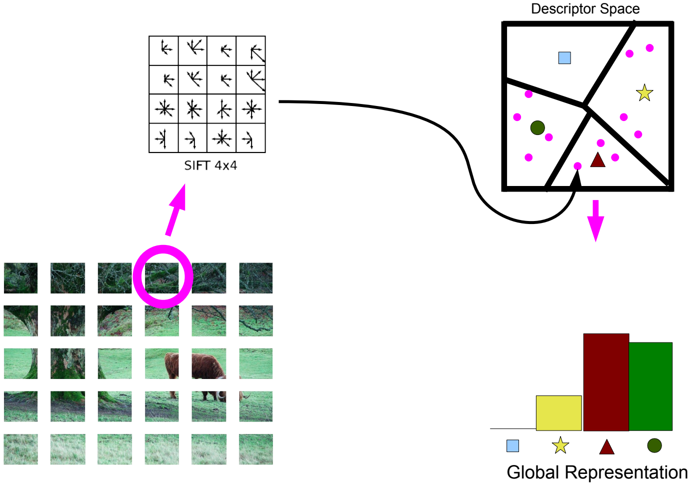
BoVW: Image representation



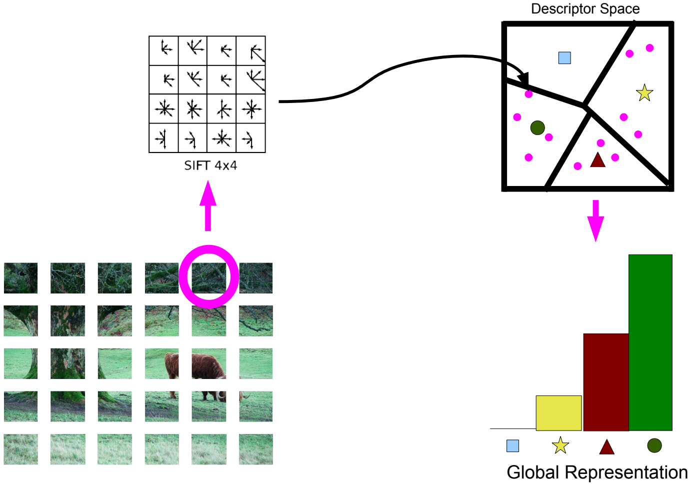
BoVW: Image representation



BoVW: Image representation



BoVW: Image representation



1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

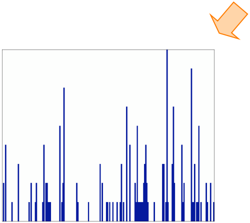
4 Visual words for higher-level image representation
- Constructing a vocabulary of visual words
- Classic Bags-of-visual-words representation
- Recent advances

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

Spatial pyramid representation [Lazebnik et al. 2006]

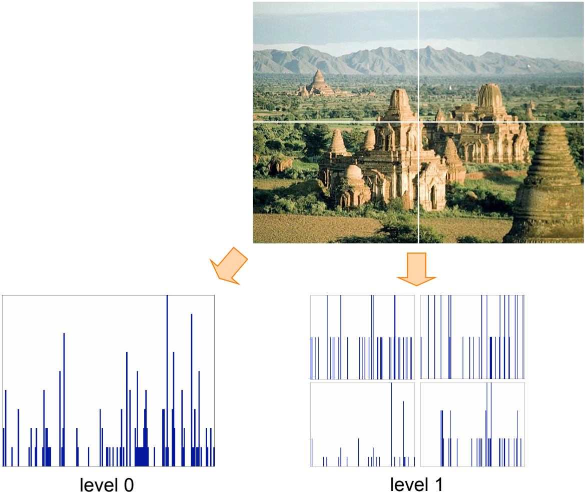
Locally orderless representation at several levels of spatial resolution



level 0

Slide credit: Svetlana Lazebnik

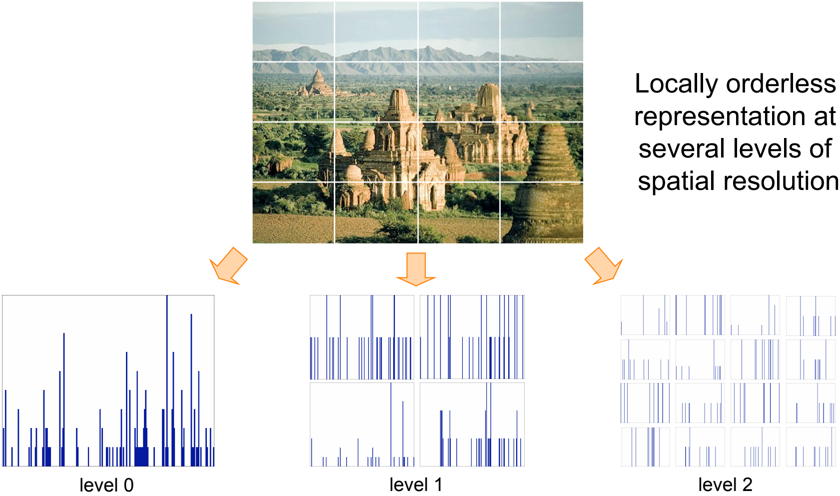
Spatial pyramid representation



Locally orderless representation at several levels of spatial resolution

Slide credit: Svetlana Lazebnik

Spatial pyramid representation

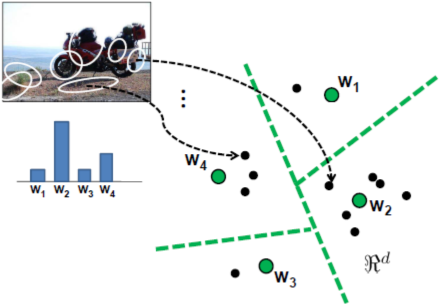


Locally orderless representation at several levels of spatial resolution

Slide credit: Svetlana Lazebnik

Fisher encoding [Perronnin et al. 2010]

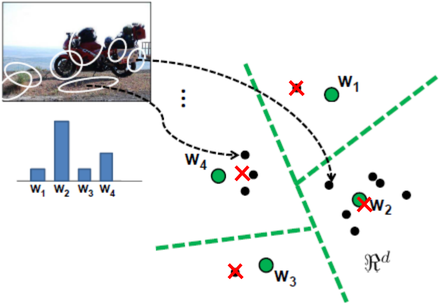
- BoVW is only about **counting** the number of local descriptors assigned to each region
- Why not including other statistics?



Slide credit: Kristen Grauman

Fisher encoding

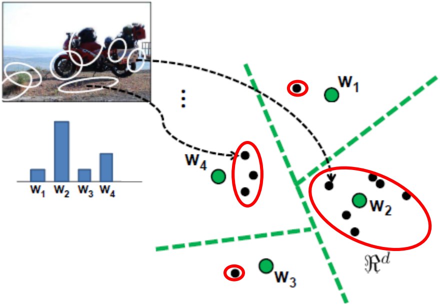
- BoVW is is only about **counting** the number of local descriptors assigned to each region
- Mean of local descriptors



Slide credit: Kristen Grauman

Fisher encoding

- BoVW is is only about **counting** the number of local descriptors assigned to each region
- Variance of local descriptors



Slide credit: Kristen Grauman

More on bag-of-visual-words

- J. Sivic and A. Zisserman 2003. Video Google: A text retrieval approach to object matching in videos. *Proceedings of ICCV*.
- G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray. 2004. Visual categorization with bags of keypoints. *Proceedings of ECCV*.
- S. Lazebnik, C. Schmid and J. Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proceedings of CVPR*.
- F. Perronnin, J. Sanchez and T. Mensink. 2010. Improving the fisher kernel for large-scale image classification. *Proceedings of ECCV*.

Outline of the tutorial

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision
- Emotion analysis
- Object recognition
- Visual attributes

6 Going multimodal: Visual features in NLP

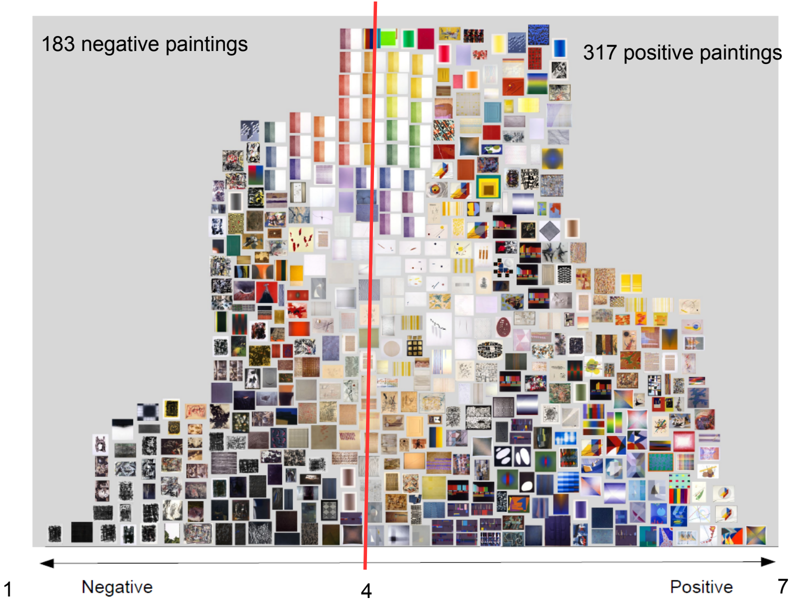
Emotion analysis: Images evoke emotions



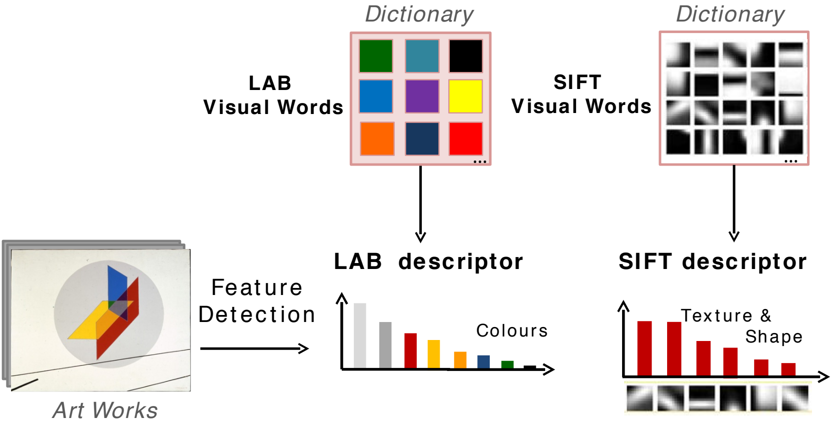



Slide credit: Yanulevskaya et al. 2012

Emotion analysis: The dataset



Emotion analysis: The method



Slide credit: Yanulevskaya et al. 2012

Emotion analysis: Results

- Task: Divide the dataset into train and test and use a classifier (SVM) to distinguish between positive and negative paintings

Accuracy:

0.63 Random

0.76 LAB visual words

0.73 SIFT visual words

0.78 LAB + SIFT

Slide credit: Yanulevskaya et al. 2012

More on emotion analysis

- J. Machajdik and A. Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. *Proceedings of ACMM*.
- V. Yanulevskaya, J. Uijlings, E. Bruni, E. Zamboni, F. Bacci, D. Melcher and N. Sebe. 2012. In the Eye of the Beholder: Employing Statistical Analysis and Eye Tracking for Analyzing Abstract Paintings. *Proceedings of ACMM*.

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

- Emotion analysis
- Object recognition
- Visual attributes

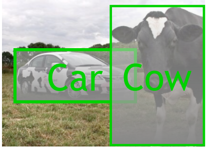
6 Going multimodal: Visual features in NLP

Object recognition

- Image classification: assigning a class label to the image

Car: present
Cow: present
Bike: not present
Horse: not present
...

- Object localization: define the location and the category

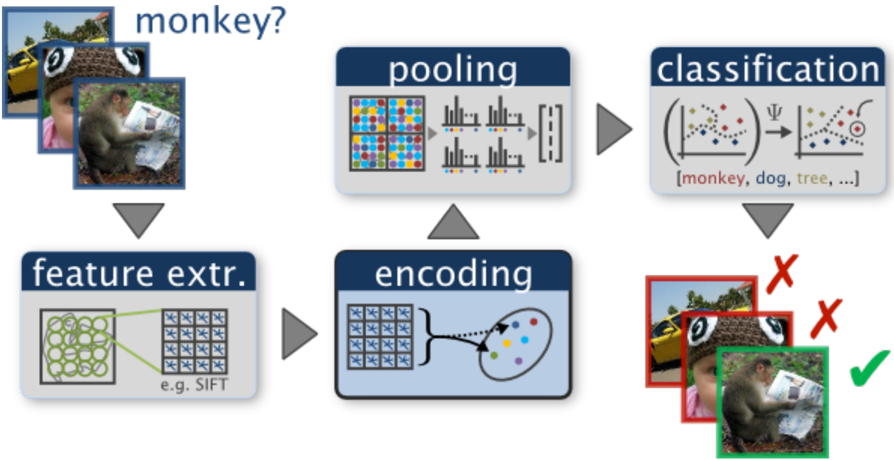


Location

Category

Slide credit: Cordelia Schmid

Object recognition: Typical pipeline



Slide credit: Chatfield et al. 2011

Pascal VOC 2012: Statistics

	Training	Testing
Images	11,540	10,994
Objects	27,450	27,078

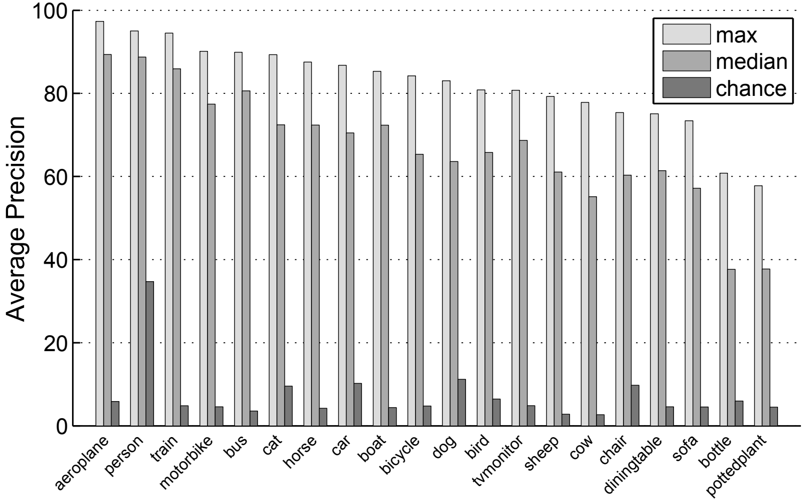
- 20 categories
- Minimally, around 600 training objects per category
- Around 2,000 cars, 1,500 dogs and 8,500 people
- Approximately equal distribution across training and test datasets

Slide credit: Andrew Zisserman

Pascal VOC 2012: Submitted systems

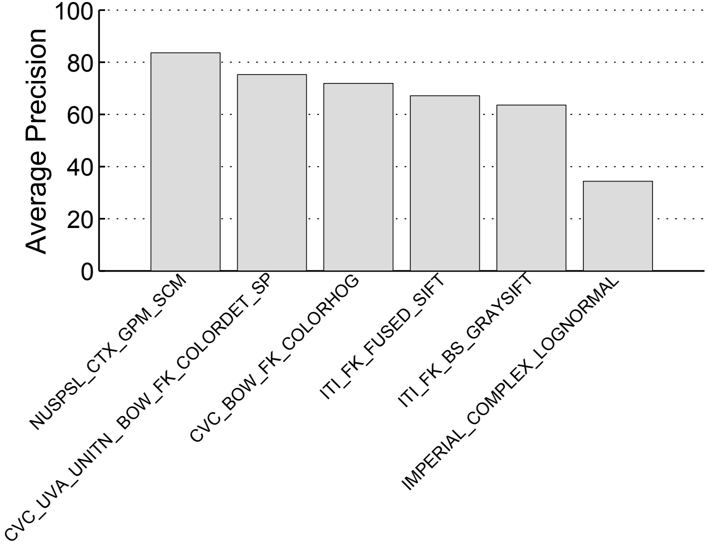
- 7 systems, 5 groups
- Methods
  - Features: Dense SIFT, HoG, colour
  - Encodings: spatial pyramid, BoVW, Fisher vector
  - Classifier: SVM

Pascal VOC 2012: Results



Aeroplanes vs. bottles

Pascal VOC 2012: Results



More on object recognition

- J. Yang, J.G. Jiang, A. Hauptmann and C.W. Ngo. 2007. Evaluating bag-of-visual-words representations in scene classification. *MIR*.
- Results of the Pascal VOC Challenge 2012: `http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/index.html`

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

- Emotion analysis
- Object recognition
- Visual attributes

6 Going multimodal: Visual features in NLP

Visual attributes

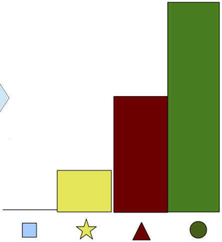
A. Farhadi, I. Endres, D. Hoiem and D. Forsyth. 2009. Describing objects by their attributes. *Proceedings of CVPR*.

Object-centric classification

Input Image

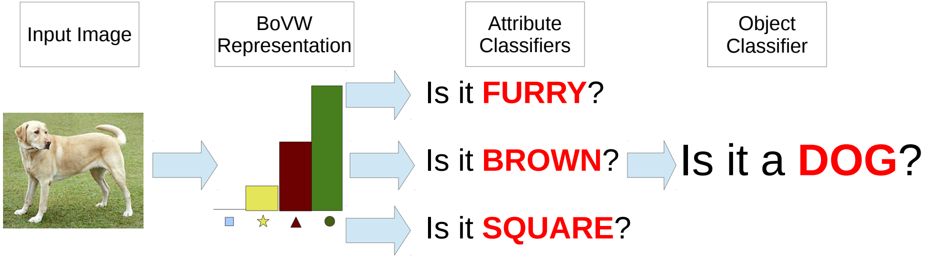
BoVW Representation

Object Classifier

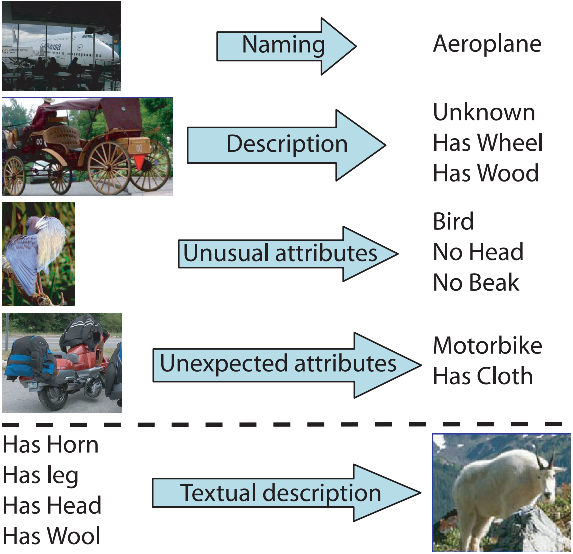


Is it a **DOG**?

Attribute-centric classification



The many uses of attributes



Farhadi et al., CVPR 2009, Fig. 1

Data sets

http://vision.cs.uiuc.edu/attributes/

a-Pascal Pascal VOC 2008 data set, 20 categories (people, dog, bus, horse…), 150-to-1K images per category (5K for people), used for training and testing

a-Yahoo 12 categories similar to those in a-Pascal but different (statue, wolf, carriage, centaur…), used for testing

Base features

- BoVW spatial pyramid histograms from bounding boxes using:
  - HoG descriptors (good for parts)
  - Canny Edges (Canny, 1986; good for shapes)
  - Textons (good for materials)
  - LAB color features (good for materials)
- 9751-dimensional vectors (mostly HoG features)

Attributes

- 64 AmazonTurk-annotated “semantic” attributes:

  Shapes: is 2D boxy, is 3D boxy, is cylindrical…

  Parts: has head, has leg, has window…

  Materials: is furry, has glass, is shiny…

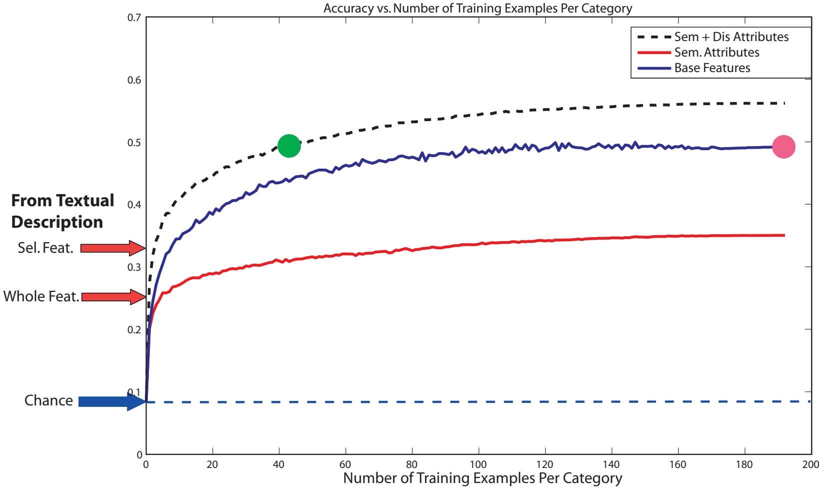
- 1K automatically selected “discriminative” attributes
  - Selected to maximize discrimination between random subsets of classes or attributes

Learning

- Base feature selection by picking features that are best at within-category attribute learning
  - If you learn *has wheels* from cars, motorbikes, buses vs. horses, cats, bottles, you might learn *metallic* instead!
- Object classifiers trained on base features (standard approach) or on vectors of automatically assigned attributes
- Various learning algorithms used: linear SVMs, regularized logistic regression, nearest neighbour classification
- See the paper for details and more experiments

Identifying “new” objects

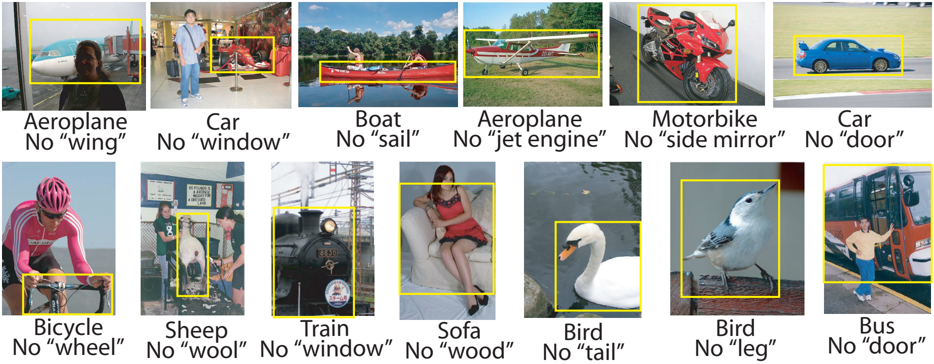
Farhadi et al., CVPR 2009, Fig. 9



Attribute classifiers trained on a-Pascal, predicted attributes used as features to train a-Yahoo object classifier (1NN)

Reporting missing attributes

68.2% accuracy



Farhadi et al., CVPR 2009, Fig. 6

Reporting atypical attributes

47.3% accuracy

Farhadi et al., CVPR 2009, Fig. 7

More on attributes

- C.H. Lampert, H. Nickisch and S. Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. *Proceedings of CVPR*.
- A. Farhadi, I. Endres and D. Hoiem. 2010. Attribute-centric recognition for cross-category generalization. *Proceedings of CVPR*.
- T.L. Berg, A.C. Berg and J. Shih. 2010. Automatic attribute discovery and characterization from noisy Web data. *Proceedings of ECCV*.
- D. Parikh and K. Grauman. 2011. Relative attributes. *Proceedings of ICCV*. Best paper award.
- Tutorial on attributes at CVPR, June 2013: `http://filebox.ece.vt.edu/~parikh/attributes/`
- C. Silberer, V. Ferrari and M. Lapata. Models of semantic representation with visual attributes. ACL 2013!

Outline of the tutorial

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

Generating image descriptions

Multimodal distributional semantics

Visual selectional preference

Generating image descriptions

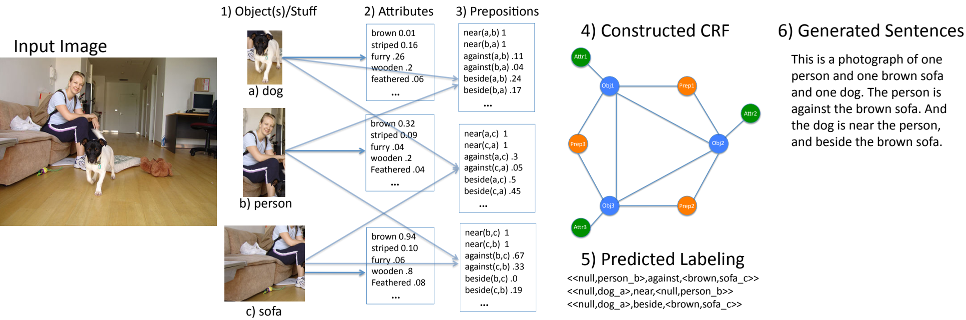
G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg and T. Berg. 2011. Baby Talk: Understanding and generating simple image descriptions. *Proceedings of CVPR*

We need image description *generation* (not just *retrieval*)

Figure 1. Our system automatically generates the following descriptive text for this example image: *"This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair. The green grass is by the chair, and near the potted plant."*

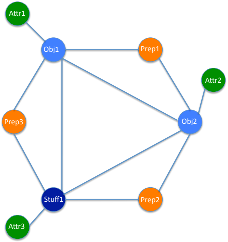
Kulkarni et al., 2011, Fig. 1

The BabyTalk pipeline



Kulkarni et al., 2011, Fig. 2

Generating a labeled graph describing image contents



- Images represented by graphs with nodes labeled with *objects* and *stuff*, *attributes* of objects/stuff and *prepositions* connecting objects/stuff
- Label assignment casted as Conditional Random Field energy minimization problem with both image- and text-based features
- CRF and image-based feature parameters trained on 153 images with (up-to-5) sentence descriptions from UIUC PASCAL data set

Graph from Kulkarni et al., 2011, Fig. 3

Image-derived features

Object detectors Based on Felzenszwalb et al.'s *deformable part models*, trained for 24 categories on PASCAL 2010 and ImageNet data

Stuff detectors Based on Farhadi et al.'s basic features, SVM trained to recognize sky, road, building, tree, water and grass on ImageNet

Attribute classifiers SVM trained on Flickr, Google, ImageNet and Farhadi et al.'s images with attribute labels, 21 attributes (blue, furry, wooden, shiny. . .)

Preposition functions Hand-coded heuristics, e.g., $near(a, b)$ value is given by normalized minimum distance between the regions of a and b

Text-derived features

- Text features provide *smoothing* of image-based hypotheses about objects, their attributes and preposition-expressed relations
- Co-occurrence counts for *attribute-object|stuff* and *object|stuff-preposition-object|stuff*
- Linearly combined counts from parsed Flickr image description corpus and Google

Generation

- From graph labels to sentences
  - E.g., from *(white cloud) in (blue sky)* to *There is a white cloud in the blue sky*
- Two simple generation strategies
  - N-gram language model used only to insert function words between nouns, adjectives and prepositions in the graph labels
  - Hand-coded templates

Language model decoding vs. templates

Templated Generation: This is a photograph of one furry sheep.
Simple Decoding: the furry sheep it.

Templated Generation: Here we see two cows and one tree. The first cow is by the tree. The second cow is by the tree.
Simple Decoding: the cow and by the tree. the cow and by the tree.

Templated Generation: Here we see three persons, one sky, one grass and one train. The first colorful person is underneath the clear sky, and beside the second colorful person, and within the shiny train. The second colorful person is underneath the clear sky, and by the shiny train. The green grass is near the clear sky. The third black person is underneath the clear sky, and by the green grass, and within the shiny train. The shiny train is by the clear sky, and beside the green grass.

Simple Decoding: the colorful person is underneath the clear sky. the colorful person who beside the colorful person. the colorful person is underneath the clear sky. the green grass and near the clear sky. the colorful person is within the shiny train. the black person is underneath the clear sky. the black person and by the green grass. the shiny train and by the clear sky. the colorful person and by the shiny train. the shiny train and beside the green grass. the black person is within the shiny train.

Kulkarni et al., 2011, Fig. 5

Quantitative performance

Method	w/o	w/ synonym
Human	0.50	0.51
Language model-based generation	0.25	0.30
Template-based generation	0.15	0.18
Meaning representation (triples)	0.20	0.30

Table 1. Automatic Evaluation: BLEU score measured at 1

Method	Score
Quality of image parsing	2.85
Language model-based generation	2.77
Template-based generation	3.49

Table 2. Human Evaluation: possible scores are 4 (perfect without error), 3 (good with some errors), 2 (many errors), 1 (failure)

BLEU scores computed on 847 UIUC PASCAL images and associated sentences

Good description examples

This is a photograph of one sky, one road and one bus. The blue sky is above the gray road. The gray road is near the shiny bus. The shiny bus is near the blue sky.

There are two aeroplanes. The first shiny aeroplane is near the second shiny aeroplane.

There are one cow and one sky. The golden cow is by the blue sky.

This is a picture of one sky, one road and one sheep. The gray sky is over the gray road. The gray sheep is by the gray road.

Here we see one road, one sky and one bicycle. The road is near the blue sky, and near the colorful bicycle. The colorful bicycle is within the blue sky.

Here we see two persons, one sky and one aeroplane. The first black person is by the blue sky. The blue sky is near the shiny aeroplane. The second black person is by the blue sky. The shiny aeroplane is by the first black person, and by the second black person.

Kulkarni et al., 2011, Fig. 4

Bad description examples 1

Missing detections:

Here we see one pottedplant.

This is a picture of one dog.

Incorrect detections:

There are one road and one cat. The furry road is in the furry cat.

This is a picture of one tree, one road and one person. The rusty tree is under the red road. The colorful person is near the rusty tree, and under the red road.

Incorrect attributes:

This is a photograph of two sheeps and one grass. The first black sheep is by the green grass, and by the second black sheep. The second black sheep is by the green grass.

This is a photograph of two horses and one grass. The first feathered horse is within the green grass, and by the second feathered horse. The second feathered horse is within the green grass.

Kulkarni et al., 2011, Fig. 6

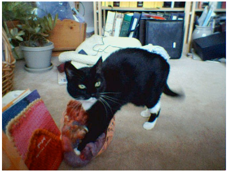
Bad description examples 2

Counting is hard!

There are two cows and one person. The first brown cow is against the brown person, and near the second cow. The brown person is beside the second cow.

This is a picture of four persons. The first colorful person is by the second pink person, and by the third colorful person. The second pink person is by the third colorful person, and by the fourth person.

Just all wrong!



There are one potted plant, one tree, one dog and one road. The gray potted plant is beneath the tree. The tree is near the black dog. The road is near the black dog. The black dog is near the gray potted plant.

This is a photograph of one person and one sky. The white person is by the blue sky.

Kulkarni et al., 2011, Fig. 6

More on image description generation/adaptation

- B. Yao, X. Yang, L. Lin, M. Lee and S.-C. Zhu. 2010. I2t: Image parsing to text description. *Proceedings of IEEE*
- A. Farhadi, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier and D. Forsyth. 2010. Every picture tells a story: Generating sentences for images. *Proceedings of ECCV*
- M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg and H. Daume III. 2012. Midge: Generating image descriptions from computer vision detections. *Proceedings of EACL*
- R. Mason. 2013. Domain-independent captioning of domain-specific images. *Proceedings of NAACL Student Research Workshop*
- Y. Feng and M. Lapata. 2013. Automatic caption generation for news images. *IEEE Trans. Pattern Anal. Mach. Intell. 35(4)*

1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

- Generating image descriptions
- Multimodal distributional semantics
- Visual selectional preference

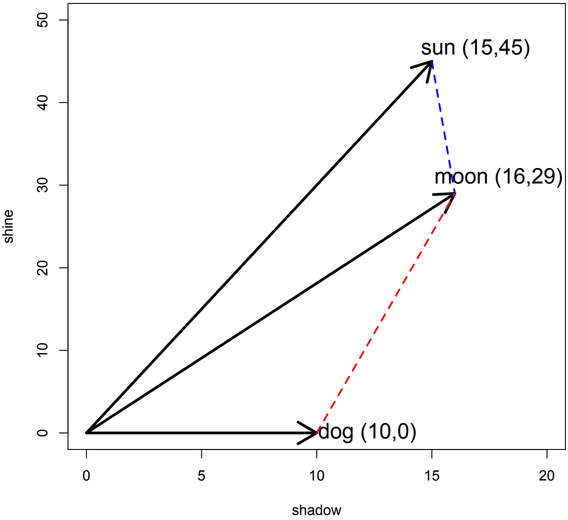
Multimodal distributional semantics

- Our work:
  - E. Bruni, G.B. Tran and M. Baroni. 2011. Distributional semantics from text and images. *Proceedings of GEMS*
  - E. Bruni, G. Boleda, M. Baroni and N.K. Tran. 2012. Distributional semantics in Technicolor. *Proceedings of ACL*
  - E. Bruni, J. Uijlings, M. Baroni and N. Sebe. 2012. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. *Proceedings of ACM-MM*
  - E. Bruni, N.K. Tran and M. Baroni. Submitted. Multimodal distributional semantics
- More:
  - Y. Feng and M. Lapata. 2010. Visual information in semantic representation. *Proceedings of NAACL*
  - C.W. Leong and R. Mihalcea. 2011. Going beyond text: A hybrid image-text approach for measuring word relatedness. *Proceedings of IJCNLP*

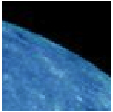
Distributional semantics

The geometry of meaning (e.g., Turney and Pantel, 2010)

	shadow	shine
moon	16	29
sun	15	45
dog	10	0

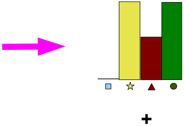


Multimodal distributional semantics

	planet	night		
moon	10	22	22	0
sun	14	10	15	0
dog	0	4	0	20

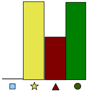
Distributional semantics from images


moon



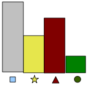
Distributional semantics from images

moon



+

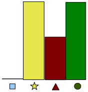
moon



Distributional semantics from images

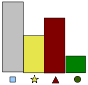
Labeled images

moon



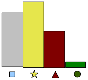
+

moon



+

moon

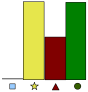


Instance counts

Distributional semantics from images

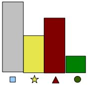
Labeled images

moon



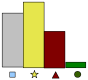
+

moon



+

moon



Instance counts

=

	□	☆	▲	●
moon	31	65	56	28

Total counts

Distributional semantics from images

sun	23	60	89	27

Total counts

Distributional semantics from images

	□	☆	▲	●
sun	23	60	89	27

Total counts

	□	☆	▲	●
dog	101	5	22	87

Total counts

Distributional semantics from images

				
moon	31	65	56	28
sun	23	60	89	27
dog	101	5	22	87

What is this good for?

Task 1 Predicting human **semantic relatedness** judgments

Task 2 **Concept categorization**, i.e. grouping words into classes based on their semantic relatedness (*car* ISA *vehicle*; *banana* ISA *fruit*)

Task 3 Find **typical color** of concrete objects (**cardboard is brown**, **tomato is red**)

Task 4 Distinguish **literal vs. non-literal** usages of color adjectives (**blue uniform** vs. **blue note**)

Task 1: Semantic relatedness data sets

- Data
  - WordSim353 dataset
    - 353 word pairs (coverage: 252)
    - 16 subjects rate each pair on a 10-point scale, ratings averaged
    - **dollar/buck: 9.22**, **professor/cucumber: 0.31**
  - MEN dataset (created by us)
    - 3,000 word pairs, tags in image datasets
    - crowdsourcing: subjects see two word pairs and pick the pair containing most related words
    - each word pair is rated 50 times, score = selected / 50
    - **cold/frost: 0.9**, **eat/hair: 0.1**

- Method
  - for each model, compute cosine between word vectors
  - score: Spearman correlation against the human ratings

Task 1: Semantic relatedness results

Bruni, N.K. Tran and Baroni (submitted)

	Window 2		Window 20	
Model	*MEN*	*WS*	*MEN*	*WS*
Text	0.73	0.70	0.68	0.70
Image	0.43	0.36	0.43	0.36
Fusion	**0.78**	**0.72**	**0.76**	**0.75**

Spearman correlation of the models on MEN and WordSim (all coefficients significant with $p < 0.001$).

Task 2: Concept categorization data sets

- Data
  - Battig (for training)
    - 77 concepts from 10 different classes
    - **bird (eagle, owl...) vegetable (broccoli, potato...)**
  - Almuhareb-Poesio (for testing)
    - 231 concepts from 21 different classes
    - **vehicle (airplane, car...) time (aeon, future...)**
- Method
  - cluster the words based on their pairwise cosines in the semantic space (using the CLUTO toolkit)

Task 2: Concept categorization results

Bruni, N.K. Tran and Baroni (submitted)

	Window 2	Window 20
Text	0.73	0.65
Image	0.26	0.26
Fusion	**0.74**	**0.69**

Percentage purities of the models on Almuhareb-Poesio.

Task 3: Find typical color of concrete objects

- Data and task
  - spot typical color of 52 concrete objects: **cardboard is brown**, **coal is black**, **forest is green**
  - typical colors assigned by two judges by consensus
    - Berlin and Kay (1969)'s basic color adjectives: **black, blue, brown, green, grey, orange, pink, purple, red, white, yellow**
- Method
  - rank color adjective vectors by similarity to the noun vectors
  - good models will rank right color high

Task 3: Find typical color of concrete objects

Bruni, Boleda, Baroni and N.K. Tran 2012

Model	*Score*
TEXT_{30K}	3 (11)
LAB_{128}	**1** (27)
SIFT_{40K}	3 (15)
TEXT+LAB_{128}	**1** (27)
TEXT+SIFT_{40K}	2 (17)

Median rank of correct color and # of top matches

Task 3: Examples

word	*gold*	*LAB*	*SIFT*	TEXT
cauliflower	white	green	yellow	orange
cello	brown	brown	black	blue
deer	brown	green	blue	red
froth	white	brown	black	orange
gorilla	black	black	red	grey
grass	green	green	green	green
pig	pink	pink	brown	brown
sea	blue	blue	blue	grey
weed	green	green	yellow	purple

Task 4: Literal vs. non-literal

- Data and task
  - distinguish literal and non-literal usages of color adjectives: **blue uniform**, **blue shark, blue note**
  - 342 adjective-noun pairs, 227 literal, 115 non-literal, as decided by two judges by consensus

- Method
  - compute cosine between color adjective vector and noun vector
  - prediction: higher similarity of color and noun vectors for literal uses

Task 4: Literal vs. non-literal

Model	*Score*
TEXT_{30K}	0.53***
LAB_{128}	0.25*
SIFT_{40K}	0.57***
TEXT+LAB_{128}	0.36***
TEXT+SIFT_{40K}	**0.73*****

Average difference in normalized adj-noun cosines
in literal vs. non-literal conditions with t-test significance

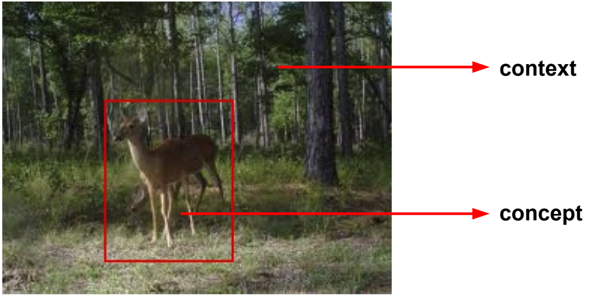
The illustrated distributional hypothesis

Current development

The meaning of a visually depicted concept is (can be approximated by, derived from) the set of **contexts** in which it occurs in images

The illustrated distributional hypothesis

	Localization:		
Area	*No*	*Manual*	*Automatic*
Concept	NA	0.39	0.36
Context	NA	0.50	0.51
Concept+Context	0.47	**0.54**	**0.54**



1 Why image analysis?

2 Annotated image datasets

3 Extraction of low-level features from images

4 Visual words for higher-level image representation

5 Example applications in computer vision

6 Going multimodal: Visual features in NLP

Generating image descriptions

Multimodal distributional semantics

Visual selectional preference

Visual selectional preference

S. Bergsma and R. Goebel. 2011. Using visual information to predict lexical preference. *Proceedings of RANLP*

What would you rather eat?

- migas?
- zeolite?
- carillons?
- a ficus?
- a mamey?
- manioc?

What would you rather eat?

Bergsma and Goebel, 2011, Fig. 1

Discriminative selectional preference

Bergsma et al. 2008

- Train a classifier for each verb to predict which nouns are acceptable objects:

$$y^v = \vec{\lambda}^v \cdot \vec{\Phi}^v(n)$$

- Requires large corpus to extract features such as co-occurrence of noun with other verbs
- For out-of-corpus-vocabulary nouns, use simple string-shape features

Enrich out-of-vocabulary noun representation with visual features

E.g. "*migas*" →  → $\mathbf{\Phi}^v(n)$

- Images retrieved from Flickr or Google Image Search, 6 images per noun
- Features: RGB-color and SIFT visual words (64- and 512-item color and 100- and 1000-item SIFT vocabularies, concatenated and fed to classifier)

Image from Shane Bergsma's slides

Task

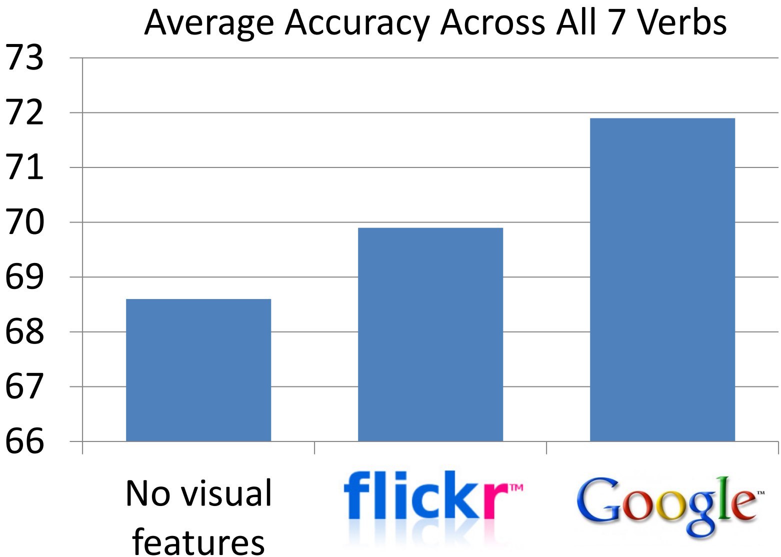
- Seven verbs: eat, inform, hit, kill, park, hunt and shoot down
- Training examples range from 500 to 10,000 per-verb, test instances from 50 to 1,000
  - Positive examples have AQUAINT corpus $PMI(v, n) > 0$, for negative examples $PMI(v, n) < 0$
- Baseline uses string features only
  - When using (co-)occurrence counts from the Web, visual features do not significantly improve over using textual data only!

Accuracy across verbs

Verb	Baseline	+Google features
eat	**68.3**	**79.5**
inform	*68.0*	*68.2*
hit	68.7	68.7
kill	**67.7**	**68.5**
park	69.9	69.9
hunt	**67.6**	**76.5**
shoot down	*70.0*	*72.0*

From Shane Bergsma's slides

Google Image better than Flickr



From Shane Bergsma's slides

The more images the better

From Shane Bergsma's slides

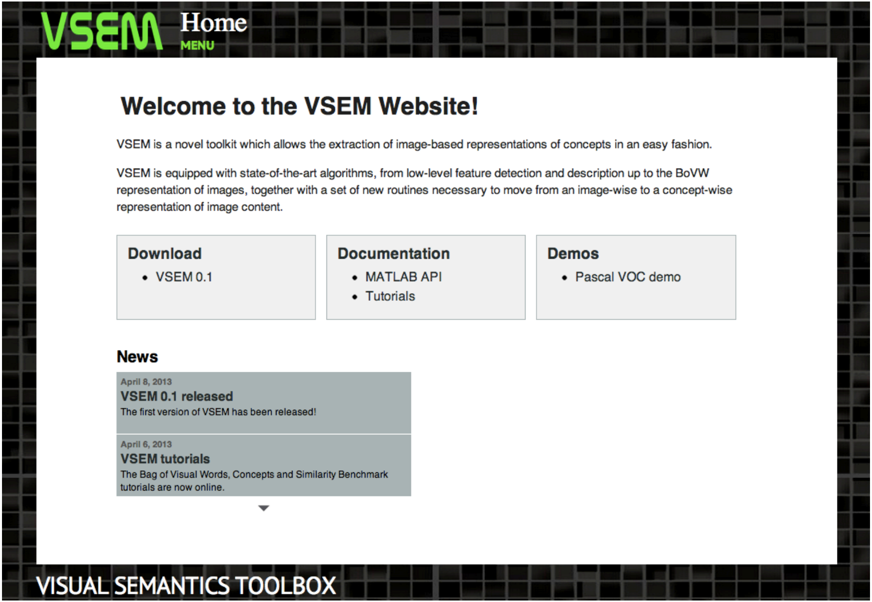
Both types of visual features help

Accuracy On The Verb "*Eat*"

Features	Accuracy
All Features	79.5
-Color Histogram	78.4
-SIFT Keypoint	78.1
-Color & -SIFT	68.3

From Shane Bergsma's slides

Come to our system demonstration!



SUPPLEMENTARY MATERIALS

Canny edges

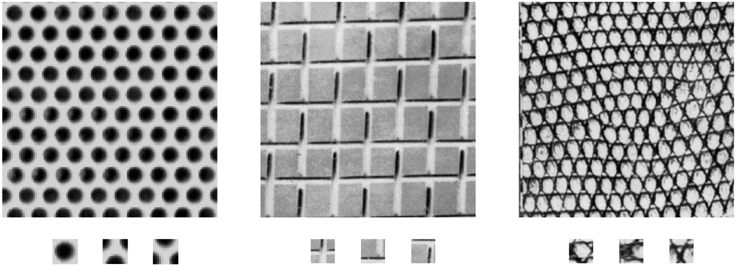
Slide credit: Kristen Grauman

Canny edges

Slide credit: Kristen Grauman

Textons

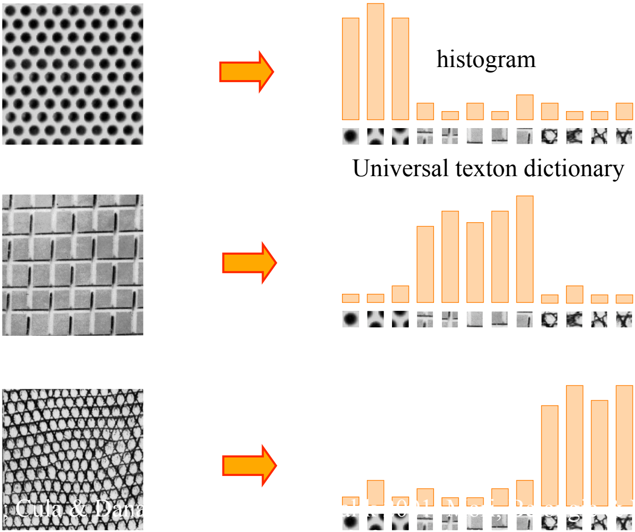
- Texture is characterized by the repetition of basic elements or **Textons**



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

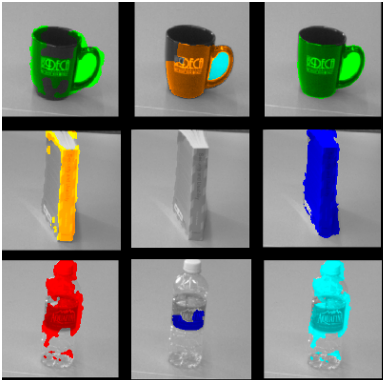
Slide credit: Cordelia Schmid

Textons



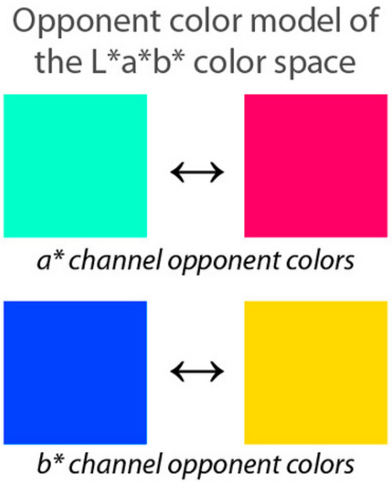
Slide credit: Cordelia Schmid

Color

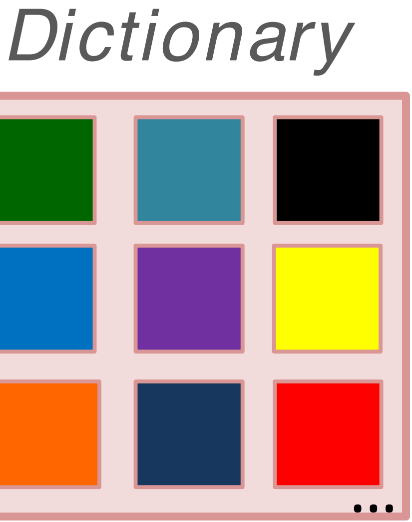


Slide credit: Dieter Fox

LAB


Opponent color model of
the L*a*b* color space
a* channel opponent colors
b* channel opponent colors

LAB



Slide credit: Victoria Yanuleskaya