6.819 / 6.869: Advances in Computer Vision

MIT COMPUTER VISION

Early vision: Image Features

Histogram of Oriented Gradients (HOG)

Carl Vondrick

Lecture TR 9:30AM – 11:00AM
(Room 34-101)

Website:
http://6.869.csail.mit.edu/fa15/

Thought experiment: let's build a person detector (HW4).
Why is this difficult?


variation in illumination


variation in appearance


variation in pose, viewpoint


occlusion & clutter

Slide credit:
Deva Ramanan

Classic "nuisance factors" for general object recognition

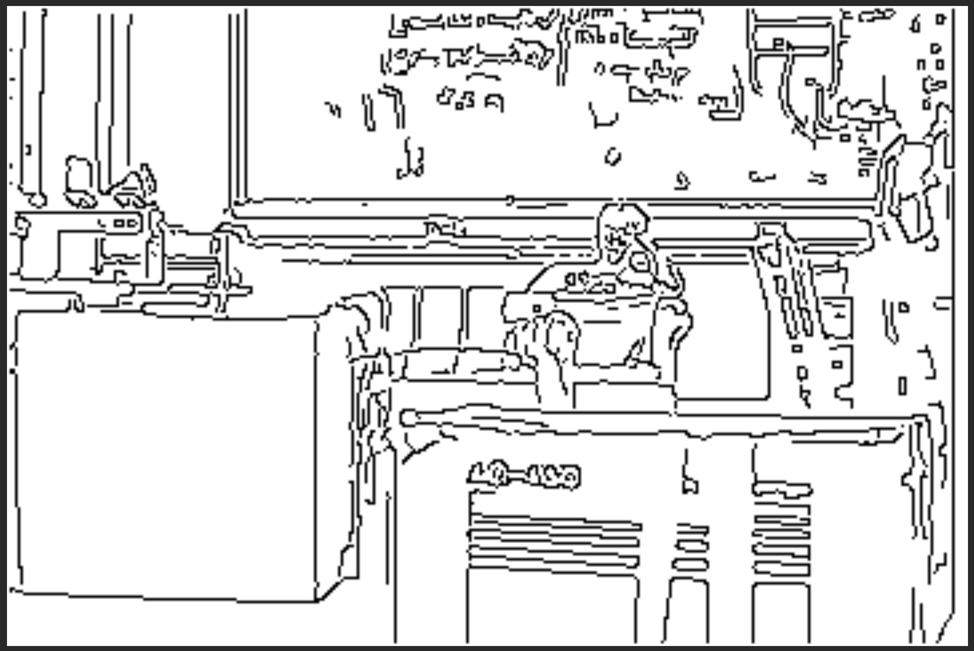
Image intensities

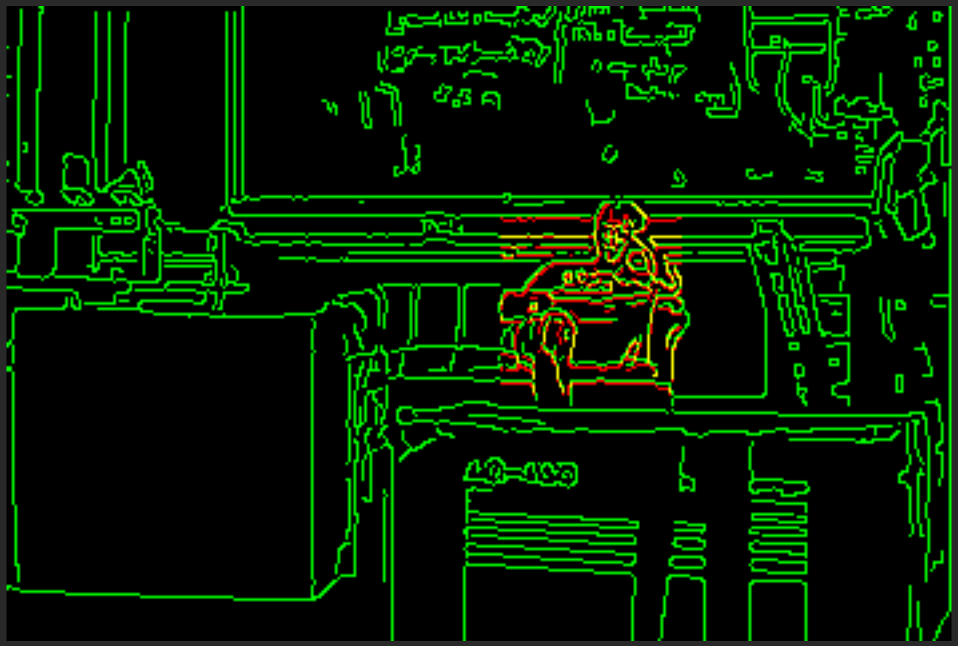
Is this a good enough feature?

Slide credit:
Deva Ramanan

Main idea: use "invariant features"

edges!

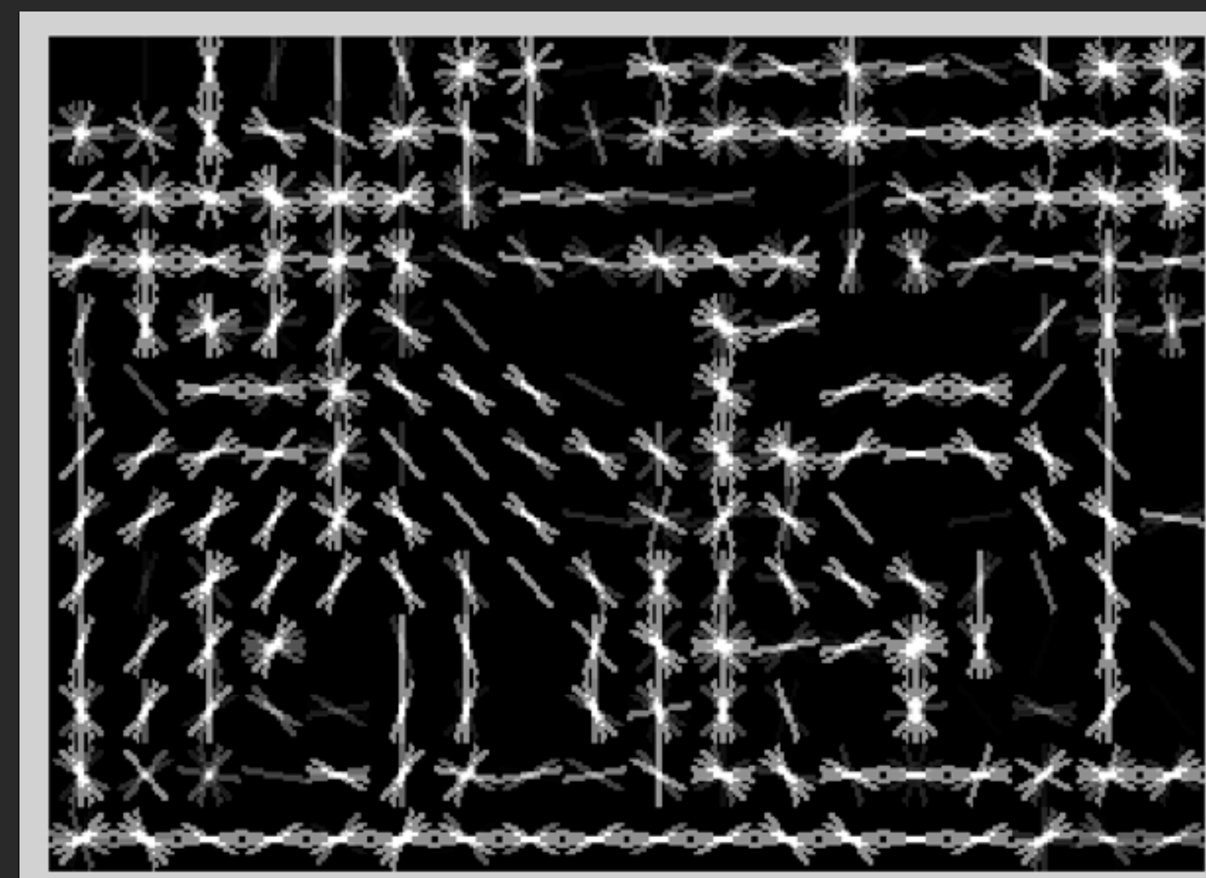




Slide credit:
Deva Ramanan

Image features:

Histograms of oriented gradients (HOG)



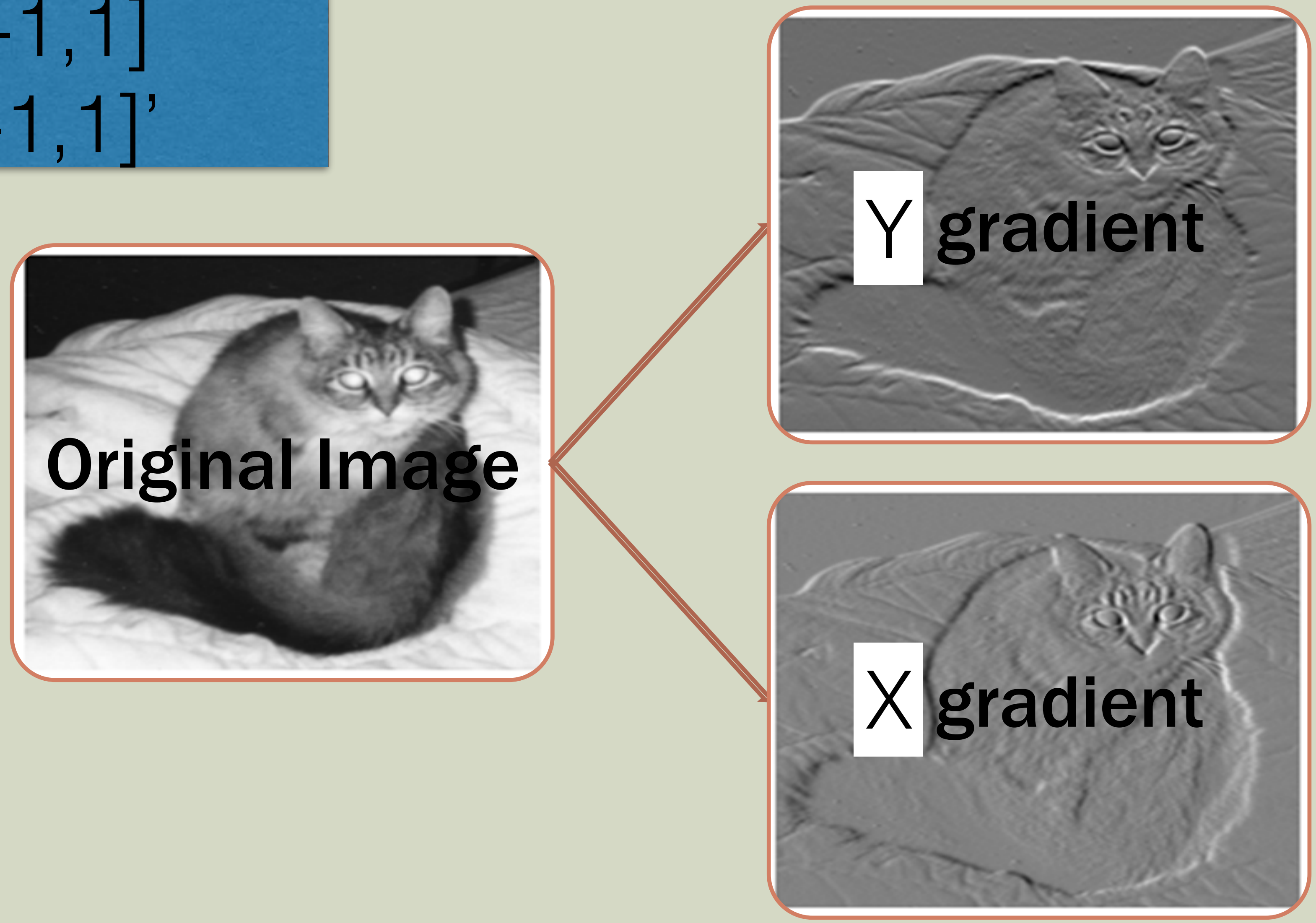
Bin gradients from 8x8 pixel neighborhoods into 9 orientations

(Dalal & Triggs 05)

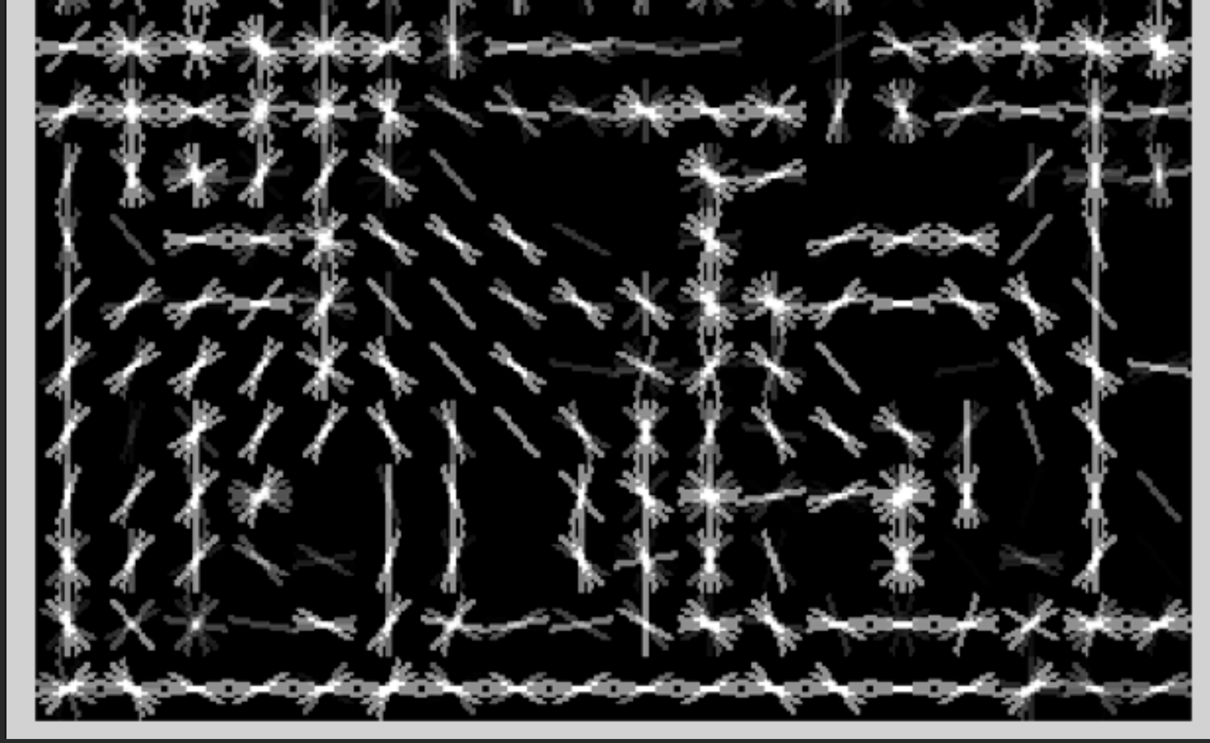
Slide credit: Deva Ramanan

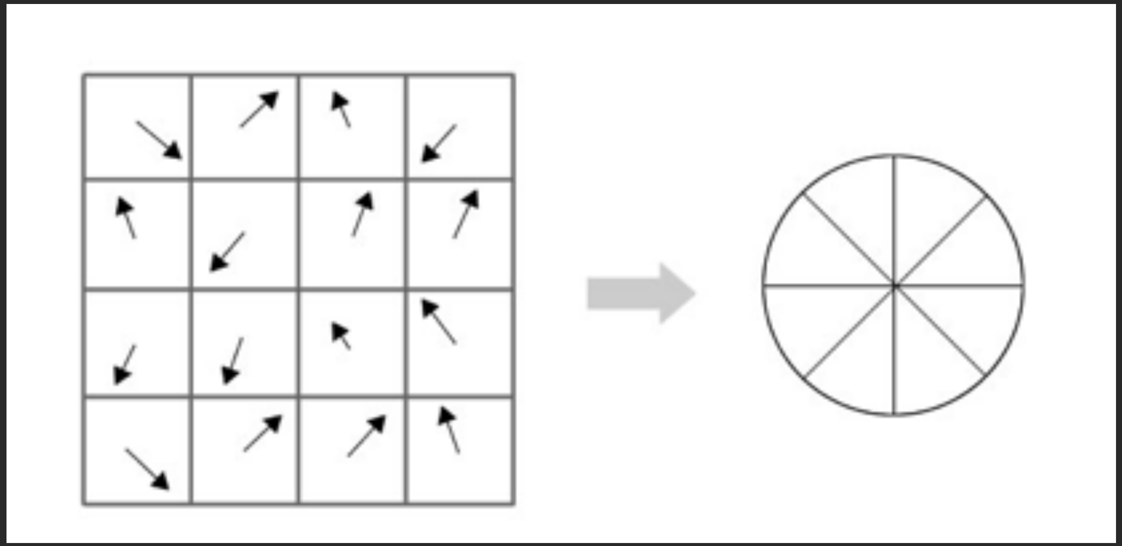
(Simplified) HOG construction

- Convolve the image with discrete derivative mask
  - [-1,1]
  - [-1,1]’



Slide credit:
Deva Ramanan





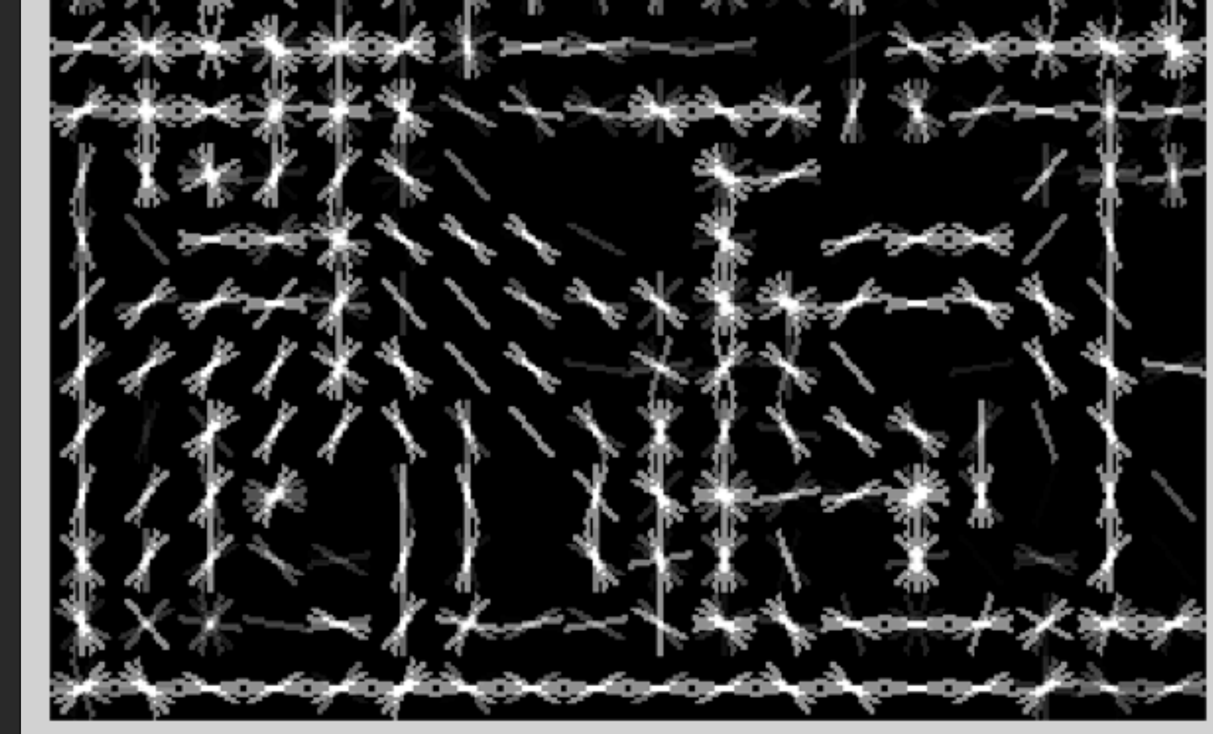
Quantize each gradient into one of $n_o = 9$ orientations

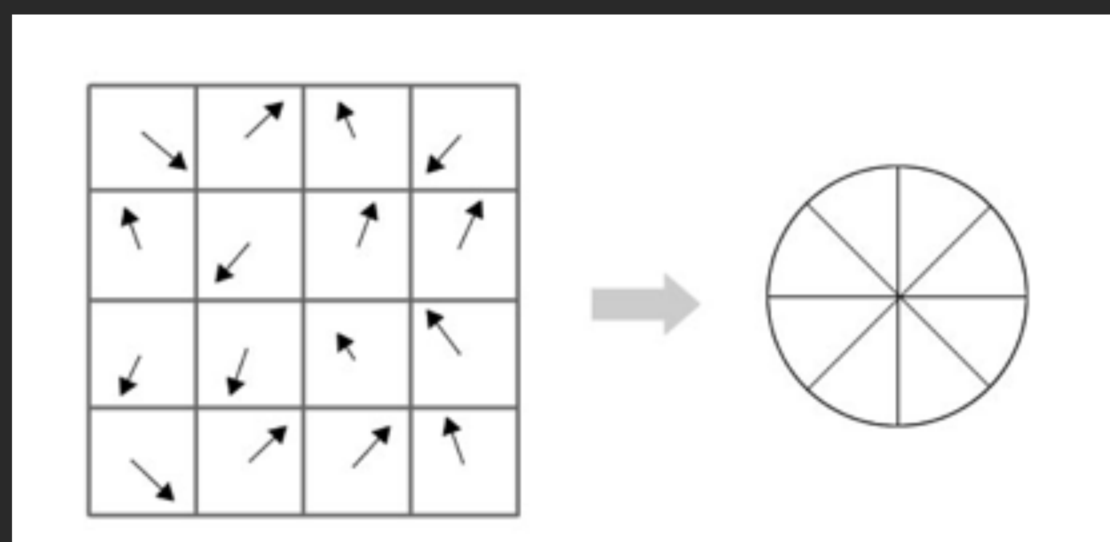
Do we want to put gradients 180 apart in same or different bins?

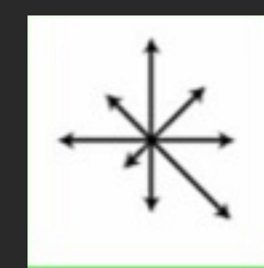
What should be the angle range of each bin?

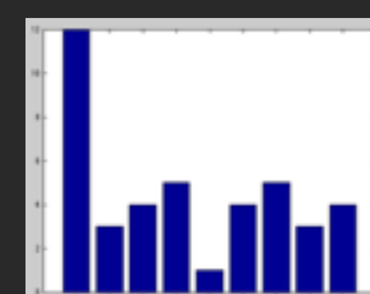
[H x W] ->[H x W x 9]
“orientation channel array”

Slide credit:
Deva Ramanan







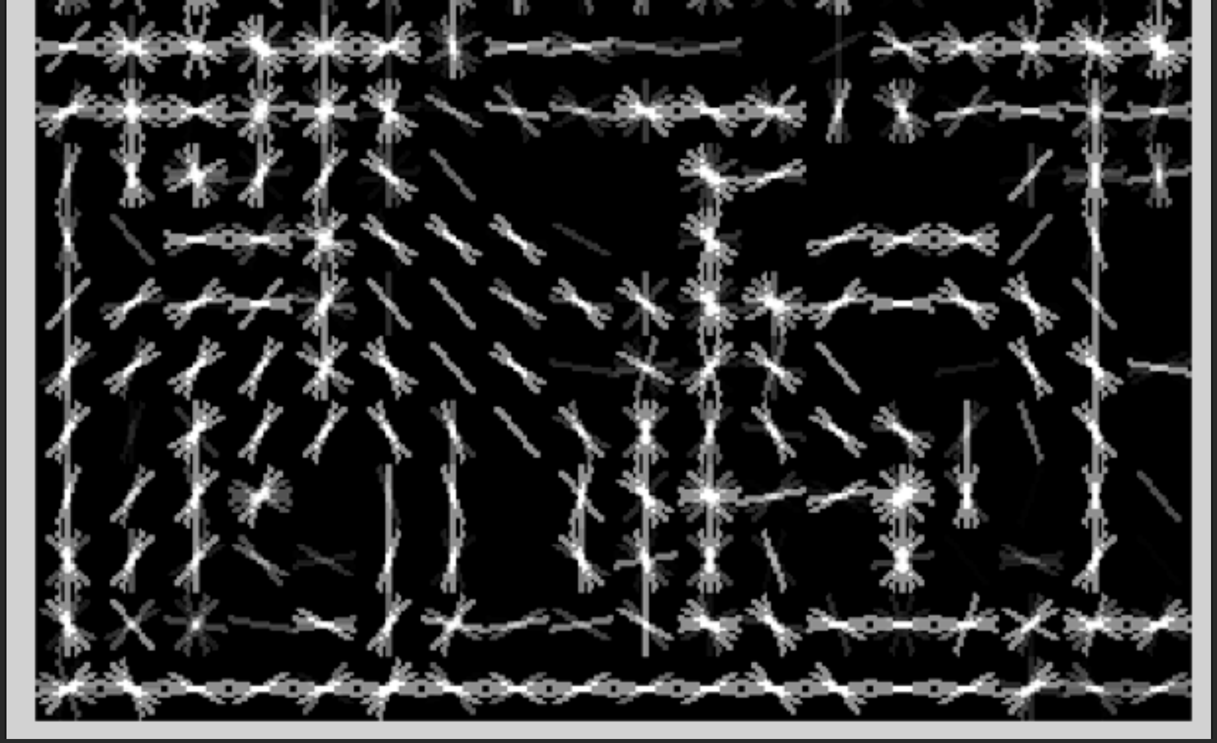


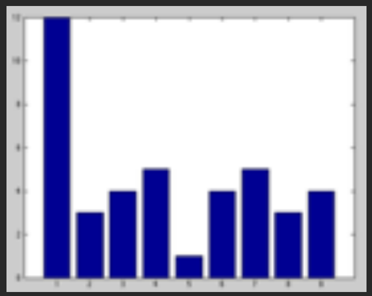
[H x W x 9] – > [H/8 x W/8 x 9]

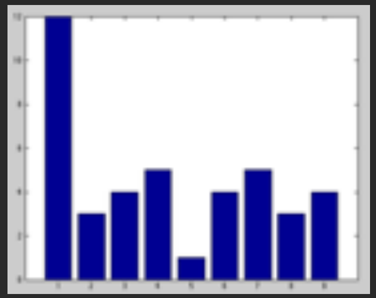
Count up orientation bins over 8x8 pixel neighborhoods. (im2col)

Get some spatial invariance (sort of)…

Slide credit:
Deva Ramanan







[H/8 x W/8 x 9] - > [H/8 x W/8 x 9]

Re-normalize 9 numbers so that their sum is 1

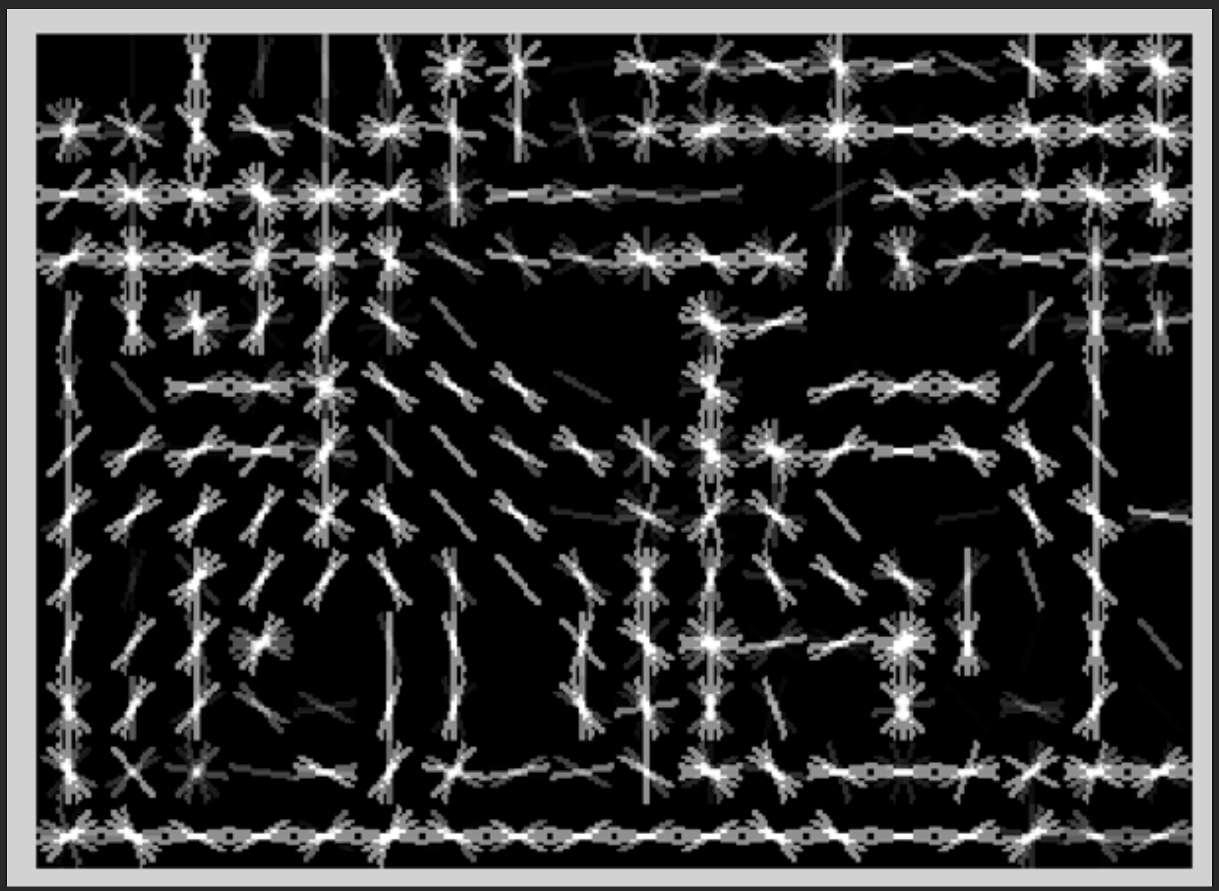
Get some lighting invariance (sort of)...

Slide credit:
Deva Ramanan

Histograms of oriented gradients (HOG)

(note that actual HOG construction is a bit more intricate)

1. Work with raw gradients instead of thresholded gradients
2. Normalize with respect to histograms of 2x2 neighborhoods



Bin gradients from 8x8 pixel neighborhoods into 9 orientations

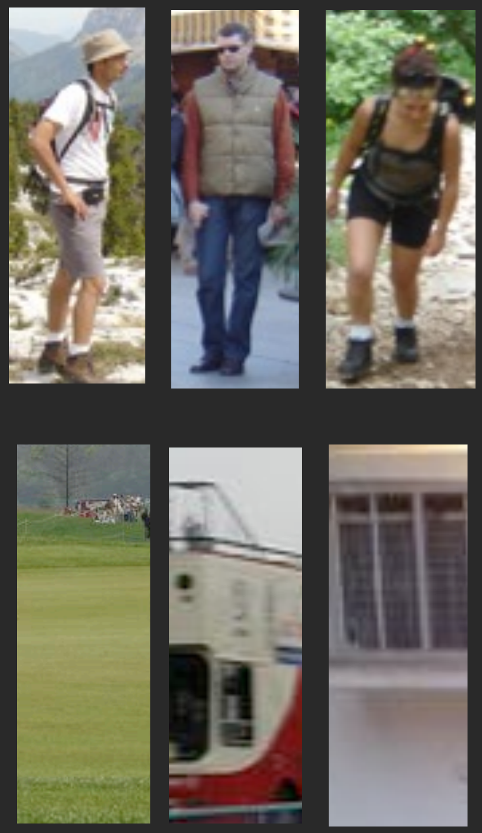
Slide credit: Deva Ramanan

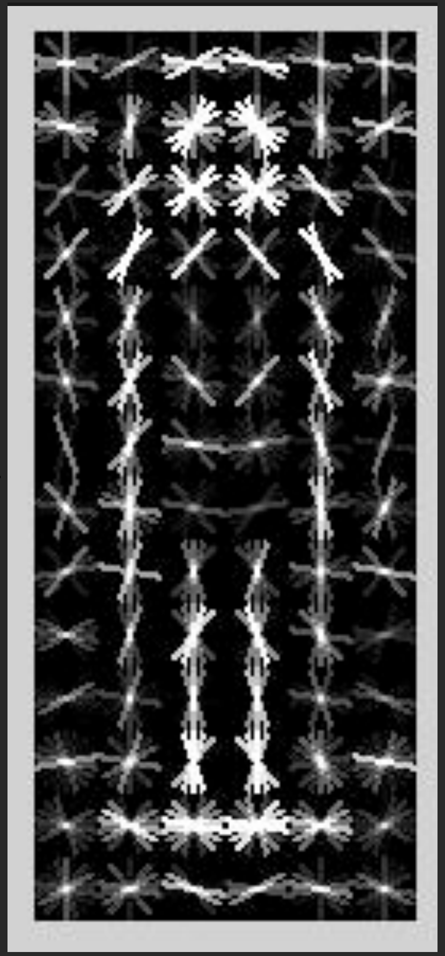
Image [H x W] -> Image Descriptor = [H/8 x W/8 x 9]

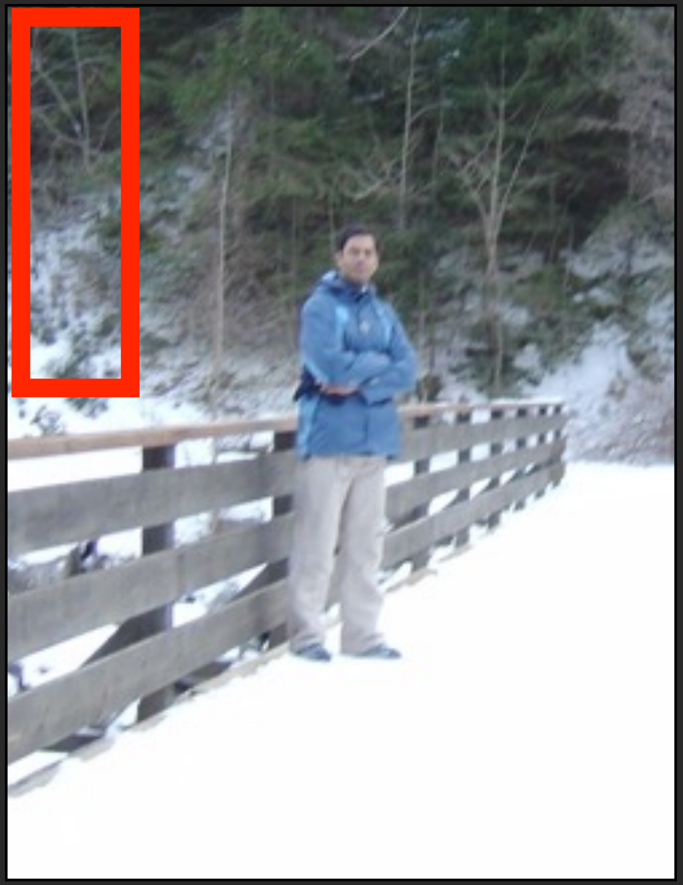
Template classifiers

pos

neg







w

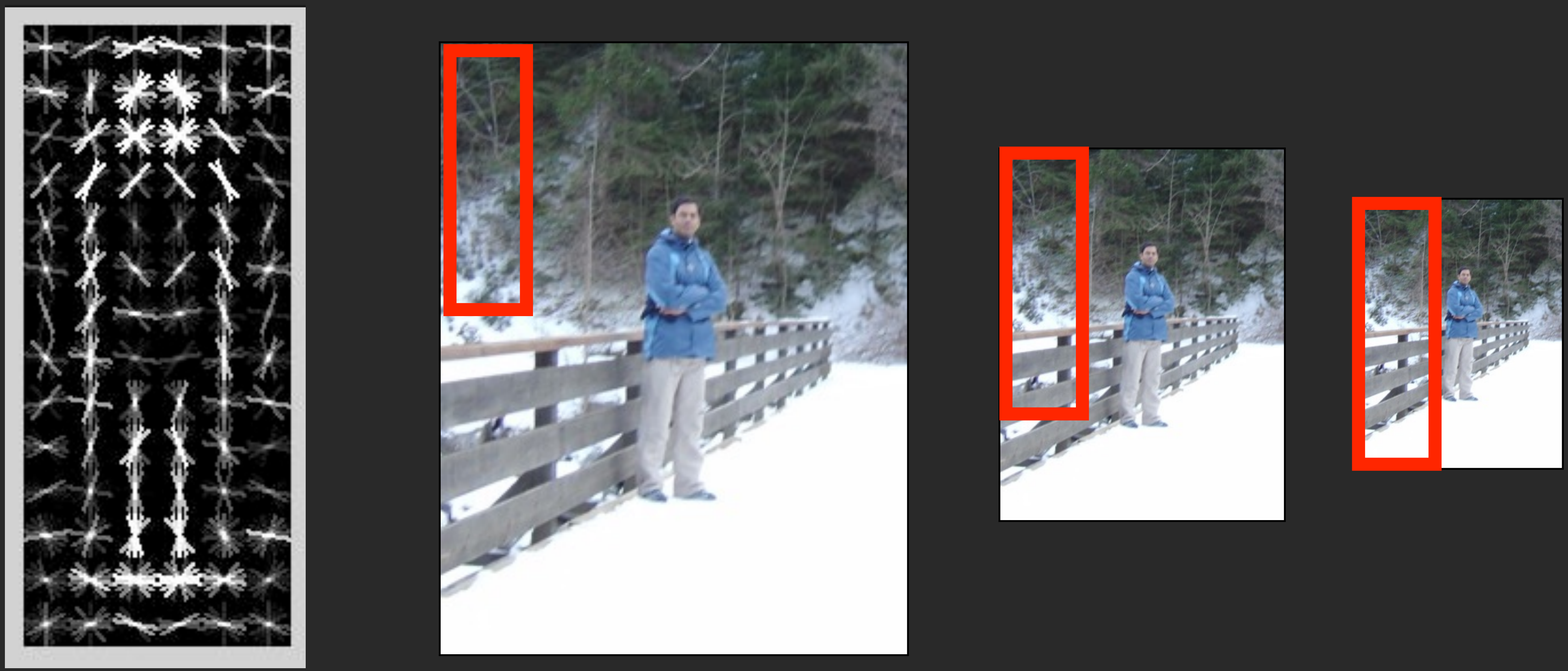
w = weights for orientation and spatial bins 

Slide credit:
Deva Ramanan

$$w \cdot x > 0$$

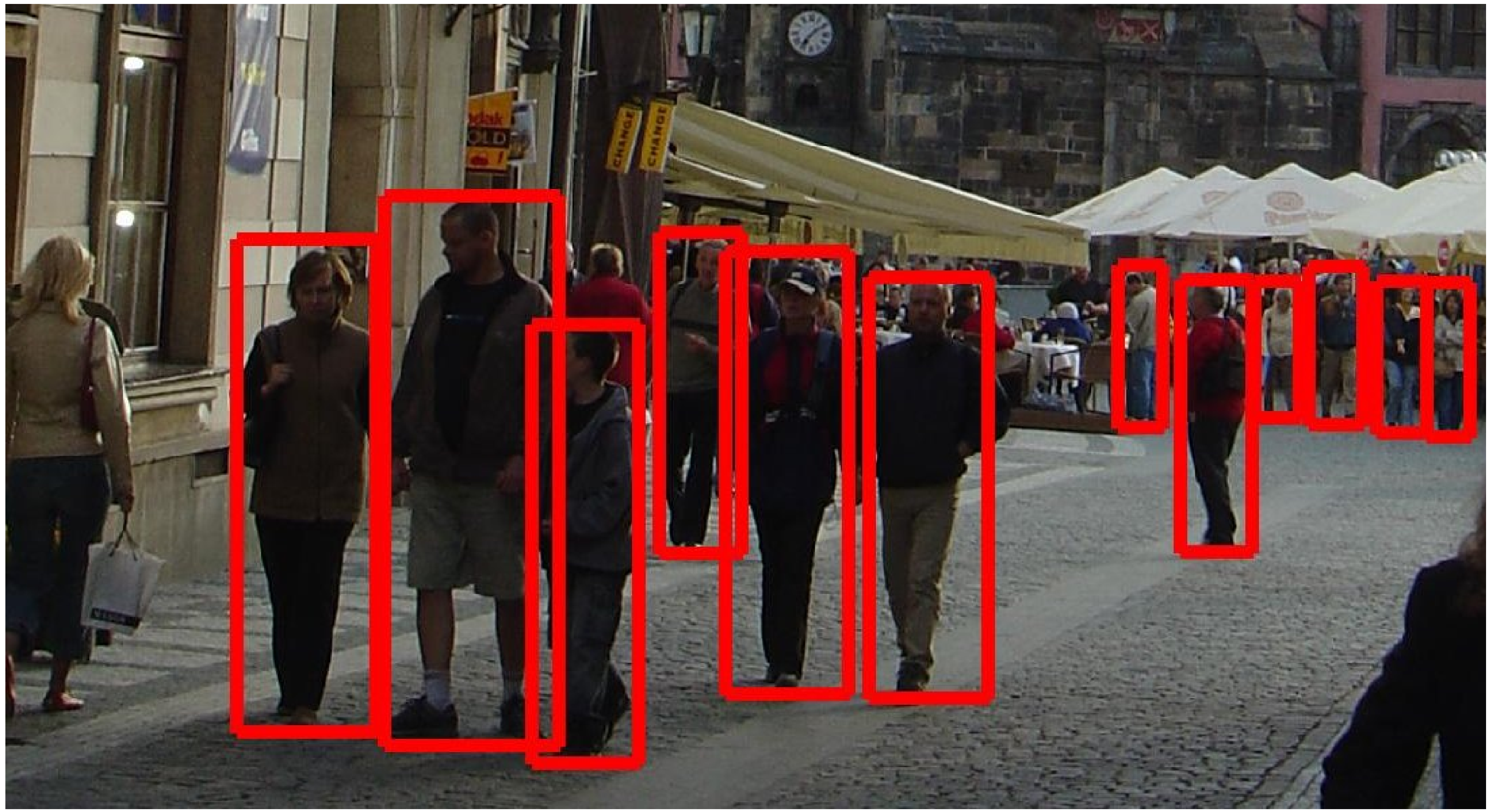
Train with a linear classifier (perceptron, logistic regression, SVMs...)

Search over scales



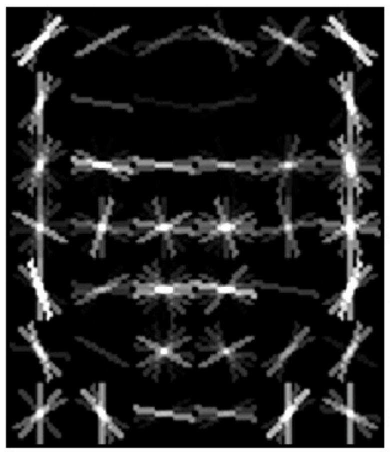
Slide credit:
Deva Ramanan

Pedestrian detection

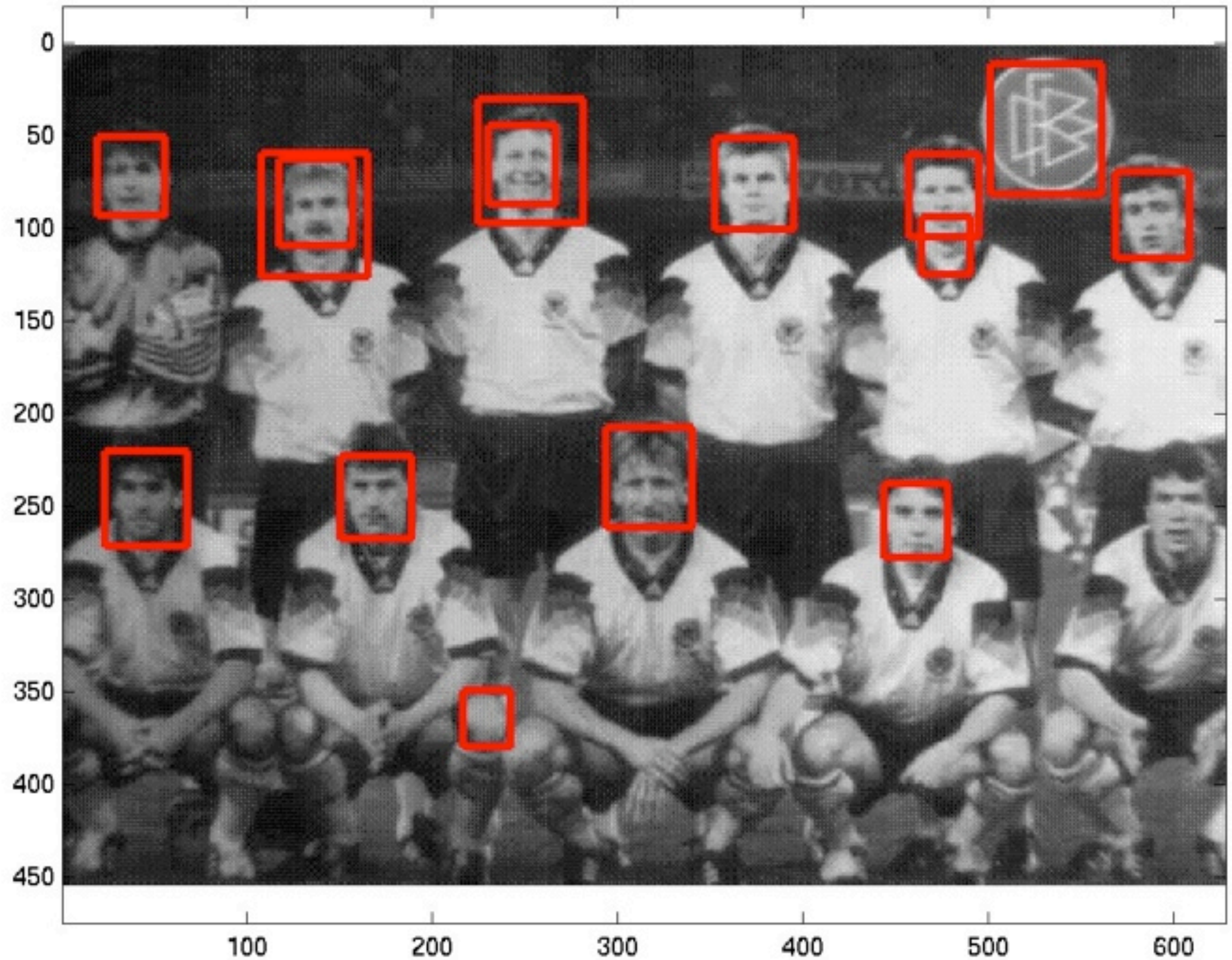


Slide credit:
Deva Ramanan

Face detection

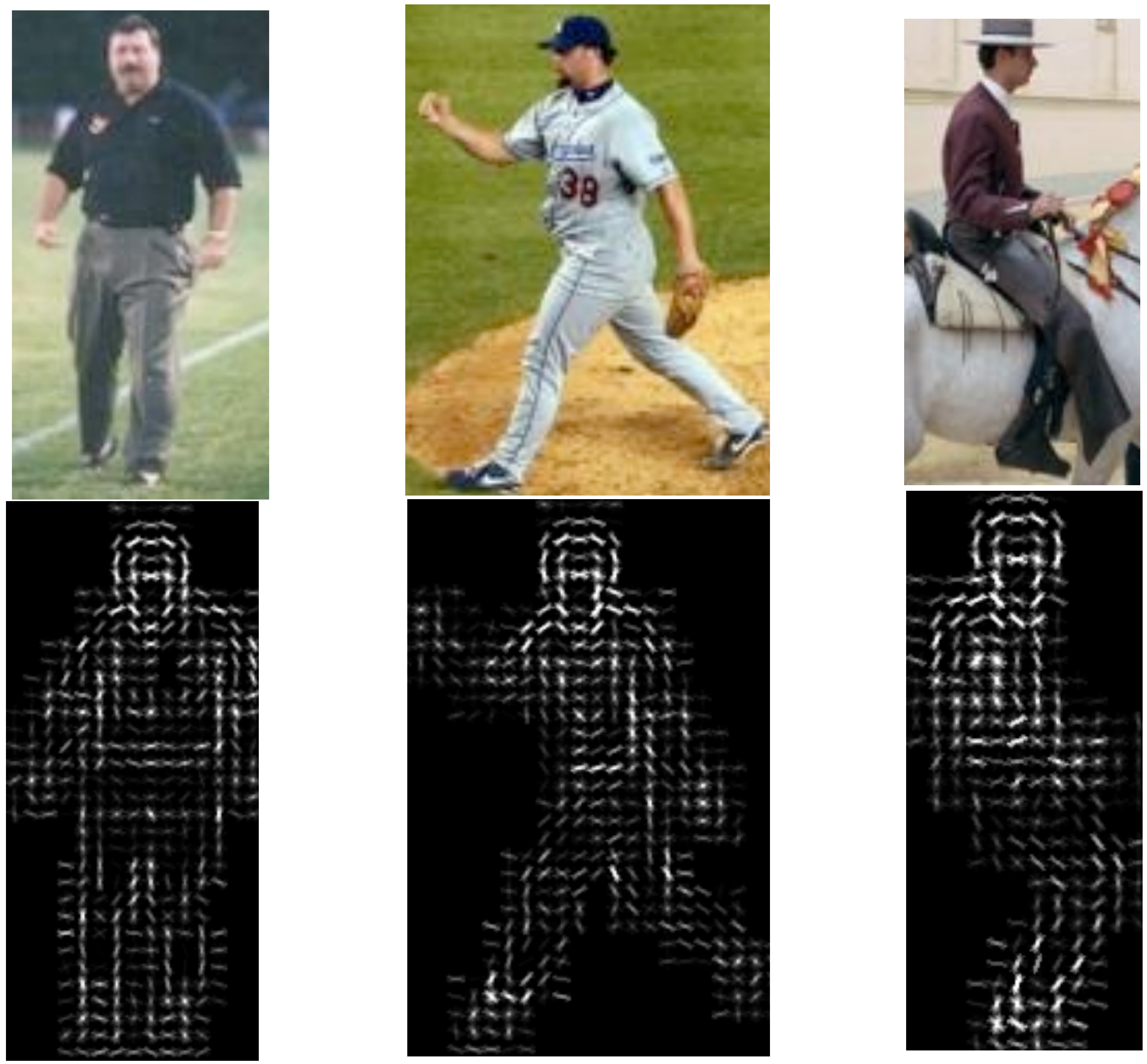


template



Slide credit:
Deva Ramanan

Object subcategories



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

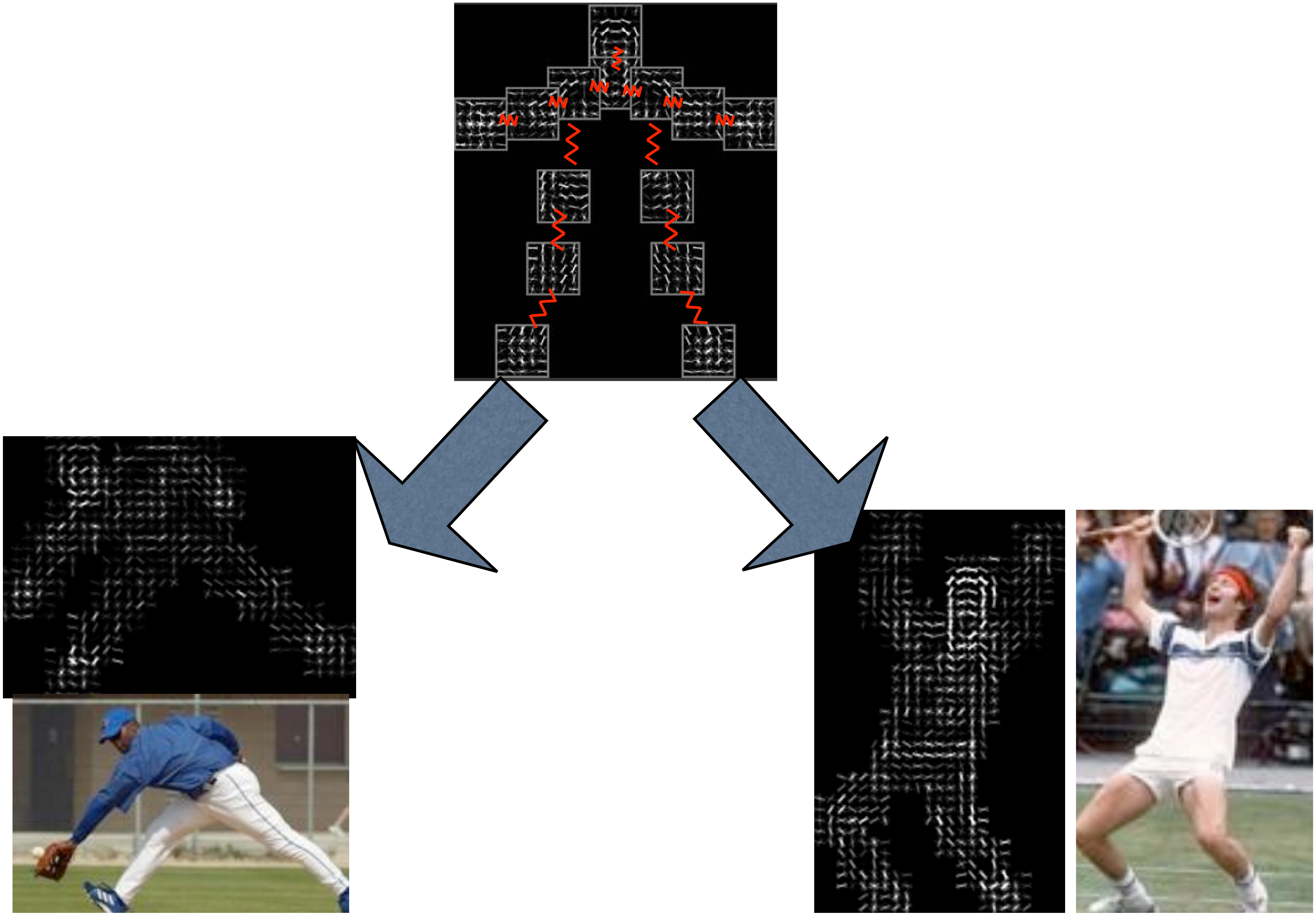
Slide credit:
Deva Ramanan

Object subcategories

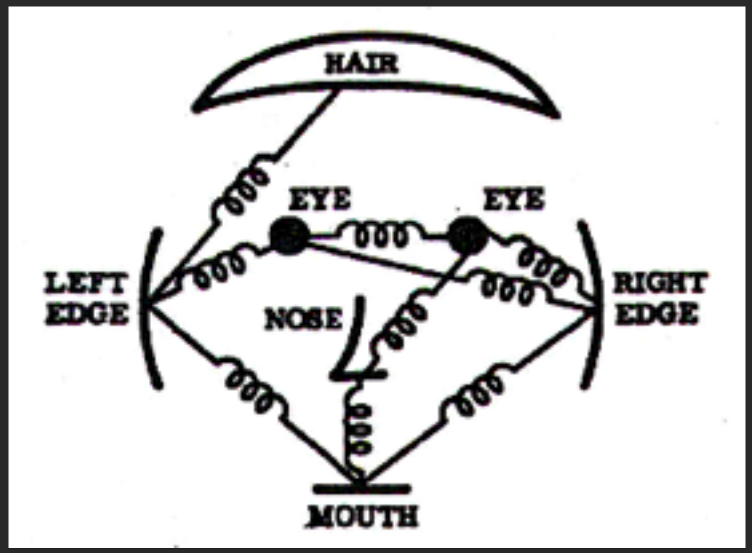
We need lots of templates, and will likely have little data of 'yoga twist' poses

Slide credit: Deva Ramanan

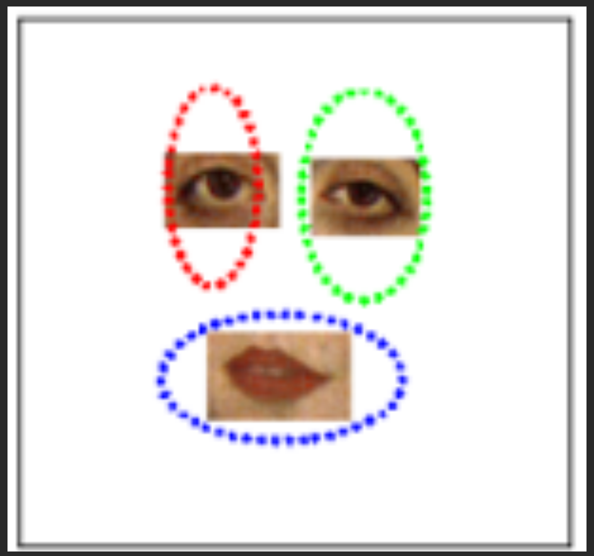
Slide credit:
Deva Ramanan



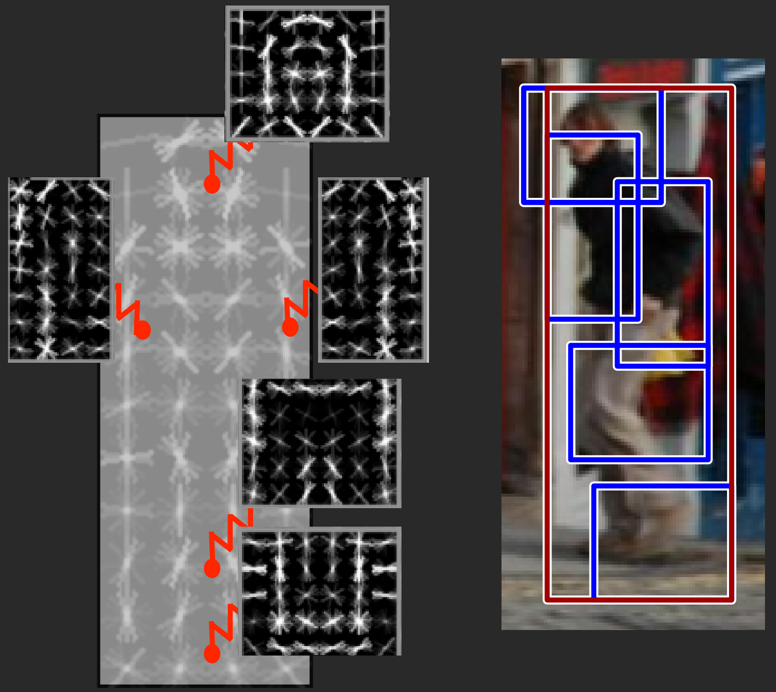
History over 40 years



Pictorial structures



Constellation models



Deformable part models

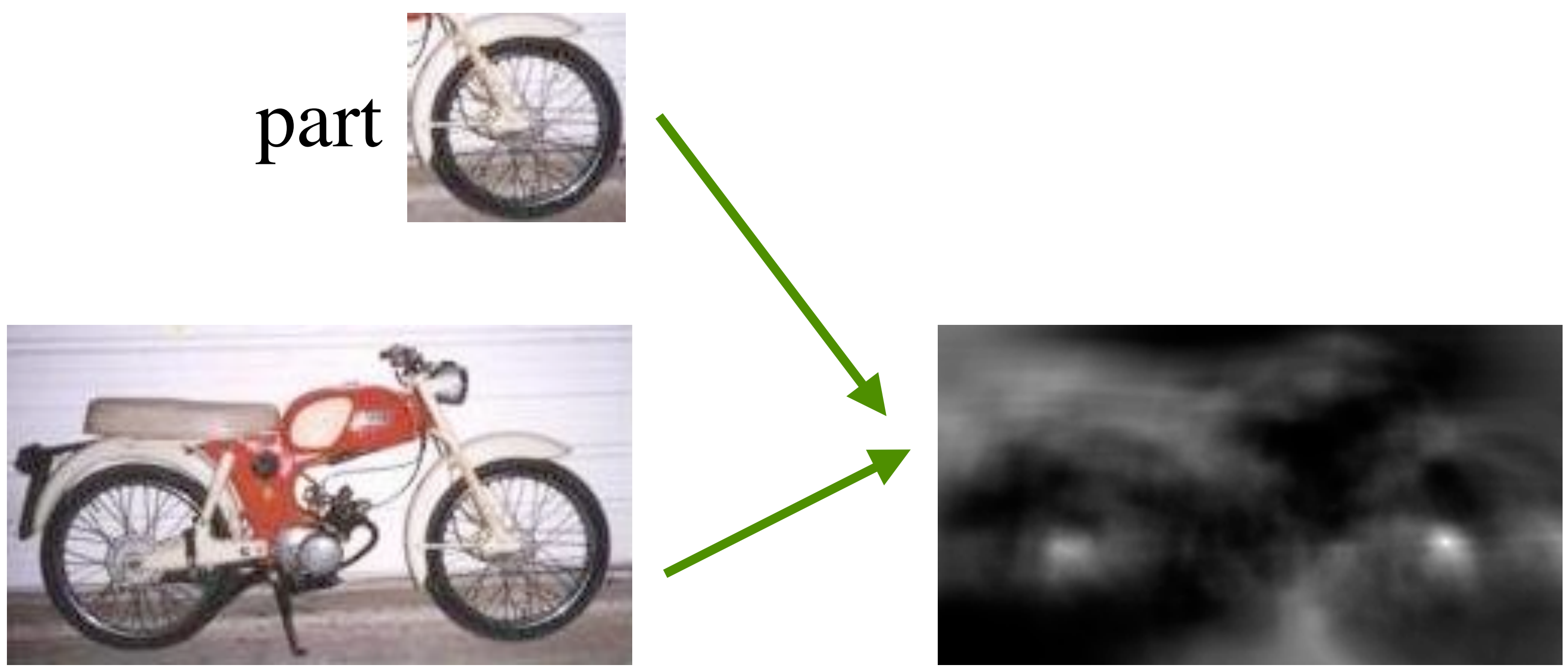
Model encodes local appearance + pairwise geometry

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)

Slide credit: Deva Ramanan

Local evidence + global decision

- Parts have a match quality at each image location.
- Local evidence is noisy.
  - Parts are detected in the context of the whole model.



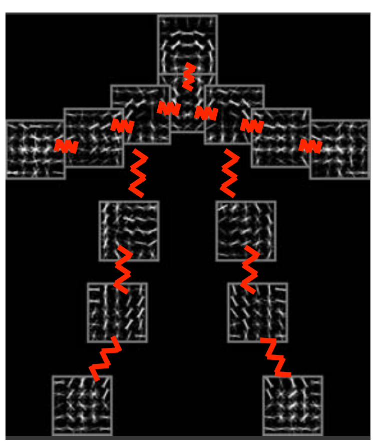
test image match quality

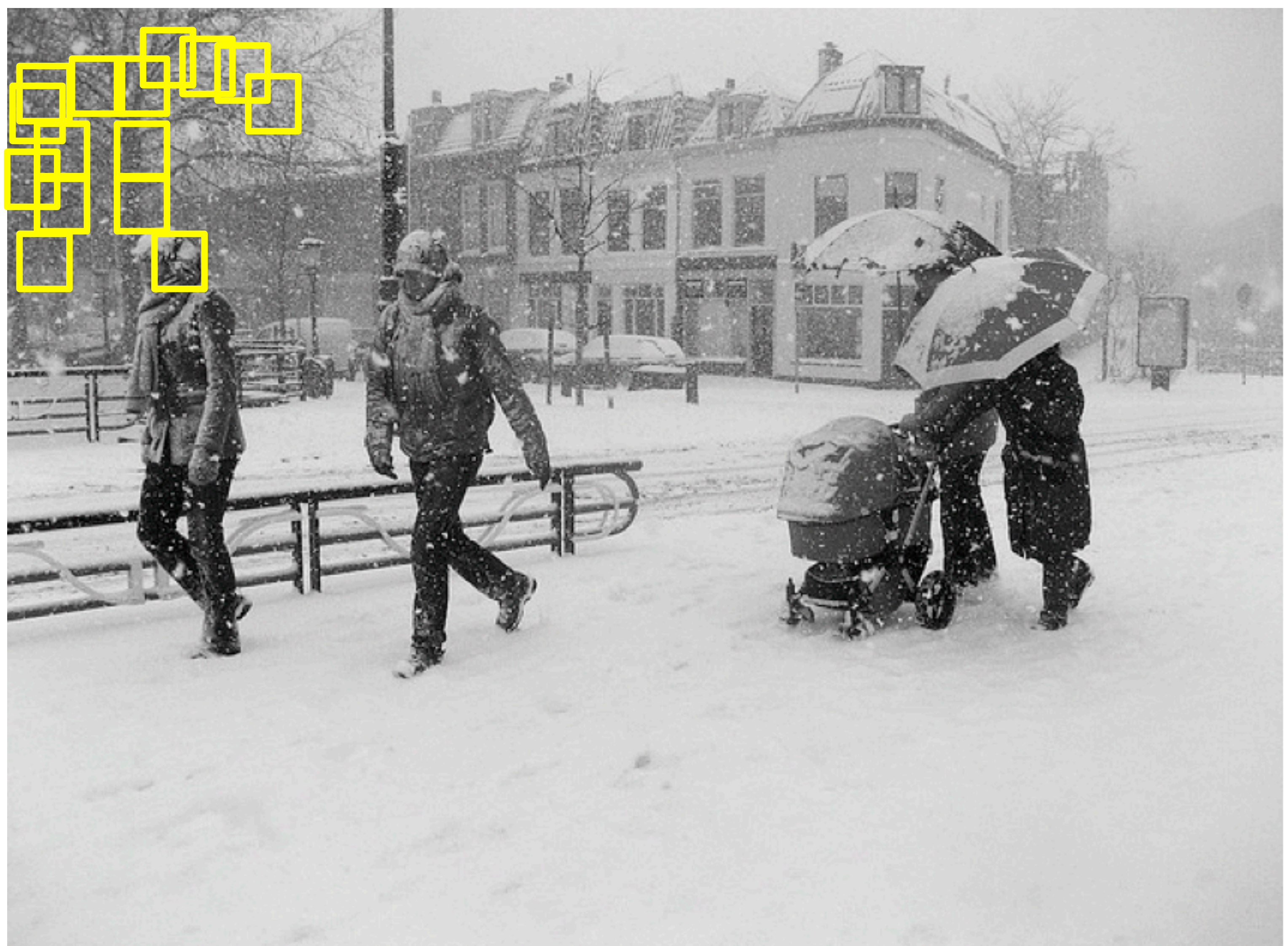
Original PS paper used a vector of filter outputs (“jet”) to define feature
Turns out that HOG works much better

Slide credit:
Deva Ramanan

$$S(l) = \sum_{i \in V} Local(l_i) + \sum_{ij \in E} Pair(l_i, l_j)$$

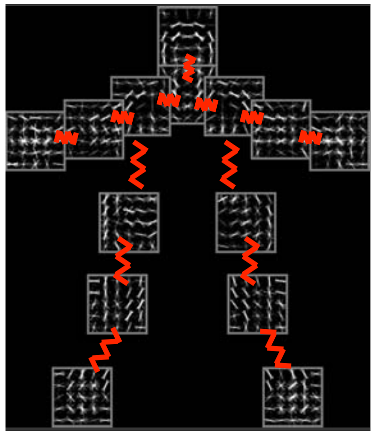
$$l_i = (x_i, y_i)$$

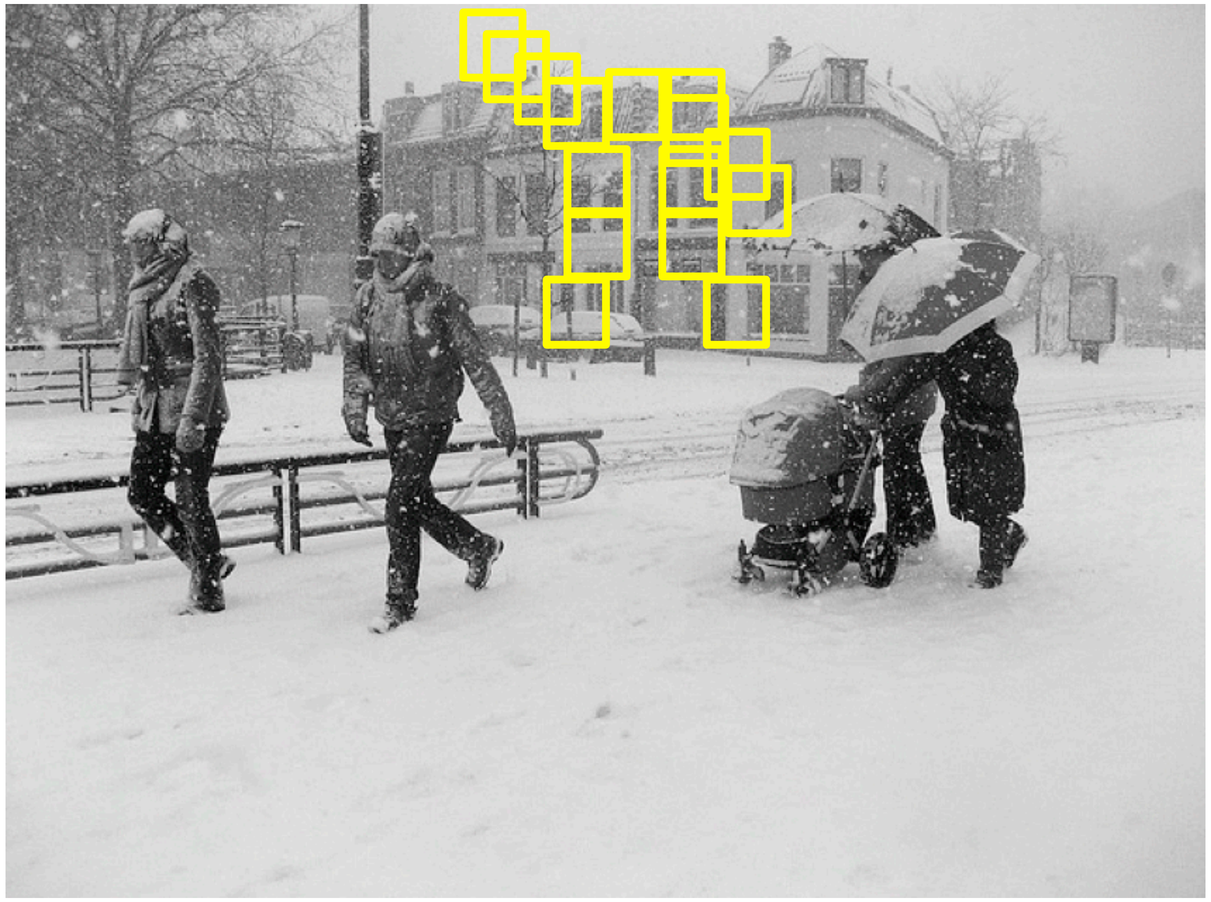


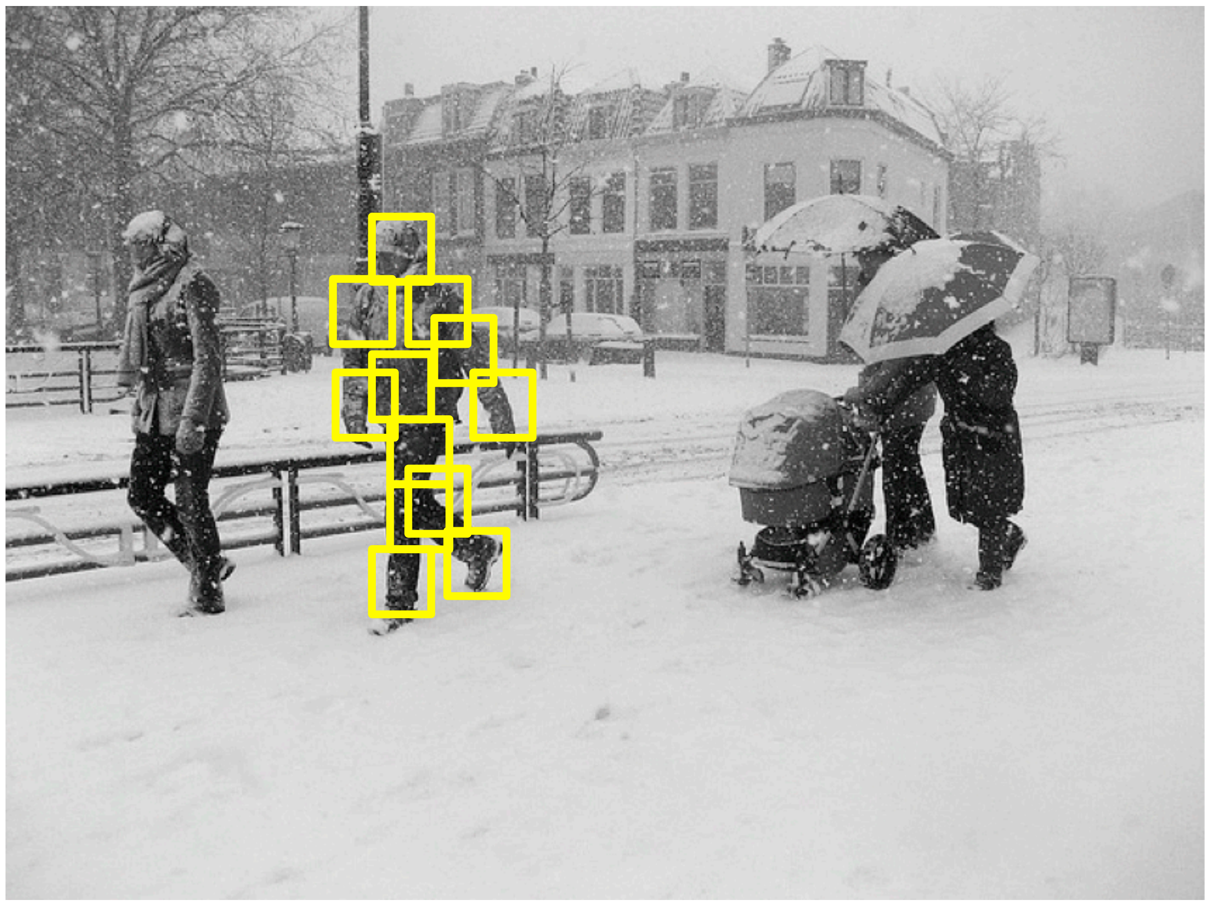


Slide credit: Deva Ramanan

Scanning window detection

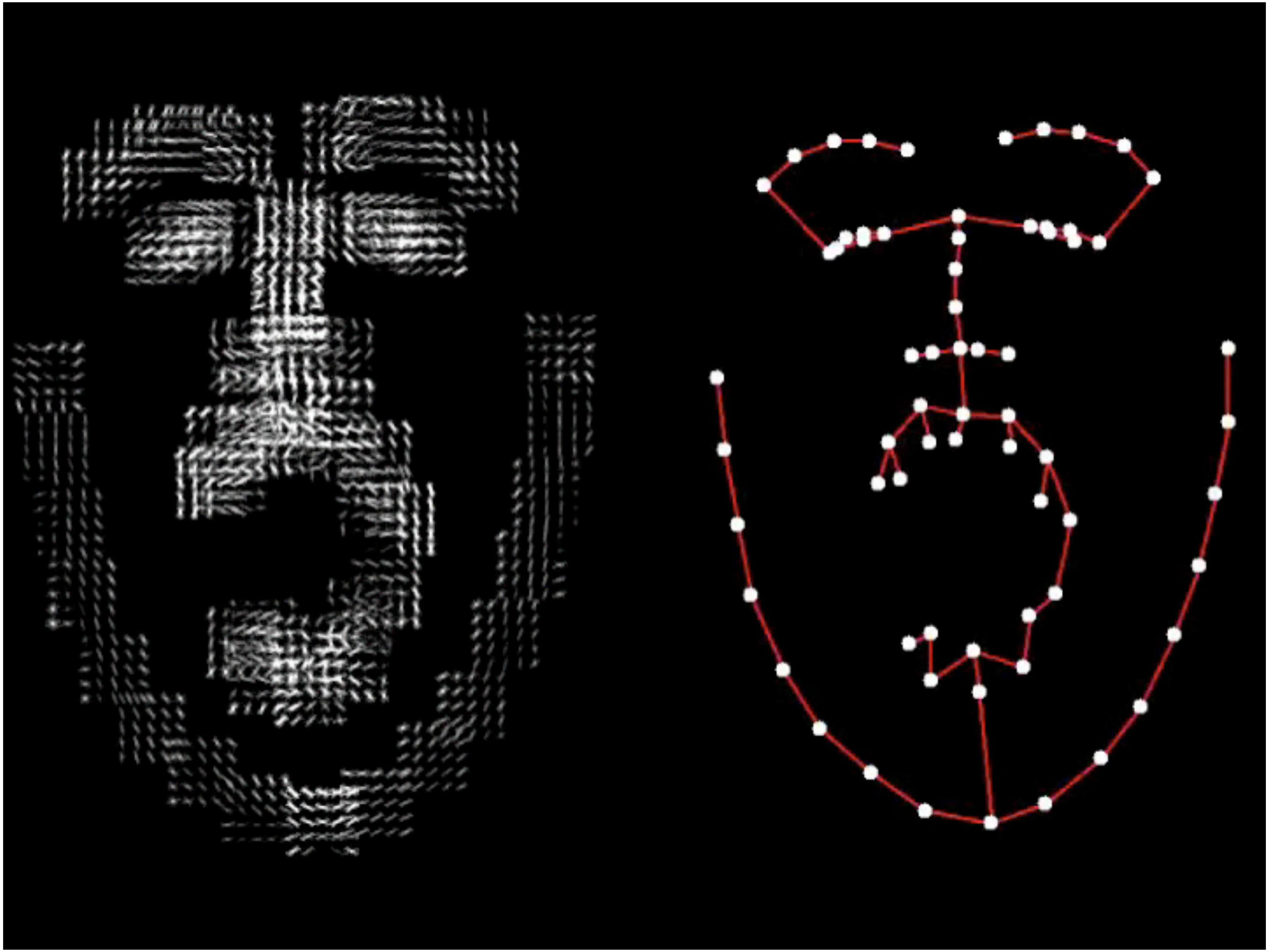






K parts with L possible positions: efficiently score all L^K configurations

Slide credit:
Deva Ramanan



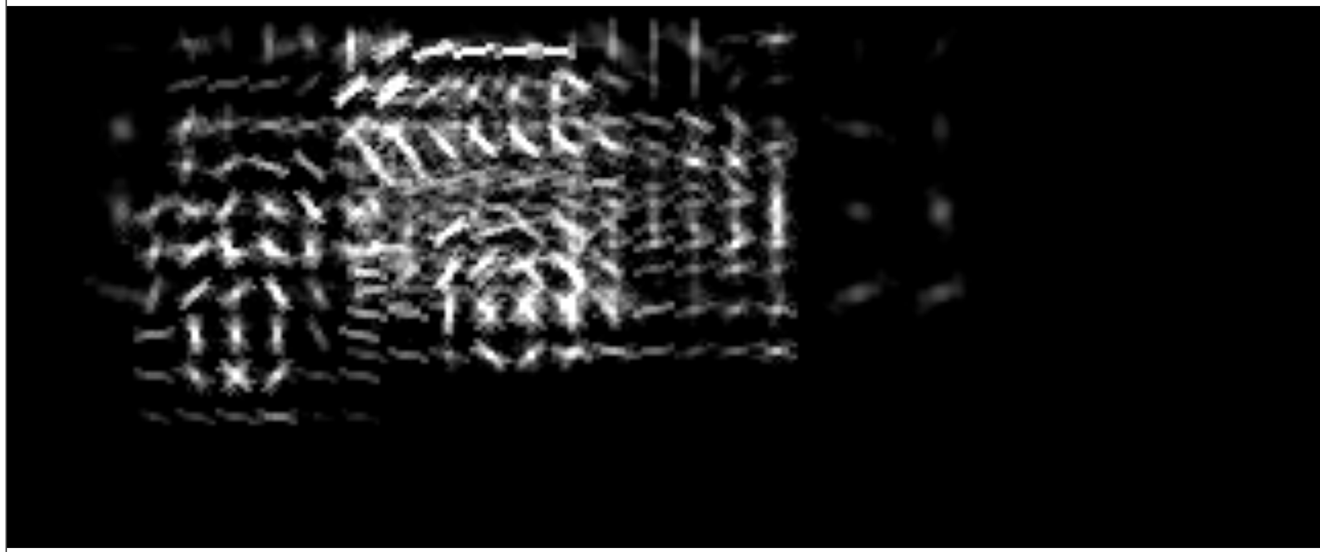
Slide credit:
Deva Ramanan

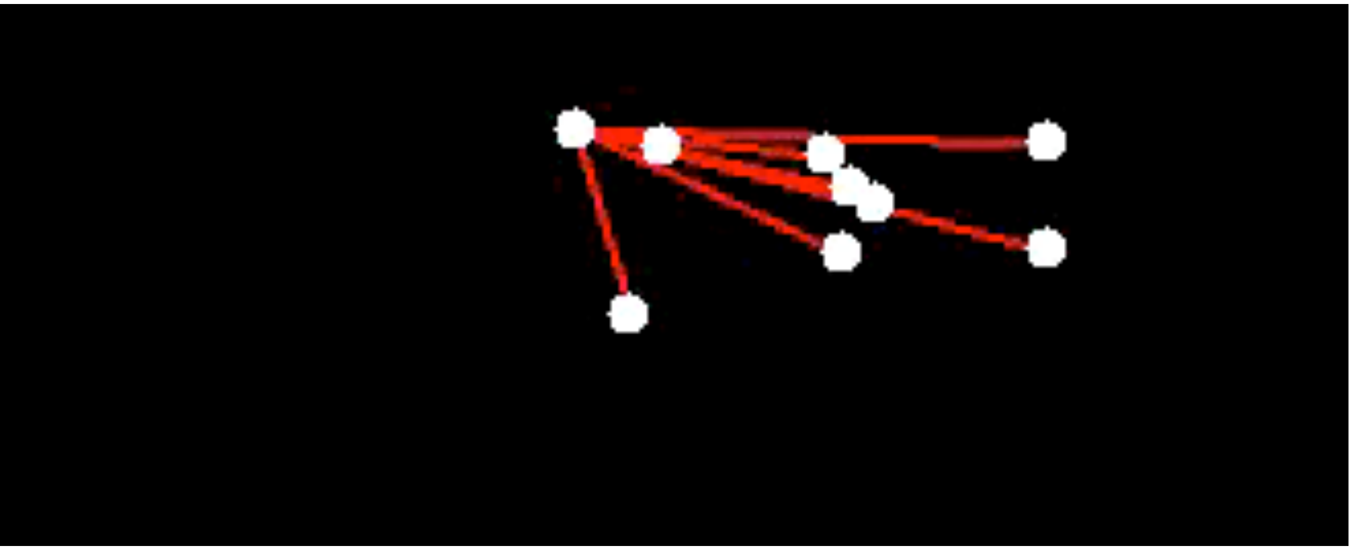
Facial analysis

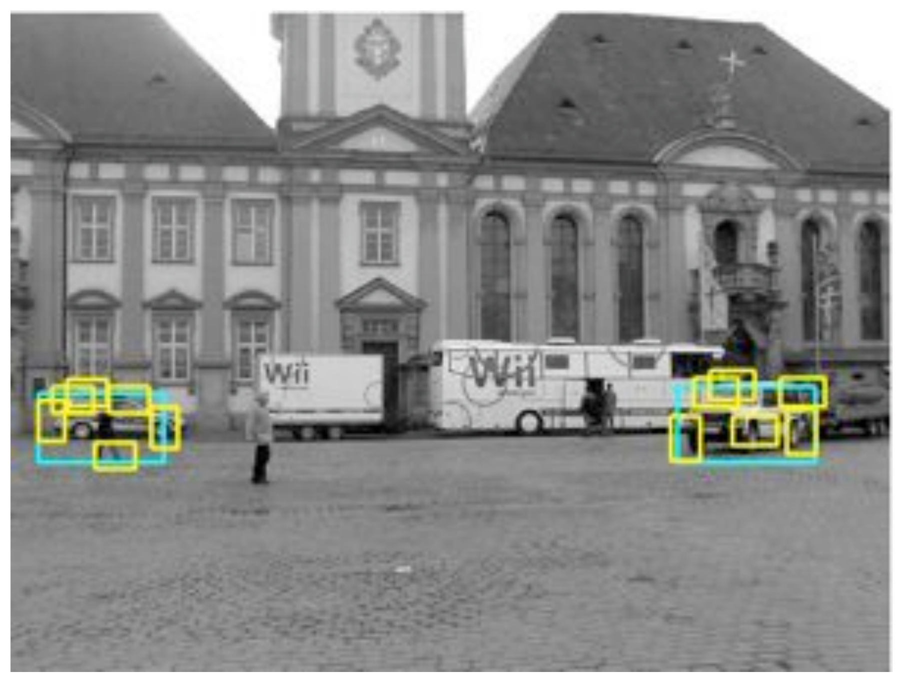
About as accurate as Google Picassa

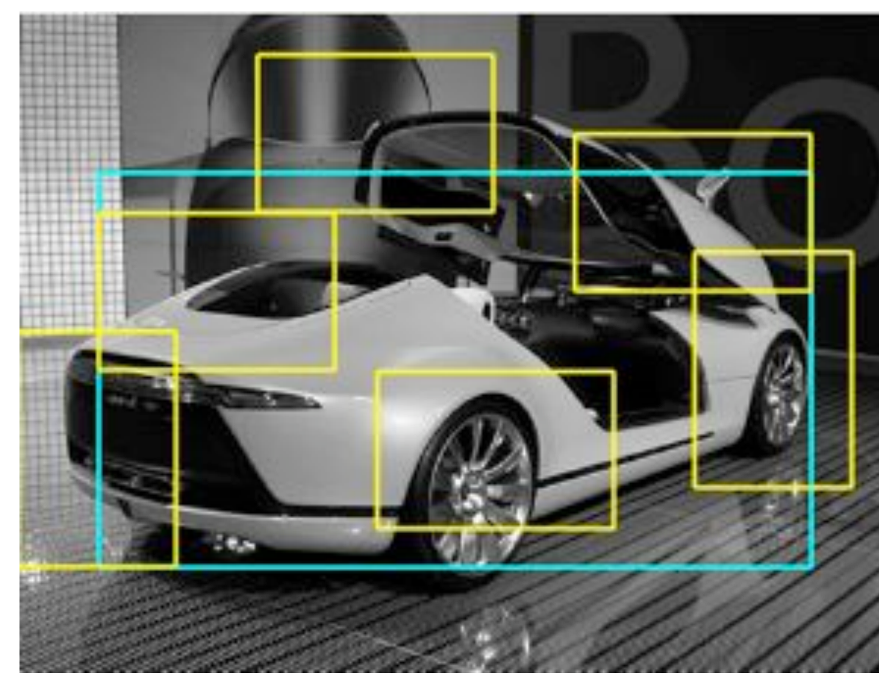
Slide credit:
Deva Ramanan

Example object models



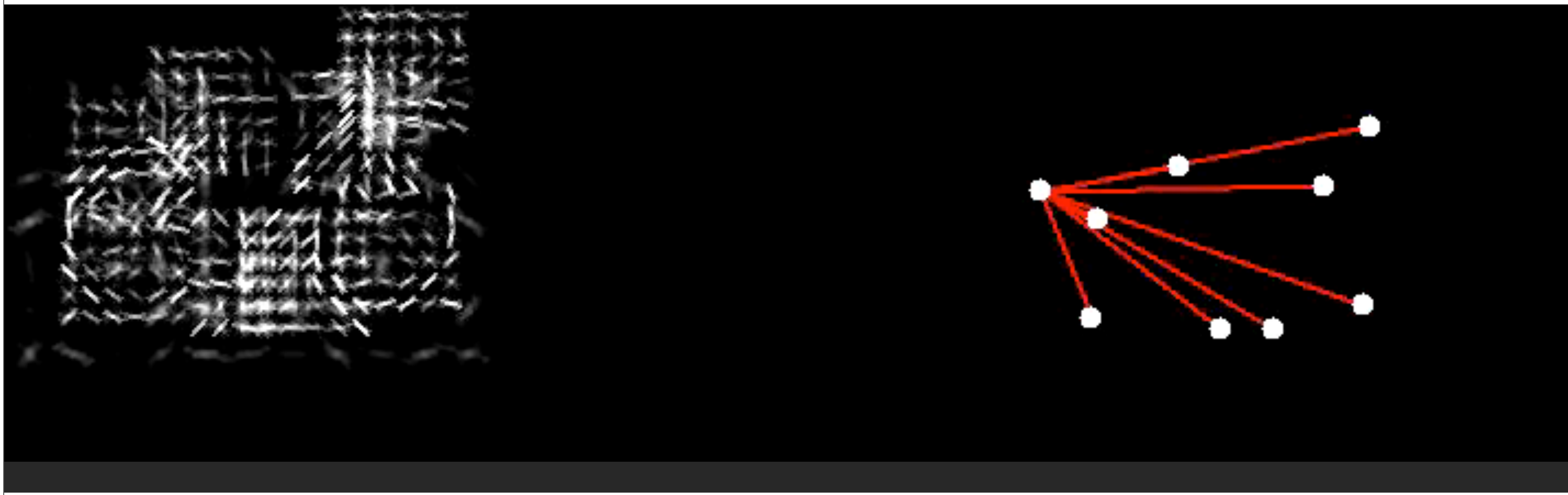


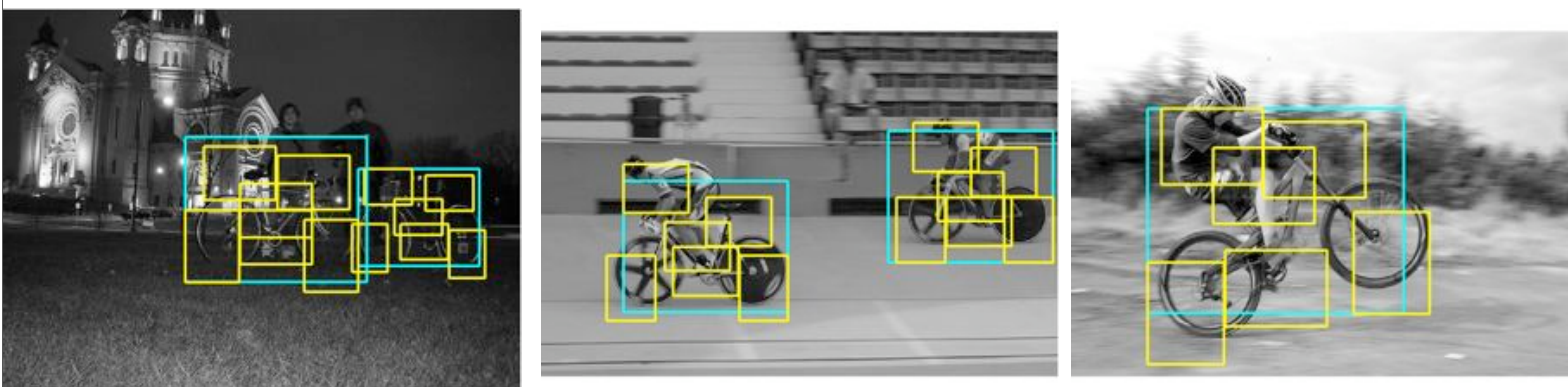




Slide credit: Deva Ramanan

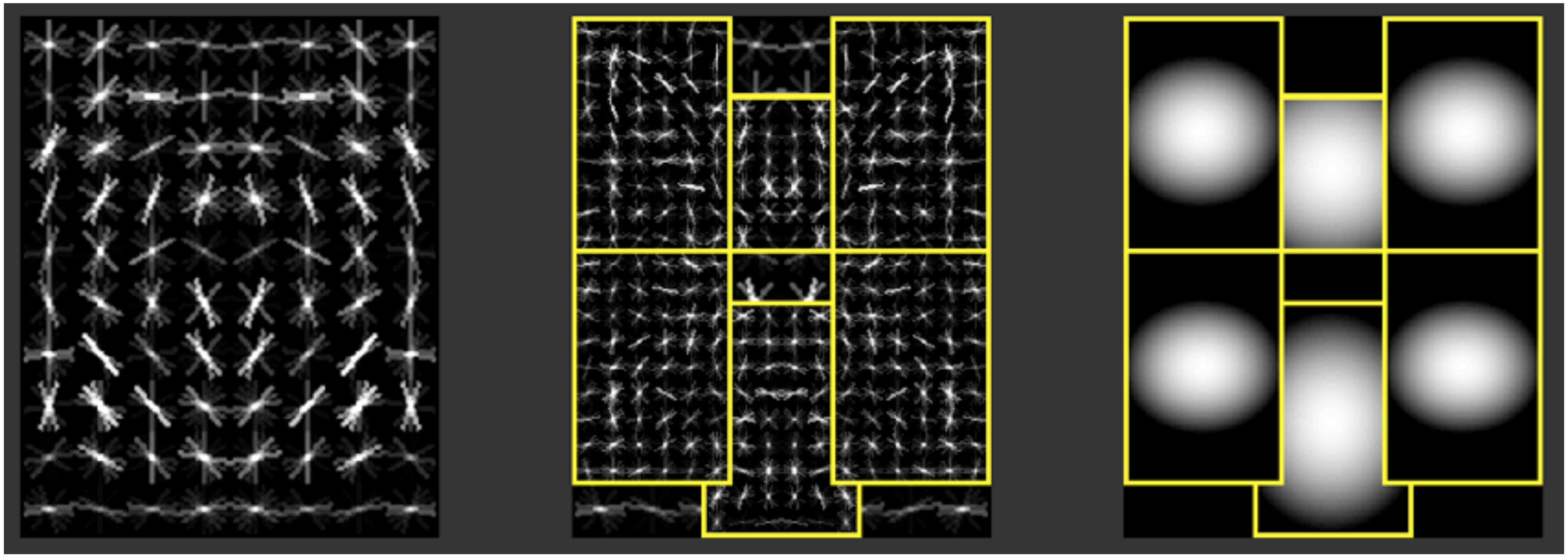
Example object models

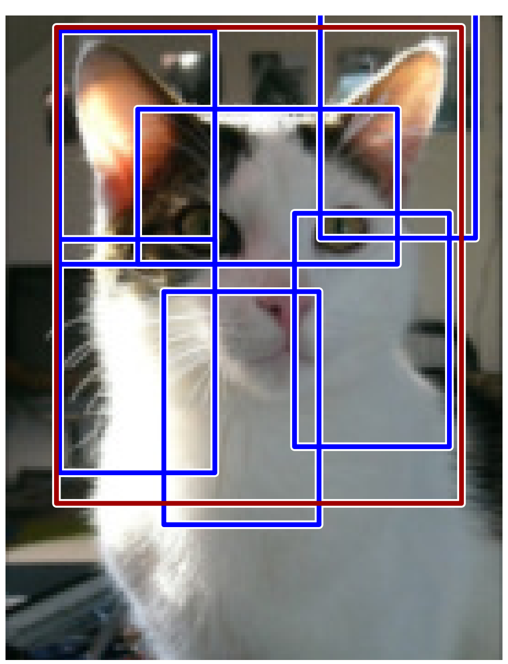


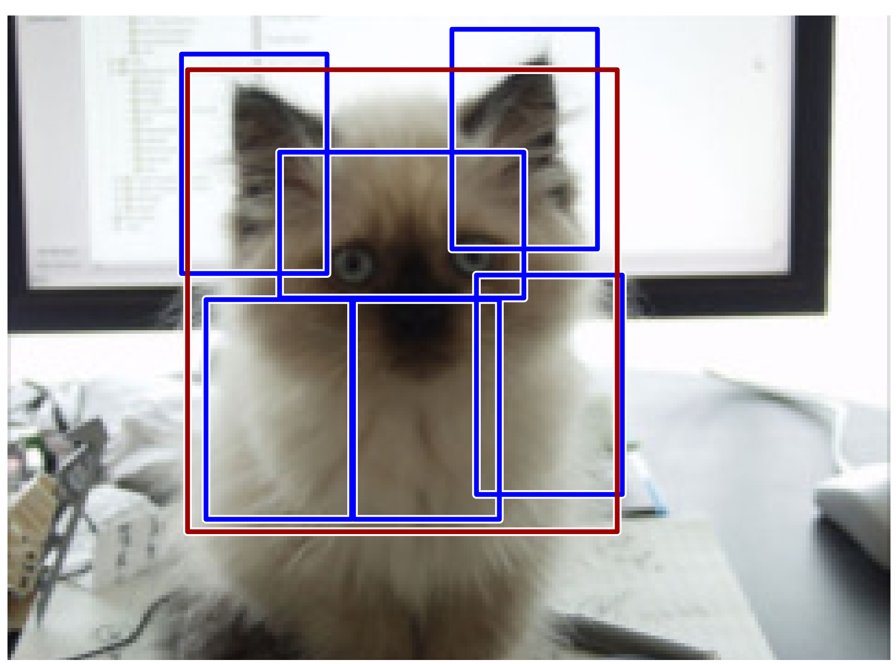


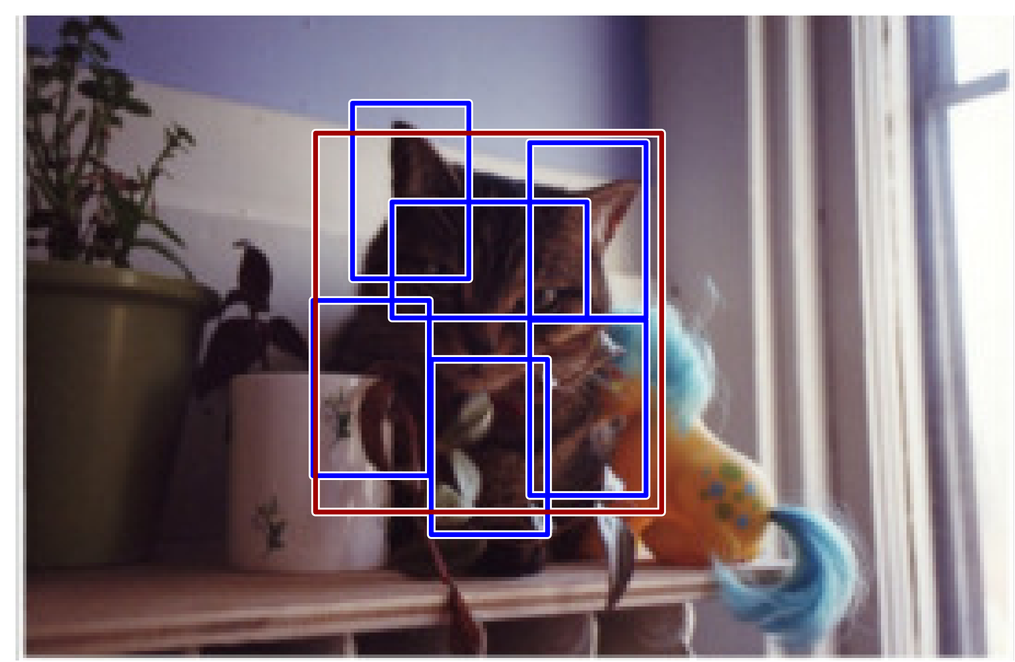
Slide credit: Deva Ramanan

Example object models







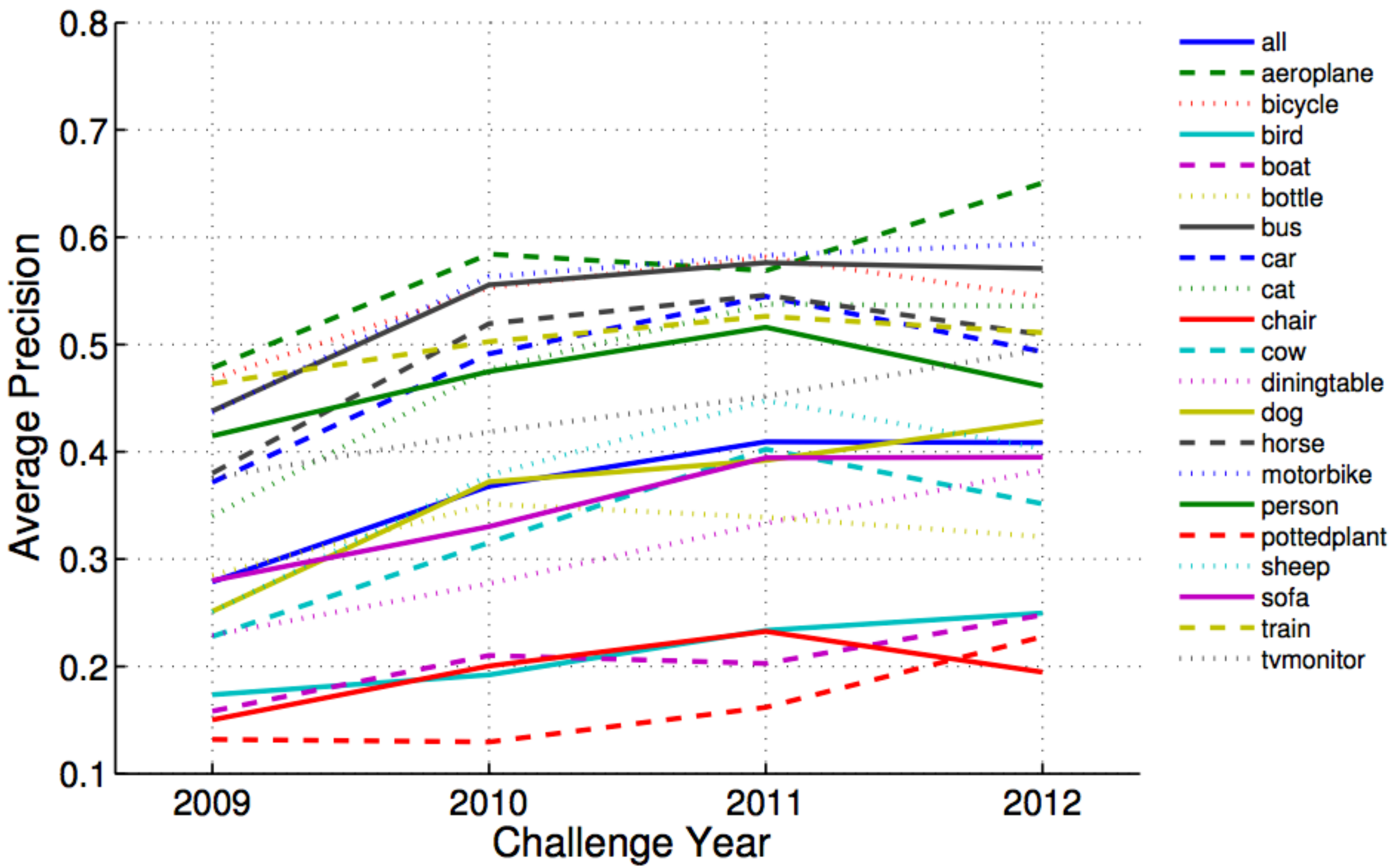


Slide credit: Deva Ramanan

Code

http://www.cs.berkeley.edu/~rbg/latent/index.html

- HOG feature extraction
- DPM training and inference code
- Object detection models for several categories
- …but not state-of-the-art anymore!



Everingham et al.


Chair


Car

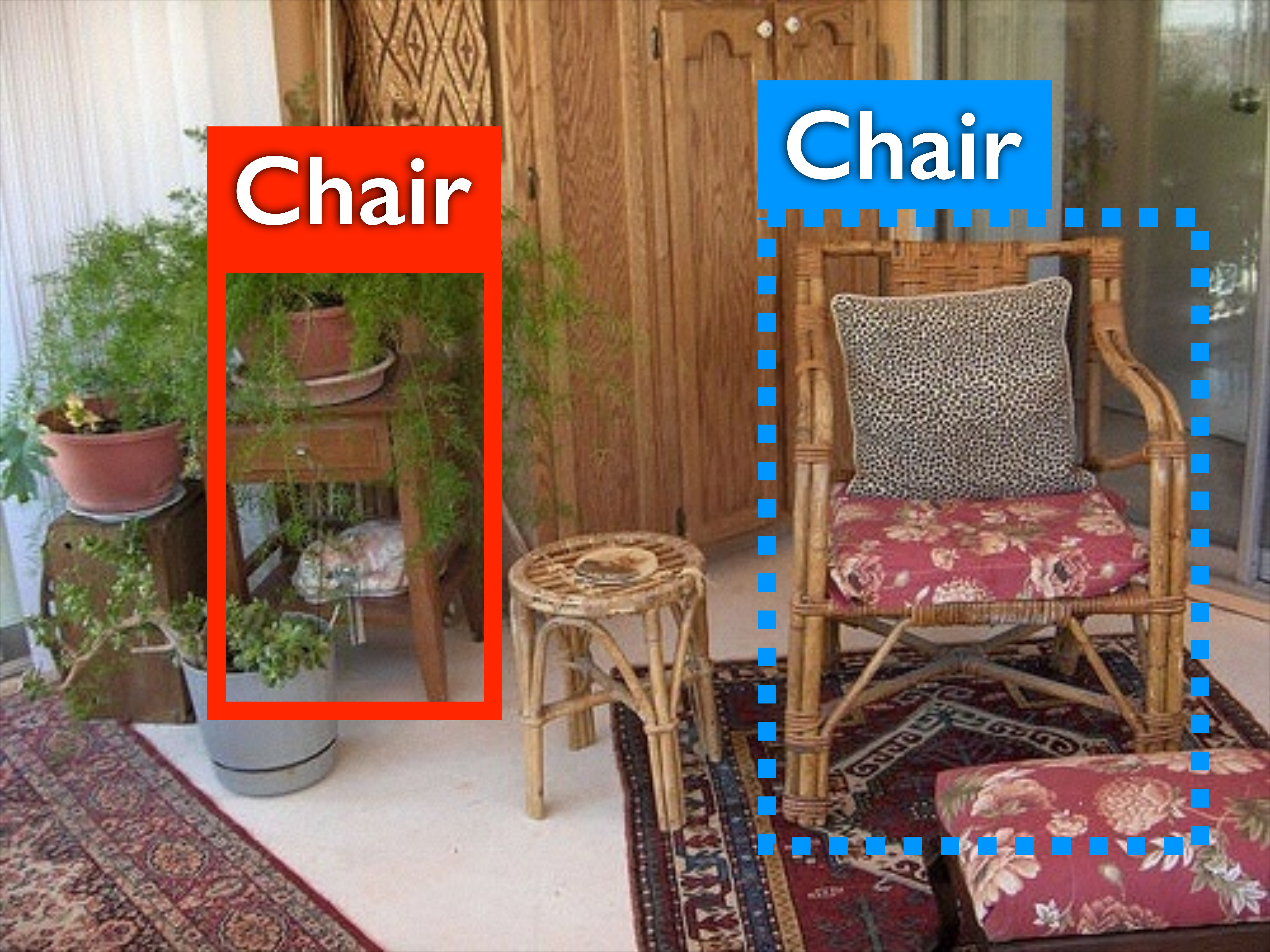

U.S. NAVY


Aeroplane


Aeroplane
Aeroplane


Person


Chair


Chair
Chair


Car

What information does HOG have?

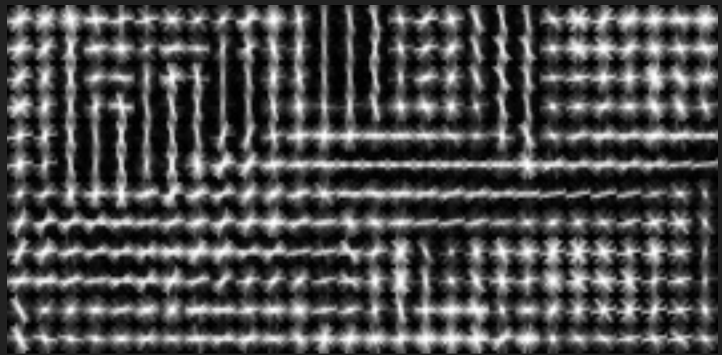
What information does HOG have?

Image

What information does HOG have?

Image

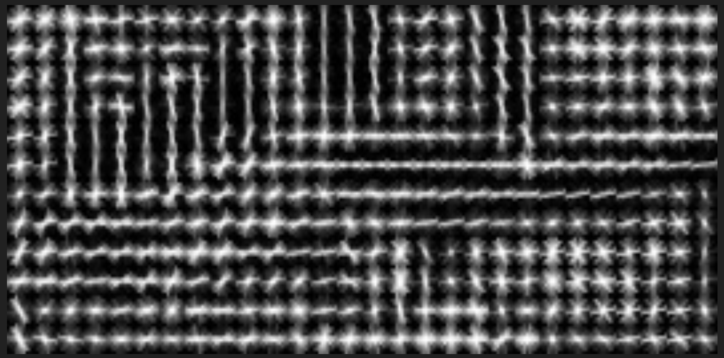
HOG

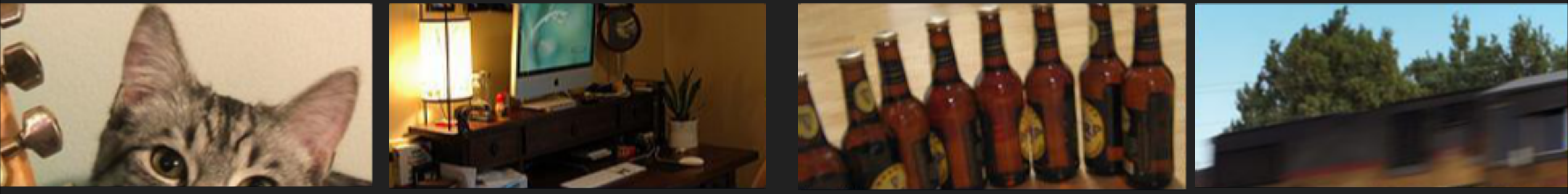


What information does HOG have?

Image

HOG



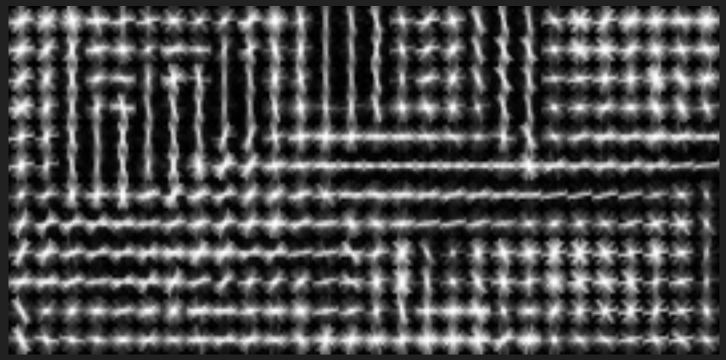


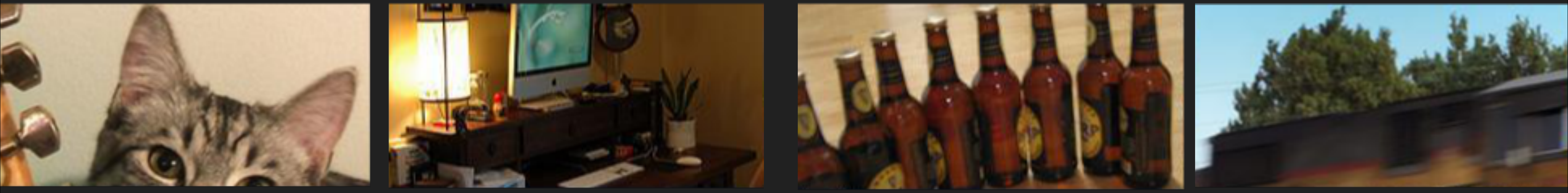
Nearest Neighbors

What information does HOG have?

Image

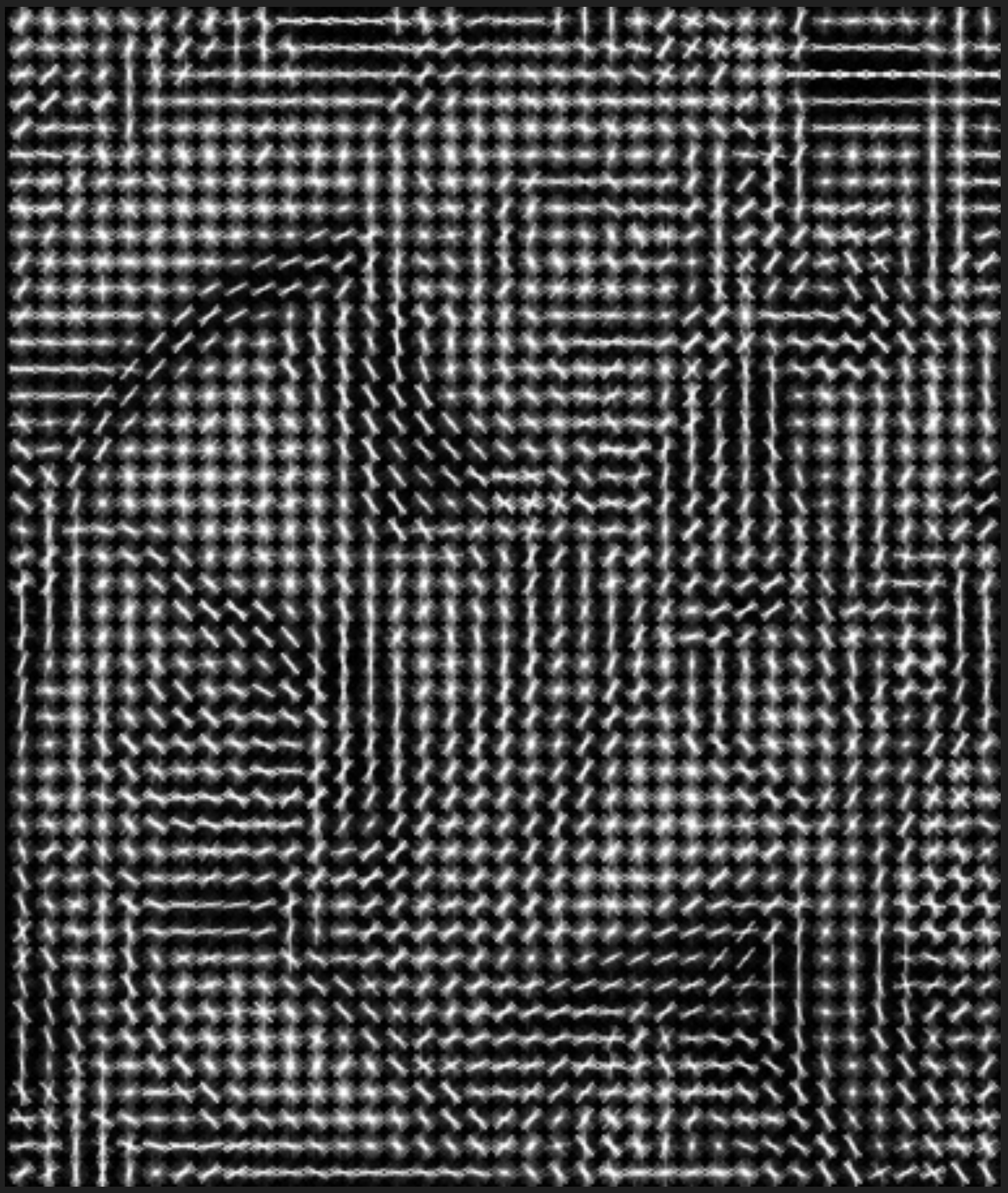
HOG



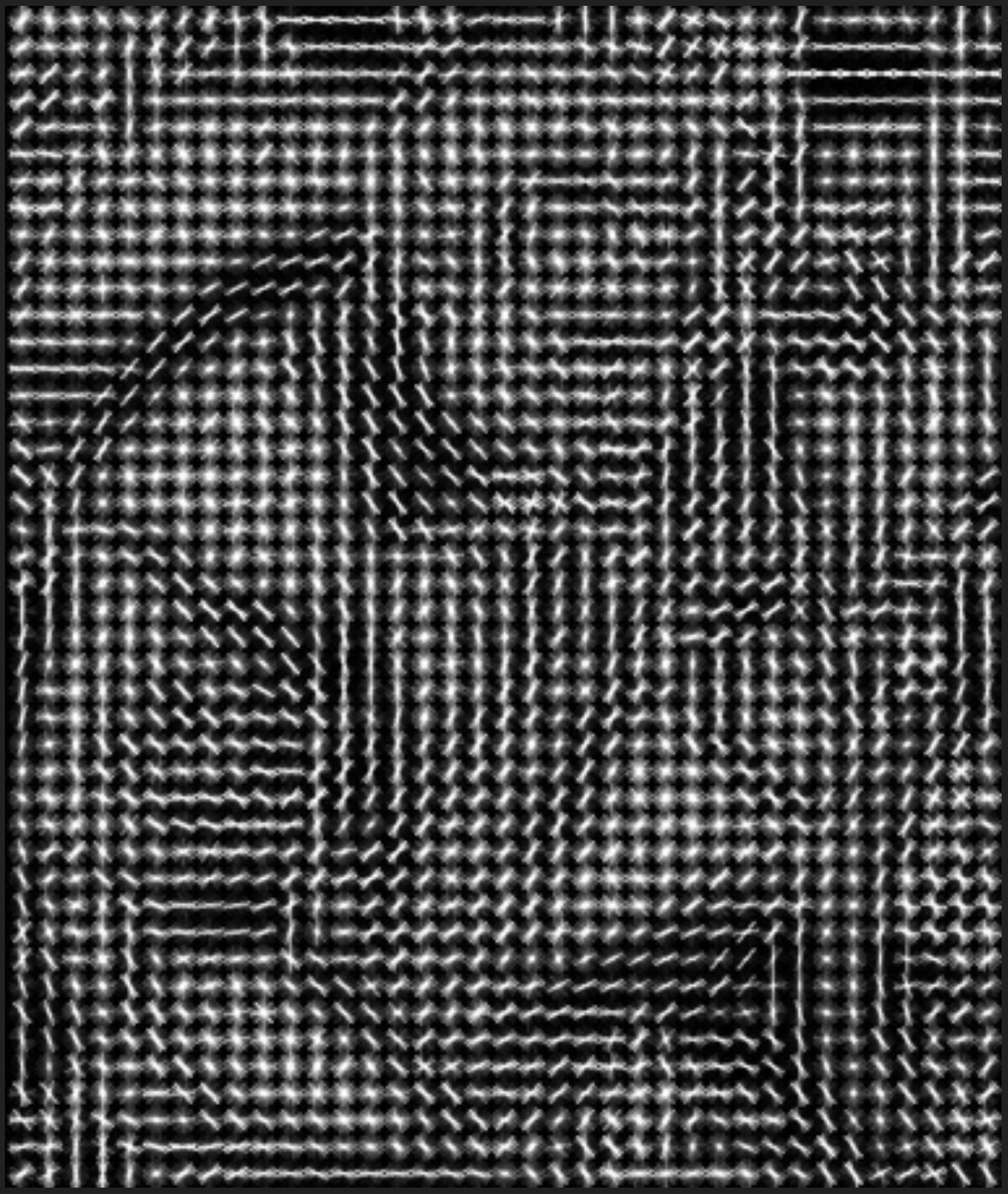


Nearest Neighbors

What information is lost?

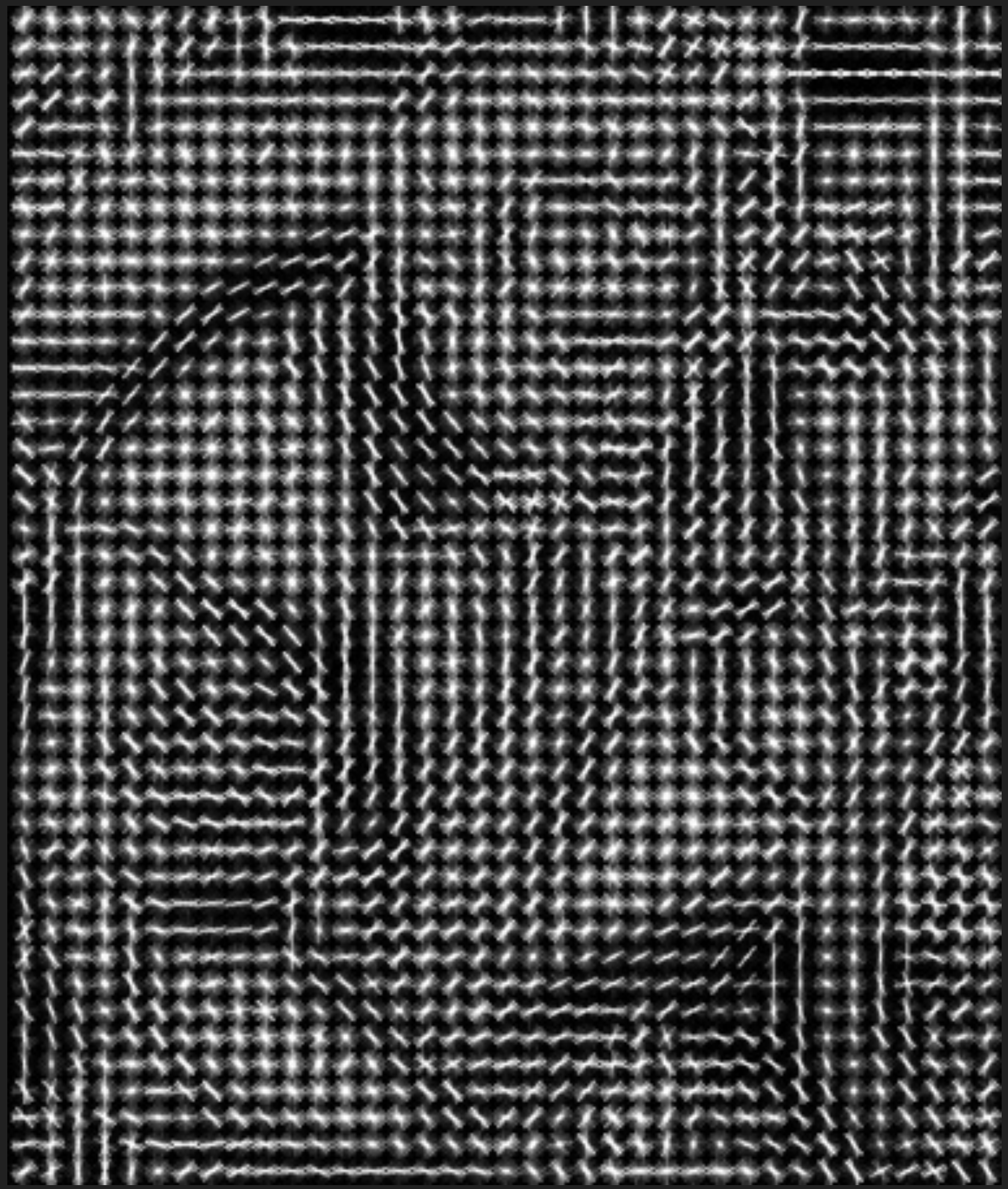


What information is lost?

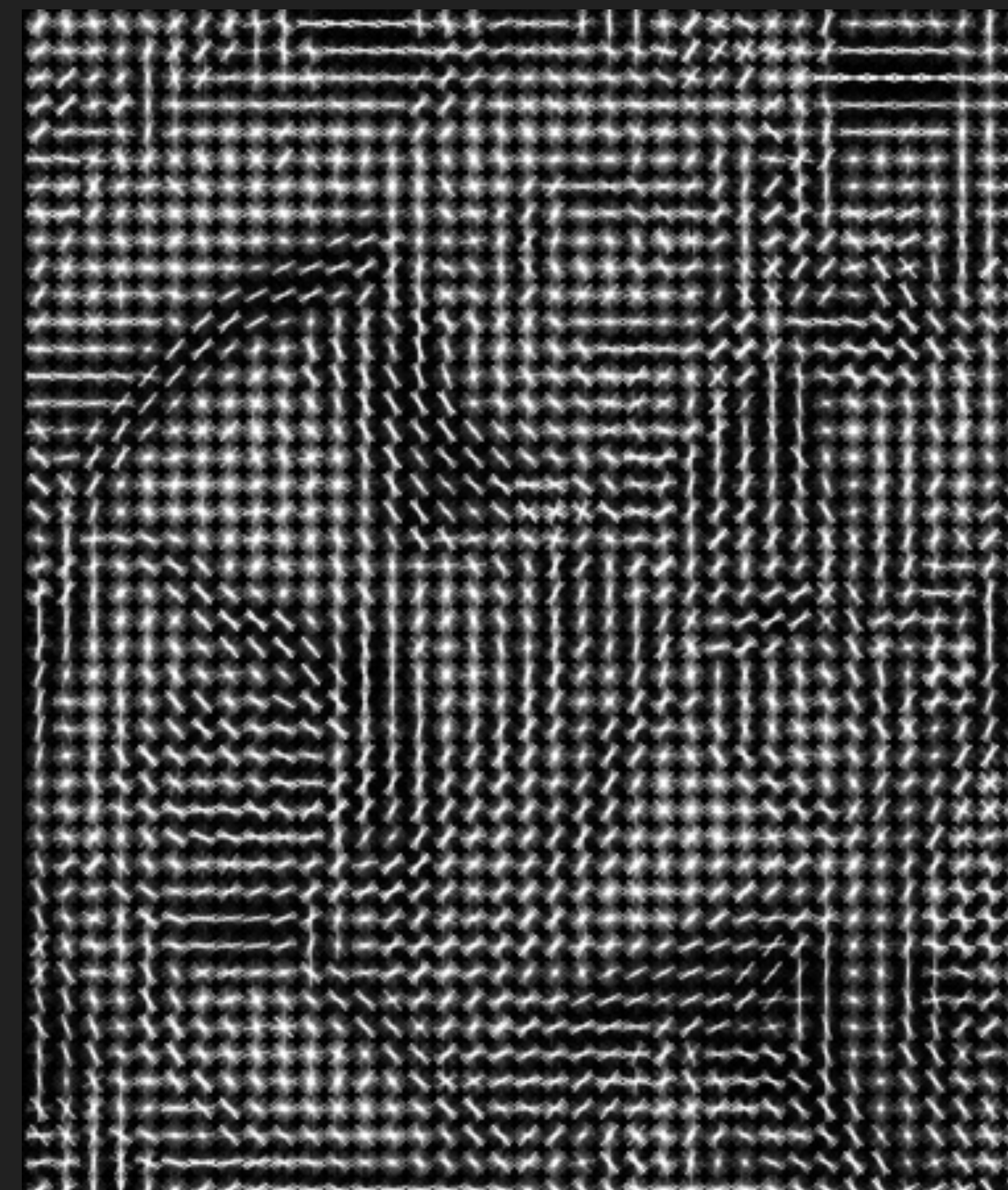


What information is lost?

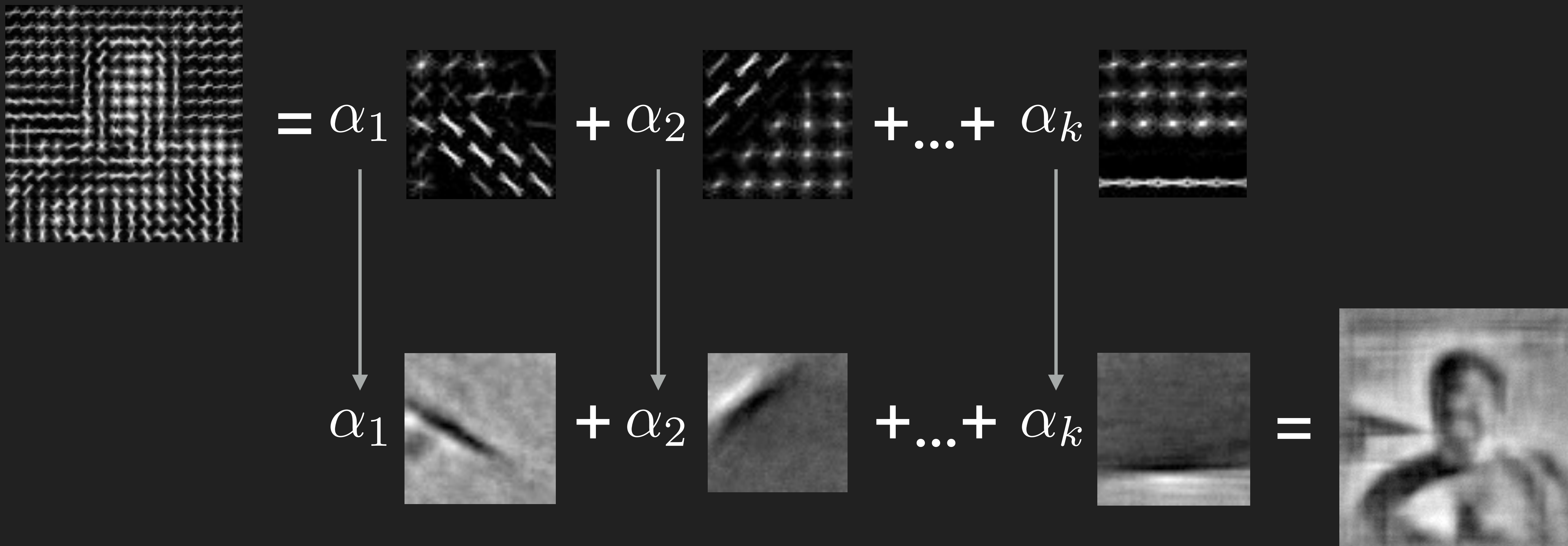
$$\min_{x \in \mathbb{R}^d} ||\phi(x) - y||_2^2$$



What information is lost?

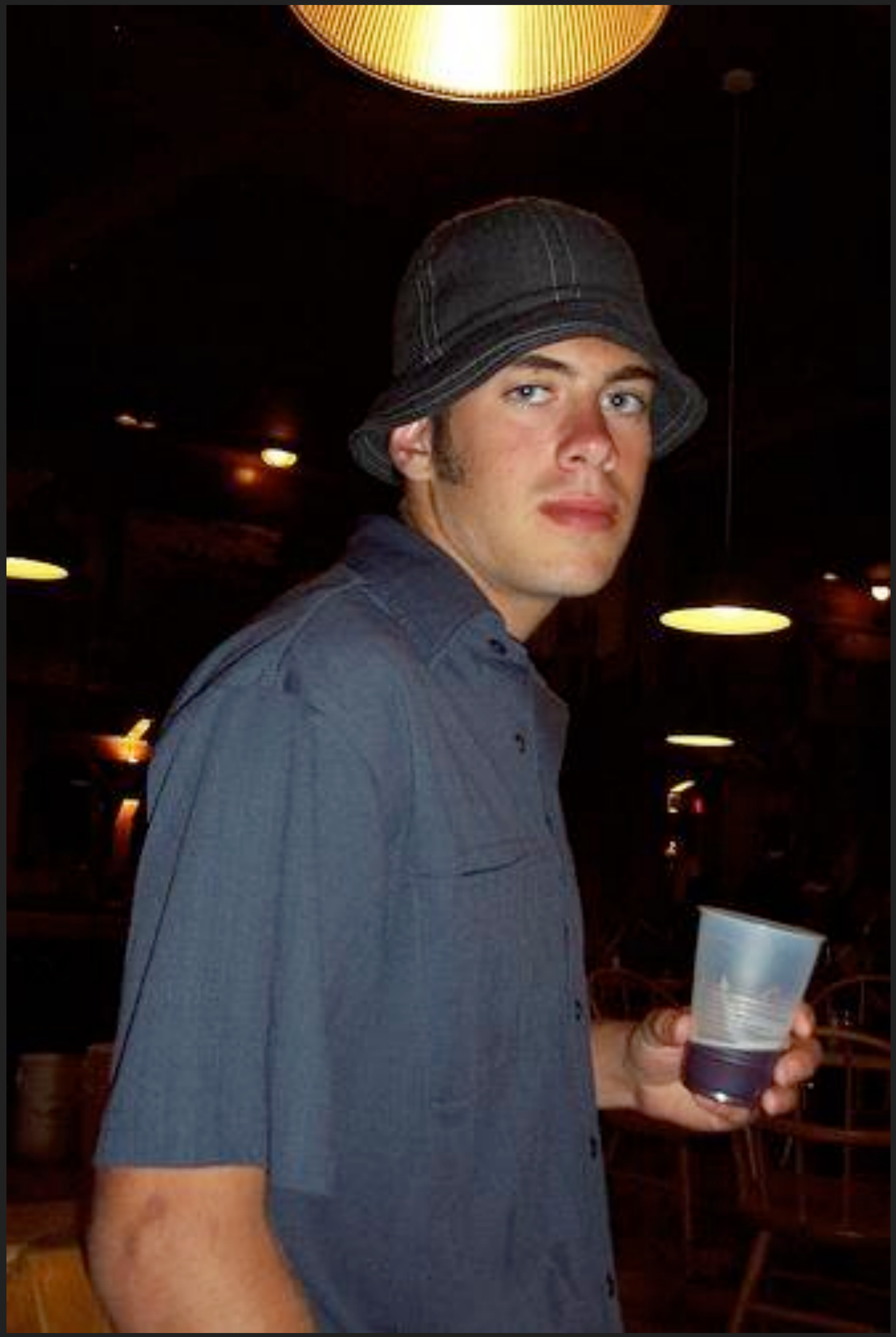


Method: Paired Dictionary



$$\hat{y} = f(x) = V\hat{\alpha}$$

$$\text{where} \quad \hat{\alpha} = \arg\min_{\alpha} ||x - U\alpha||_2^2 \quad \text{s.t.} \quad ||\alpha||_1 \leq \lambda$$



Human Vision

vs

Human Vision

HOG Vision


Car


Image
HOG
SVM

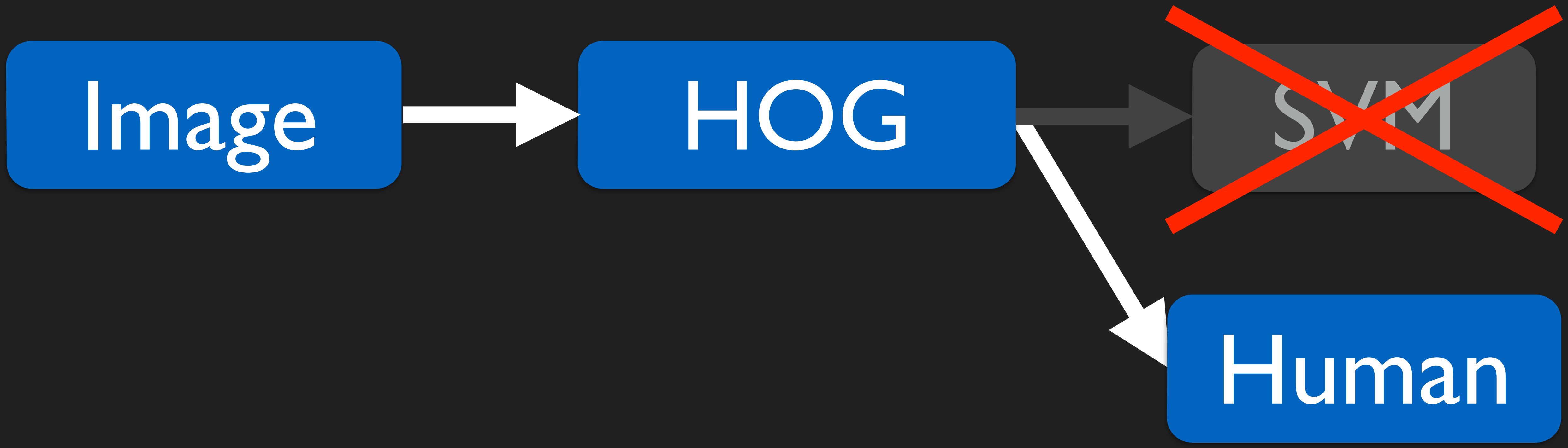
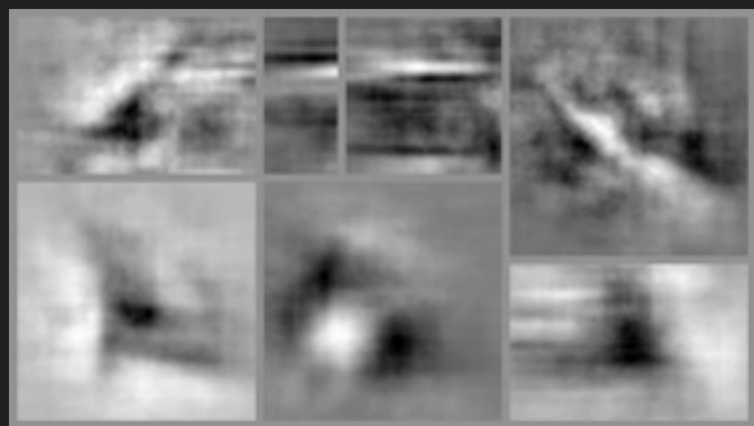


Image
HOG
SVM
Human

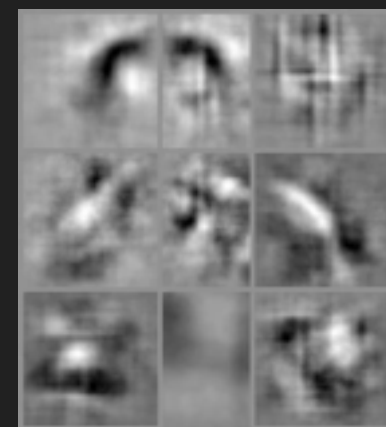
The HOGgles Challenge

Clap your hands when you see a person

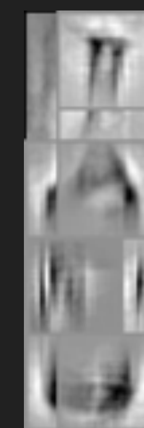
Visualizing Learned Models



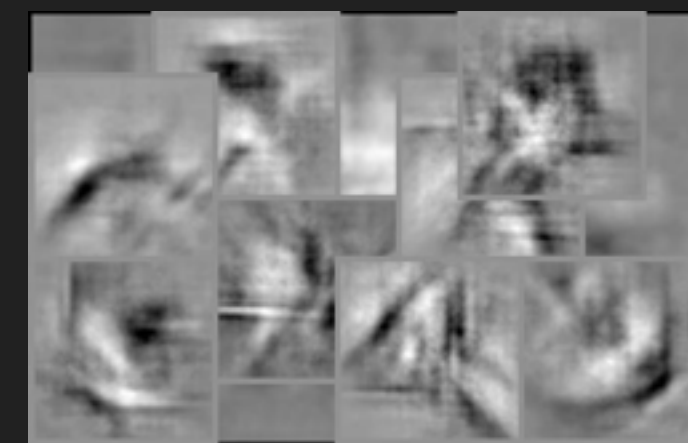
Car



Person



Bottle

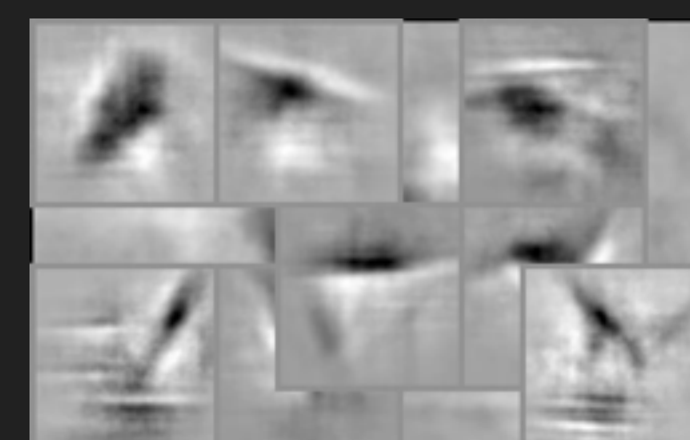


Bicycle

Motorbike

Chair

TV



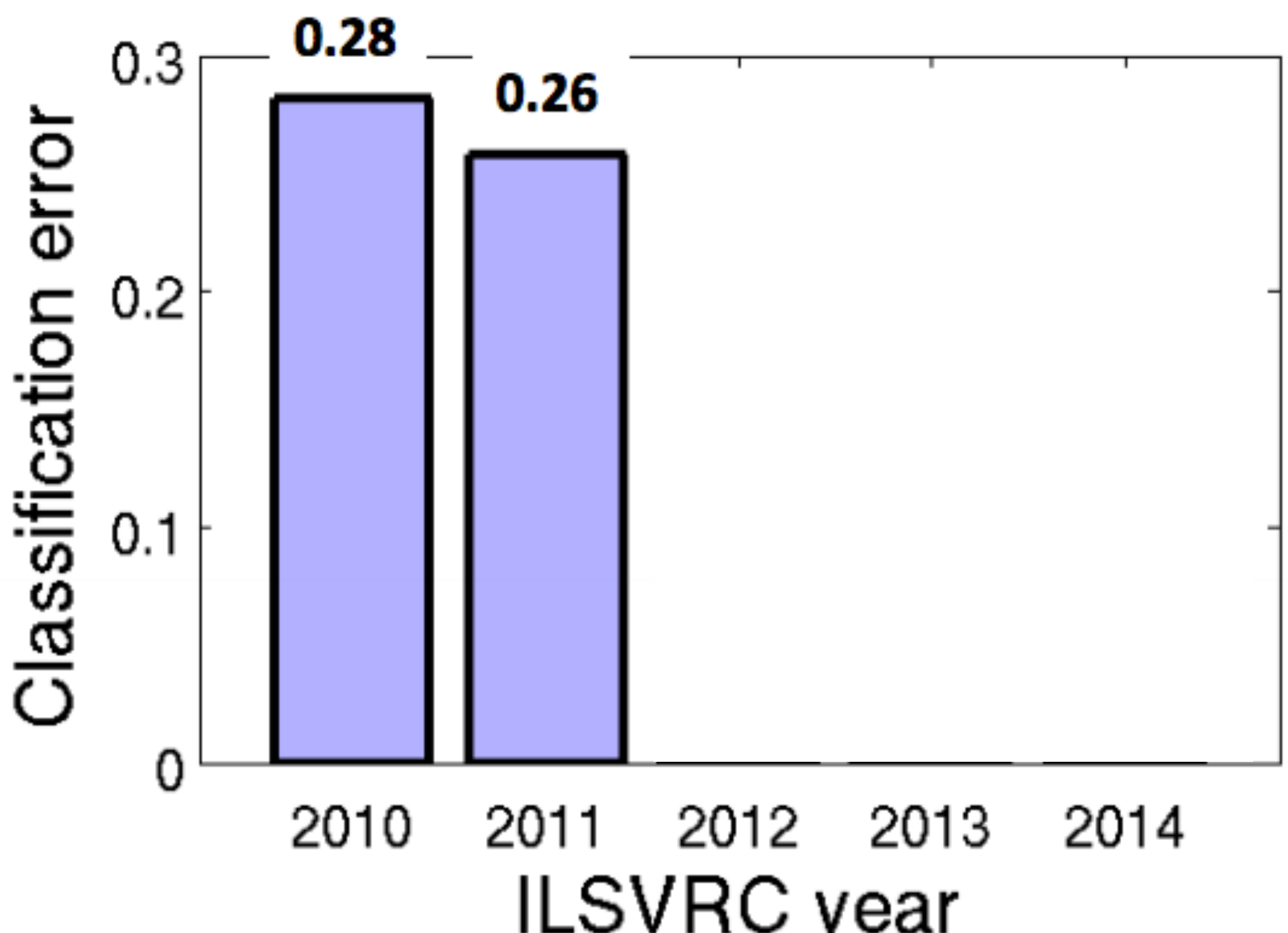
Horse

Why did the detector fail?

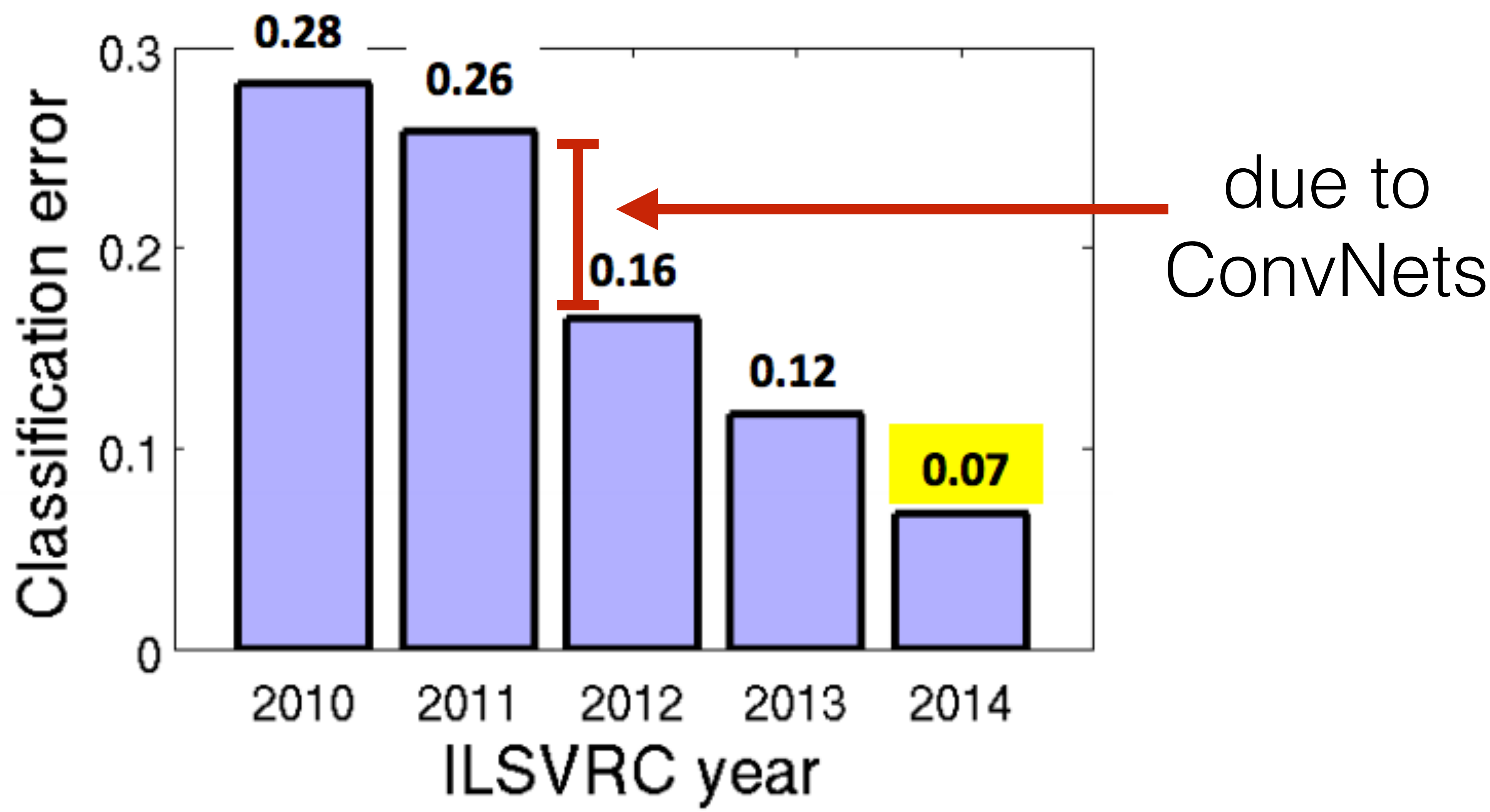
Code

mit.edu/hoggles

- HOG feature extraction
- Code to visualize HOG: `vis = invertHOG(feat)`
- Training code to create your own visualizations



[Russakovsky, et al.]



[Russakovsky, et al.]

Deformable Part Models are Convolutional Neural Networks

Ross Girshick[1] Forrest Iandola[2] Trevor Darrell[2] Jitendra Malik[2]

[1]Microsoft Research [2]UC Berkeley

rbg@microsoft.com {forresti,trevor,malik}@eecs.berkeley.edu