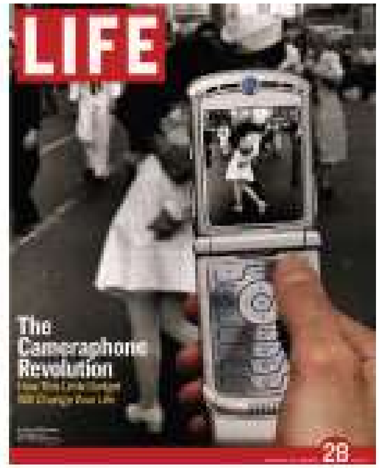
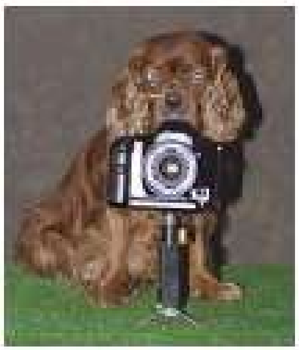
Words and Pictures and Beyond: Mining weakly-labeled web images and videos for automatic concept learning

Pinar Duygulu
Hacettepe University






Slide credit: Svetlana Lazebnik

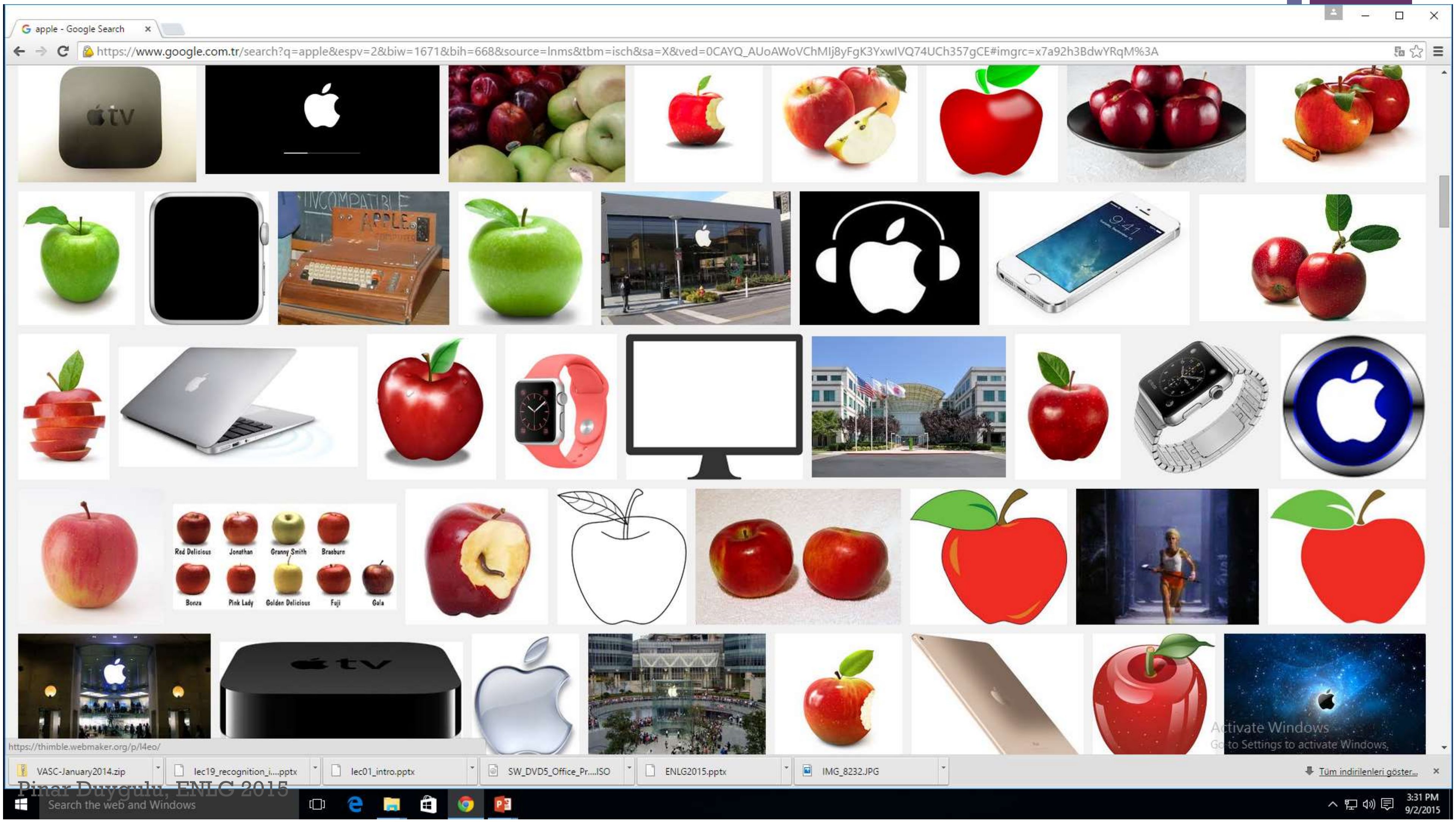
Massive amounts of visual data

- For YouTube alone
  - More than 1 billion unique users
  - Hundreds of millions of hours are watched every day
  - 300 hours of video are uploaded every minute

http://www.youtube.com/yt/press/statistics.html

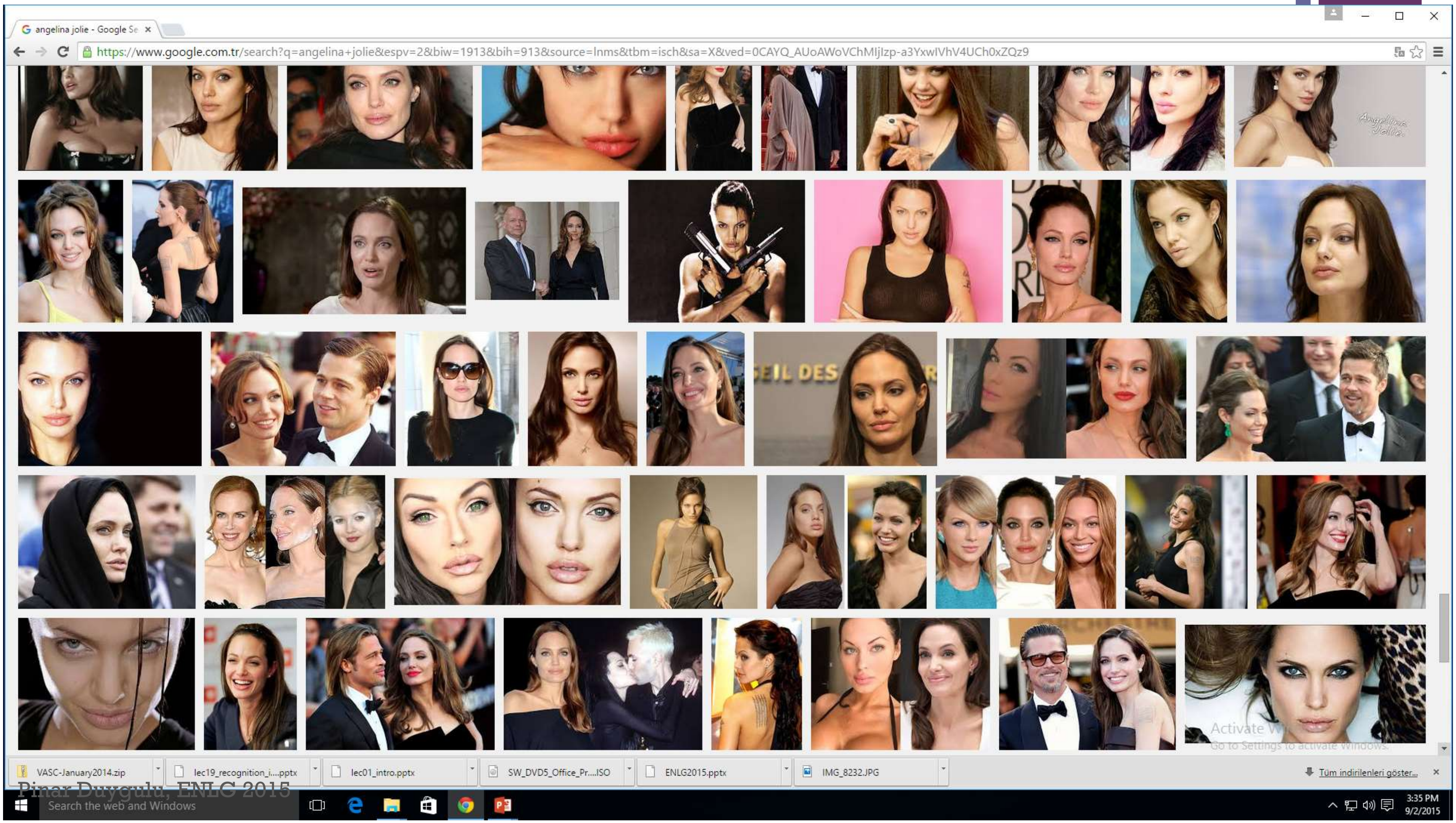
Access images through text search

Query : Apple



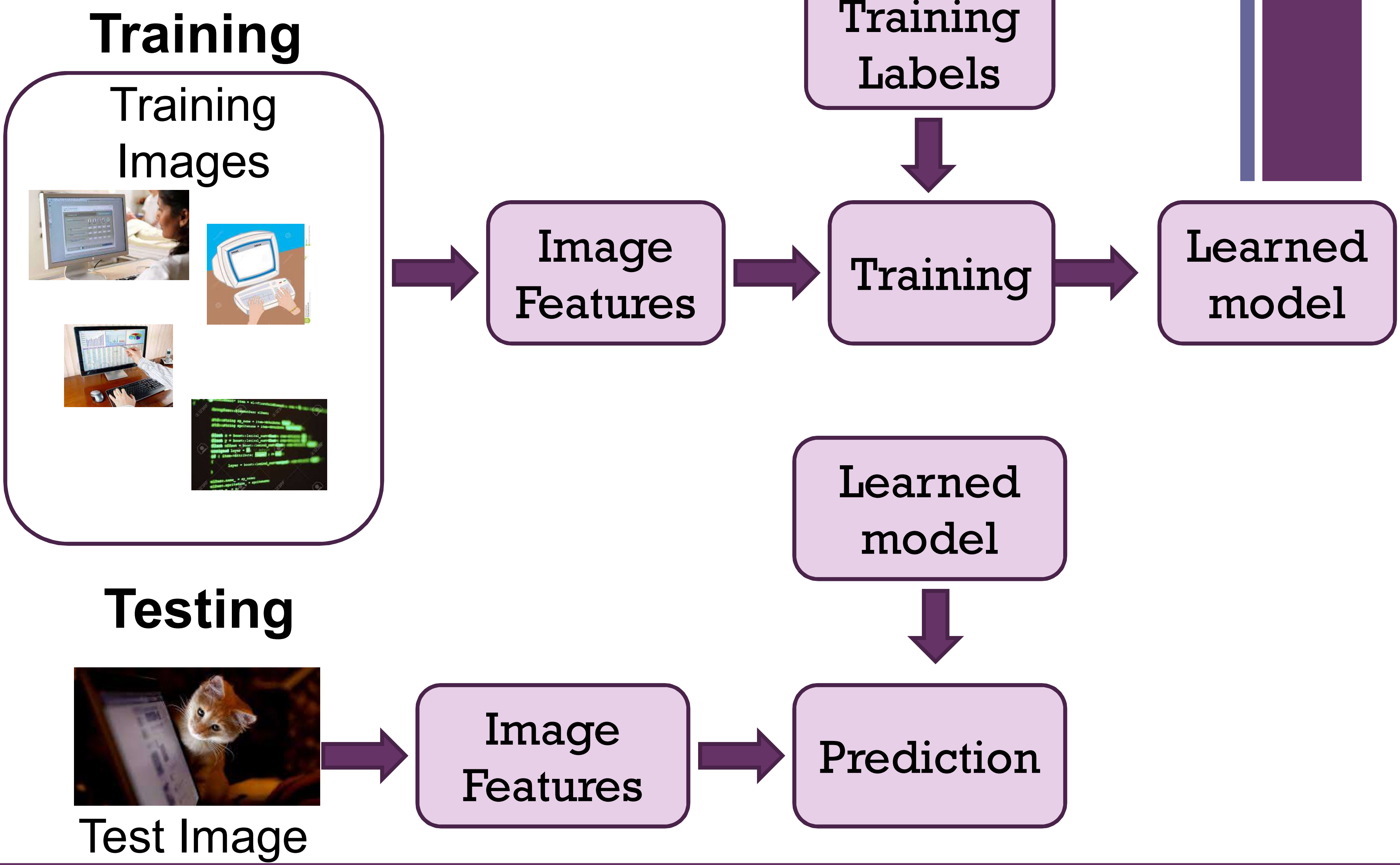
+ Access images through text search

Query : Angelina Jolie




~10,000 to 30,000

+ Learning Models

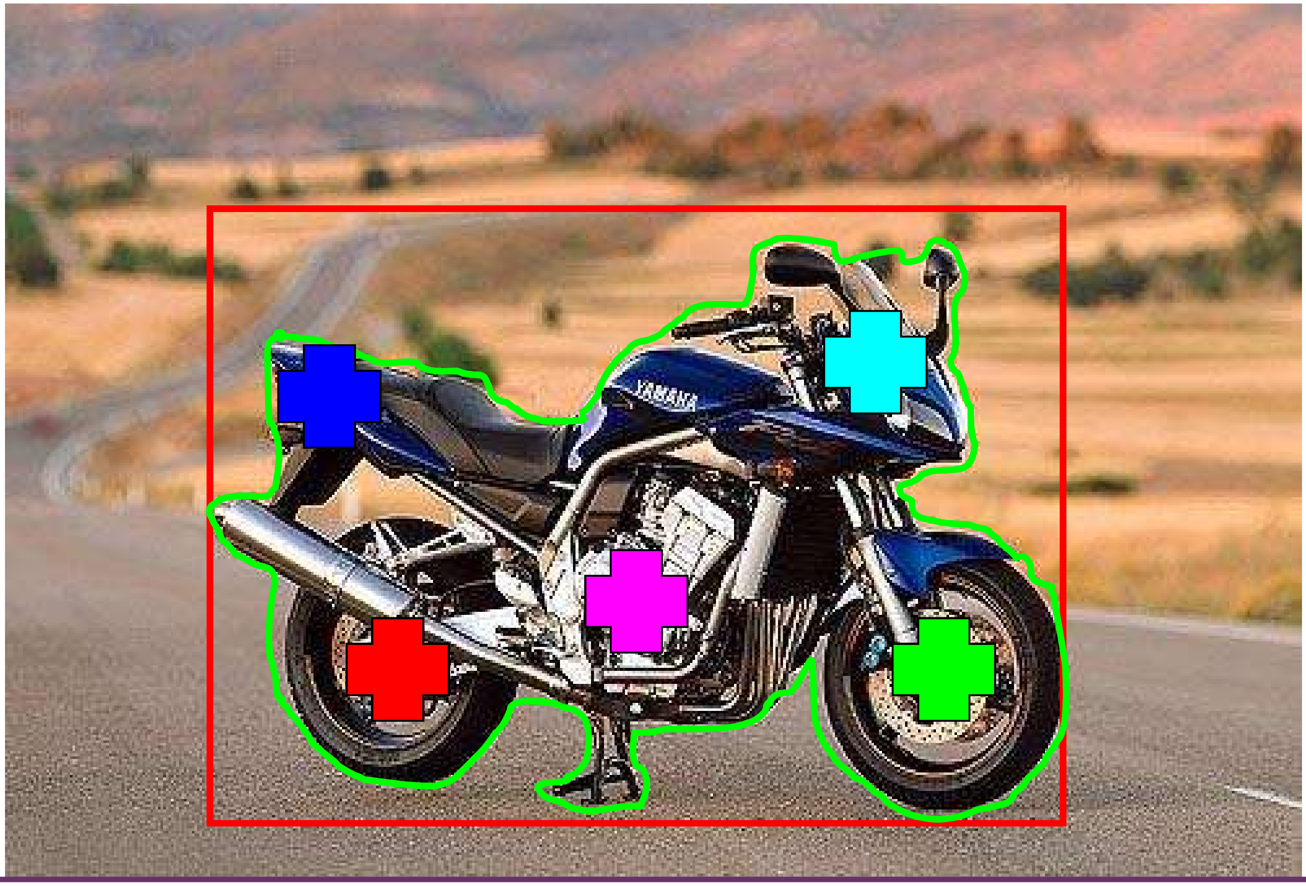


Slide credit: Derek Hoiem

+ Labeling required for supervision

Images in the training set must be annotated with the "correct answer" that the model is expected to produce

Contains a motorbike



Slide credit: Svetlana Lazebnik

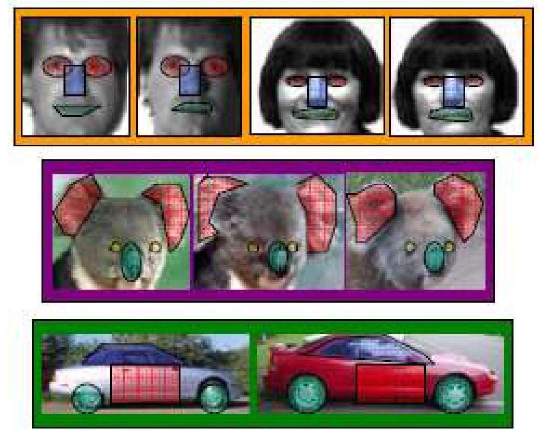
+

Spectrum of supervision

Less **More**

Unsupervised

"Weakly" supervised

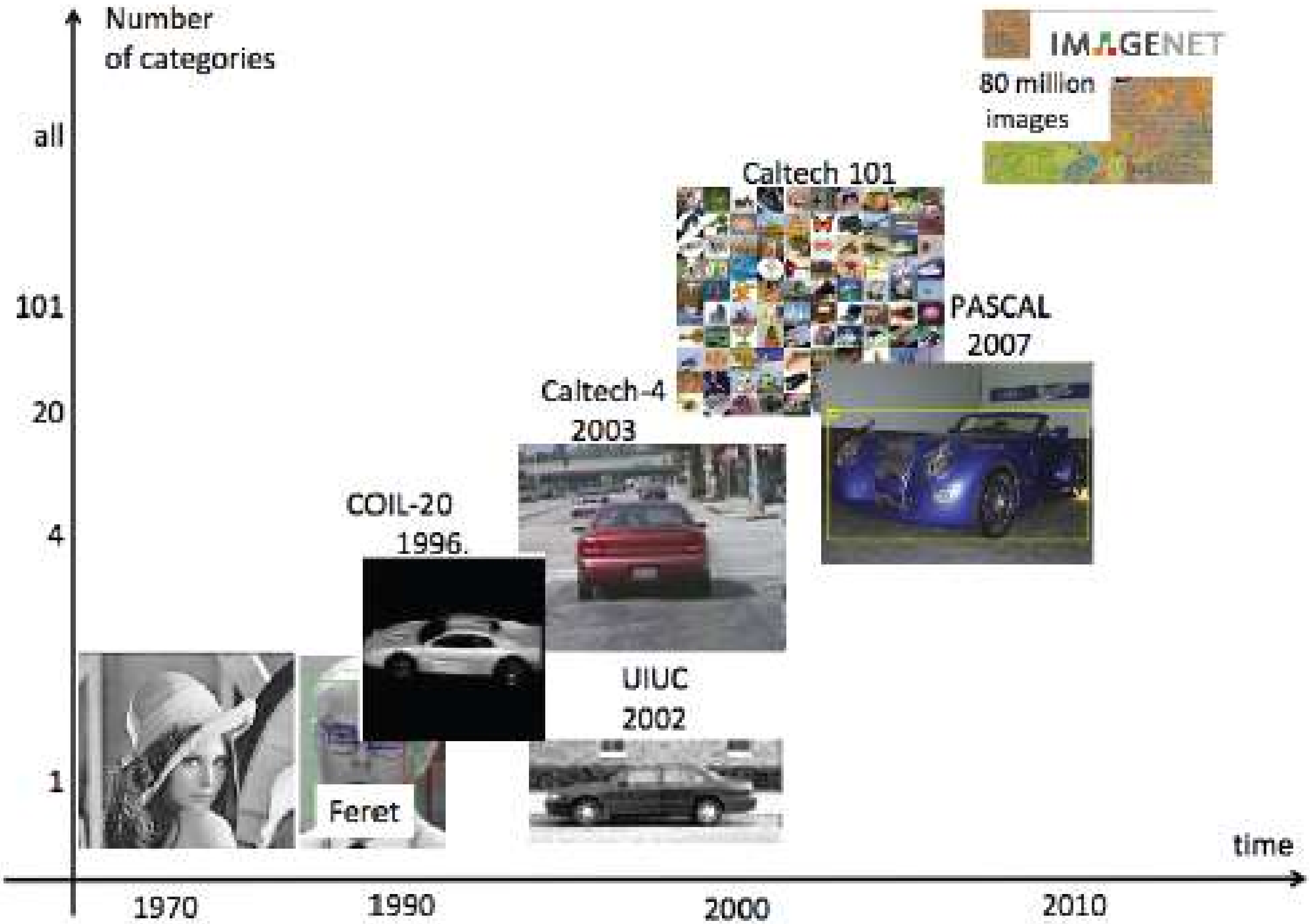


Fully supervised

Definition depends on task

Slide credit: Svetlana Lazebnik

+ Available datasets

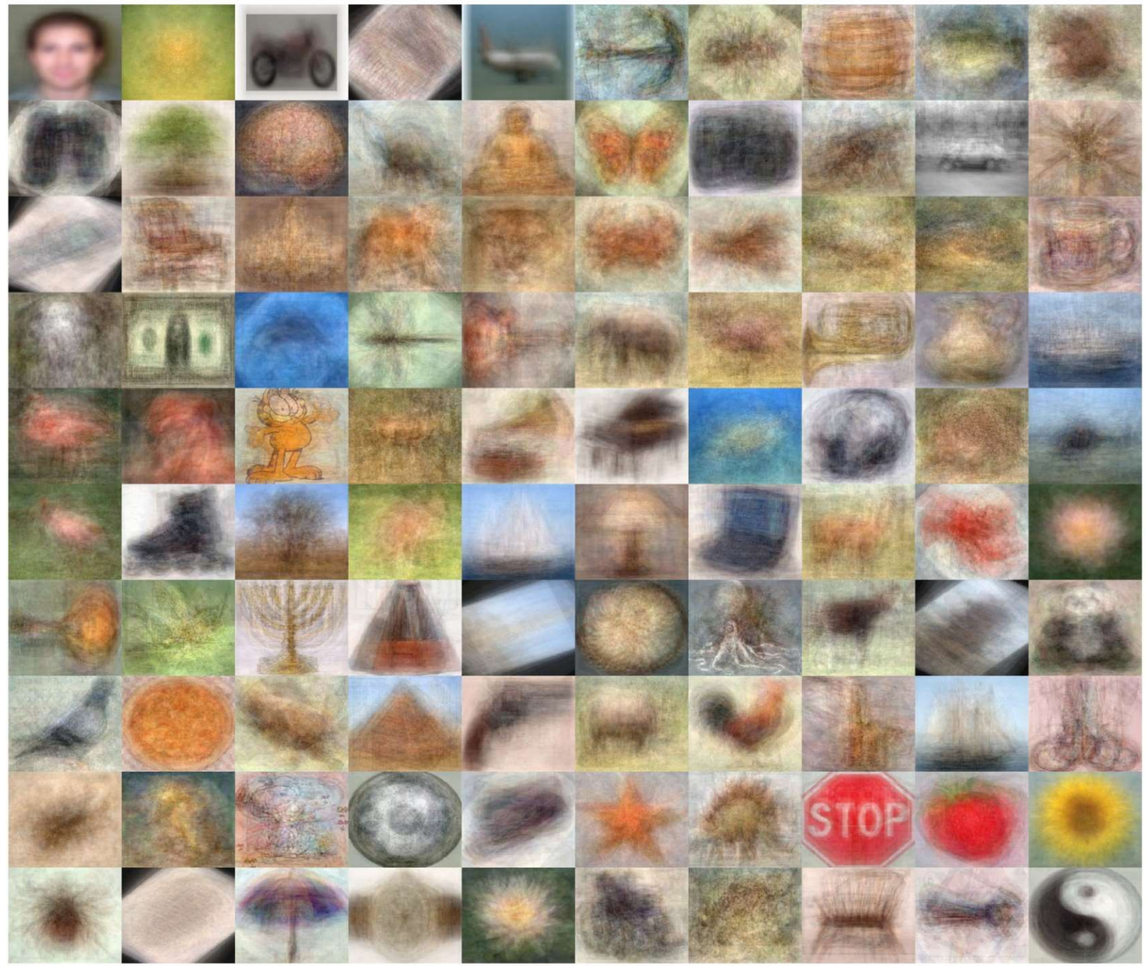


From "The Promise and Perils of Benchmark Datasets and Challenges", D. Forsyth, A. Efros, F.-F. Li, A. Torralba and A. Zisserman, Talk at "Frontiers of Computer Vision"

+ Caltech 101 and 256

Fei-Fei, Fergus, Perona, 2004

Caltech-101: Intra-class variability



Griffin, Holub, Perona, 2007

Slide credit: Svetlana Lazebnik

+ The PASCAL Visual Object Classes Challenge (2005-2012)

- **Challenge classes:**
  Person: person
  Animal: bird, cat, cow, dog, horse, sheep
  Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
  Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

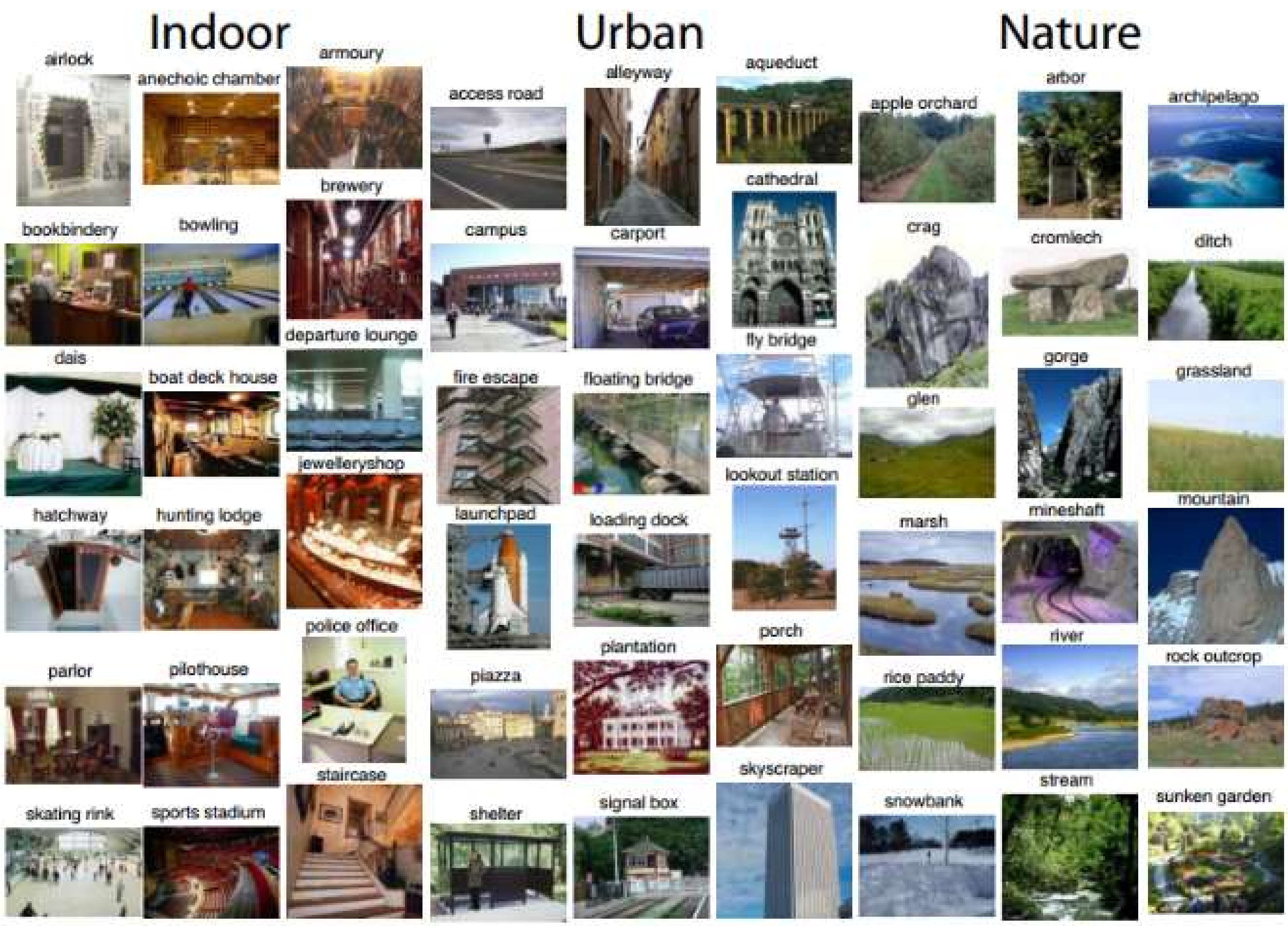
- **Dataset size (by 2012):**
  11.5K training/validation images, 27K bounding boxes, 7K segmentations

- Classification, detection, segmentation, person layout

Slide credit: Svetlana Lazebnik

+ Sun Dataset

~900 scene categories (~400 well-sampled), 130K images

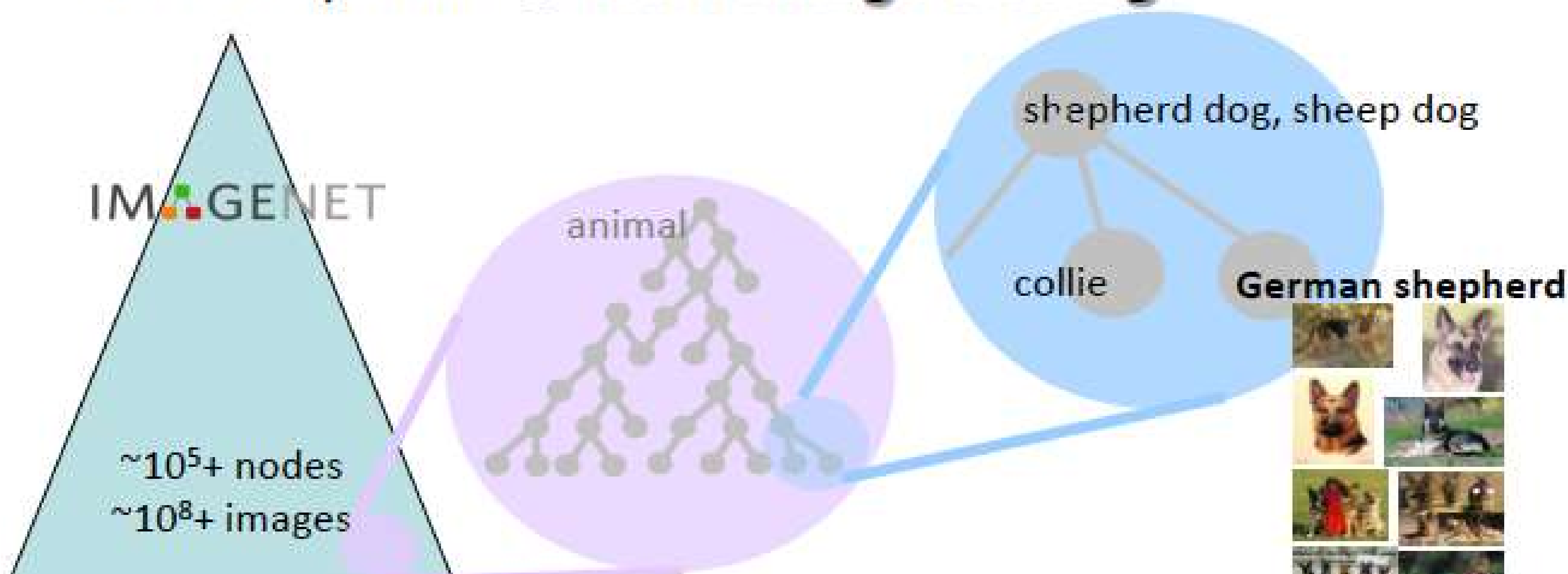


J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "SUN Database: Large-scale Scene Recognition from Abbey to Zoo," CVPR 2010

IMAGENET

$10^{6\text{-}7}$
images

- An ontology of images based on WordNet
- ImageNet currently has
  - ~15,000 categories of visual concepts
  - 10 million human-cleaned images (~700im/categ)
  - Free to public @ **www.image-net.org**



Slide credit: Fei-fei Li

Deng, Dong, Socher, Li & Fei-Fei, CVPR 2009

+MS COCO

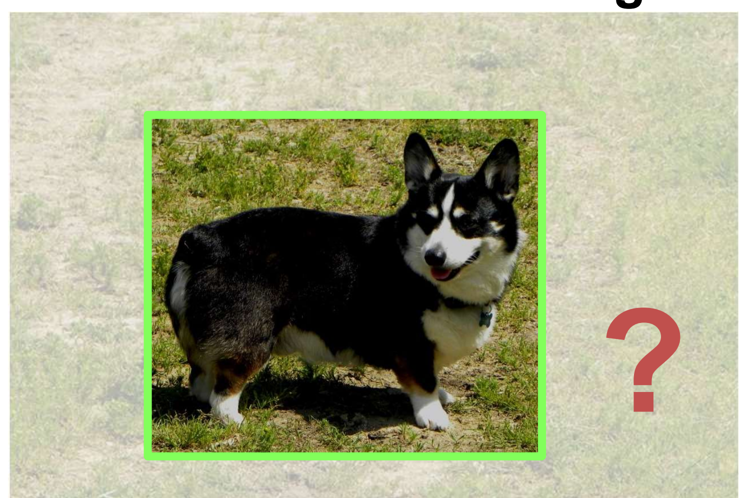
Over 77,000 worker hours (8+ years)

- 70-100 object categories (things not stuff)
- 330,000 images (~150k first release)
- 2 million instances (400k people)
- Every instance is segmented
- 7.7 instances per image (3.5 categories)
- Key points
- 5 sentences per image

http://mscoco.org

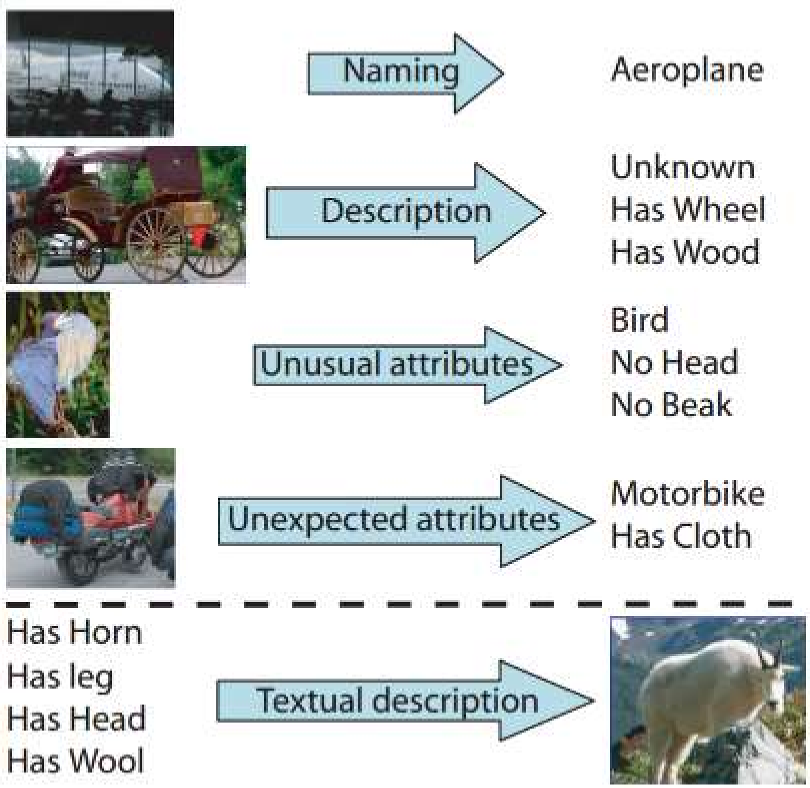
+ Fine grained recognition

What breed is this dog?

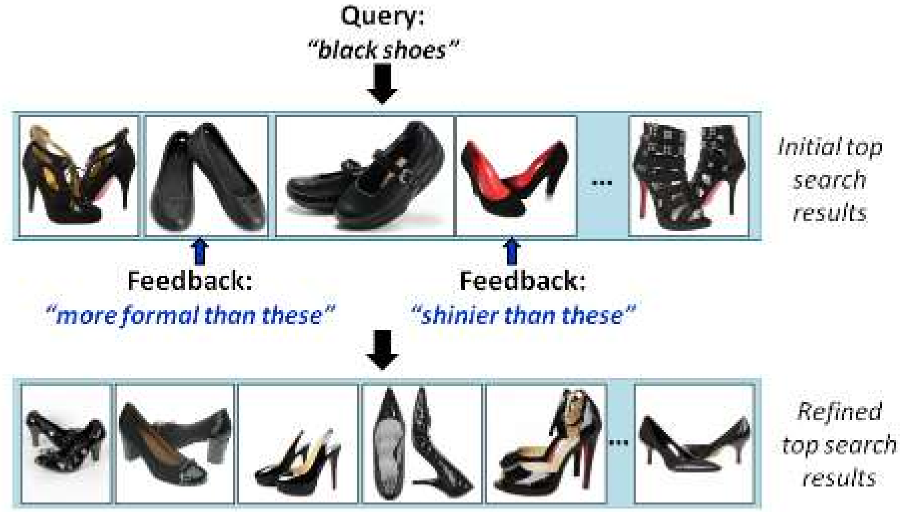


Slide credit: Svetlana Lazebnik

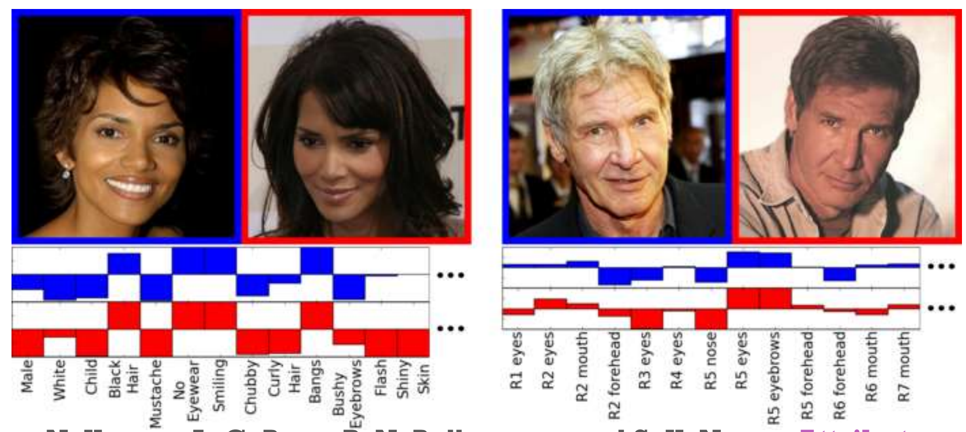
+ Attribute based recognition



A. Farhadi, I. Endres, D. Hoiem, and D Forsyth, **Describing Objects by their Attributes**, CVPR 2009



A. Kovashka, D. Parikh and K. Grauman, **WhittleSearch: Image Search with Relative Attribute Feedback**, CVPR 2012



N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, **Attribute and Simile Classifiers for Face Verification**, ICCV 2009

Slide credit: Svetlana Lazebnik

What is in this picture?

Fuzzy black thing with

a face-like part

-- maybe an animal?

Green, textured region
– maybe tree?

Tags:

leaves, endangered, green, i love nature, chennai, nilgiri langur, monkey, forest, wildlife, perch, black, wallpaper, ARK OF WILDLIFE, topv111, WeeklySurvivor, top20HallFame, topv333, 100v10f, captive, simian

Slide credit Tamara Berg

Consumer Photo Collections

Flickr – 6+ billion photographs, millions uploaded per day

Over the hills and far away

Road, Hills, Germany, Hoffenheim, Outstanding Shots, specland, Baden-Wuerttemberg

Heavenly

Peacock, AlbinoPeacock, WhiteBeauty, Birds, Wildlife, FeathredaleWildlifePark,

End of the world - Verdens Ende - T lighthouse 1

Verdens ende, end of the world, norway, lighthouse, ABigFave, vippefyr, wood, coal

Slide credit Tamara Berg

+ Museum and Library Collections

- Fine Arts Museum of San Francisco (82,000 images)

bowl stemmed small Irridescent glass

Woman of Head Howard H G Mrs Gift America North bust States United Sculpture marble

New York Public Library

Digital Collection

The new board walk, Rockaway, Long Island

Part of New England, New York, east New Iarsey and Long Iland.

Slide credit Tamara Berg

+ Consumer Products

Soft and glossy patent calfskin trimmed with natural vachetta cowhide, open top satchel for daytime and weekends, interior double slide pockets and zip pocket, seersucker stripe cotton twill lining, kate spade leather license plate logo, imported.
2.8" drop length
14"h x 14.2"w x 6.9"d

Katespade.com

It's the perfect party dress. With distinctly feminine details such as a wide sash bow around an empire waist and a deep scoopneck, this linen dress will keep you comfortable and feeling elegant all evening long.

* Measures 38" from center back, hits at the knee.
* Scoopneck, full skirt.
* Hidden side zip, fully lined.
* 100% Linen. Dry clean.

bananarepublic.com

Slide credit Tamara Berg

+ Video

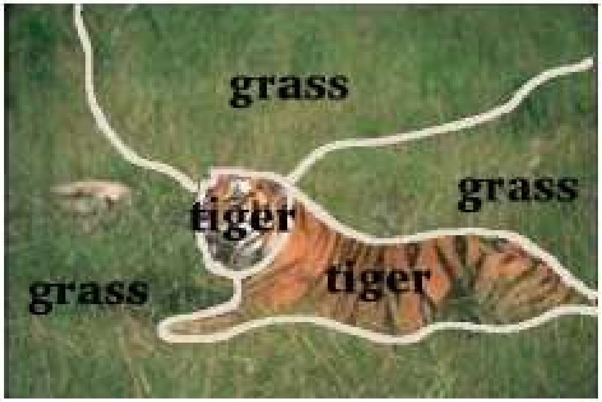
OUTSIDE IN THE RAIN THE SENATOR WEARING HIS UH BASEBALL CAP A BOSTON RED SOX CAP AS HE TALKED TO HIS SUPPORTERS HERE IN THE RAIN THE UH SENATOR THEY'RE DOING HIS BEST TO TRY TO MAKE HIS CASE THAT HE WILL BE THE MAN FOR THE MIDDLE CLASS AND UH TRY TO CONVINCE HIS SUPPORTERS TO EXPRESS THEIR SUPPORT THROUGH A VOTE ON TUESDAY IN THERE WE ARE TWENTY FOUR HOURS FROM THE GREAT MOMENT THAT THE WORLD IN AMERICA IS WAITING FOR IT I NEED TO YOU IN THESE HOURS TO GO OUT AND DO THE HARD WORK NOT ON THOSE DOORS MAKE THOSE PHONE CALLS TO TALK TO FRIENDS TAKE PEOPLE TO THE POLLS HELP US CHANGE THE DIRECTION OF THIS GREAT NATION FOR THE BETTER CAN YOU IMAGINE A UH SENATOR BEGINNING HIS DAY IN FLORIDA TODAY

TrecVid 2006 – video frames with speech processing output

Slide credit Tamara Berg

A novel approach for object recognition

Object recognition on large scale is linking image regions with words



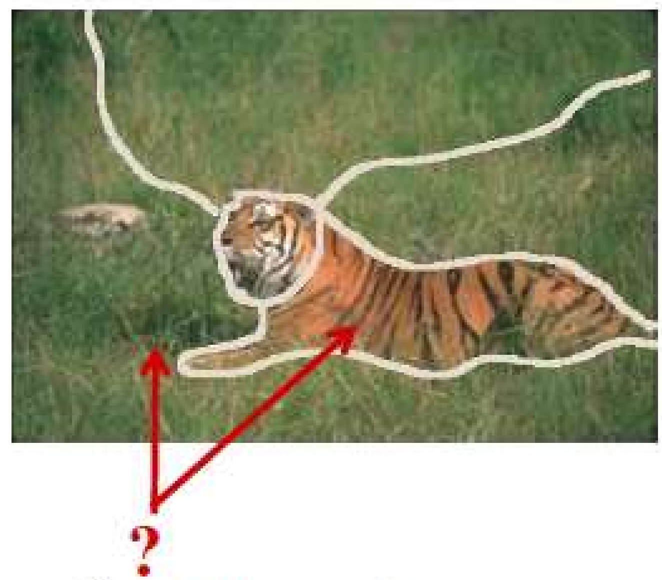
Use joint probability of words and Images in large data sets.

tiger grass cat

P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object Recognition as Machine Translation", ECCV 2002

+ Annotation versus recognition

tiger grass cat

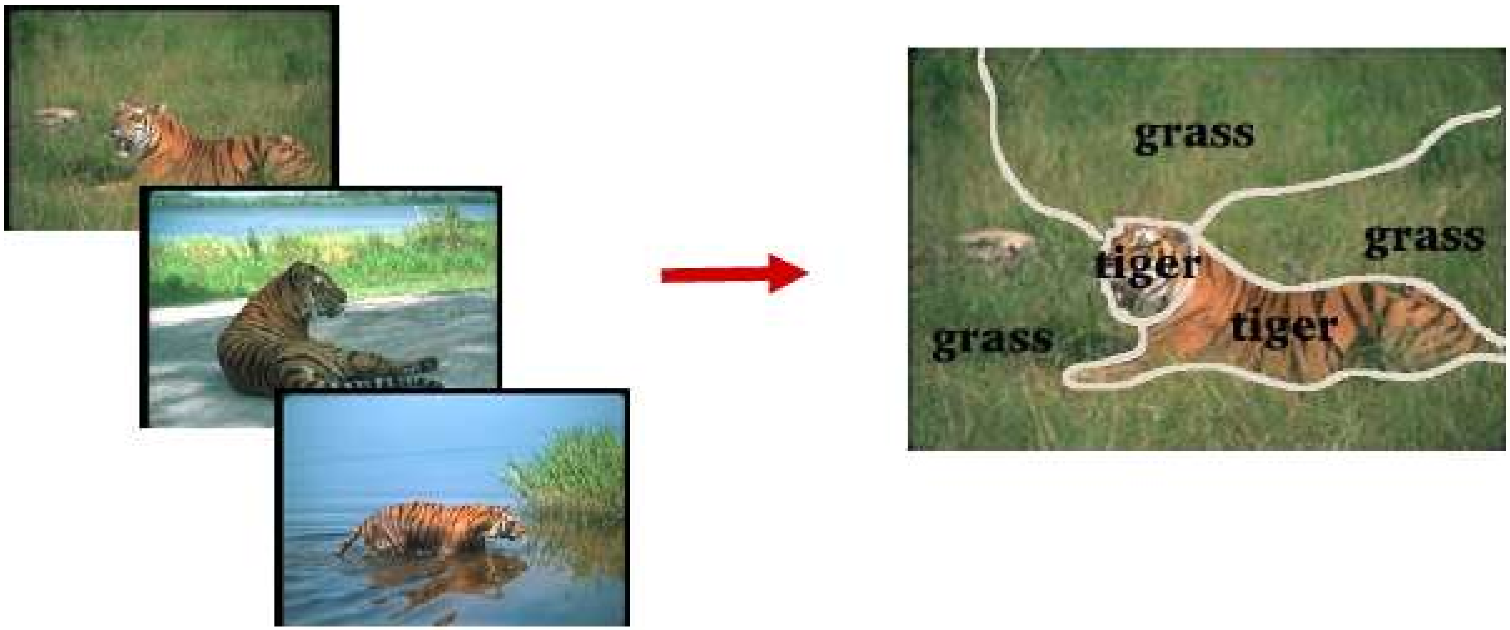


tiger grass cat

Cannot be learned from

P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object Recognition as Machine Translation", ECCV 2002

Making use of large volumes



P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object Recognition as Machine Translation", ECCV 2002

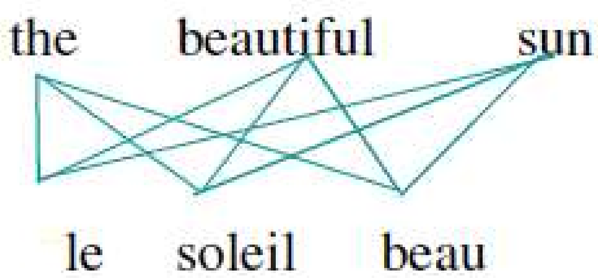
+ Statistical Machine Translation

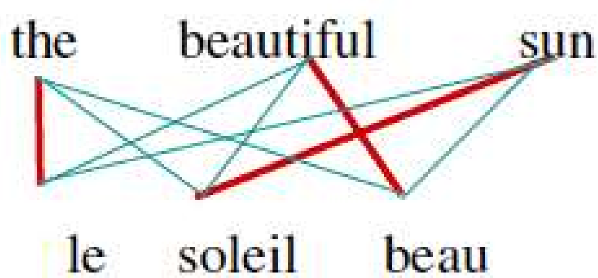
Data : aligned sentences
But word correspondences
are unknown

•Given the correspondences,
we can estimate the translation
p(sun | soleil)
•Given the probabilities, we can
estimate the correspondences

Solution: enough data + EM

Brown et. al 1993





P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object Recognition as Machine Translation", ECCV 2002

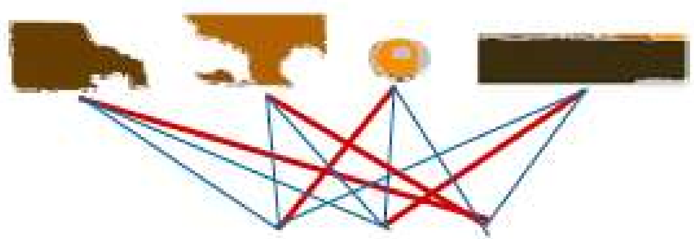
+

Data :

Words are associated with the images

But correspondences between image regions and words are unknown

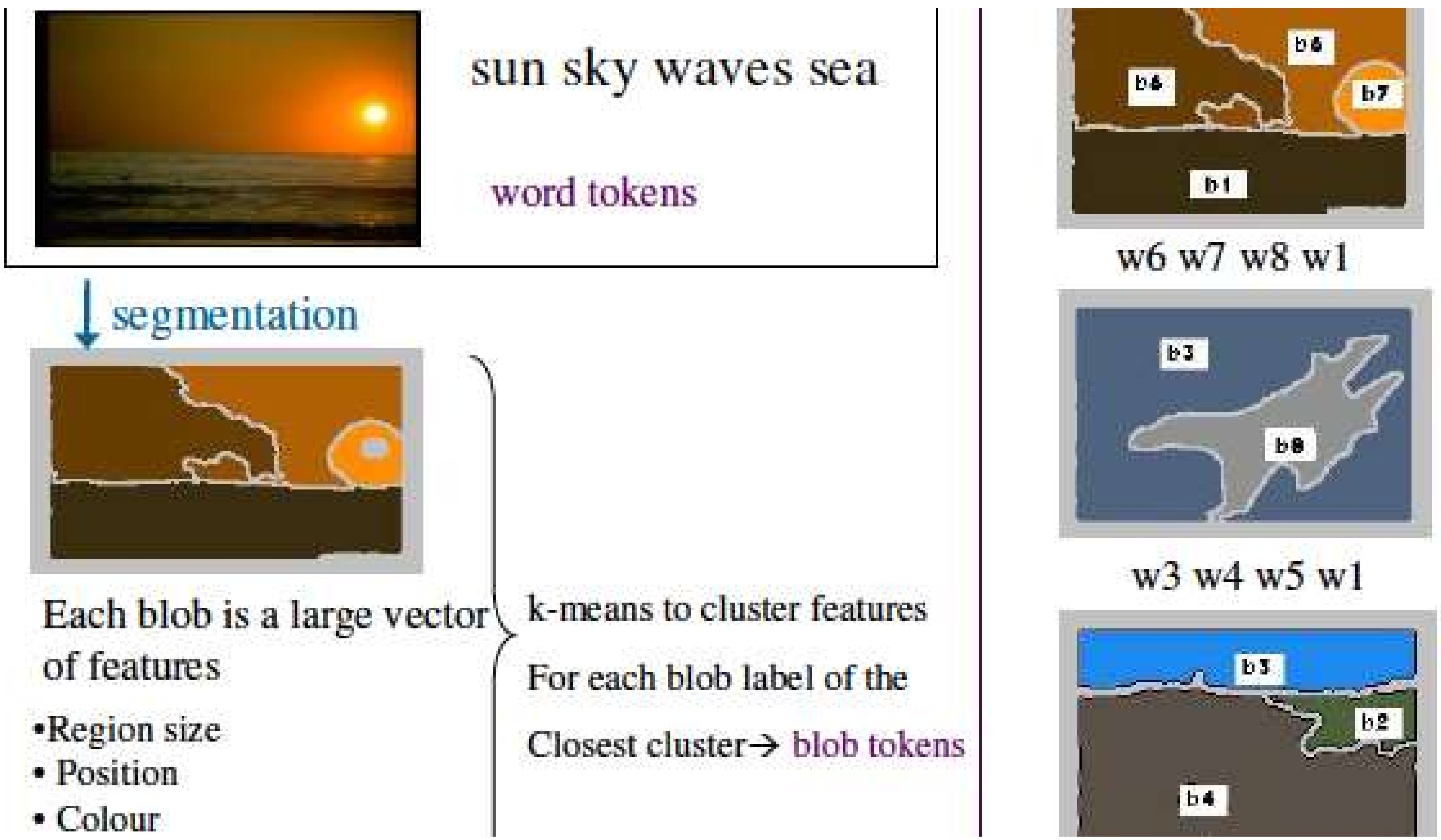
"sun sea sky"



"sun sea sky"

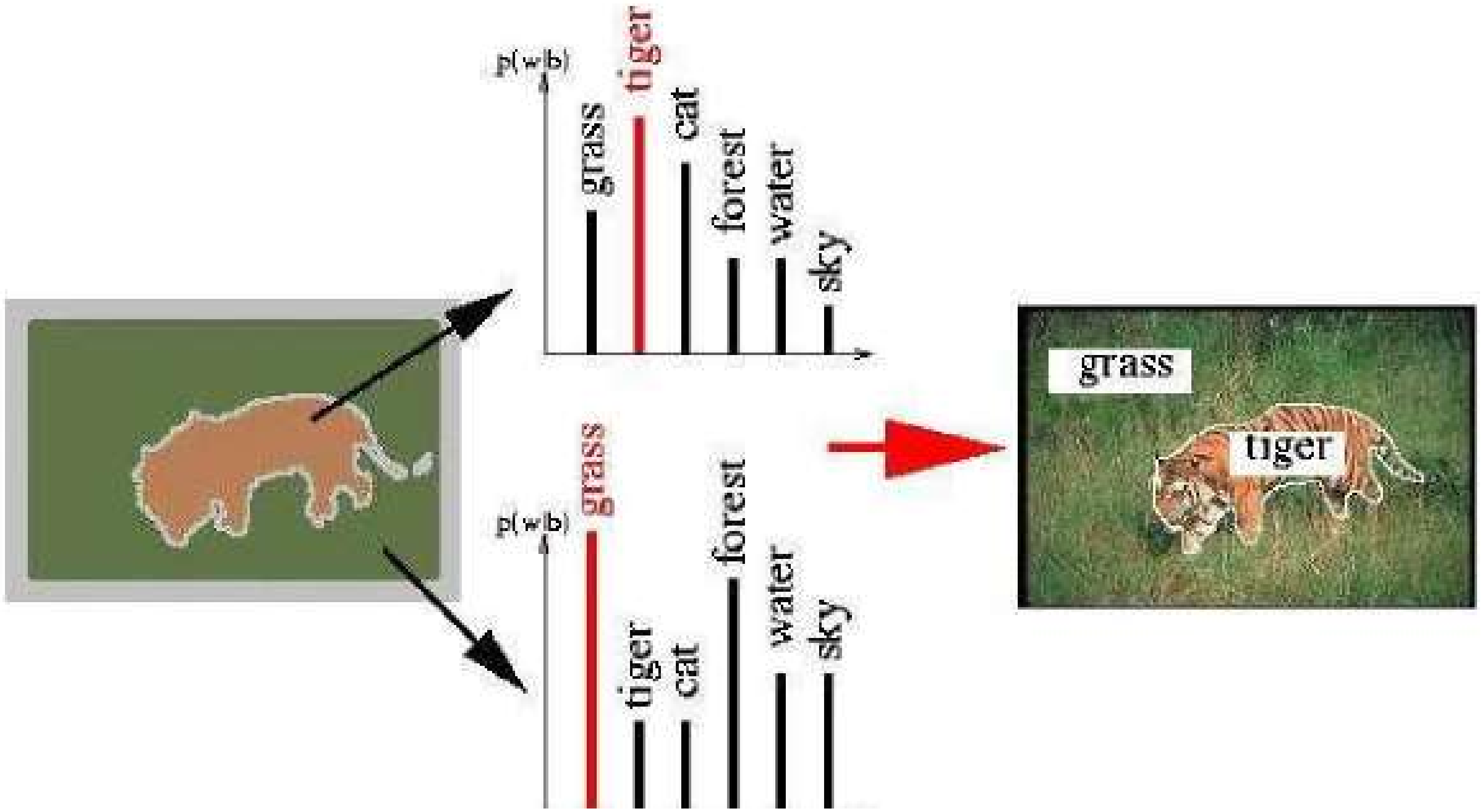
P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object Recognition as Machine Translation", ECCV 2002

+ Input representation



P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object Recognition as Machine Translation", ECCV 2002

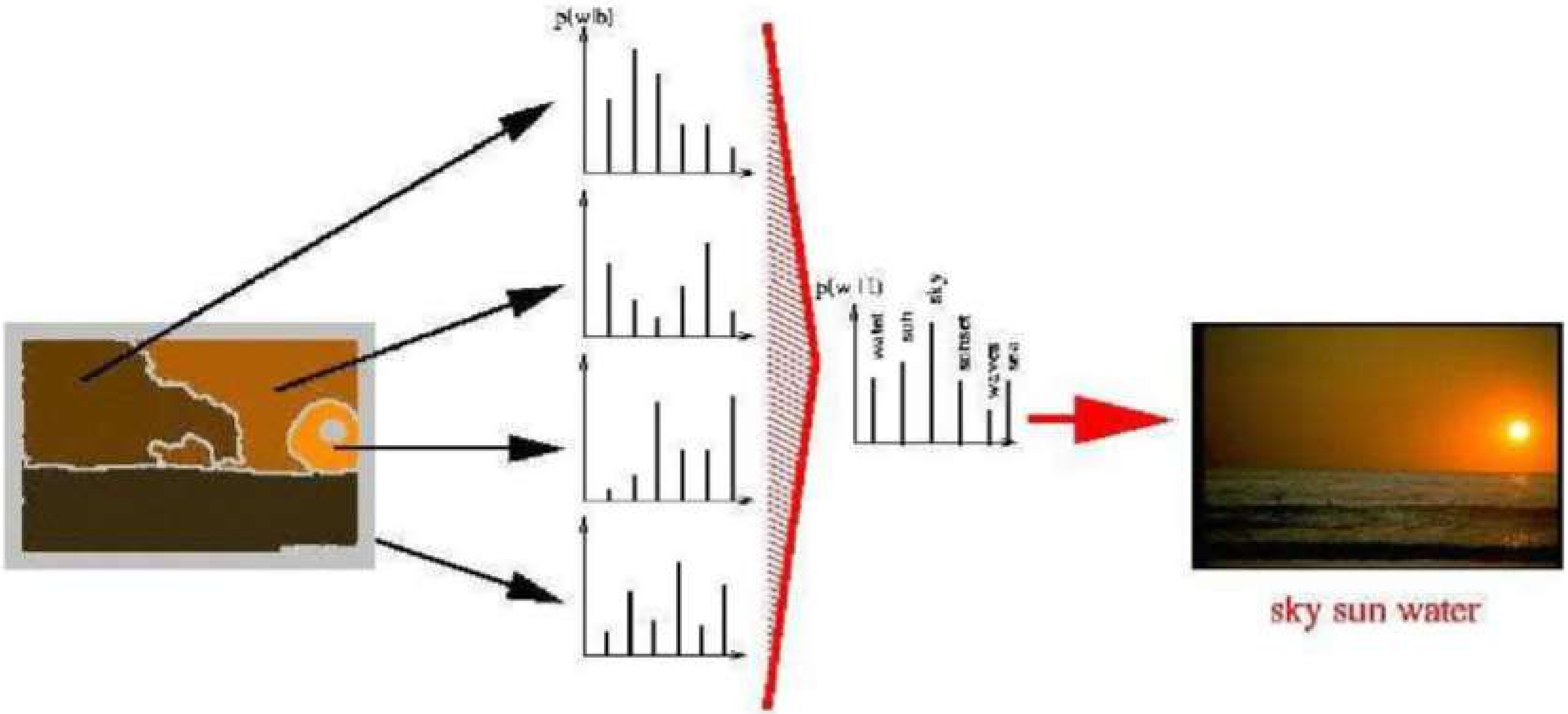
Region naming



P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object Recognition as Machine Translation", ECCV 2002

Barnard et al. JMLR, 2005

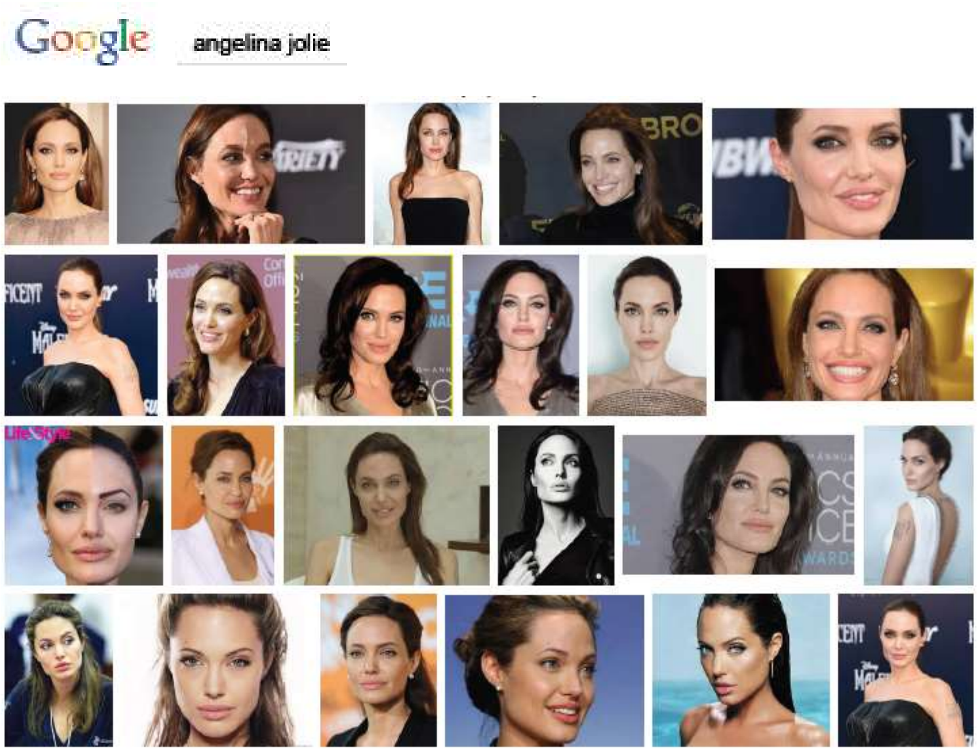
+ Auto Annotation



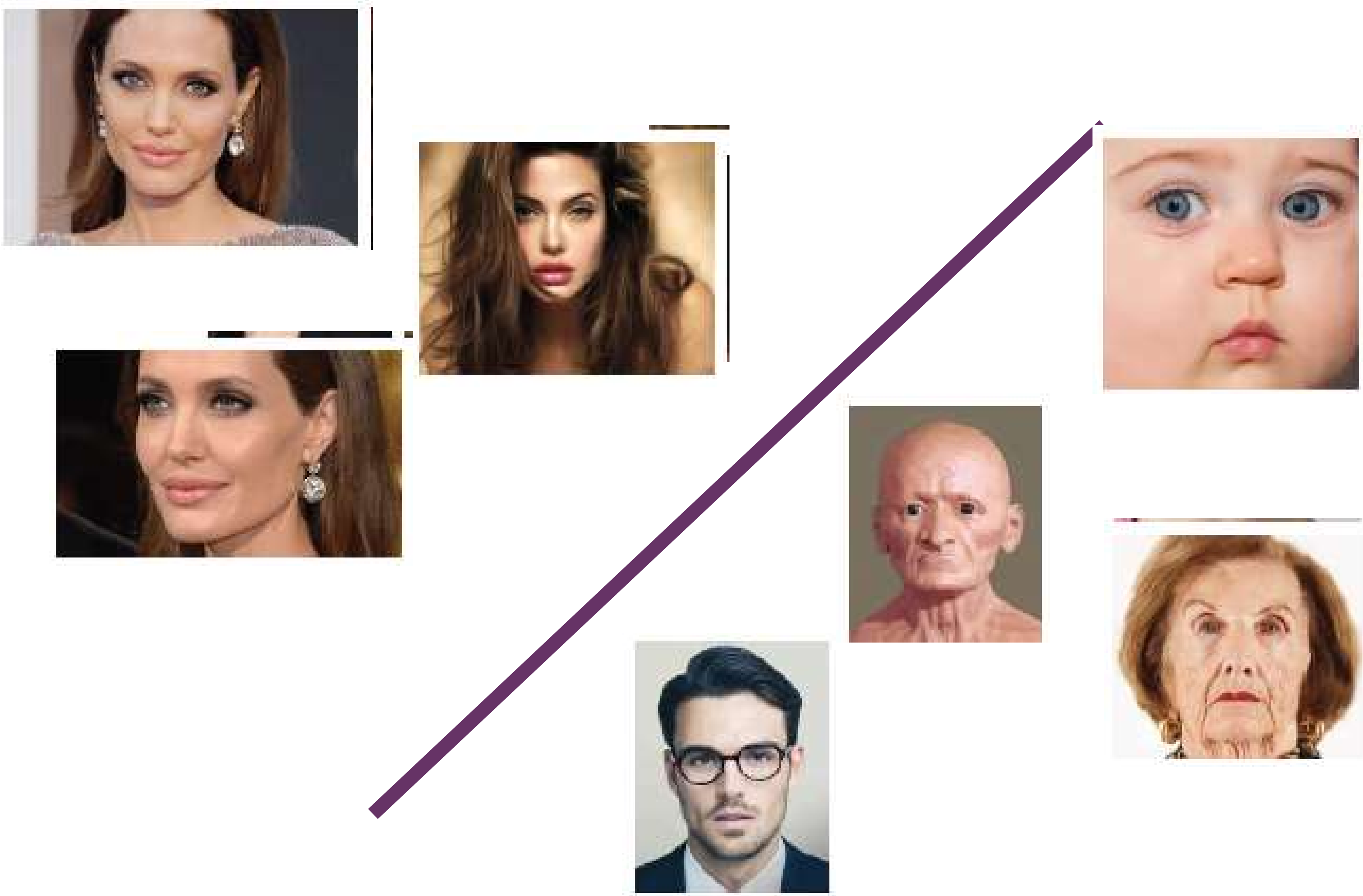
P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object Recognition as Machine Translation", ECCV 2002

+ Labeling for how many?

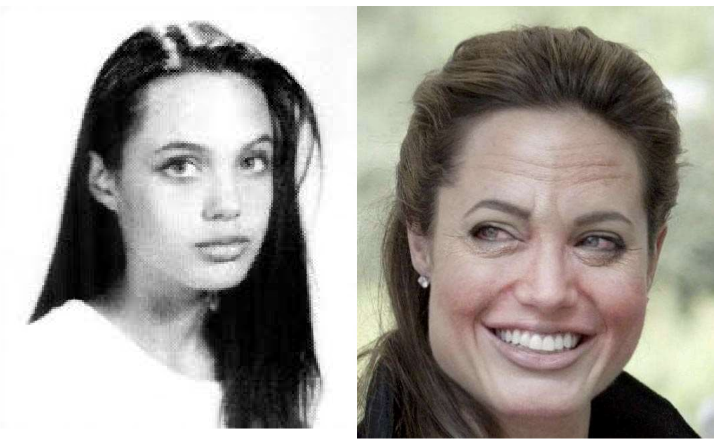
Search web for faces of a query name

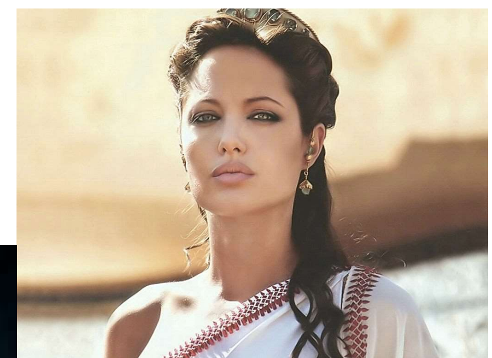


Use this set to learn models



Variations and sub-categories



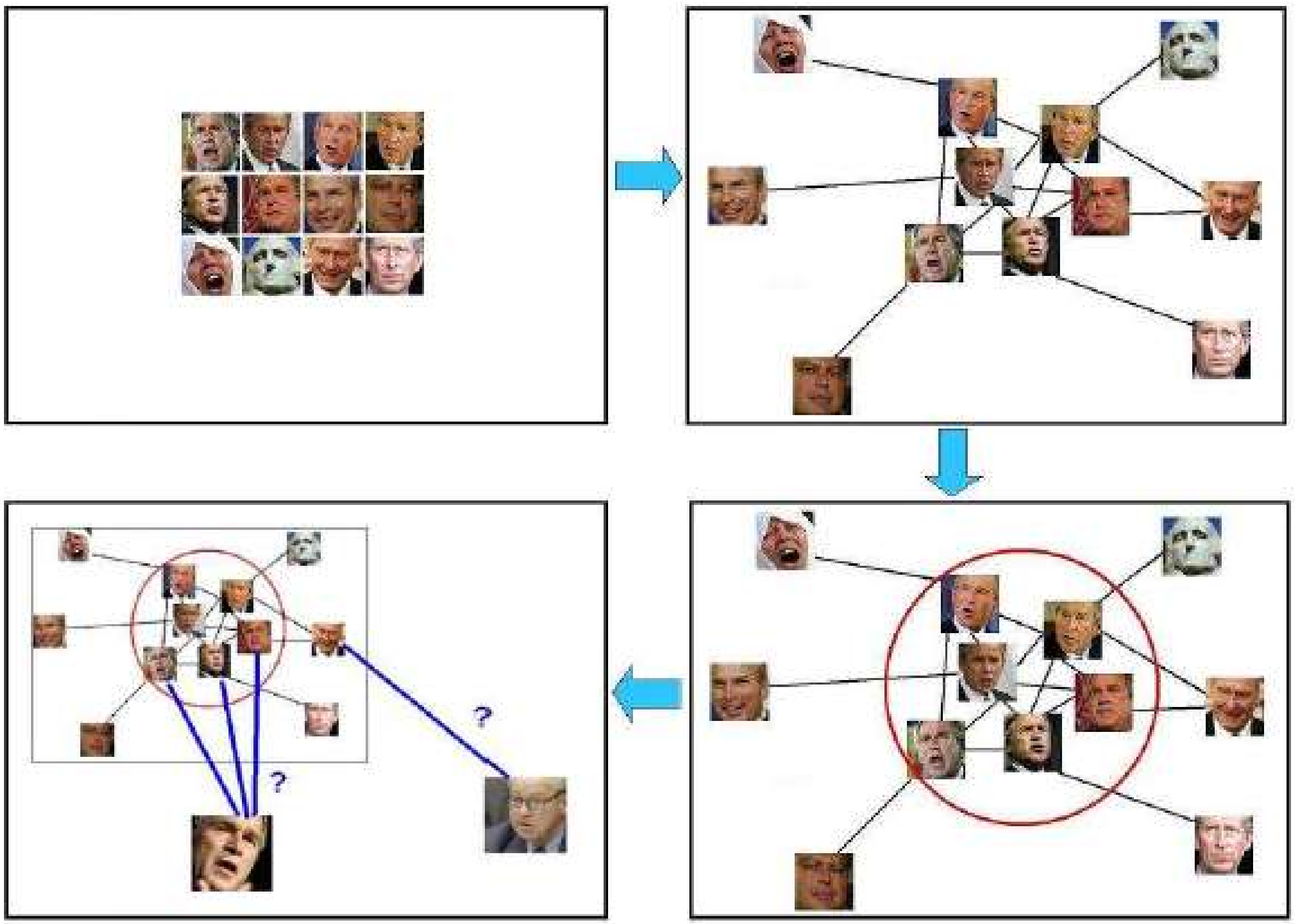


Irrelevant people

+ Single Dominant Category

Query : George W. Bush

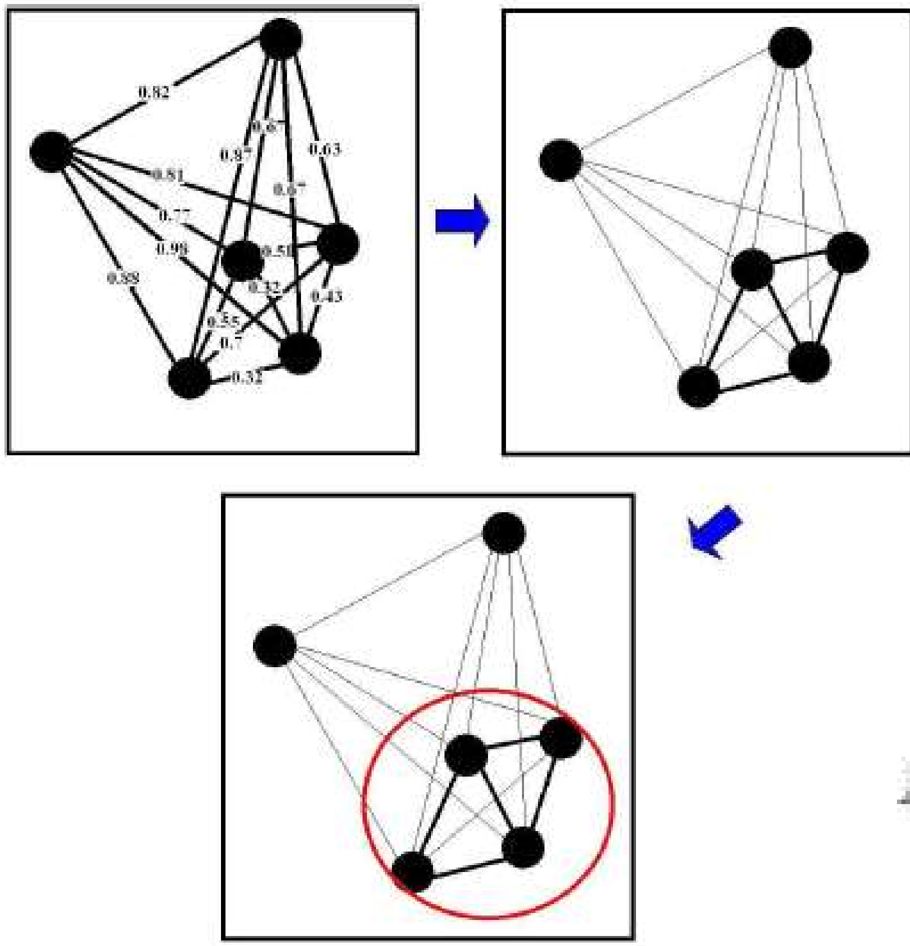
Naming faces



Among the faces associated with a name find the correct subset : The most similar subset of faces

Ozkan, D., Duygulu, P., "Interesting Faces: A Graph Based Approach for Finding People in News", Pattern Recognition, 2010
Ozkan, D., Duygulu, P., "A Graph Based Approach for Naming Faces in News Photos", CVPR, 2006
Ozkan, D., Duygulu, P., "Finding People Frequently Appearing in News", CIVR, 2006

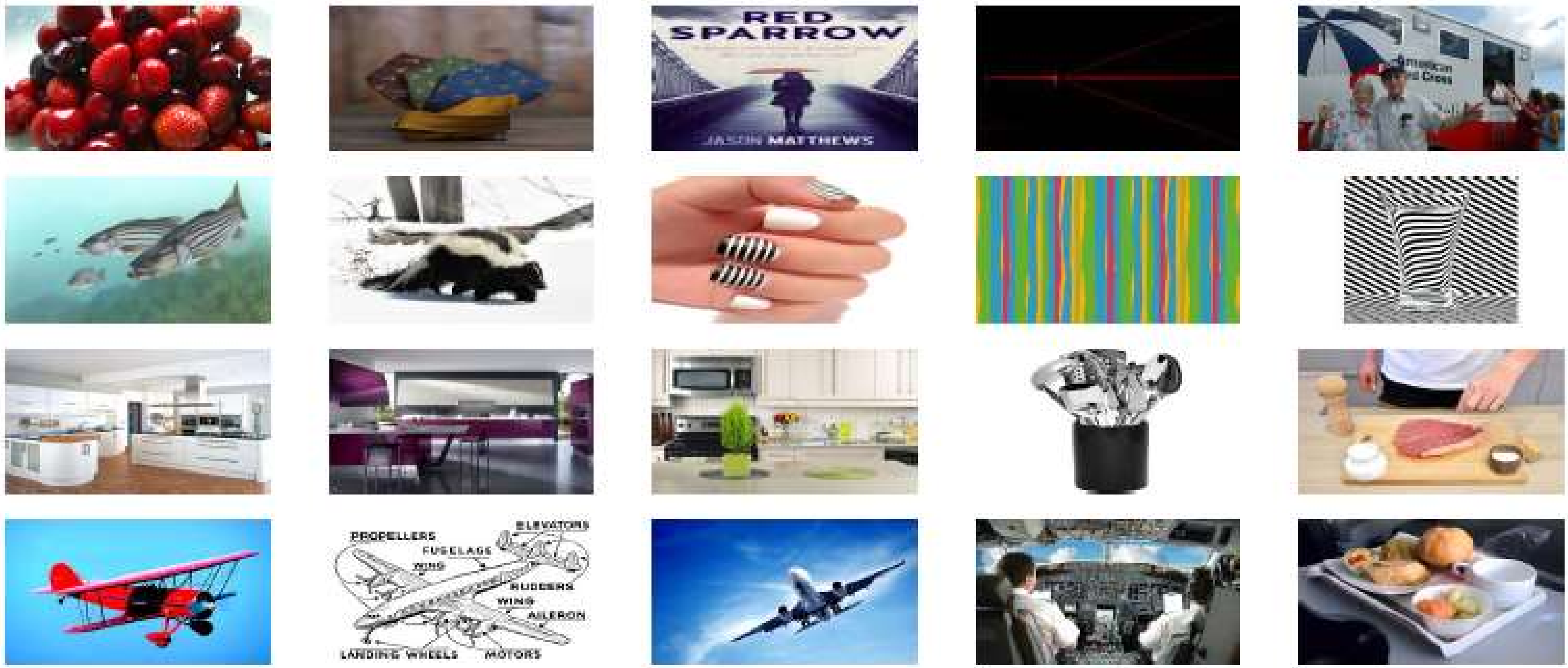
+ Finding Densest component



$$f(S) = \frac{|E(S)|}{|S|},$$

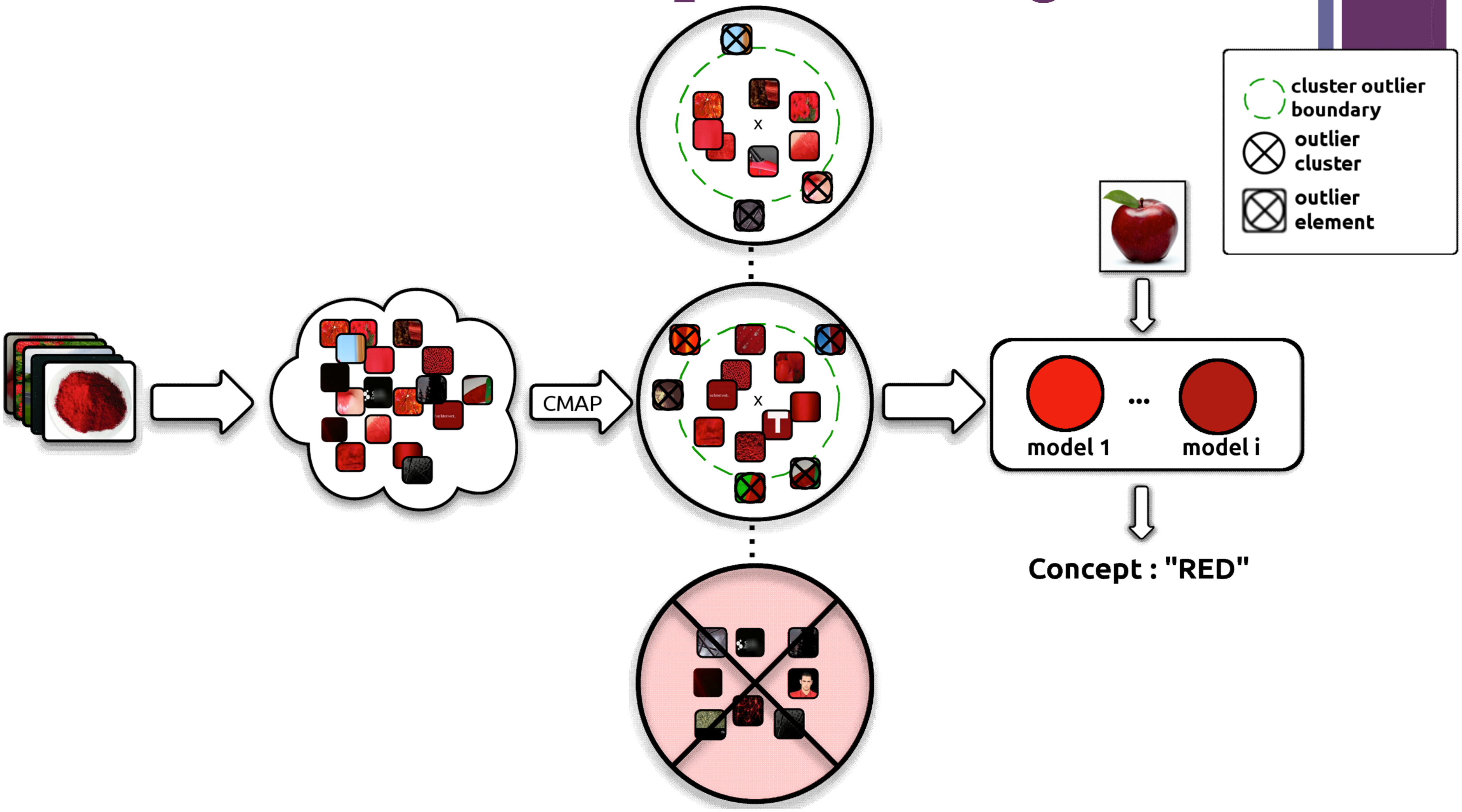
Node with the minimal degree is removed at each iteration (Charikar, 2000)

Multiple meanings/variations



The concepts are observed in different forms requiring grouping and irrelevant elements to be eliminated.

CMAP for Concept Learning



Golge, E., Duygulu, P., "Concept Maps: Mining Noisy Web Data for Concept Learning ", ECCV 2014

+ RSOM

Look **activation statistics** of each SOM unit in learning phase

Latter learning iterations are more **reliable**

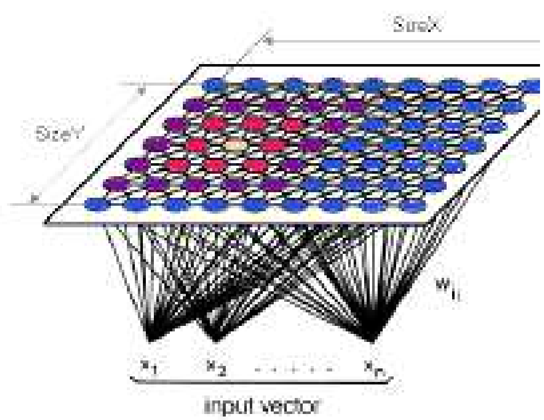
IF a unit is **activated**

REARLY → OUTLIER

FREQUENTLY → SALIENT

Winner activations

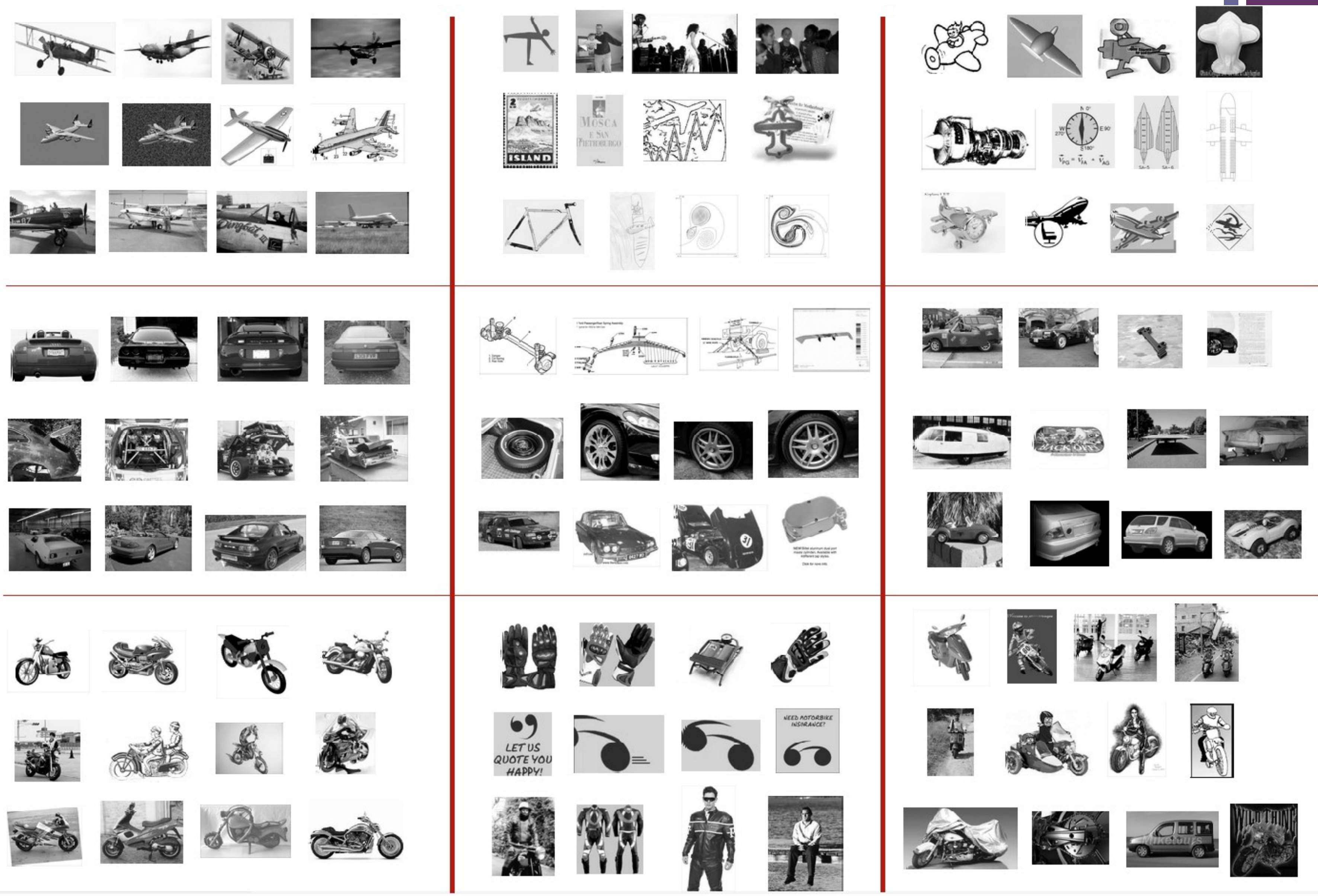
Neighbor activations



+ Color and Texture Attributes

+ Scene Concepts

+ Objects


ISLAND
MOSCA
E SAN
PIETROBURGO
LET US
QUOTE YOU
HAPPY!
NEED MOTORBIKE
INSURANCE?
WILD THING

Pi

+ Faces

+

AME

Association through Model Evolution

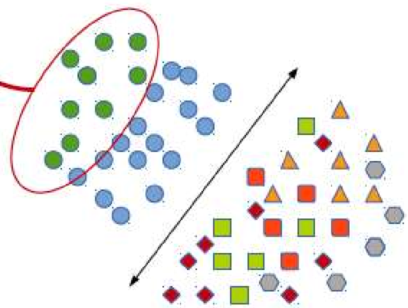
Capture **discriminative** and **representative** category images through **iterative data cleansing**

Separate **category instances** versus **random images**.

Golge, E., Duygulu, P., "FAME: Face Association through Model Evolution", CVPRW 2015

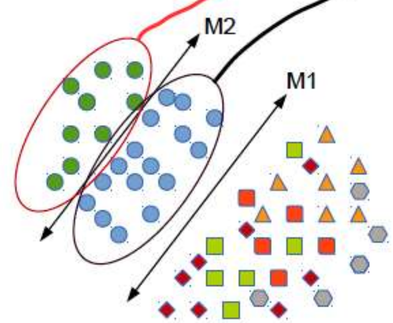
Step1

- Discerning category from random set
  - Learn a linear model **M1** between **CC** and **RS**.
  - Take the most confidently classified instances as tl
CR.



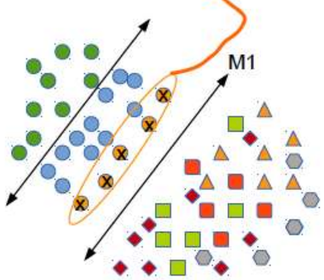
Step2

- Discerning category references from others
  - Learn linear model **M2** between **CR** and **others**.



Step3

- Define **SI** against **CR**.
- **Eliminate** SI.



AME's method overview

- First discern **category candidates** (CC) from **random set** (RS).
- Define **category references**(CR).
- Second discern **CR** from **CC**.
- Define **spurious instances** (SI) against **CR** and **eliminate**.
- **Re-Iterate**

+ Features

- Learn frequent pattern on the data
- Learning Pipeline (similar to [1]);

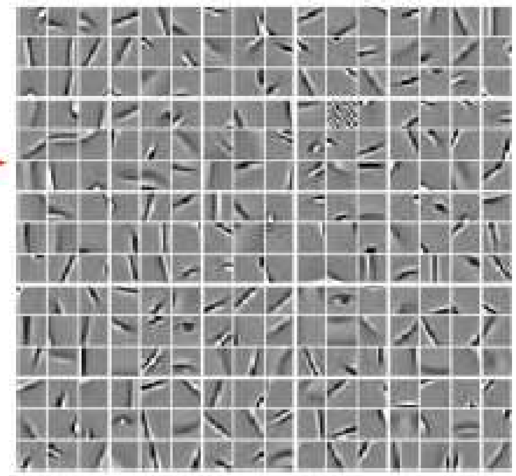
1. Scrap random nxn patches from the images.

<u>*Over Collected Patches;*</u>

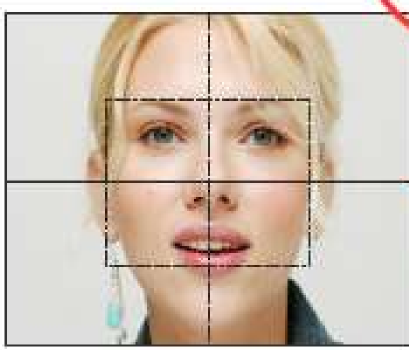
2. Contrast normalization
3. ZCA Whitening
4. K-means for <u>C words</u>

<u>*Over Whole Image;*</u>

5. Spatial (Max or Avg) **Pooling** by C words

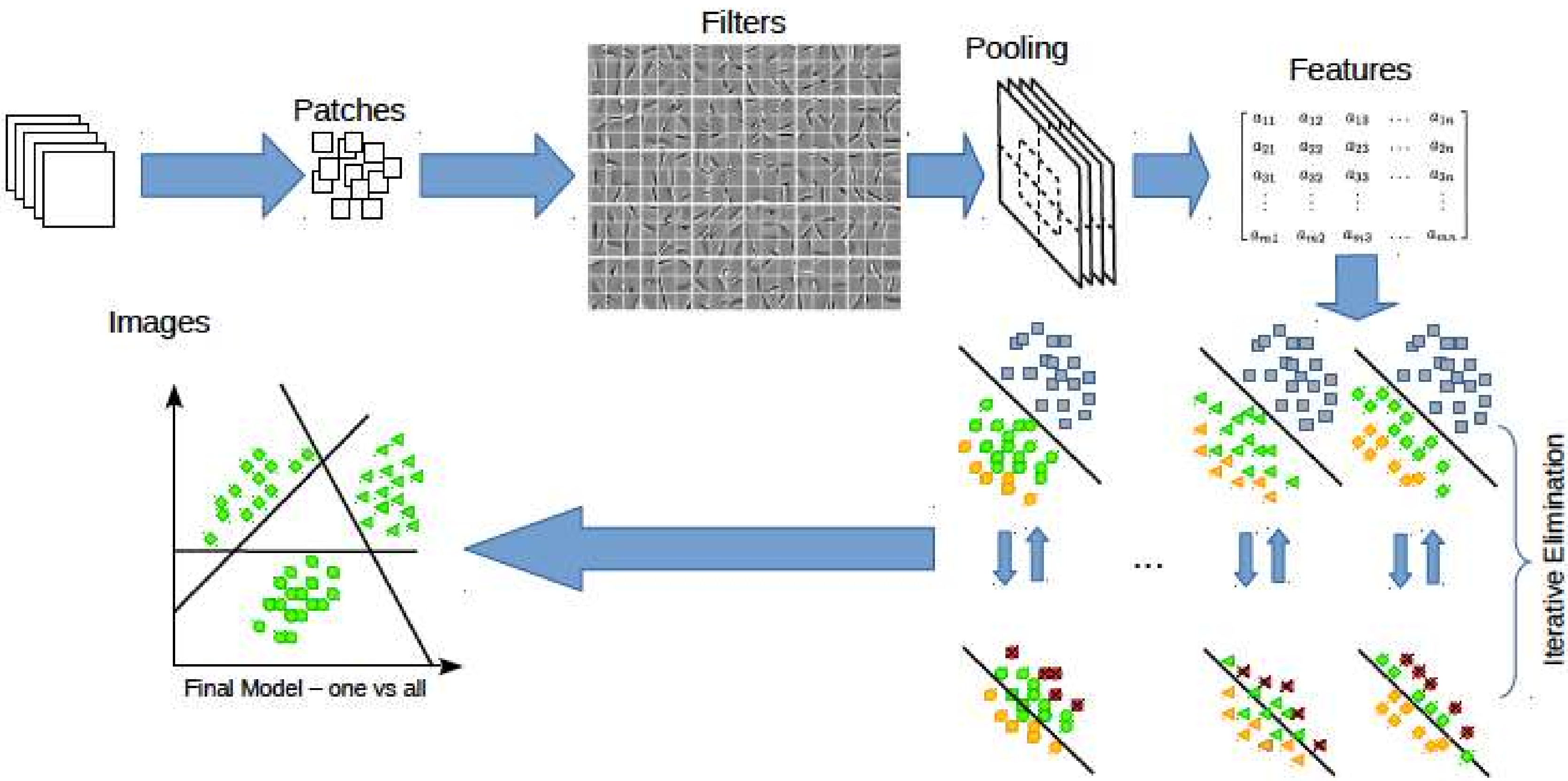


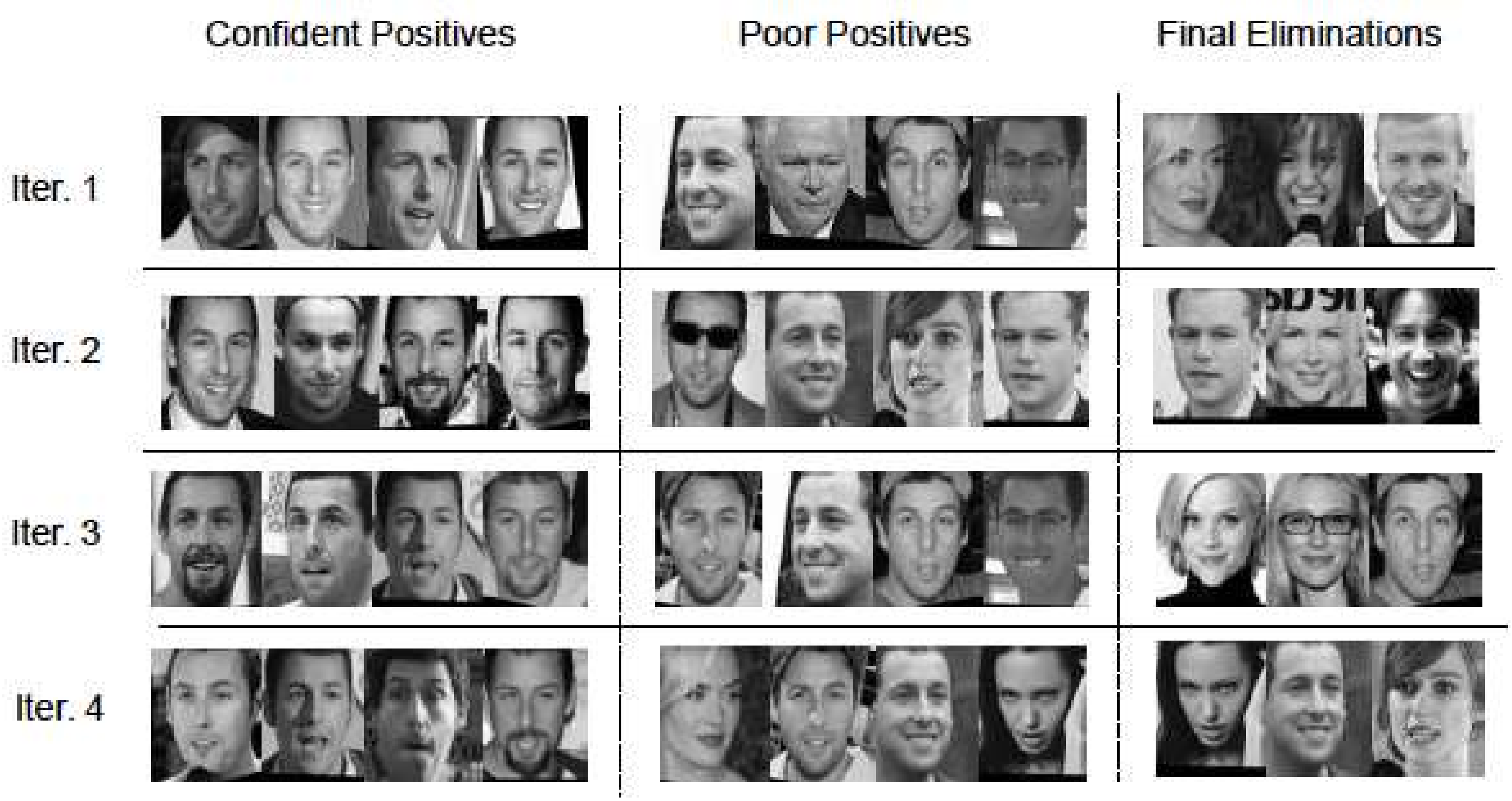
Learned Visual Words

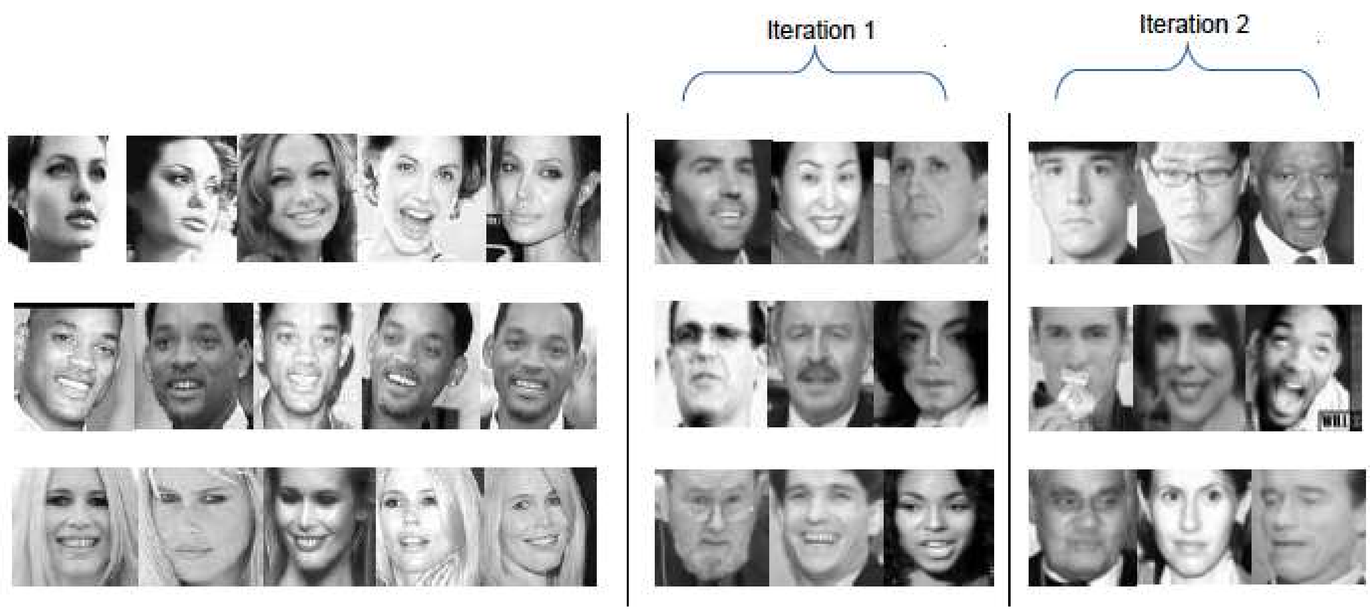


={ 5 x C words }
dimension for each img

FAME: Face Association Through Model Evolution







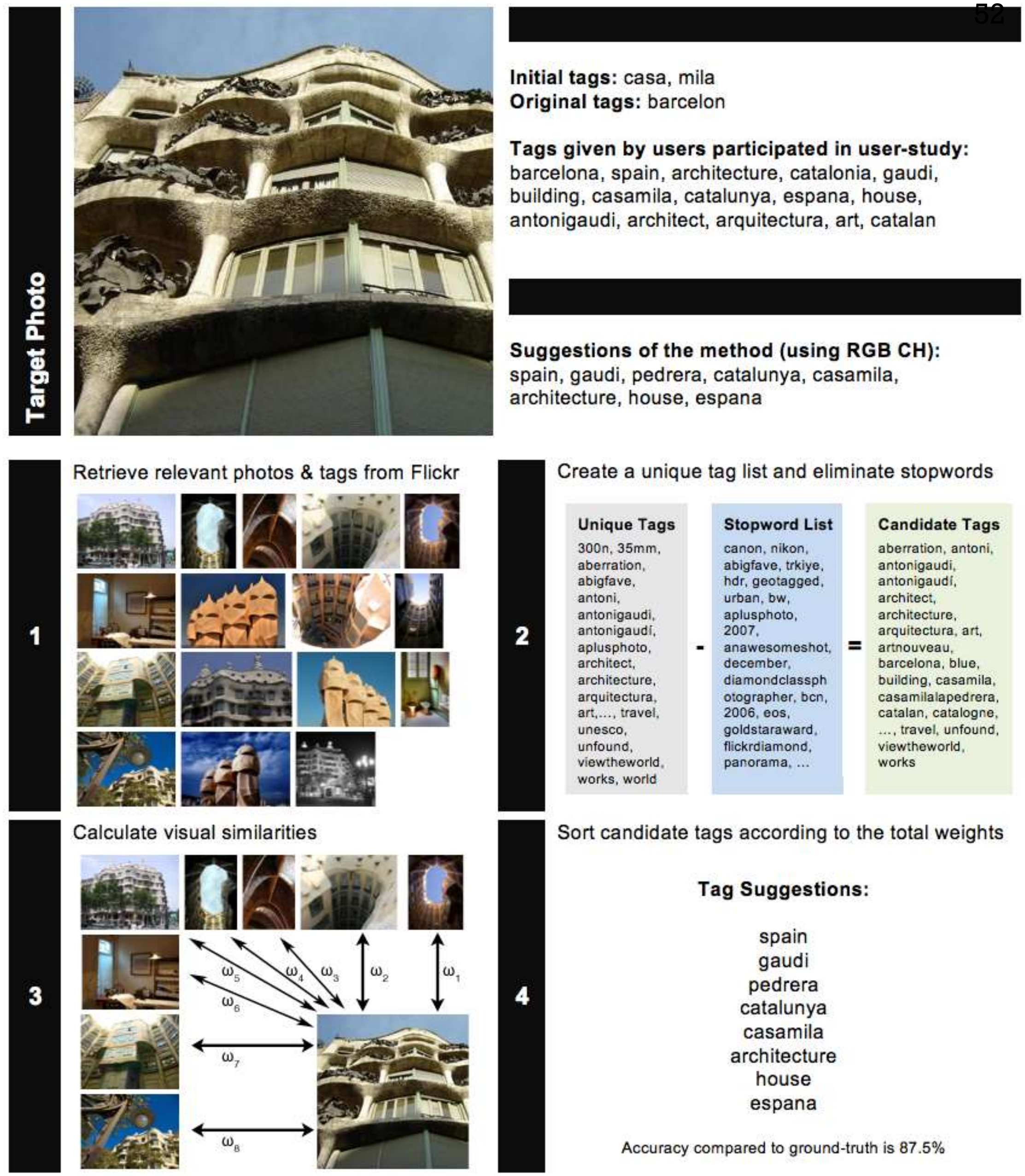
Pinar Duygulu, ENLG 2015

+ TagSuggestr

- Given a few initial tags predict more

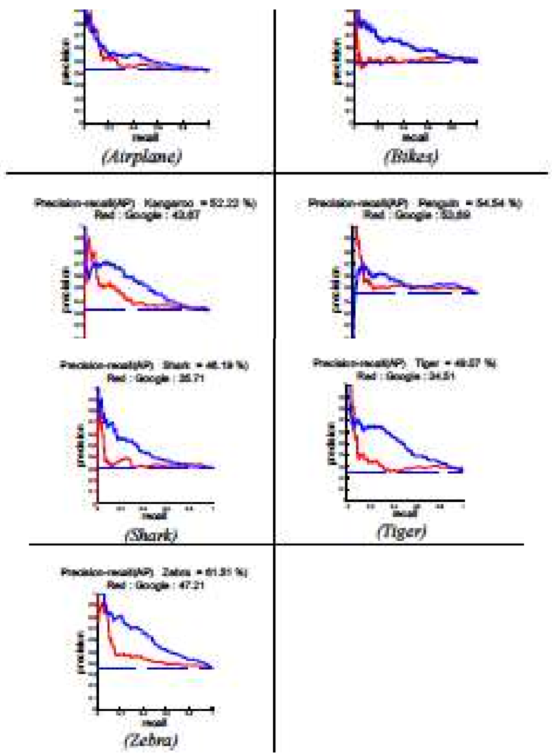
Give more weights to the visually similar images

$$W = \begin{array}{c} \\ I_1 \\ I_2 \\ \\ I_m \end{array} \begin{array}{c} \begin{array}{ccccc} t_1 & t_2 & t_3 & & t_n \end{array} \\ \left|\begin{array}{ccccc} 1 & 0 & 1 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ & & & & \\ 0 & 0 & 0 & \cdots & 1 \end{array}\right| \end{array} \begin{array}{c} \\ \times\ \omega_1 \\ \times\ \omega_2 \\ \\ \times\ \omega_m \end{array} \Bigg\} \sum = \begin{array}{c} t_1\ t_2\ t_3 \qquad t_n \\ \boxed{\ \ \ \ \ \ \ \ \ \ \ } \end{array}$$

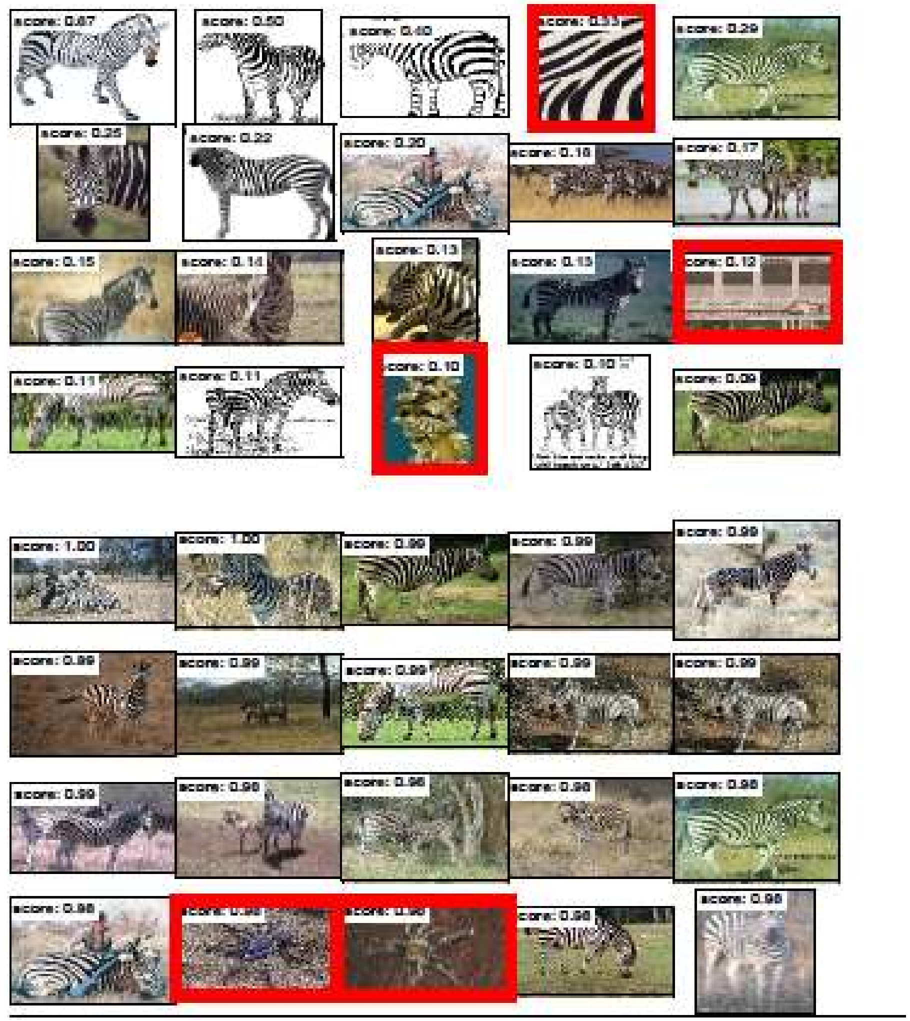


Sevil, S., Kucuktunc, O., Duygulu, P., Can. F.,"Automatic Tag Expansion using Visual Similarity for Photo Sharing Websites", MTAP 2010

Multiple Instance Learning for re-ranking

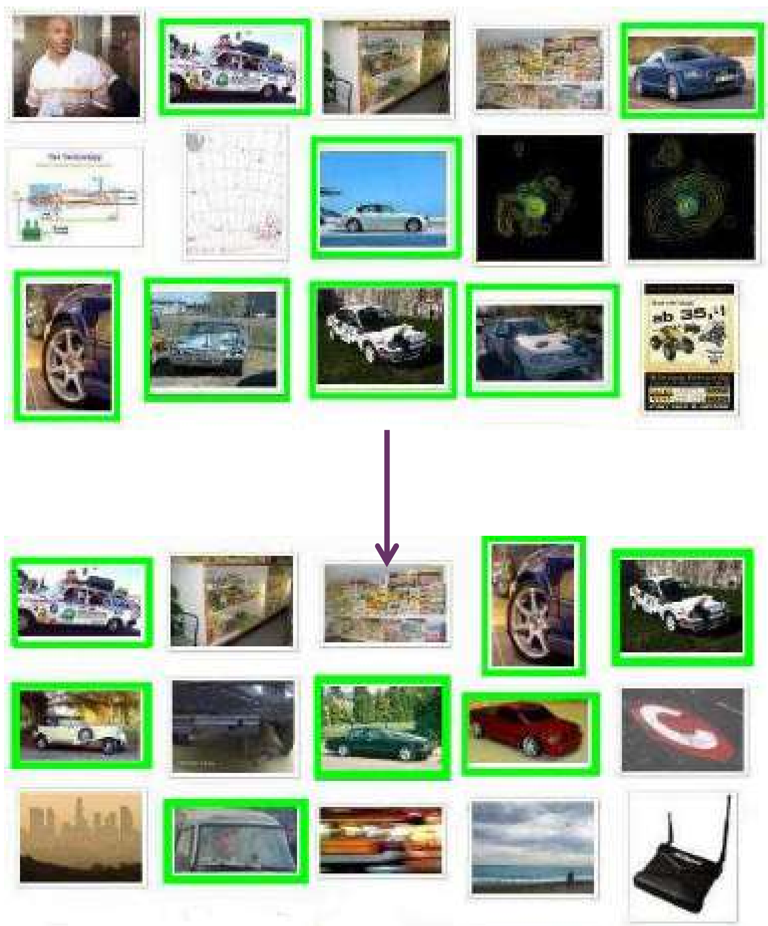


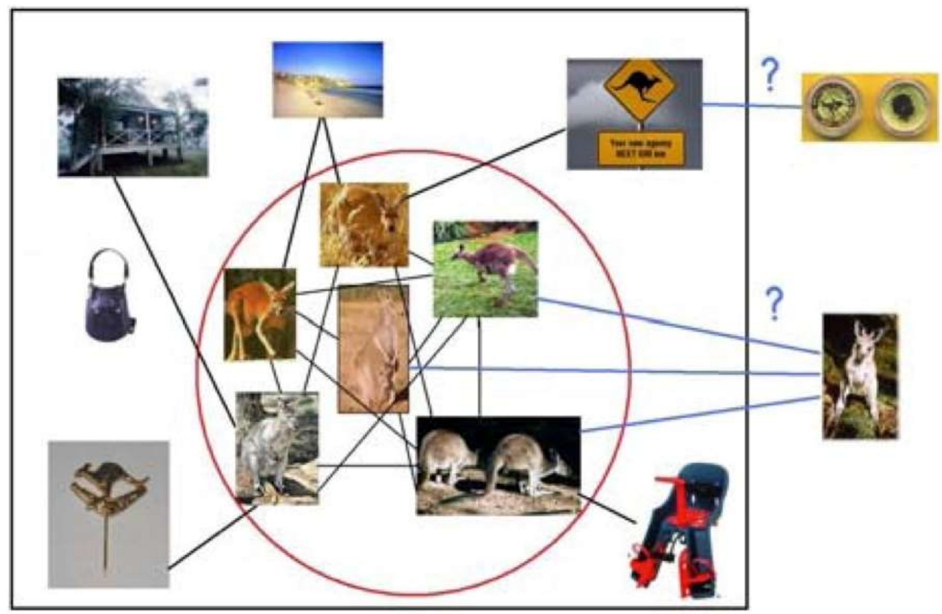
On the dataset by Schroff, F., ICCV 2007
"Harvesting Image Databases from the Web".



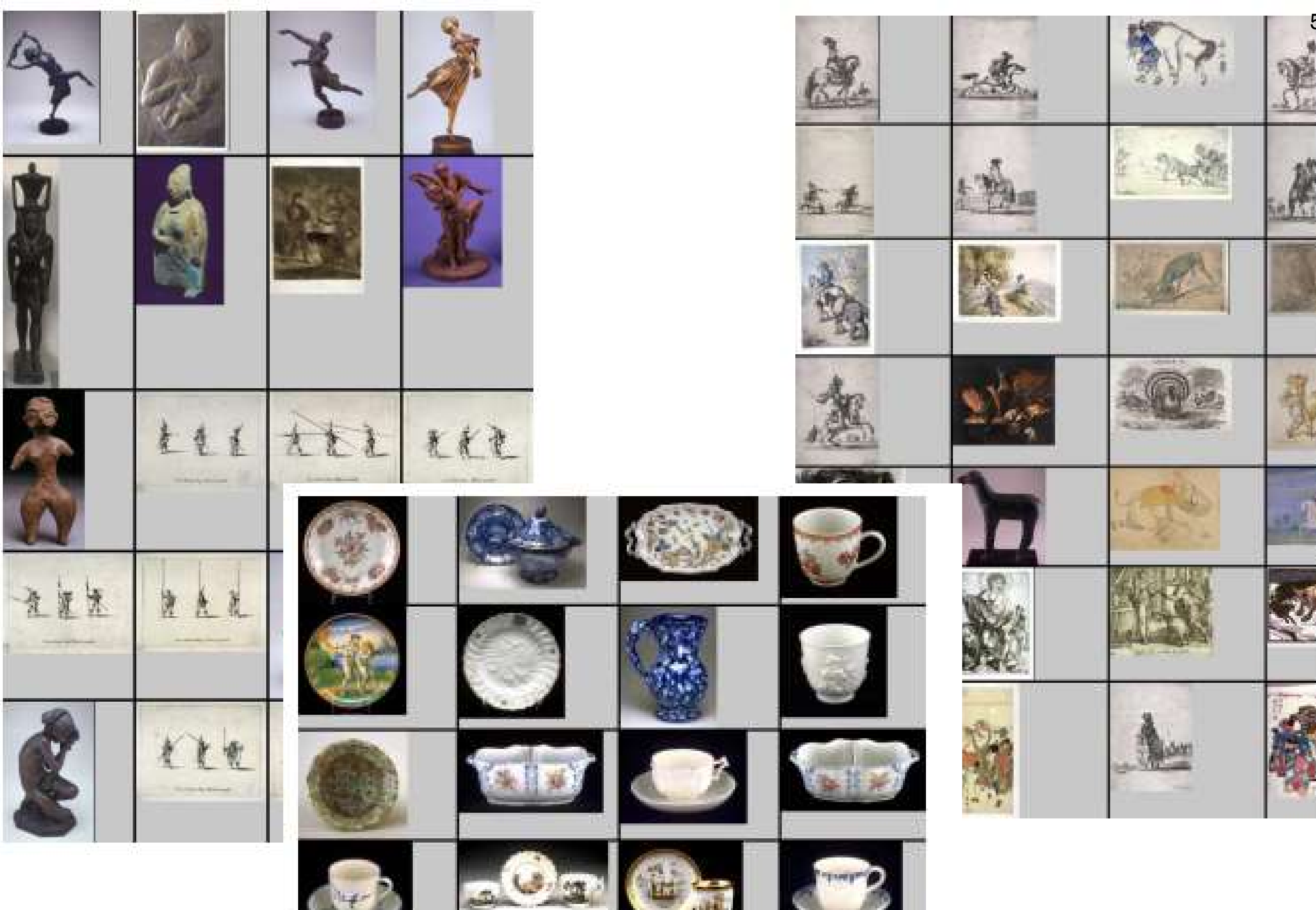
Sener, F., Ikizler-Cinbis, N., Duygulu, P., "Multiple Instance Learning for re-ranking of Web image search results", SIU 2012

Image Re-ranking





Zitouni, H., Sevil, S. G., Ozkan, D., Duygulu, P., "Re-ranking of Image Search Results using a Graph Algorithm", ICPR 2008



K. Barnard, P. Duygulu, D. Forsyth, "Clustering Art", CVPR 2001

Auto Illustration

large importance attached fact old dutch century more command whale ship was per son was divided officer word means fat cutter time made days was general vessel whale hunting concern british title old dutch official present rank such more good american officer boat night watch ground command ship deck grand political sea men mast way professional superior

"The large importance attached to the harpooneer's vocation is evinced by the fact, that originally in the old Dutch Fishery, two centuries and more ago, the command of a whale-ship was not wholly lodged in the person now called the captain, but was divided between him and an officer called the Specksynder. Literally this word means Fat-Cutter; usage, however, in time made it equivalent to Chief Harpooneer. In those days, the captain's authority was restricted to the navigation and general management of the vessel; while over the whale-hunting department and all its concerns, the Specksynder or Chief Harpooneer reigned supreme. In the British Greenland Fishery, under the corrupted title of Specksioneer, this old Dutch official is still retained, but his former dignity is sadly abridged. At present he ranks simply as senior Harpooneer; and as such, is but one of the captain's more inferior subalterns. Nevertheless, as upon the good conduct ..."

K. Barnard, P. Duygulu, D. Forsyth, "Clustering Art", CVPR 2001

+

Query on "president"

Association problem

Pinar Duygulu and Alex Hauptmann, What's news, what's not? Associating News videos with words, CIVR 2004

Concepts or Free text

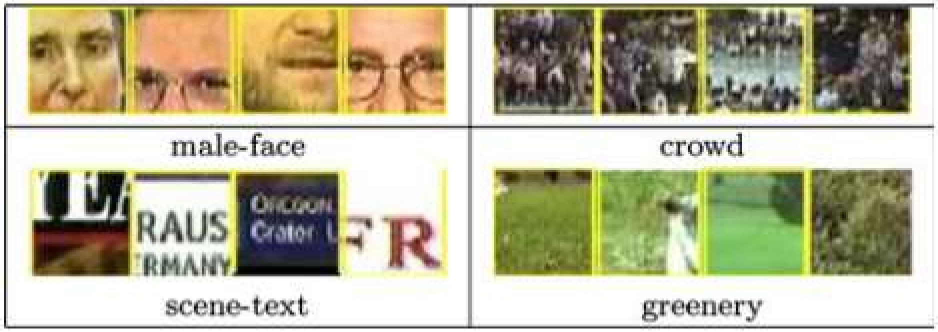
Concepts

Requires manual annotation
Noisy
Limited set of vocabulary

Speech transcripts and closed captions

Available for almost all the videos
Free text which usually does not correspond to the visual cues
Text is not associated with the frames

…despite heroic efforts many of the worlds wild creatures are doomed the loss of species is now the same as when the great dinosaurs become extinct will these creatures become the dinosaurs of our time today…



	
snow road car	*building graphics*
outdoors car road male-news-subject snow	building graphics outdoors graphics-and-text scene-text
	
female-person overlaid-text *head-and-shoulder road face* *windows single-person-female* *reporters daytime-outdoor*	*cityscape politics* *runway overlaid-text* *daytime-outdoor* *building outdoor*
overlaid-text face daytime-outdoor outdoor head-and-shoulder building female-person vehicle	outdoor daytime-outdoor overlaid-text face building sky crowd suits

ASR : weather headline weather thunderstorm texas arkansas cold pressure shower lake ...
PREDICTED : temperature weather thunderstorm pressure shower southeast forecast snow coast lake ...

ASR : florida home home home game
PREDICTED : ball technology play sport game baseball

+ What do these people do?

running


walking

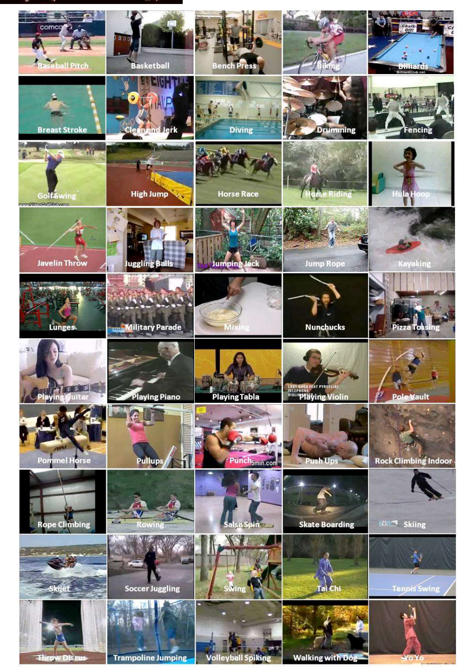
throwing


crouching

Ikizler, N. Duygulu, P. "Human Action Recognition Using Distribution of Oriented Rectanguar Patches", Proc. 2nd Workshop on Human Motion: Understanding, Modeling, Capture and Animation, In conjunction with ICCV2007
Ikizler, N. ve Duygulu P., "Histogram of Oriented Rectangles: A New Pose descriptor for Human Action Recognition", Image and Vision Computing, volume 27, Issue 10, pages 1515-1526, September 2009

Available Datasets

Dataset #Clas
KTH 6
Weizmann 9
IXMAS
Hollywood
UCF Sports
Hollywood2
UCF YouTube 11
MSR
Olympic
UCF50
HMDB51



http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

+ Videos in the wild

- Unrestricted type of events with various activities

Harlem Shake : http://www.youtube.com/watch?v=4hpEnLtqUDg

Multimedia Event Detection Birthday event

Blowing candles

What, where and who? Classifying events by scene and object recognition

L-J Li & L. Fei-Fei, ICCV 2007

Beyond Labels

car

pink car

car on road

Little pink smart car parked on the side of a road in a London shopping district.

Slide credit Tamara Berg

Baby Talk: Understanding and Generating Simple Image Descriptio

"This picture shows one person, one grass, one chair, and one pottec

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, Tamara L Berg, CVPR 2011

Slide credit Tamara Berg

+

“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.”

Some good results

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.

This is a picture of two dogs. The first dog is near the second furry dog.

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.

Slide credit Tamara Berg

+ Some bad results

Missed detections:

Here we see one potted plant.

This is a picture of one dog.

False 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

Slide credit Tamara Berg

+ Us vs Humans

"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."

↕

H1: A Lemonaide stand is manned by a blonde child with a cookie.
H2: A small child at a lemonade and cookie stand on a city corner.
H3: Young child behind lemonade stand eating a cookie.

- Sounds unnatural

UIUC pascal sentence dataset
Rashtchian, Young, Hodosh and Hockenmaier
NAACL HLT 2010

Slide credit Tamara Berg

Composing captions
guessing game

a) monkey playing in the tree canopy, Monte Verde in the rain forest

b) capuchin monkey in front of my window

c) monkey spotted in Apenheul Netherlands under the tree

d) a white-faced or capuchin in the tree in the garden

e) the monkey sitting in a tree, posing for his picture

Slide credit Tamara Berg

Captions in the Wild

+ http://tamaraberg.com/sbucaptions

The Egyptian cat statue by the floor clock and perpetual motion machine in the pantheon


Little girl and her dog in northern Thailand. They both seemed interested in what we were doing


Interior design of modern white and brown living room furniture against white wall with a lamp hanging


Man sits in a rusted car buried in the sand on Waitarere beach


Our dog Zoe in her bed

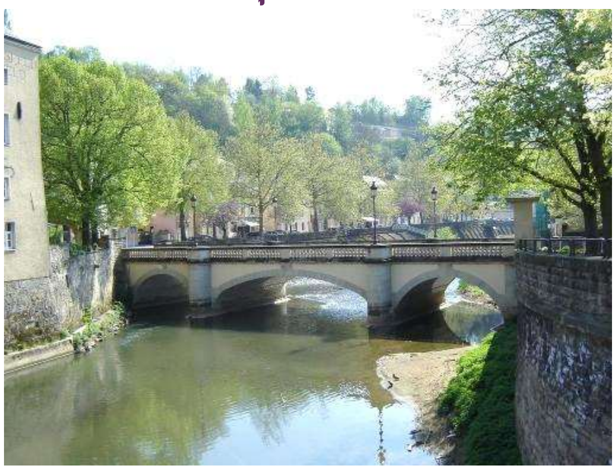

Emma in her hat looking super cute

+Harness the Web

Ordonez et al, NIPS 2011

Global Matching

Captioned Photo Dataset
1 *million* captioned images!



The bridge over the lake on Suzhou Street.

Bridge to temple in Hoan Kiem lake.

A walk around the lake near our house with Abby.

Transfer Whole Caption(s)

e.g. "The bridge over the lake on Suzhou Street."

Smallest house in paris between red (on right) and beige (on left).

Hangzhou bridge in West lake.

The daintree river by boat.

Slide credit Tamara Berg

Transfer pieces of Captions

Kuznetsova et al, ACL 2012

Object appearance → NP: the dirty sheep

Object pose → VP: meandered along a desolate road

Scene appearance → PP: in the highlands of Scotland

Region appearance & relationship → PP: through frozen grass

Example Composed Description: the dirty sheep meandered along a desolate road in the highlands of Scotland through frozen grass

Slide credit Tamara Berg

+

Object NPs

birds
the bird

Actions VPs

are standing
looking for food

Stuff PPs

in water
over water

Scene PPs

in the ocean
near Salt Pond

Position 1	Position 2	Position 3	Position 4
birds	over water	are standing	in the ocean

Dynamic Programming

Position1	Position2	Position3	Position4
birds	birds	birds	birds
the bird	the bird	the bird	the bird
are standing	are standing	are standing	are standing
…	…	…	…
in the ocean	in the ocean	in the ocean	in the ocean

+ ReferItGame

http://referitgame.com

Collecting referring expressions for objects in real world photos

Kazemzadeh et al, EMNLP 2014

+ ReferitGame Dataset

Collected: 130,525 expressions, referring to 96,654 objects, in 19,894 photographs

"picture on the wall"

"picture"

"picture"

"big gated window on right of white section"

"black big window right"

"brown railings on right"

"red guy left sitting"

"leftmost bottom guy"

"red shirt on left"

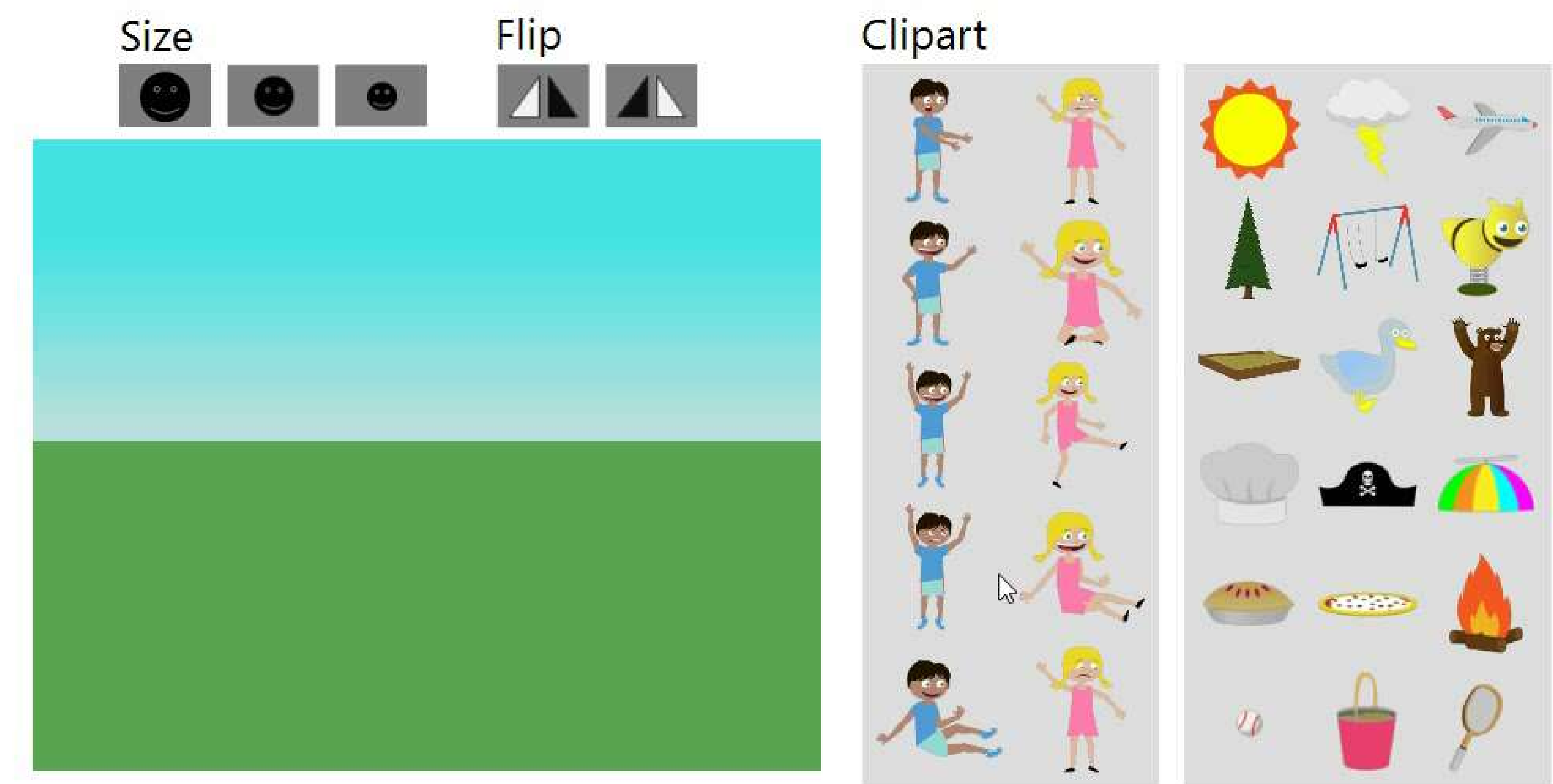
Kazemzadeh et al, EMNLP 2014

Abstract Scenes Dataset

Create a children's illustration!

Please help us create an illustration for a children's story book by creating a realistic scene from the clipart below. Use your imagination! Clipart may be added by dragging the clipart onto the scene, and removed by dragging it off. The clipart may be resized or flipped, and each clipart may only be added once. Please use at least 6 pieces of clipart in each scene. You will be asked to complete 3 different scenes. Press "Next" when finished with the current scene and "Done" when all are finished. Thanks!

Scene 1/3

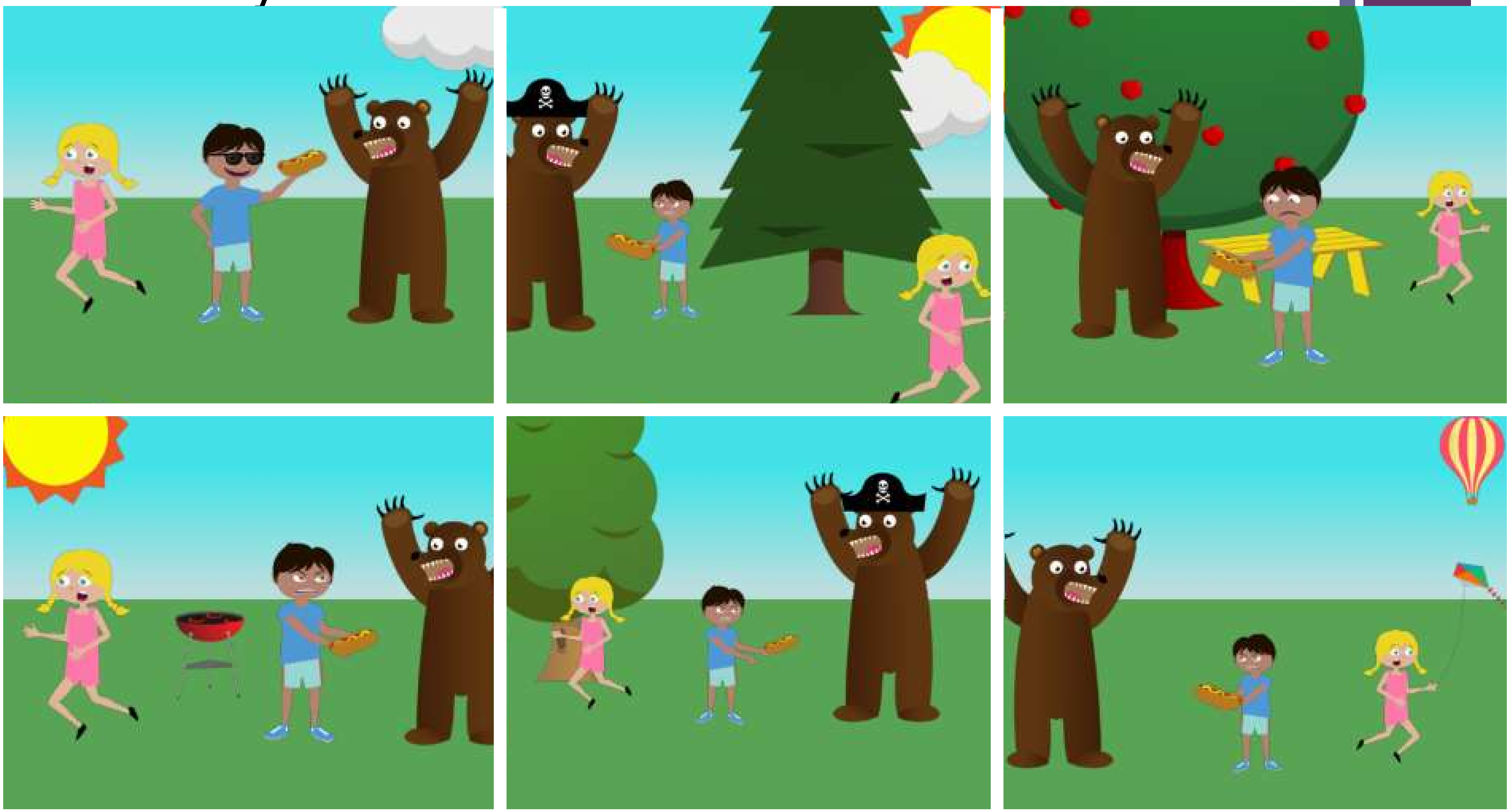


Generating sentences

Jenny loves to play soccer but she is worried that Mike will kick the ball too hard.

Mike and Jenny play outside in the sandbox. Mike is afraid of an owl that is in the tree.

Mike fights off a bear by giving him a hotdog while jenny runs away.



Visual features