Machine Learning saves Computer Vision

Dima Damen

Origins of Computer Science
December 2014

University of BRISTOL

Overview

- What is computer vision?
- Early attempts
- The need for machine learning
- Success Stories
  - Viola&Jones Face Detector
  - Pictorial Structures
  - Background Subtraction
- Have we been saved??

What is computer vision?

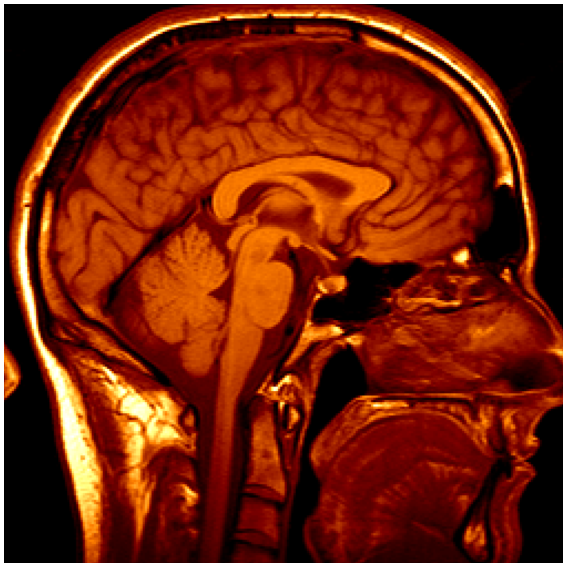
- A digital image is just a bunch of samples (pixels) and quantised values (colour)

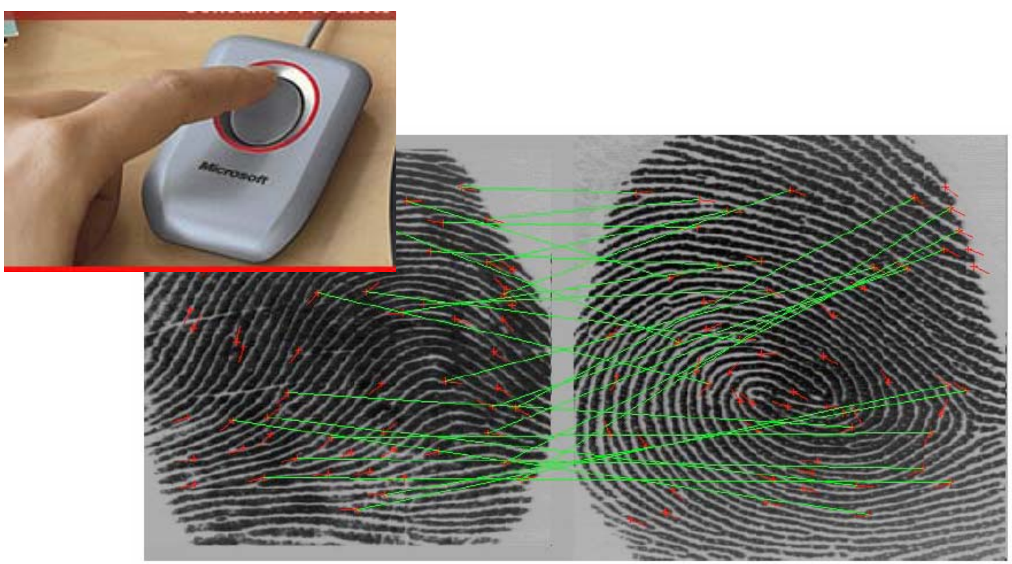
55	58	59	63	65	71	73	75	78	80	84	86	90	93	97	101	102	94	95	111	116	120	121
56	63	64	66	68	71	73	75	78	80	84	86	90	93	97	101	95	98	100	102	104	122	125
59	63	64	66	68	70	70	71	73	75	78	80	84	86	90	93	101	102	110	115	119	124	126
60	63	64	66	68	71	73	75	78	80	84	86	90	93	97	101	104	108	112	116	120	126	129
62	64	67	69	71	73	74	78	81	84	87	90	94	97	101	106	110	114	120	124	127	133	136
65	66	69	71	72	76	78	81	84	88	91	94	99	103	107	110	117	125	131	136	140	144	148
66	68	71	73	75	79	81	85	87	92	96	100	104	109	116	73	37	57	73	89	107	125	137
69	71	74	76	78	82	85	89	93	96	101	106	110	116	113	29	1	6	6	8	13	19	27
71	74	77	78	82	87	90	93	96	101	105	110	115	125	81	18	8	6	7	7	10	10	13
73	76	79	81	85	89	93	97	101	105	109	115	121	124	43	23	16	5	7	5	10	13	14
74	77	81	85	89	93	97	101	105	110	116	121	129	90	5	7	12	17	8	6	9	13	14
77	81	85	88	91	96	100	104	110	115	121	126	132	108	55	10	7	26	16	12	9	14	17
81	84	87	92	95	101	106	110	115	121	126	131	135	140	140	42	13	32	29	26	14	13	15
83	85	90	94	98	103	108	114	119	126	130	136	142	147	142	37	17	31	33	37	17	15	28
84	87	92	97	101	106	112	119	125	131	136	142	149	157	141	31	23	22	12	14	14	19	18
87	90	95	100	105	111	116	112	122	134	144	154	163	178	144	33	33	22	8	9	10	23	14
91	95	99	105	110	119	106	28	28	43	57	75	93	118	92	32	28	20	12	10	9	22	16
95	99	104	109	115	126	70	8	5	7	7	9	11	14	25	27	11	25	20	19	8	19	14
98	102	108	114	121	114	34	19	6	6	7	10	14	12	23	18	9	29	20	19	7	13	19
101	107	113	117	127	74	7	14	14	9	6	11	13	13	19	19	15	29	21	20	8	8	27
105	111	116	122	127	114	69	11	24	16	10	10	15	15	17	20	12	32	20	18	8	10	33
109	113	120	126	130	138	108	19	29	29	18	9	15	16	19	23	11	33	20	15	8	25	38
111	116	121	128	132	140	101	16	29	37	31	10	14	17	21	21	9	32	21	12	13	38	41

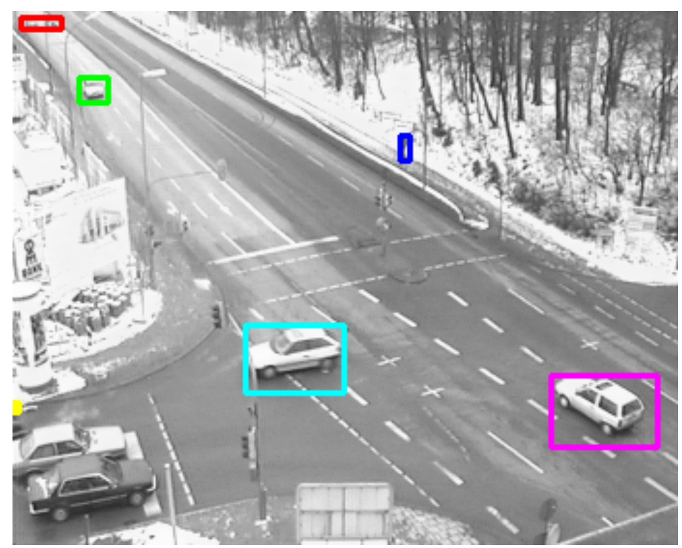
What is computer vision?

- Can we make computer understand
  - images? [photos, medical, …]
  - videos? [tv broadcast, youtube, …]
- Looks easy… but… !

What is computer vision?



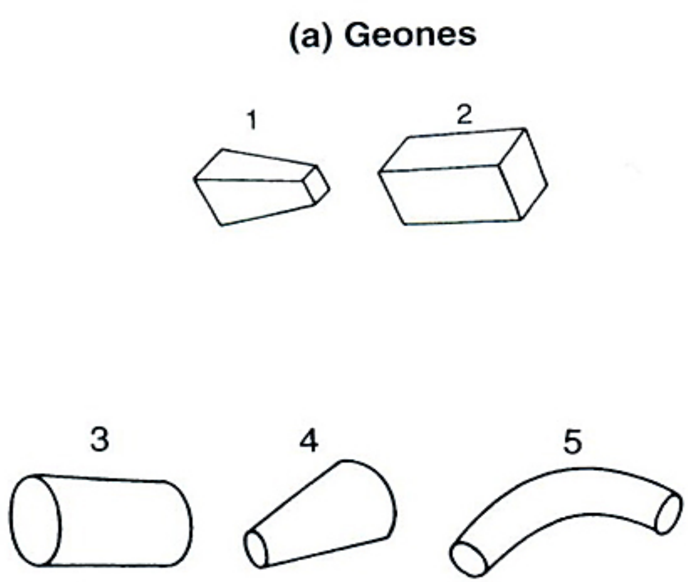




Computer Vision

- Early computer vision methods tried to model the world, without using training data

(RBC – Recognition by Components)

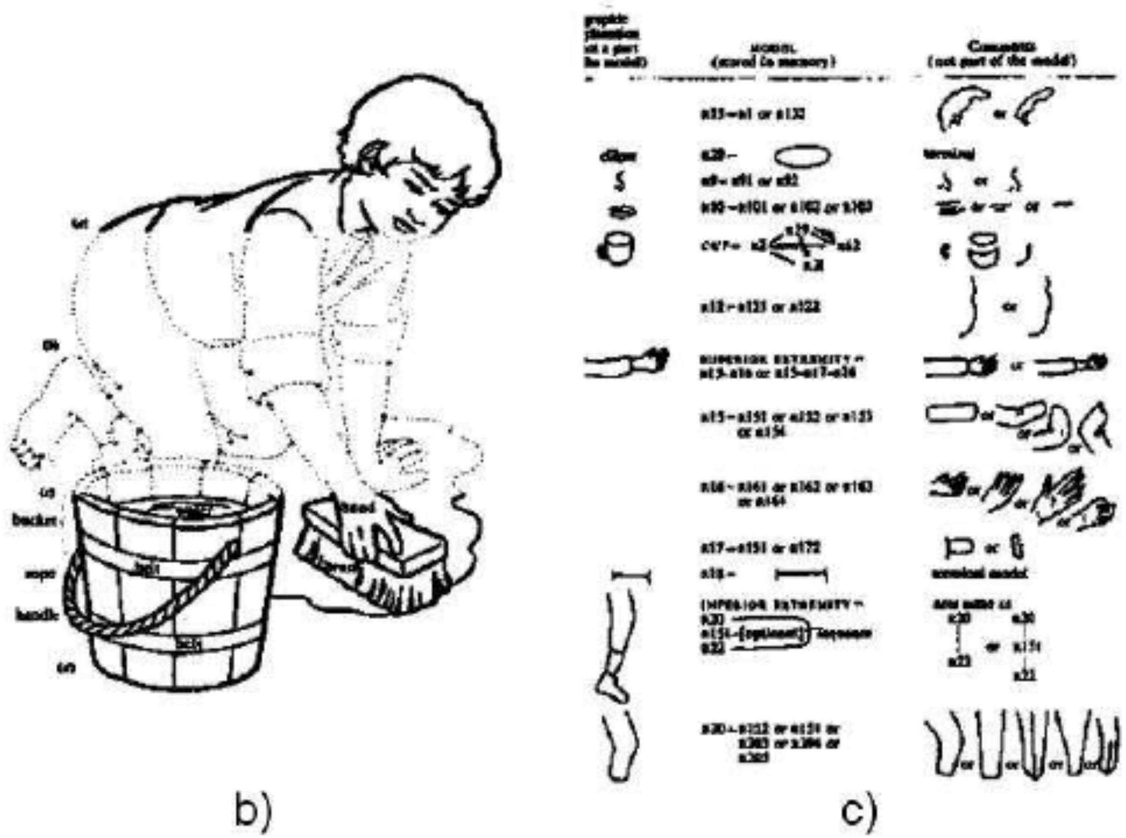


Biederman, I. (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev. 1987;94(2):115-47.

Computer Vision

- Early computer vision methods tried to model the world, without using training data

(curves)

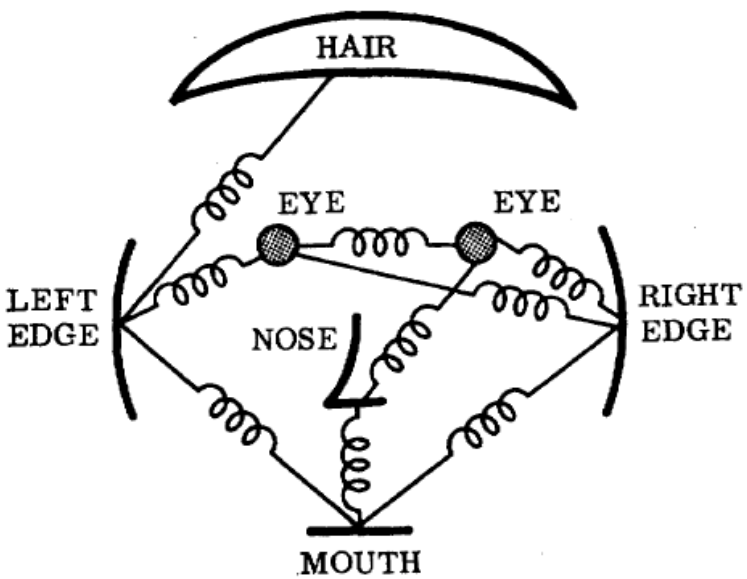


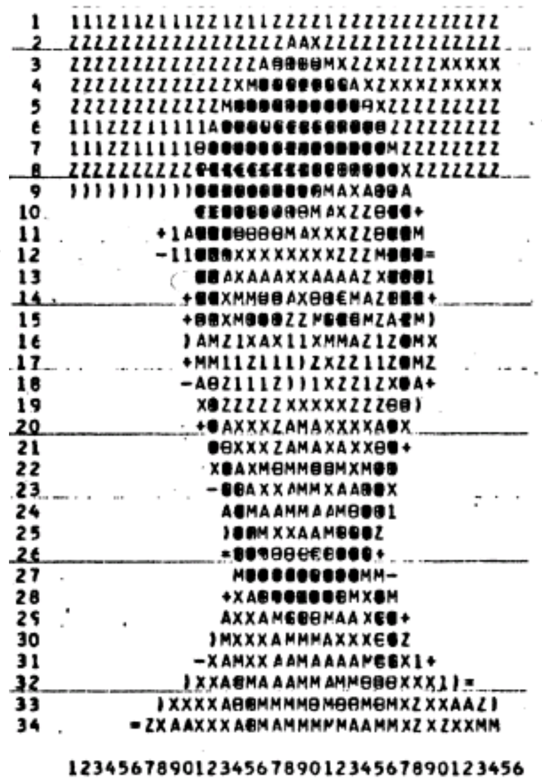
A. Guzman (1971). Analysis of curved line drawings using context and global information. Machine Intelligence 6

Computer Vision

- Early computer vision methods tried to model the world, without using training data

(Part-Based Models)





Original picture.

HAIR WAS LOCATED AT (7, 23)
L/EDGE WAS LOCATED AT (17, 13)
R/EDGE WAS LOCATED AT (17, 26)
L/EYE WAS LOCATED AT (14, 17)
R/EYE WAS LOCATED AT (14, 23)
NOSE WAS LOCATED AT (20, 20)
MOUTH WAS LOCATED AT (22, 20)

Fischler, M.A.; Elschlager, R.A. (1973). "The Representation and Matching of Pictorial Structures". IEEE Transactions on Computers: 67.

What is Machine Learning?

- Algorithms for learning from data
- In 1959, Arthur Samuel defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed" [Wiki]

The need for machine learning

- Vision can be formulated as a learning problem
- Formulation: binary classification

Features $x =$ x_1 x_2 x_3 $\cdots$ x_N x_{N+1} x_{N+2} $\cdots$ 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)

The need for machine learning

- Methods that use training data quickly outperformed modelling approaches (1990+).
- Machine learning is now a core part of computer vision.
- Nearly every machine learning algorithm has been used in one way or another in computer vision.
- Visual data (images and videos) is a new source for machine learning scientists.

Success Stories

Several success stories have paved the way:

1. Viola & Jones Face Detector (2001)

2. Pictorial Structures (2001)

Success Stories

- Face Detection – the Viola & Jones Face Detector

BBC NEWS

LIVE BBC NEWS CHANNEL

News services
Your news when you want it

News Front Page
World
UK
England
Northern Ireland
Scotland
Wales
Business
Market Data
Your Money
Economy
Companies
Politics
Health
Education
Science & Environment
Technology
Entertainment

Last Updated: Monday, 6 February 2006, 14:29 GMT

E-mail this to a friend   Printable version

Face-hunting cameras boost Nikon

Japanese camera maker Nikon has tripled its profits on the back of strong sales of digital cameras that automatically focus on human faces.

Face recognition cameras like the Coolpix L1 are popular

Operating profit for the three months to 31 December was 19.8bn yen ($167m; £95m), up from 5.9bn yen in 2004.

Nikon said that sales of compact digital cameras had been boosted by the success of new face recognition models.

It had also seen strong sales of its digital "SLR" cameras with interchangeable lenses and bodies

SEE ALSO:
- Nikon to focus on digital cameras
  12 Jan 06 | Business
- Digital trouble hits Nikon shares
  10 Feb 04 | Business
- Why digital cameras = better photographers
  20 Jan 04 | Magazine
- R.I.P. 35mm Camera
  15 Jan 04 | Magazine

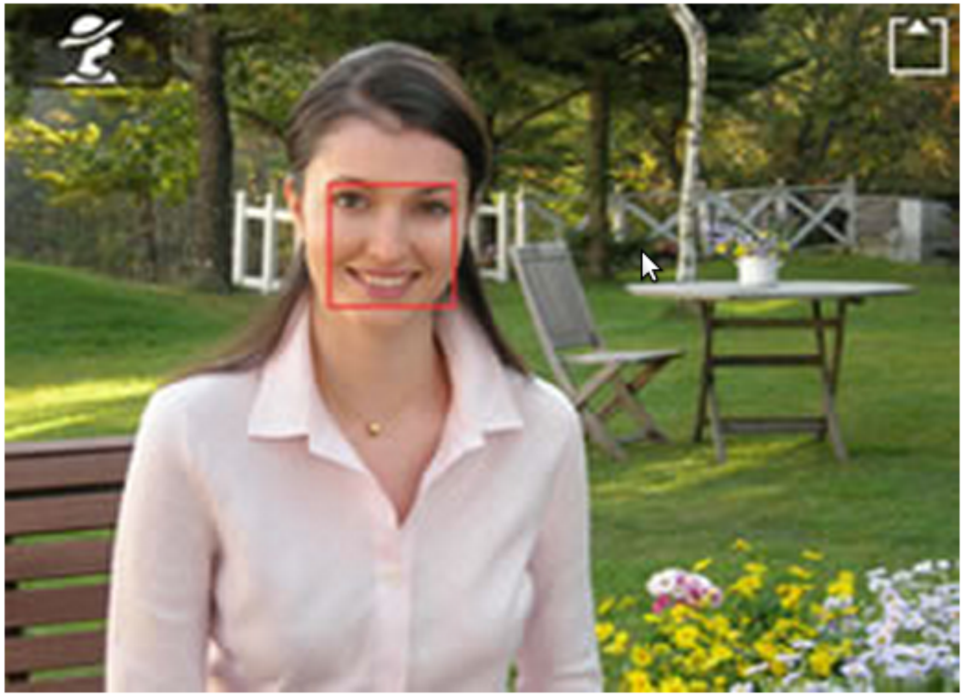
RELATED INTERNET LINKS:
- Nikon

The BBC is not responsible for the content of external internet sites

TOP BUSINESS STORIES
- Unemployment dips to 2.47

Success Stories

- Face Detection – the Viola & Jones Face Detector



Sample image: Subject as seen on the COOLPIX 5900 camera's color LCD and when using Nikon's Face-priority AF function

Success Stories

Case I: Viola & Jones Face Detector

Paul Viola
MIT (1996-2000)
MERL (2001-2002)
Microsoft (2002 - now)

Michael Jones
Compaq (-2000)
MERL (2001-now)

Case I: Viola & Jones Face Detector

Robust Real-time Object Detection

Paul Viola
viola@merl.com
Mitsubishi Electric Research Labs
201 Broadway, 8th FL
Cambridge, MA 02139

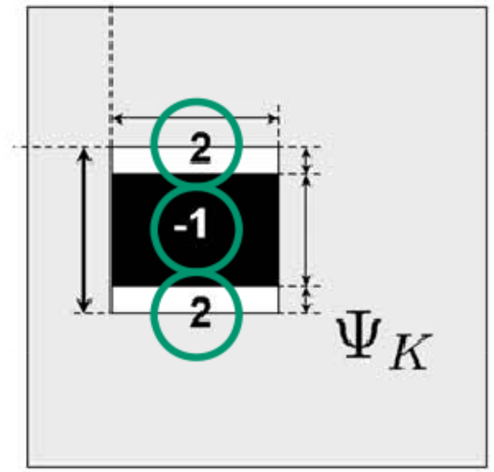
Michael Jones
mjones@crl.dec.com
Compaq CRL
One Cambridge Center
Cambridge, MA 02142

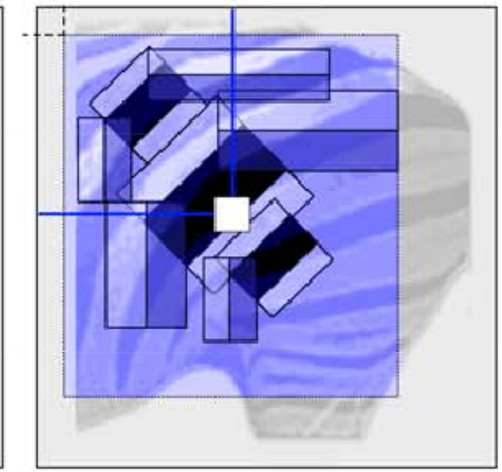
Abstract

This paper describes a visual object detection framework that is capable of processing images extremely rapidly while achieving high detection rates. There are three key contributions. The first is the introduction of a new image representation called the "Integral Image" which allows the features used by our detector to be computed very quickly. The second is a learning algorithm, based on AdaBoost, which selects a small number of critical visual features and yields extremely efficient classifiers [6]. The third contribution is a method for combining classifiers in a "cascade" which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions. A set of experiments in the domain of face detection are presented. The system yields face detection performace comparable to the best previous systems [18, 13, 16, 12, 1]. Implemented on a conventional desktop, face detection proceeds at 15 frames per second.

Case I: Viola & Jones Face Detector

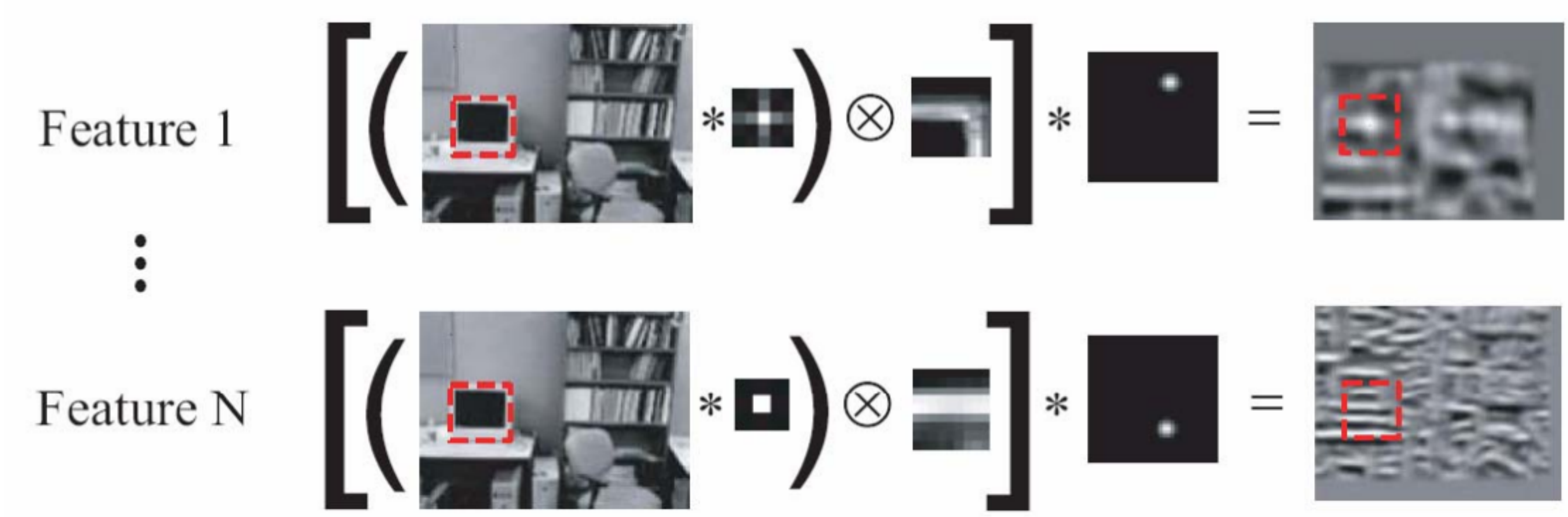
Haar wavelets and Integral Images



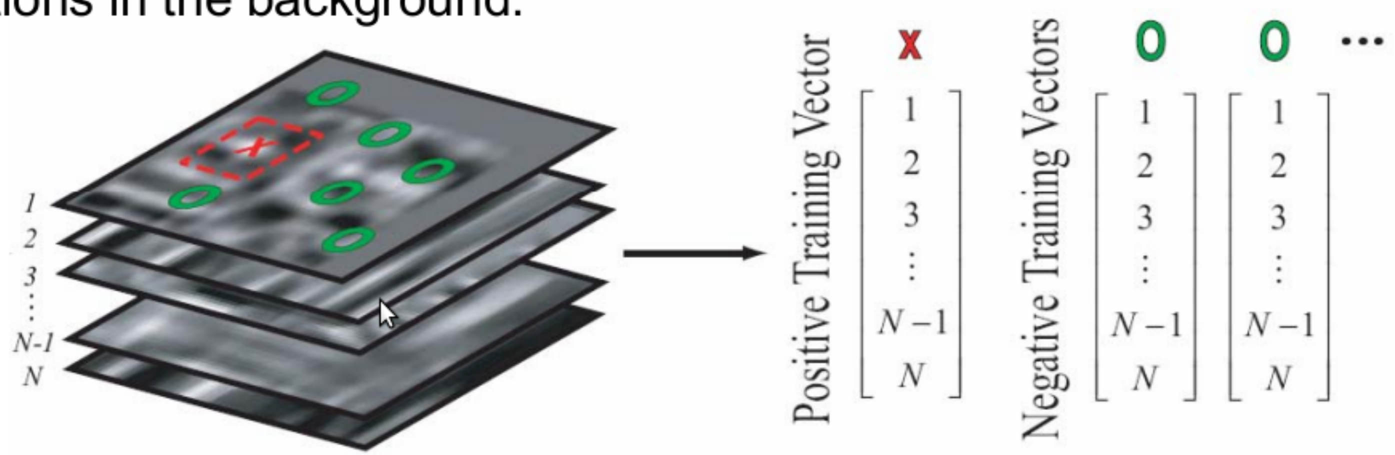


Case I: Viola & Jones Face Detector

First we evaluate all the N features on all the training images.



Then, we sample the feature outputs on the object center and at random locations in the background:



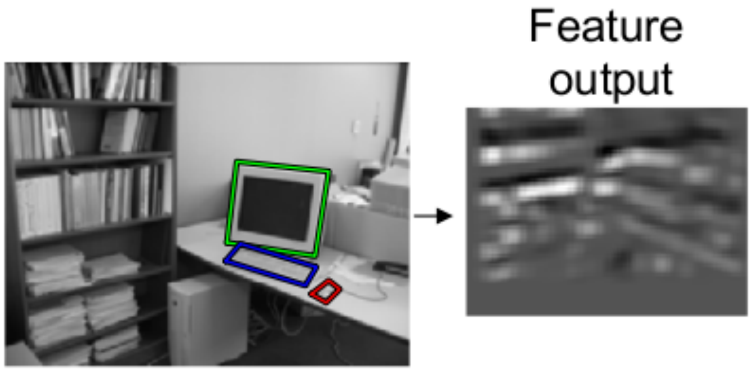
Case I: Viola & Jones Face Detector

Training Data + 10,000 negative examples were selected by randomly picking sub-windows from 9500 images which did not contain faces

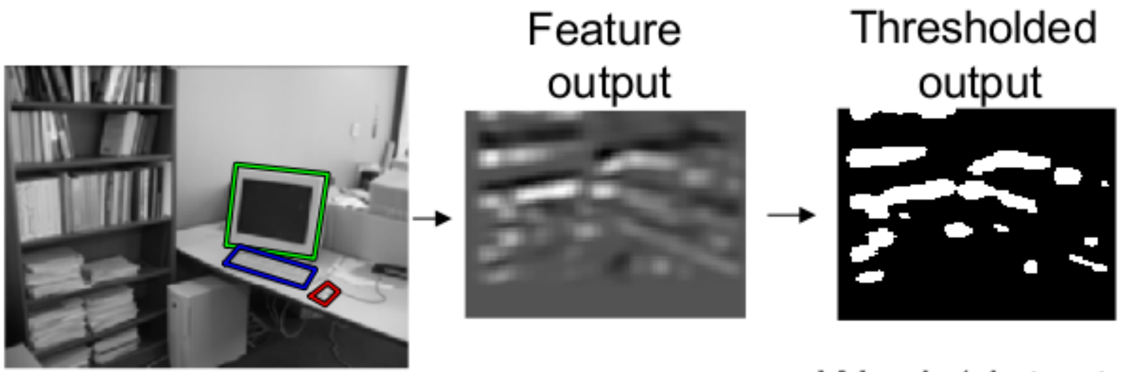
Case I: Viola & Jones Face Detector

- AdaBoost Classification – a method for supervised learning
- Weak classifiers: classifiers that perform slightly better than chance. (error < 0.5)
- Boosting is an iterative algorithm that repeatedly constructs a hypothesis aimed at correcting mistakes of the previous hypothesis
- Strong classifier: has an error rate ε
- Introduced by Freund & Shapire (1995)

Case I: Viola & Jones Face Detector

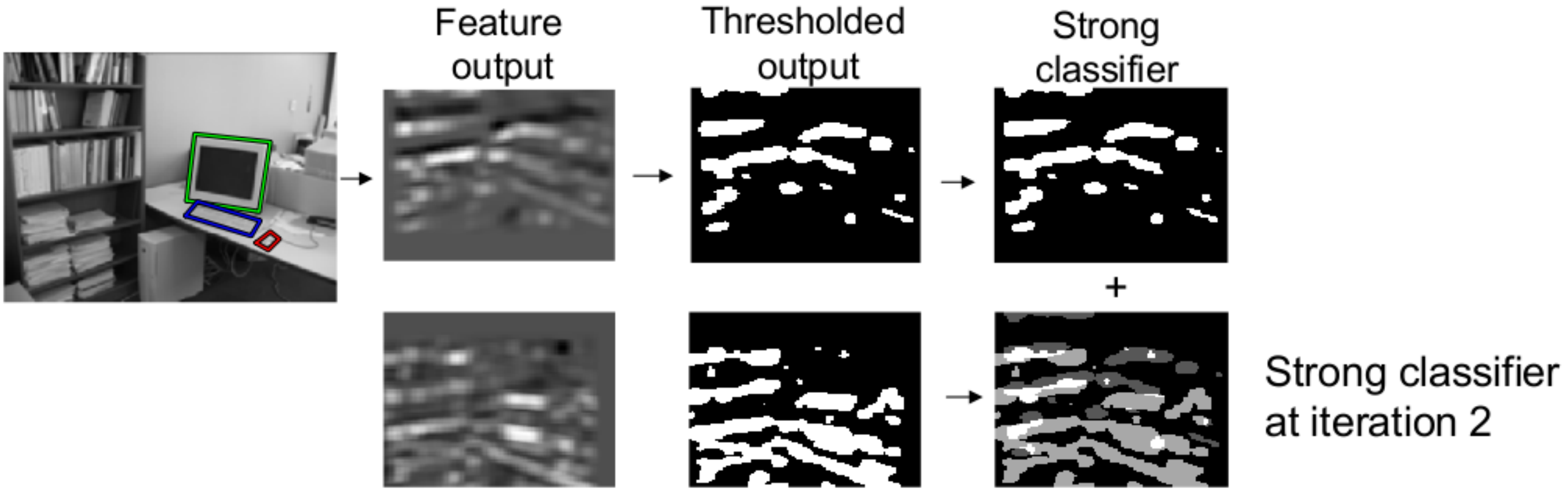


Case I: Viola & Jones Face Detector

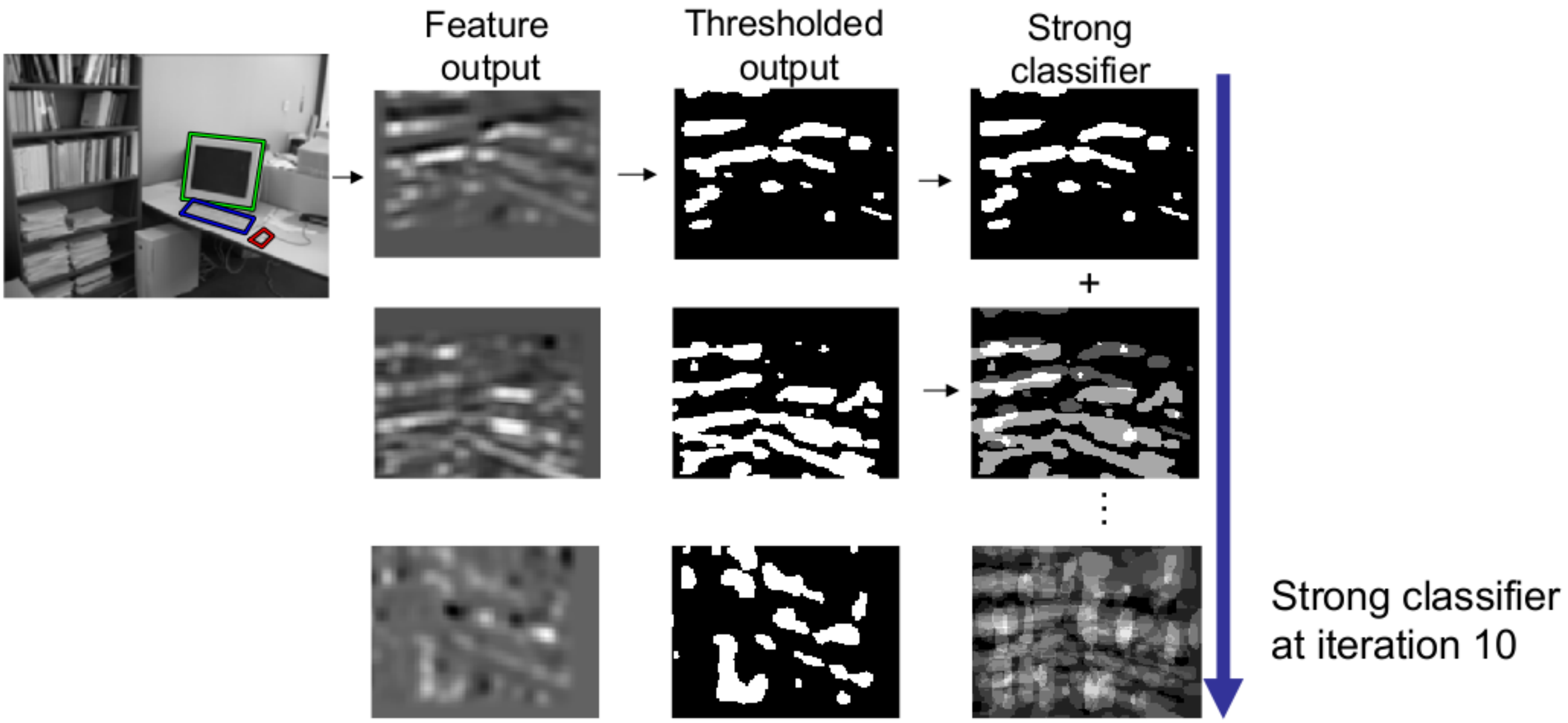


Weak ‘detector’
Produces many false alarms.

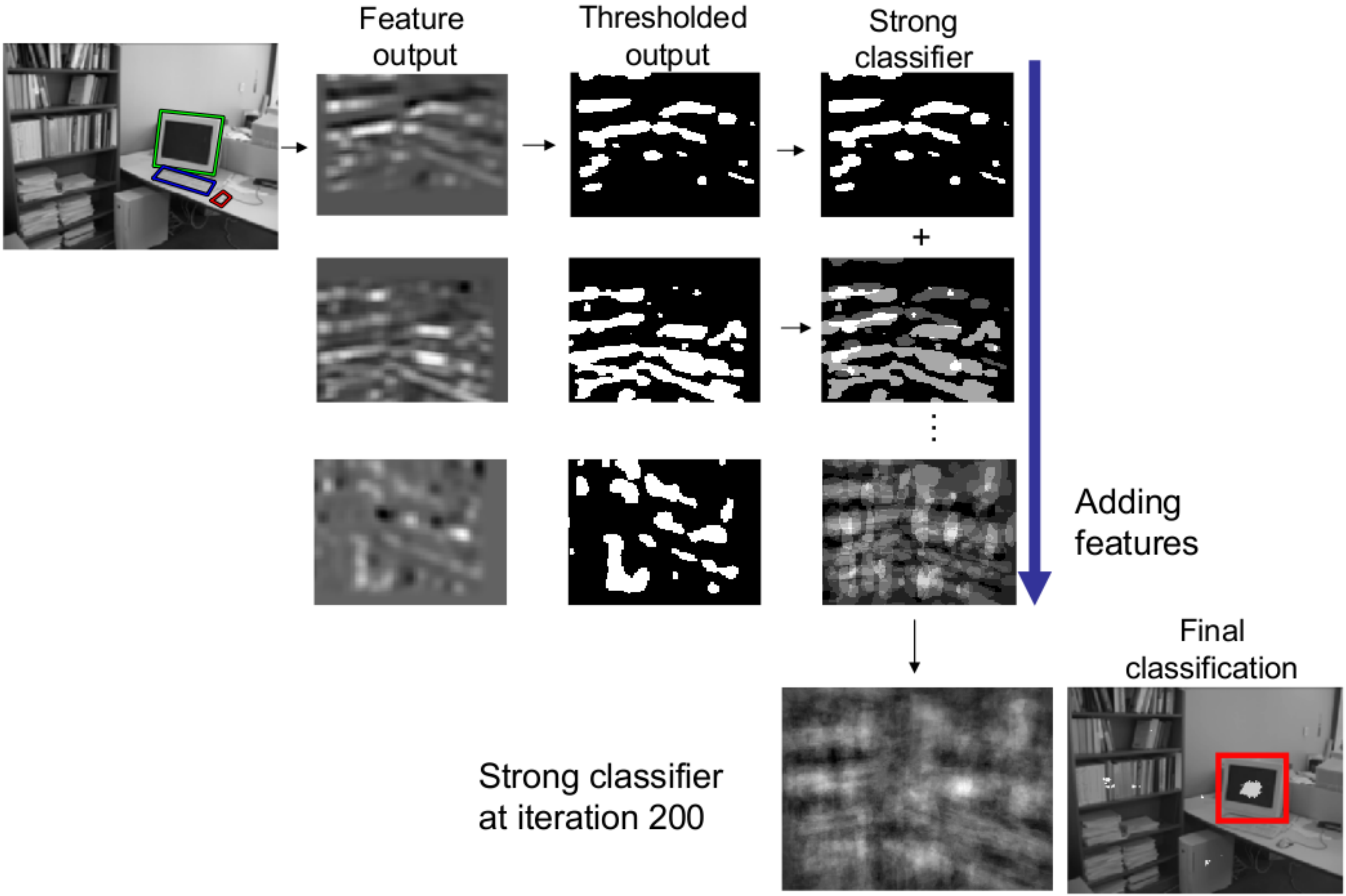
Case I: Viola & Jones Face Detector



Case I: Viola & Jones Face Detector

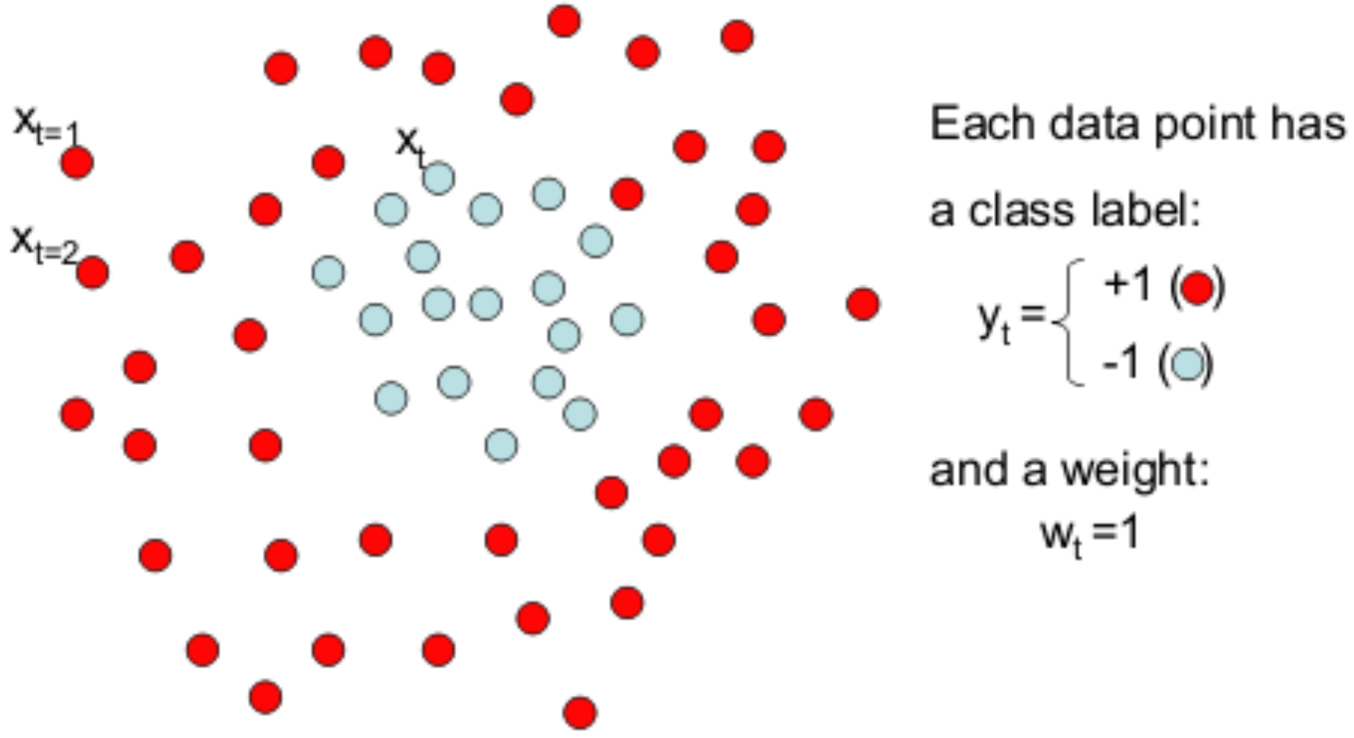


Case I: Viola & Jones Face Detector



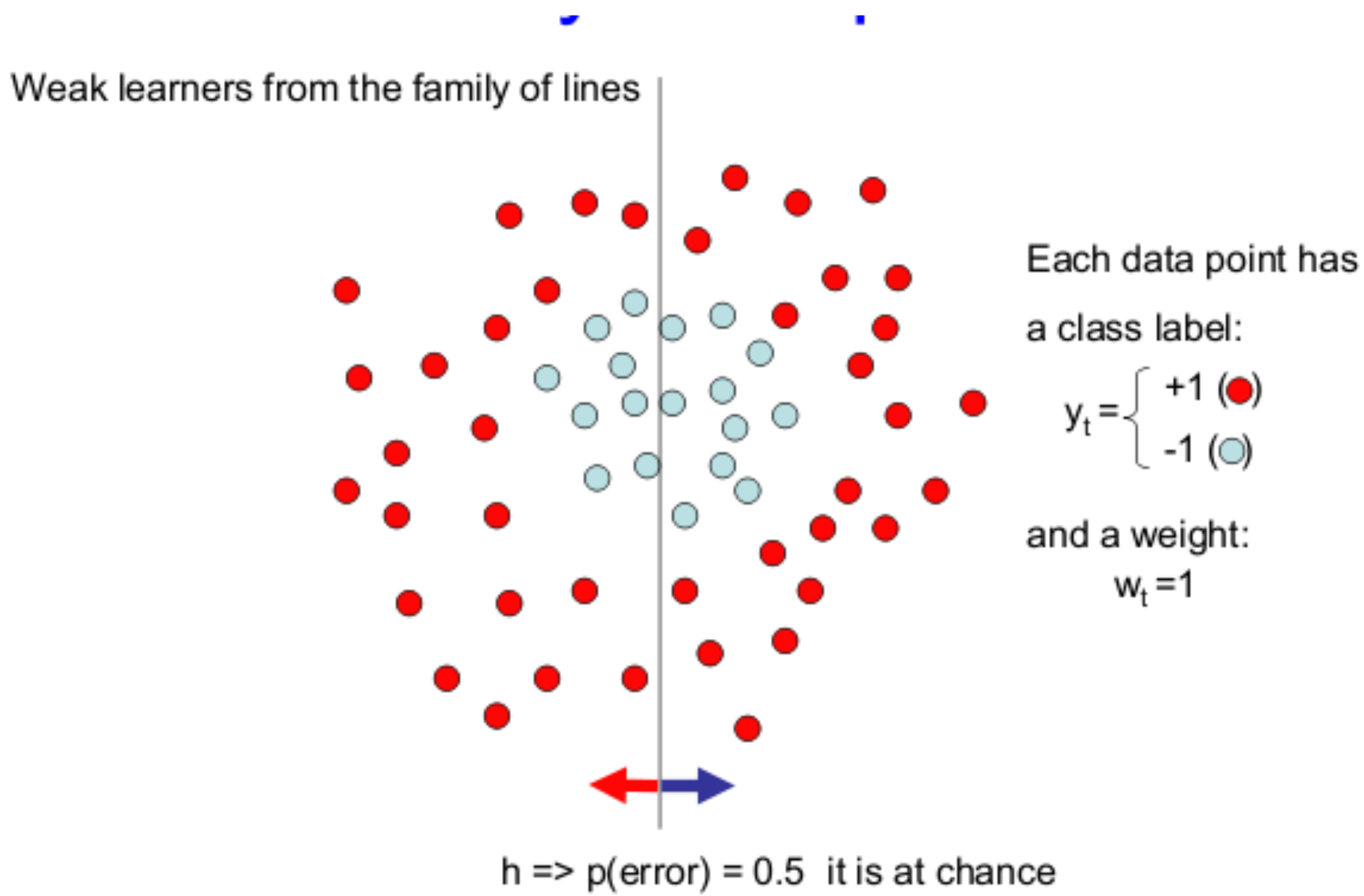
Case I: Viola & Jones Face Detector

Boost Classification



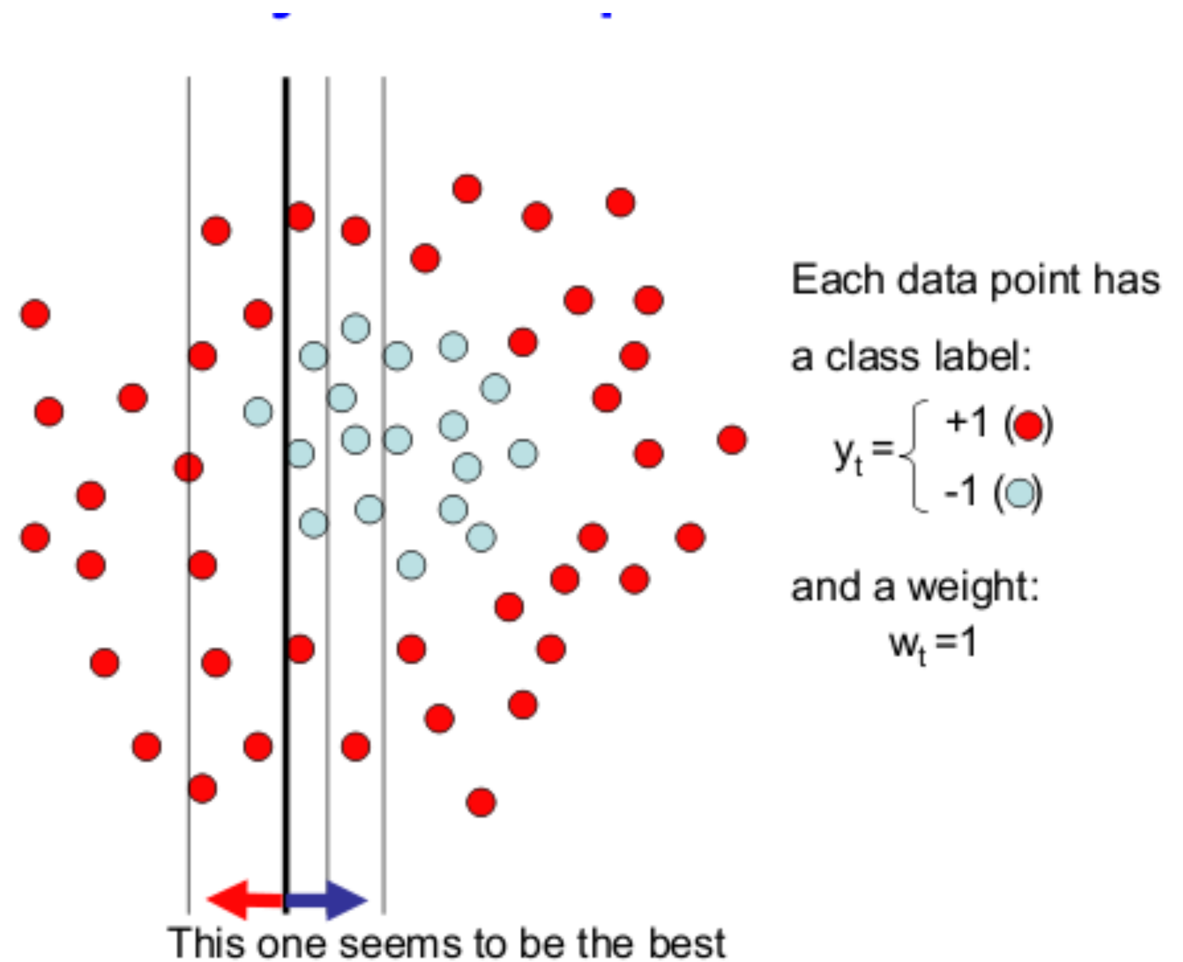
Case I: Viola & Jones Face Detector

Boost Classification



Case I: Viola & Jones Face Detector

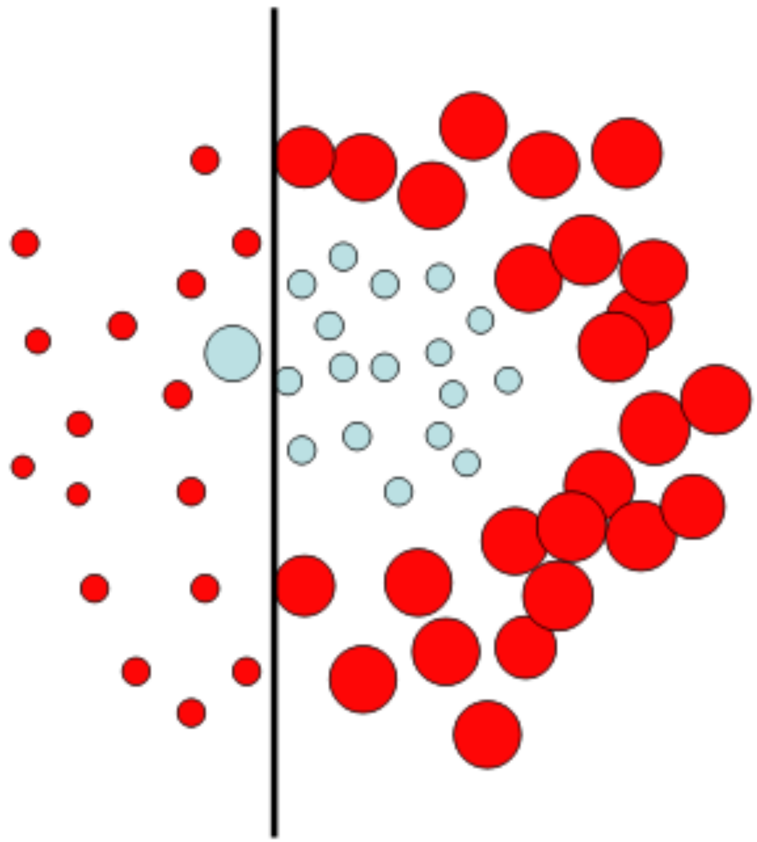
Boost Classification



This is a '**weak classifier**': It performs slightly better than chance.

Case I: Viola & Jones Face Detector

AdaBoost Classification



Each data point has a class label:

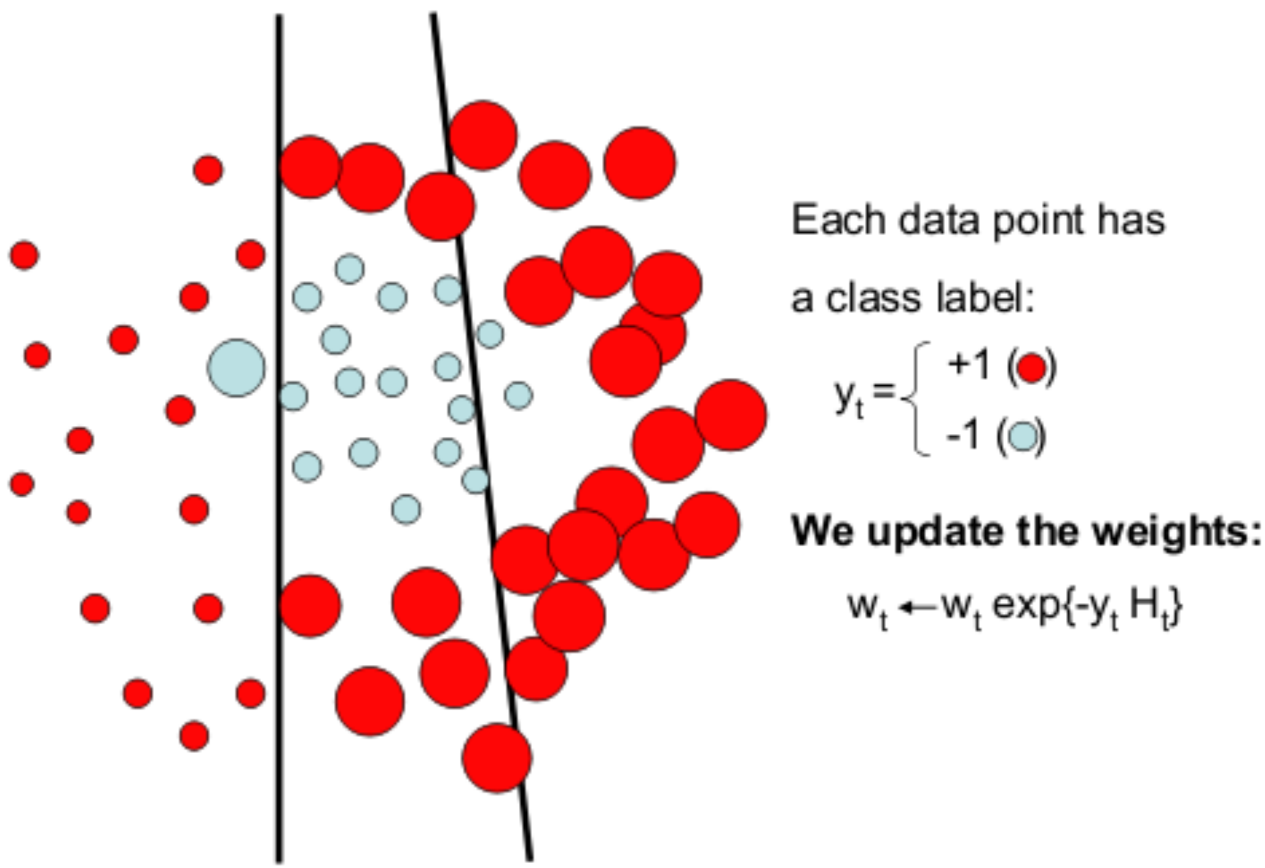
$$y_t = \begin{cases} +1 & (\text{red}) \\ -1 & (\text{light blue}) \end{cases}$$

We update the weights:

$$w_t \leftarrow w_t \exp\{-y_t H_t\}$$

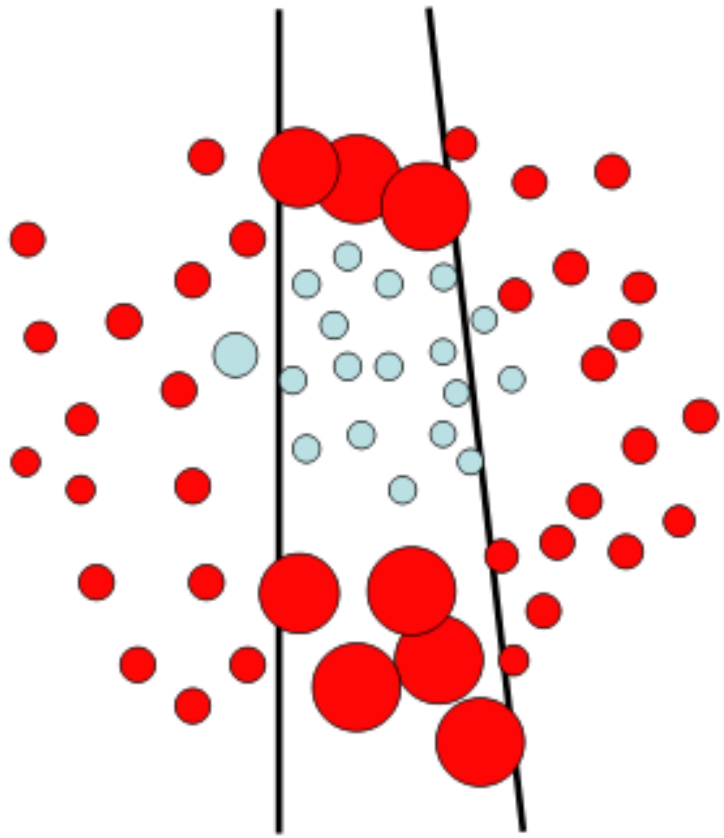
Case I: Viola & Jones Face Detector

Boost Classification



Case I: Viola & Jones Face Detector

Boost Classification



Each data point has

a class label:

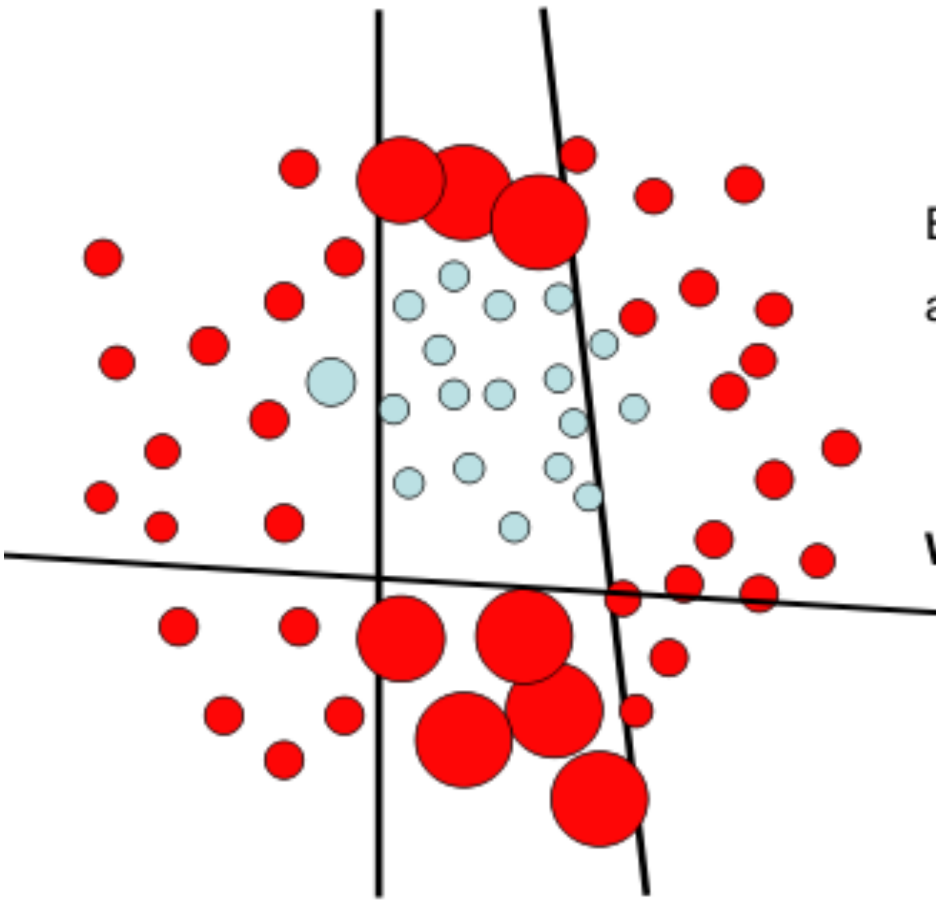
$$y_t = \begin{cases} +1 & (\text{red}) \\ -1 & (\text{blue}) \end{cases}$$

We update the weights:

$$w_t \leftarrow w_t \exp\{-y_t H_t\}$$

Case I: Viola & Jones Face Detector

Boost Classification



Each data point has

a class label:

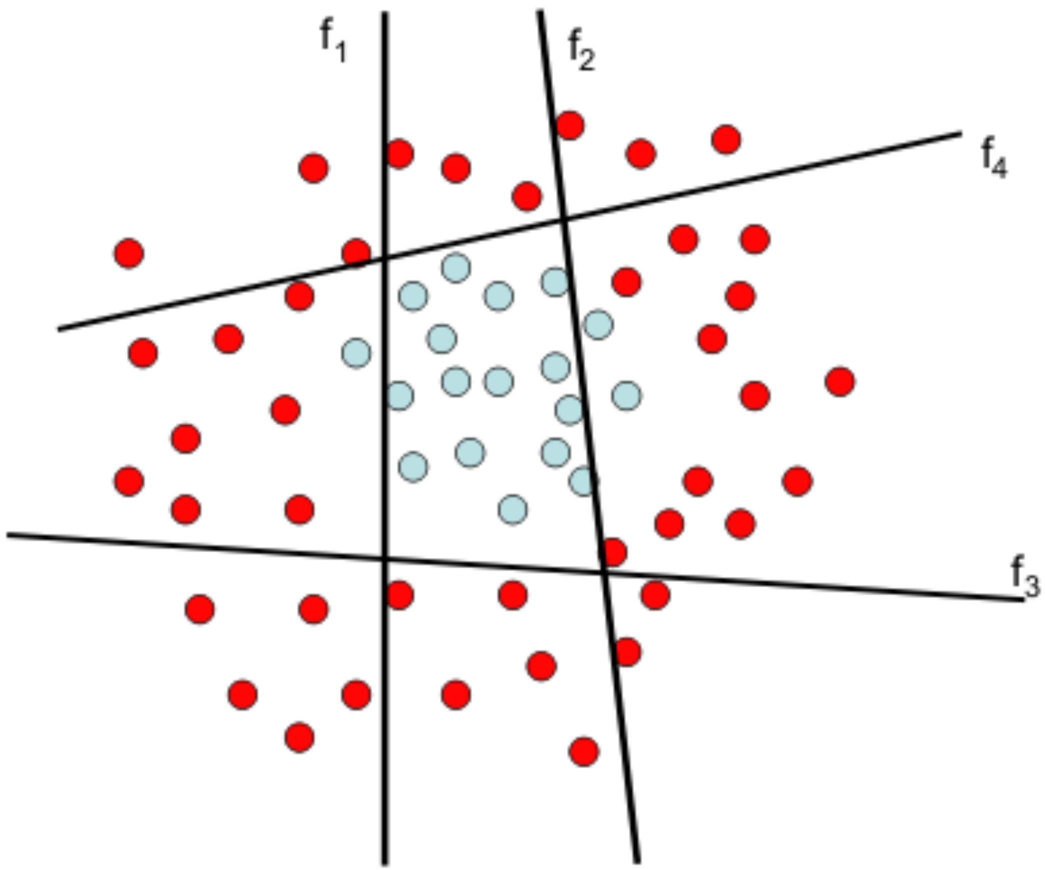
$$y_t = \begin{cases} +1 \; (\;) \\ -1 \; (\;) \end{cases}$$

We update the weights:

$$w_t \leftarrow w_t \exp\{-y_t H_t\}$$

Case I: Viola & Jones Face Detector

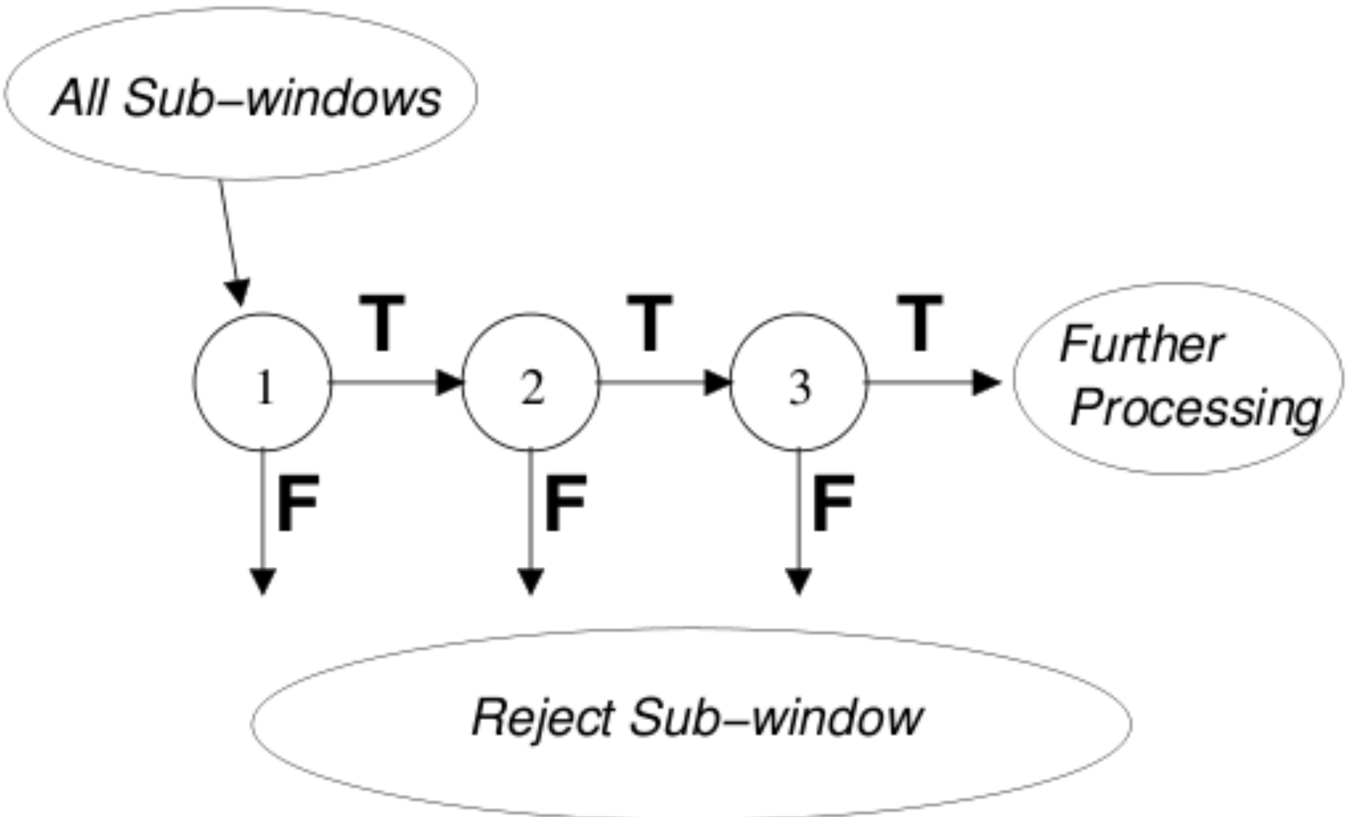
Boost Classification



The strong (non- linear) classifier is built as the combination of all the weak (linear) classifiers.

Case I: Viola & Jones Face Detector

Cascade of classifiers



Case I: Viola & Jones Face Detector

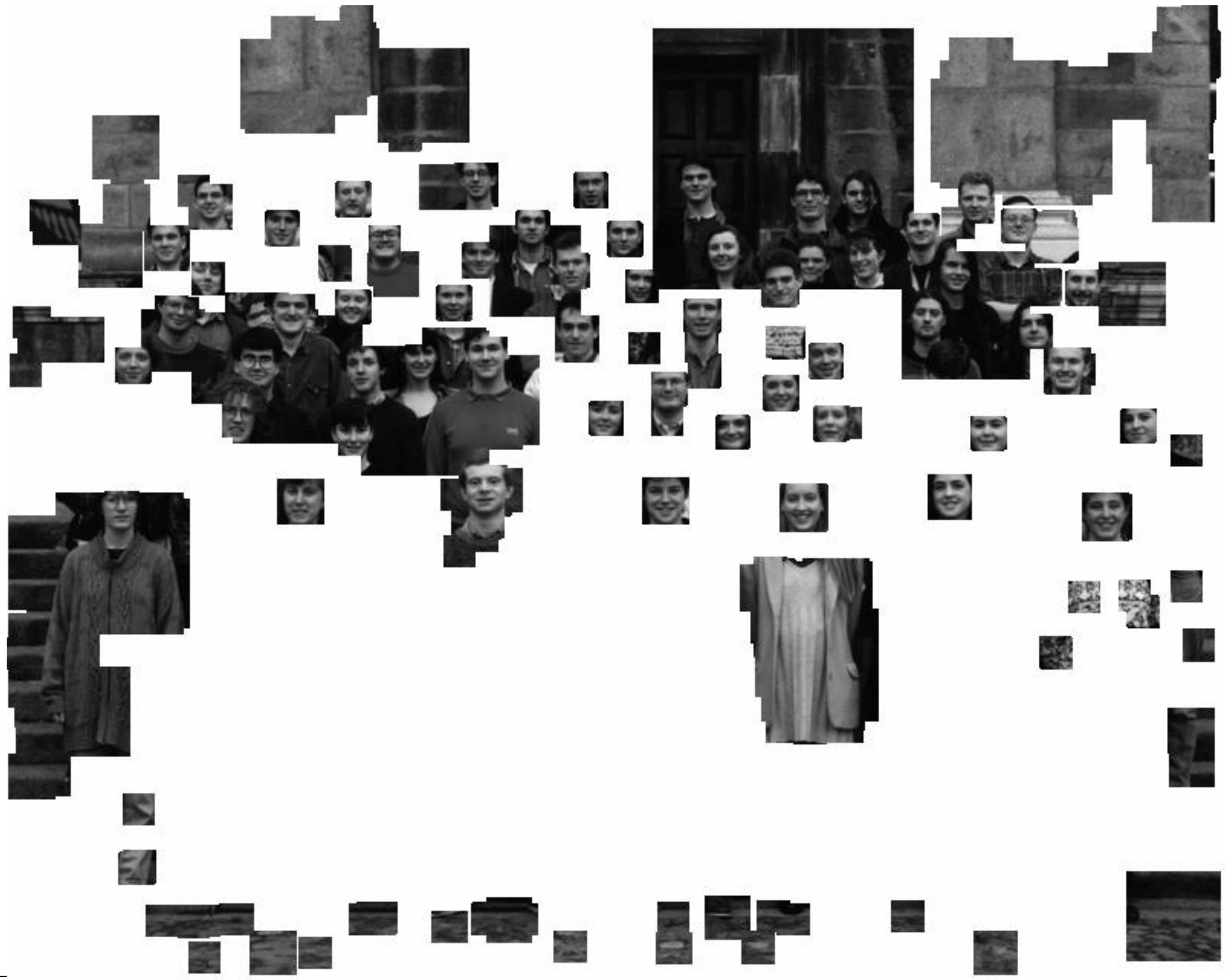
Cascade of classifiers

Case I: Viola & Jones Face Detector

Cascade of classifiers

Case I: Viola & Jones Face Detector

Cascade of classifiers

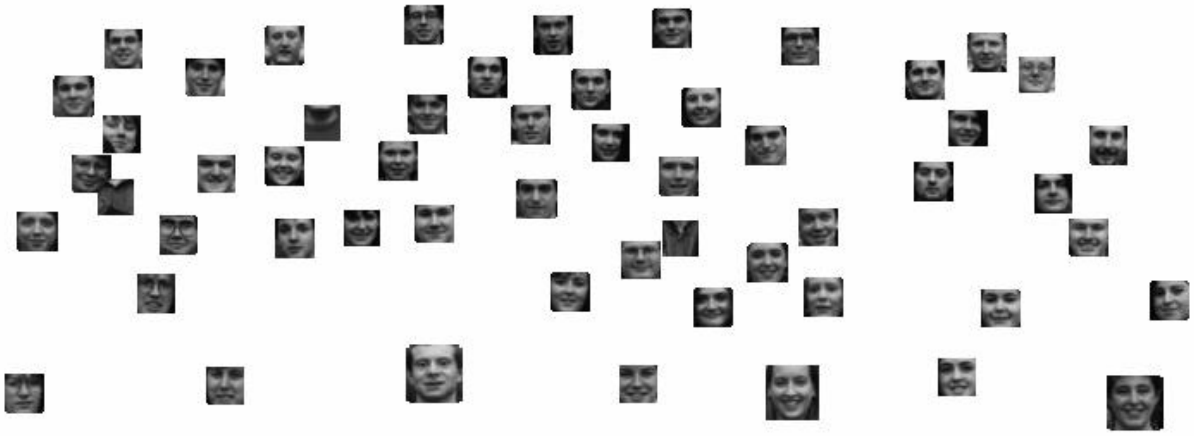


Case I: Viola & Jones Face Detector

Cascade of classifiers

Case I: Viola & Jones Face Detector

Cascade of classifiers



Case I: Viola Jones Face Detector

- The processing time of a 384 by 288 pixel image on a conventional personal computer (back in 2001) about 0.067 seconds.
- Free implementation of it available as part of OpenCV

Success Stories

Pictorial Structures…

Pedro Felzenszwalb
MIT (1999-2003)
Cornell University
Chicago University
Brown University (2011-now)

Daniel Huttenlocher
Cornell University

Case II: Pictorial Structures

Efficient Matching of Pictorial Structures *

Pedro F. Felzenszwalb
Artificial Intelligence Laboratory
MIT
Cambridge, MA 02139
pff@ai.mit.edu

Daniel P. Huttenlocher
Computer Science Department
Cornell University
Ithaca, NY 14853
dph@cs.cornell.edu

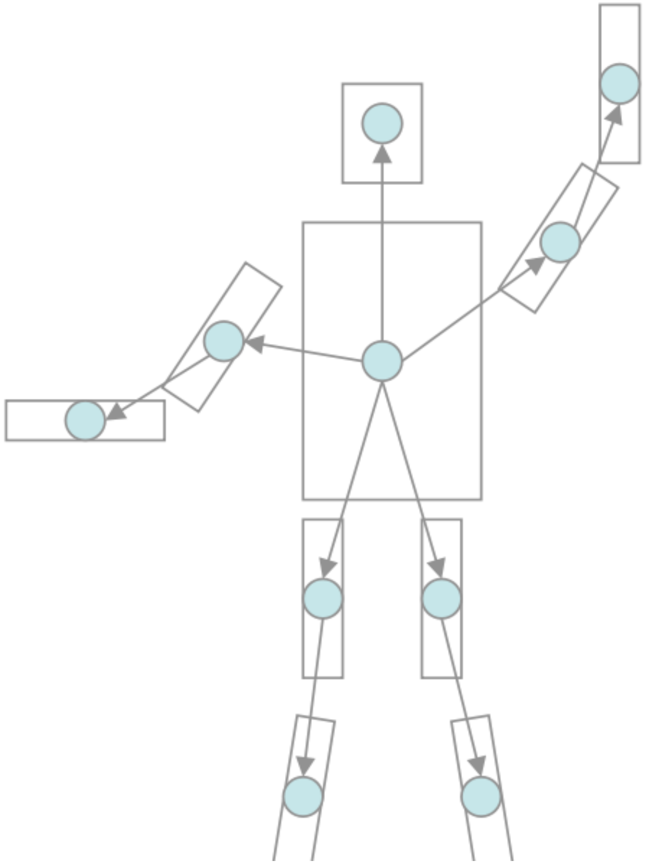
Abstract

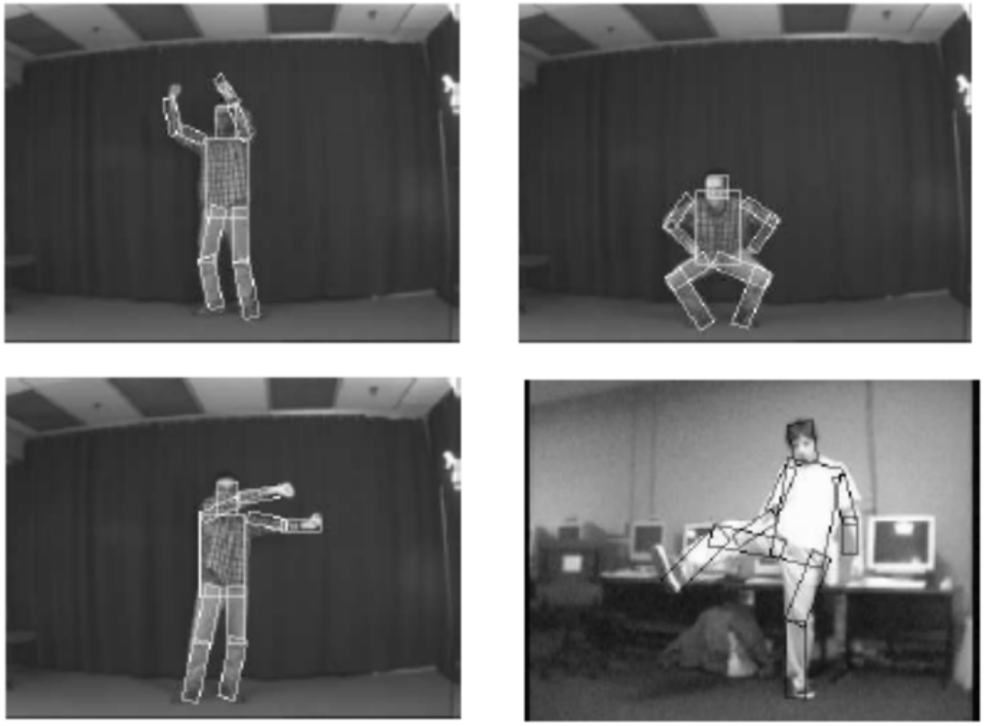
A pictorial structure is a collection of parts arranged in a deformable configuration. Each part is represented using a simple appearance model and the deformable configuration is represented by spring-like connections between pairs of parts. While pictorial structures were introduced a number of years ago, they have not been broadly applied to matching and recognition problems. This has been due in part to the computational difficulty of matching pictorial structures to images. In this paper we present an efficient algorithm for finding the best global match of a pictorial structure to an image. The running time of the algorithm is optimal and it it takes only a few seconds to match a model with five to ten parts. With this improved algorithm, pictorial structures provide a practical and powerful framework for qualitative descriptions of objects and scenes, and are suitable for many generic image recognition problems. We illustrate the approach using simple models of a person and a car.

is providing a Bayesian interpretation of the problem, in terms of MAP estimation. The running time of our algorithm is optimal, in the sense that it runs as quickly as simply matching each part separately, without accounting for the relationships between parts. In practice the algorithm is also fast, finding the globally best match of a pictorial structure to an image in just a few seconds.

Pictorial structures provide a powerful framework for qualitative descriptions of objects and scenes, making them suitable for many generic image recognition problems. In [8] and in [7], pictorial structures were used to form generic models of a human face. Simple generic appearance models were used for parts such as the eyes, mouth, etc., and the connections between parts ensured that the geometric arrangement of the parts was face-like. In [16], pictorial structures were used to model generic scene concepts such as a waterfall, a snowy montain, or a sunset. For example, a waterfall was modeled as a bright white region (water) in the middle of darker regions (rocks). The method

Case II: Pictorial Structures





Case II: Pictorial Structures

- Model is represented by a graph $G = (V, E)$.
  - $V = \{v_1, \ldots, v_n\}$ are the parts.
  - $(v_i, v_j) \in E$ indicates a connection between parts.
- $m_i(l_i)$ is the cost of placing part i at location l_i.
- $d_{ij}(l_i, l_j)$ is a deformation cost.
- Optimal location for object is given by $L^* = (l_1^*, \ldots, l_n^*)$,

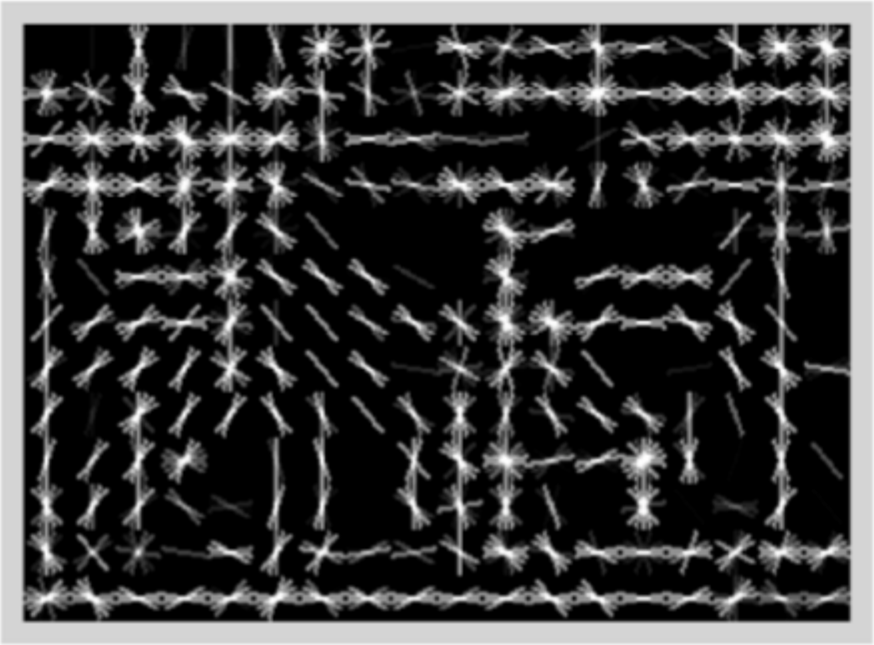
$$L^* = \underset{L}{\text{argmin}} \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i,v_j)\in E} d_{ij}(l_i, l_j) \right)$$

  - n parts and h locations gives h^n configurations.
  - If graph is a tree we can use dynamic programming.

Case II: Pictorial Structures

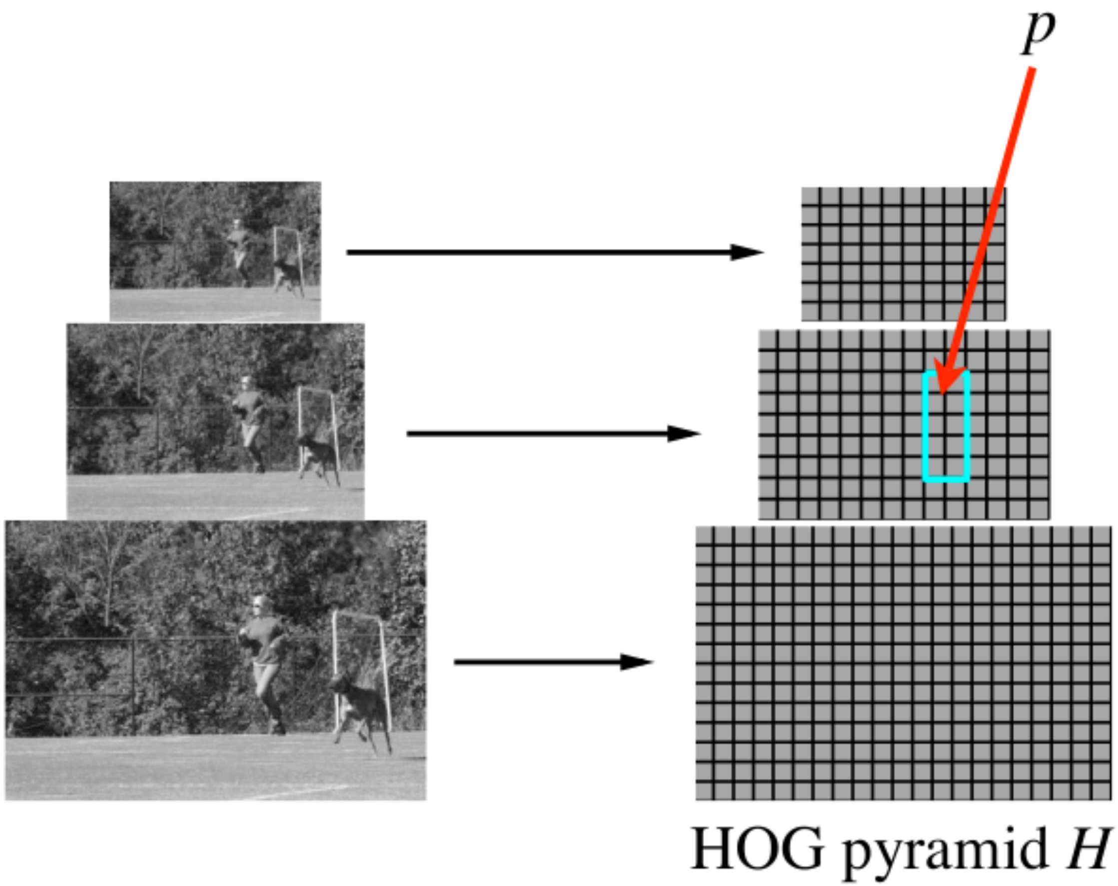
Parts can also be learnt from training data!

A complete framework for learning and detection of discriminative part-based models was proposed…

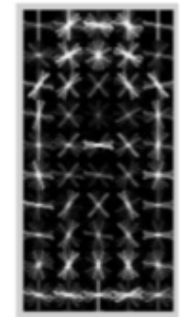


Histogram of Gradients (HoG) features

Case II: Pictorial Structures



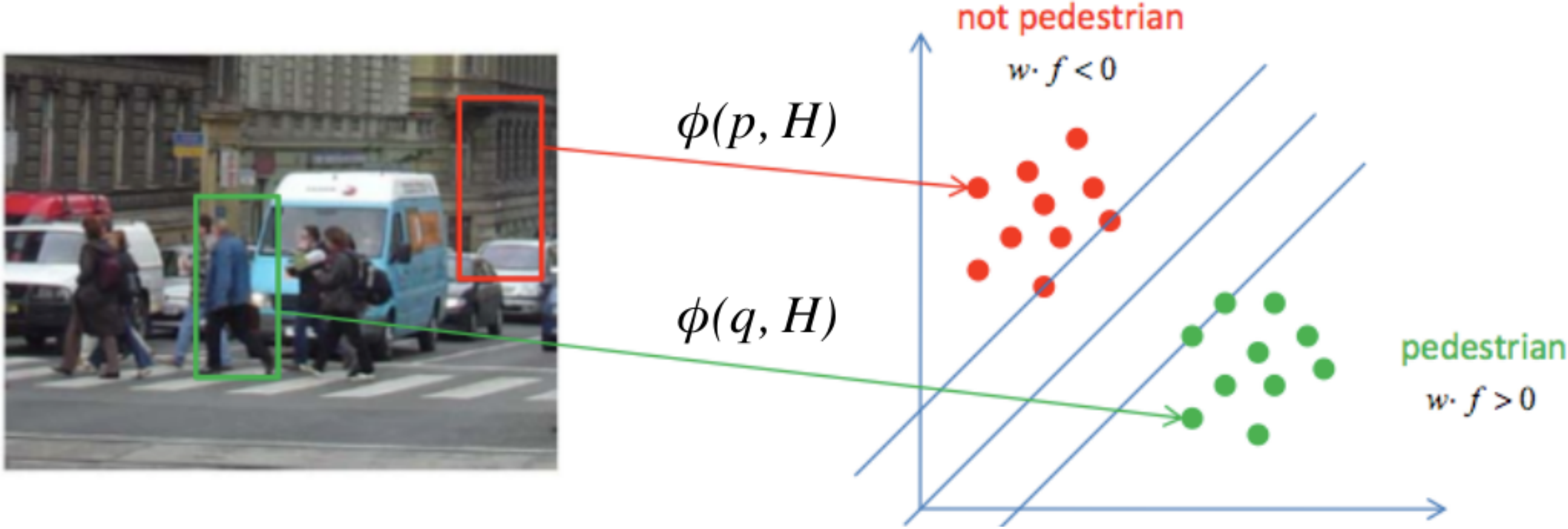
Filter F



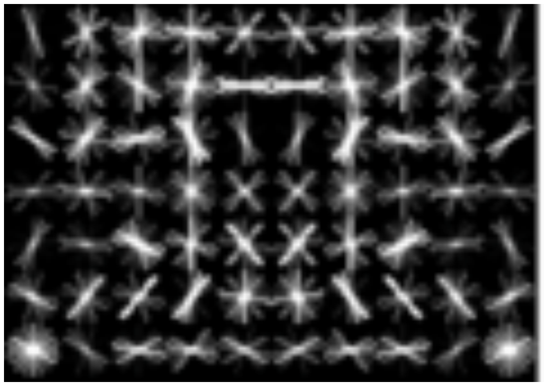
Score of F at position p is
$F \cdot \phi(p, H)$

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

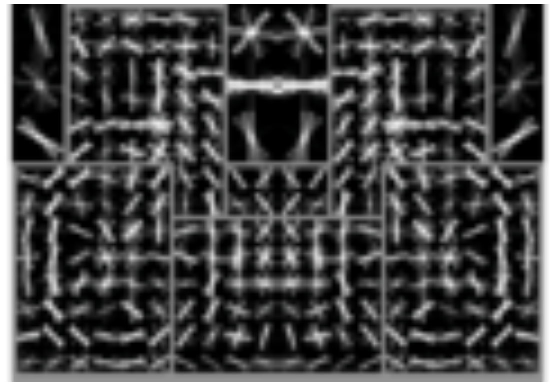
Case II: Pictorial Structures



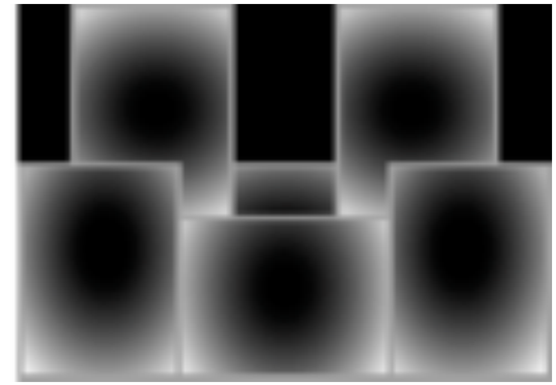
Case II: Pictorial Structures



root filter



part-based filters



deformable model

"data term" "spatial prior"

$$\mathrm{score}(p_0, \ldots, p_n) = \boxed{\sum_{i=0}^{n} F_i \cdot \phi(H, p_i)} - \boxed{\sum_{i=1}^{n} d_i \cdot (dx_i^2, dy_i^2)}$$

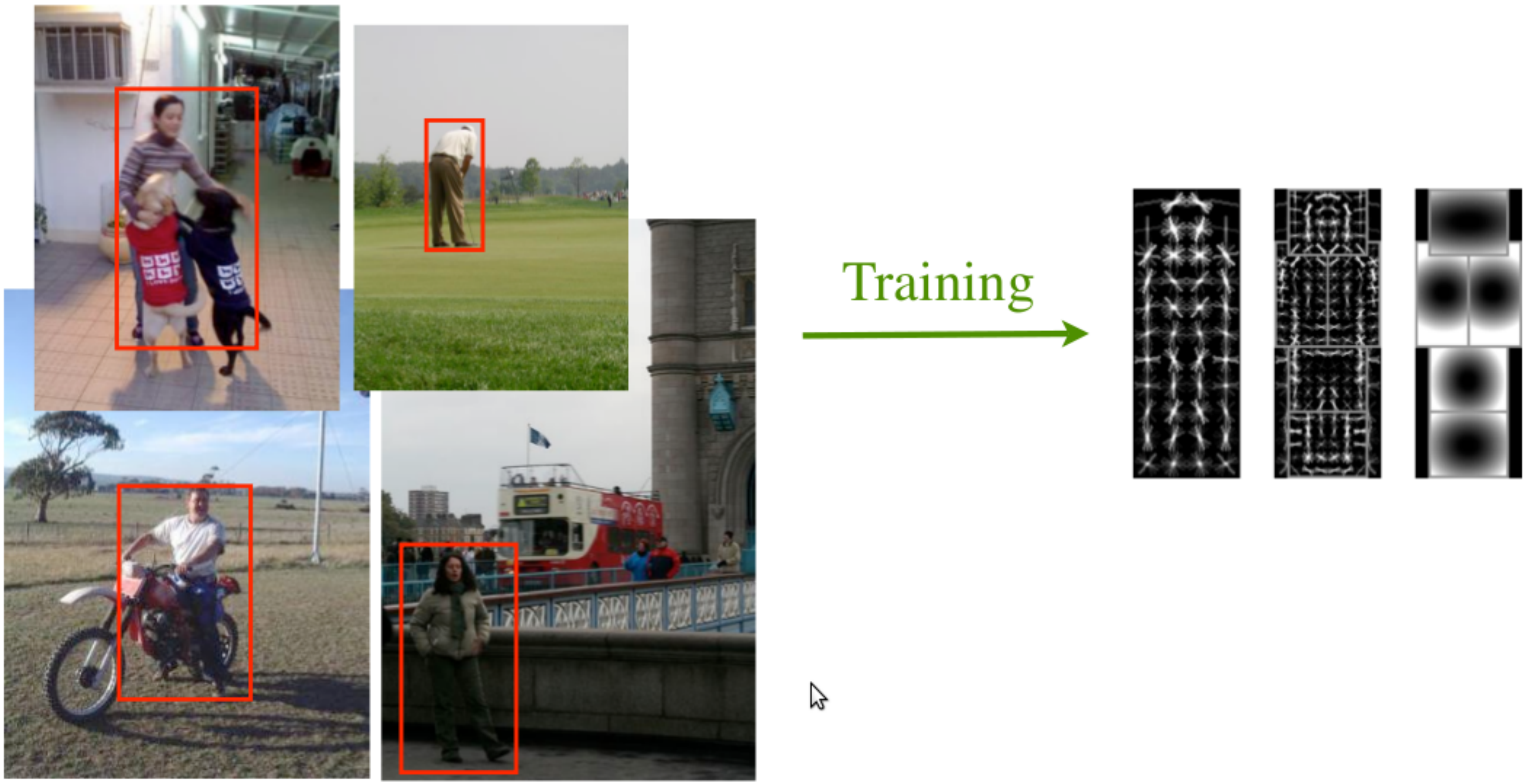
filters

displacements

deformation parameters

Case II: Pictorial Structures

Machine learning methods are needed for training



The need for machine learning

- The PASCAL challenge

PASCAL (2006)
- 5,304 images
- 9,507 objects

The need for machine learning

Pascal 2006 – Car category

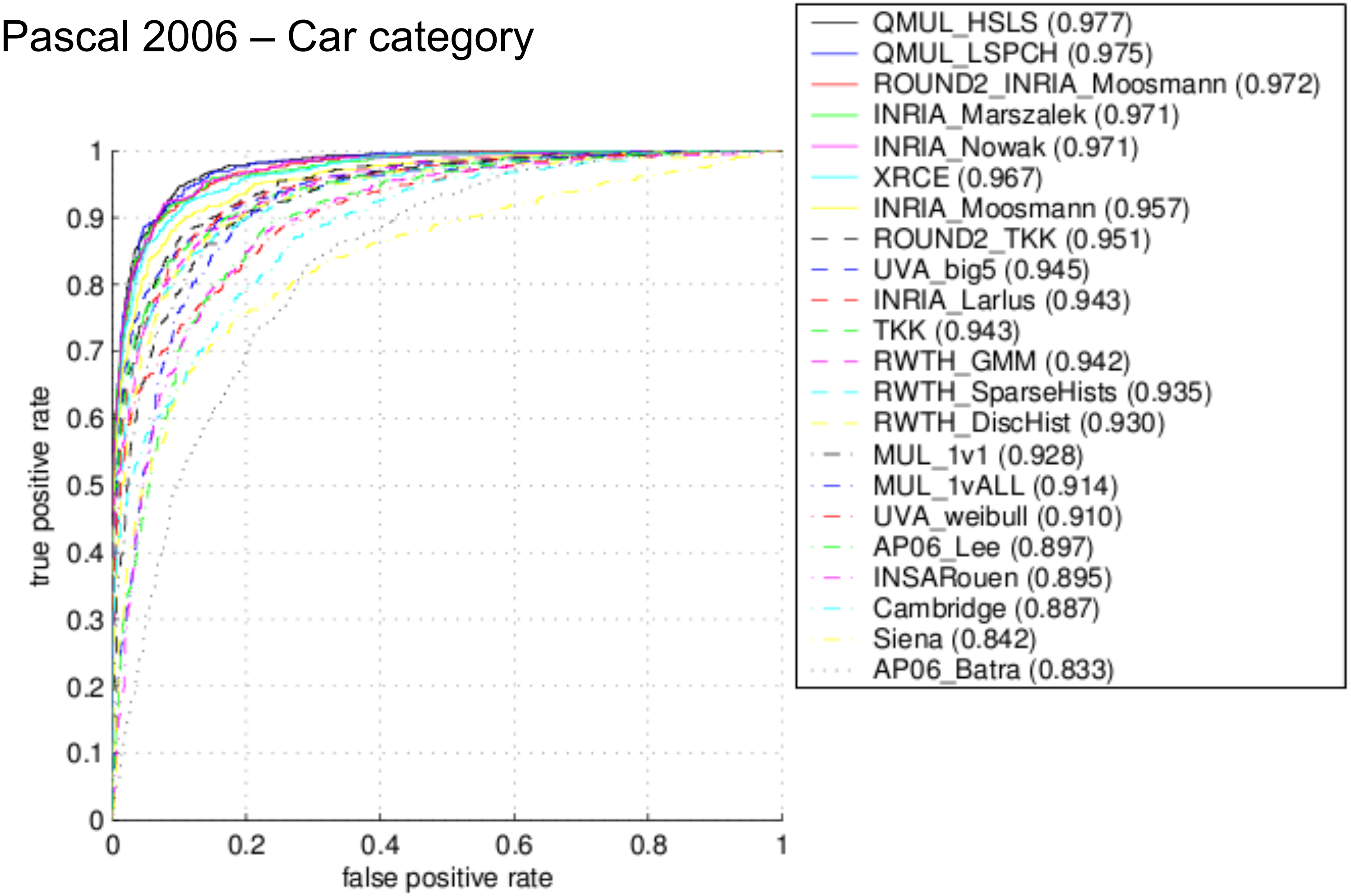


Figure 5: Competition 1.3: car (all entries)

The need for machine learning

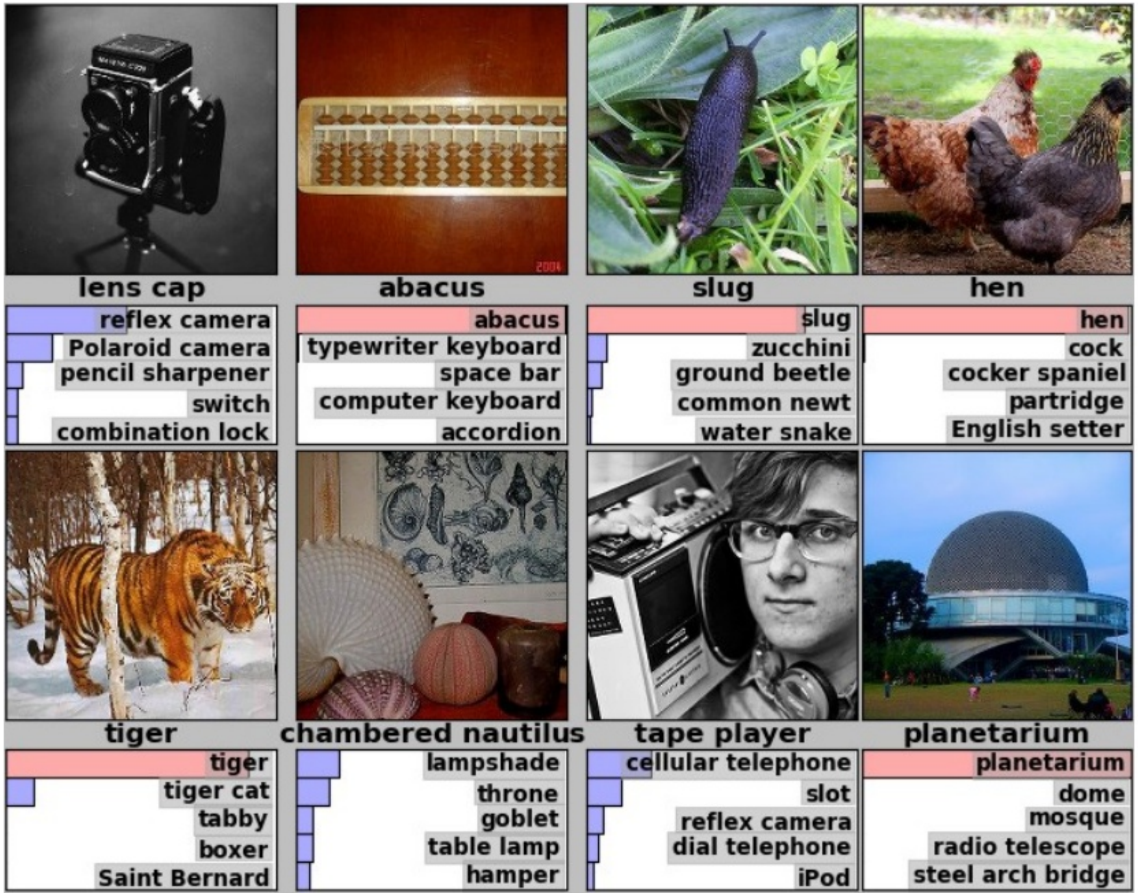
IMAGENET

- 10,000,000 labeled images depicting 10,000+ object categories

The need for machine learning

IMAGENET

Validation classification



Have we been saved?

Image size:
800 × 600

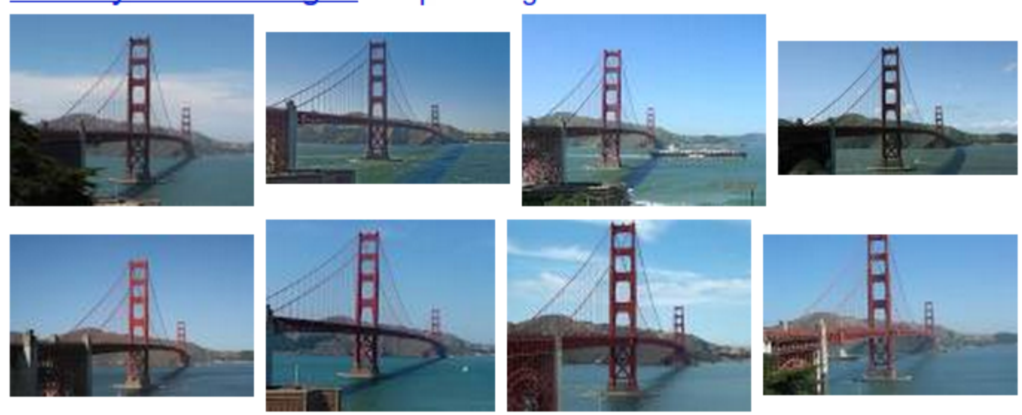
No other sizes of this image found.

Best guess for this image: ***golden gate bridge***

Golden Gate Bridge
www.goldengatebridge.org/
Golden Gate Bridge Highway and Transportation District.

Golden Gate Bridge - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Golden_Gate_Bridge
The **Golden Gate Bridge** is a suspension bridge spanning the Golden Gate, the opening of the San Francisco Bay into the Pacific Ocean. As part of both U.S. **...** 14 images

Visually similar images - Report images



But…

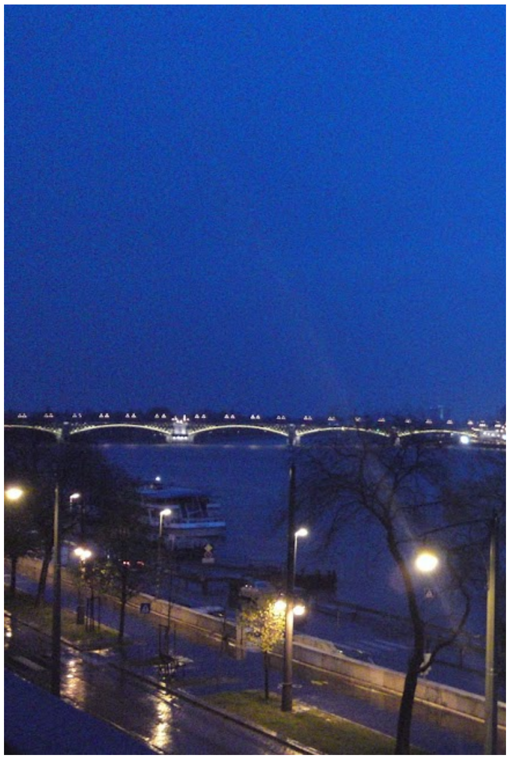


Image size:
1152 × 648

No other sizes of this image found.

Visually similar images - Report images

Conclusion

Vision problems have been increasingly solved using statistical inference

Training data and standardised datasets are a common practice in computer vision

But… might not work in unforseen situations

… Different results for different datasts

… Computational complexity is still a bottleneck for real-time performance

References

- Bosch, Anna and Zisserman, Andrew and Munoz, Xavier (2008). Scene Classification Using a Hybrid Generative/Discriminative Approach. TPAMI, 30(4): 712-727
- Boykov, Yuri and Kolmogorov, Vladimir (2004). An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. TPAMI, 26(9).
- Chen Goldberg (online 2013). On Viola & Jones. Presentation
- Fei-Fei, Li (Online 2013). Machine Learning in Computer Vision. Presentation
- Felzenszwalv, Pedro and Huttenlocher, Dan (2000). Efficient Matching of Pictorial Structures. Computer Vision and Pattern Recognition (CVPR).
- Flach, Peter (2012). Machine Learning – the Art and Science of Making Sense of Data. Cambridge University Press
- Joshi, Ajay and Cherian, Anoop and Shivalingam, Ravishankar (Online 2013). Machine Learning in Computer Vision – A Tutorial. Presentation
- Kokkinos, Iasonas (2012). Machine Learning for Computer Vision. Online Presentation
- Matas, Jiri and Sochman, Jan (Online 2013). AdaBoost. Centre for Machine Perception, Prague. Presentation
- Viola, Paul and Jones, Michael (2001). Robust Real-time Object Detection. Second Intl. Workshop on Statistical and Computational Theories of Vision.
- Hinton, G. E., Osindero, S. and Teh, Y. (2006) A fast learning algorithm for deep belief nets. Neural Computation, 18, pp 1527-1554.