MIT CSAIL

6.869: Advances in Computer Vision

MIT COMPUTER VISION

Lecture 19

Object recognition

Object detection
- Viola-Jones
- Part based models (Hog)
- iHog
- Region based methods (objectness)
- fast RCNN

Vision and language
- word2vec, LSTM
- captioning, seq2seq model
- VQA
- speech + vision- movies

Object segmentation
- CRF- context
- segnet: encoder-decoder-
- mask RCNN

Various topics
- CLEVR
- 3D from single images
- face recognition, AAM
- GANs

Action recognition
- 3d-convNets
- Rgb+motion estimation
- Datasets for activity recognition

Instances vs. categories

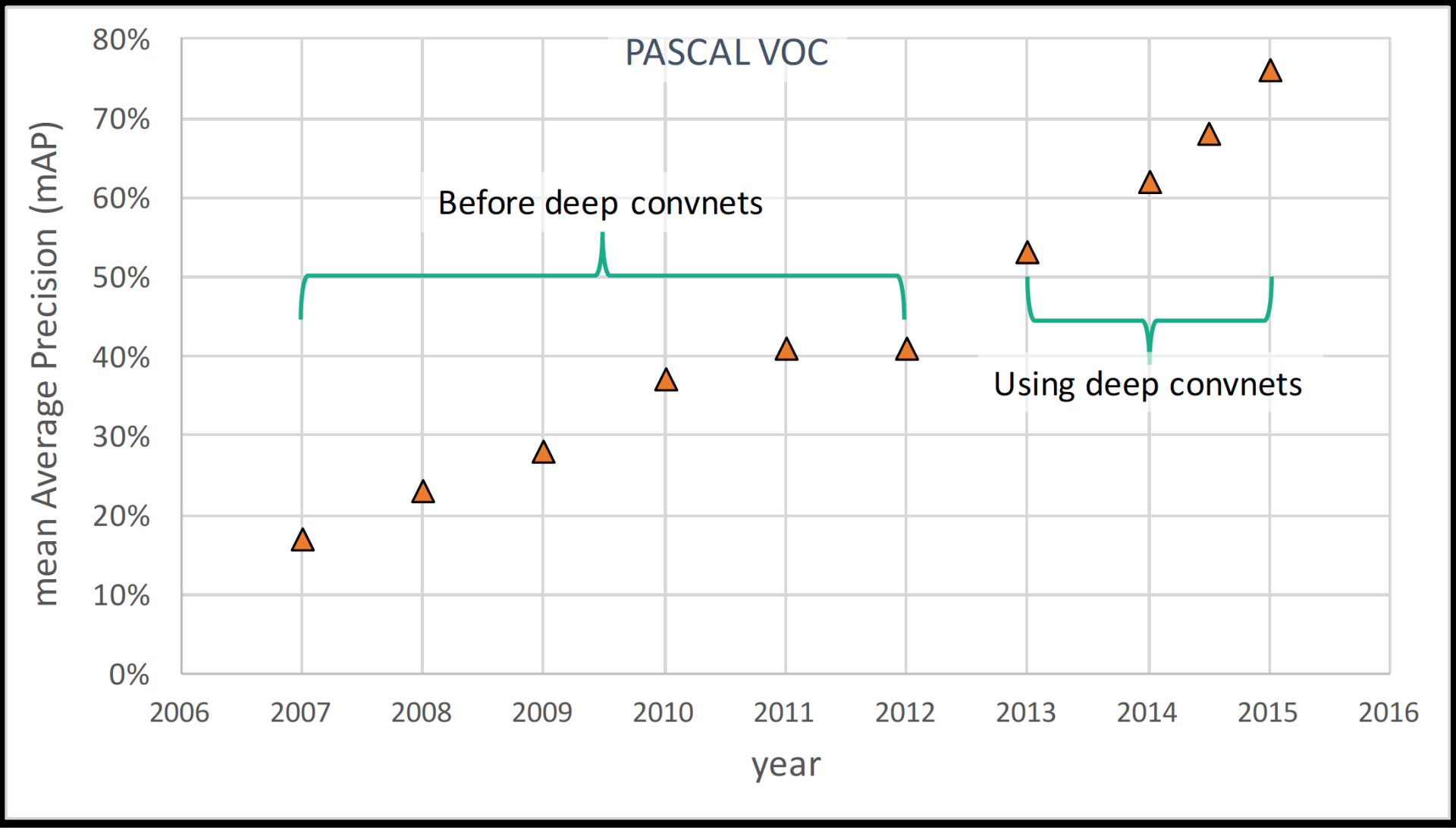
Instances Find these two toys

Can nail it

Categories Find a bottle:

Can't do
unless you do not
care about few errors…

Object detection renaissance (2013-present)



Slide credit: Ross Girshick

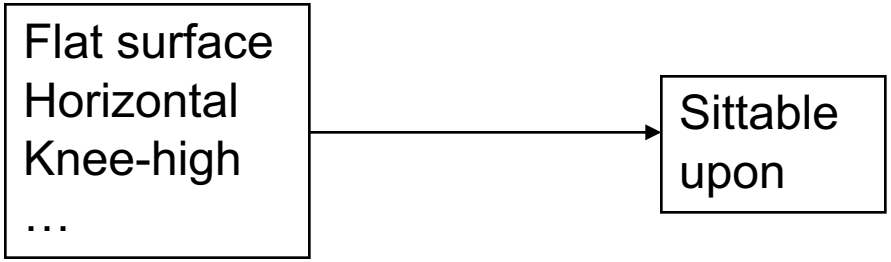
Why do we care about recognition?

Perception of function: We can perceive the 3D shape, texture, material properties, without knowing about objects. But, the concept of category encapsulates also information about what can we do with those objects.

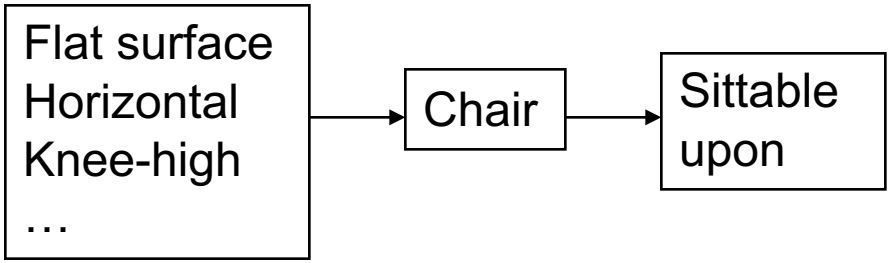
"We therefore include the perception of function as a proper –indeed, crucial- subject for vision science", *from Vision Science, chapter 9, Palmer*.

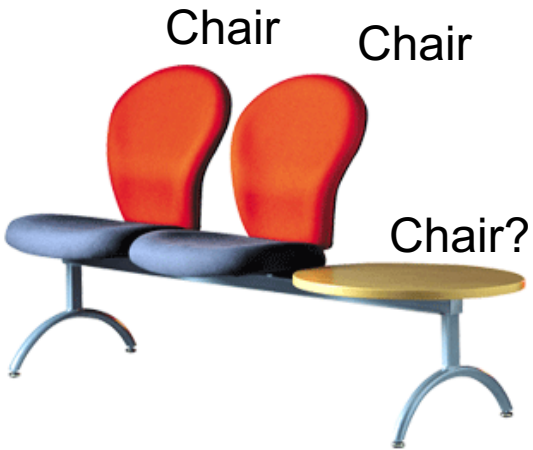
The perception of function

- Direct perception (affordances): Gibson



- Mediated perception (Categorization)

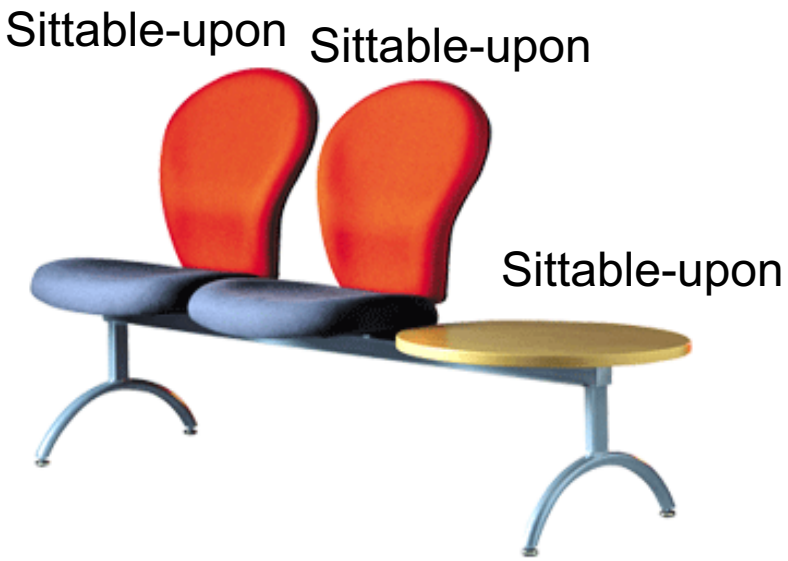




Direct perception

Some aspects of an object function can be perceived directly

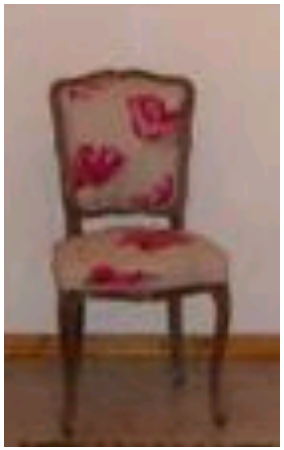
- Functional form: Some forms clearly indicate to a function (“sittable-upon”, container, cutting device, …)



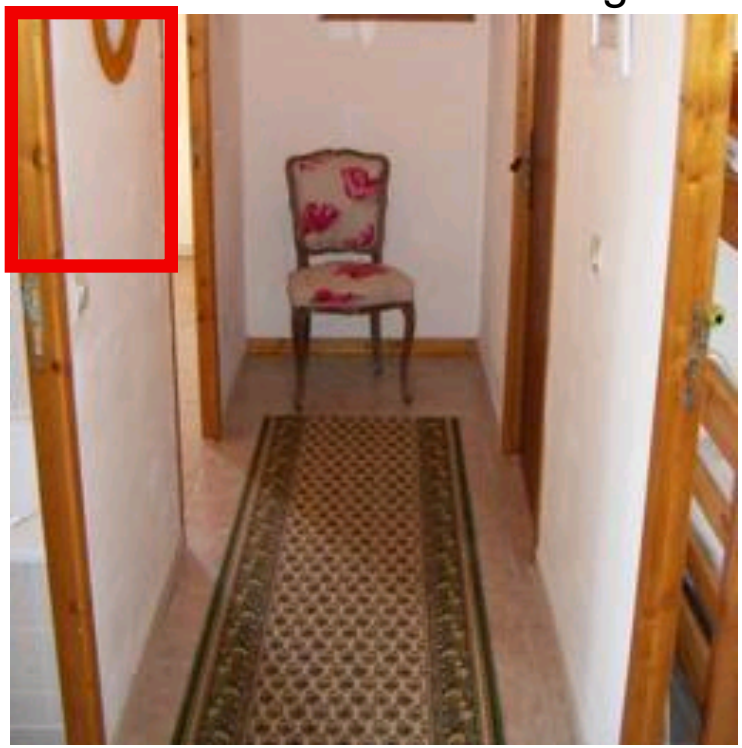
It does not seem easy to sit-upon this…

Object recognition
Is it really so hard?

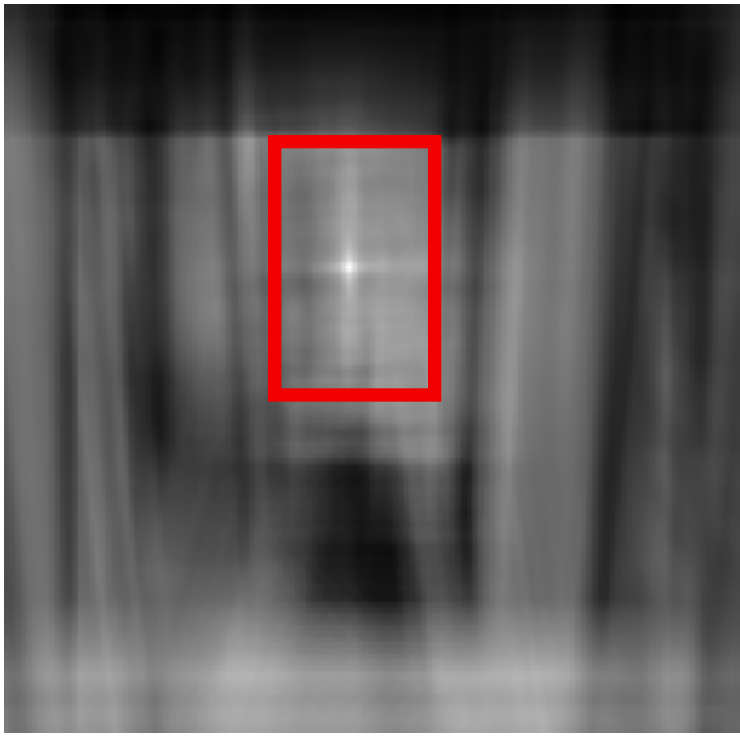
This is a chair

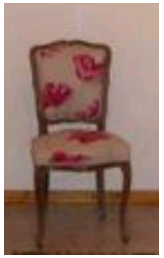


Find the chair in this image



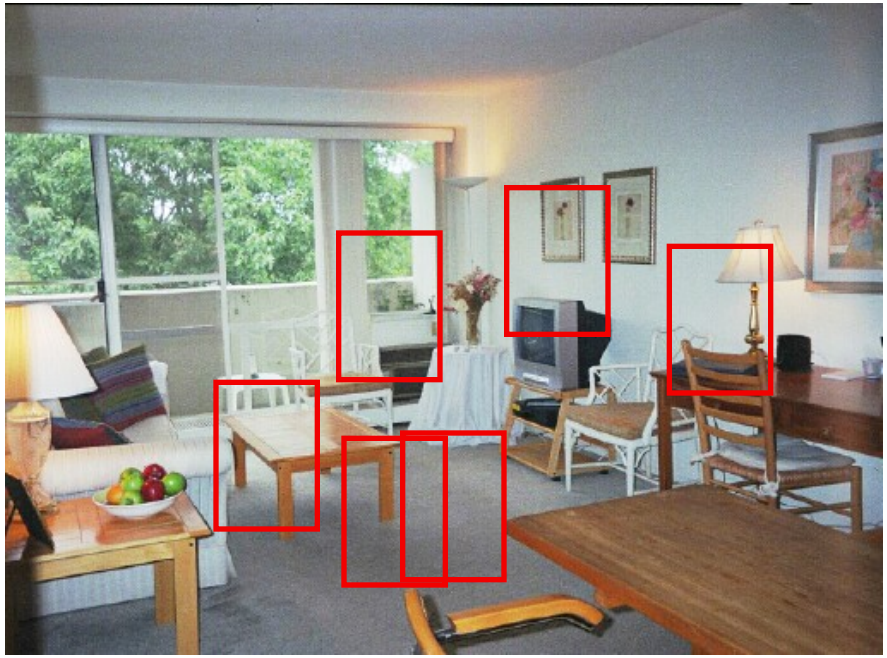
Output of normalized correlation

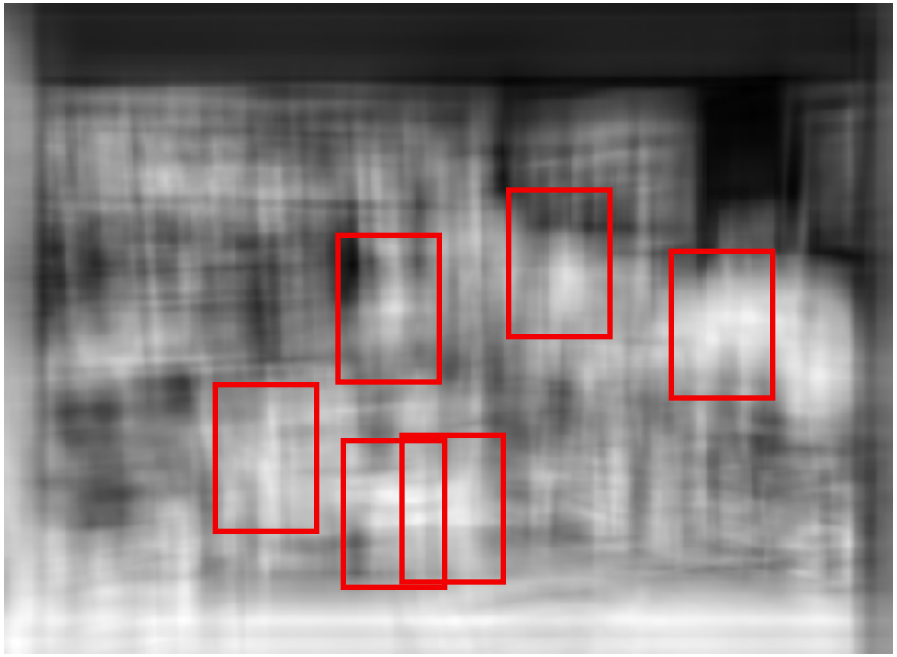




Object recognition
Is it really so hard?

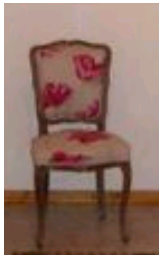
Find the chair in this image





Pretty much garbage
Simple template matching is not going to make it

My biggest concern while making this slide was:
how do I justify 50 years of research, and this course, if this experiment did work?



Object recognition
Is it really so hard?

Find the chair in this image

A “popular method is that of template matching, by point to point correlation of a model pattern with the image pattern. These techniques are inadequate for three-dimensional scene analysis for many reasons, such as occlusion, changes in viewing angle, and articulation of parts.” Nivatia & Binford, 1977.

A bit of history…

So, let's make the problem simpler: Block world

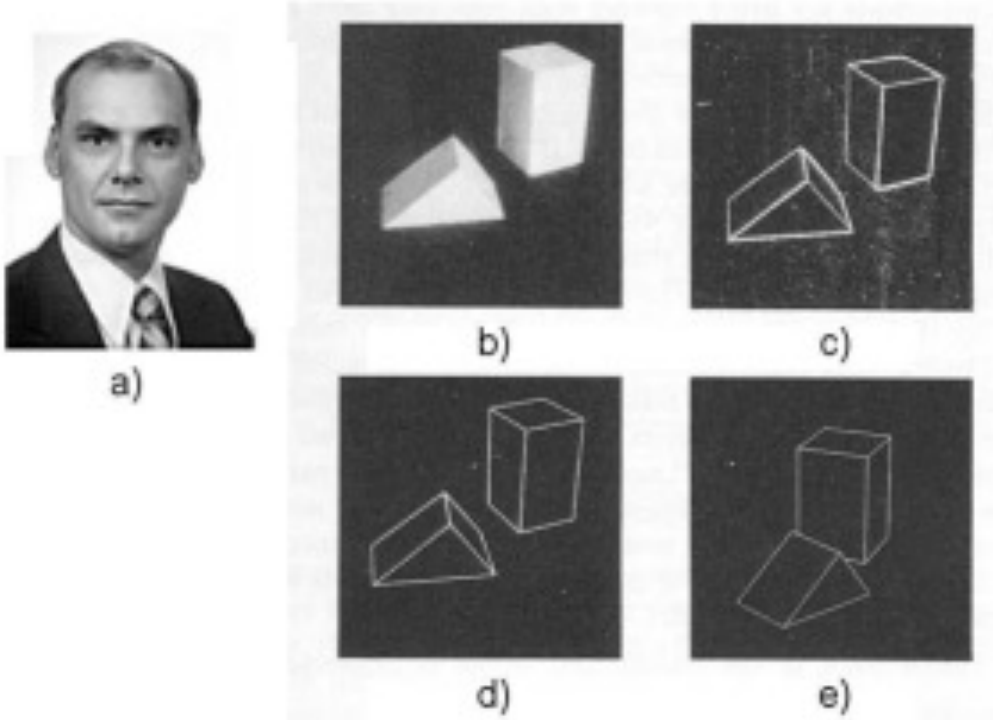


Fig. 1. A system for recognizing 3-d polyhedral scenes. a) L.G. Roberts. b)A blocks world scene. c)Detected edges using a 2x2 gradient operator. d) A 3-d polyhedral description of the scene, formed automatically from the single image. e) The 3-d scene displayed with a viewpoint different from the original image to demonstrate its accuracy and completeness. (b) - e) are taken from [64] with permission MIT Press.)

Nice framework to develop fancy math, but too far from reality…

Object Recognition in the Geometric Era: a Retrospective. Joseph L. Mundy. 2006

Binford and generalized cylinders

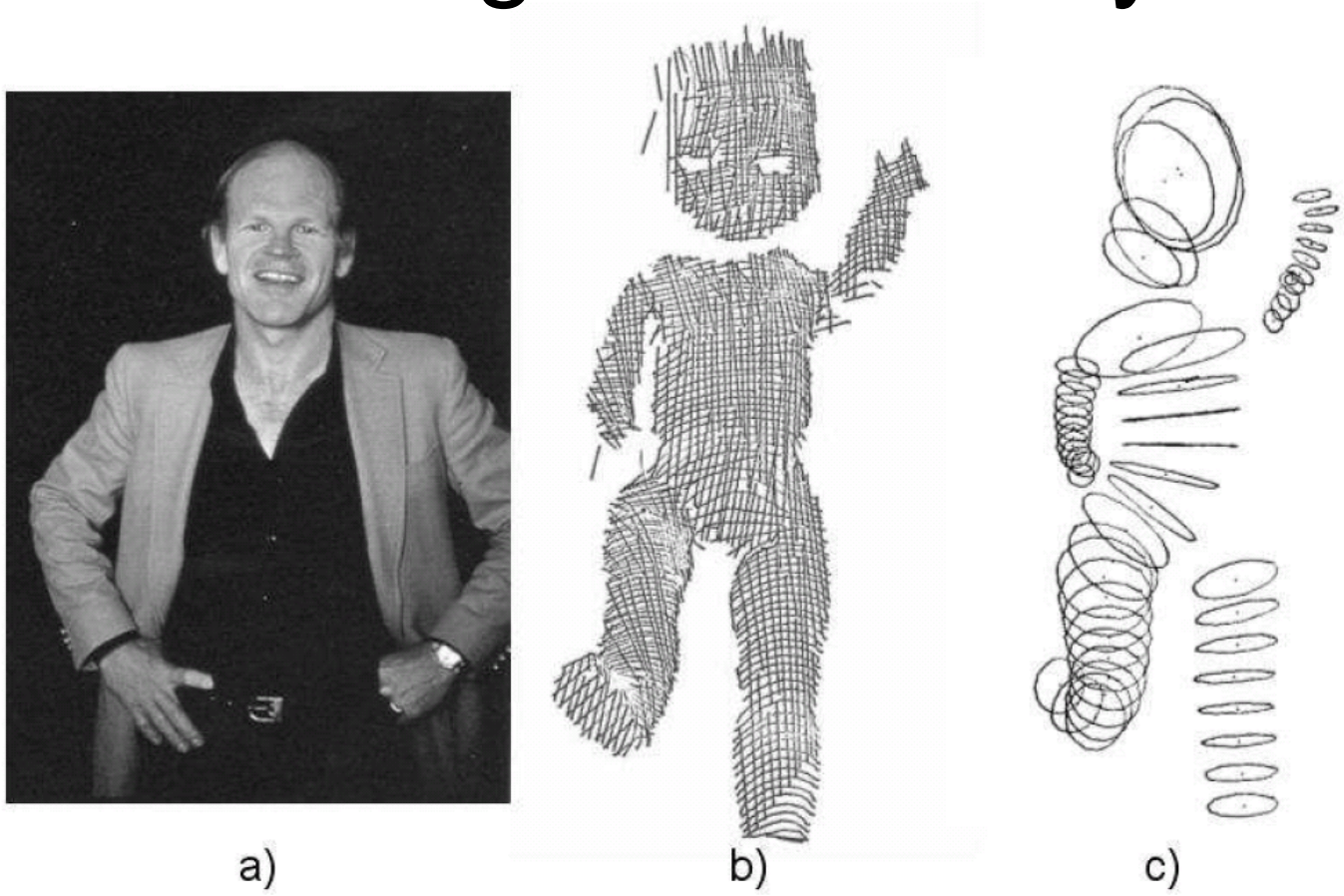
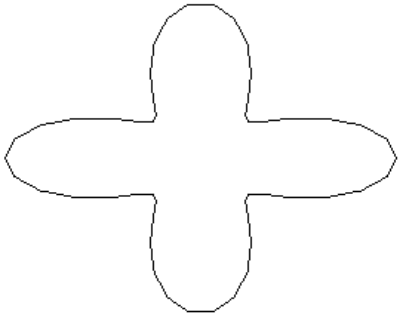


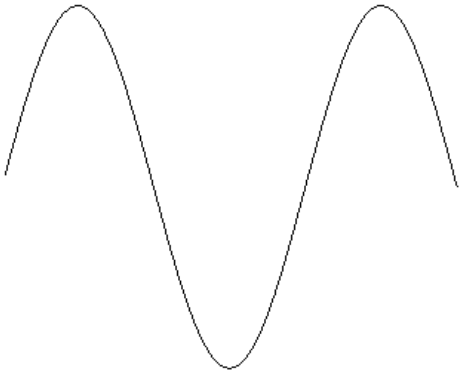
Fig. 3. The representation of objects by assemblies of generalized cylinders. a) Thomas Binford. b) A range image of a doll. c) The resulting set of generalized cylinders. (b) and c) are taken from Agin [1] with permission.)

Object Recognition in the Geometric Era: a Retrospective. Joseph L. Mundy. 2006

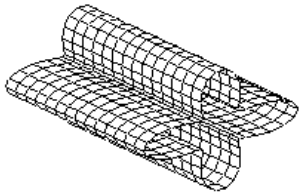
Binford and generalized cylinders



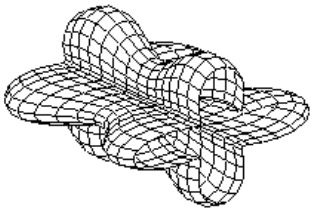
(a) Cross section.



(b) Sweeping rule.



(c) True cylinder

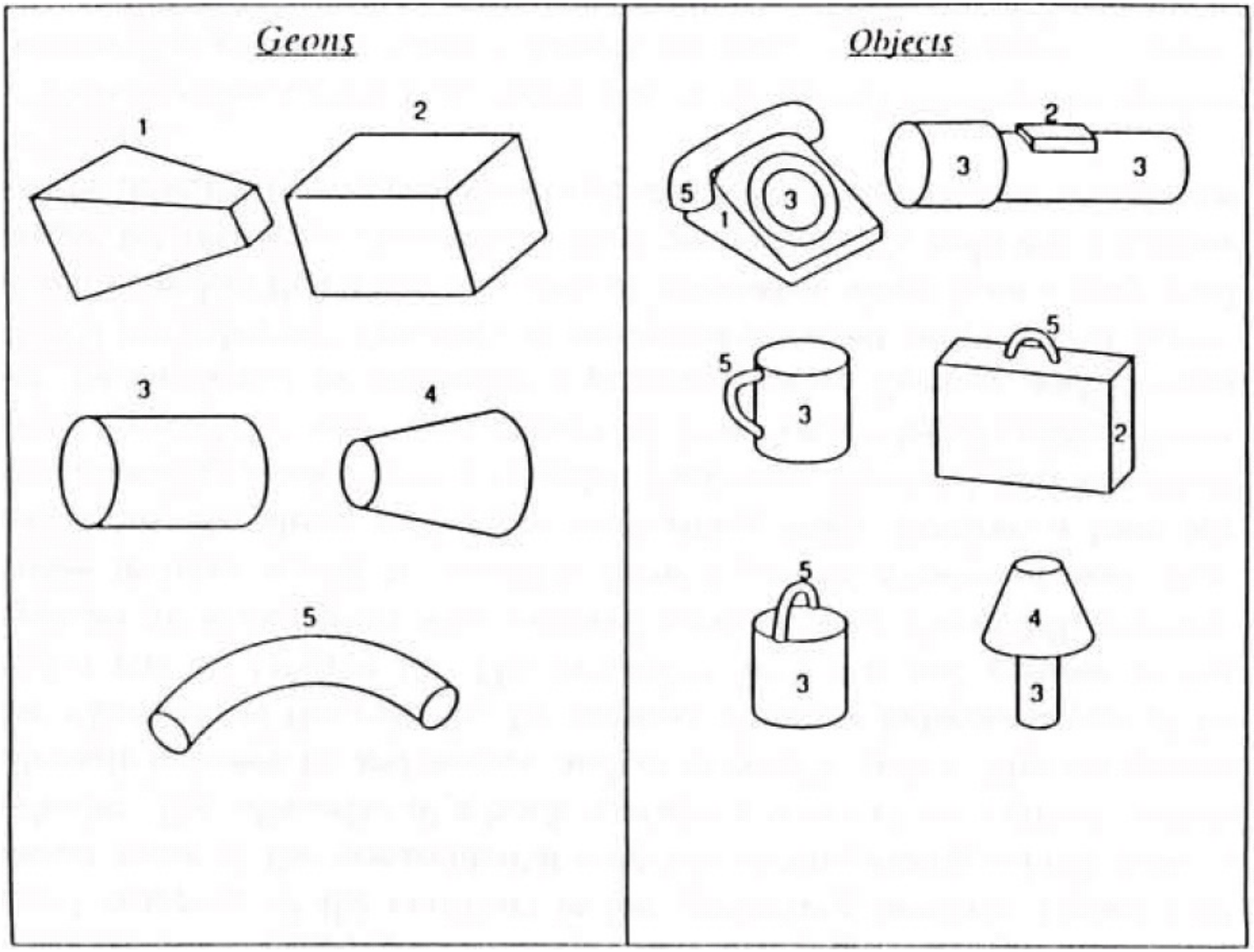


(d) Generalized cylinder

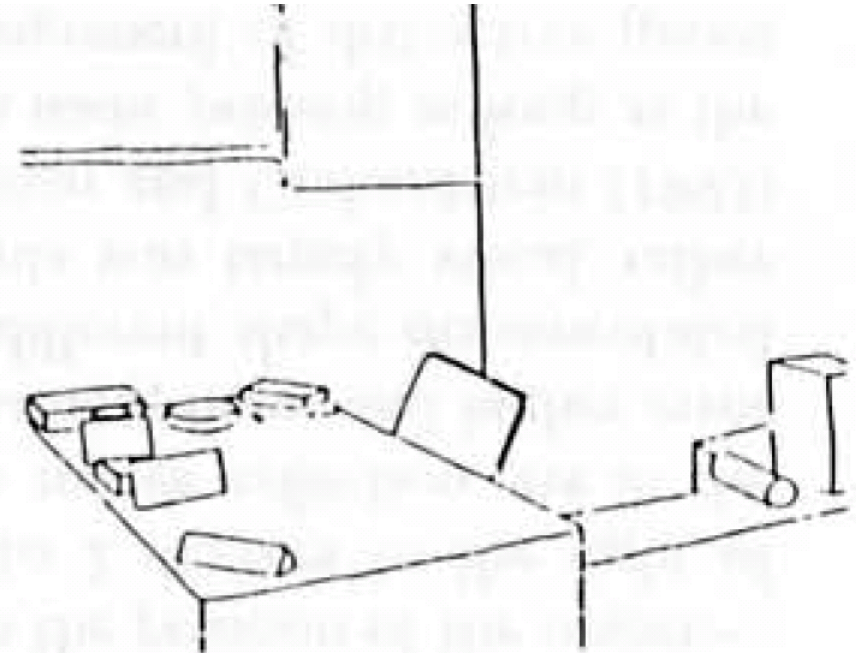
Recognition by components

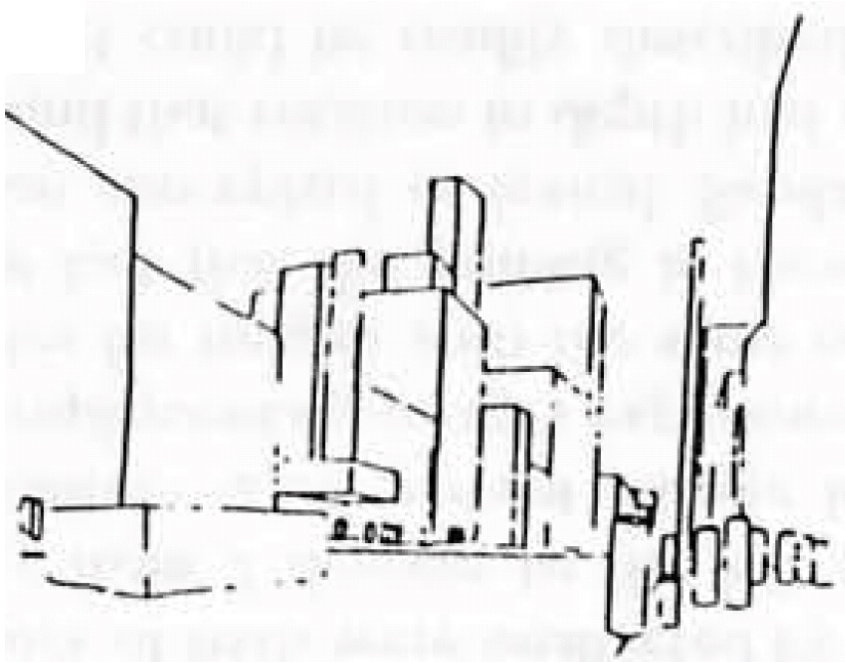
Irving Biederman
Recognition-by-Components: A Theory of Human Image Understanding.
Psychological Review, 1987.

Objects and their geons



Scenes and geons





Mezzanotte & Biederman

What is missing?

The notion of geometric structure.

Although they were aware of it, the previous works put more emphasis on defining the primitive elements than modeling their geometric relationships.

Parts and Structure approaches

With a different perspective, these models focused more on the geometry than on defining the constituent elements:

- Fischler & Elschlager 1973
- Yuille ‘91
- Brunelli & Poggio ‘93
- Lades, v.d. Malsburg et al. ‘93
- Cootes, Lanitis, Taylor et al. ‘95
- Amit & Geman ‘95, ‘99
- Perona et al. ‘95, ‘96, ’98, ’00, ’03, ‘04
- Felzenszwalb & Huttenlocher ’00, ’04
- Crandall & Huttenlocher ’05, ’06
- Leibe & Schiele ’03, ’04
- Many papers since 2000

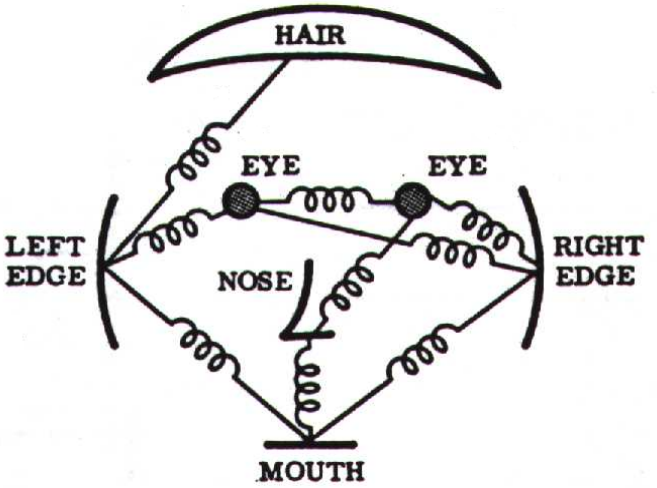


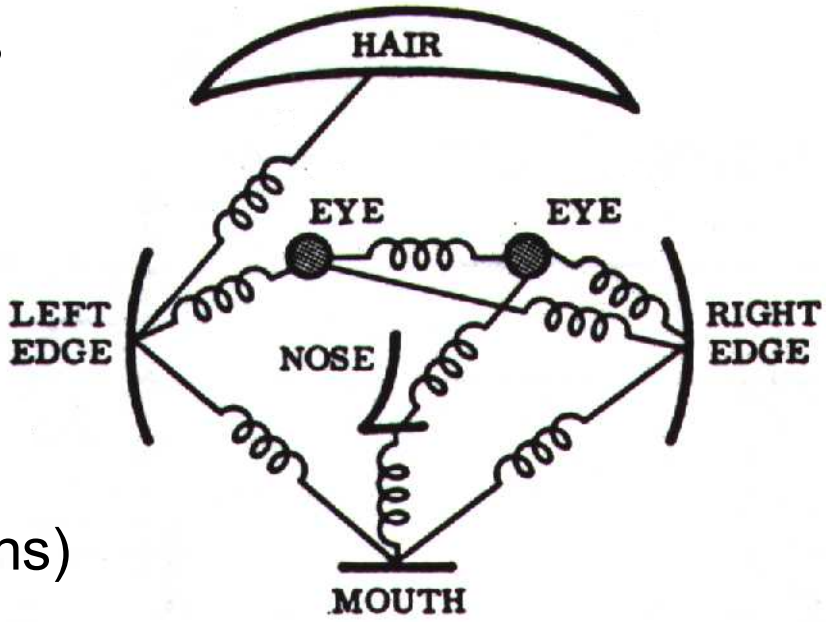
Figure from [Fischler & Elschlager 73]

Representation

- Object as set of parts
  - Generative representation

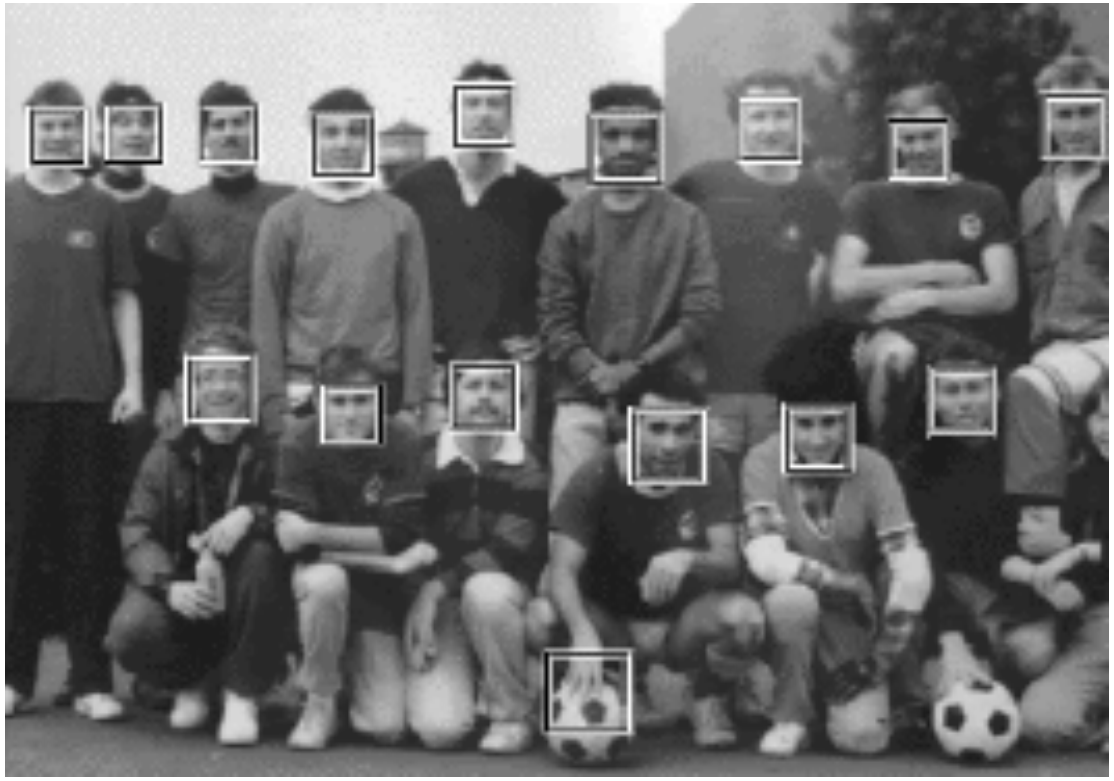
- Model:
  - Relative locations between parts
  - Appearance of part

- Issues:
  - How to model location
  - How to represent appearance
  - Sparse or dense (pixels or regions)
  - How to handle occlusion/clutter



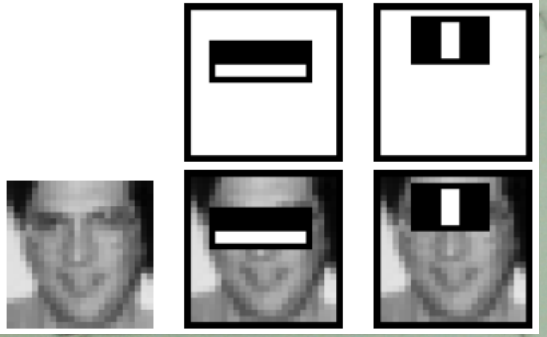
We will discuss these models more in depth later

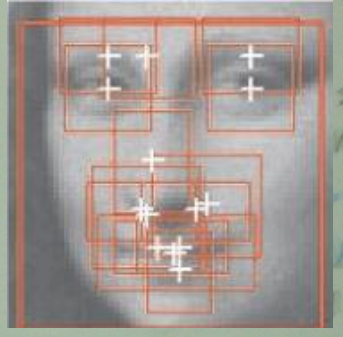
Face detection and the success of learning based approaches

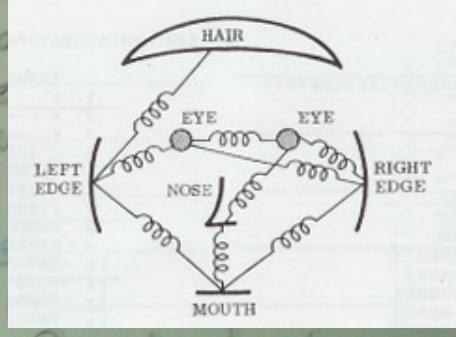


- The representation and matching of pictorial structures Fischler, Elschlager (1973).
- Face recognition using eigenfaces M. Turk and A. Pentland (1991).
- Human Face Detection in Visual Scenes - Rowley, Baluja, Kanade (1995)
- Graded Learning for Object Detection - Fleuret, Geman (1999)
- Robust Real-time Object Detection - Viola, Jones (2001)
- Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images - Heisele, Serre, Mukherjee, Poggio (2001)
- ….

Face detection



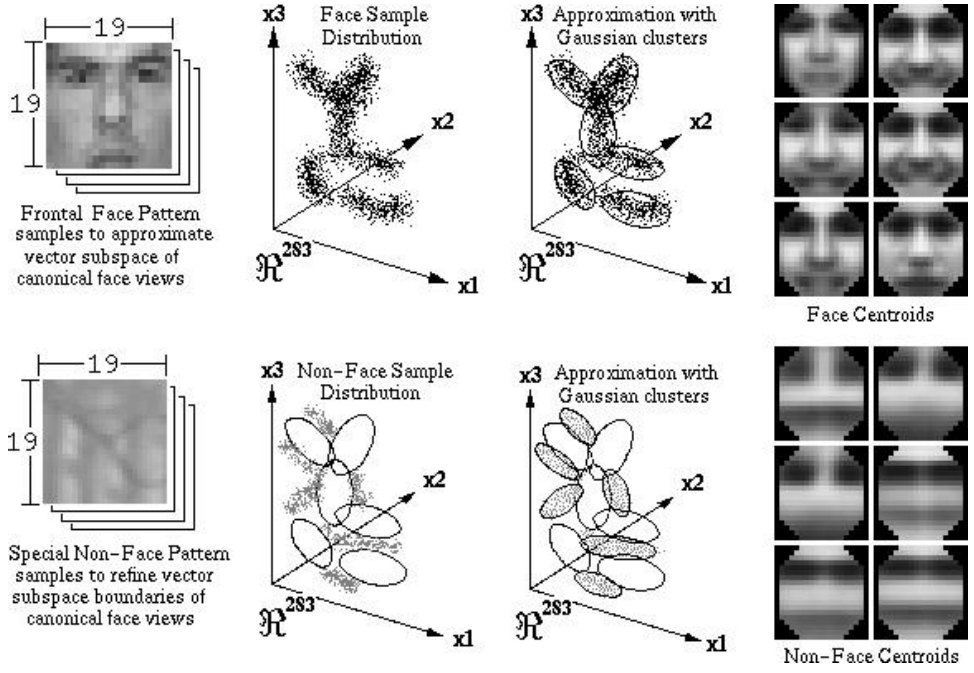


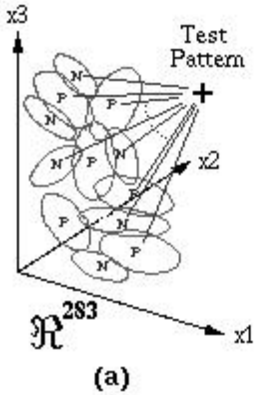


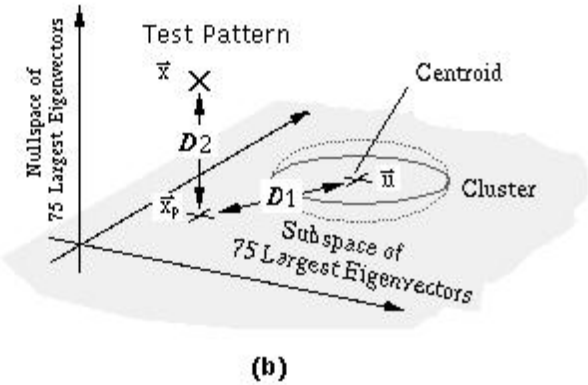
- The representation and matching of pictorial structures Fischler, Elschlager (1973)
- Face recognition using eigenfaces M. Turk and A. Pentland (1991).
- Human Face Detection in Visual Scenes - Rowley, Baluja, Kanade (1995)
- Graded Learning for Object Detection - Fleuret, Geman (1999)
- Robust Real-time Object Detection - Viola, Jones (2001)
- Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images - Heisele, Serre, Mukherjee, Poggio (2001)
- ….

Distribution-Based Face Detector

- Learn face and nonface models from examples [Sung and Poggio 95]
- Cluster and project the examples to a lower dimensional space using Gaussian distributions and PCA
- Detect faces using distance metric to face and nonface clusters







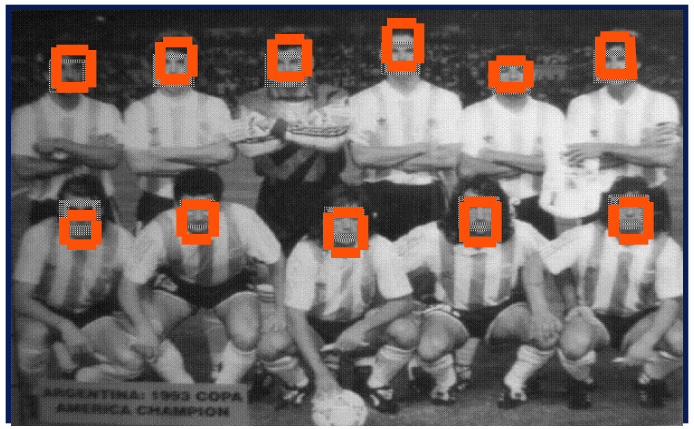
Distribution-Based Face Detector

- Learn face and nonface models from examples [Sung and Poggio 95]

Training Database

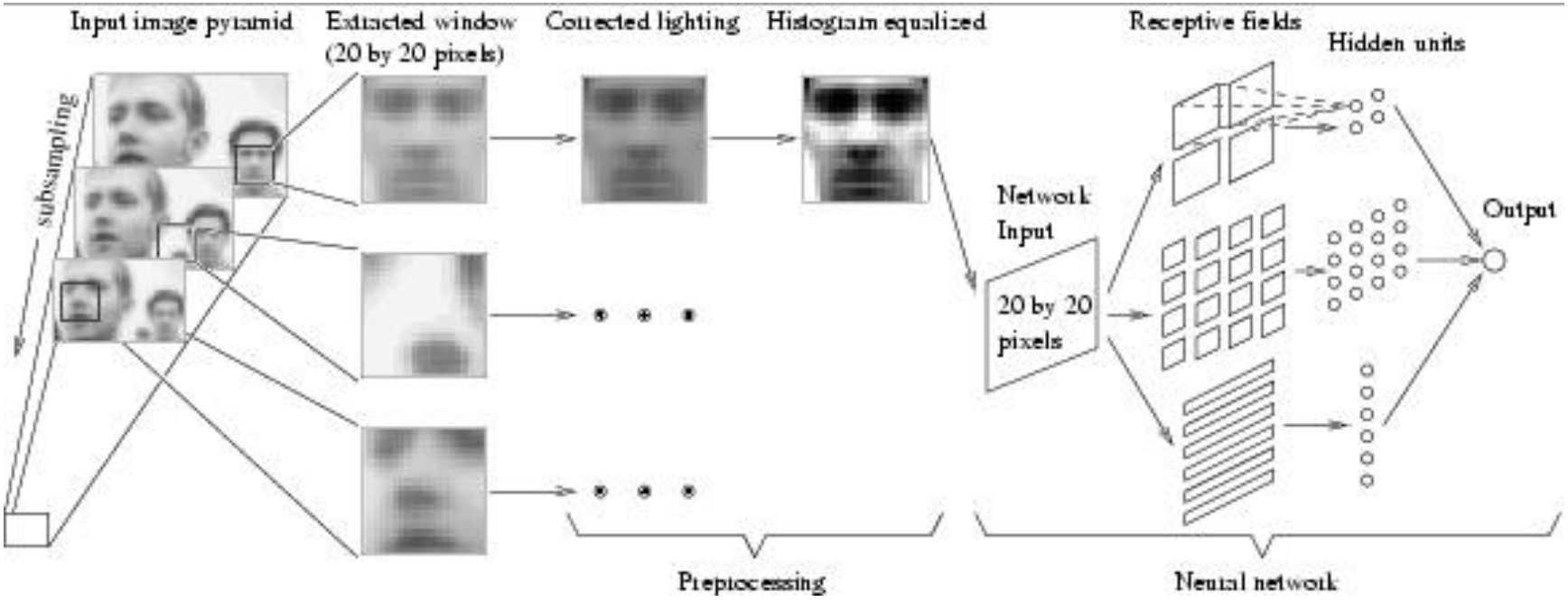
1000+ Real, 3000+ *VIRTUAL*

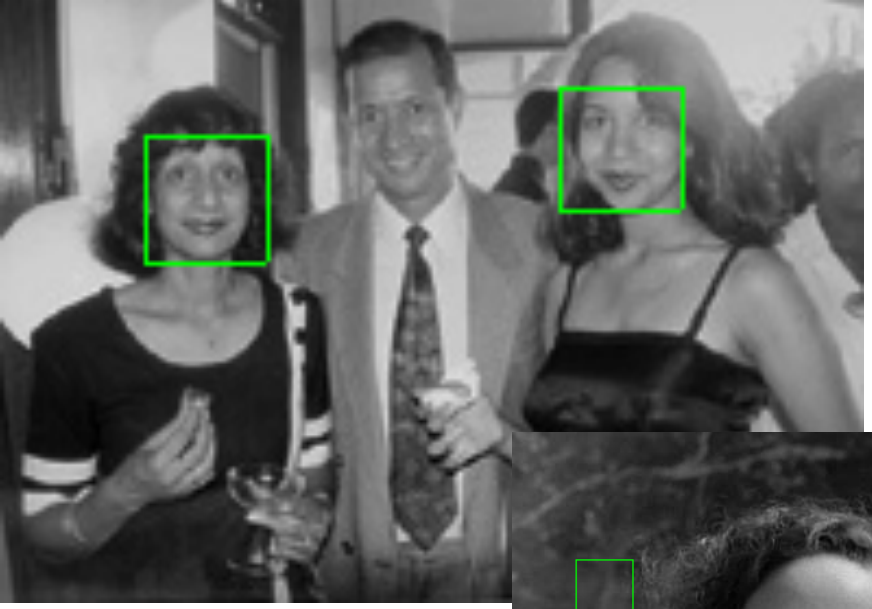
50,0000+ Non-Face Pattern

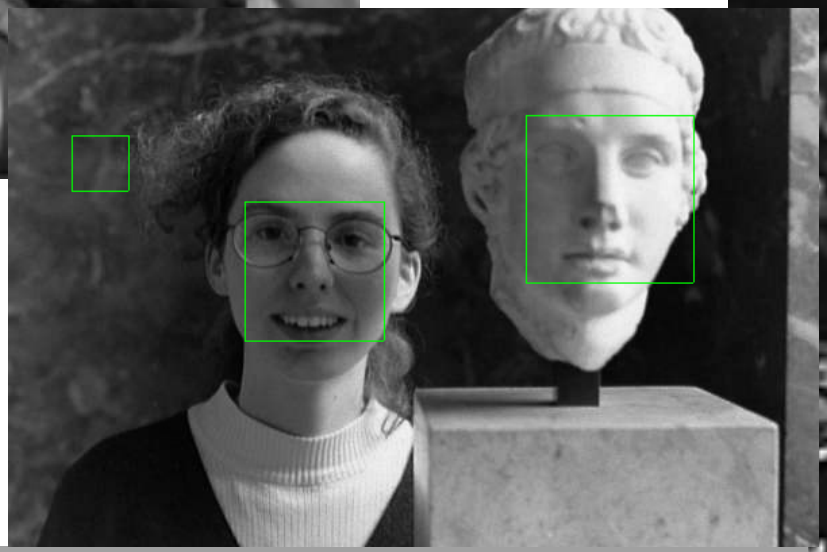


Neural Network-Based Face Detector

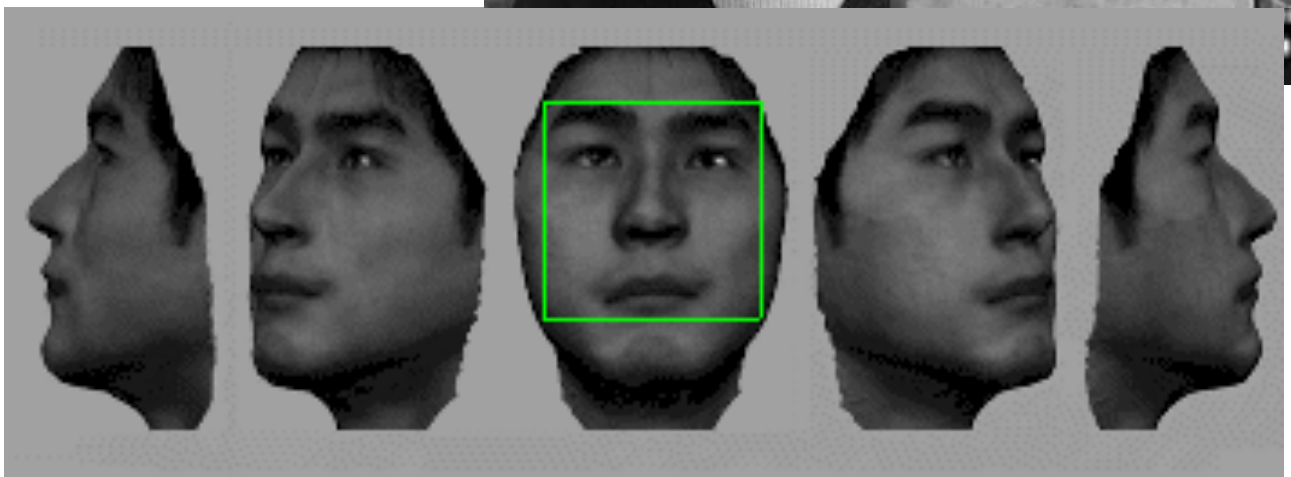
- Train a set of multilayer perceptrons and arbitrate a decision among all outputs [Rowley et al. 98]








STAR TREK
DEEP SPACE NINE
THE WAY OF THE WARRIOR
A Novel by Diane Carey

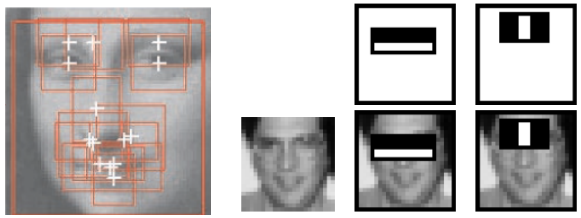


Families of recognition algorithms

Bag of words models

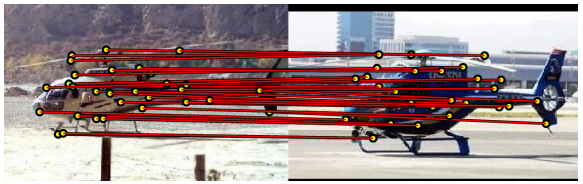
Csurka, Dance, Fan, Willamowski, and Bray 2004
Sivic, Russell, Freeman, Zisserman, ICCV 2005

Voting models



Viola and Jones, ICCV 2001
Heisele, Poggio, et. al., NIPS 01
Schneiderman, Kanade 2004
Vidal-Naquet, Ullman 2003

Shape matching
Deformable models



Berg, Berg, Malik, 2005
Cootes, Edwards, Taylor, 2001

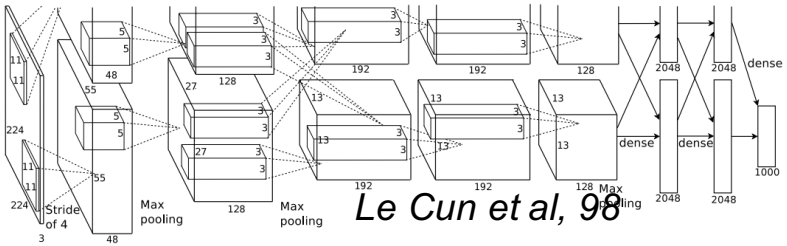
Rigid template models

Sirovich and Kirby 1987
Turk, Pentland, 1991
Dalal & Triggs, 2006

Constellation models

Fischler and Elschlager, 1973
Burl, Leung, and Perona, 1995
Weber, Welling, and Perona, 2000
Fergus, Perona, & Zisserman, CVPR 2003

Neural networks



Le Cun et al, 98

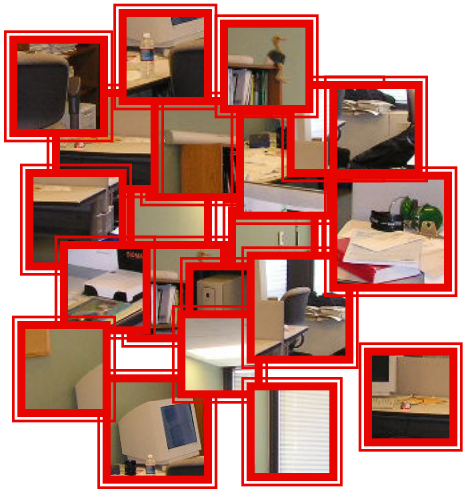
Discriminative methods

Object detection and recognition is formulated as a classification problem.

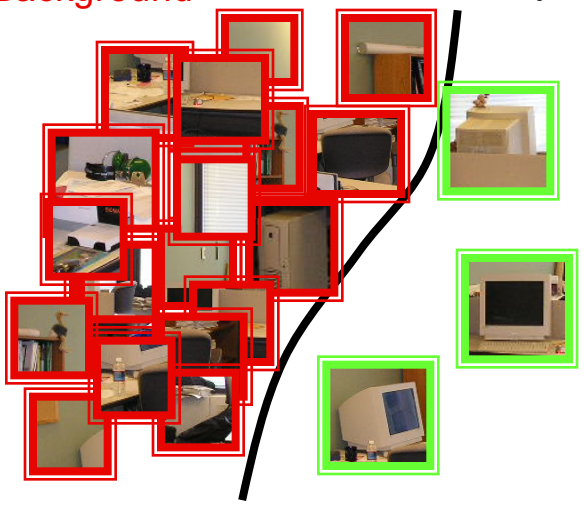
The image is partitioned into a set of overlapping windows

… and a decision is taken at each window about if it contains a target object or not.

Where are the screens?



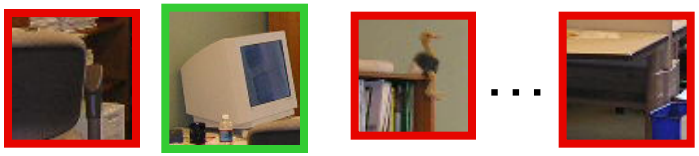
Bag of image patches



In some feature space

Formulation

- Formulation: binary classification



Features	x =	x_1	x_2	x_3	⋯	x_N	x_{N+1}	x_{N+2}	⋯	x_{N+M}
Labels	y =	-1	+1	-1		-1	?	?		?

Training data: each image patch is labeled as containing the object or background

Test data

- Classification function

$$\hat{y} = F(x)$$ Where $F(x)$ belongs to some family of functions

- Minimize misclassification error

(Not that simple: we need some guarantees that there will be generalization)

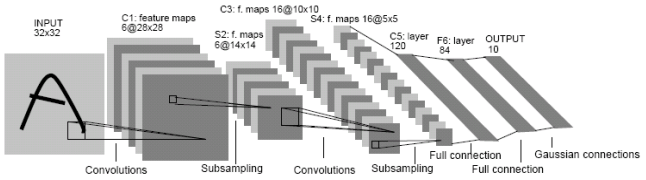
Discriminative methods

Nearest neighbor

10^6 examples

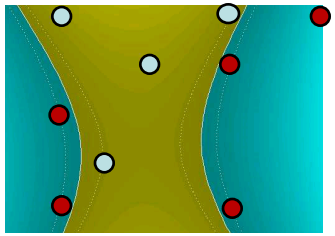
Shakhnarovich, Viola, Darrell 2003
Berg, Berg, Malik 2005
…

Neural networks



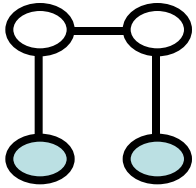
LeCun, Bottou, Bengio, Haffner 1998
Rowley, Baluja, Kanade 1998
…

Support Vector Machines and Kernels



Guyon, Vapnik
Heisele, Serre, Poggio, 2001
…

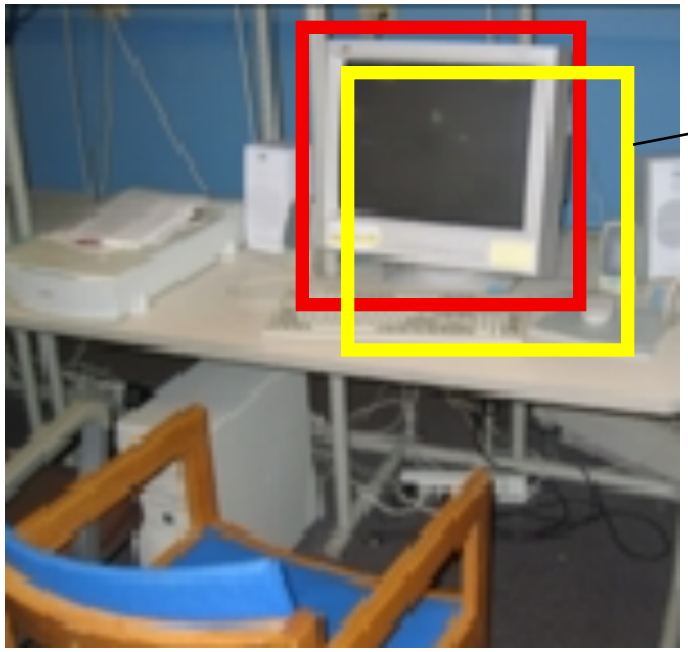
Conditional Random Fields



McCallum, Freitag, Pereira 2000
Kumar, Hebert 2003
…

Evaluation

When do we have a correct detection?



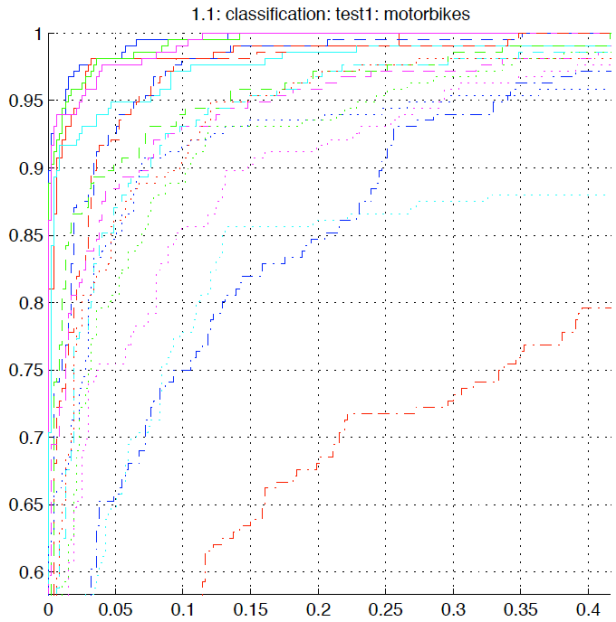
Is this correct?

$$\frac{\text{Area intersection}}{\text{Area union}} > 0.5$$

- ROC
- Precision-recall

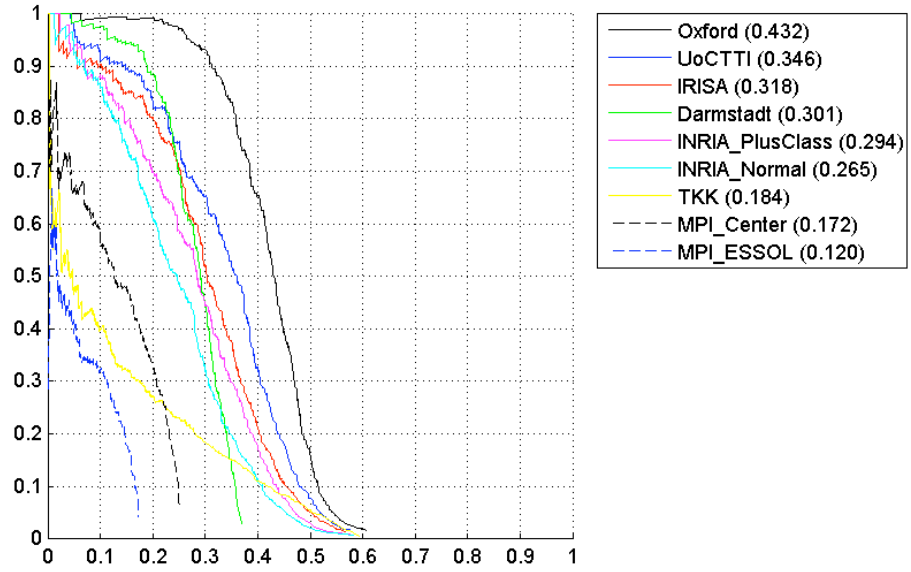
ROC and Precision-Recall

Detection rate



False alarm rate

Precision



Recall

Plots from PASCAL competition

Rapid Object Detection Using a Boosted Cascade of Simple Features

Paul Viola Michael J. Jones
Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA

Most of this work was done at Compaq CRL before the authors moved to MERL

Manuscript available on web:

http://citeseer.ist.psu.edu/cache/papers/cs/23183/http:zSzzSzwww.ai.mit.eduzSzpeoplezSzviolazSzresearchzSzpublicationszSzICCV01-Viola-Jones.pdf/viola01robust.pdf

Coarse-to-Fine Face Detection

François Fleuret * Donald Geman †

June 2000

for other objects in various subsets.

Finally, in defense of limited goals, nobody has yet demonstrated that objects from even one generic class under constrained poses can be rapidly detected without errors in complex, natural scenes; visual selection by humans occurs within two hundred milleseconds and is virtually perfect.

Acknowledgements: We are grateful to Yali Amit for many suggestions during a

*Avant-Projet IMEDIA, INRIA-Rocquencourt, Domaine de Voluceau, B.P.105, 78153 Le Chesnay. Email:Francois.Fleuret@inria.fr. Supported in part by the CNET.

†Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003. Email:geman@math.umass.edu. Supported in part by ONR under contract N00014-97-1-0249 and ARO under MURI grant DAAH04-96-1-0445.

What is novel about this approach?

- Feature set (… is huge about 16,000,000 features)
- Efficient feature selection using AdaBoost
- New image representation: Integral Image
- Cascaded Classifier for rapid detection
  - Hierarchy of Attentional Filters

What is new is the combination of these ideas. This yields the fastest known face detector for gray scale images.

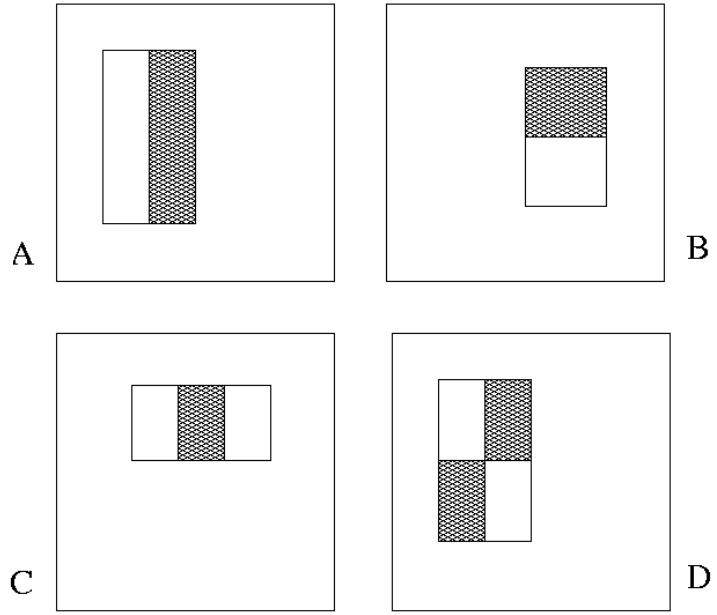
Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001

Image Features

"Rectangle filters"

Similar to Haar wavelets

Differences between sums of pixels in adjacent rectangles



$$h_t(x) = \begin{cases} +1 & \text{if } f_t(x) > \theta_t \\ -1 & \text{otherwise} \end{cases}$$

$$160{,}000 \times 100 = 16{,}000{,}000$$

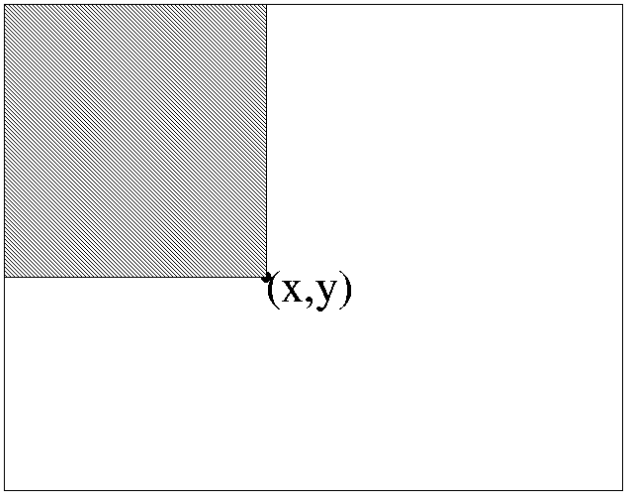
Unique Features

Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001

Integral Image

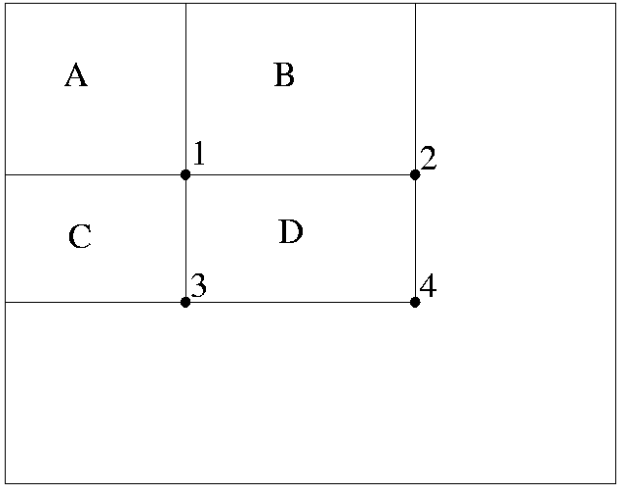
- Define the Integral Image

$$I'(x,y) = \sum_{\substack{x' \le x \\ y' \le y}} I(x',y')$$



- Any rectangular sum can be computed in constant time:

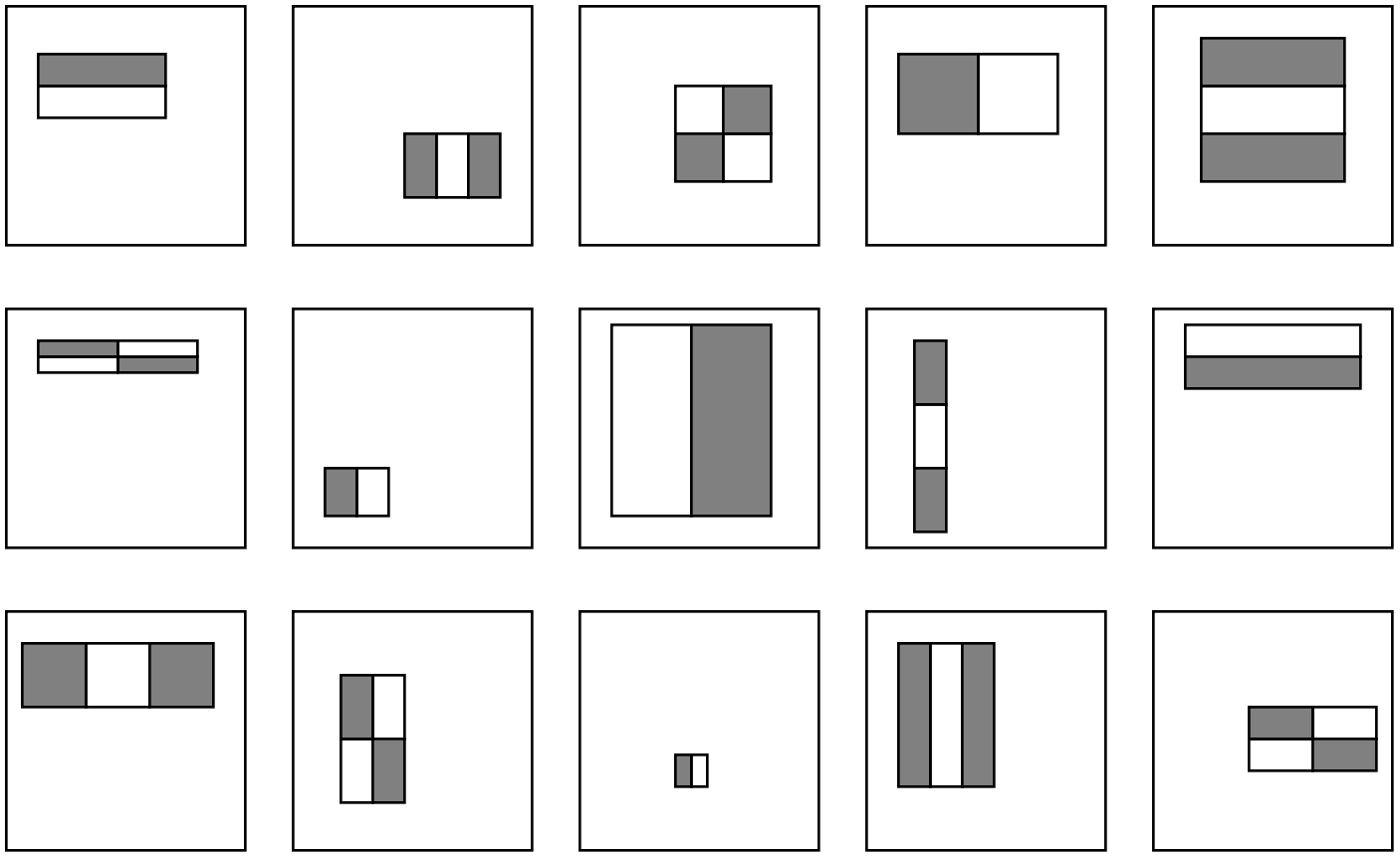
$$\begin{aligned} D &= 1 + 4 - (2 + 3) \\ &= A + (A + B + C + D) - (A + C + A + B) \\ &= D \end{aligned}$$



- Rectangle features can be computed as differences between rectangles

Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001

Huge “Library” of Filters



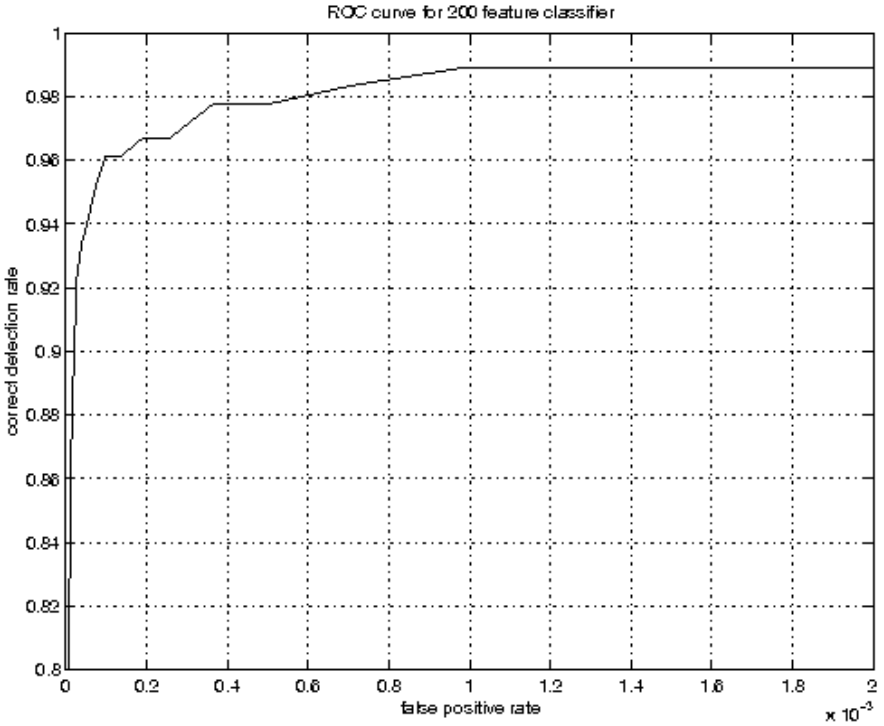
Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001

Example Classifier for Face Detection

A classifier with 200 rectangle features was learned using AdaBoost

95% correct detection on test set with 1 in 14084 false positives.

Not quite competitive.
Need to add more features,
but then that slows it down.



ROC curve for 200 feature classifier

Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001

Boosting

- Defines a classifier using an additive model:

$$F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) + \ldots$$

Strong classifier

Features vector

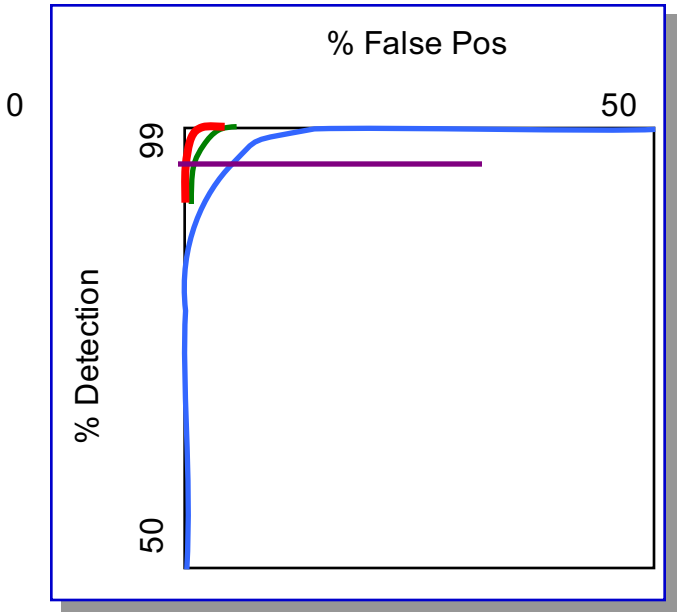
Weight

Weak classifier

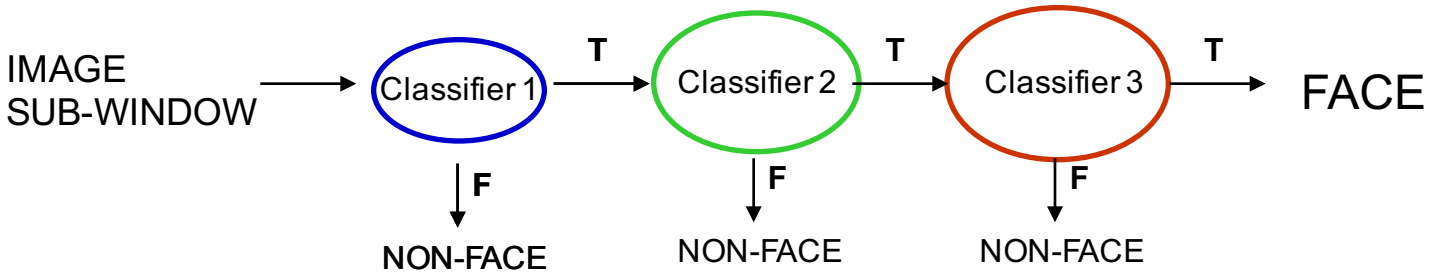
Fast and accurate classifier using a cascade

Fleuret and Geman 2001, Viola and Jones 2001

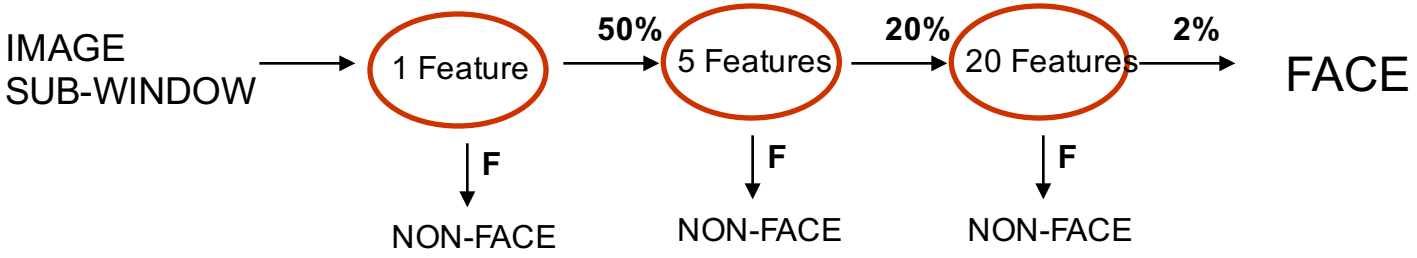
- Given a nested set of classifier hypothesis classes



- Cascade



Cascaded Classifier



- A 1 feature classifier achieves 100% detection rate and about 50% false positive rate.
- A 5 feature classifier achieves 100% detection rate and 40% false positive rate (20% cumulative)
  - using data from previous stage.
- A 20 feature classifier achieve 100% detection rate with 10% false positive rate (2% cumulative)

Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001

A Real-time Face Detection System

Training faces: 4916 face images (24 x 24 pixels) plus vertical flips for a total of 9832 faces

Training non-faces: 350 million sub-windows from 9500 non-face images

Final detector: 38 layer cascaded classifier
The number of features per layer was 1, 10, 25, 25, 50, 50, 50, 75, 100, …, 200, …

Final classifier contains 6061 features.

Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001

Speed of Face Detector

Speed is proportional to the average number of features computed per sub-window.

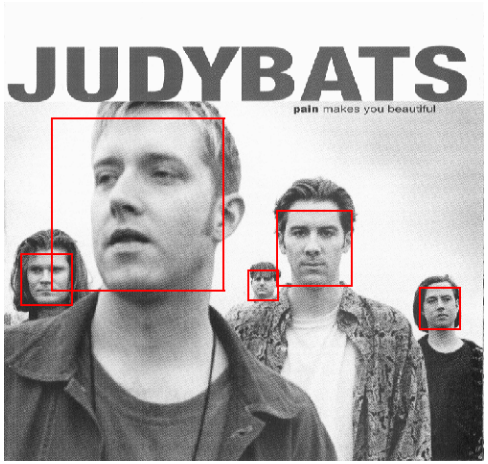
On the MIT+CMU test set, an average of 9 features out of a total of 6061 are computed per sub-window.

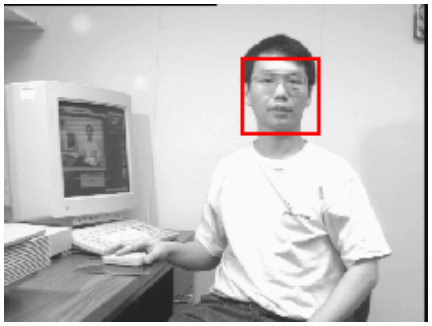
On a 700 Mhz Pentium III, a 384x288 pixel image takes about 0.067 seconds to process (15 fps).

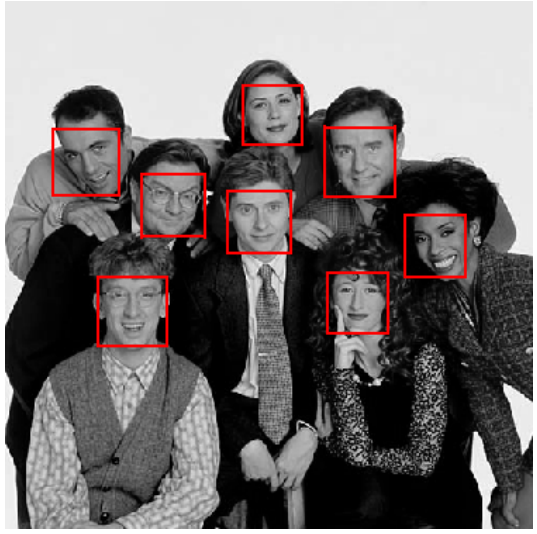
Roughly 15 times faster than Rowley-Baluja-Kanade and 600 times faster than Schneiderman-Kanade.

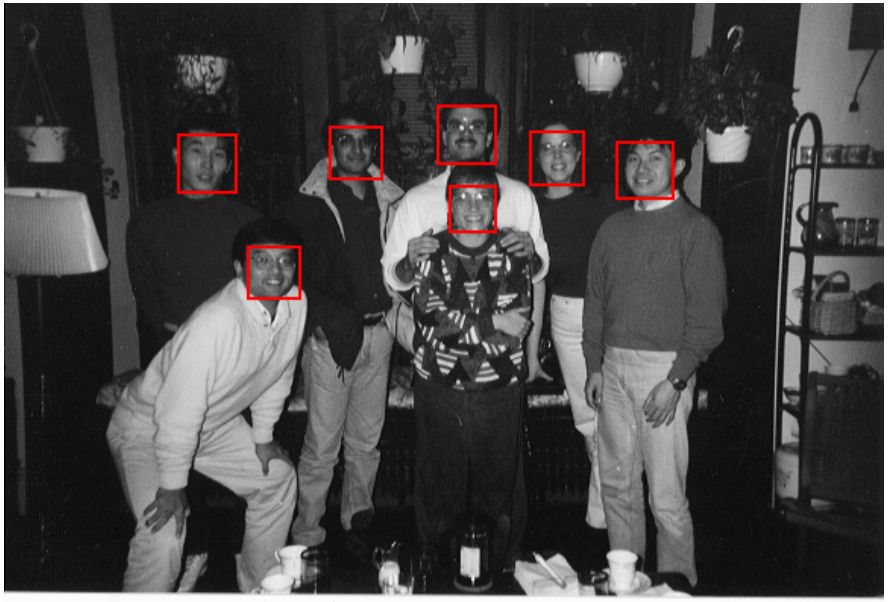
Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001

Output of Face Detector on Test Images

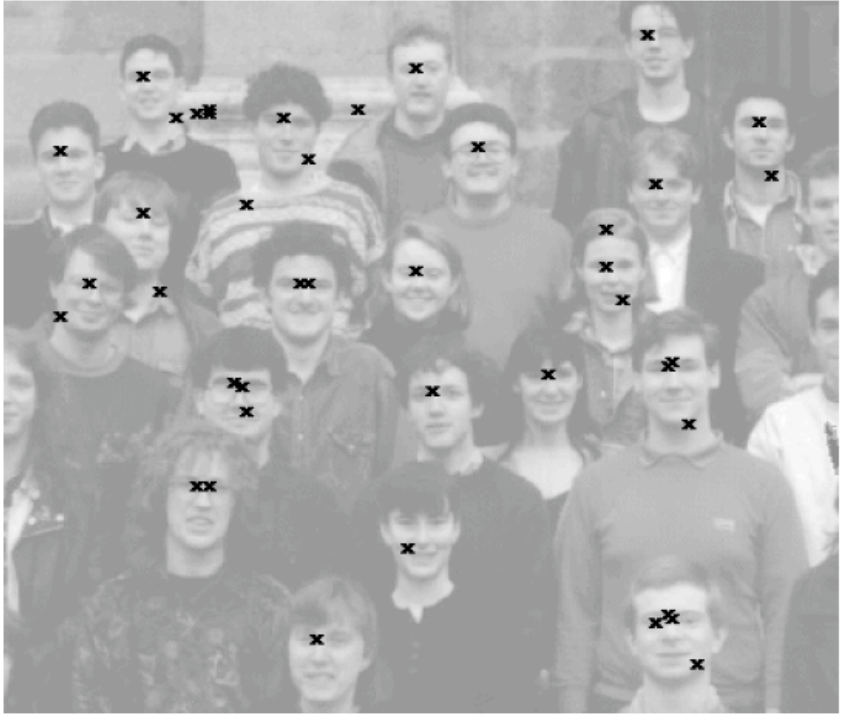


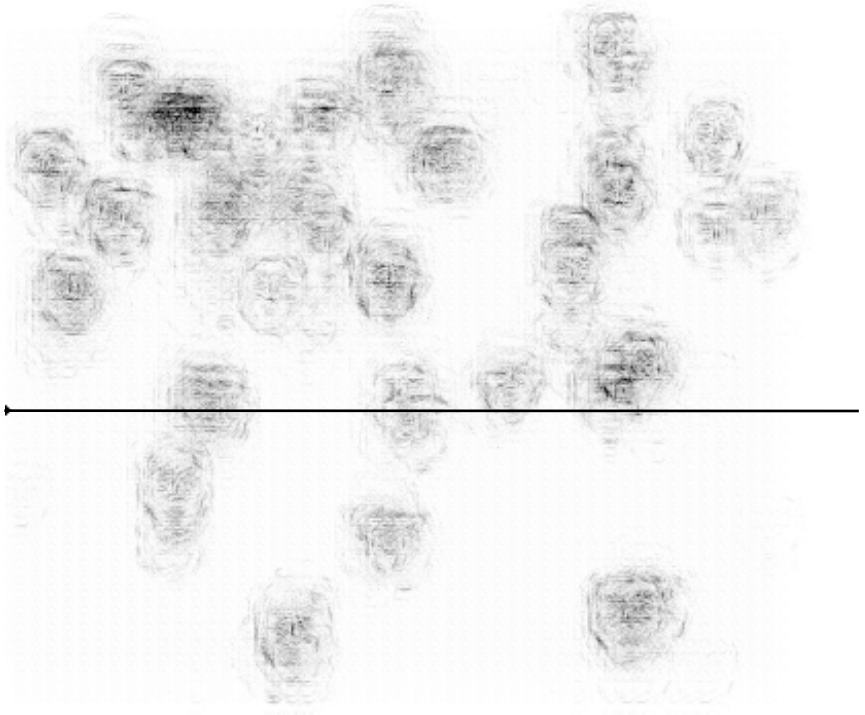


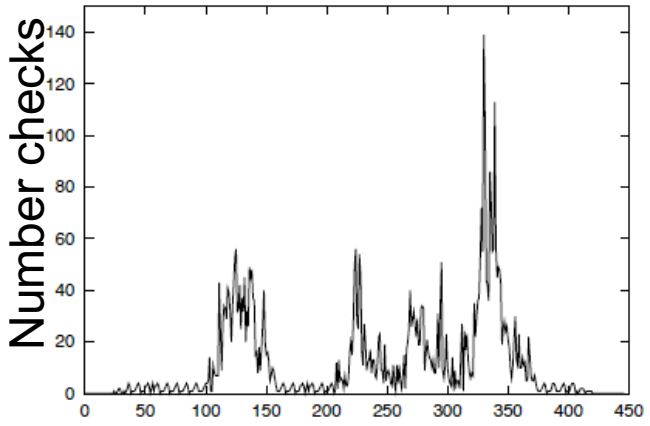




Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001





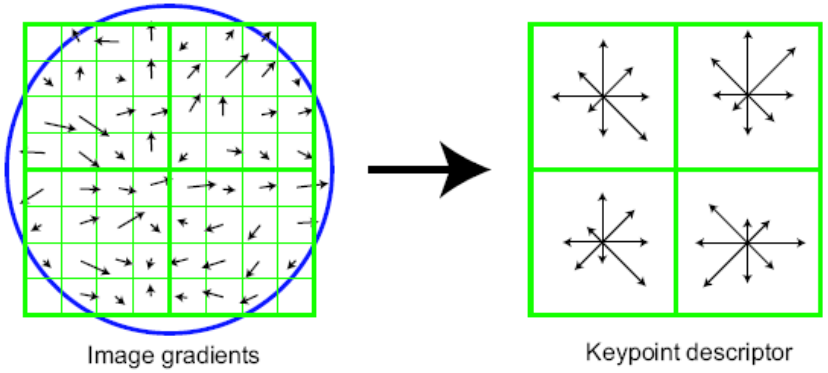


Fleuret and Geman 2001

Histograms of oriented gradients

Histograms of oriented gradients

SIFT, D. Lowe, ICCV 1999



Shape context

Belongie, Malik, Puzicha, NIPS 2000

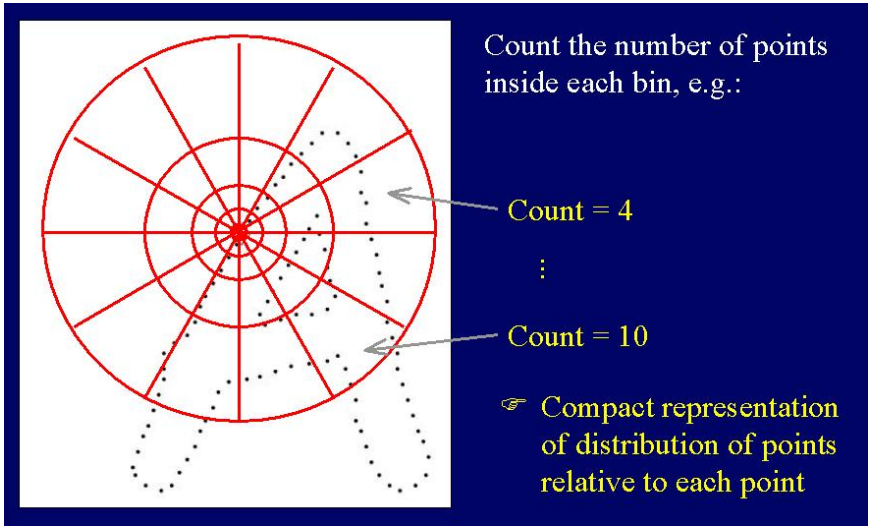
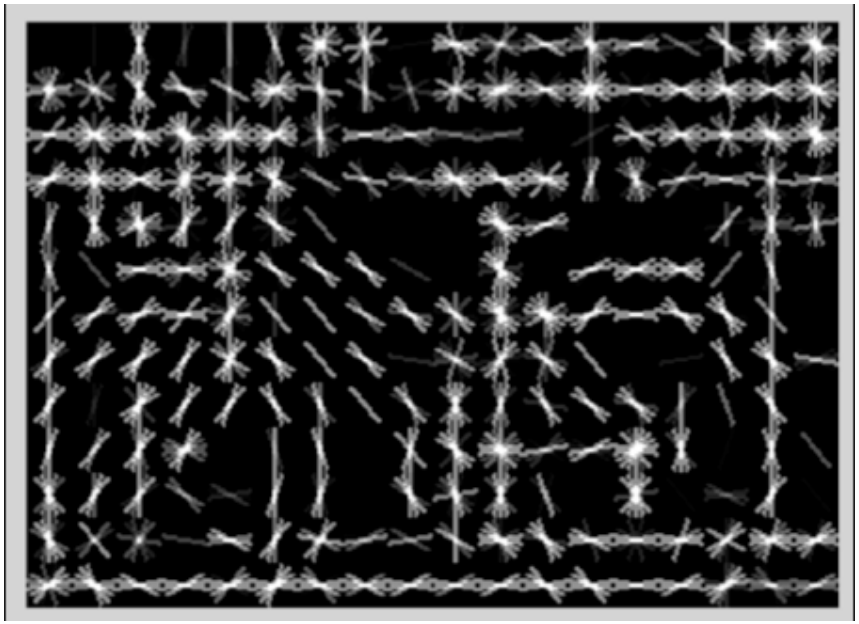


Image features:

Histograms of oriented gradients (HOG)



Bin gradients from 8x8 pixel neighborhoods into 9 orientations

(Dalal & Triggs *CVPR* 05)

Source: Deva Ramanan

Histograms of Oriented Gradients for Human Detection

Navneet Dalal and Bill Triggs

INRIA Rhône-Alps, 655 avenue de l'Europe, Montbonnot 38334, France

{Navneet.Dalal,Bill.Triggs}@inrialpes.fr, http://lear.inrialpes.fr

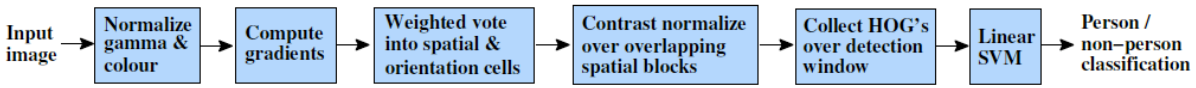
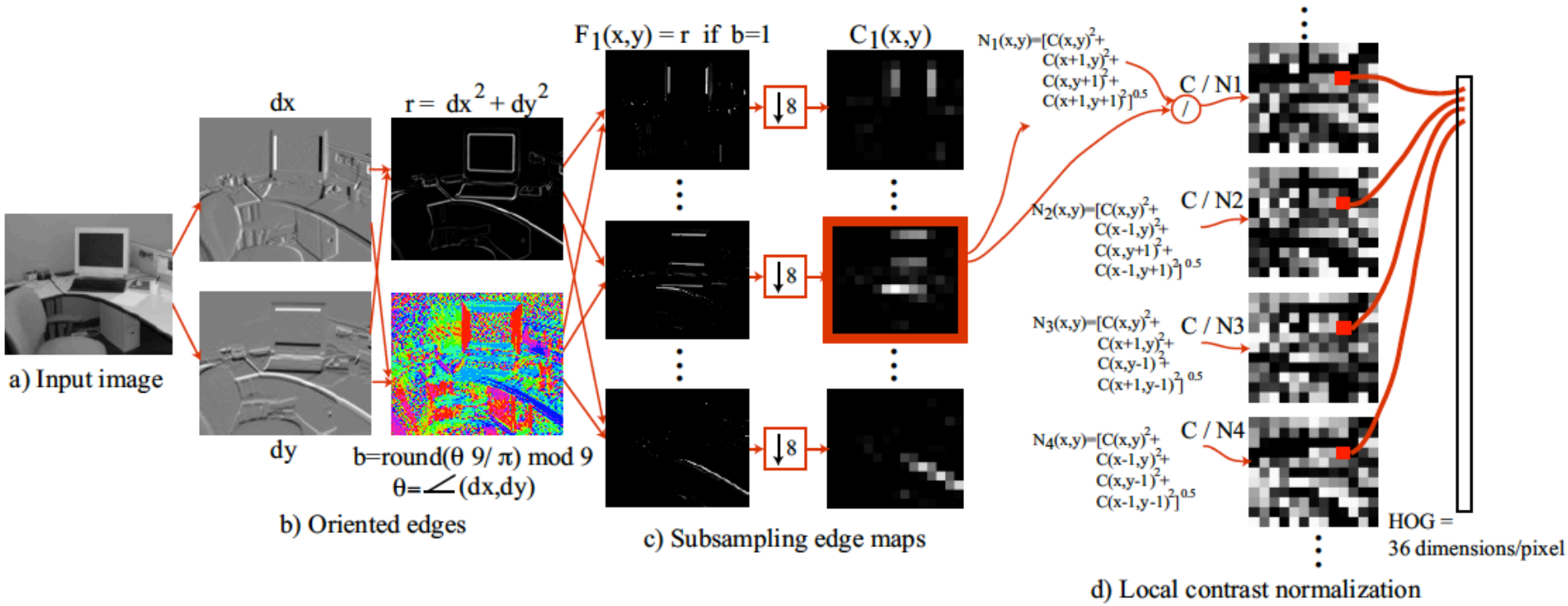


Figure 1. An overview of our feature extraction and object detection chain. The detector window is tiled with a grid of overlapping blocks in which Histogram of Oriented Gradient feature vectors are extracted. The combined vectors are fed to a linear SVM for object/non-object classification. The detection window is scanned across the image at all positions and scales, and conventional non-maximum suppression is run on the output pyramid to detect object instances, but this paper concentrates on the feature extraction process.



SVM

A Support Vector Machine (SVM) learns a classifier with the form:

$$H(x) = \sum_{m=1}^{M} a_m y_m k(x, x_m)$$

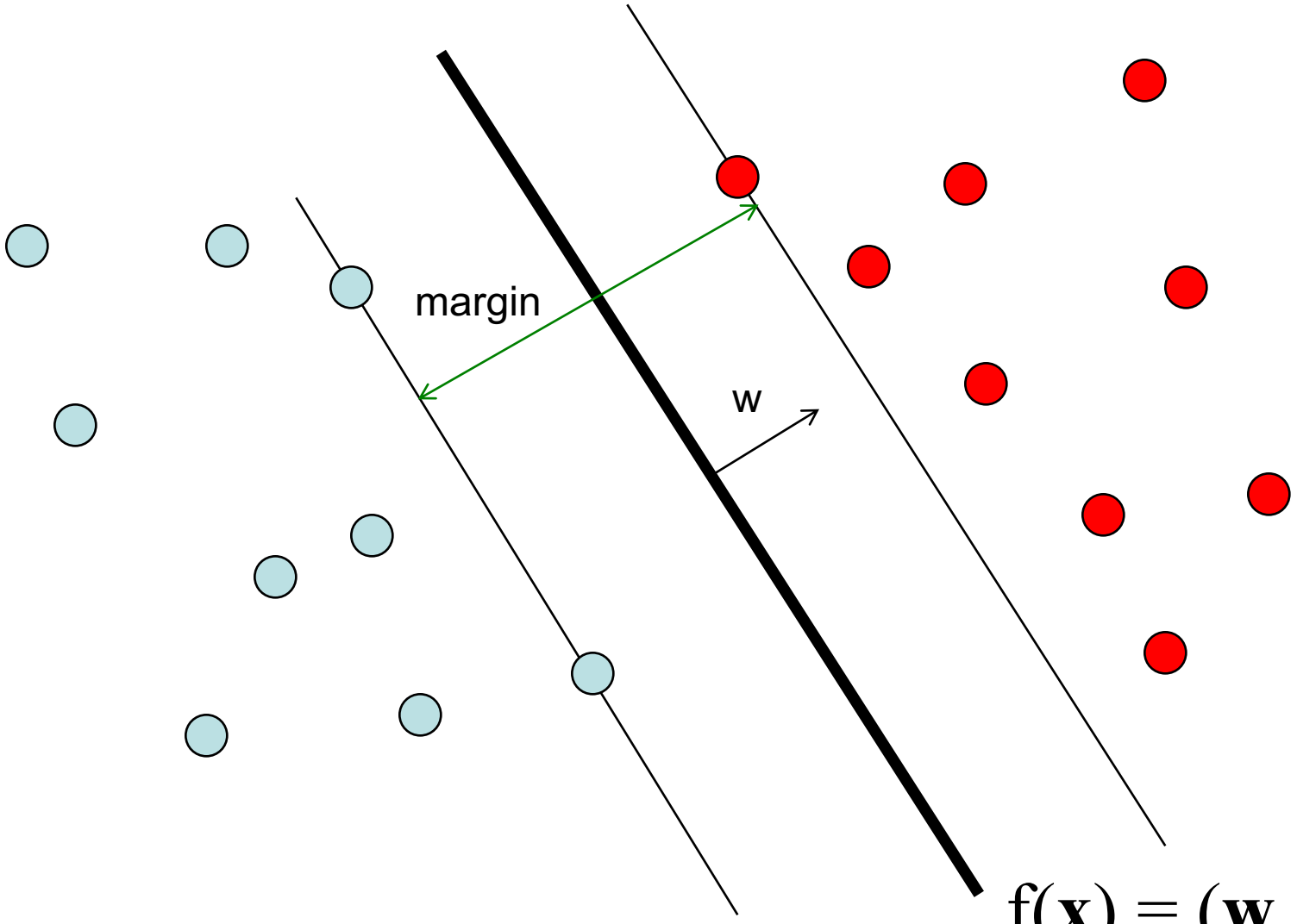
Where $\{x_m, y_m\}$, for m = 1 . . .M, are the training data with x_m being the input feature vector and y_m = +1,-1 the class label. k(x, x_m) is the kernel and it can be any symmetric function satisfying the Mercer Theorem.

The classification is obtained by thresholding the value of H(x).

There is a large number of possible kernels, each yielding a different family of decision boundaries:

- Linear kernel: $k(x, x_m) = x^T x_m$
- Radial basis function: $k(x, x_m) = \exp(-|x - x_m|^2/\sigma^2)$.
- Histogram intersection: $k(x,x_m) = \text{sum}_i(\min(x(i), x_m(i)))$

Linear SVM



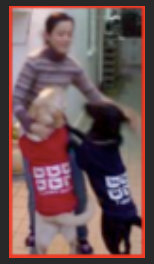
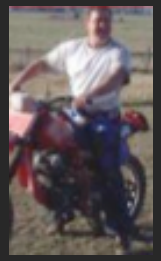
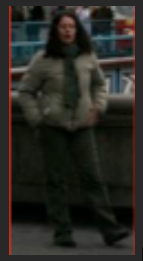
$$f(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x} + b)$$

Scanning-window templates

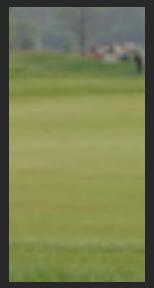
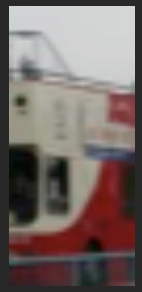
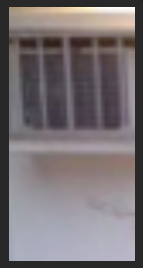
Dalal and Triggs CVPR05 (HOG)

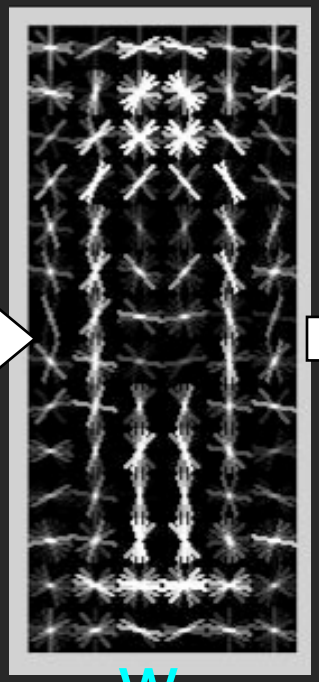
Papageorgiou and Poggio ICIP99 (wavelets)

pos

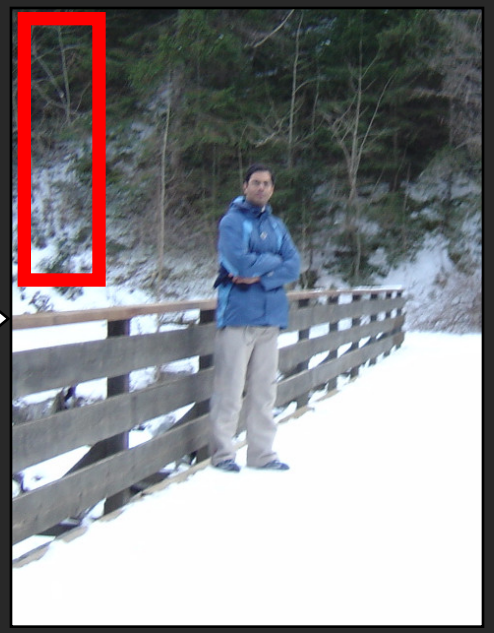




neg

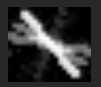






W



w = weights for orientation and spatial bins



$w \cdot x > 0$

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

Source: Deva Ramanan

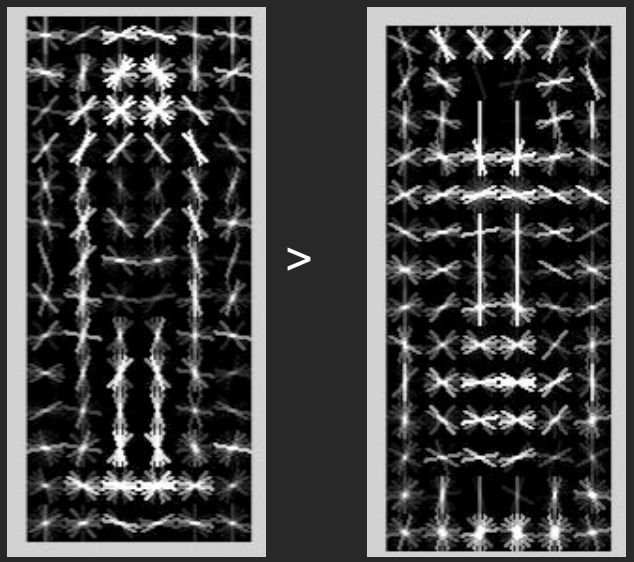
How to interpret positive and negative weights?

$$w \cdot x > 0$$

$$(w_{pos} - w_{neg}) \cdot x > 0$$

$$w_{pos} \cdot x > w_{neg} \cdot x$$

Pedestrian template



Pedestrian background template

w_{pos}, w_{neg} = weighted average of positive, negative support vectors

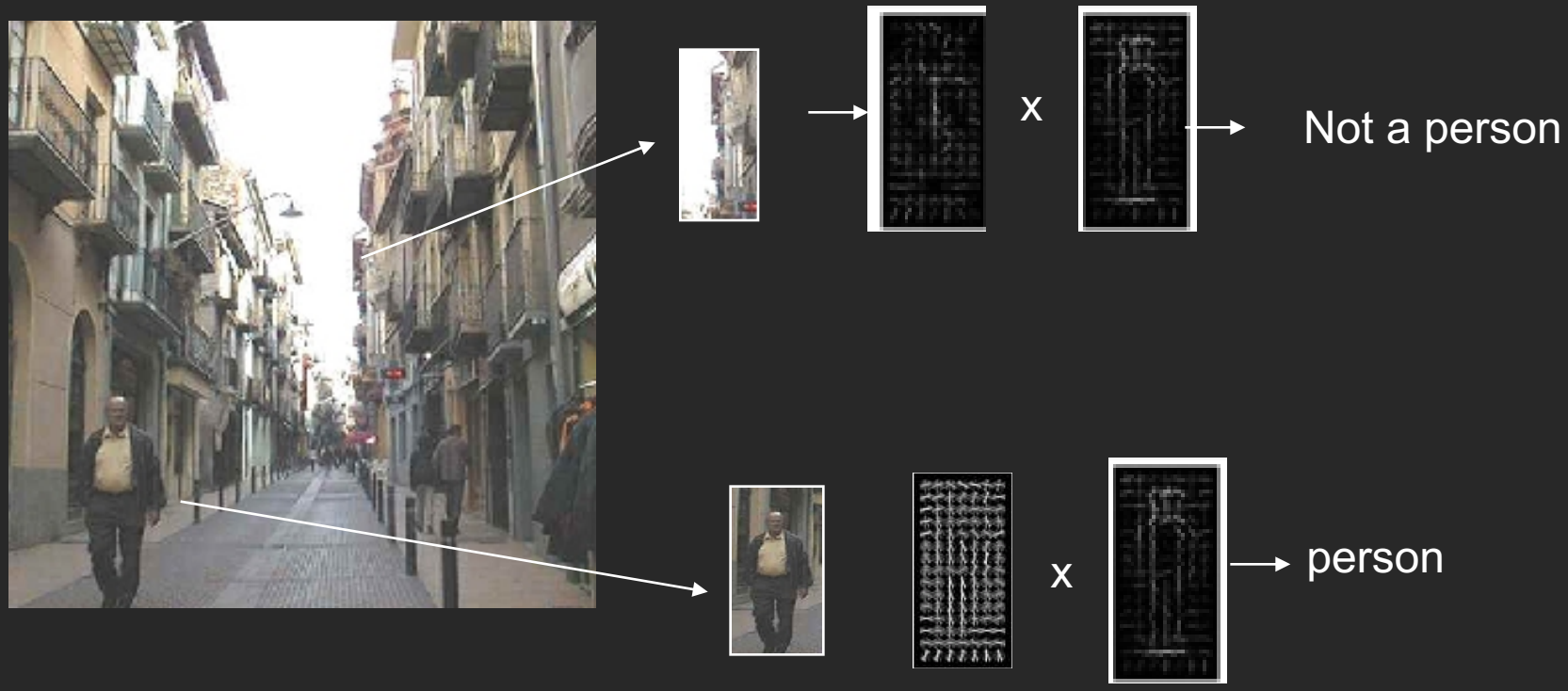
Right approach is to compete pedestrian, pillar, doorway... models

Background class is hard to model - easier to penalize particular vertical edges

Source: Deva Ramanan

Histograms of oriented gradients

Dalal & Trigs, 2006



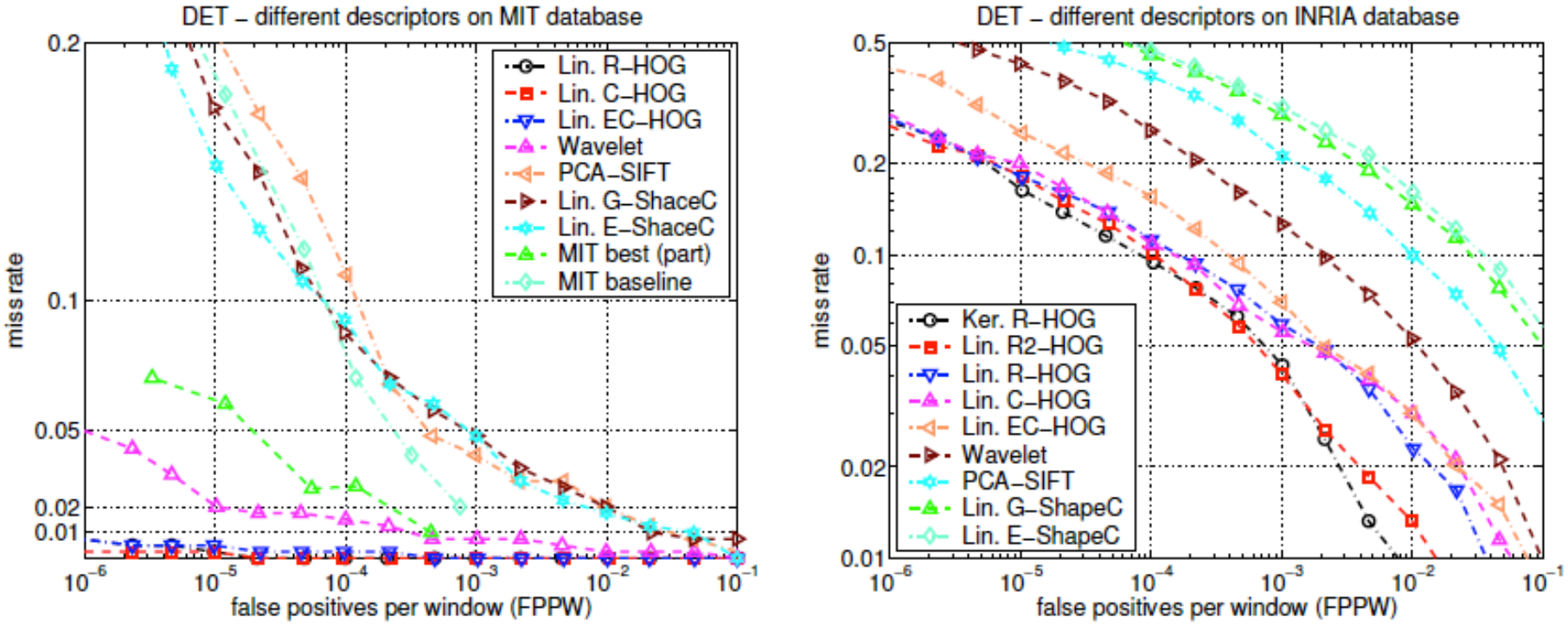


Figure 3. The performance of selected detectors on (left) MIT and (right) INRIA data sets. See the text for details.

Representation

- Object as set of parts
  - Generative representation

- Model:
  - Relative locations between parts
  - Appearance of part

- Issues:
  - How to model location
  - How to represent appearance
  - Sparse or dense (pixels or regions)
  - How to handle occlusion/clutter

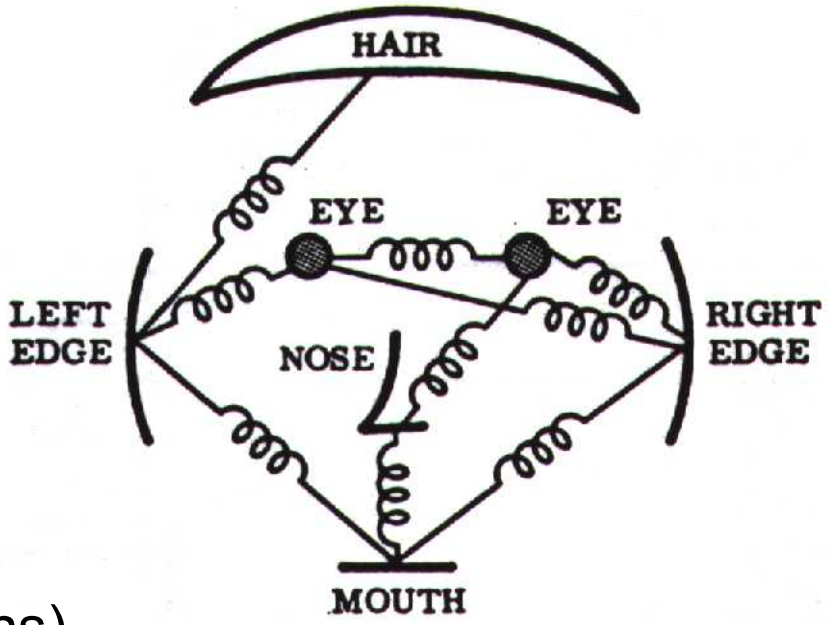


Figure from [Fischler & Elschlager 73]

The Representation and Matching of Pictorial Structures

MARTIN A. FISCHLER AND ROBERT A. ELSCHLAGER

Abstract—The primary problem dealt with in this paper is the following. Given some description of a visual object, find that object in an actual photograph. Part of the solution to this problem is the specification of a descriptive scheme, and a metric on which to base the decision of "goodness" of matching or detection.

We offer a combined descriptive scheme and decision metric which is general, intuitively satisfying, and which has led to promising experimental results. We also present an algorithm which takes the above descriptions, together with a matrix representing the intensities of the actual photograph, and then finds the described object in the matrix. The algorithm uses a procedure similar to dynamic programming in order to cut down on the vast amount of computation otherwise necessary.

One desirable feature of the approach is its generality. A new programming system does not need to be written for every new description; instead, one just specifies descriptions in terms of a certain set of primitives and parameters.

There are many areas of application: scene analysis and description, map matching for navigation and guidance, optical tracking, stereo compilation, and image change detection. In fact, the ability to describe, match, and register scenes is basic for almost any image processing task.

Index Terms—Dynamic programming, heuristic optimization, picture description, picture r
tation.

Manuscript received November 30, 1971; revised May 22, 1972, and August 21, 1972.

The authors are with the Lockheed Palo Alto Research Laboratory, Lockheed Missiles & Space Company, Inc., Palo Alto, Calif. 94304.

INT

THE PRIMARY
paper is the follo
a visual object, fi
graph. The object mig
complicated, such as an
can be linguistic, picto
photograph will be cal
dimensional array of gr
being sought is called t

This ability to find a
equivalently, to match
scenes, is basic for alm
Application to such are
tion, map matching for

Martin A. Fischler (S'57–M'58) was born in New York, N. Y., on February 15, 1932. He received the B.E.E. degree from the City College of New York, New York, in 1954 and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, Calif., in 1958 and 1962, respectively.

He served in the U. S. Army for two years and held positions at the National Bureau of Standards and at Hughes Aircraft Corporation during the period 1954 to 1958. In 1958 he joined the technical staff of the Lockheed Missiles & Space Company, Inc., at the Lockheed Palo Alto Research Laboratory, Palo Alto, Calif., and currently holds the title of Staff Scientist. He has conducted research and published in the areas of artificial intelligence, picture processing, switching theory, computer organization, and information theory.

Dr. Fischler is a member of the Association for Computing Machinery, the Pattern Recognition Society, the Mathematical Association of America, Tau Beta Pi, and Eta Kappa Nu. He is currently an Associate Editor of the journal *Pattern Recognition* and is a past Chairman of the San Francisco Chapter of the IEEE Society on Systems, Man, and Cybernetics.

❖

Robert A. Elschlager was born in Chicago, Ill., on May 25, 1943. He received the B.S. degree in mathematics from the University of Illinois, Urbana, in 1964, and the M.S. degree in mathematics from the University of California, Berkeley, in 1969.

Since then he has been an Associate Scientist with the Lockheed Missiles & Space Company, Inc., at the Lockheed Palo Alto Research Center, Palo Alto, Calif. His current interests are picture processing, operating systems, computer languages, and computer understanding.

Mr. Elschlager is a member of the American Mathematical Society, the Mathematical Association of America, and the Association for Symbolic Logic.

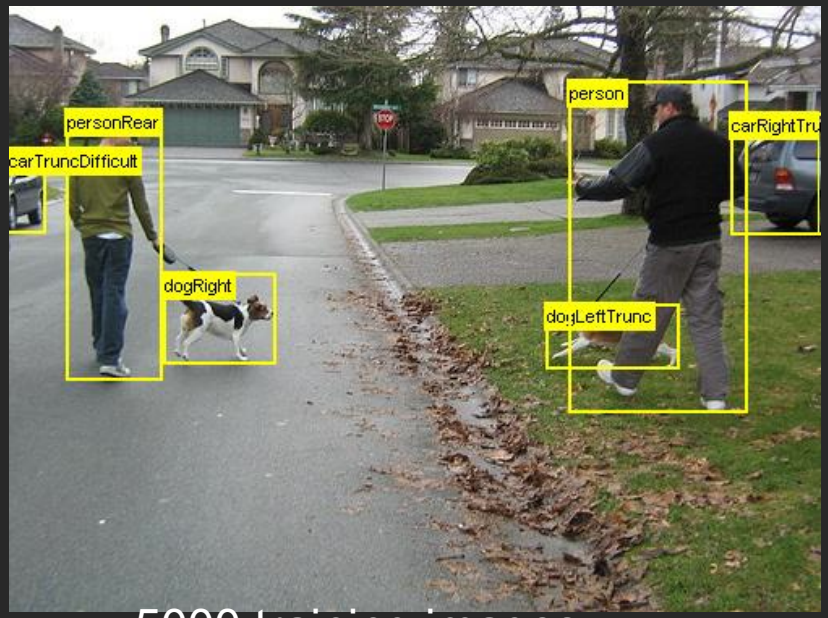
Object Detection with Discriminatively Trained Part Based Models

Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan

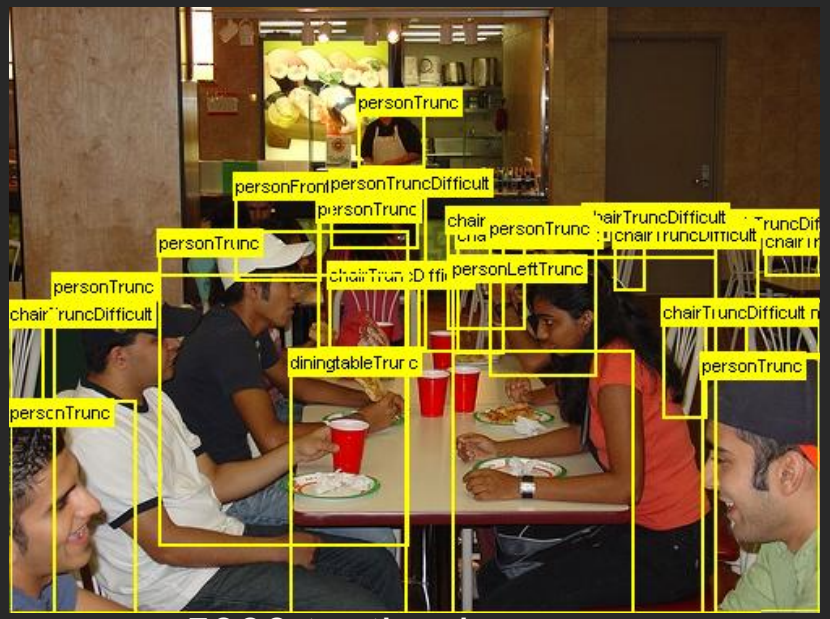
Abstract—We describe an object detection system based on mixtures of multiscale deformable part models. Our system is able to represent highly variable object classes and achieves state-of-the-art results in the PASCAL object detection challenges. While deformable part models have become quite popular, their value had not been demonstrated on difficult benchmarks such as the PASCAL datasets. Our system relies on new methods for discriminative training with partially labeled data. We combine a margin-sensitive approach for data-mining hard negative examples with a formalism we call *latent SVM*. A latent SVM is a reformulation of MI-SVM in terms of latent variables. A latent SVM is semi-convex and the training problem becomes convex once latent information is specified for the positive examples. This leads to an iterative training algorithm that alternates between fixing latent values for positive examples and optimizing the latent SVM objective function.

Index Terms—Object Recognition, Deformable Models, Pictorial Structures, Discriminative Training, Latent SVM

PASCAL Visual Object Challenge



5000 training images



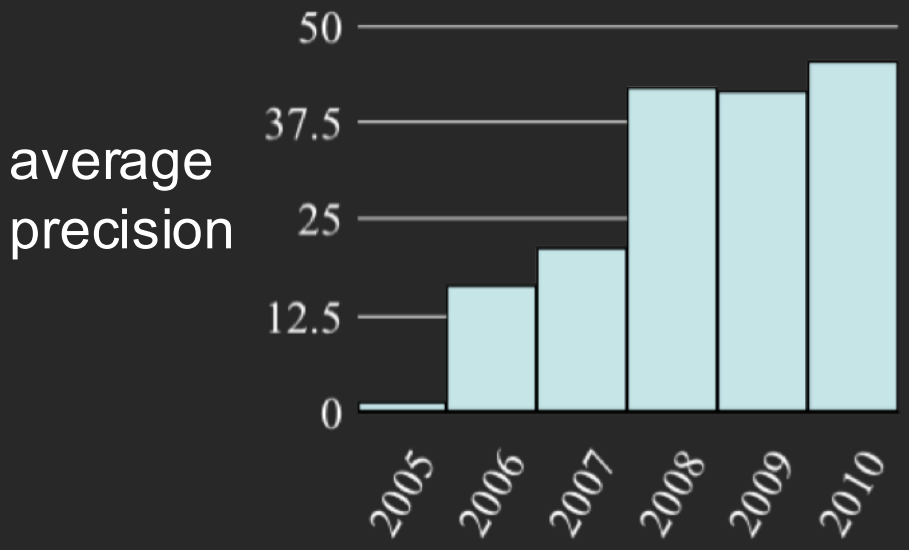
5000 testing images

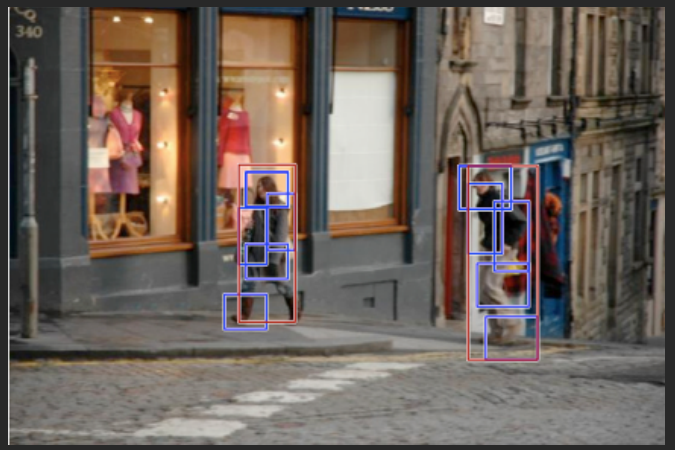
20 everyday object categories

aeroplane bike bird boat bottle bus car cat chair cow table
dog horse motorbike person plant sheep sofa train tv

Source: Deva Ramanan

5 years of PASCAL people detection



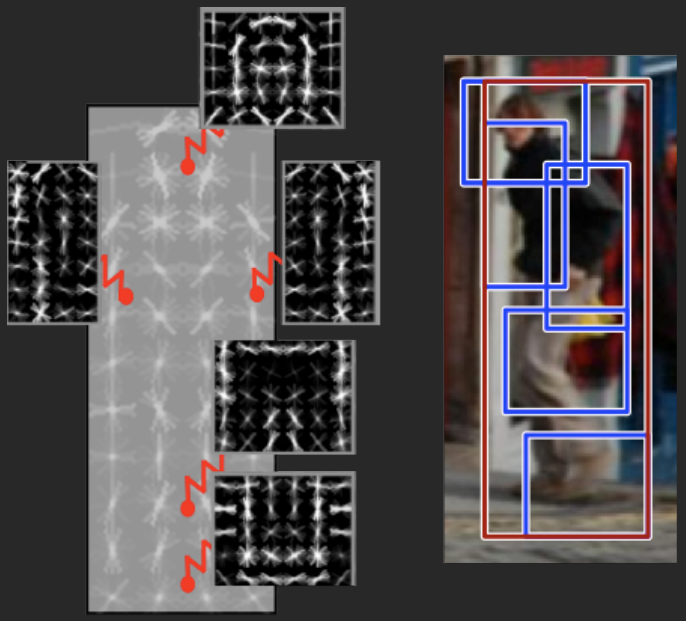


1% to 45% in 5 years

Discriminative mixtures of star models 2007-2010 Felzenszwalb, McAllester, Ramanan *CVPR* 2008
Felzenszwalb, Girshick, McAllester, and Ramanan *PAMI* 2009

Source: Deva Ramanan

Deformable part models



Model encodes local appearance + pairwise geometry

Source: Deva Ramanan

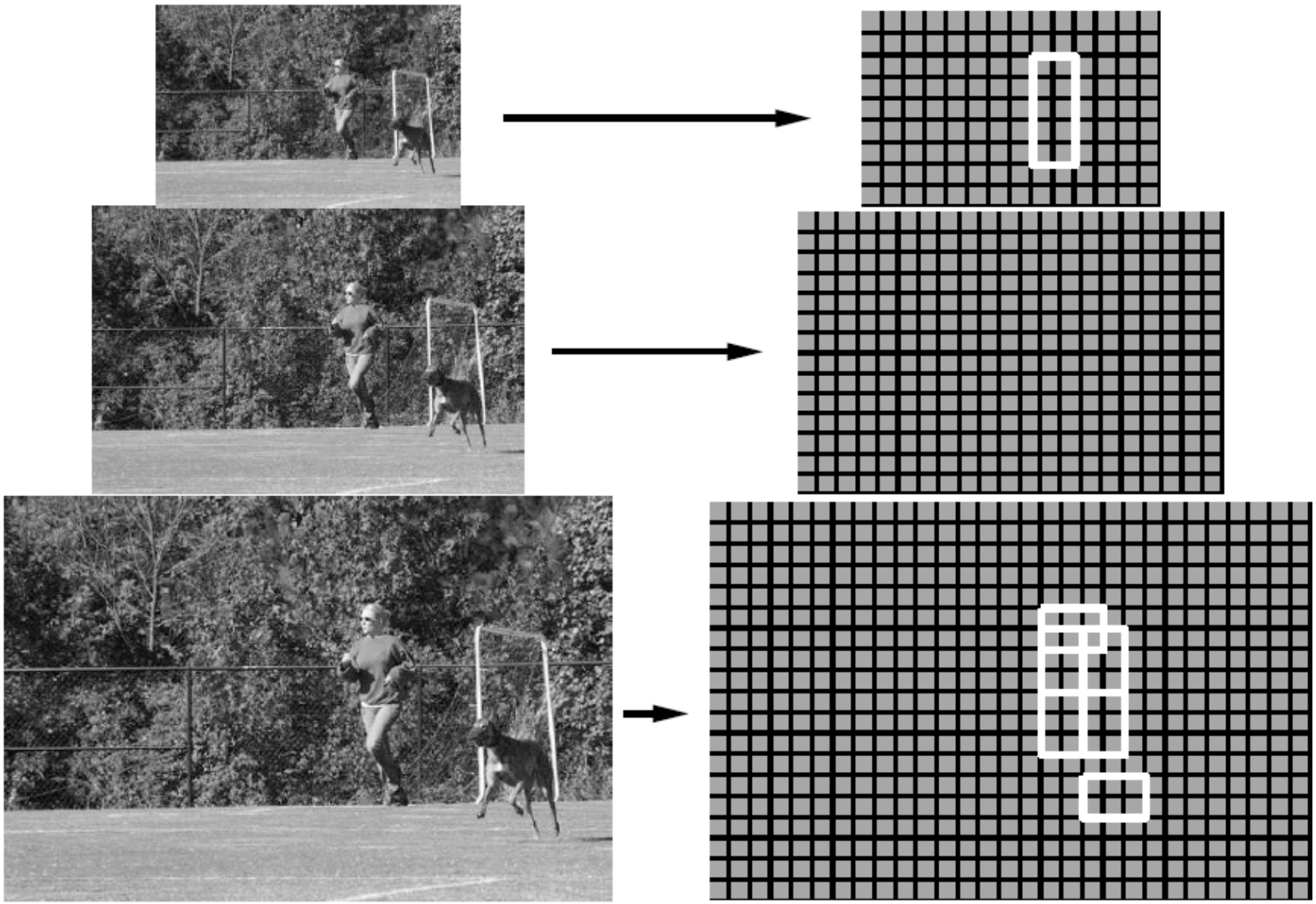
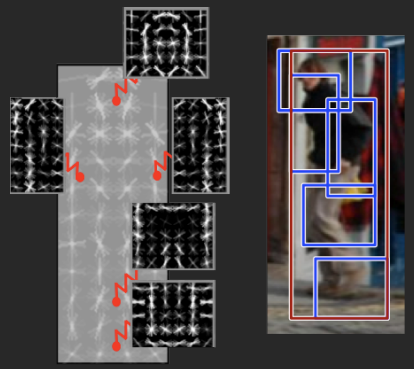


Image pyramid
Feature pyramid

Scoring function

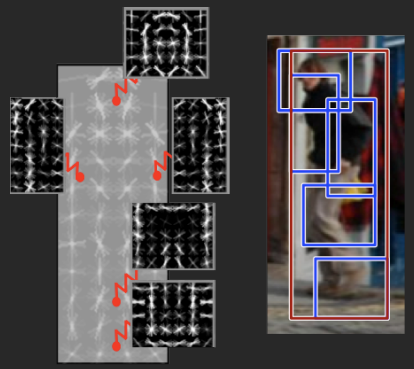


score(x,z)

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

Source: Deva Ramanan

Scoring function



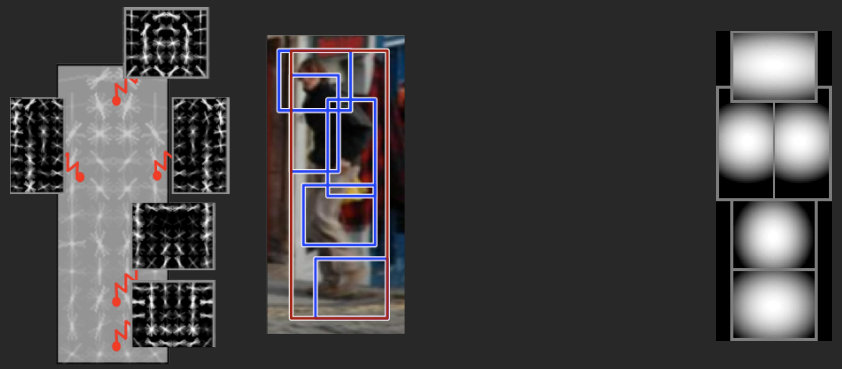
score(x,z) $= \sum_i w_i \cdot \phi(x, z_i) +$

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

part template scores

Source: Deva Ramanan

Scoring function



$$\text{score}(x,z) = \sum_i w_i \cdot \phi(x, z_i) + \sum_{i,j} w_{ij} \cdot \Psi(z_i, z_j)$$

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

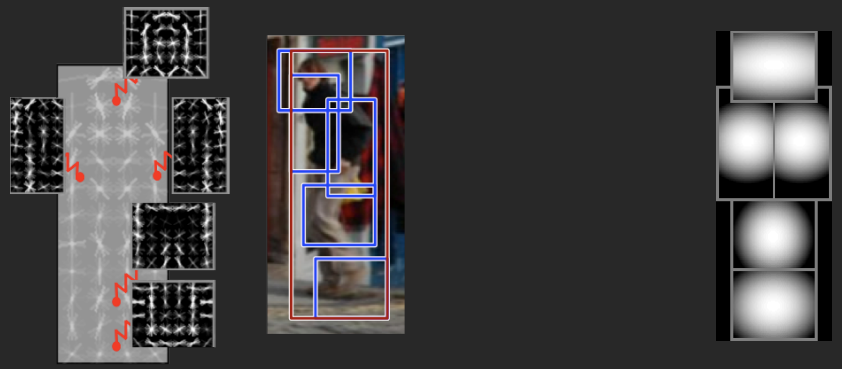
part template scores

spring deformation model

E = relational graph

Source: Deva Ramanan

Scoring function



$$\text{score}(x,z) = \sum_i w_i \cdot \phi(x, z_i) + \sum_{i,j} w_{ij} \cdot \Psi(z_i, z_j)$$

x = image $z_i = (x_i, y_i)$ $z = \{z_1, z_2 \ldots\}$	part template scores	spring deformation model

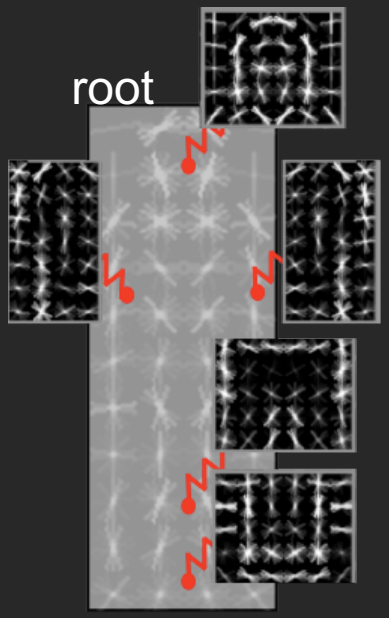
Score is linear in local templates w_i and spring parameters w_{ij}

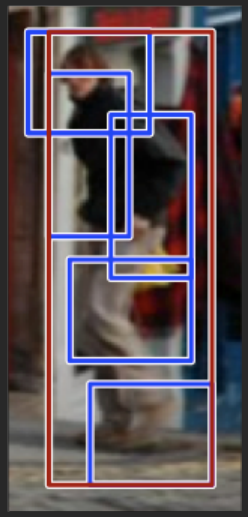
$$\text{score}(x,z) = w \cdot \Phi(x, z)$$

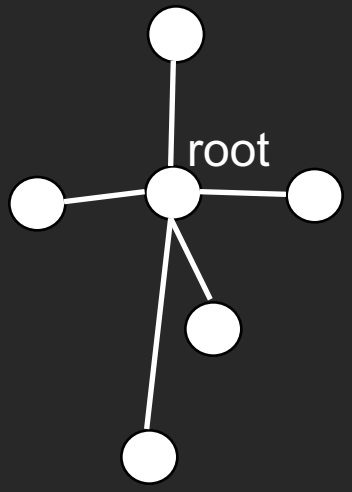
Source: Deva Ramanan

Inference: $\max_z$ score(x,z)

Felzenszwalb & Huttenlocher 05







Star model: the location of the root filter is the anchor point
Given the root location, all part locations are independent

Source: Deva Ramanan

Classification

$f_w(x)=w \cdot \Phi(x)$

$f_w(x)>0$

Source: Deva Ramanan

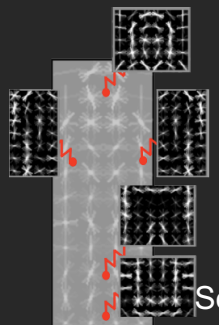
Latent-variable classification

$f_w(x) = w \cdot \Phi(x)$

$f_w(x) > 0$

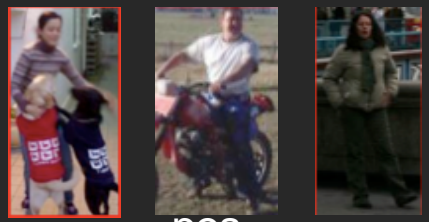
$f_w(x) = \max_z S(x,z)$

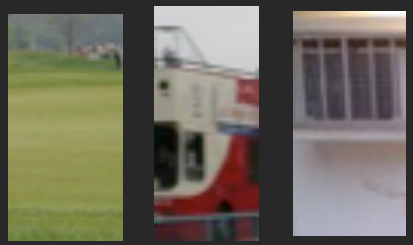
$= \max_z w \cdot \Phi(x, z)$



Source: Deva Ramanan

Latent SVMs


pos


neg

Given positive and negative training windows {xn}

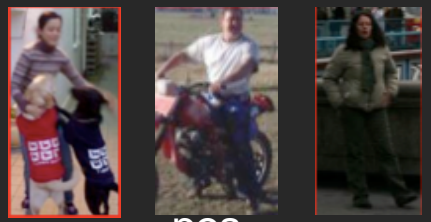
$$L(w) = ||w||^2 + \sum_{n \in \text{pos}} \max(0, 1 - f_w(x_n)) + \sum_{n \in \text{neg}} \max(0, 1 + f_w(x_n))$$

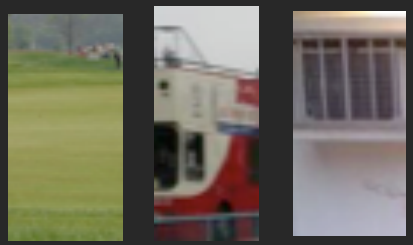
$$f_w(x) = \max_z w \cdot \Phi(x, z)$$

L(w) is “almost” convex

Source: Deva Ramanan

Latent SVMs


pos


neg

Given positive and negative training windows {xn}

$$L(w) = ||w||^2 + \sum_{n \in \text{pos}} \max(0, 1 - f_w(x_n)) + \sum_{n \in \text{neg}} \max(0, 1 + f_w(x_n))$$

$$w \cdot \Phi(x_n, z_n)$$

$$f_w(x) = \max_z w \cdot \Phi(x, z)$$

L(w) is convex if we fix latent values for positives

Source: Deva Ramanan

Coordinate descent

1) Given positive part locations, learn w with a convex program

$$w = \underset{w}{\operatorname{argmin}} L(w) \quad \text{with fixed} \quad \{z_n : n \in \text{pos}\}$$

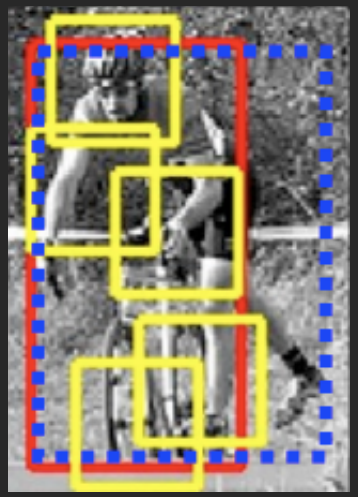
2) Given w, estimate part locations on positives

$$z_n = \underset{z}{\operatorname{argmax}}\, w \cdot \Phi(x_n, z) \quad \forall n \in \text{pos}$$

The above steps perform coordinate descent on a joint loss

Source: Deva Ramanan

Treat ground-truth labels as partially latent



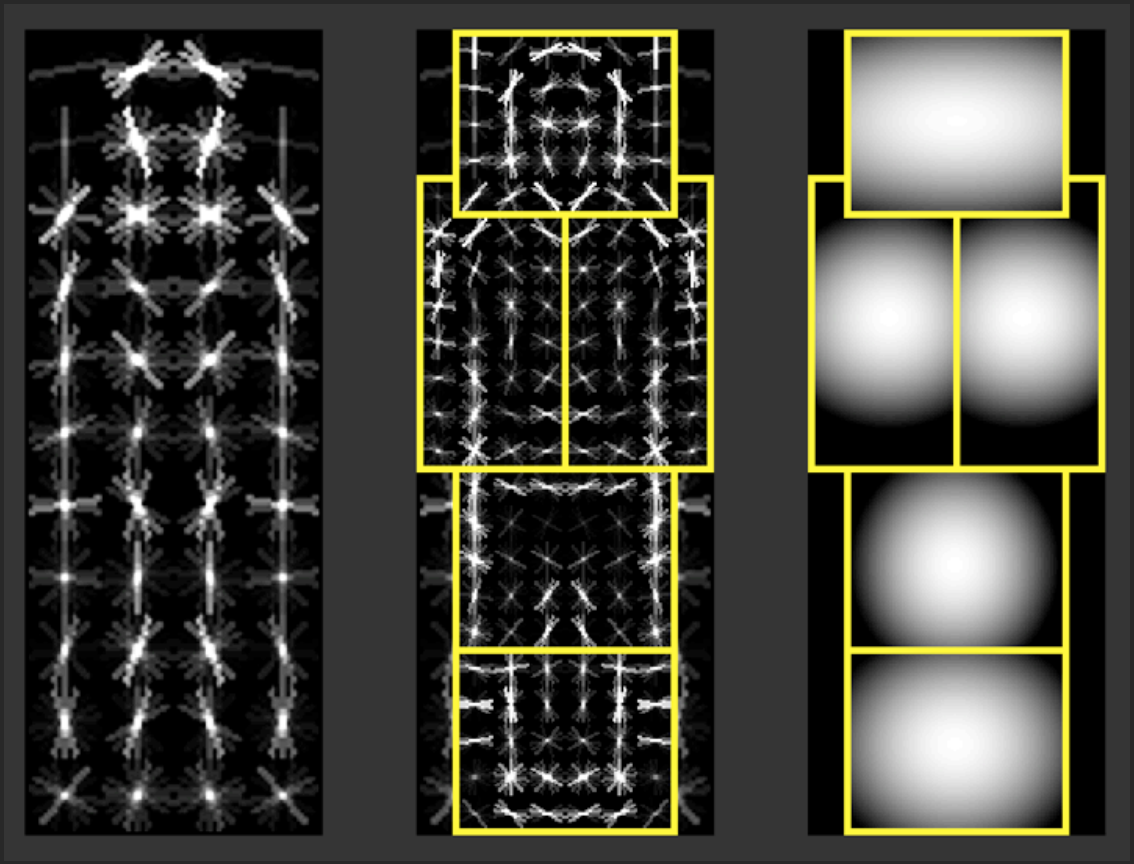
Allows for “cleaning up” of noisy labels (in blue) during iterative learning

Source: Deva Ramanan

Initialization

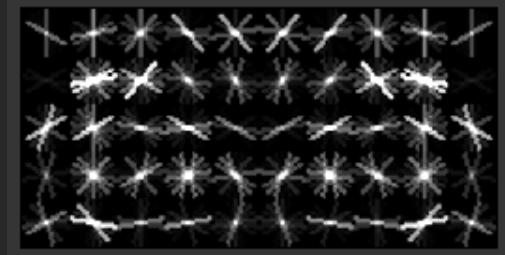
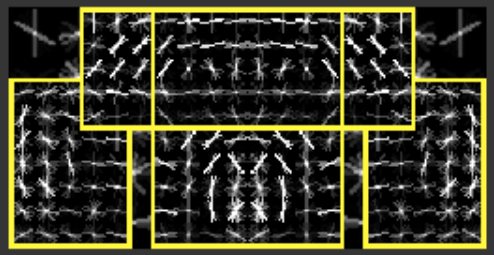
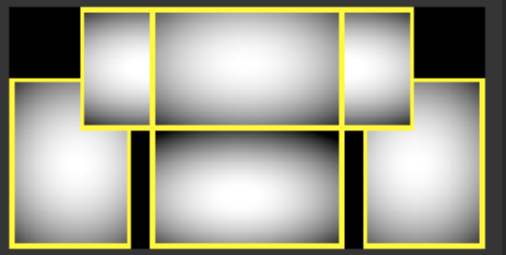
Learn root filter with SVM

Initialize part filters to regions in root filter with lots of energy

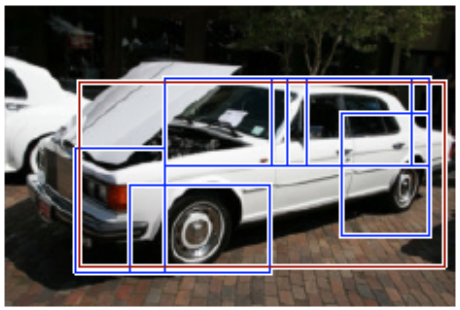
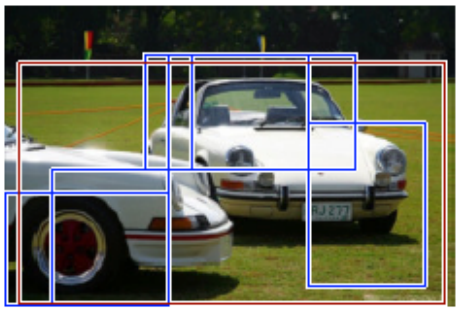
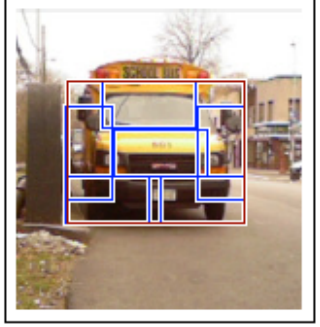


Source: Deva Ramanan

Example models

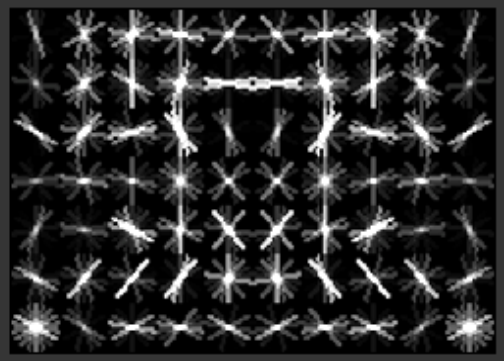




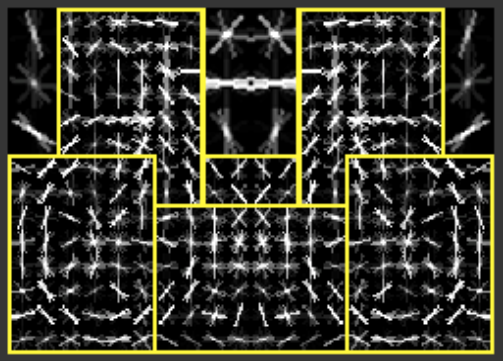


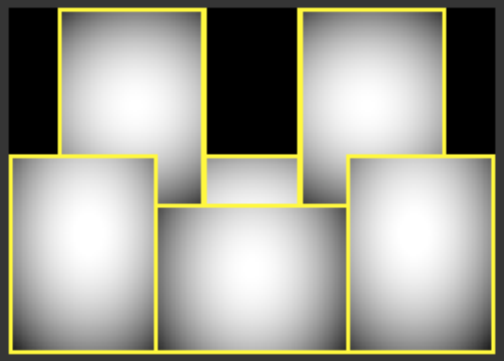




Source: Deva Ramanan

Example models

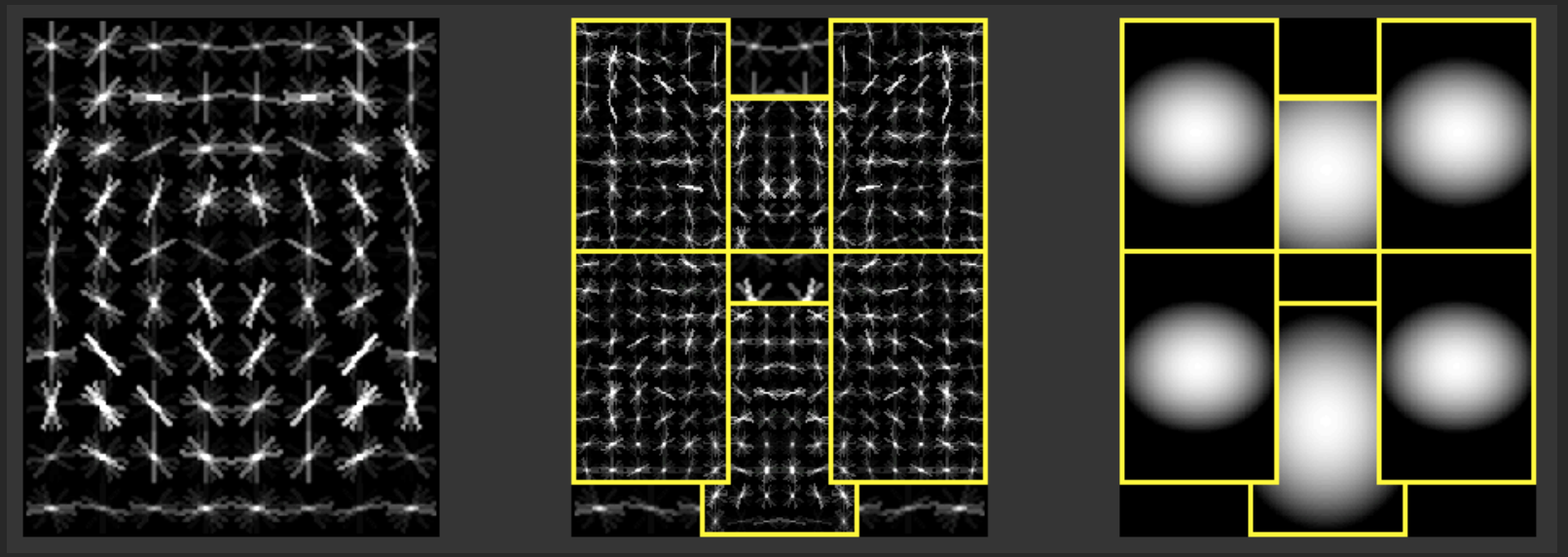






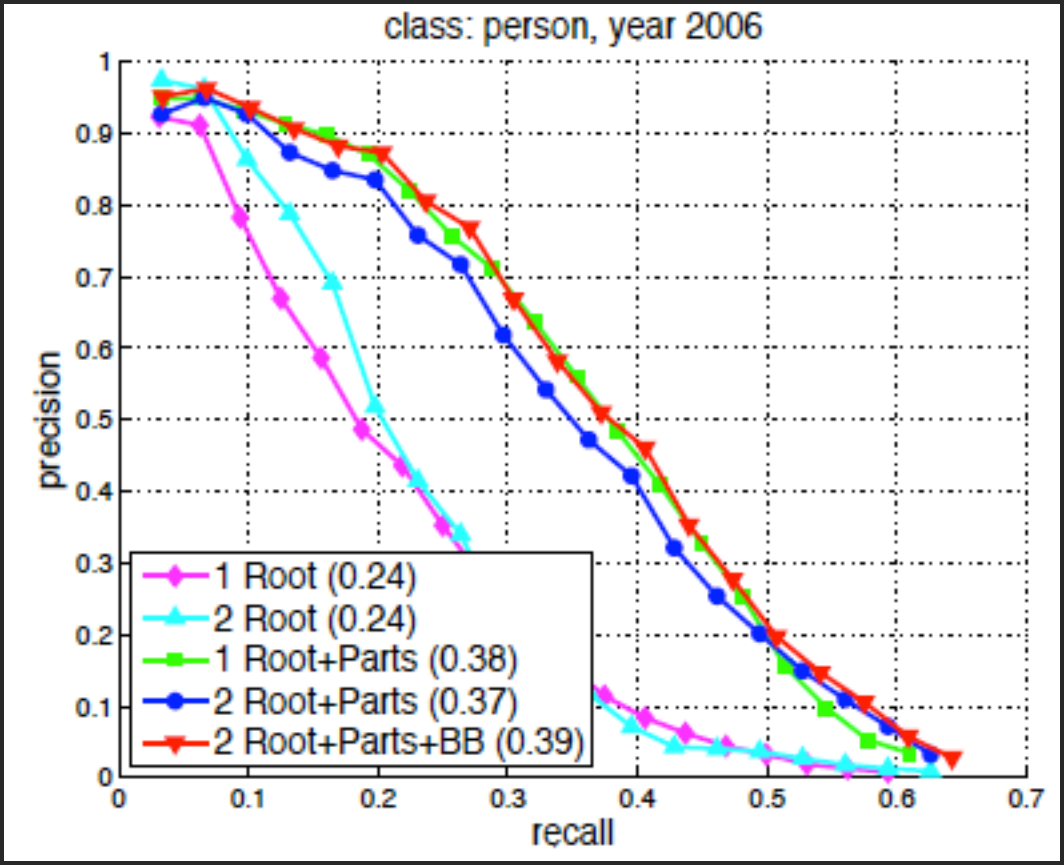
Source: Deva Ramanan

Example models



False positive due to imprecise bounding box

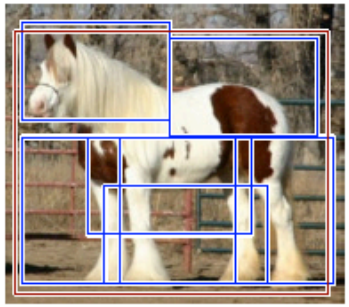
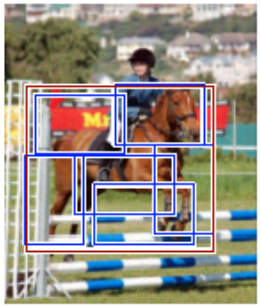
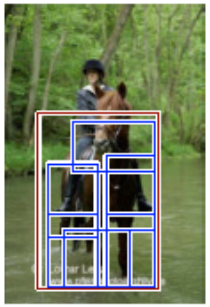
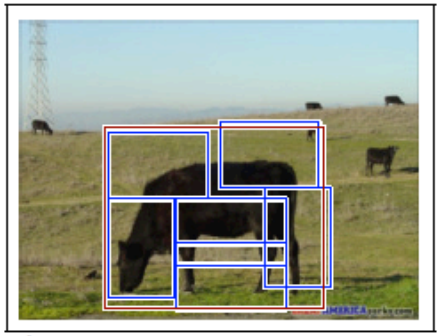
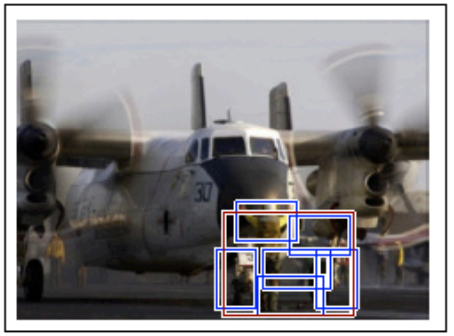
Source: Deva Ramanan



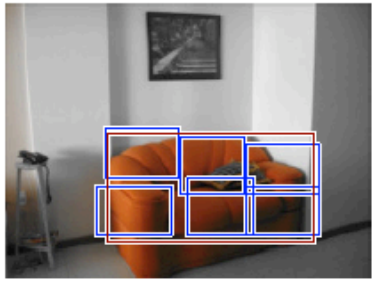
Other tricks:

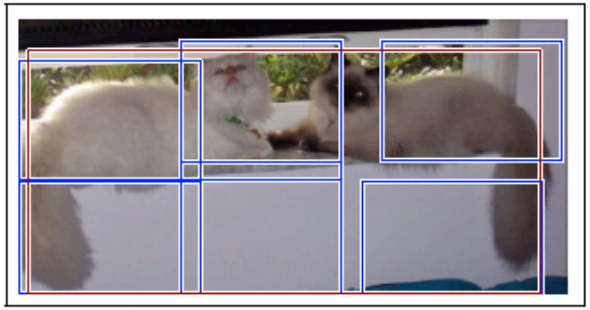
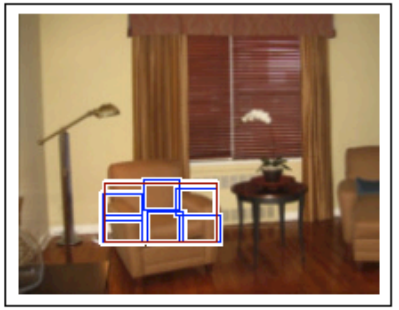
- Mining hard negative examples
- Noisy annotations

horse

sofa

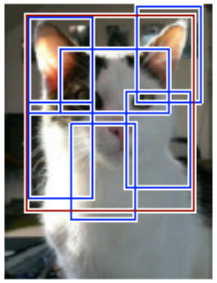
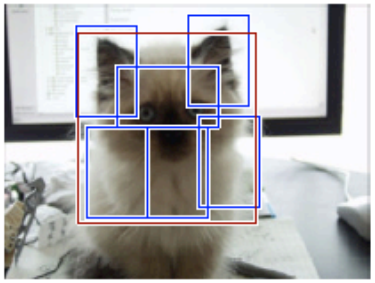
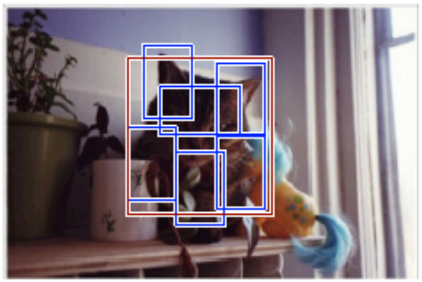
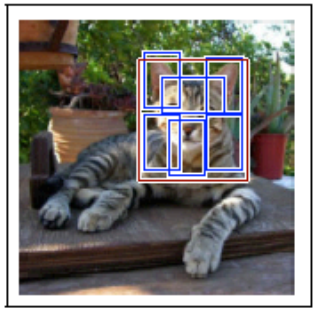
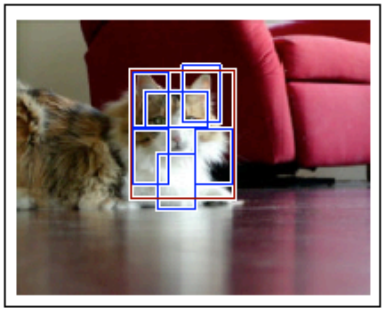


bottle




cat

Scanning window approach

Starting from object proposals

Selective Search for Object Recognition

J.R.R. Uijlings*[1,2], K.E.A. van de Sande†[2], T. Gevers[2], and A.W.M. Smeulders[2]

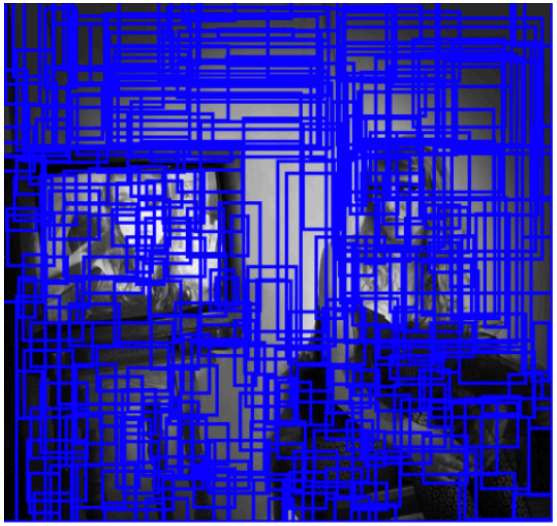
[1]University of Trento, Italy
[2]University of Amsterdam, the Netherlands

2011

http://www.huppelen.nl/publications/selectiveSearchDraft.pdf

Selective search

Input image



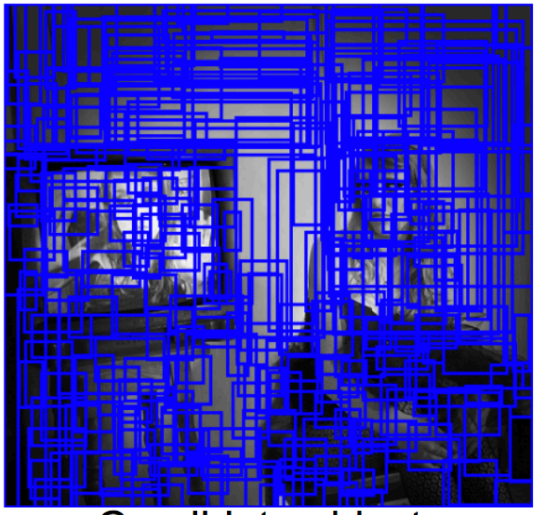
Candidate bounding boxes

Detected objects
(by applying classifier on candidate bb)

Selective search

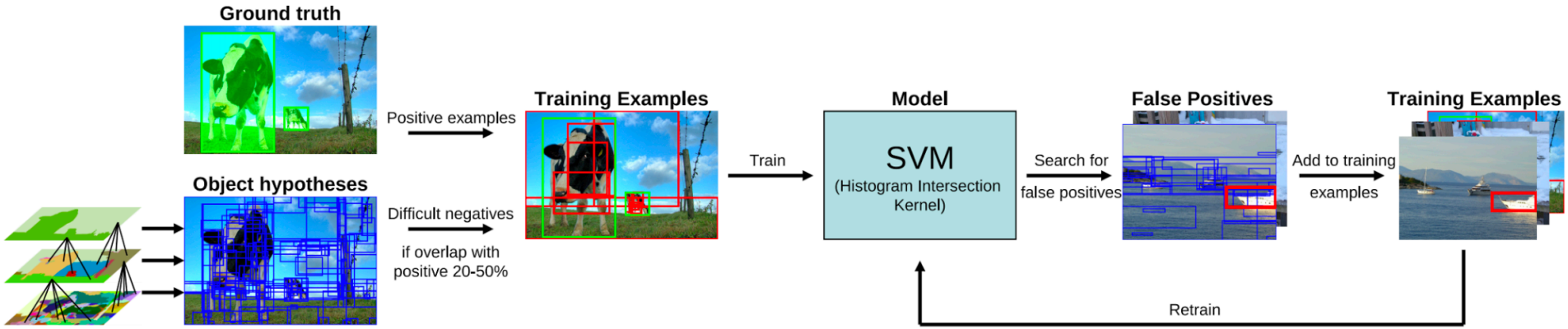
Input Image

Segmentation



Candidate objects

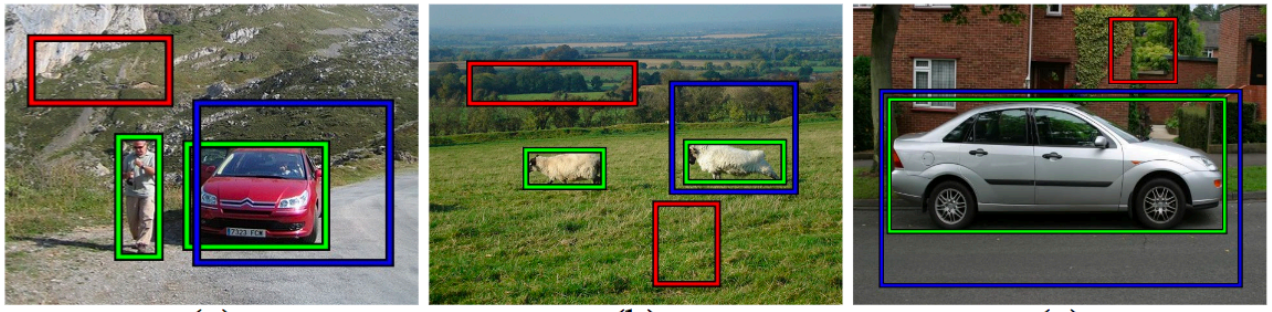
Training


Ground truth
Positive examples
Training Examples
Model
SVM
(Histogram Intersection Kernel)
Train
Search for
false positives
False Positives
Add to training
examples
Training Examples
Object hypotheses
Difficult negatives
if overlap with
positive 20-50%
Retrain

Removing the need for scanning

What is an object ?

Bogdan Alexe, Thomas Deselaers, Vittorio Ferrari
Computer Vision Laboratory, ETH Zurich
{bogdan, deselaers, ferrari}@vision.ee.ethz.ch



(a) (b) (c)

Fig. 1: **Desired behavior of an objectness measure.** *The desired objectness measure should score the blue windows, partially covering the objects, lower than the ground truth windows (green), and score even lower the red windows containing only stuff or small parts of objects.*

http://groups.inf.ed.ac.uk/calvin/objectness/

The limit of hand-cracted features

HOGgles: Visualizing Object Detection Features*

Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, Antonio Torralba
Massachusetts Institute of Technology
{vondrick,khosla,tomasz,torralba}@csail.mit.edu

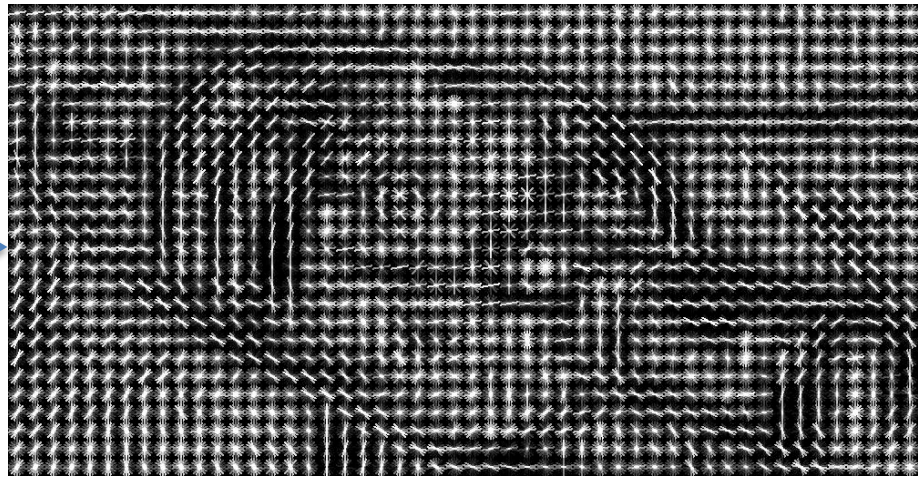
ICCV 2013

http://carlvondrick.com/ihog/iccv.pdf

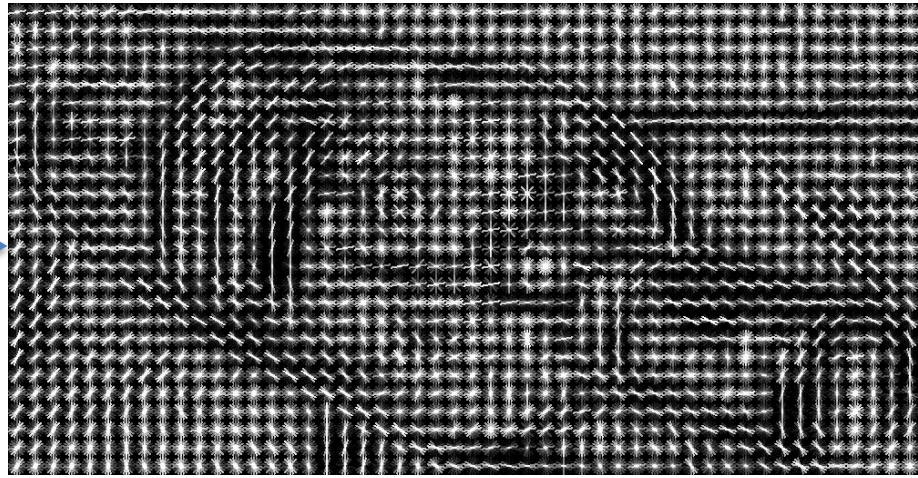

car

What does a detector sees though HOG eyes?

What does a detector sees though HOG eyes?



What does a detector sees though HOG eyes?

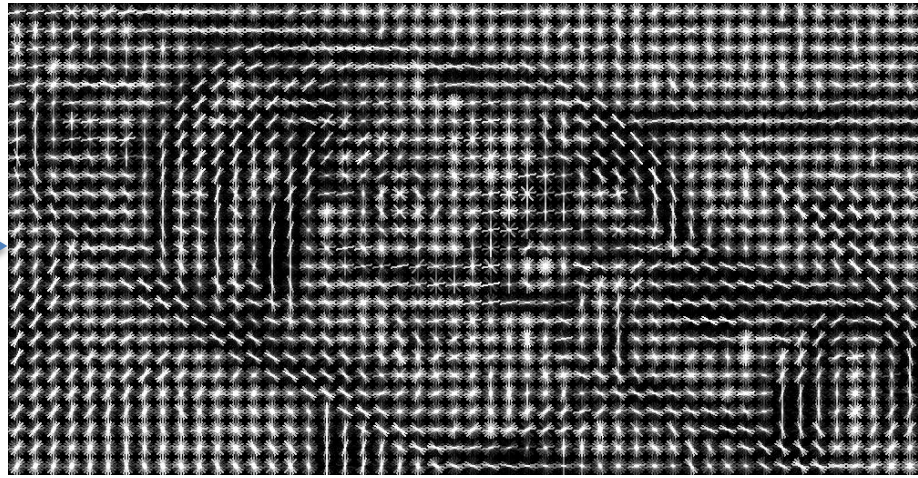


But, what is lost in this transformation?

Carl Vondrick A. Khosla

Vondrick, Khosla, Malisiewicz, Torralba. "Inverting and Visualizing Features for Object Detection."

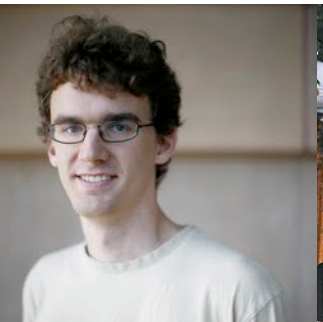
What does a detector sees though HOG eyes?



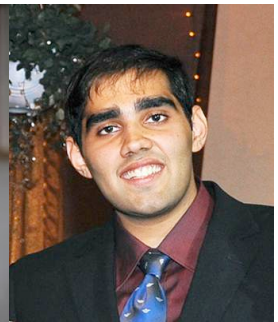
?

Can we recover the input image from HOG?

Carl Vondrick

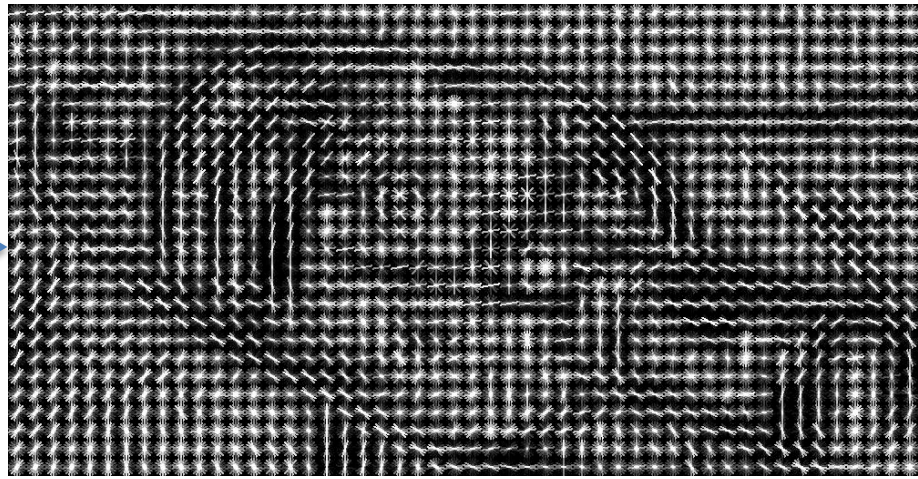


A. Khosla



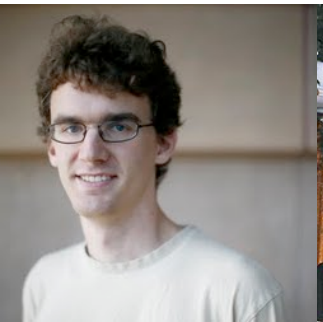
Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”

What does a detector sees though HOG eyes?

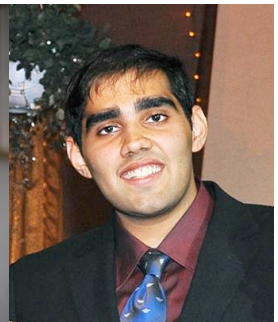


?

Carl Vondrick



A. Khosla



Can we recover the input image from HOG?

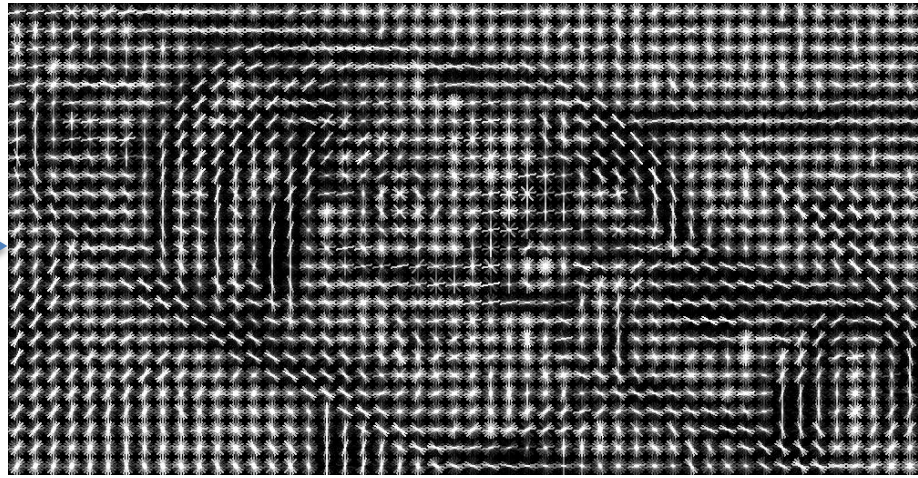
$$\phi^{-1}(y) = \underset{x \in \mathbb{R}^D}{\operatorname{argmin}} ||\phi(x) - y||_2^2$$

HOG

Recovered image

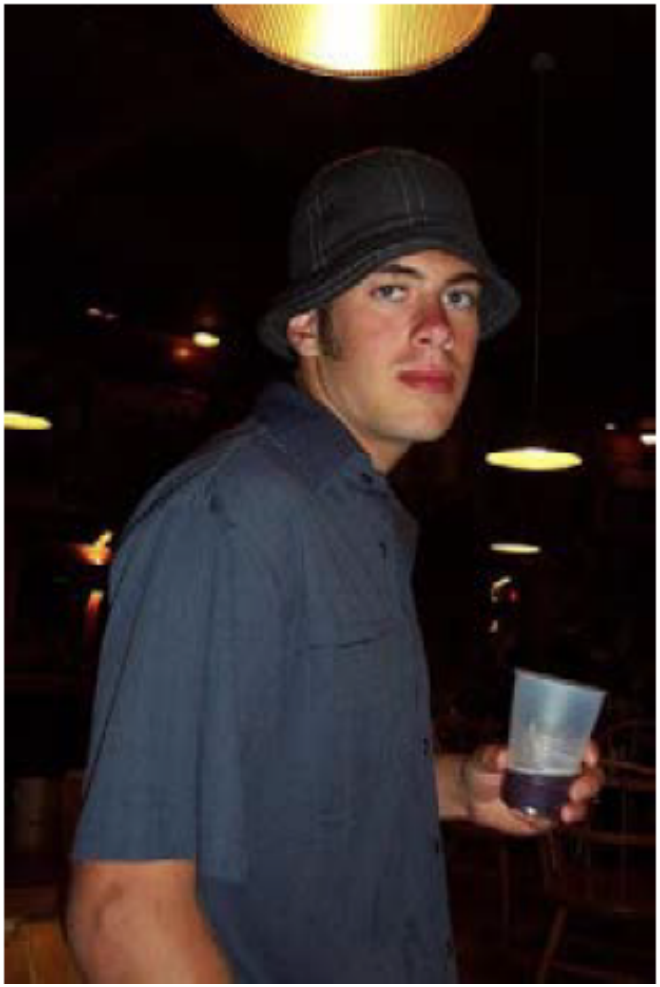
Vondrick, Khosla, Malisiewicz, Torralba. "Inverting and Visualizing Features for Object Detection."

What does a detector sees though HOG eyes?



Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”

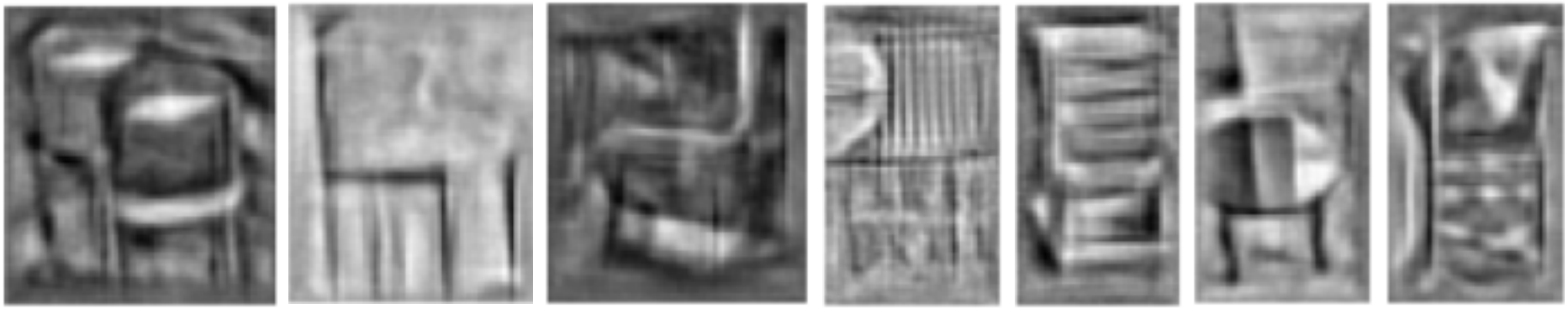
What does a detector sees though HOG eyes?



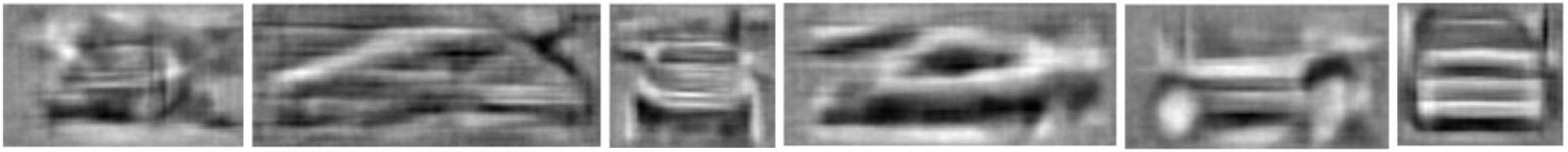
Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”

Person

Chair



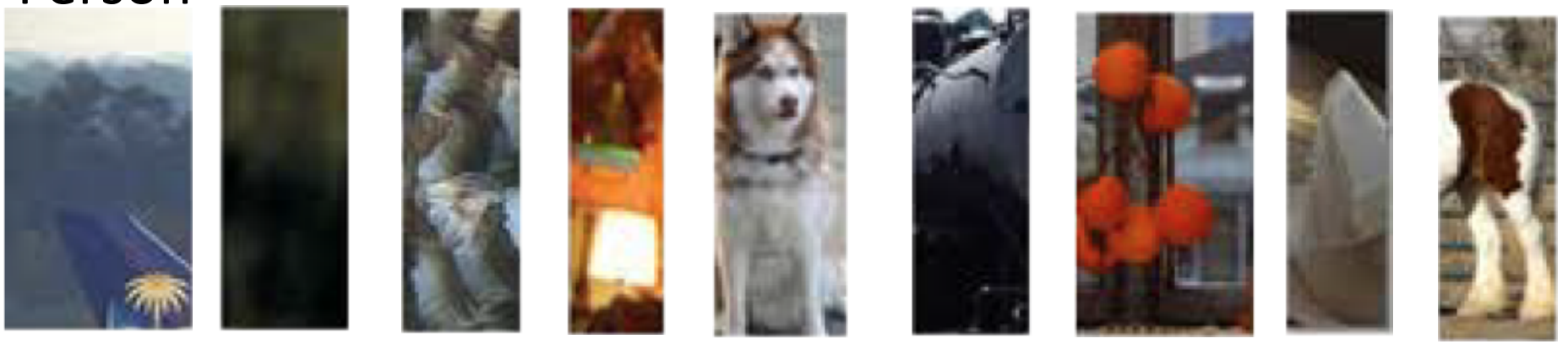
Car



Can you tell which ones are the false alarms?

Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”

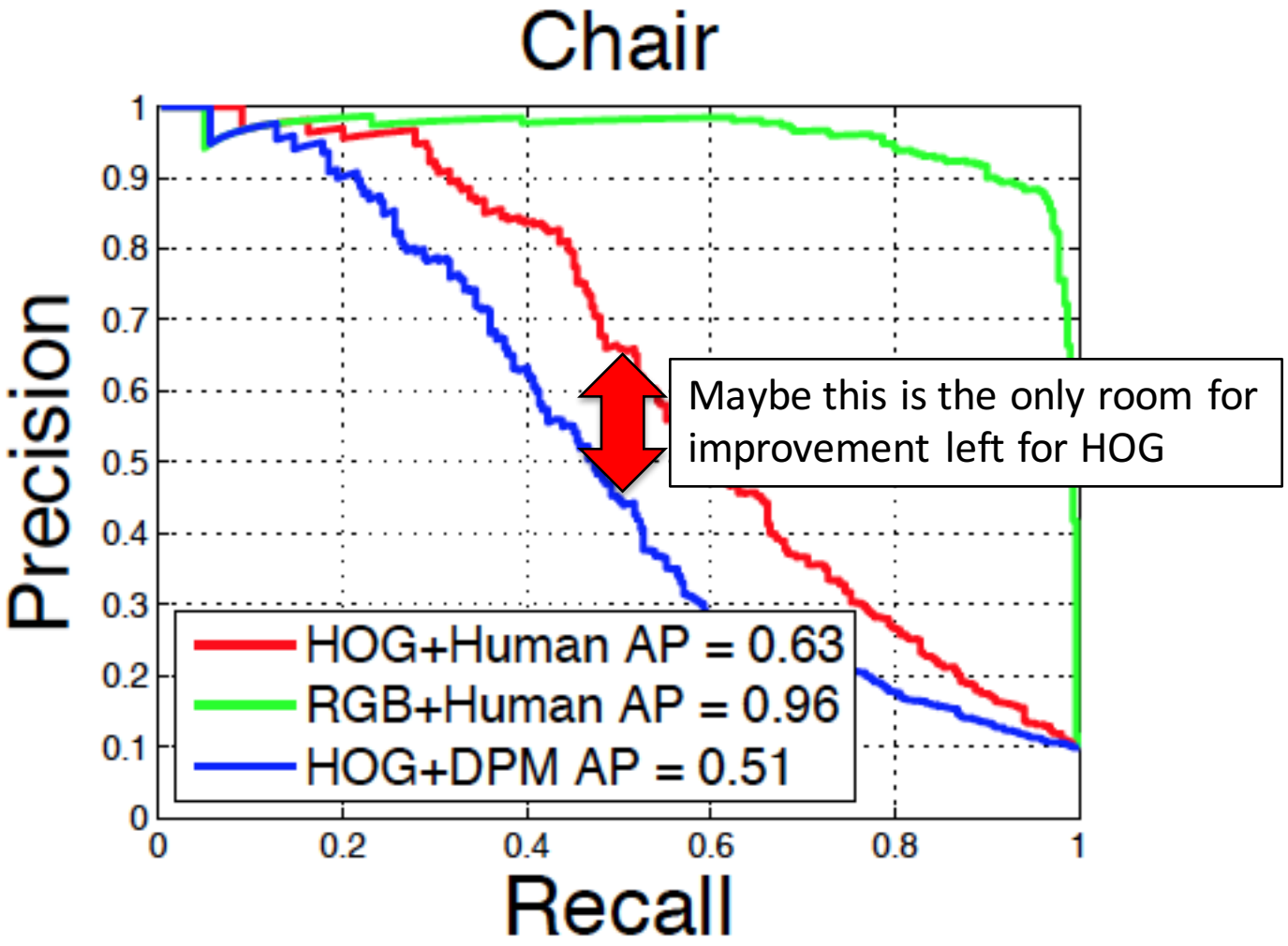
Person



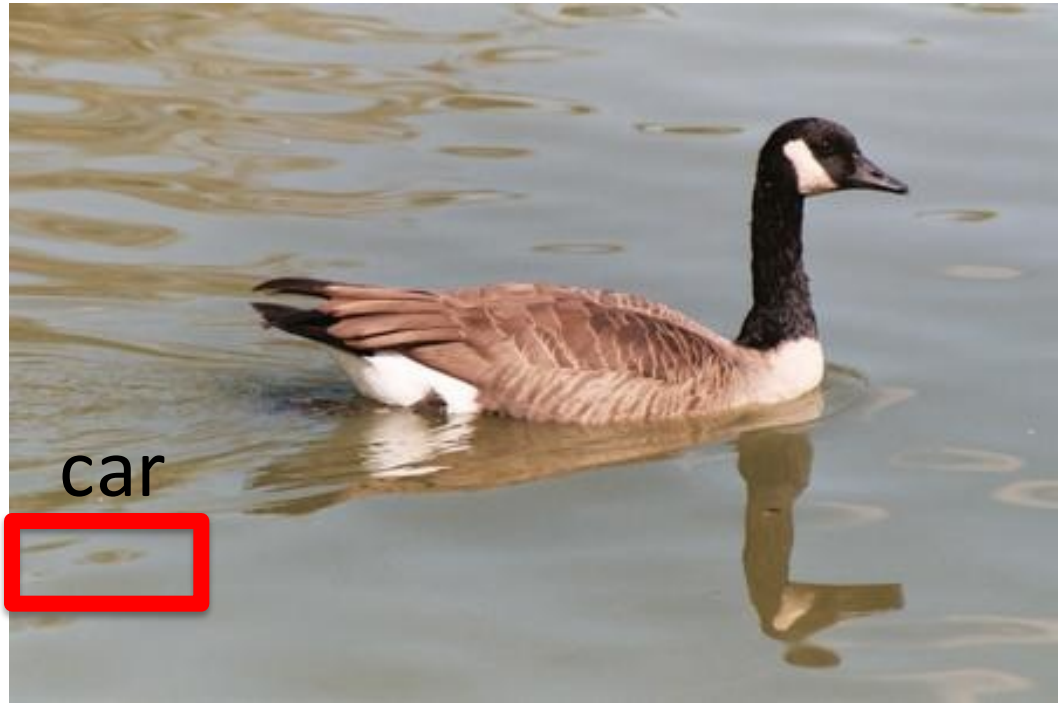
Chair

Car

Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”

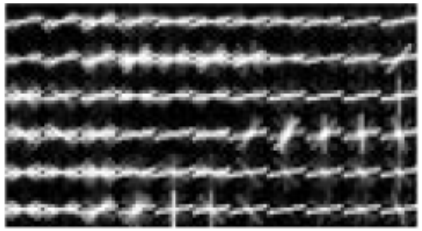


Vondrick, Khosla, Malisiewicz, Torralba. "Inverting and Visualizing Features for Object Detection."

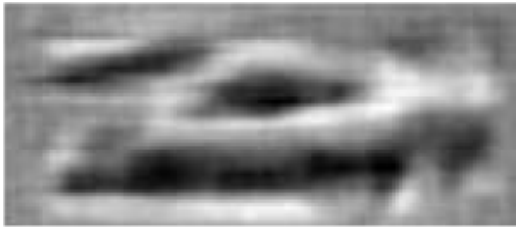


The image patch

HOG



What HOG sees

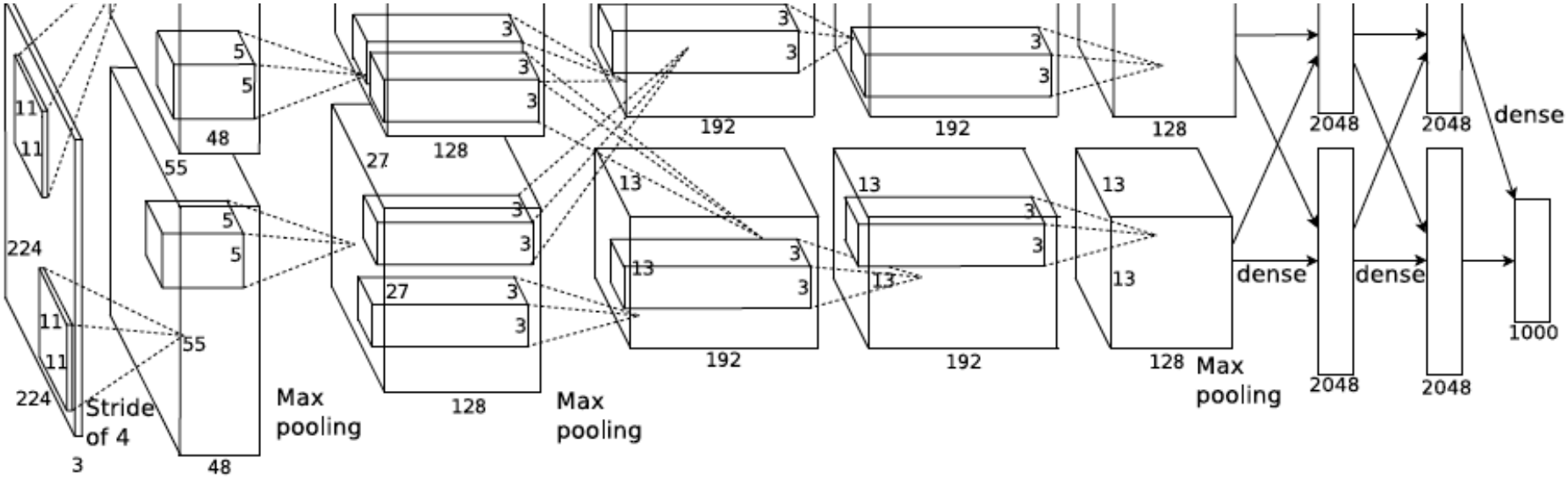


http://mit.edu/vondrick/ihog/

Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”

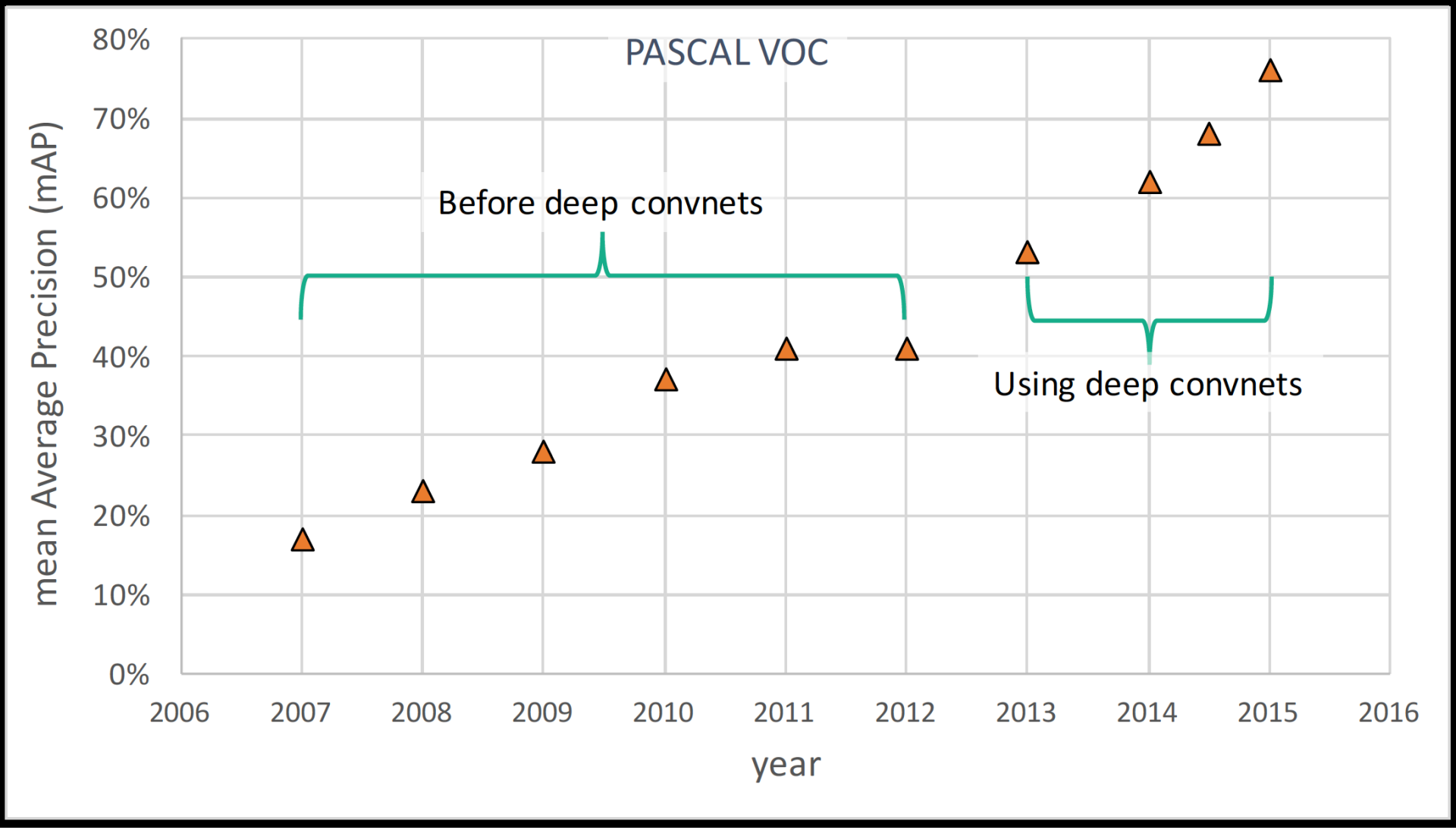
Krizhevsky et al. [NIPS2012]

- Same model as LeCun'98 but:
  - Bigger model (8 layers)
  - More data (10^6 vs 10^3 images)
  - GPU implementation (50x speedup over CPU)
  - Better regularization (DropOut)



- 7 hidden layers, 650,000 neurons, 60,000,000 parameters
- Trained on 2 GPUs for a week

Object detection renaissance (2013-present)

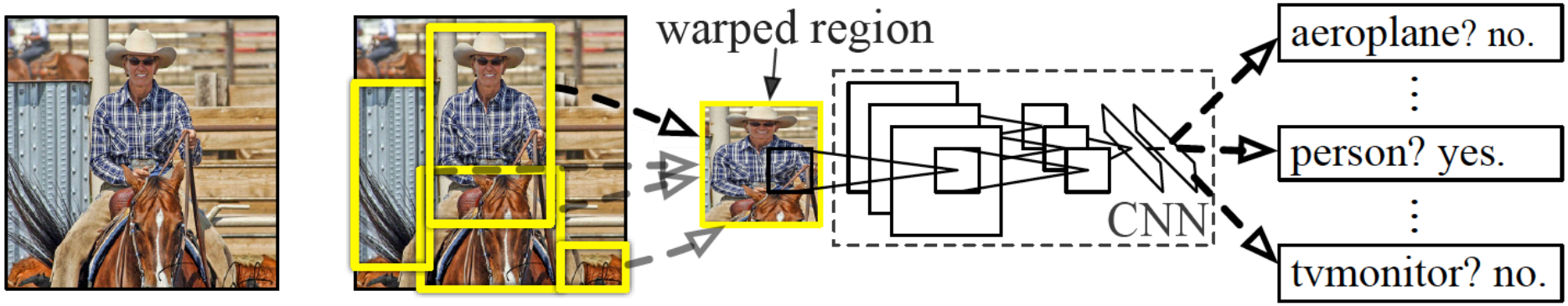


Slide credit: Ross Girshick

Rich feature hierarchies for accurate object detection and semantic segmentation

Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik
UC Berkeley
{rbg,jdonahue,trevor,malik}@eecs.berkeley.edu

CVPR 2014



https://dl.dropboxusercontent.com/s/293tu0hh9ww08co/r-cnn-cvpr.pdf?dl=0

Slow R-CNN

Input image

Girshick et al. CVPR14.

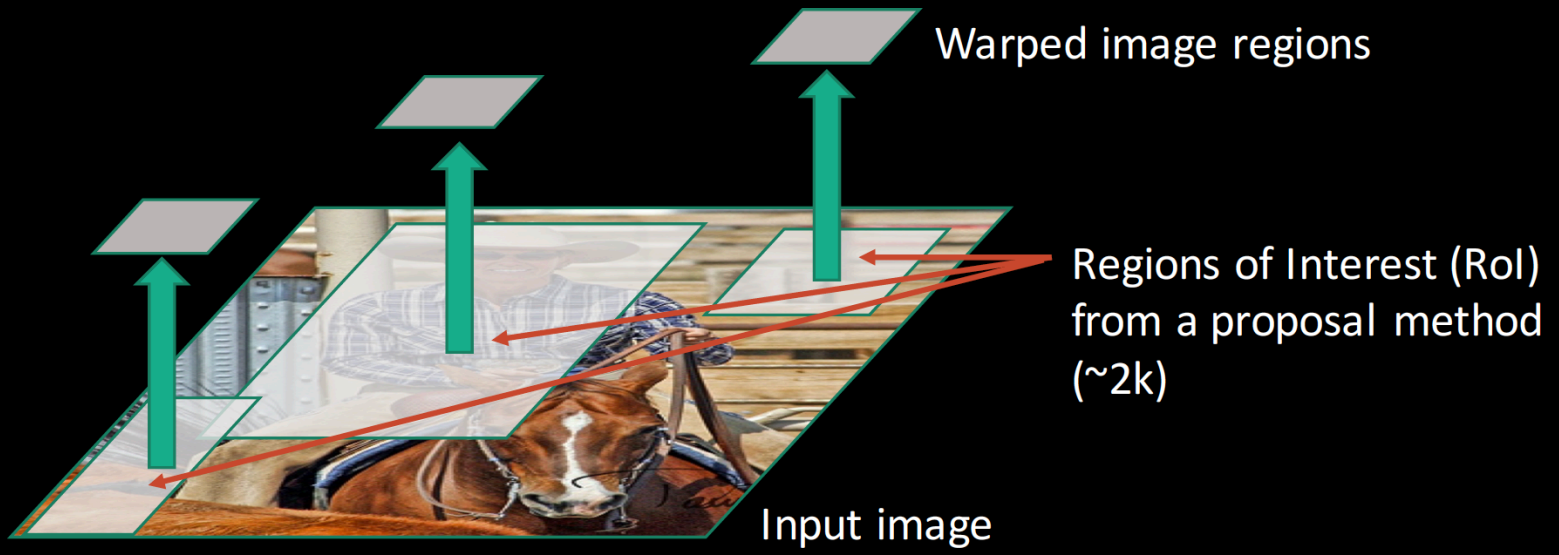
Slow R-CNN

Regions of Interest (RoI) from a proposal method (~2k)

Input image

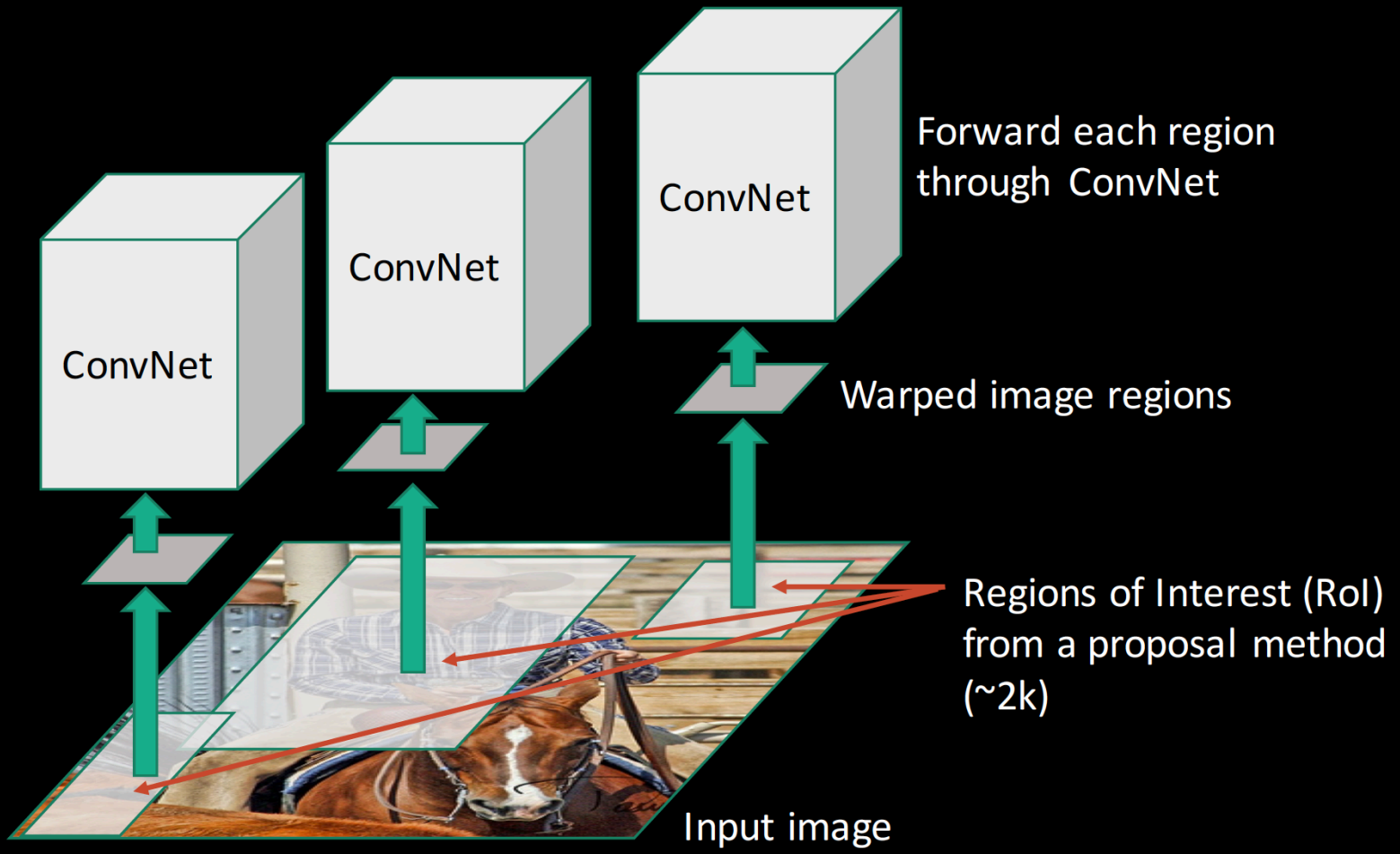
Girshick et al. CVPR14.

Slow R-CNN



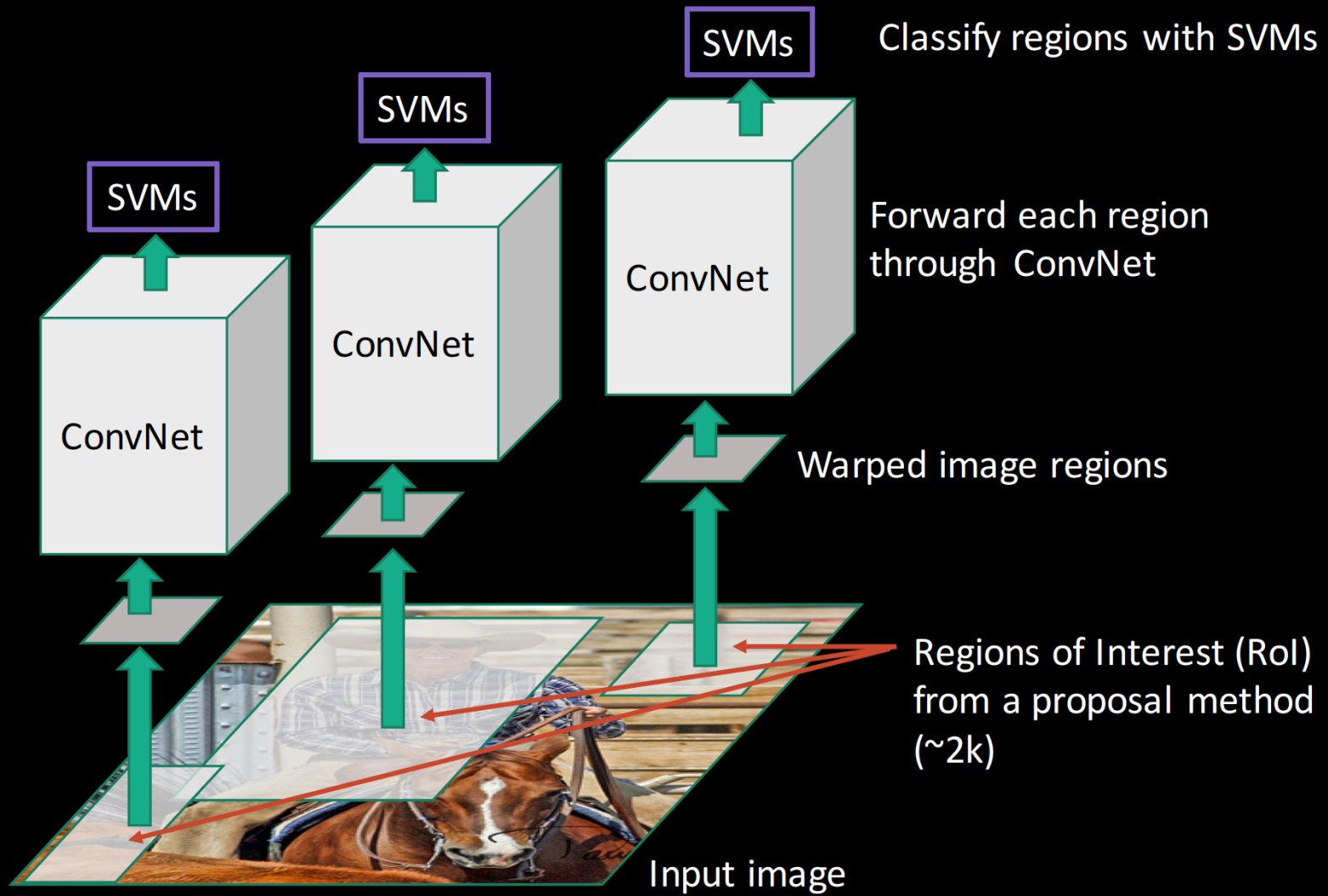
Girshick et al. CVPR14.

Slow R-CNN



Girshick et al. CVPR14.

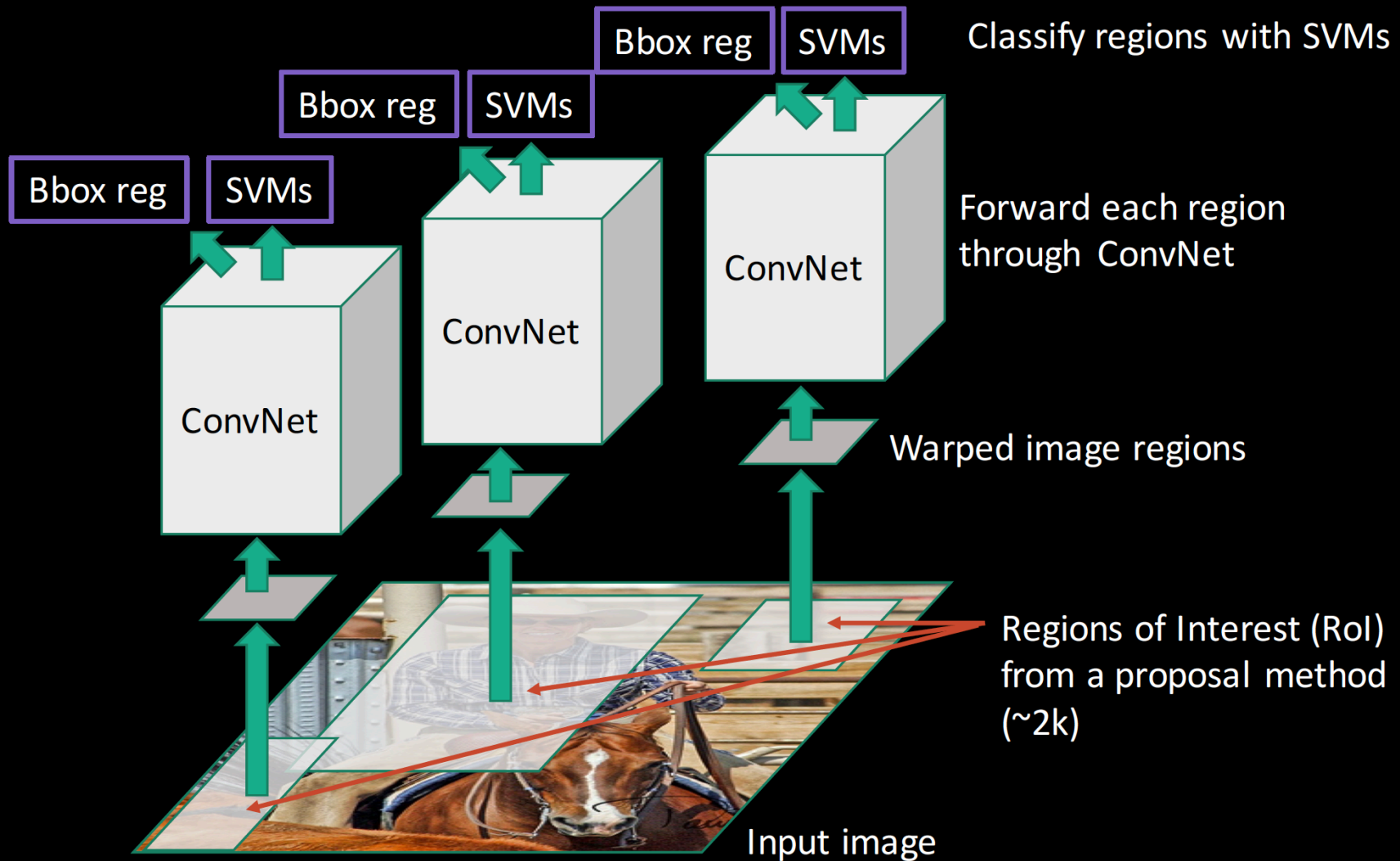
Slow R-CNN



Girshick et al. CVPR14.

Post hoc component

Slow R-CNN



Girshick et al. CVPR14.

Post hoc component

Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun

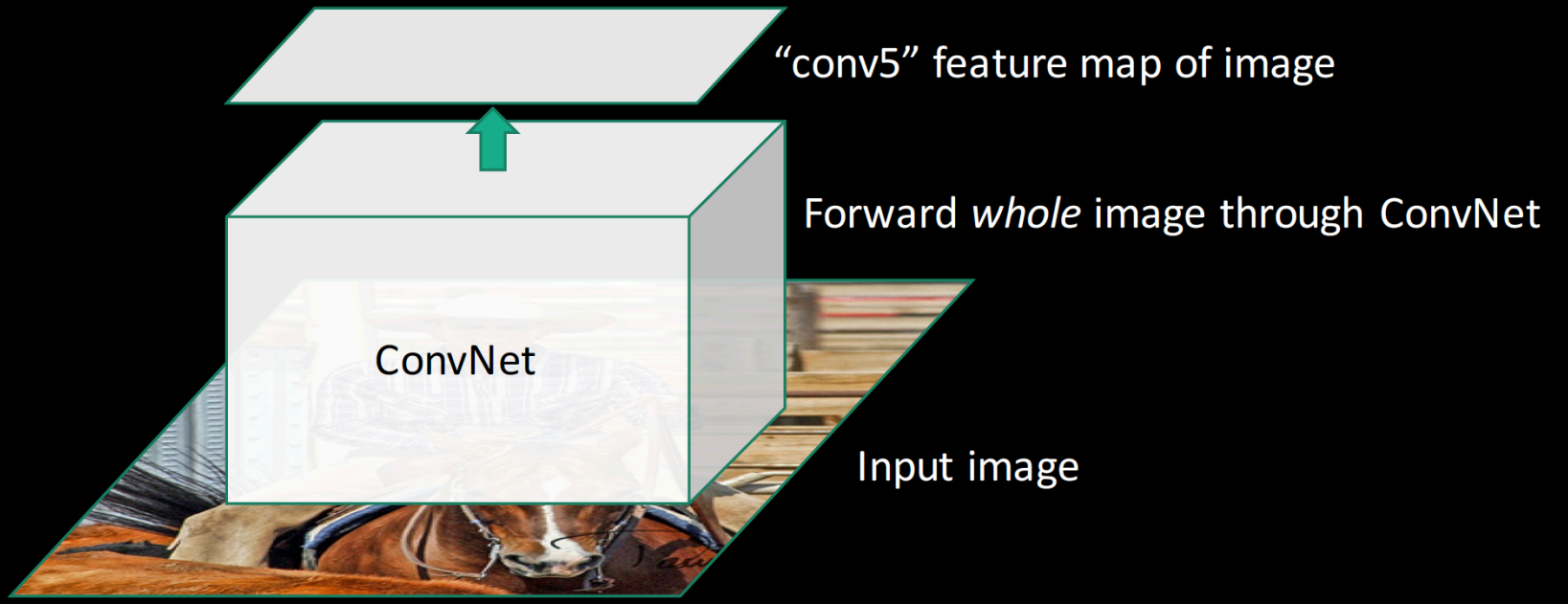
https://arxiv.org/pdf/1406.4729.pdf

SPP-net

Input image

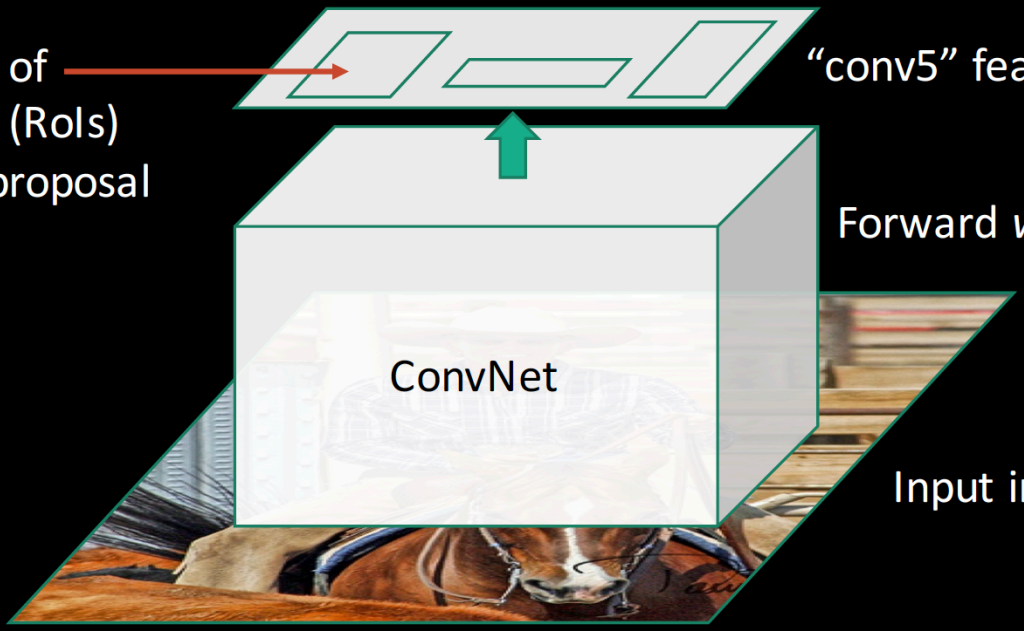
He et al. ECCV14.

SPP-net



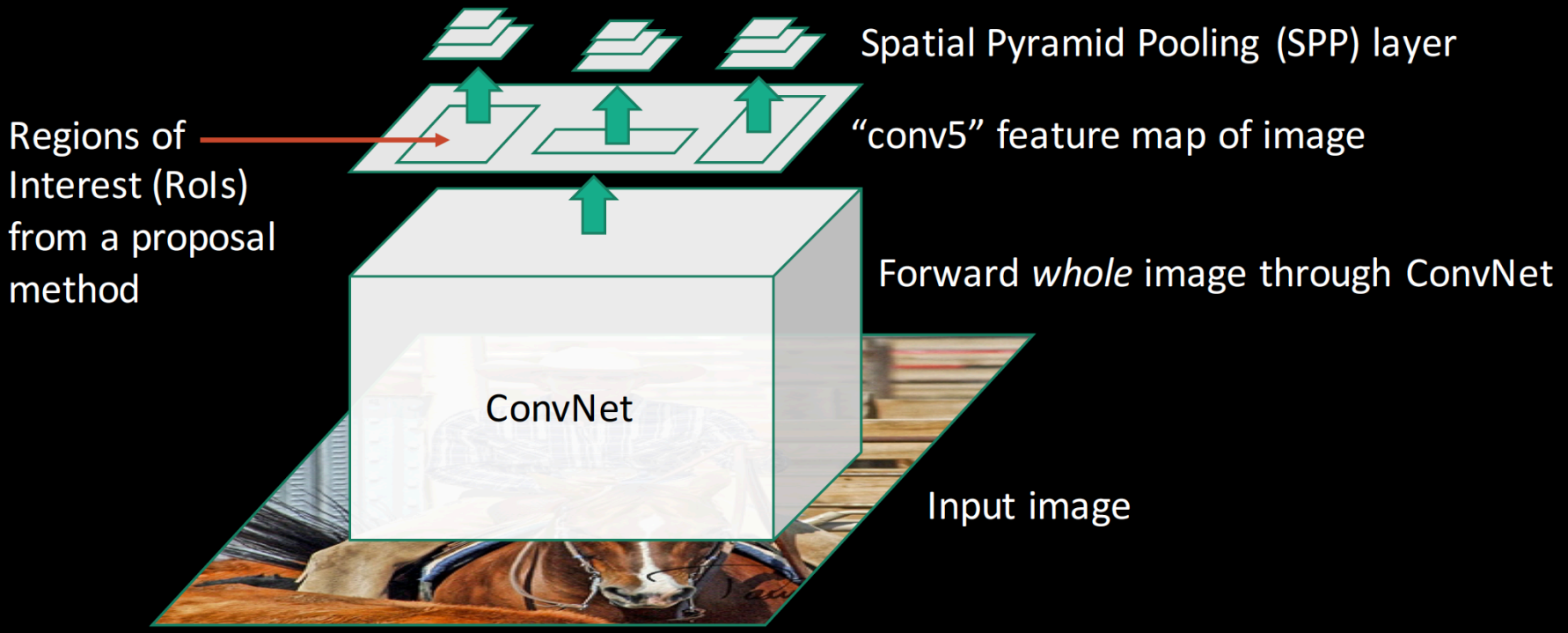
He et al. ECCV14.

SPP-net



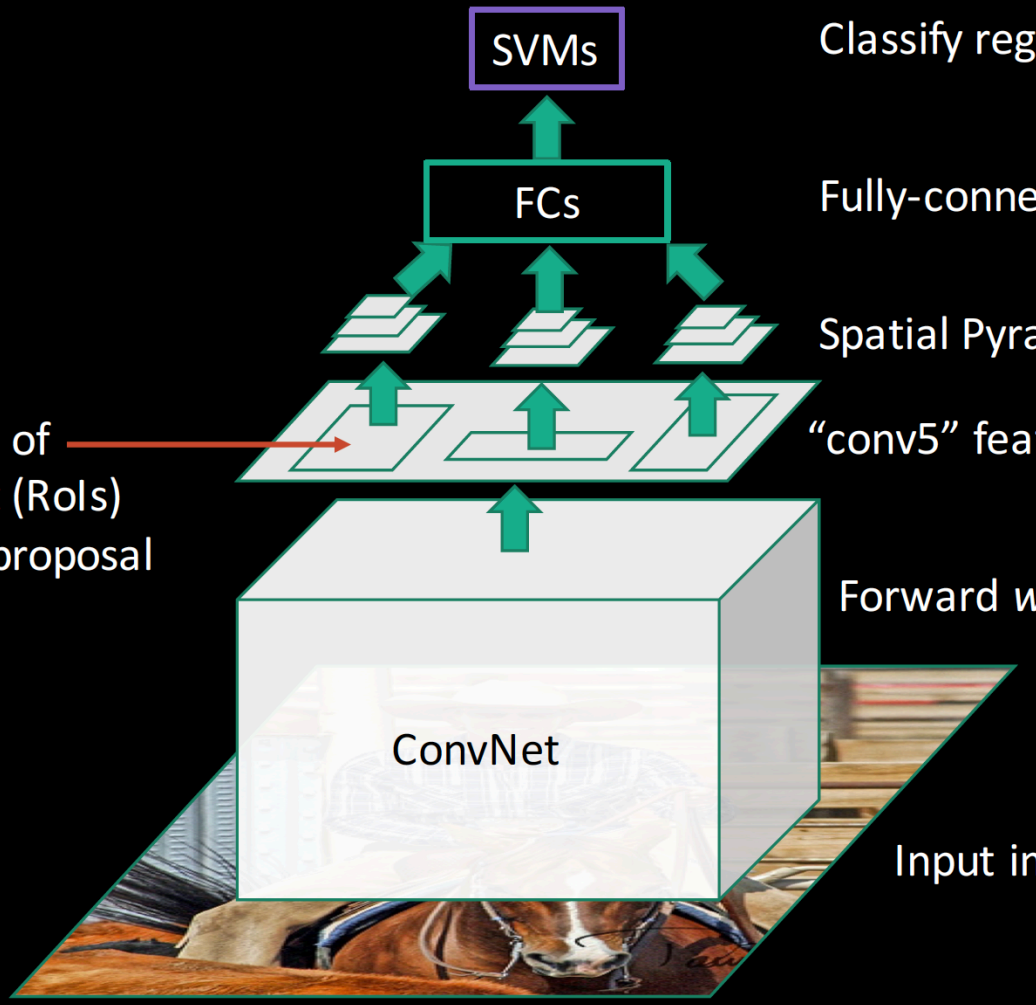
He et al. ECCV14.

SPP-net



He et al. ECCV14.

SPP-net



Classify regions with SVMs

Fully-connected layers

Spatial Pyramid Pooling (SPP) layer

"conv5" feature map of image

Regions of Interest (RoIs) from a proposal method

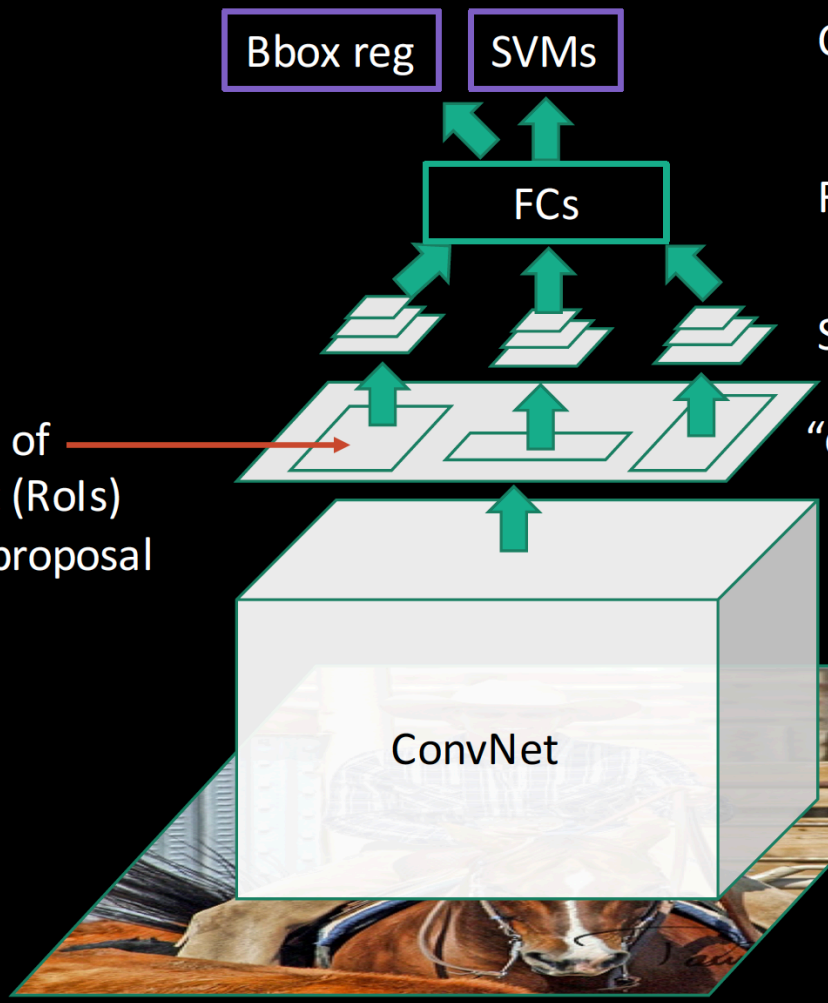
Forward *whole* image through ConvNet

Input image

He et al. ECCV14.

Post hoc component

SPP-net



Apply bounding-box regressors

Classify regions with SVMs

Fully-connected layers

Spatial Pyramid Pooling (SPP) layer

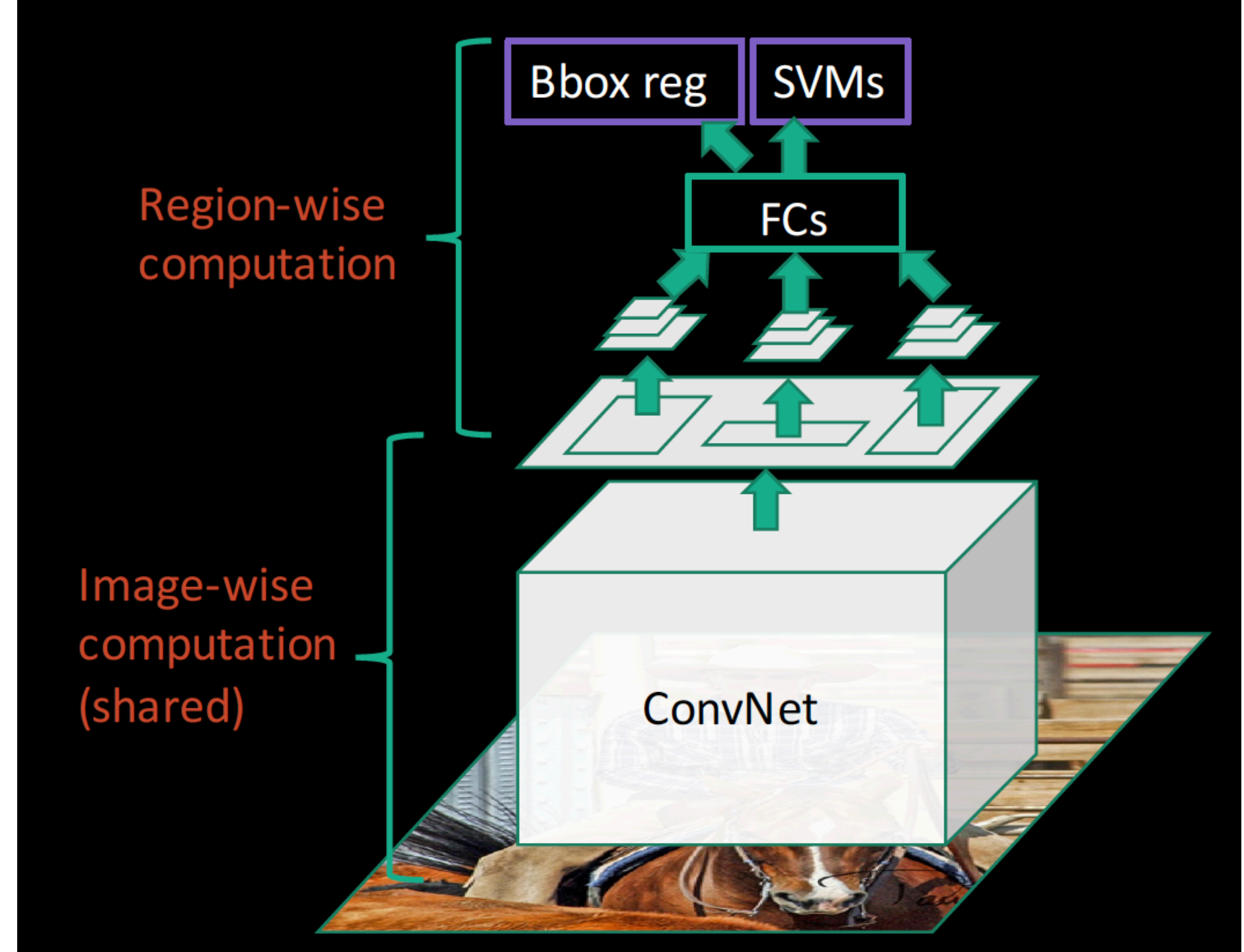
"conv5" feature map of image

Forward *whole* image through ConvNet

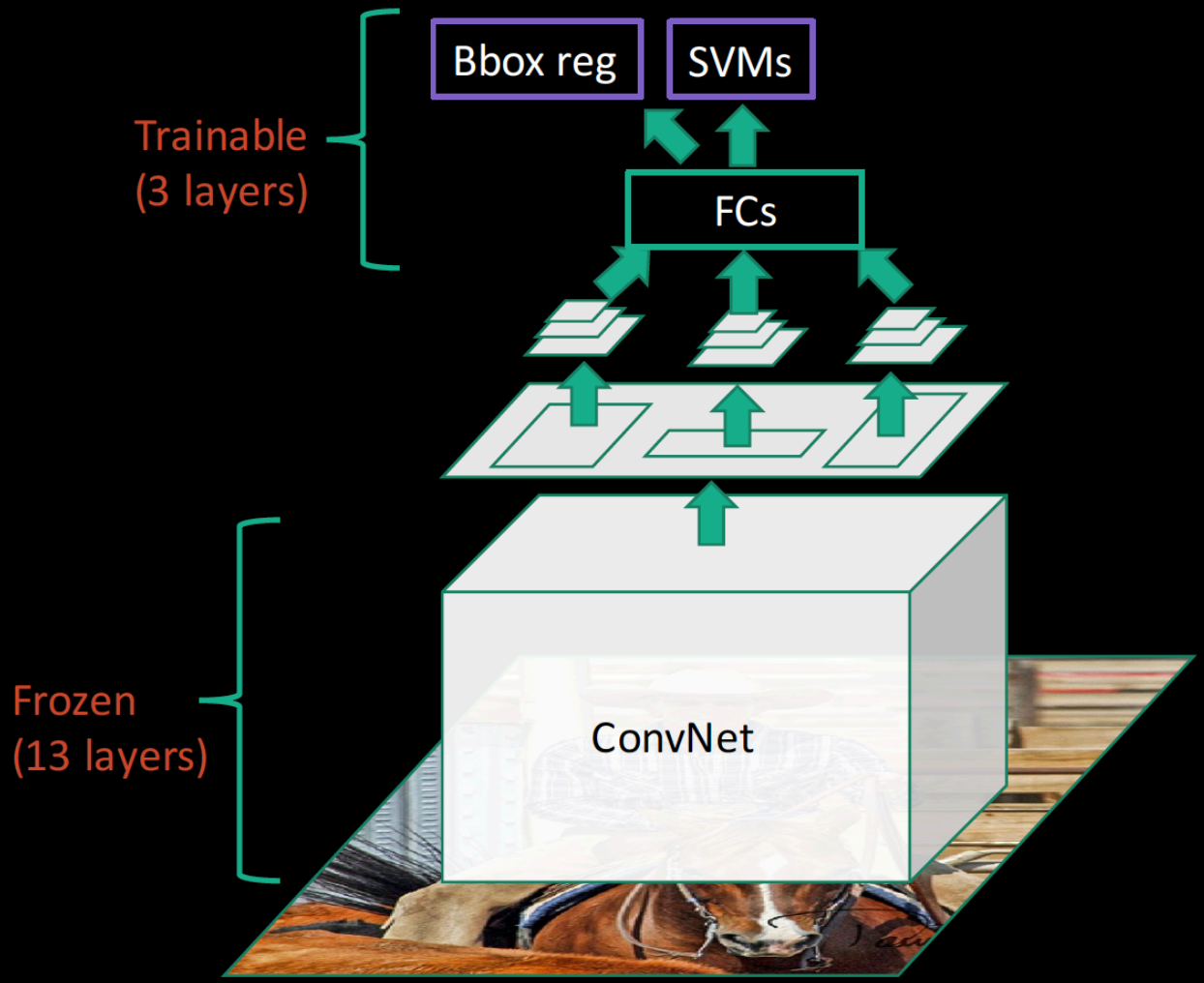
Input image

He et al. ECCV14.

Post hoc component


Bbox reg
SVMs
FCs
Region-wise computation
Image-wise computation (shared)
ConvNet

SPP-net: the main limitation



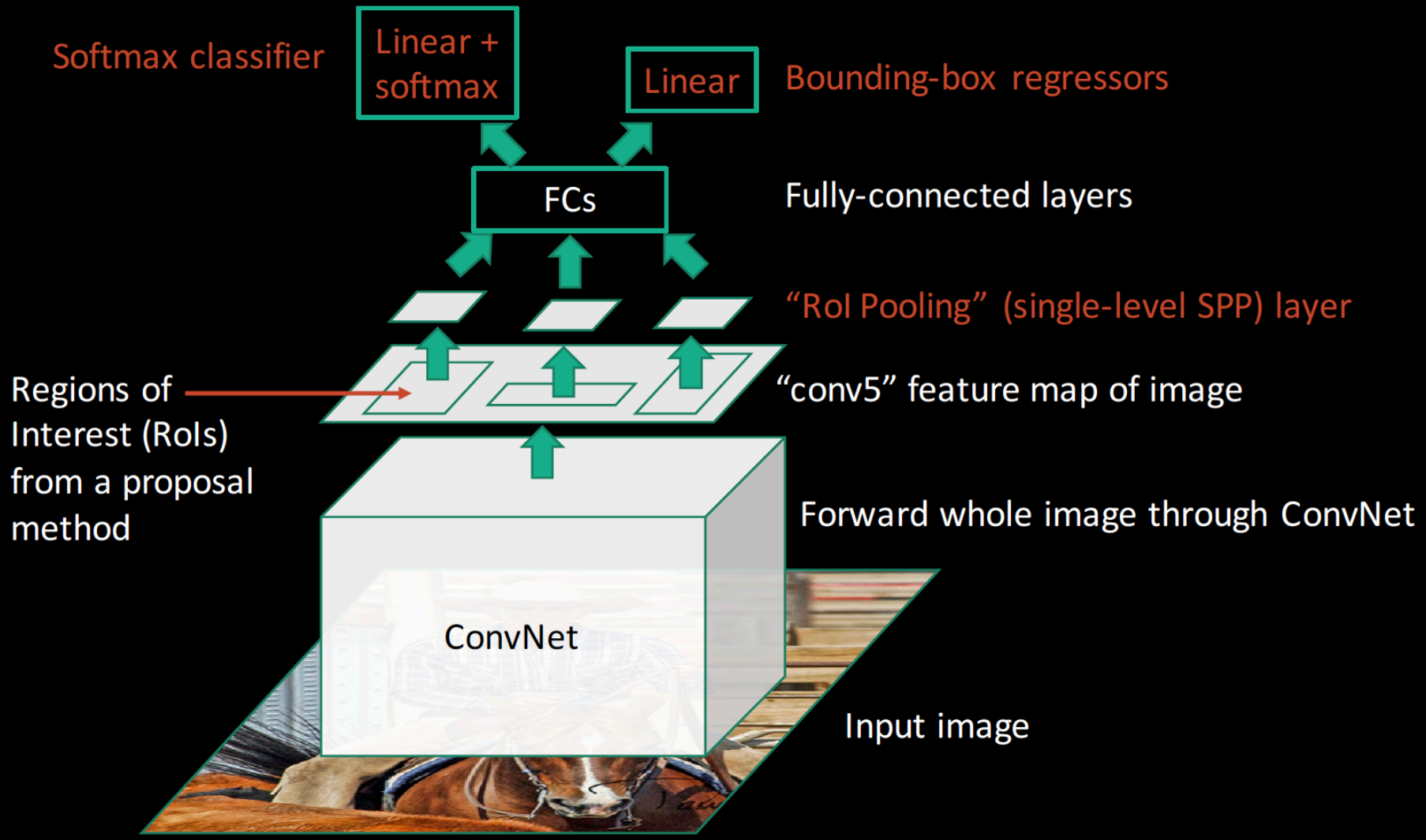
He et al. ECCV14.

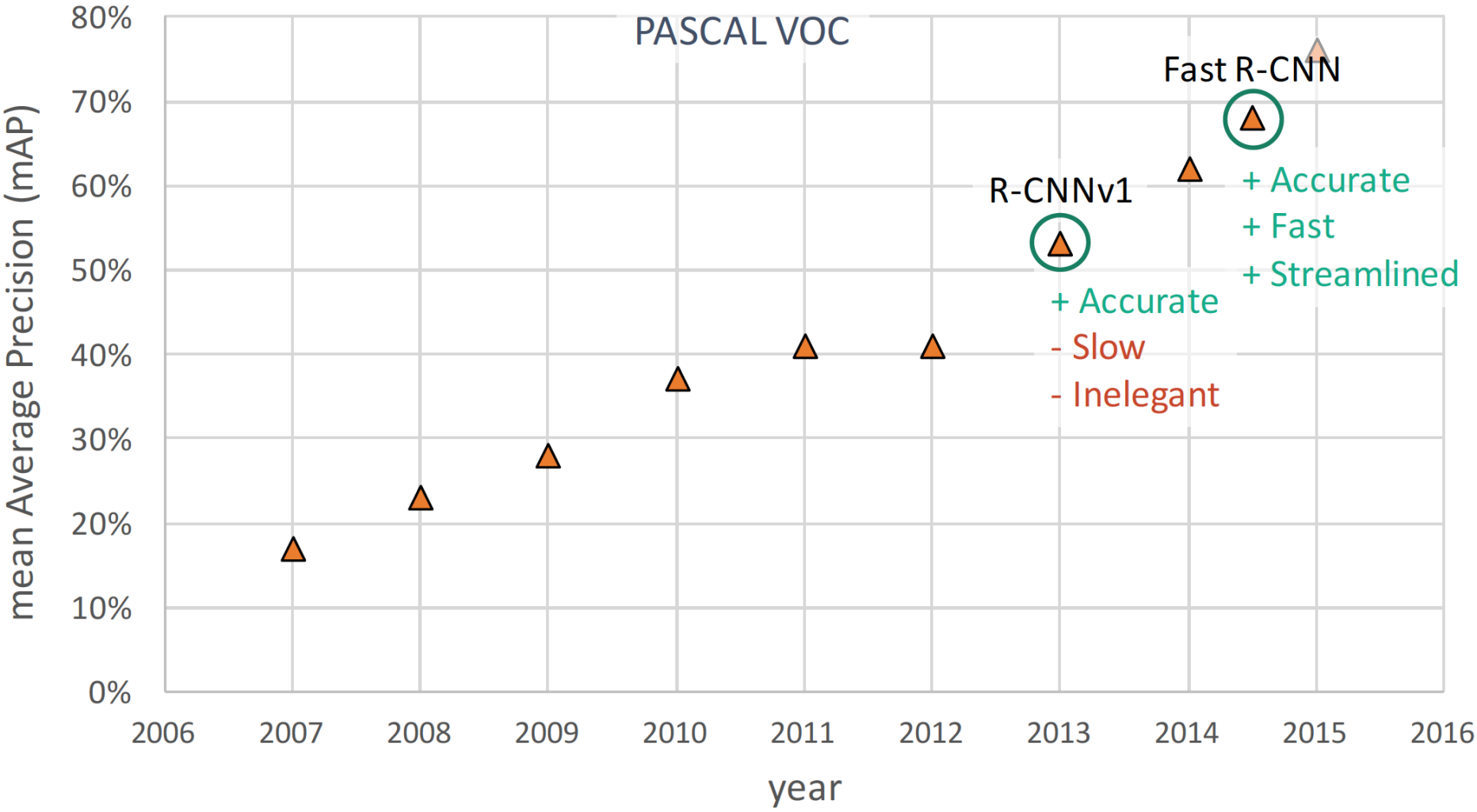
Post hoc component

Fast R-CNN

Ross Girshick
Microsoft Research
rbg@microsoft.com

Fast R-CNN (test time)

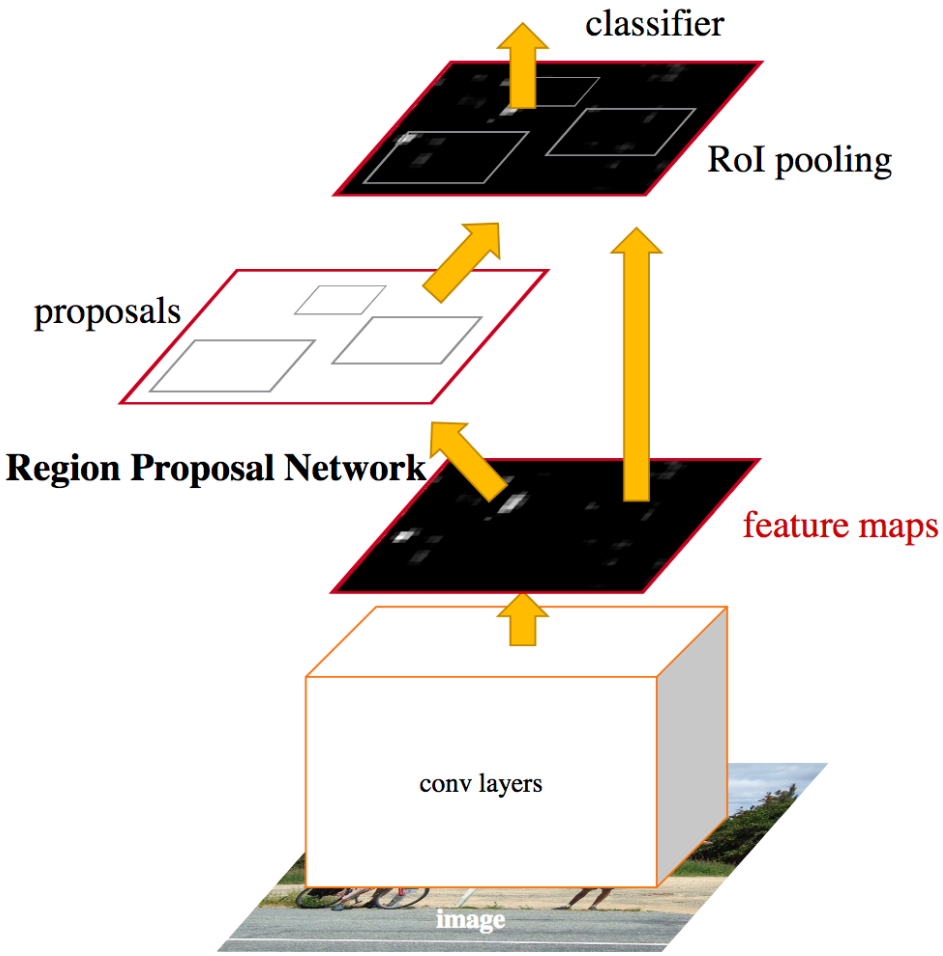




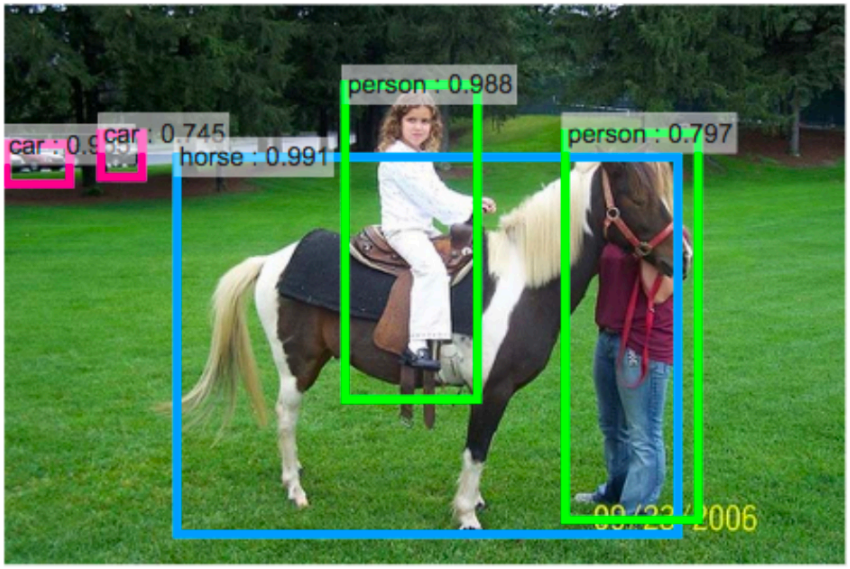
Slide credit: Ross Girshick

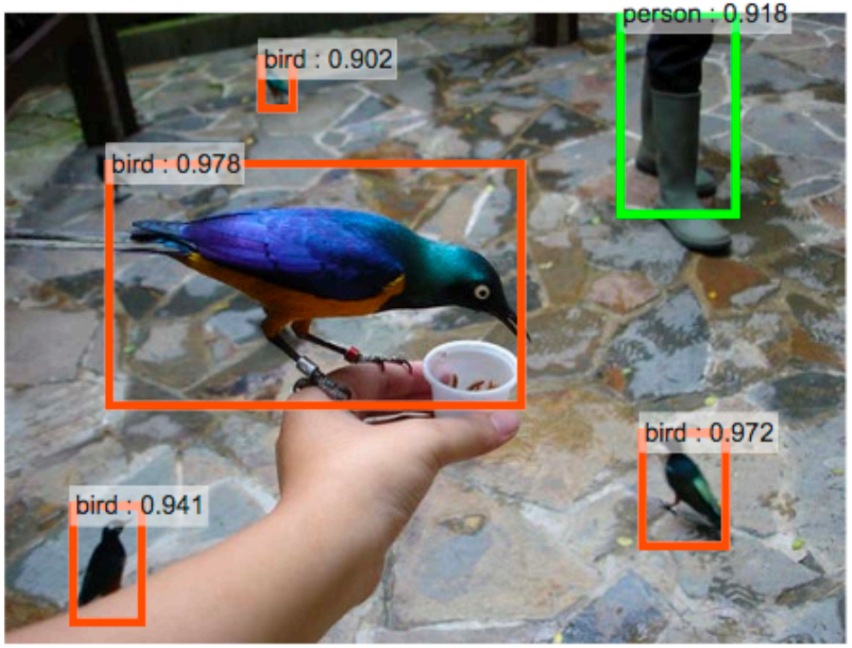
Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

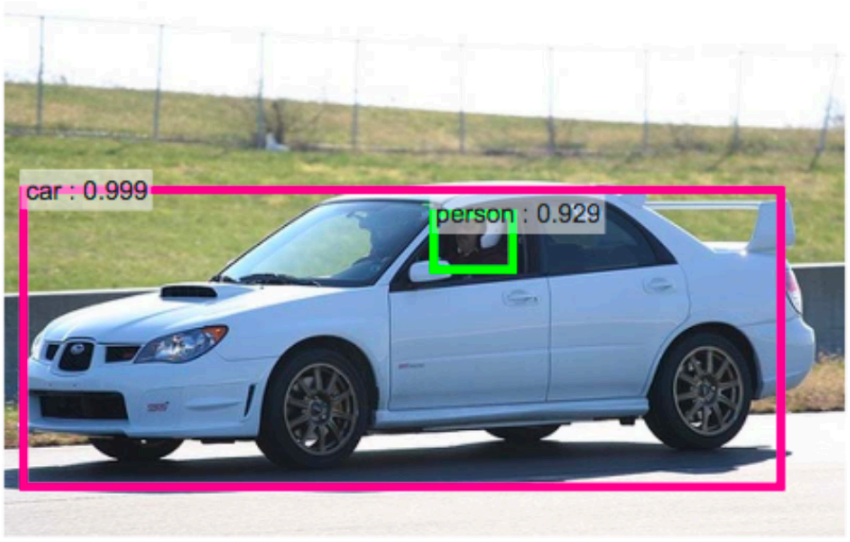
Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun

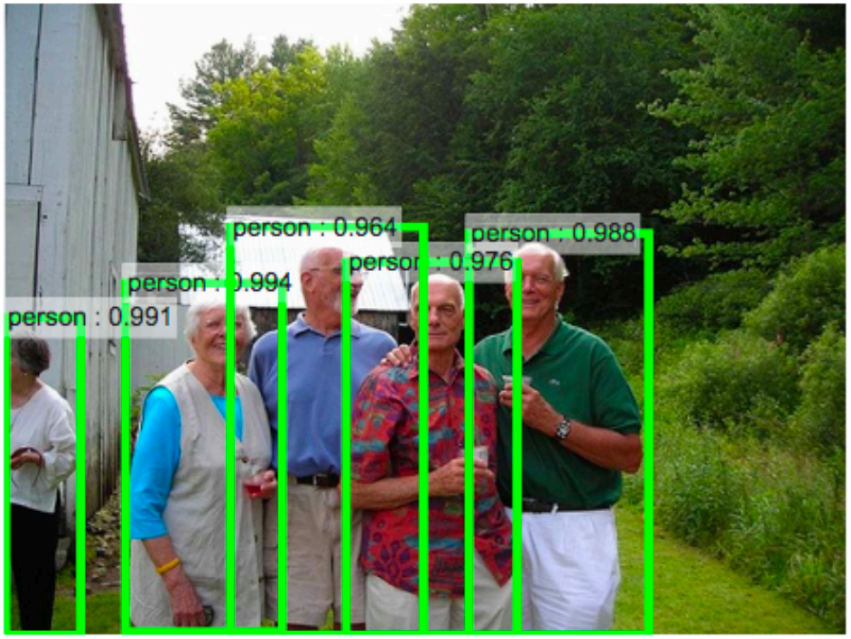


https://arxiv.org/pdf/1506.01497.pdf


car : 0.9
car : 0.745
horse : 0.991
person : 0.988
person : 0.797


bird : 0.902
person : 0.918
bird : 0.978
bird : 0.972
bird : 0.941


car : 0.999
person : 0.929


person : 0.964
person : 0.988
person : 0.994
person : 0.991