Penn Engineering

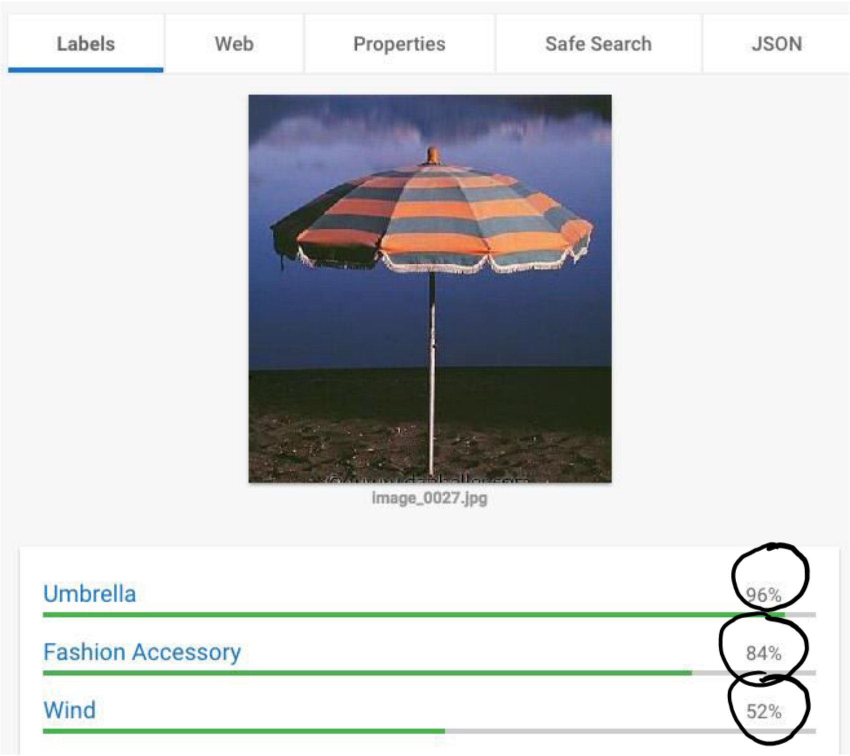
ONLINE LEARNING

Video 11.1
Kostas Daniilidis

By object recognition we might mean…

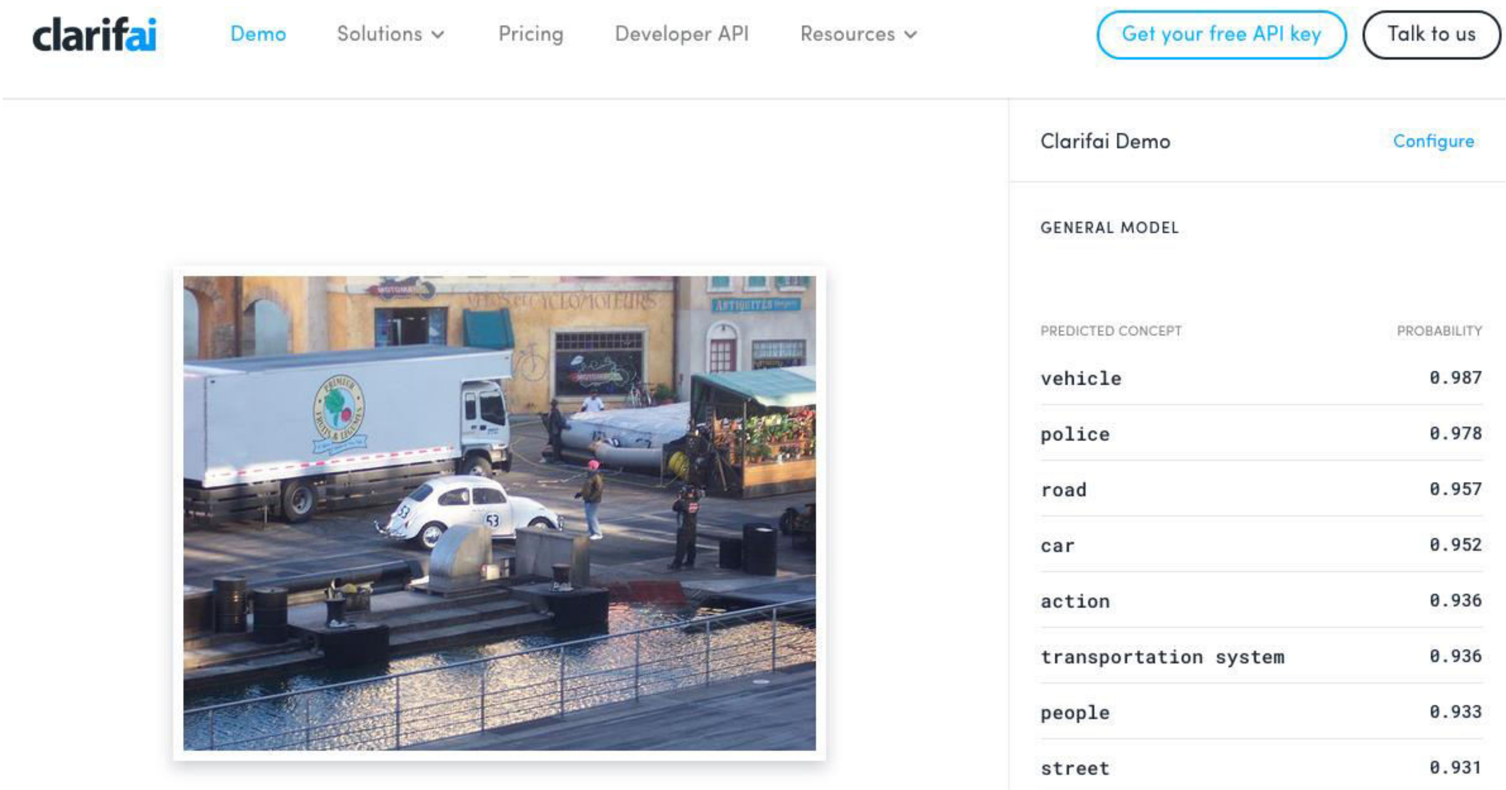
Image classification:

Input is an image and output is a set of class labels with probabilities



`https://cloud.google.com/vision/docs/drag-and-drop`

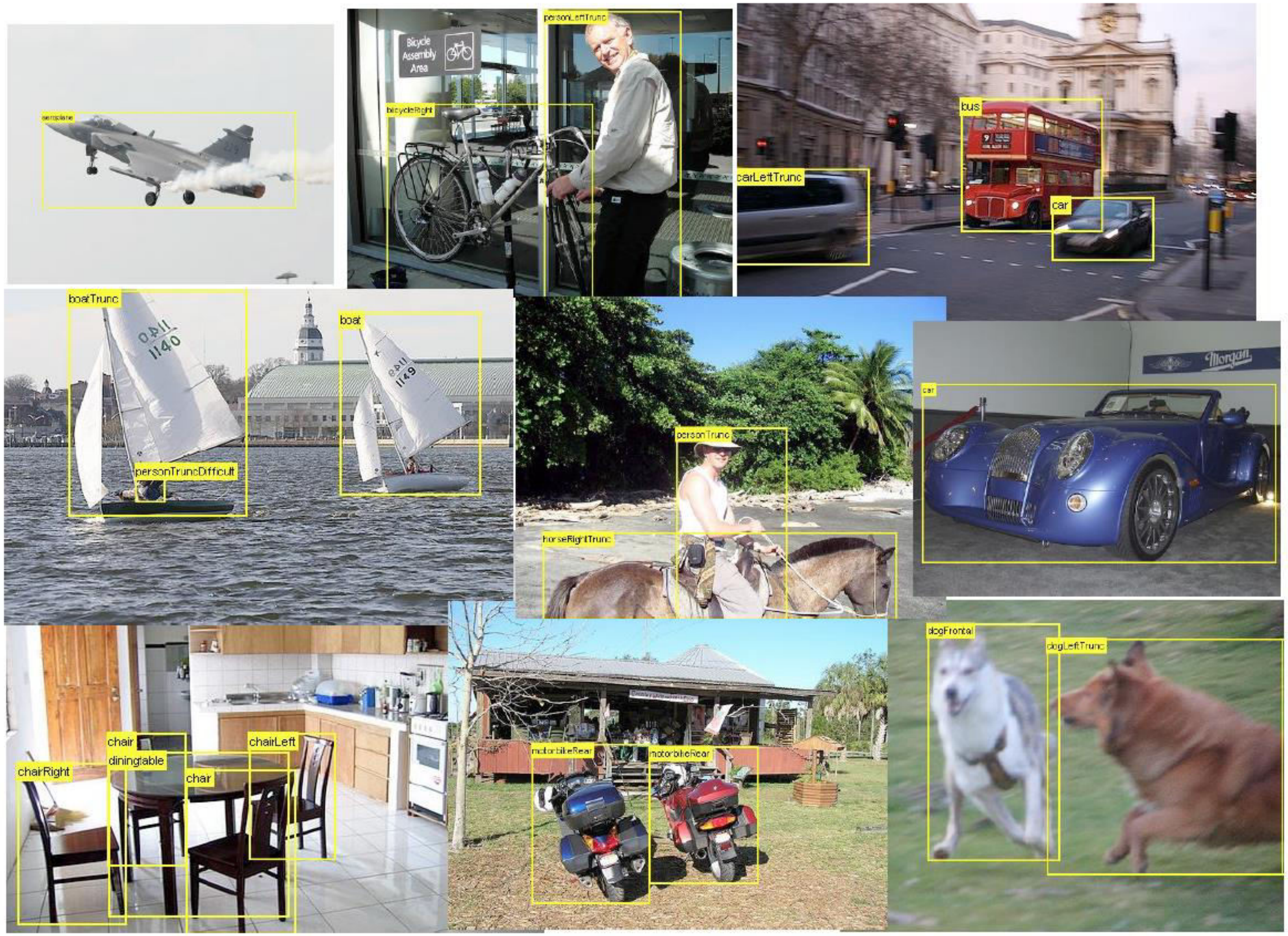
.. Or here:


clarifai
Demo
Solutions
Pricing
Developer API
Resources
Get your free API key
Talk to us
Clarifai Demo
Configure
GENERAL MODEL
PREDICTED CONCEPT
PROBABILITY
vehicle 0.987
police 0.978
road 0.957
car 0.952
action 0.936
transportation system 0.936
people 0.933
street 0.931

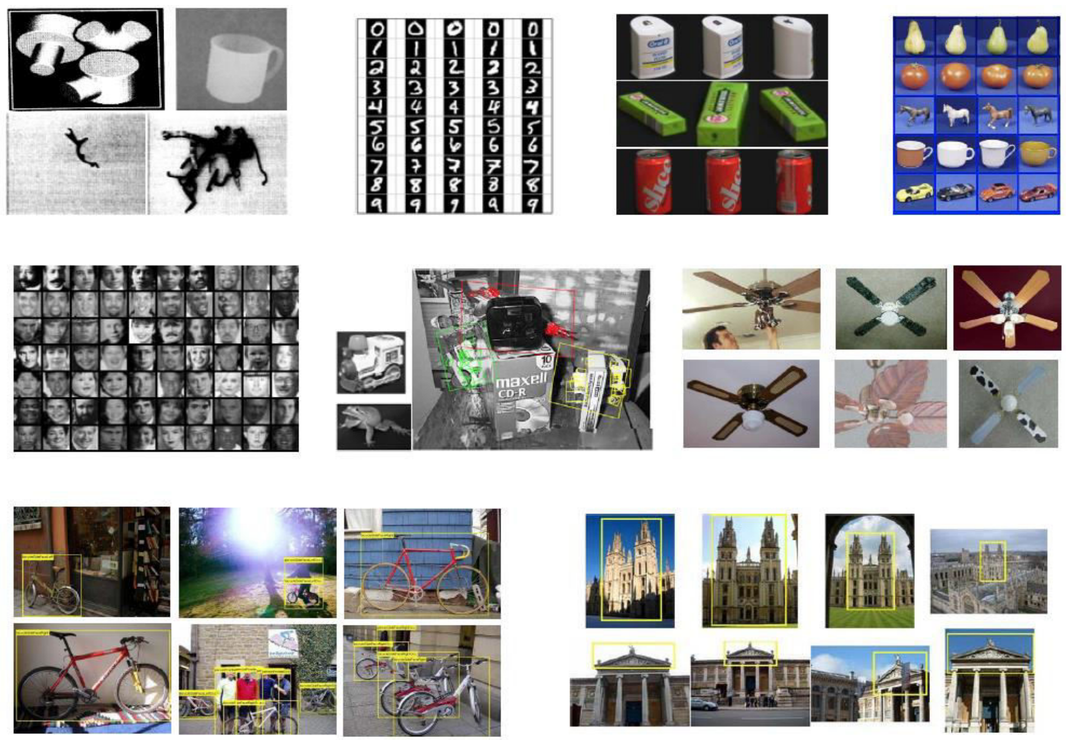
But object recognition usually means…

Object Detection and Localization:

Find where is a car in the image?


aeroplane
Bicycle Assembly Area
bicycleRight
personLeftTrunc
bus
carLeftTrunc
car
boatTrunc
boat
personTruncDifficult
personTrunc
horseRightTrunc
car
chair
chairLeft
diningtable
chairRight
char
motorbikeRear
motorbikeRear
dogFrontal
dogLeftTrunc

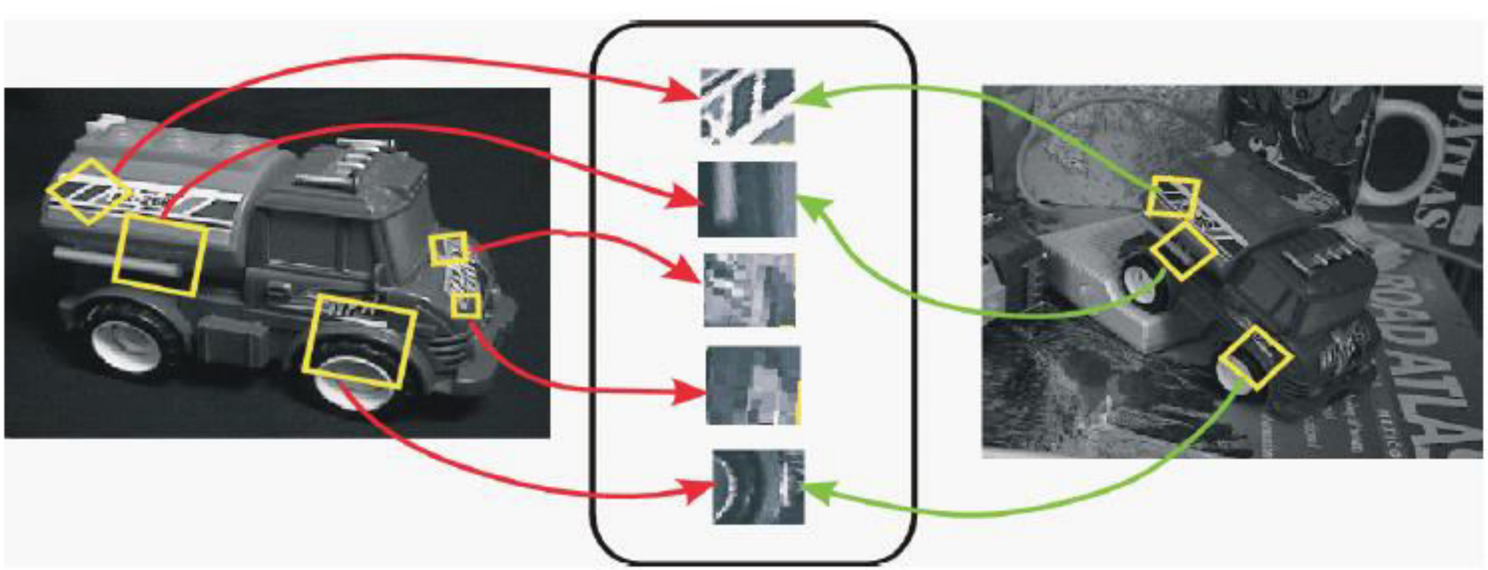
History (Grauman and Leibe, 2011)



Recognition ingredients

- Object features (for example, HOG) that is resistent to geometric and photometric nuisances.
- Learning (training) the class model (SVM) to represent intra-class variations.
- Testing on an input image and output all bounding boxes for a specific class (Sliding window)

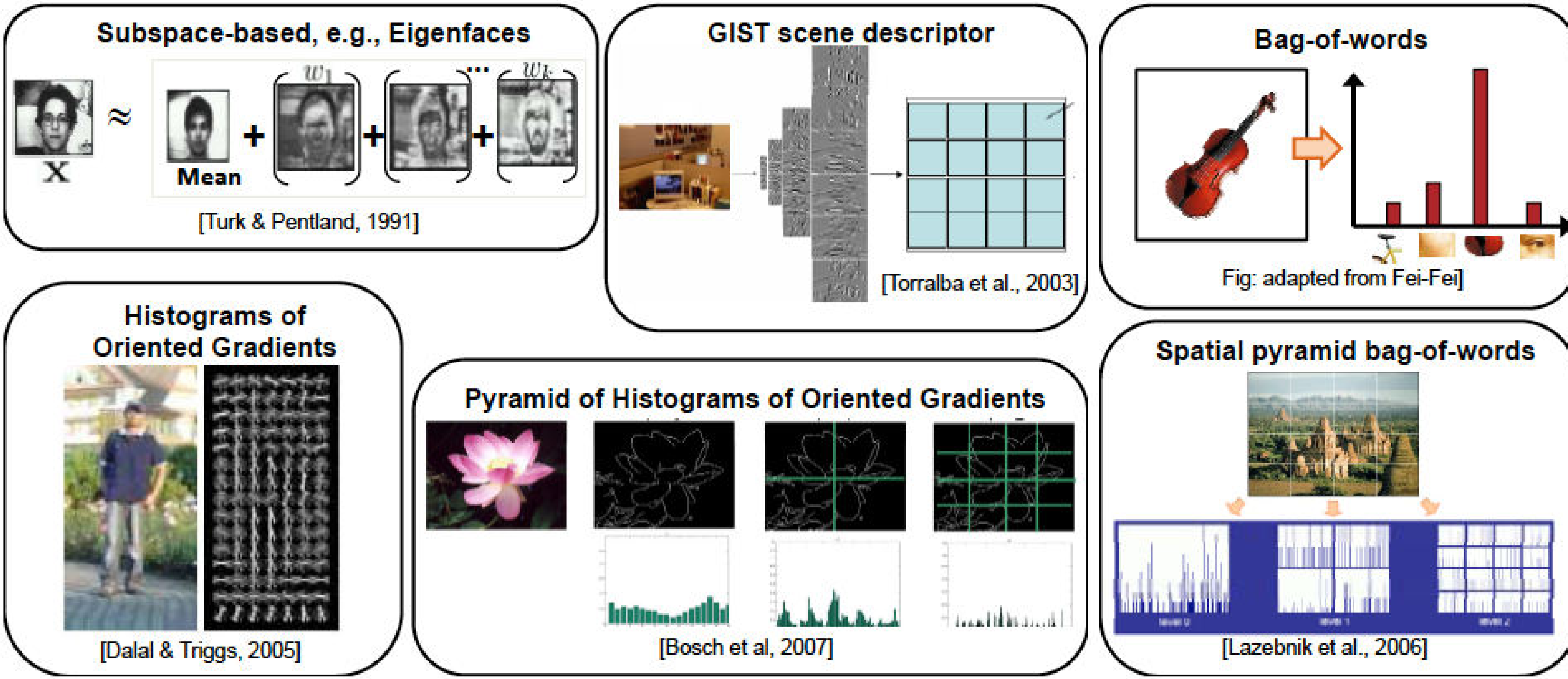
Instance recognition by matching (Lowe, 2004)



Match SIFT Features and verify geometric consistency

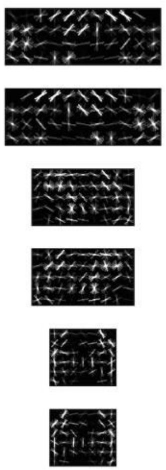
Video 11.2
Kostas Daniilidis

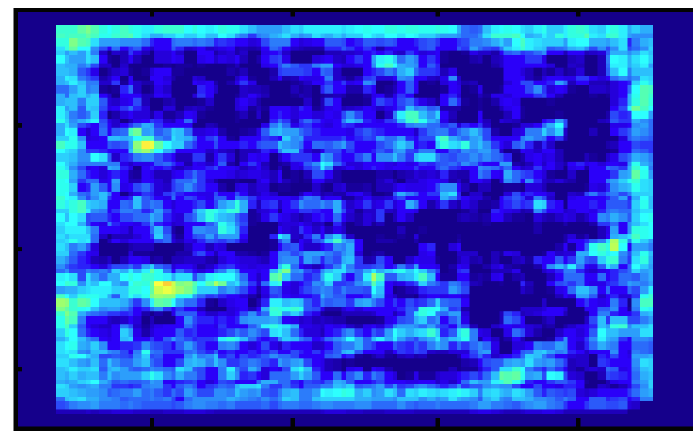
Holistic approaches (Grauman and Leibe, 2011)



Class Recognition: Sliding windows

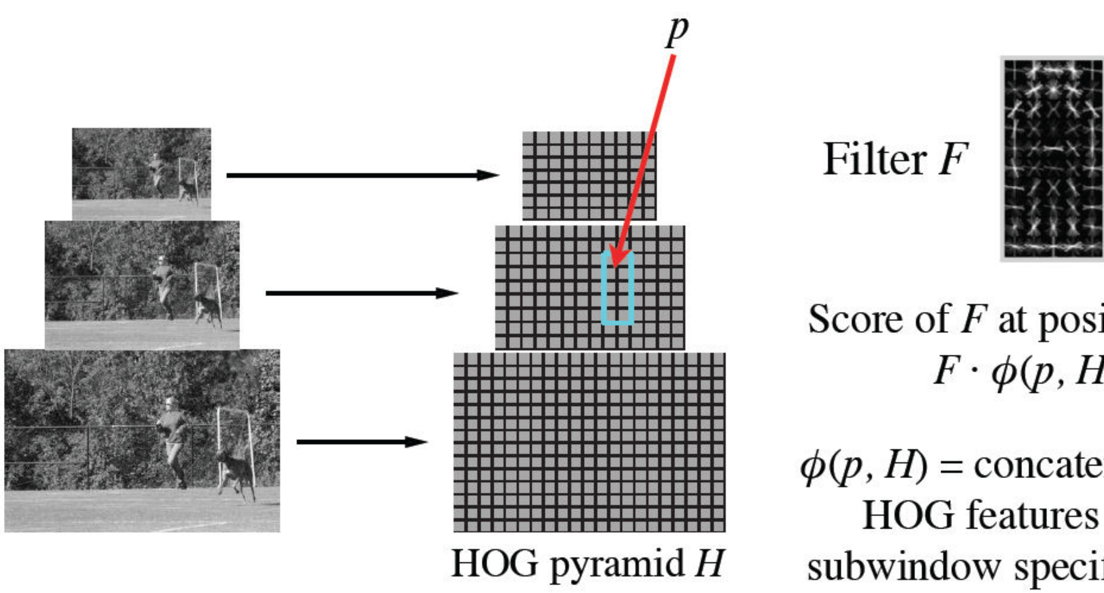
- Learn a HOG representative of a car over multiple scales
- What happens if we slide this window over an input image?





HOG filter

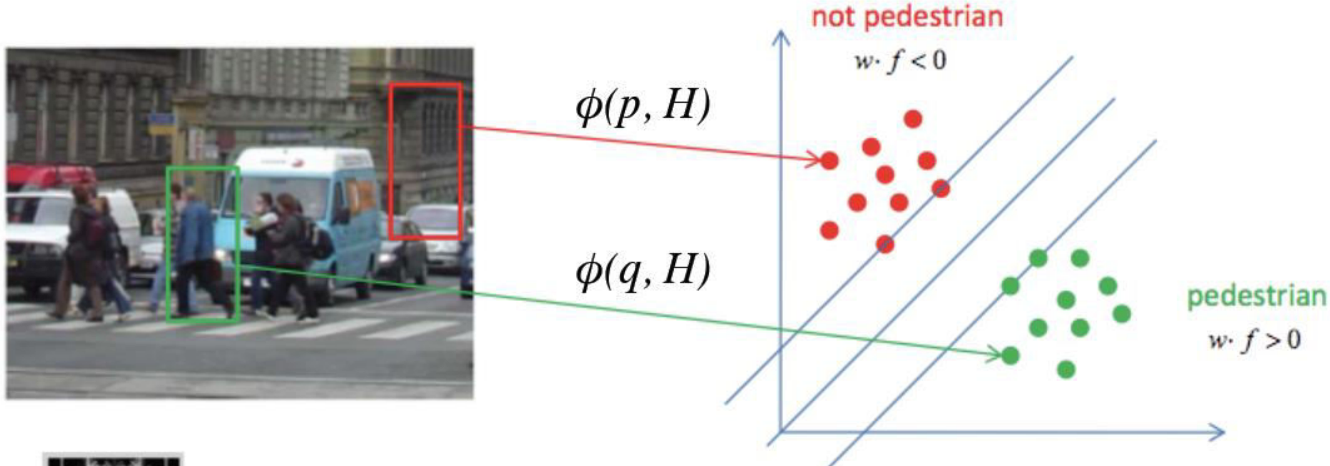
- Array of weights for features in subwindow of HOG pyramid
- Score is dot product of filter and feature vector



Score of F at position p is
$F \cdot \phi(p, H)$

$\phi(p, H)$ = concatenation of HOG features from subwindow specified by p

Dalal & Triggs: HOG + linear SVMs



Typical form of
a model

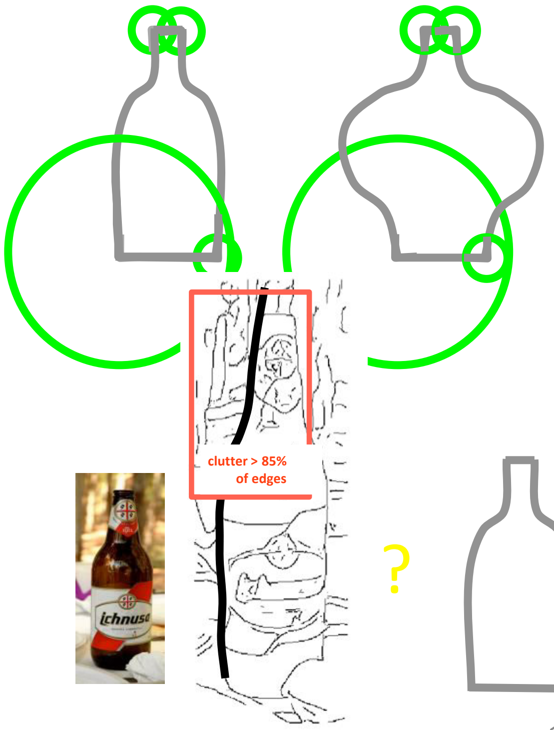
There is much more background than objects

Start with random negatives and repeat:

1) Train a model
2) Harvest false positives to define "hard negatives"

Holistic / Global Representation

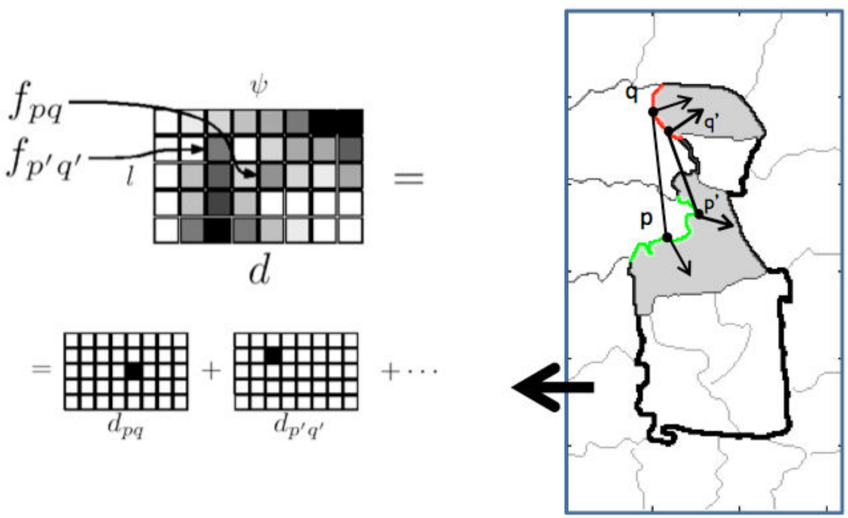
+ Gestaltism: shape is perceived as a whole:
  - not merely sum of parts
  - global dependencies



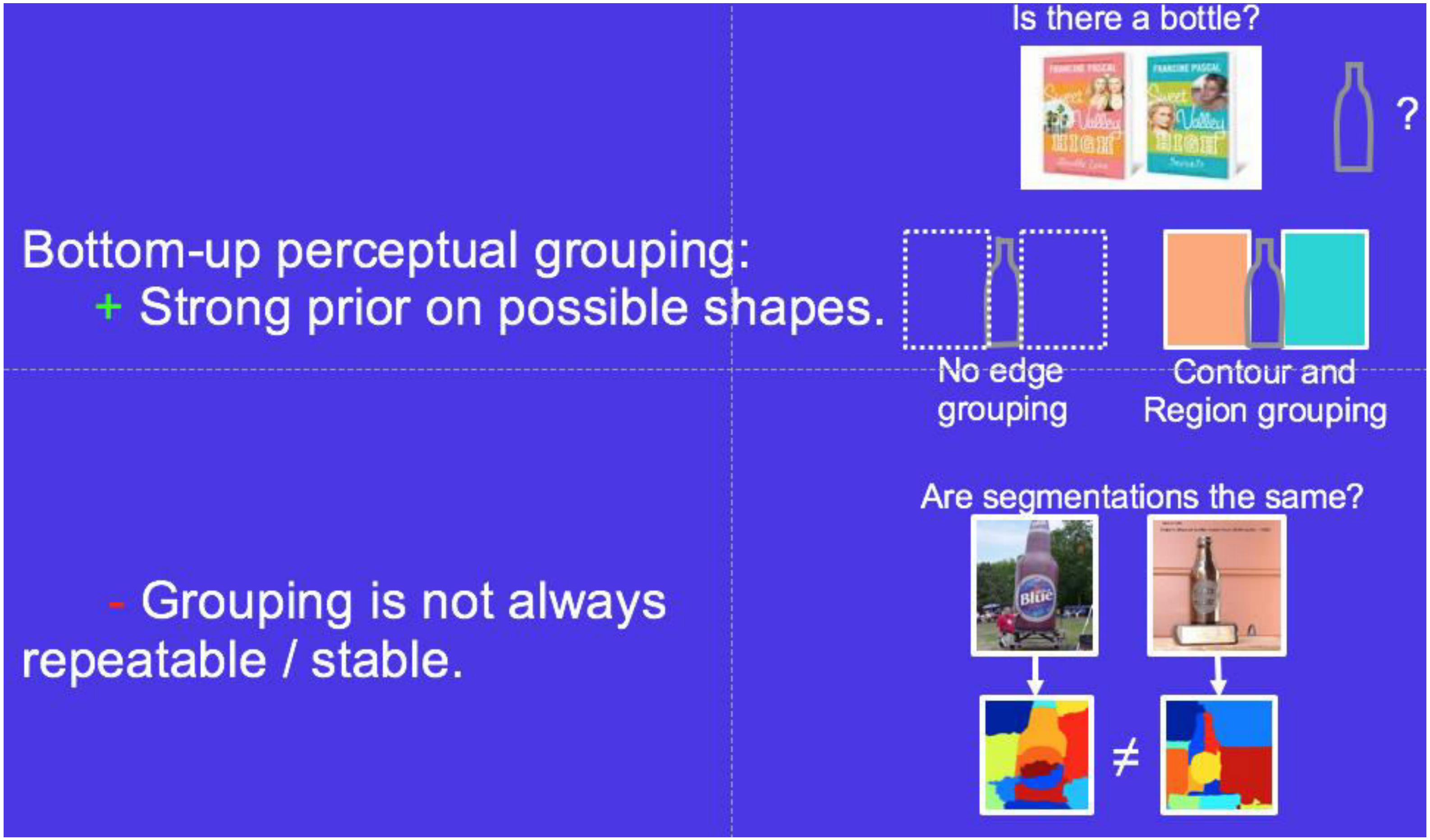
-

1
4

Localization through Shape Verification



Holistic approaches need perceptual grouping



Chordiogram

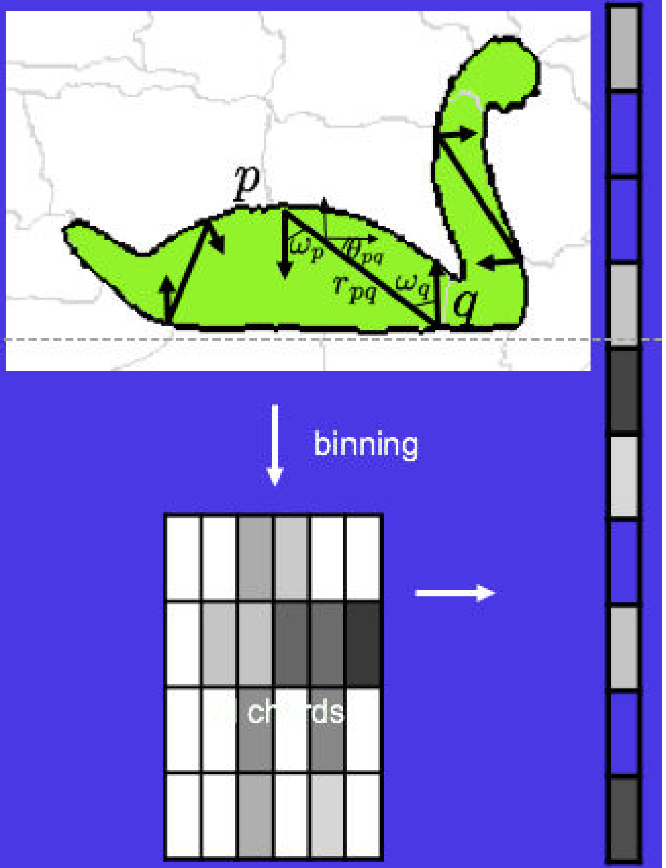
Idea: capture all global dependences among edges.

Chord: pair of boundary edges (p,q).

Chord feature: $f_{pq} = (r_{pq}, \theta_{pq}, \omega_p, \omega_q)$

Chordiogram: histogram of chord features:

$$\mathrm{ch}_k = \#\{(p,q) | f_{pq} \in bin(k)\}$$



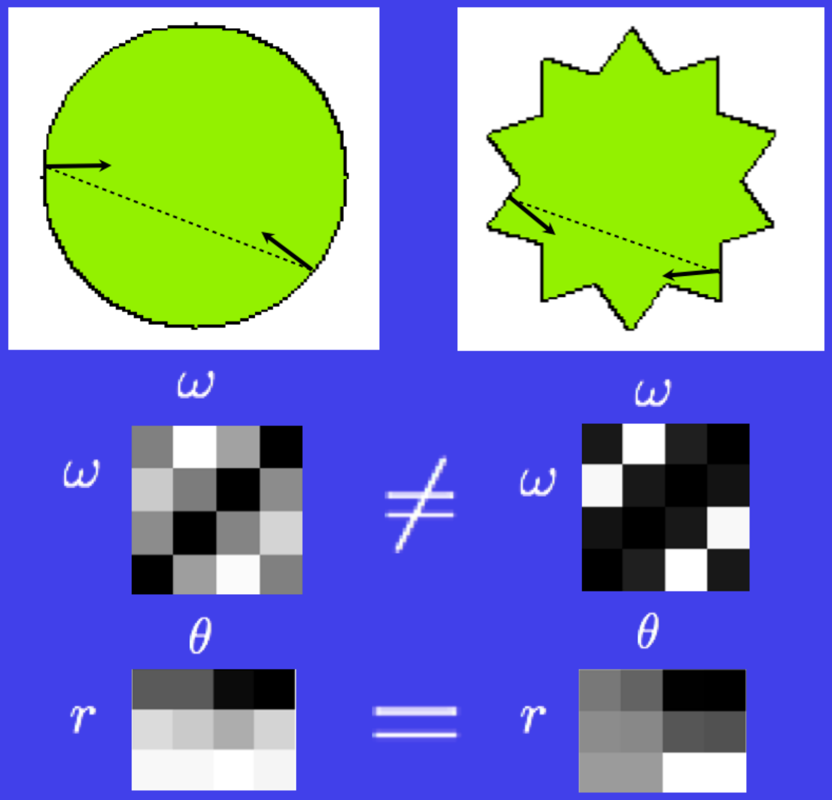
Chord Features

Chord normals capture local finer shape information:

Chordiogram over only two features:

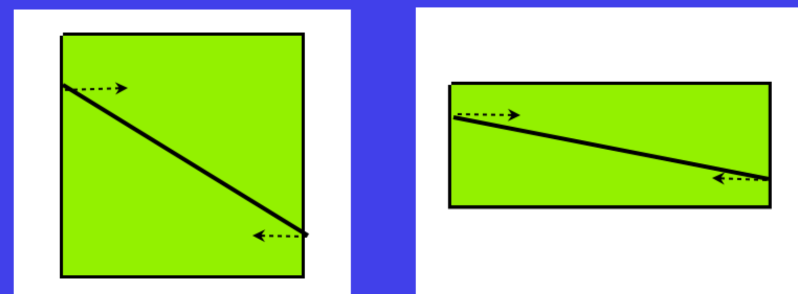
Chord normals ω

Chord length and orientation r, θ



Chord Features

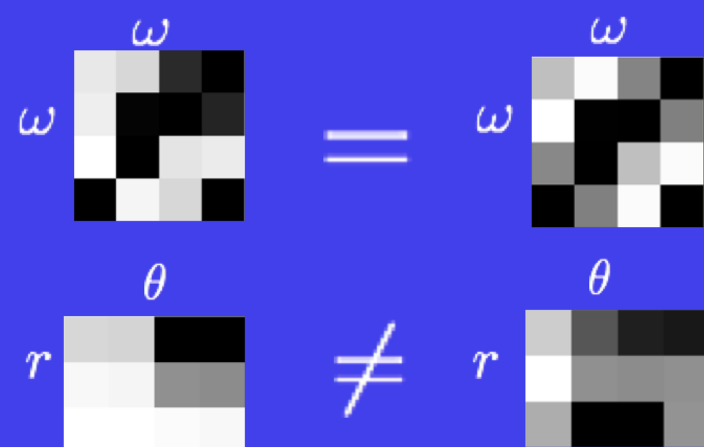
Chord length and orientation capture global coarse shape information:



Chordiogram over only two features:

Chord normals ω →

Chord length and orientation r, θ →

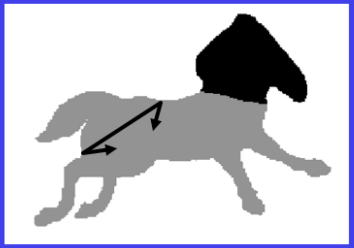


Gestalt Properties of the Chordiogram

Global:
short as well as long range chords.

ch(centaur torso)

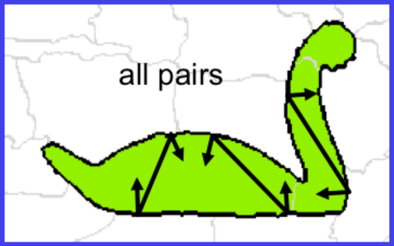
Holistic:
the chords of an edge are affected by all object parts.



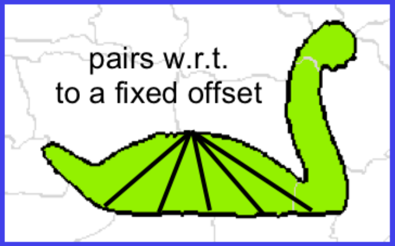
ch(horse torso)

Transformation Properties of the Chordiogram

Translation invariance:
a chord captures only relative location.

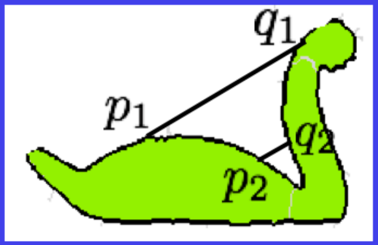


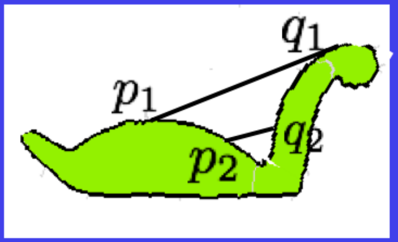
chordiogram



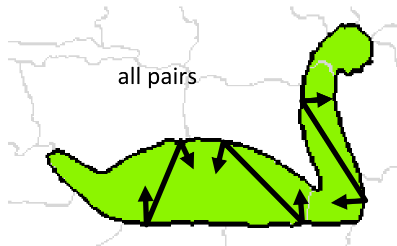
shape context

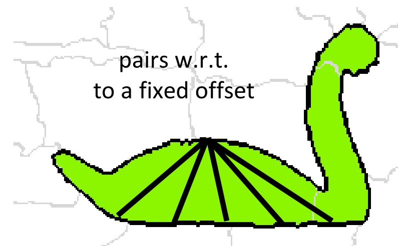
Robustness to shape variation:
lengths are quantized uniformly in log space.





Chordiogram vs Shape Context



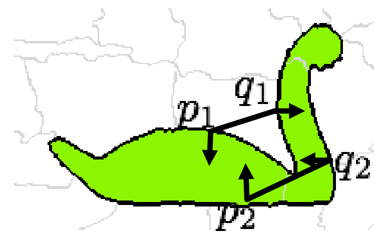


Translation invariance:	Yes	No
Global support:	Yes	No
Global configuration:	No	Yes
Notion of parts:	No	Yes

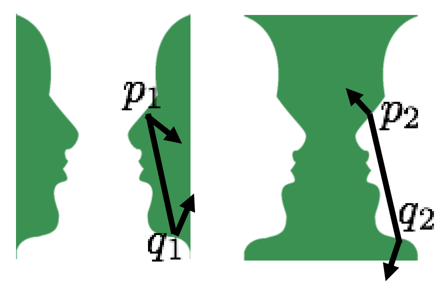
Figure/Ground Organization

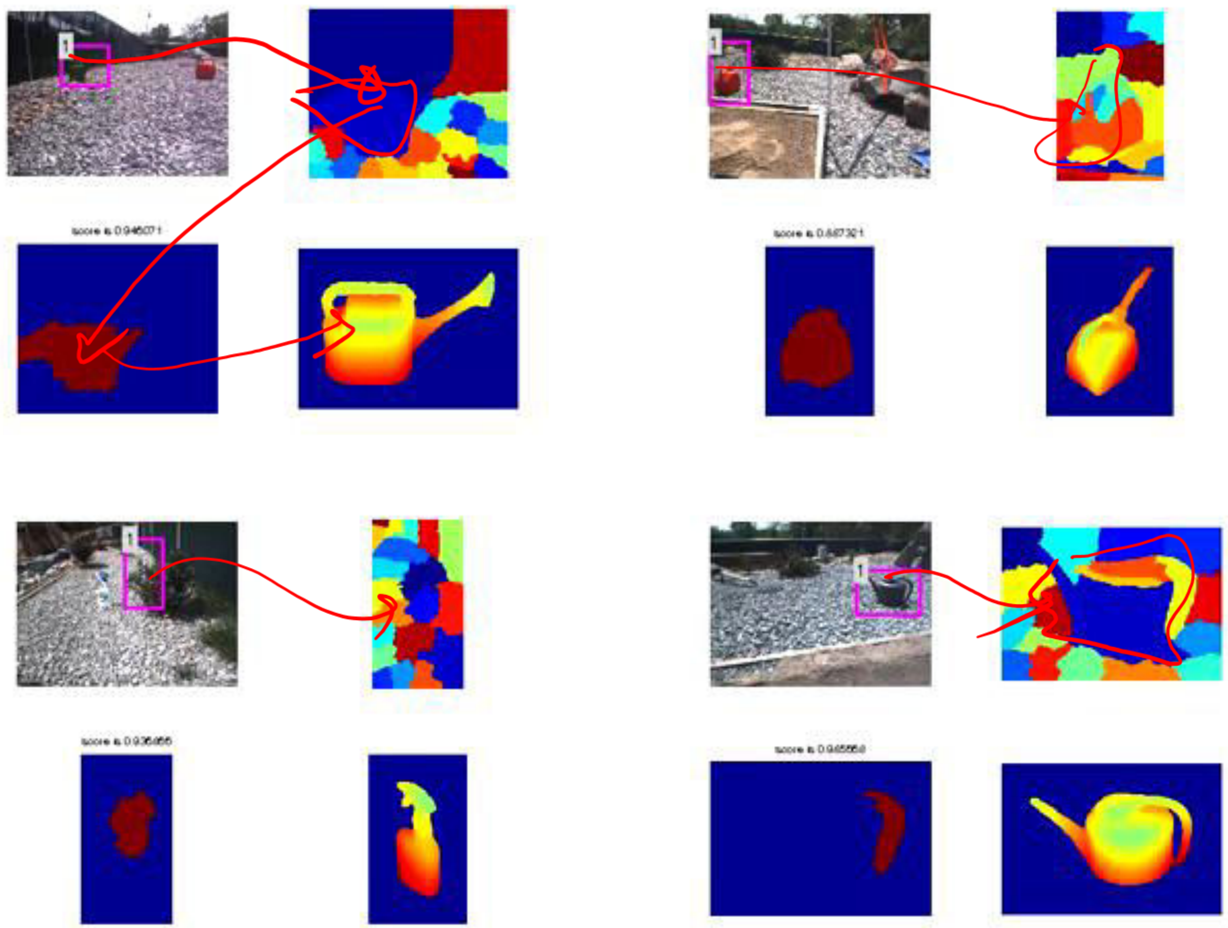
Chord normals point towards the object interior.

$\mathrm{chord}(p_1, q_1) \not\sim \mathrm{chord}(p_2, q_2)$



Contour and interior descriptor:
More discriminative than pure contour descriptors.
Encode figure/ground organization.



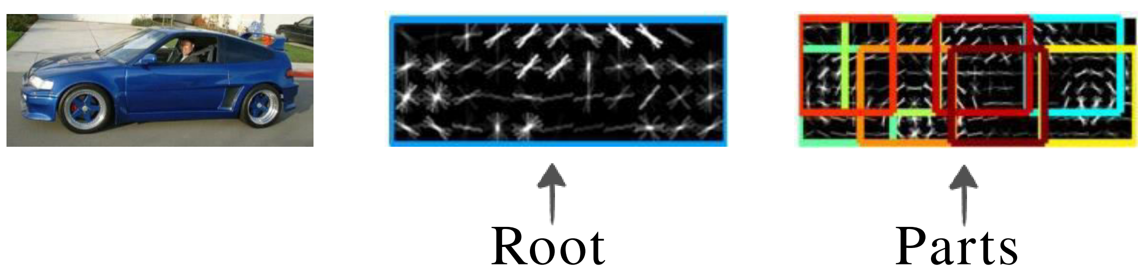


Video 11.3
Kostas Daniilidis

Deformable Part Models

- DPM consists a Root and several Parts
  - Root represents holistic object shape
  - Part captures detailed part appearance

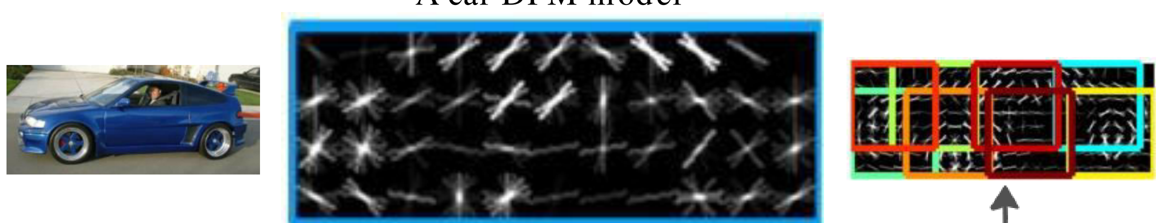
A car DPM model



Deformable Part Models

- DPM consists a Root and several Parts
  - Root represents holistic object shape
  - Part captures detailed part appearance

A car DPM model

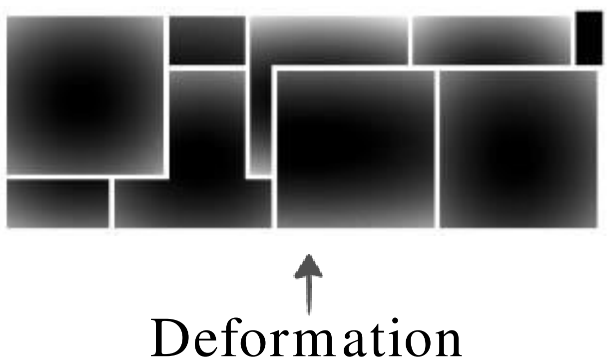


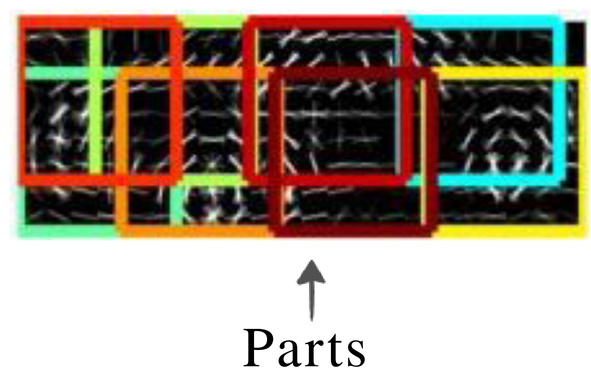
Root

Parts

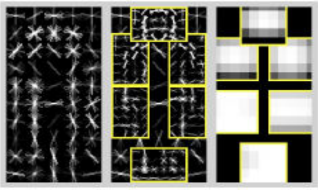
Deformable Part Models

- Flexible part configuration
  - Parts are allowed to move around the anchor points (default position)





Object hypothesis



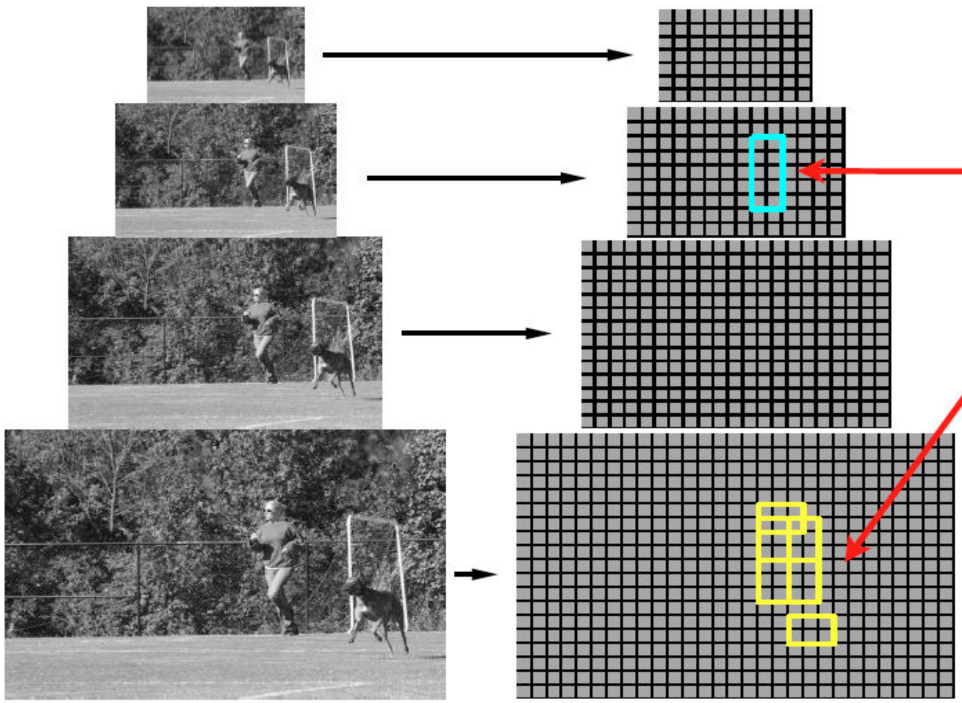


Image pyramid

HOG feature pyramid

$z = (p_0, \ldots, p_n)$

p_0 : location of root

$p_1, \ldots, p_n$: location of parts

Score is sum of filter scores minus deformation costs

Multiscale model captures features at two-resolutions

Deformable Part Models

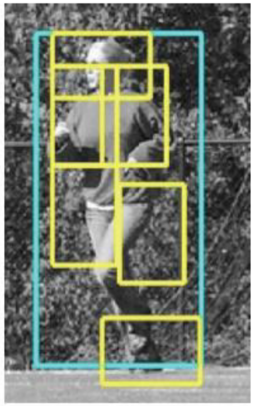
- Detector response at a given root location

$$x = (r, c, l)$$

$$\underbrace{F_0' \cdot \phi(H, x)}_{\text{Root}} + \sum_{k=1}^{n} \max_{x_k} \Big(\underbrace{F_k' \cdot \phi(H, x_k) - d_k \cdot \phi_d(\delta_k)}_{\text{Parts}} \Big) + \underbrace{b}_{\text{bias}}$$

Score of a hypothesis

$$\text{score}(p_0, \ldots, p_n) = \boxed{\underbrace{\sum_{i=0}^{n} \underbrace{F_i}_{\text{filters}} \cdot \phi(H, p_i)}_{\text{"data term"}}} - \boxed{\underbrace{\sum_{i=1}^{n} \underbrace{d_i}_{\text{deformation parameters}} \cdot \underbrace{(dx_i^2, dy_i^2)}_{\text{displacements}}}_{\text{"spatial prior"}}}$$



$$\text{score}(z) = \beta \cdot \Psi(H, z)$$

concatenation filters and deformation parameters

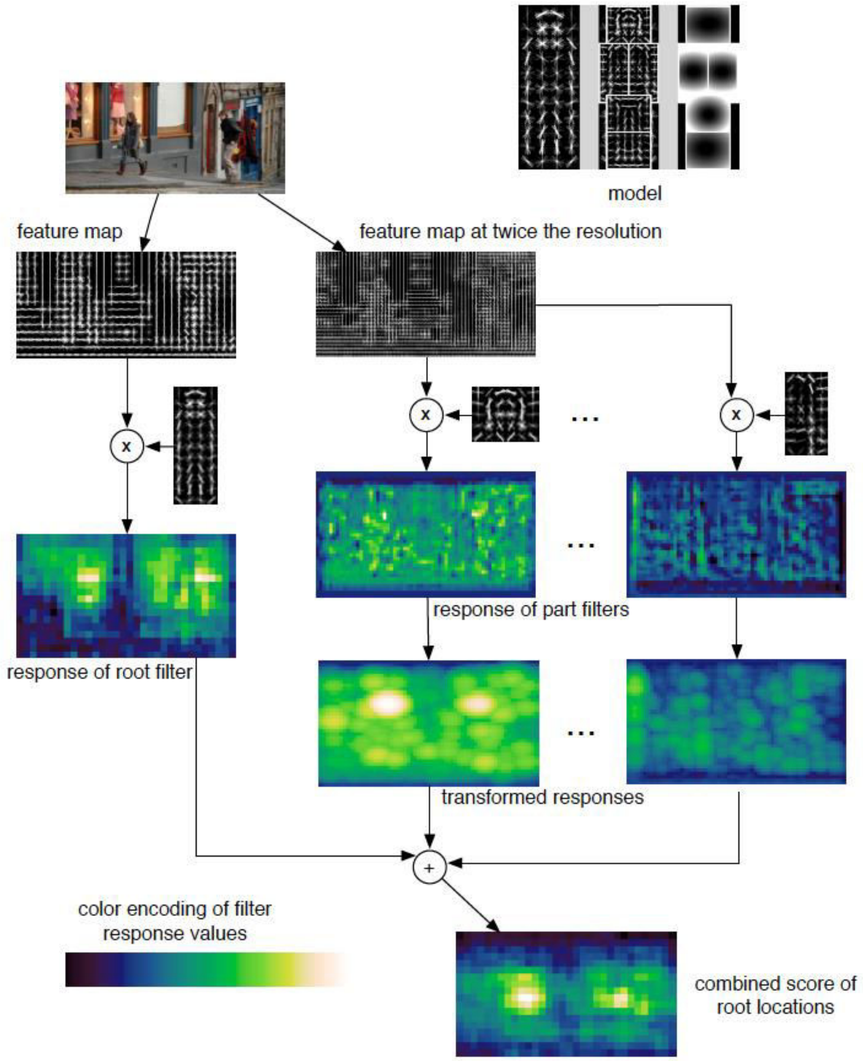
concatenation of HOG features and part displacement features

Matching

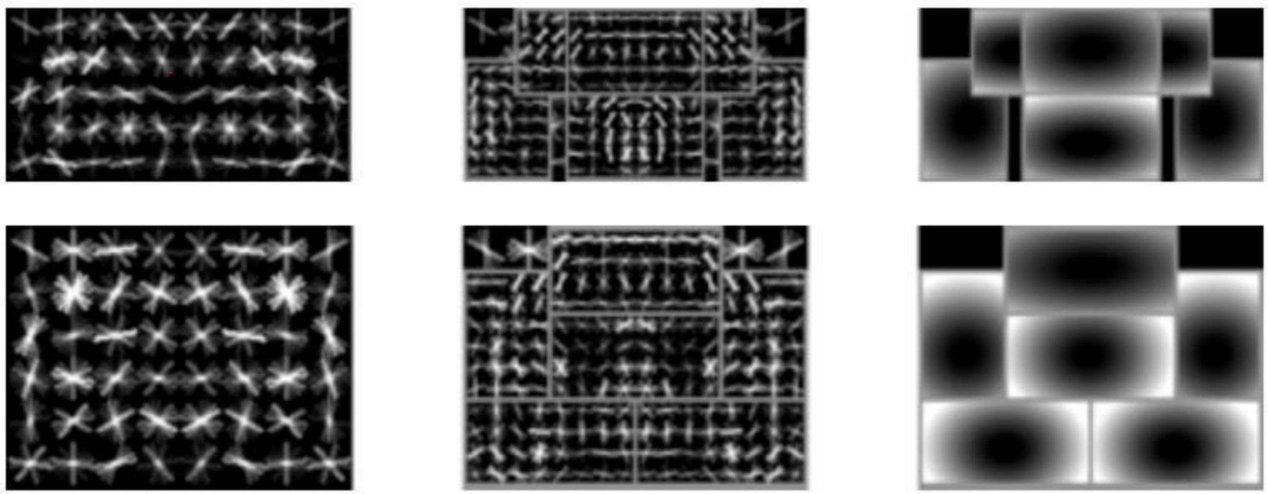
- Define an overall score for each root location
  - Based on best placement of parts

$$\text{score}(p_0) = \max_{p_1,\ldots,p_n} \text{score}(p_0, \ldots, p_n).$$

- High scoring root locations define detections
  - "sliding window approach"
- Efficient computation: dynamic programming + generalized distance transforms (max-convolution)


model
feature map
feature map at twice the resolution
x
x
x
...
...
...
response of part filters
response of root filter
transformed responses
+
color encoding of filter
response values
combined score of
root locations

Car model

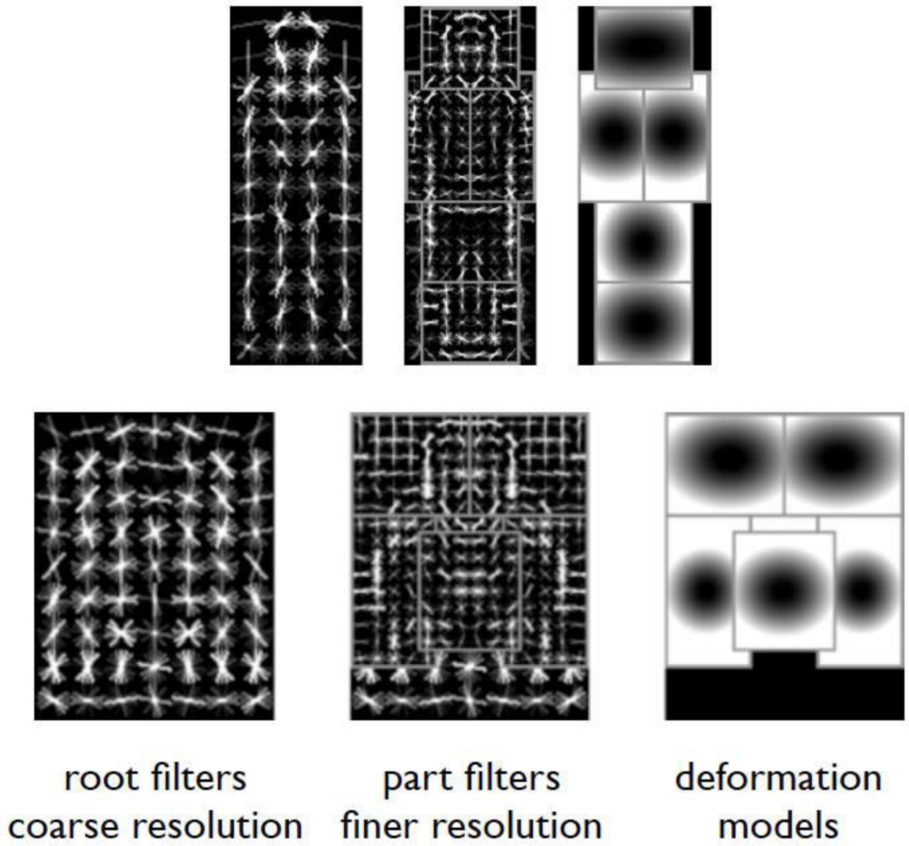


root filters
coarse resolution

part filters
finer resolution

deformation
models

Person model



Precision/Recall results on Bicycles 2008