Machine Learning in CV - Vision + Language ,Object Detection

CSE473/573

Descriptive Text

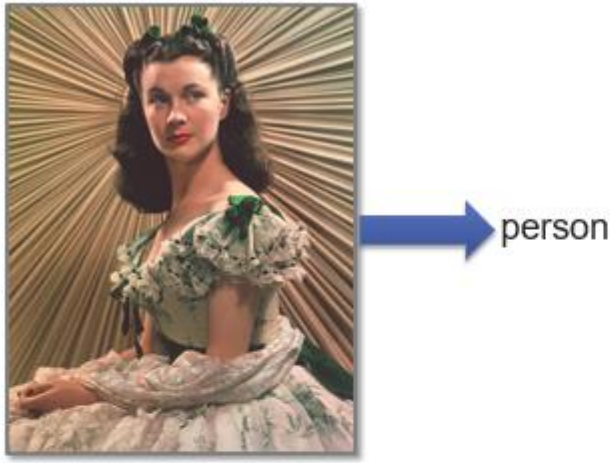
"It was an arresting face, pointed of chin, square of jaw. Her eyes were pale green without a touch of hazel, starred with bristly black lashes and slightly tilted at the ends. Above them, her thick black brows slanted upward, cutting a startling oblique line in her magnolia-white skin–that skin so prized by Southern women and so carefully guarded with bonnets, veils and mittens against hot Georgia suns"

Scarlett O'Hara described in Gone with the Wind.

Slide from T Berg.

Berg, Attributes Tutorial CVPR13

Computer Vision Algorithms…



person

car

shoe

Towards Complex Structured outputs

Towards Complex Structured outputs

pink car

Attributes of objects

Towards Complex Structured outputs

car on road

Relationships between objects

Towards Complex Structured outputs

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

… Complex structured recognition outputs

Telling the *"story of an image"*

Learning from descriptive text

"It was an arresting face, pointed of chin, square of jaw. Her eyes were pale green without a touch of hazel, starred with bristly black lashes and slightly tilted at the ends. Above them, her thick black brows slanted upward, cutting a startling oblique line in her magnolia-white skin–that skin so prized by Southern women and so carefully guarded with bonnets, veils and mittens against hot Georgia suns"

Scarlett O'Hara described in Gone with the Wind.

How does the world work?

Visually descriptive language provides:

- Information about the world, especially the visual world.
- information about how people construct natural language for imagery.
- guidance for visual recognition.

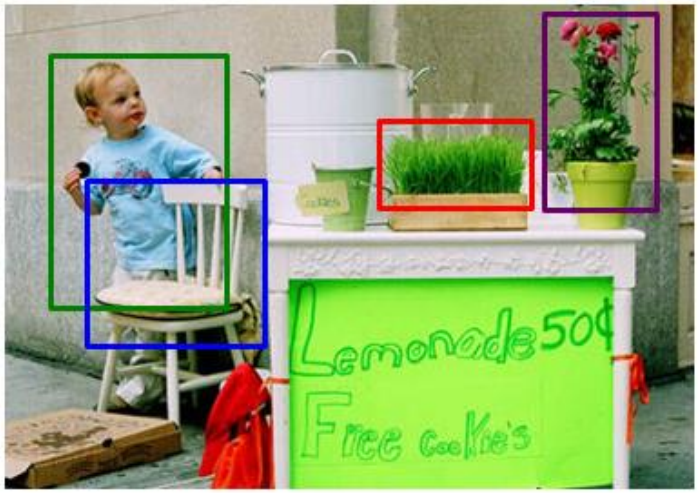
What should we recognize?

How do people describe the world?

Slide from T Berg.

Berg, Attributes Tutorial CVPR13

BabyTalk: Generating Sentences out of Images



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

Slide from T Berg.

Kulkarni et al, CVPR11

http://tamaraberg.com/papers/generation_cvpr11.pdf

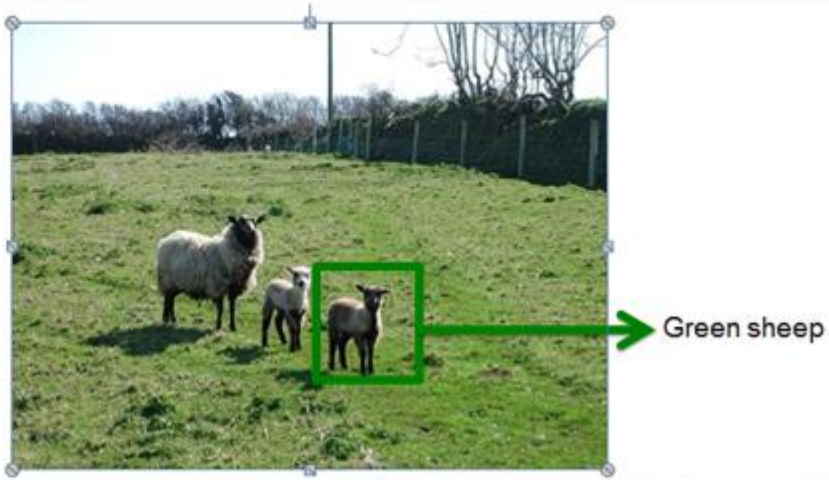
BabyTalk: Generating Sentences out of Images

"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."

Slide from T Berg.

Kulkarni et al, CVPR11

Need for Joint Visual and Lingual Model



- Vision is hard!
- World knowledge (from descriptive text) can be used to smooth noisy vision predictions!

Slide from T Berg.

Berg, Attributes Tutorial CVPR13

Methodology

- Vision -- detection and classification
- Text -- statistics from parsing lots of descriptive text
- Model (CRF) to predict best image labeling given vision and text based potentials
- Generation algorithms to compose natural language

Slide from T Berg.

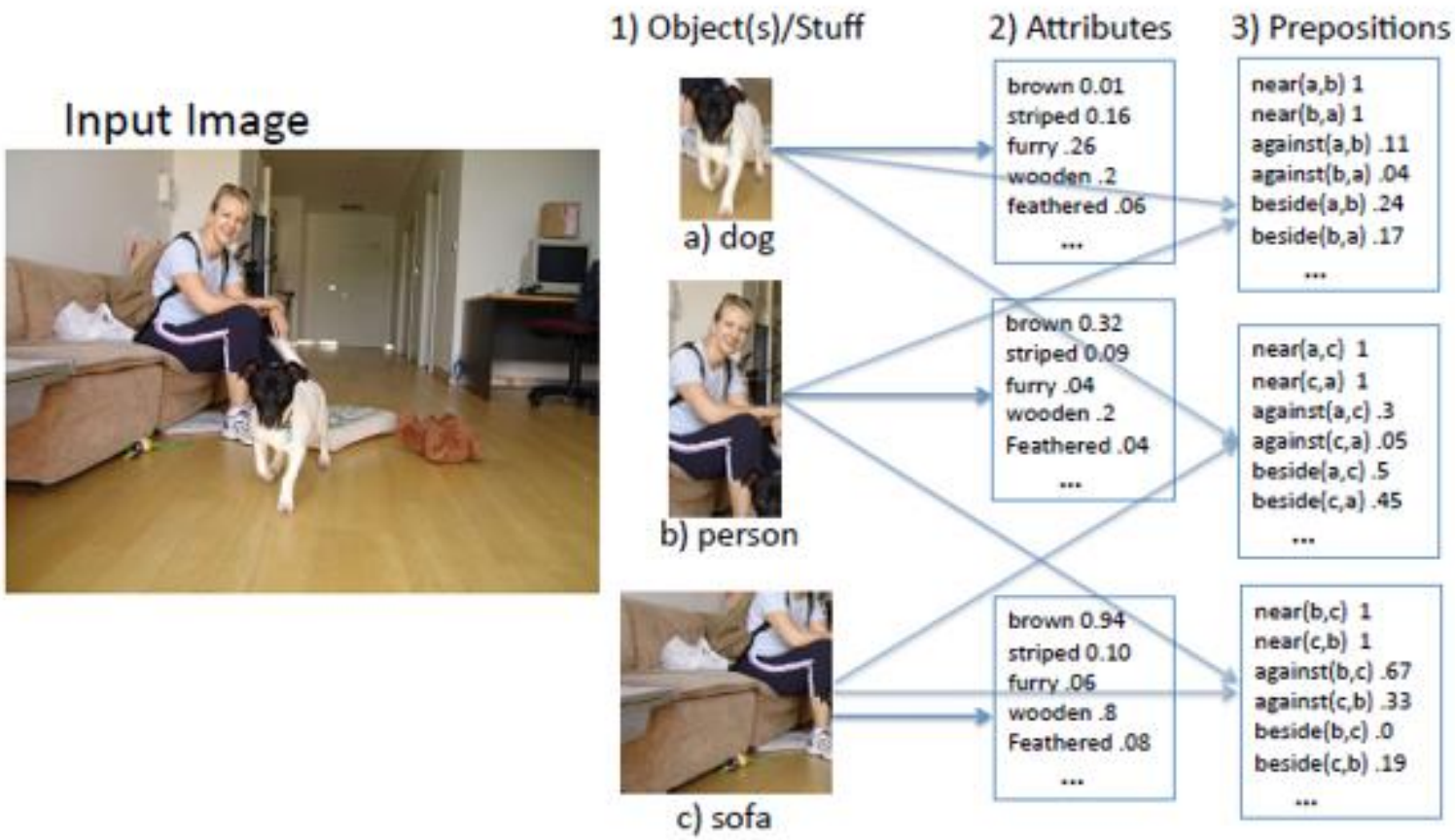
Kulkarni et al, CVPR11

BabyTalk - Capabilities

- Detects 24 objects, 6 stuff categories and 21 Visual attributes
- Augments sentence-generation using language statistics derived from large visually-descriptive corpus

"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."

Implementation Pipeline



Implementation Pipeline

4) Constructed CRF

5) Predicted Labeling

<<null,person_b>,against,<brown,sofa_c>>
<<null,dog_a>,near,<null,person_b>>
<<null,dog_a>,beside,<brown,sofa_c>>

6) Generated Sentences

This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

Experimental Results

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.

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

Experimental Results - Negative

Missing detections:

Here we see one pottedplant.

Incorrect detections:

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

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.

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.

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.

Vision + Text Applications

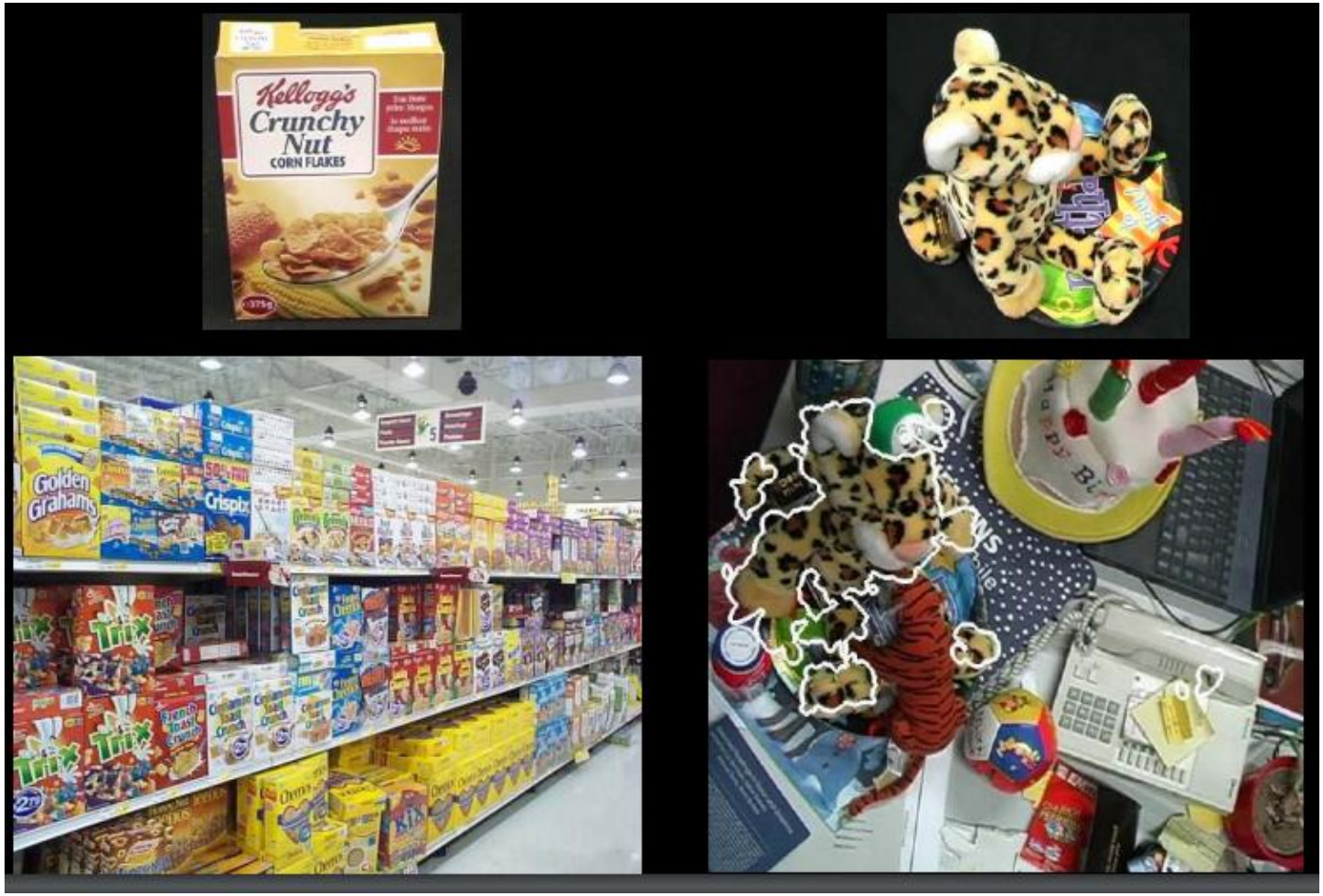
- Improvement of Image retrieval
- Surveillance
- Assisting visually impaired.

black, dog, car
"A **black dog** is sitting inside a car."

black, dog, car
"A dog is sitting inside a **black car**."

"Generating Descriptions for Images" – MS thesis, Ankush Gupta.

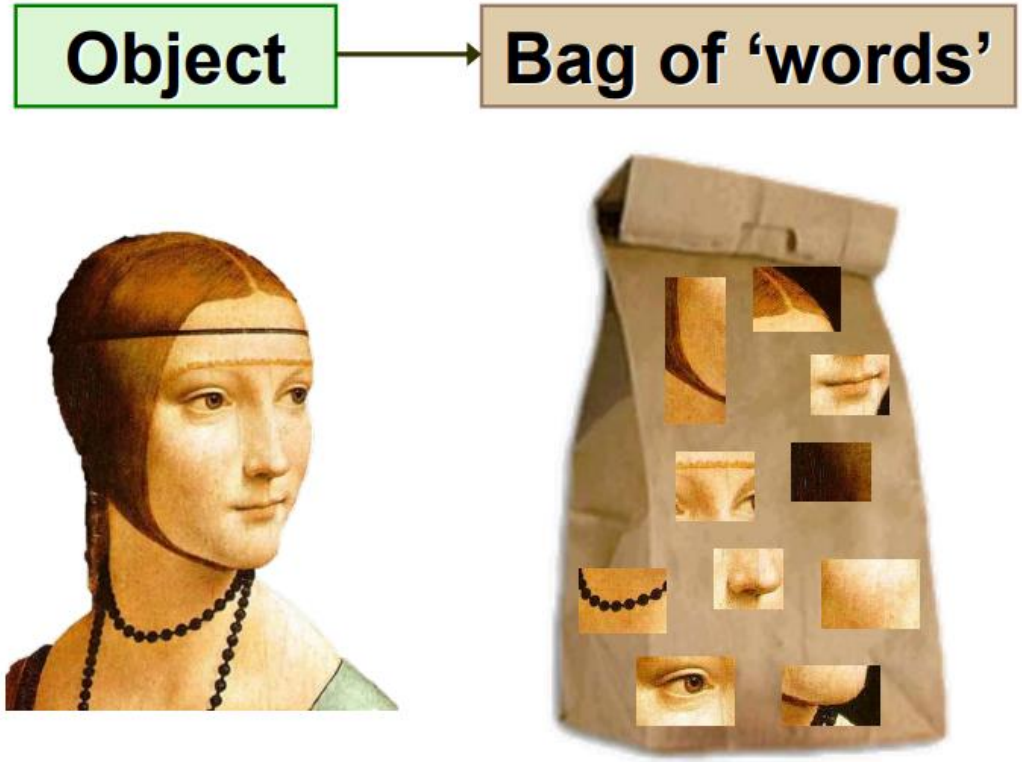
Single Object Detection



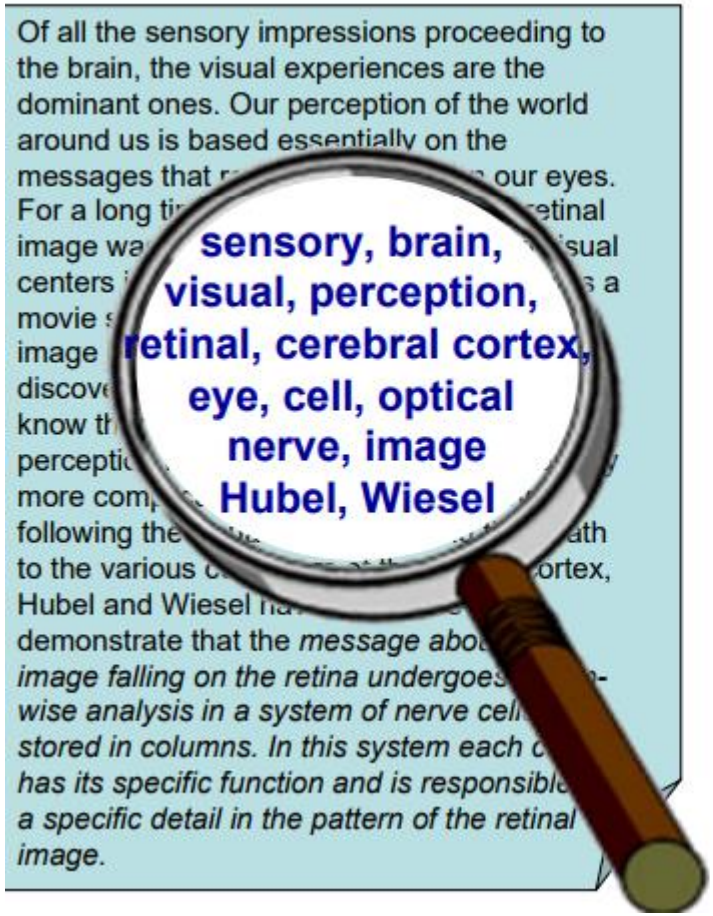
Single Object Detection

- Lowe, et al. 1999, 2003
- Mahamud and Herbert, 2000
- Ferrari, Tuytelaars, and Van Gool, 2004
- Rothganger, Lazebnik, and Ponce, 2004
- Moreels and Perona, 2005
- …

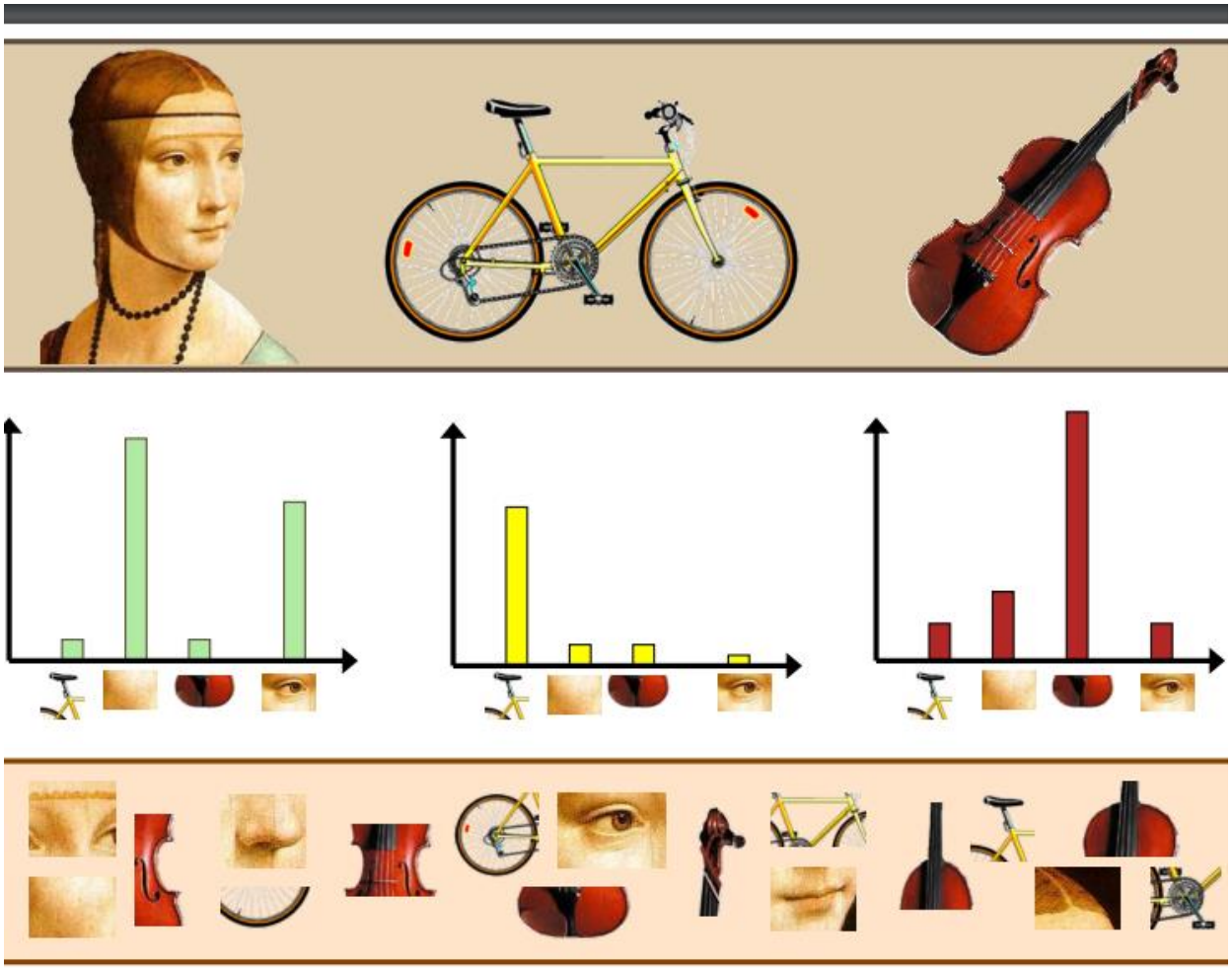
Bag of Words Model



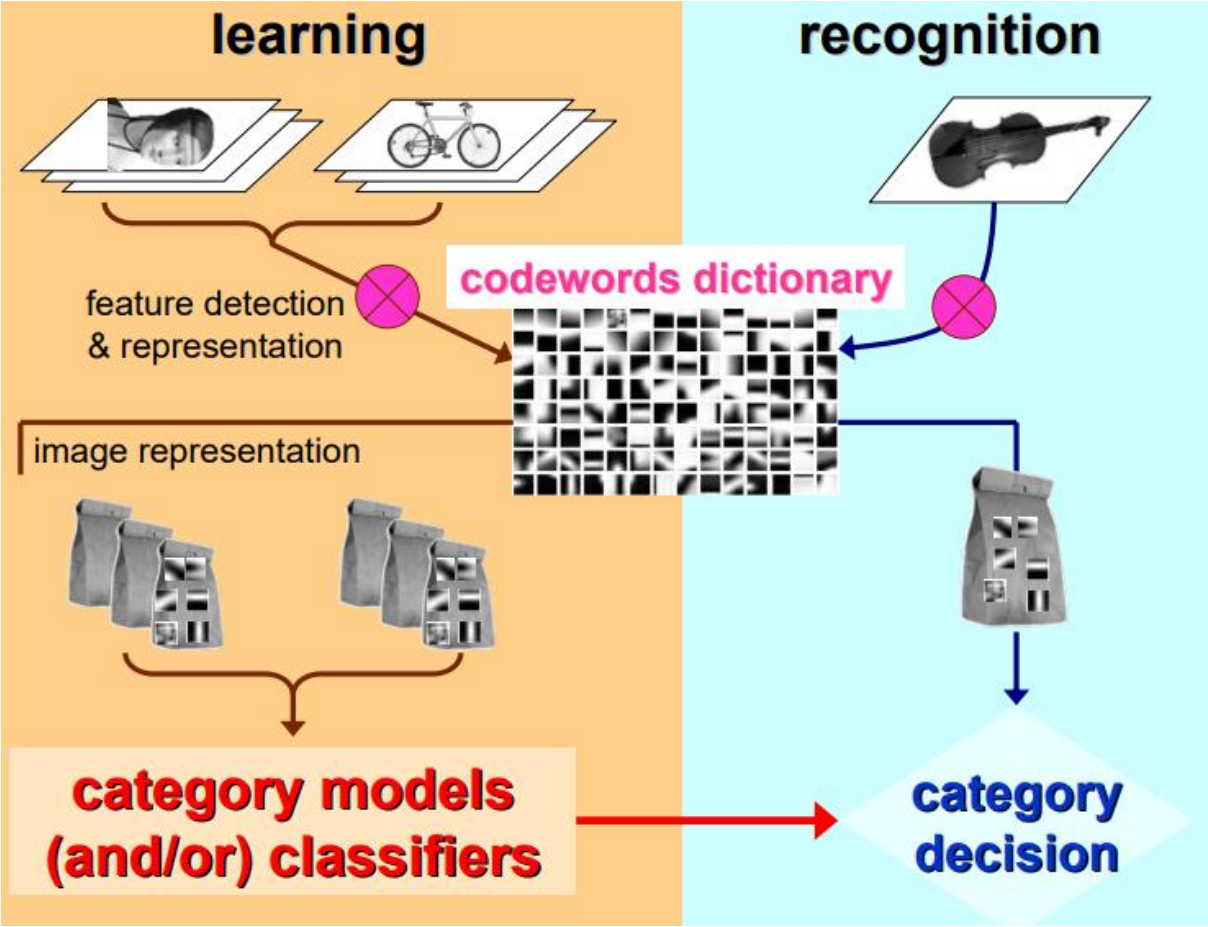
Analogy to Documents



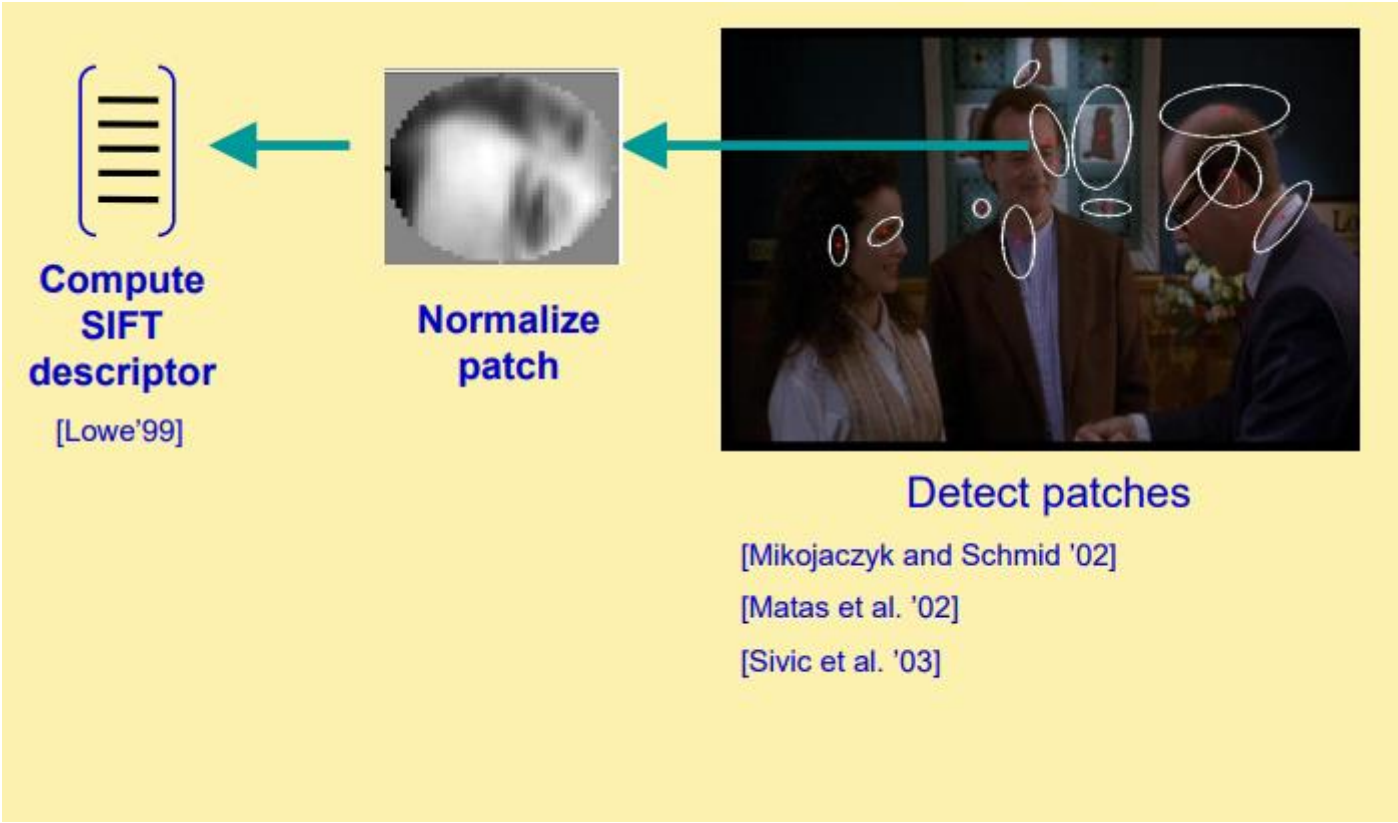
Toy Example for Bag of Words Approach



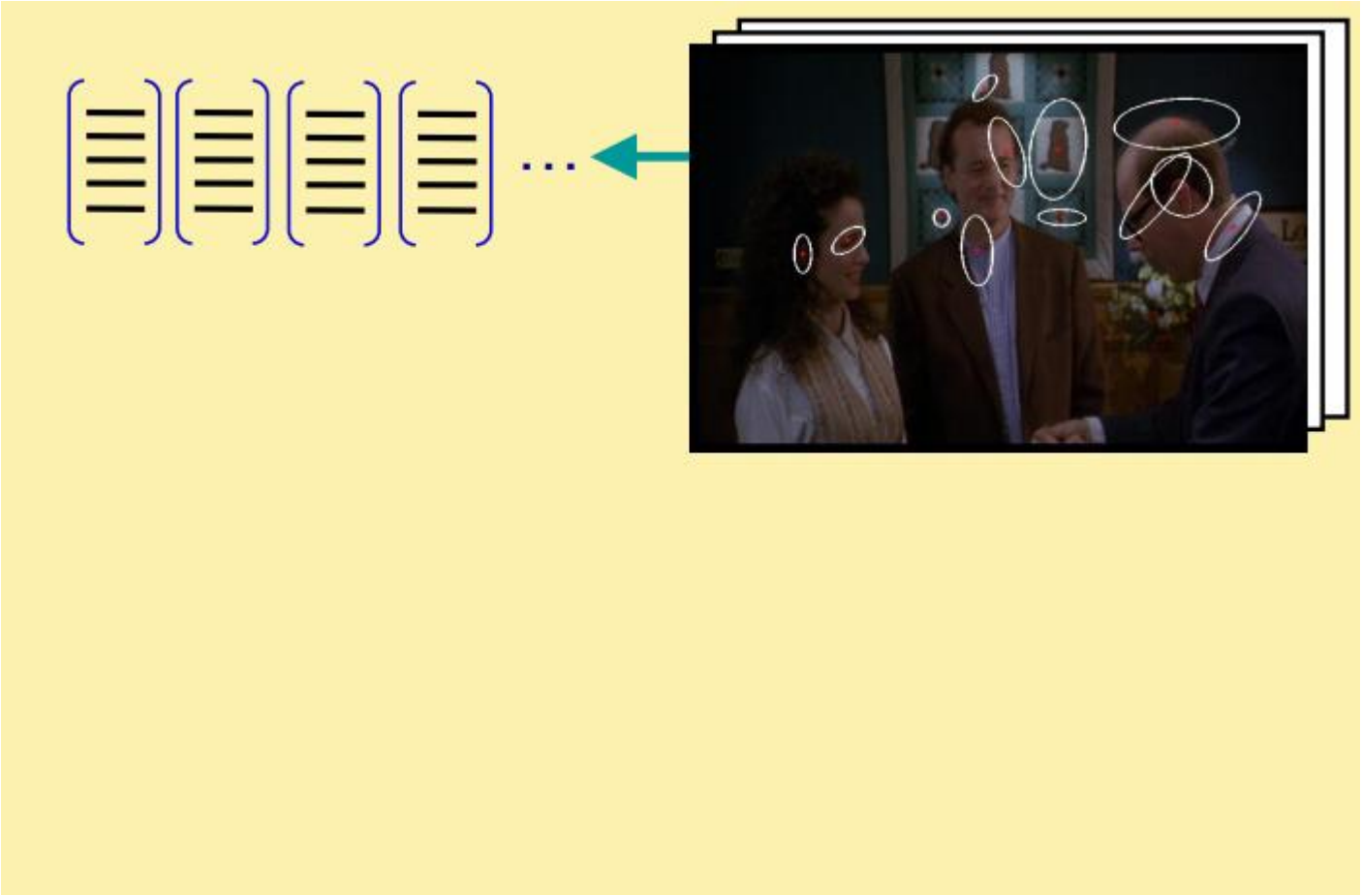
Bag of Words based detection



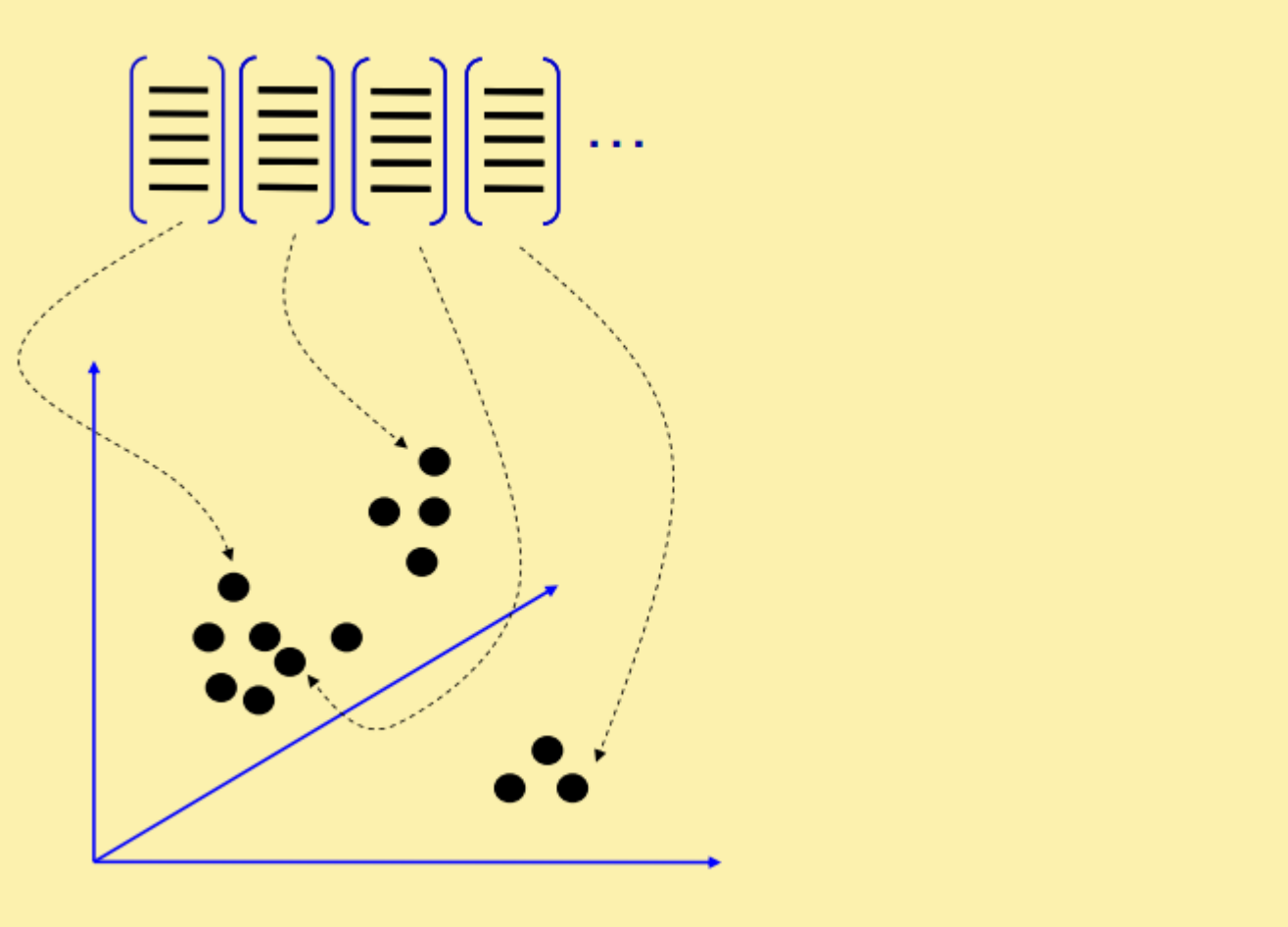
Feature detection



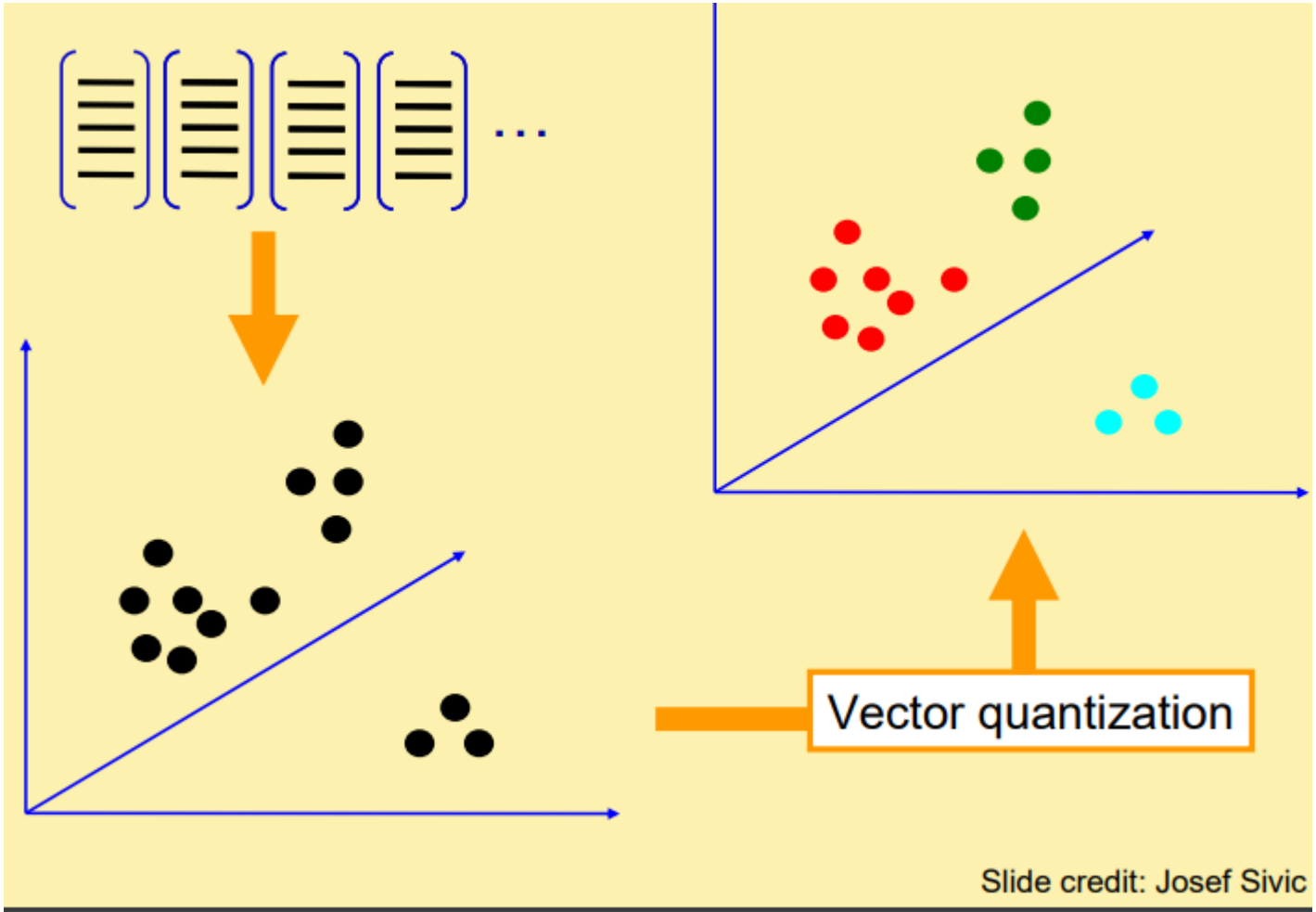
Feature detection



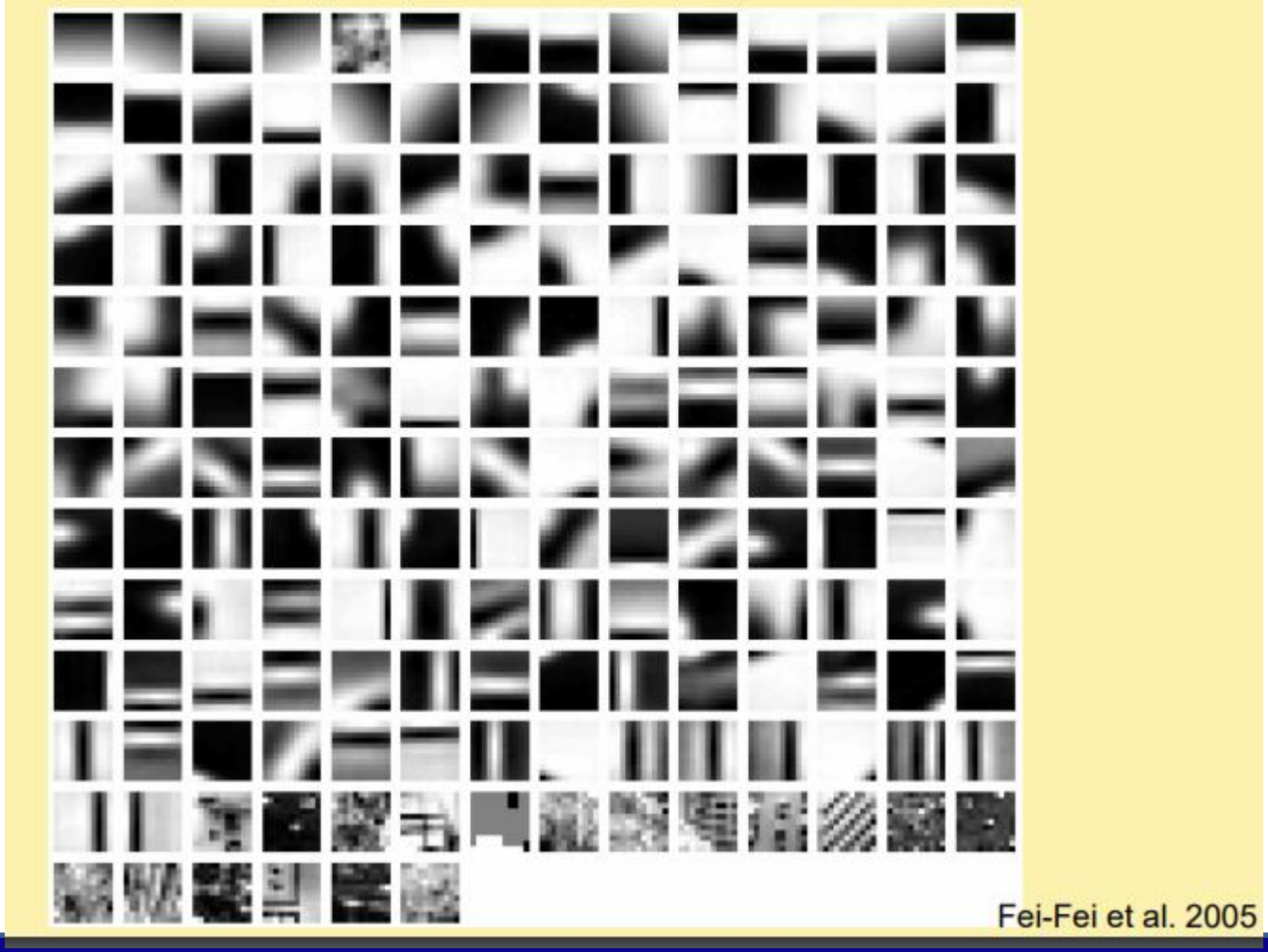
Codewords Dictionary Formation



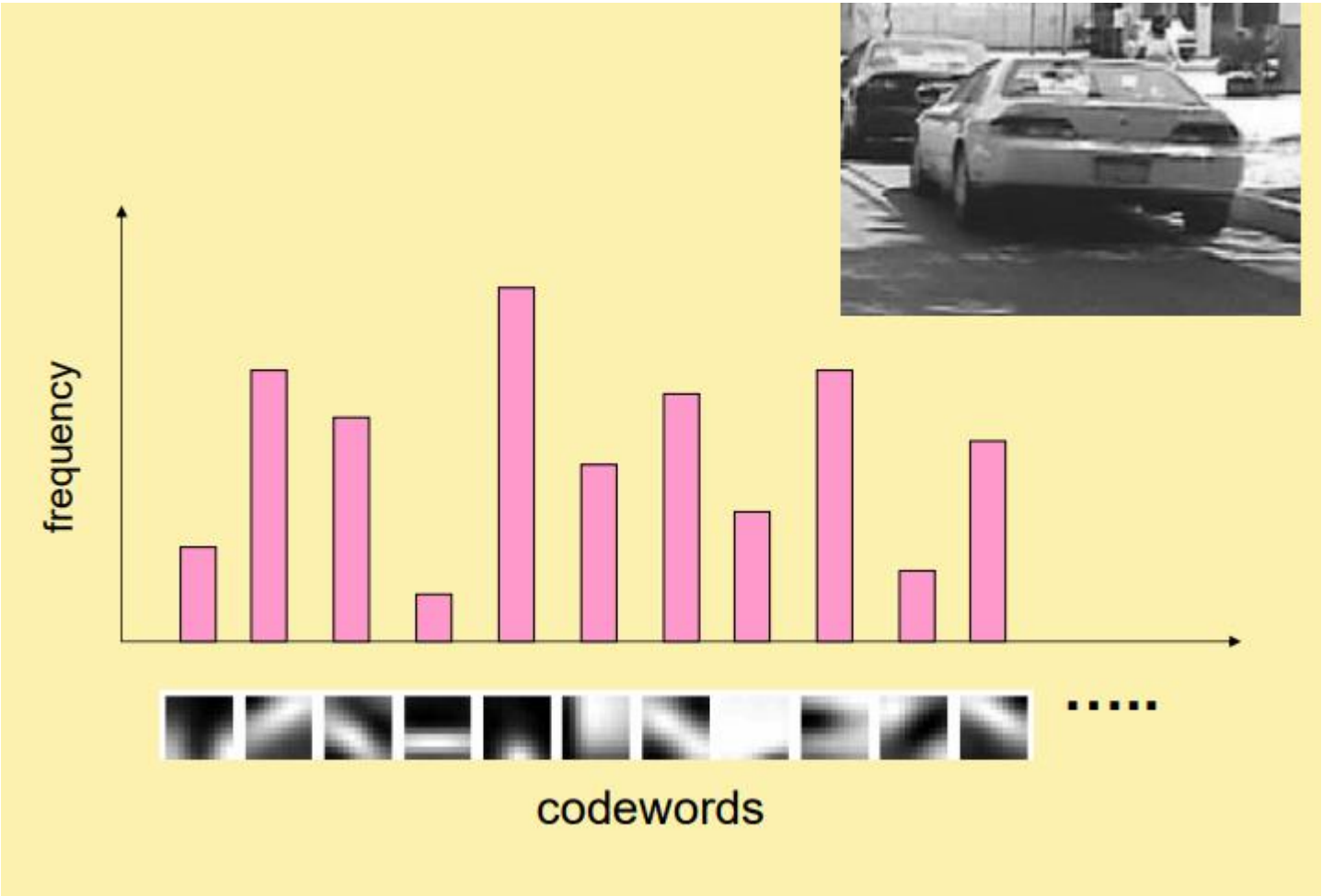
Codewords Quantization



Sample Codeword Dictionary



Bag of Words based detection



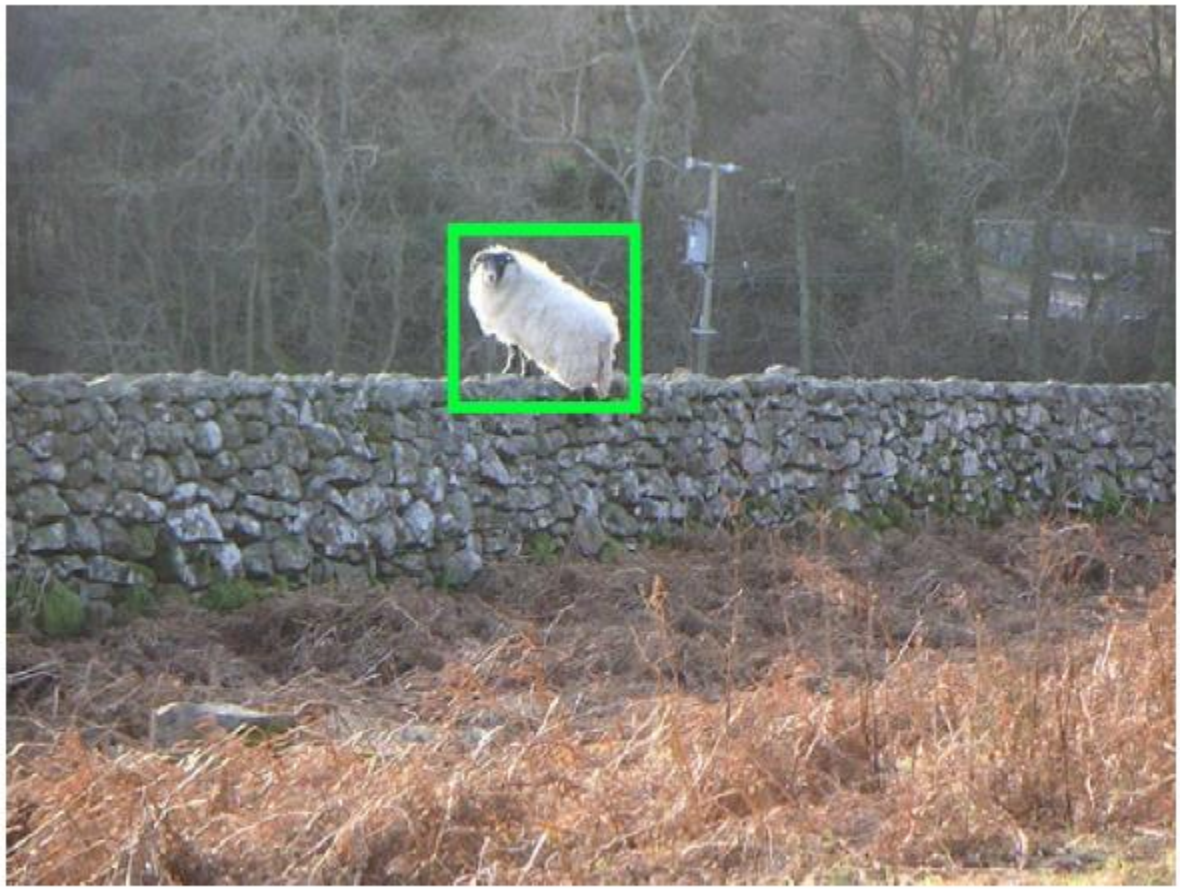
Sliding Window Classifiers

0.1

Sliding Window Classifiers

-0.2

Sliding Window Classifiers



. . .
1.5
. . .

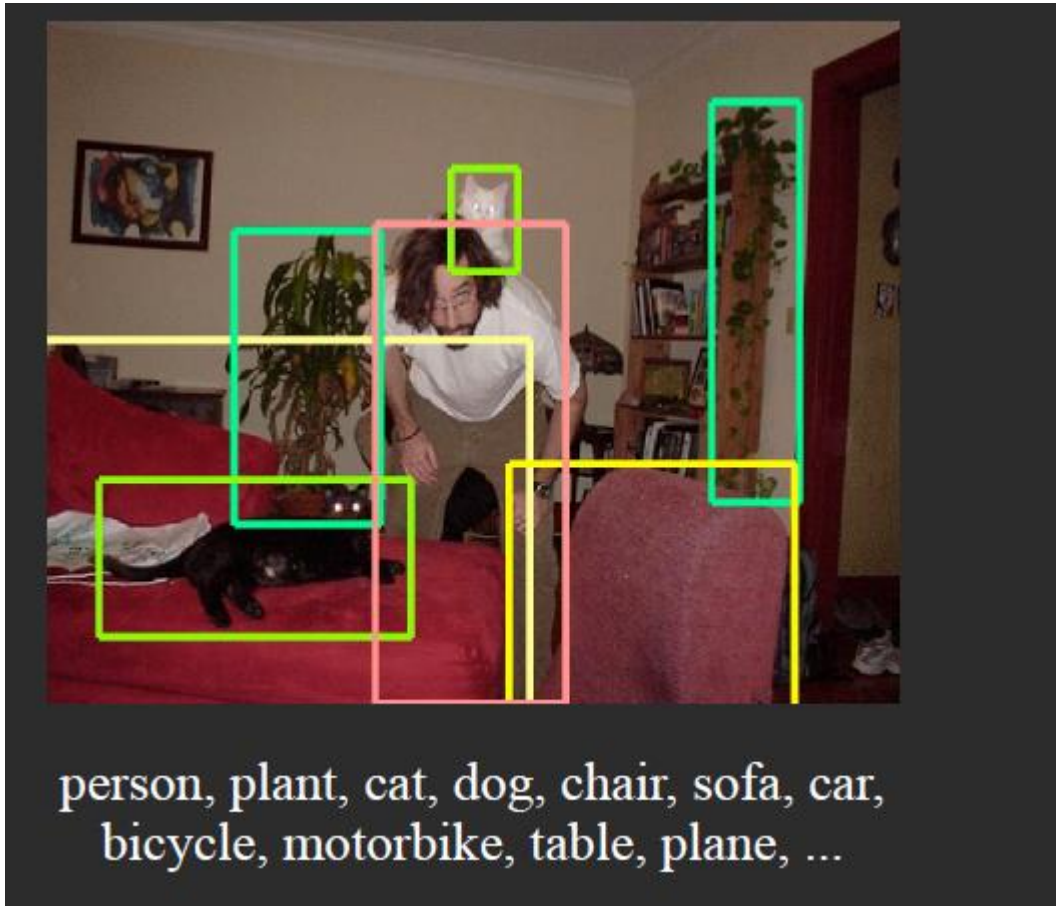
Sliding Window Classifiers

0.3

Sliding Window Classifiers

0.1
-0.2
-0.1
0.1
. . .
1.5
. . .
0.5
0.4
0.3

Goal: Detect Objects in Cluttered Images



Why is finding objects difficult?

variation in illumination

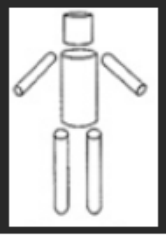
variation in appearance

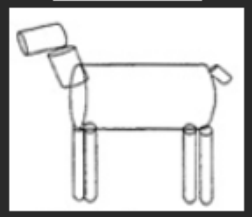
variation in pose, viewpoint

occlusion & clutter

Classic "nuisance factors" for general object recognition

Historical Approaches

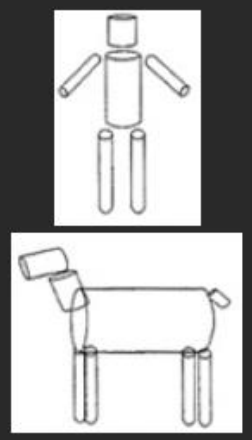




Geometric models
(1970s-1990s)

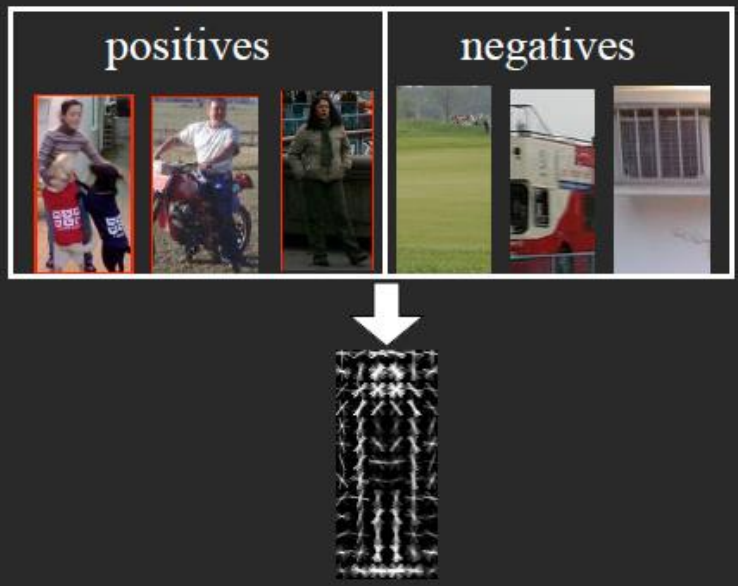
Hand-coded models

Historical Approaches



Geometric models
(1970s-1990s)

Hand-coded models



Statistical classifiers
(2000s-present)

Large-scale training
Appearance-based representations

A mix of both

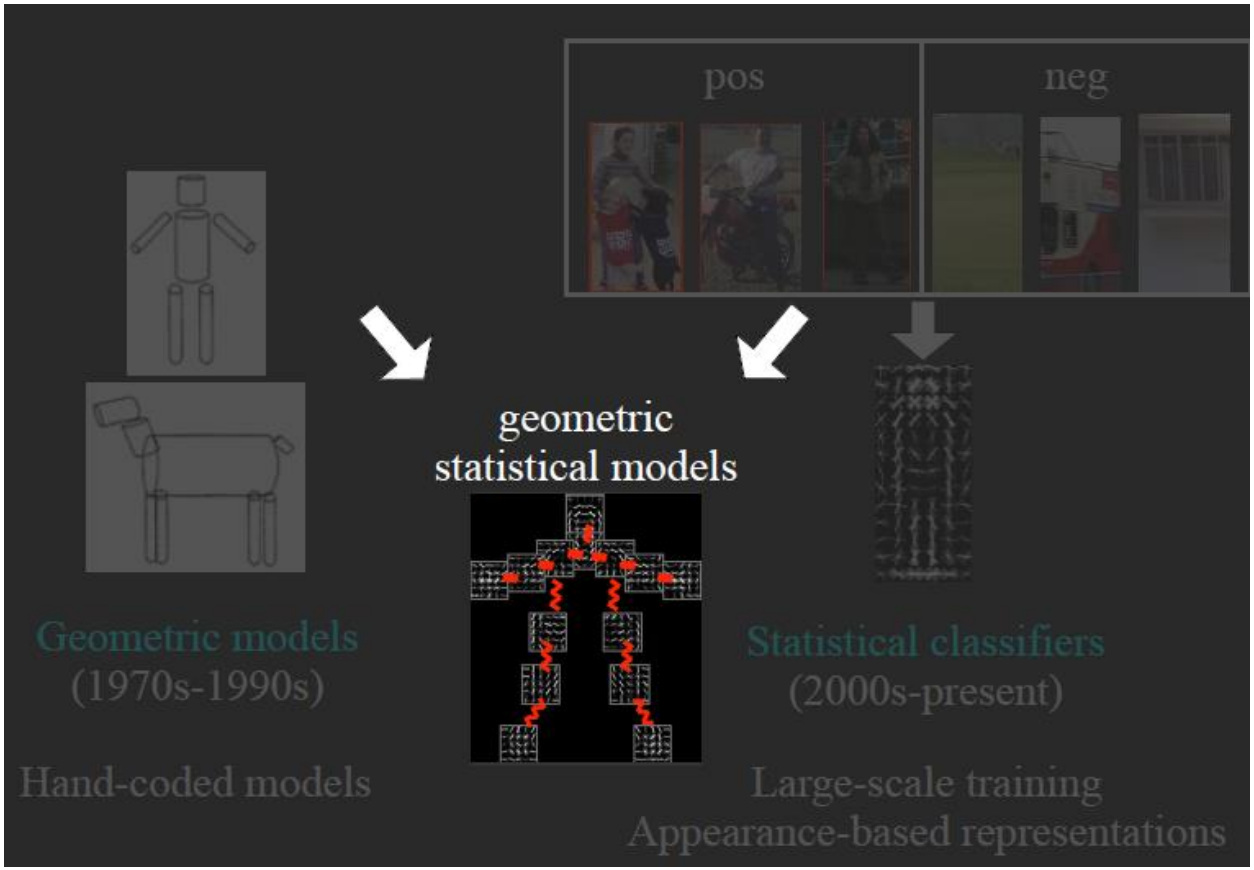
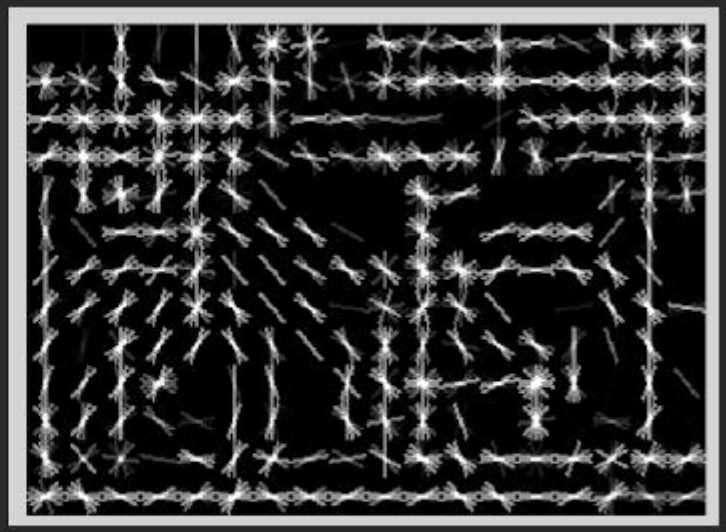

pos
neg
geometric
statistical models
Geometric models
(1970s-1990s)
Hand-coded models
Statistical classifiers
(2000s-present)
Large-scale training
Appearance-based representations

Image Features

Histograms of oriented gradients (HOG)

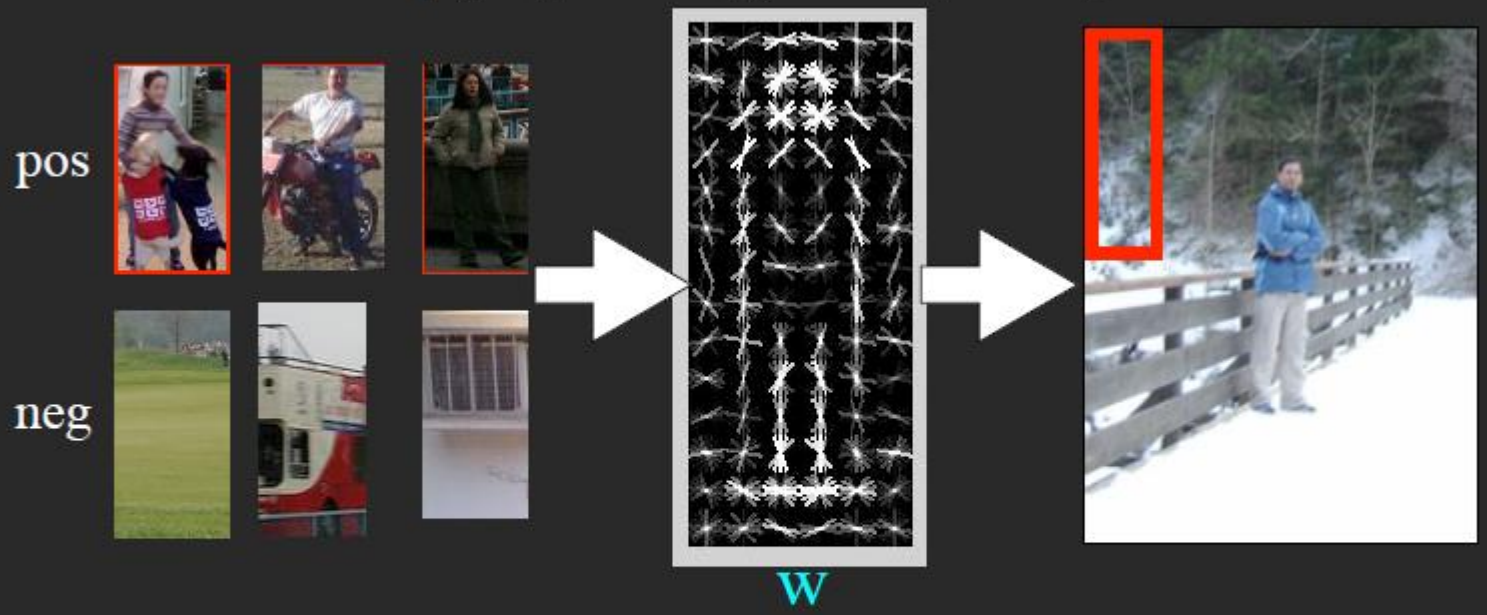


Bin gradients from 8x8 pixel neighborhoods into 9 orientations

(Dalal & Triggs *CVPR* 05)

Scanning Window Templates

Dalal and Triggs CVPR05 (HOG)
Papageorgiou and Poggio ICIP99 (wavelets)



w = weights for orientation and spatial bins 

$w \cdot x > 0$

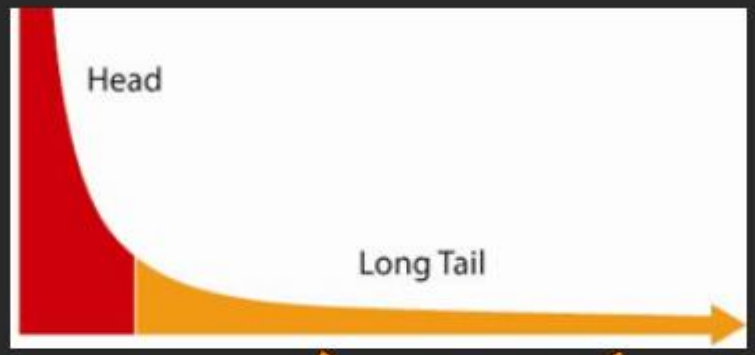
Train with a linear classifier (perceptron, logistic regression, SVMs...)

Mixtures of Templates

Train "sub-category" templates for each type of pose, body-shape, etc.

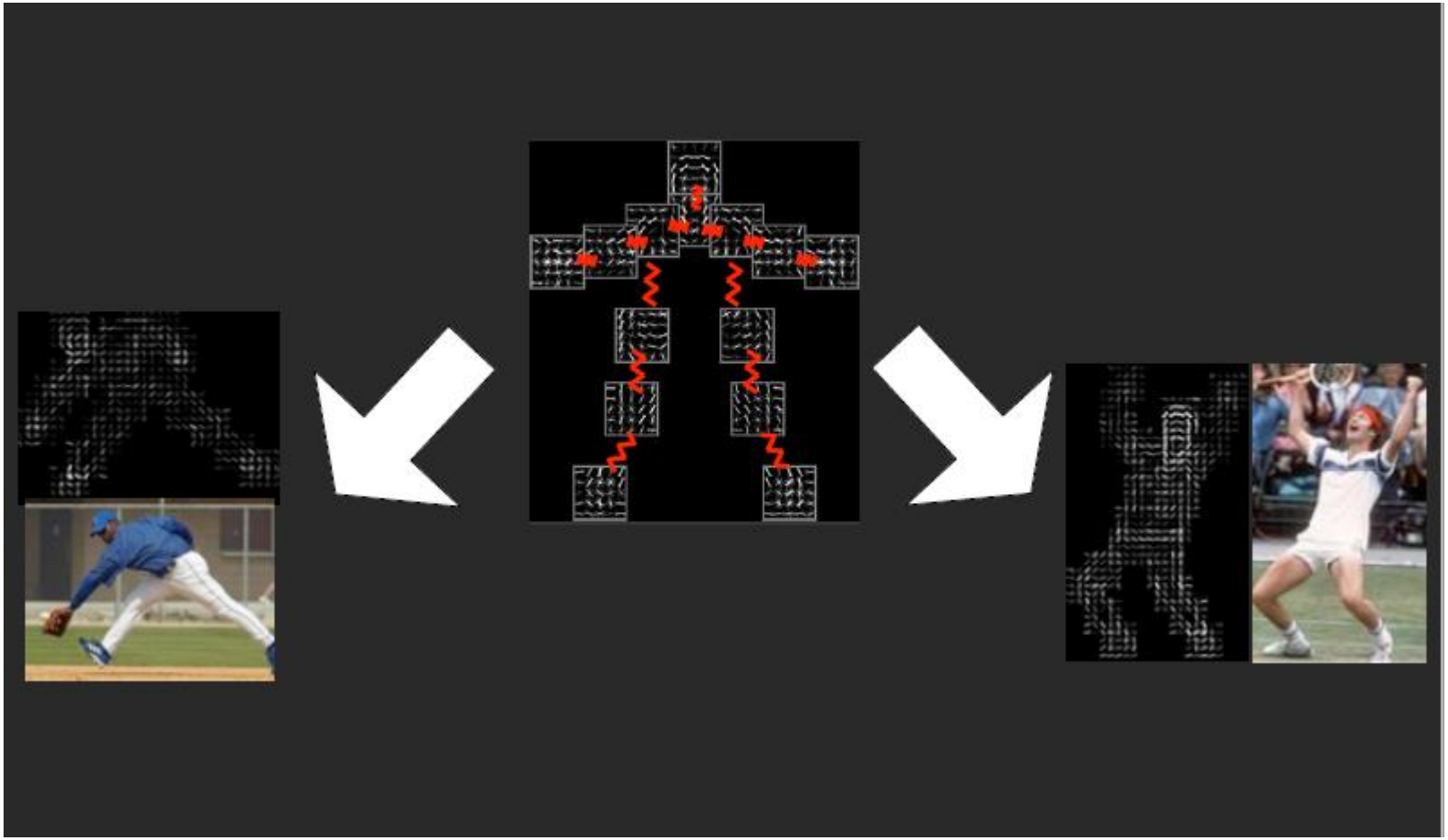
But how to handle…

Long-tail distribution of poses

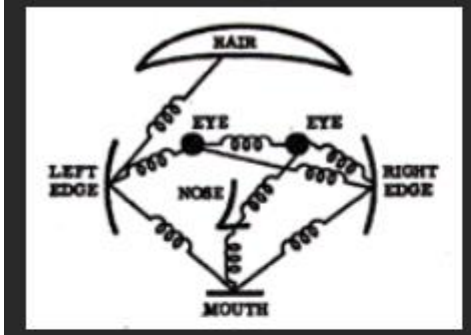


We need lots of templates, and will likely have little data of 'yoga twist' poses

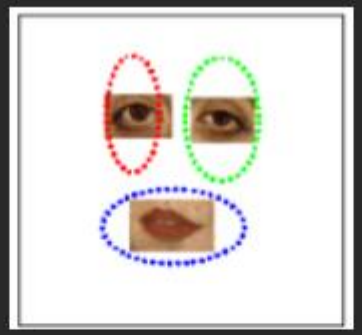
Deformable Part Models (DPMs)



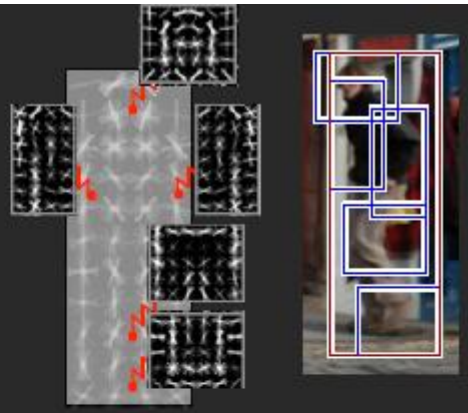
Deformable Part Models (DPMs)



Pictorial structures



Constellation models



Deformable part models

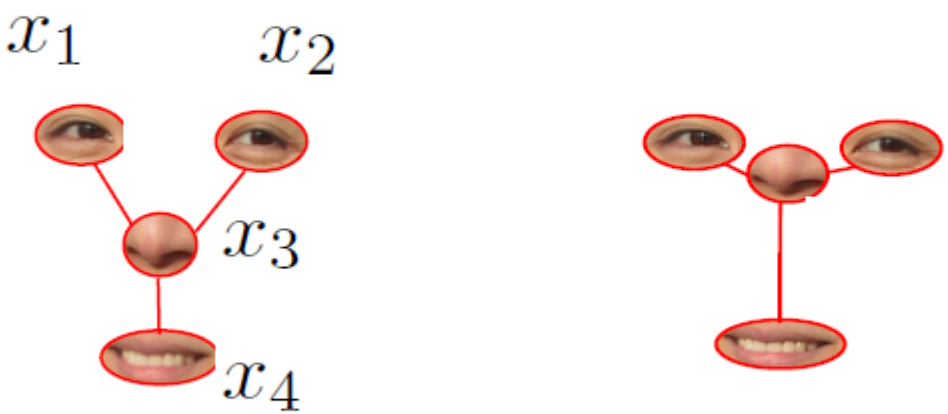
Model encodes local appearance + pairwise geometry

Pictorial Structures (Fischler & Elschlager 73, Felzenswalb and Huttenlocher 00)
Cardboard People (Yu et al 96)
Body Plans (Forsyth & Fleck 97)
Active Appearance Models (Cootes & Taylor 98)
Constellation Models (Burl et all 98, Fergus et al 03)

Deformable Part Models (DPMs)

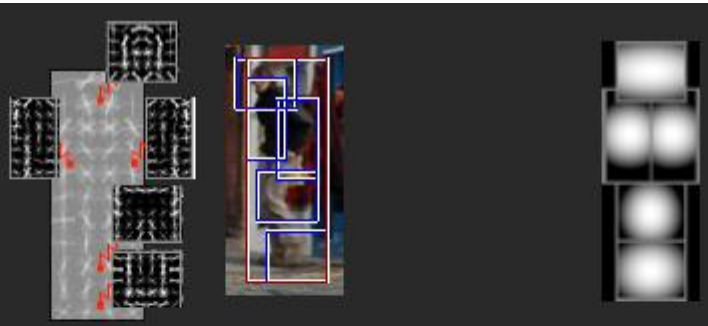
$$E(x) = \sum_i m_i(x_i) + \sum_{i,j \in E} \phi_{i,j}(x_i, x_j)$$

Local appearance **Pairwise compatibility**



P. Felzenszwalb, and D. Huttenlocher, Pictorial Structures for Object Recognition, IJCV 2005
P. Felzenszwalb, et. Al., Object Detection with Discriminatively Trained Part-Based Models, 2010

Scoring function



$$S(x,z) = \sum_i w_i \cdot \phi(x, z_i) + \sum_{i,j \in E} w_{ij} \cdot \psi(z_i, z_j)$$

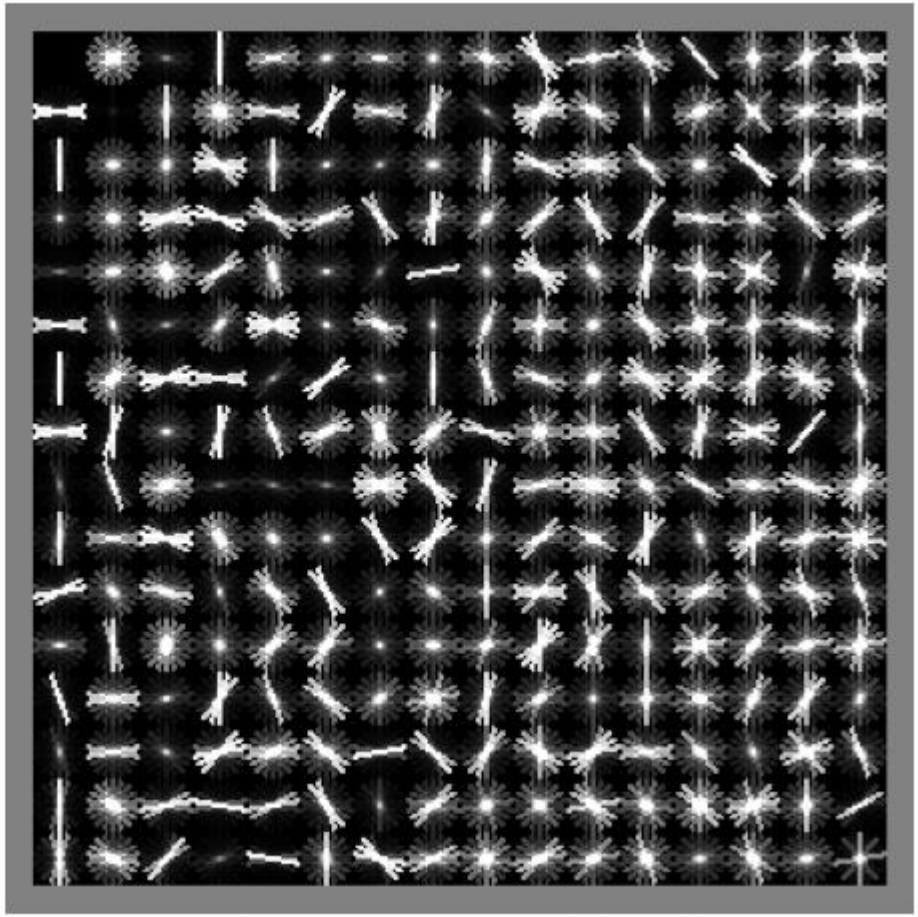
x = image
$z_i = (x_i, y_i)$
$z = \{z_1, z_2 \ldots\}$

part template scores

spring deformation model

HOG Quantization : Visual 'letters'

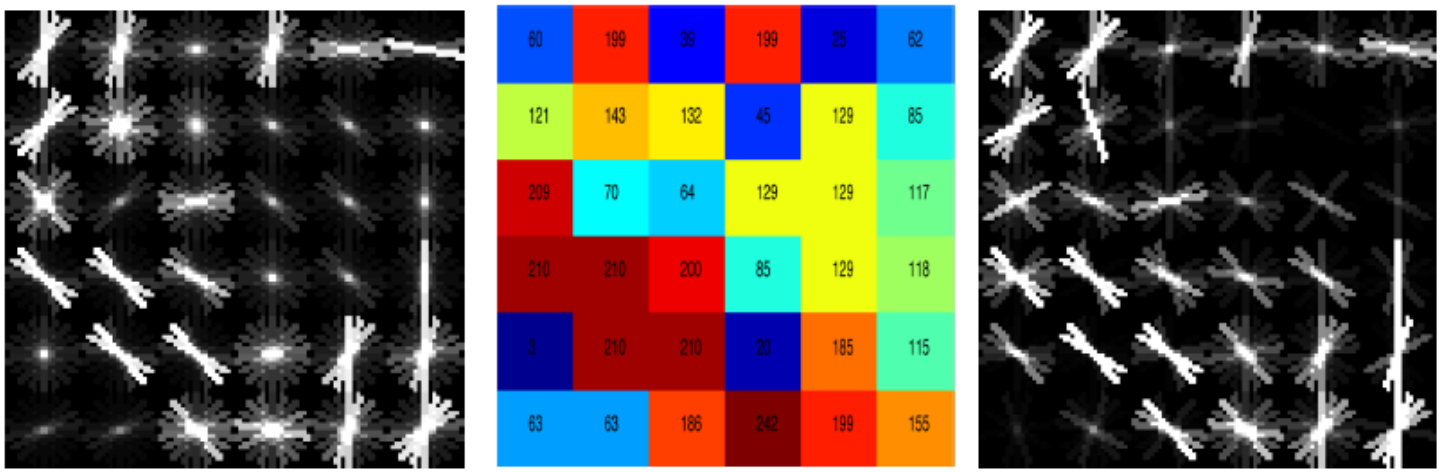
$$\mathcal{C} = \{C_1, \ldots, C_{256}\}$$



I. K., Bounding Part Scores for Rapid Object Detection with DPMs, PnA, 2012

HOG feature Quantization

HOG detail **Codebook indices** **Quantized HOG**



$$\mathbf{h}[x] \qquad i[x] = \arg\min_k d(\mathbf{h}[x], C_k) \qquad \hat{\mathbf{h}}[x] = C_{i[x]}$$

I. K., Bounding Part Scores for Rapid Object Detection with DPMs, PnA, 2012

DPM Demo


input image
0
50
100
150
200
250
300
350
400
450
100
200
300
400
500
600
7
NEWHAVEN

Object Detection – Future Direction

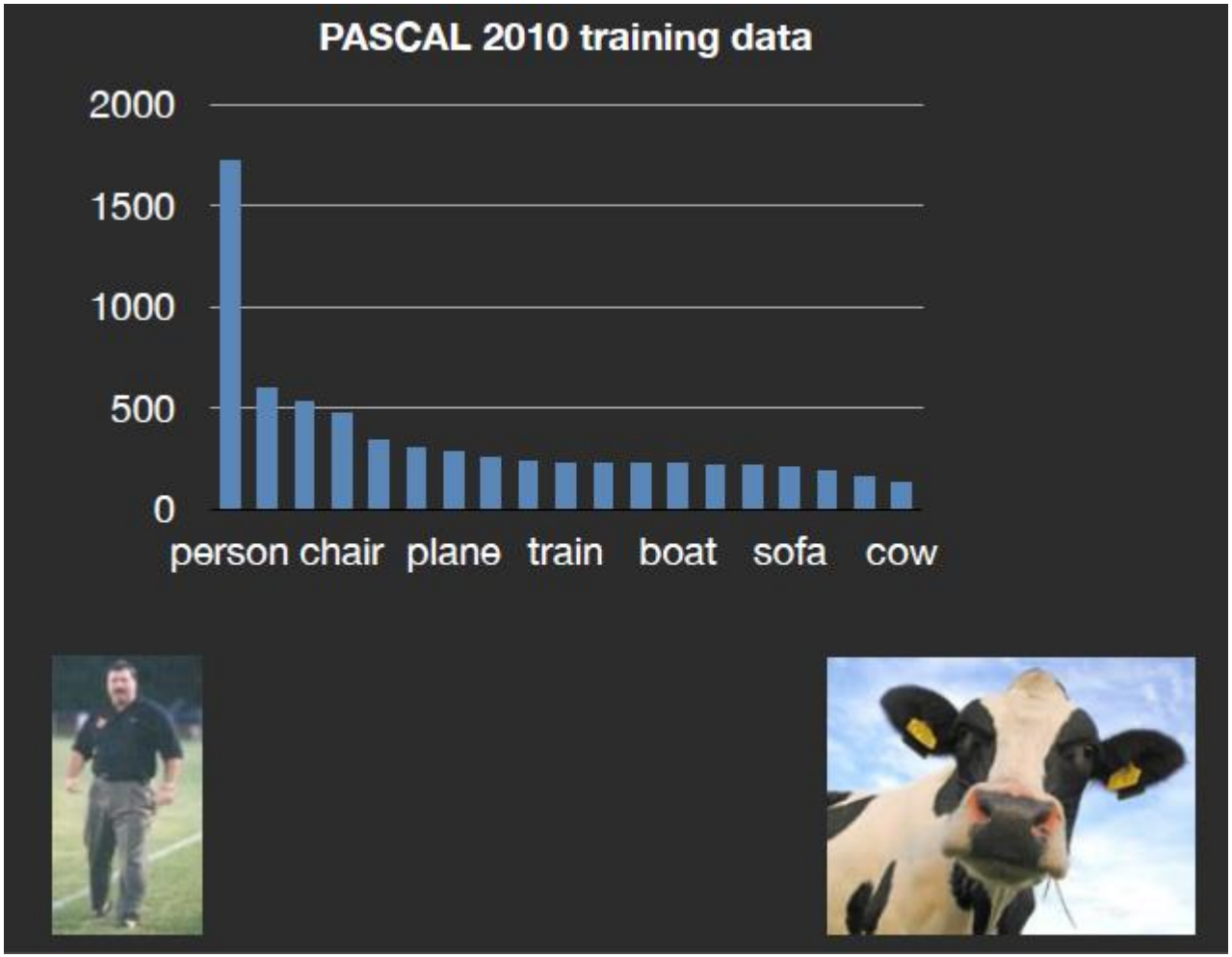
Functional prediction

If you know what can be done with a ... object, what it can be used for, you can call it whatever you please"

J. J. Gibson. *The Ecological Approach to Visual Perception*

"sittable" affordance label implies someone can sit in the future

Challenges – Long Tails



Subordinate Categories

How to sub-linearly encode fine-scale differences between object categories?

Object Representations

Patches
2011

Skeleton
1970's

Poselets
2009

As a field, we perform a human-in-the-loop search over representations, at the time-scale of years or decades

We must be able to do better!

Slides Credit

Radhakrishna Dasari