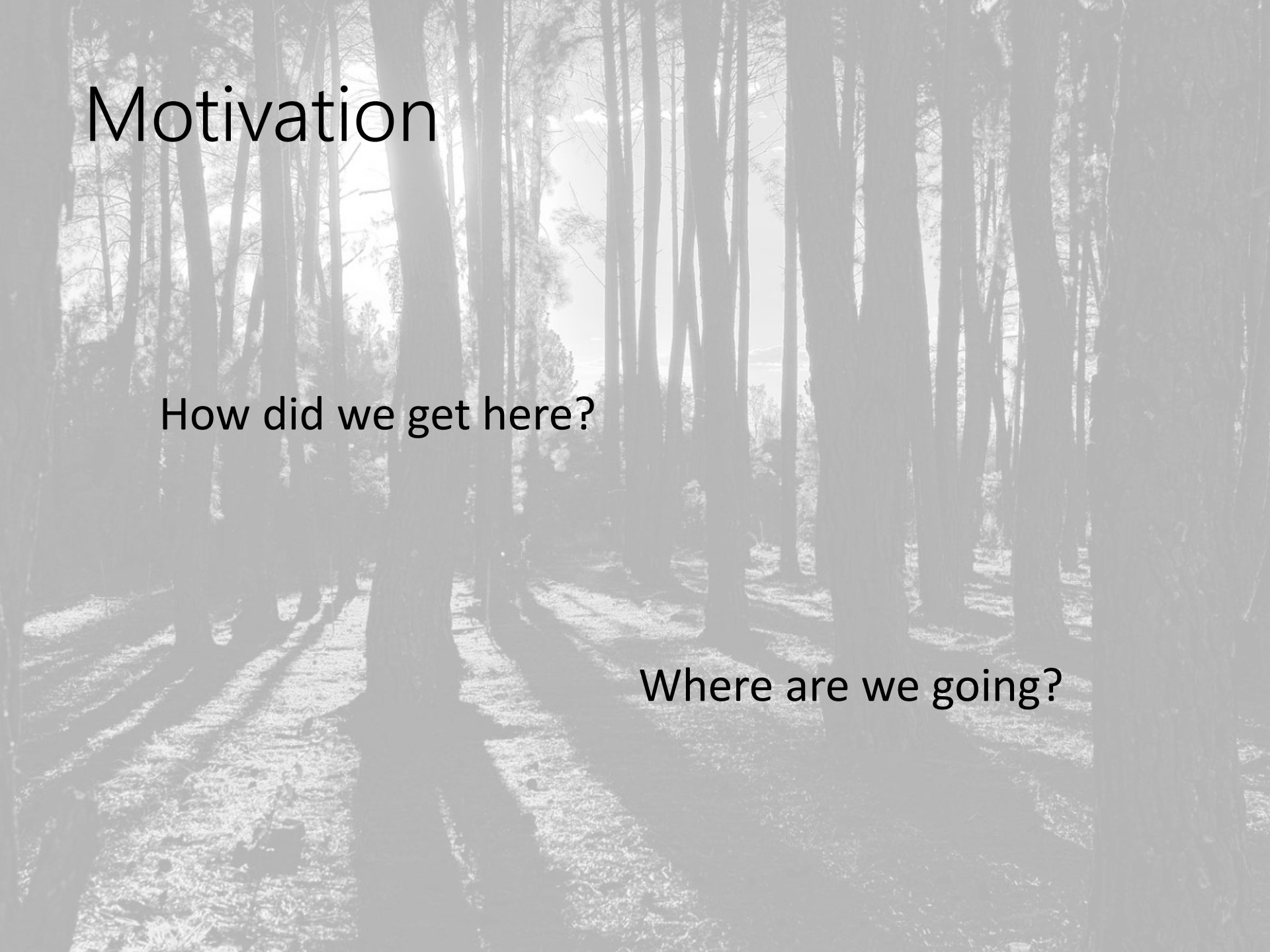
The Dark Ages:

A History of Object Recognition BDL

Larry Zitnick

facebook
Artificial Intelligence Research

Motivation



How did we get here?

Where are we going?

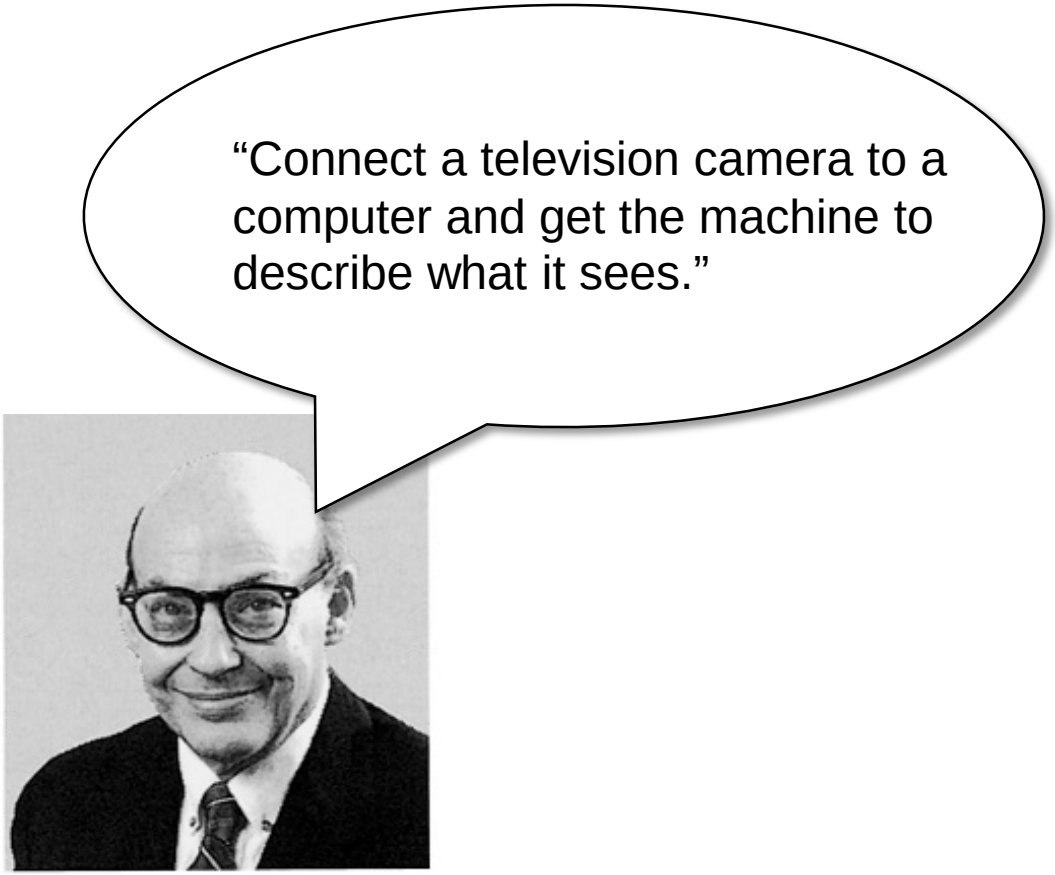
Goals

History object recognition

1960 – 2010 BDL (Before Deep Learning)

Emphasis on algorithms, why they worked or failed!

1966



Marvin Minsky

1968

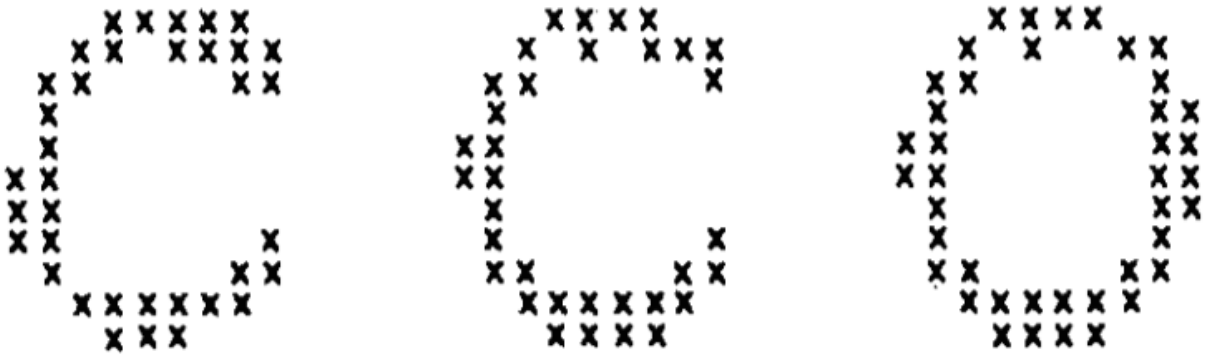


Fig. 4. Correlations in binary patterns. The center pattern, which may be considered the "unknown," differs from each of the outside patterns ("templates") by 9 bit positions. The effect of the correlations among the mismatch bits must thus be taken into account for correct identification. Although this is an artificially constructed example, instances of such neighborhood correlations frequently occur in practice.

The State of the Art in Pattern Recognition, George Nagy, 1968

1973

Picture Processing System By Computer Complex And Recognition Of Human Faces

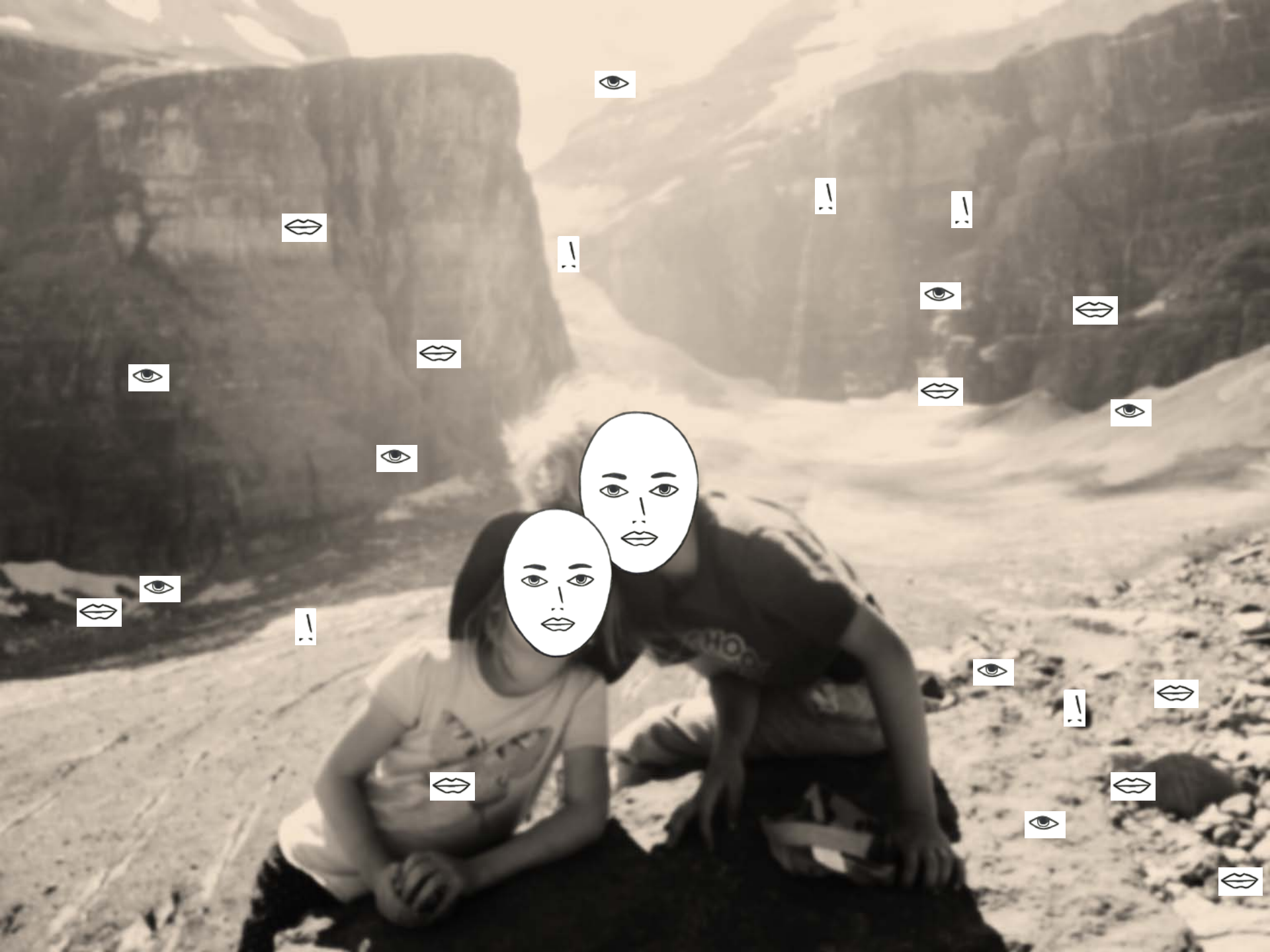
Takeo Kanade

It might be considered that recognition of faces or other natural scenes would be only a little more complex than characters, but actually, completely new aspects appear.

Figure 1-1 Pictures of human face.


Back to the 1970s…

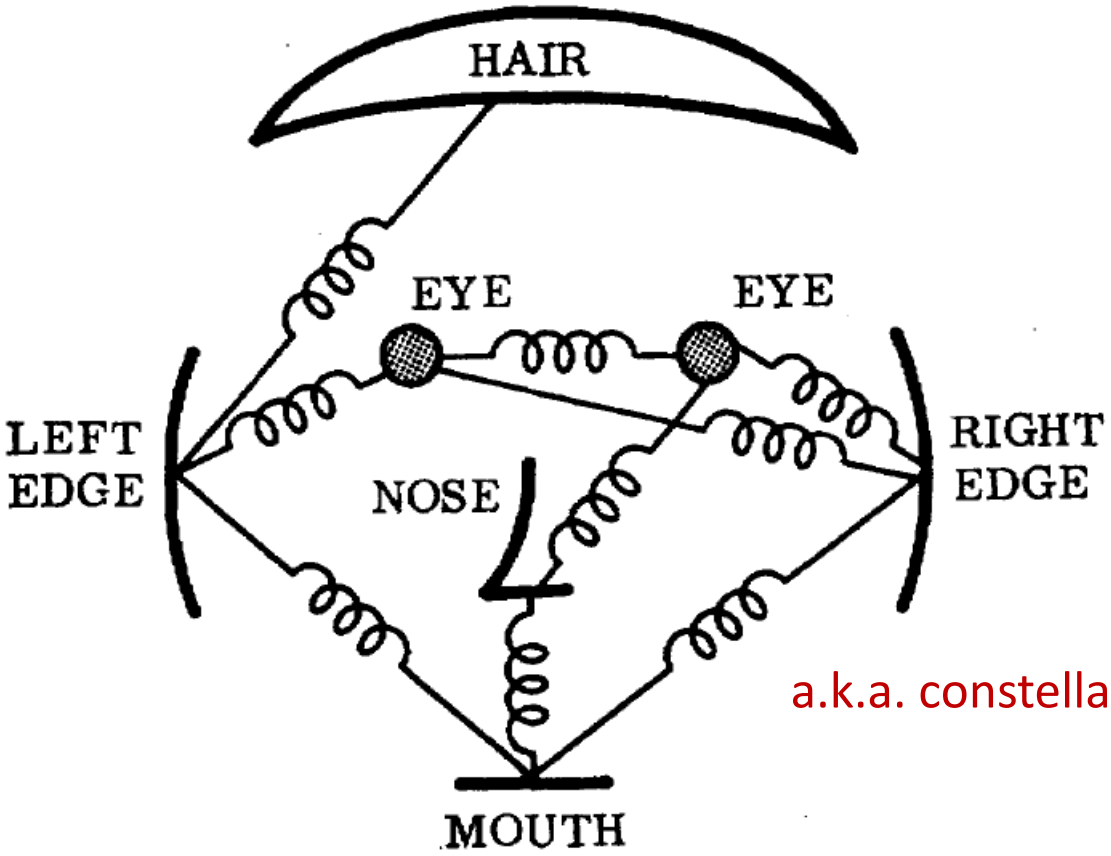
Back to the 1970s…



The “Margaret Thatcher Illusion”, by Peter Thompson

The “Margaret Thatcher Illusion”, by Peter Thompson

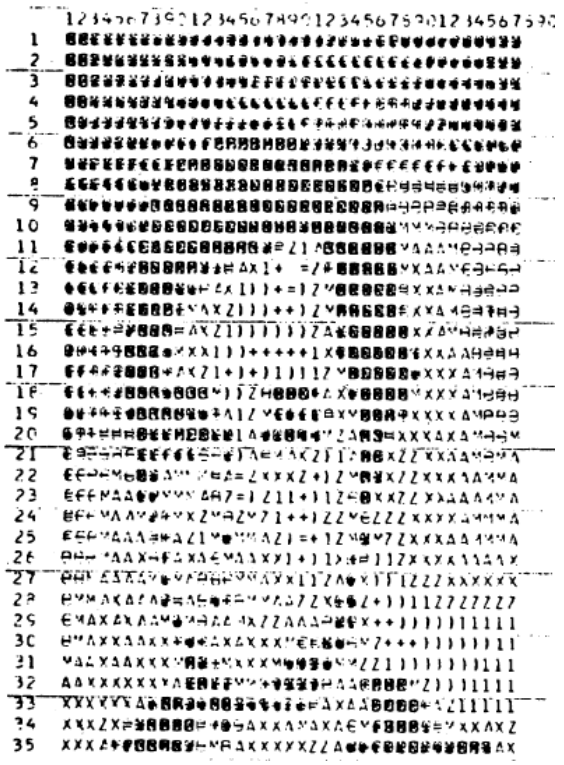
1973

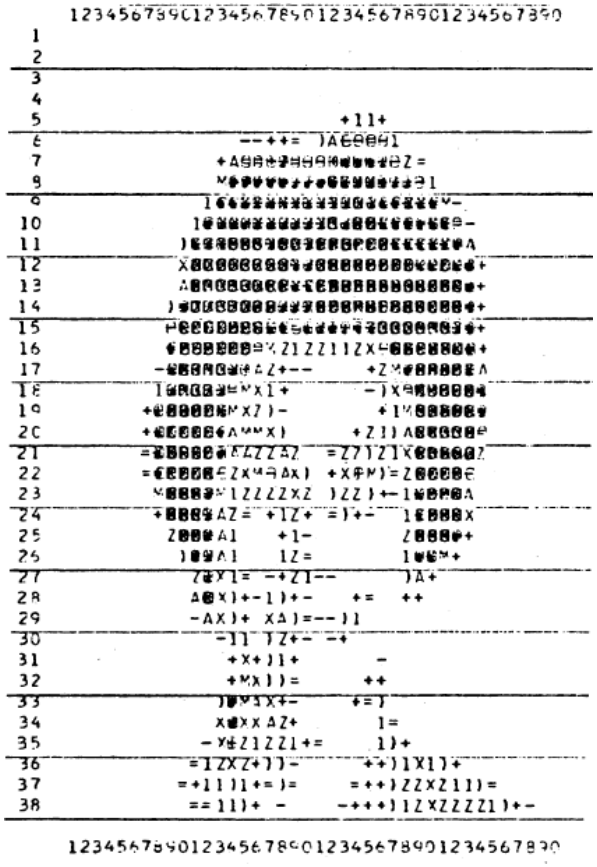


a.k.a. constellation model

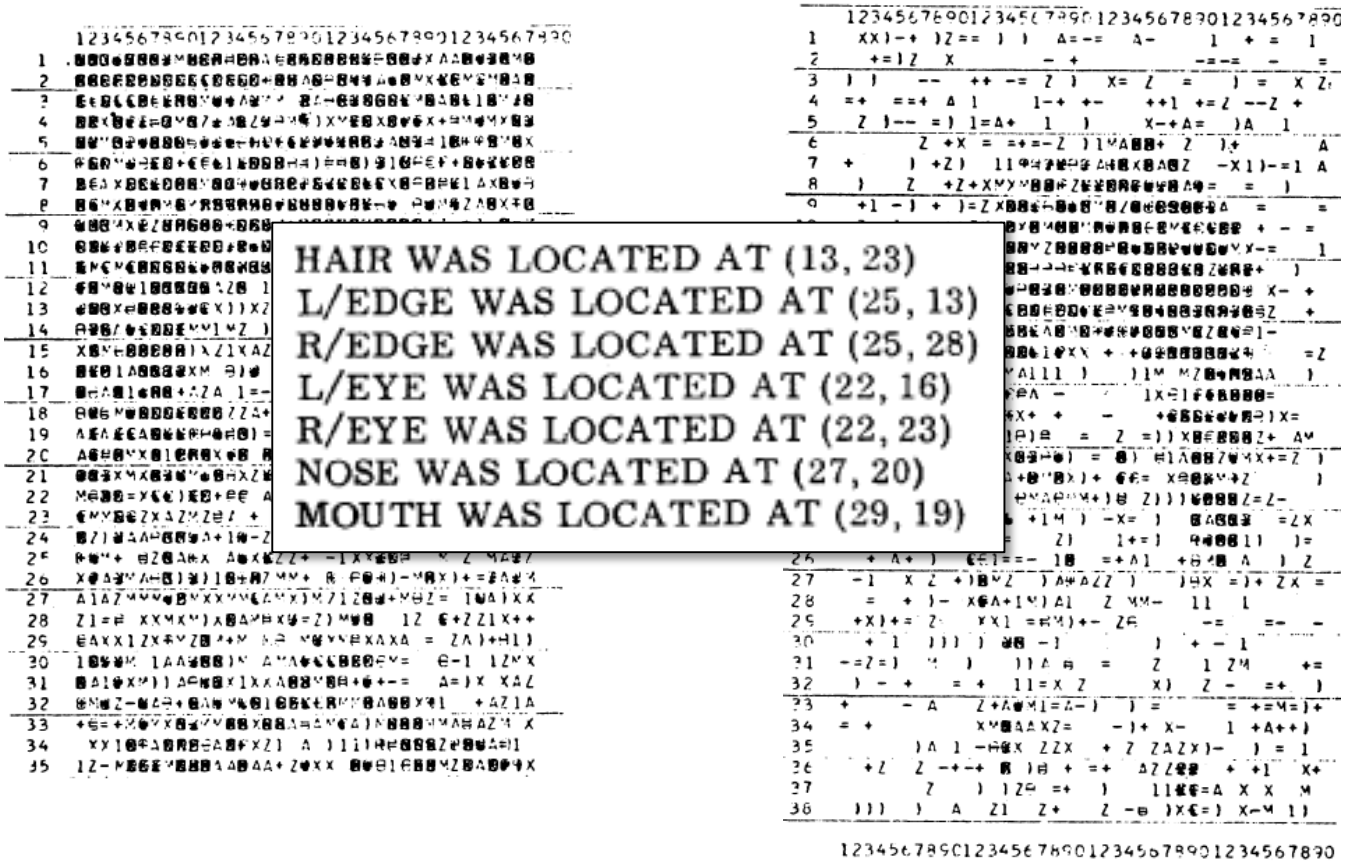
The representation and matching of pictorial structures, Fischler and Elschlager, 1973

1973



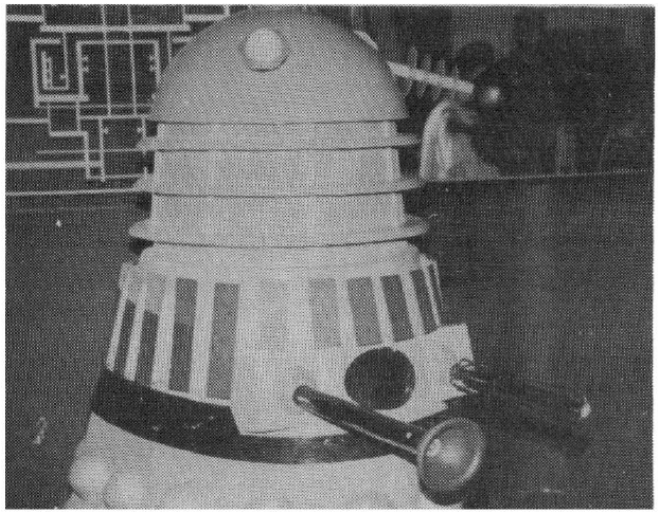


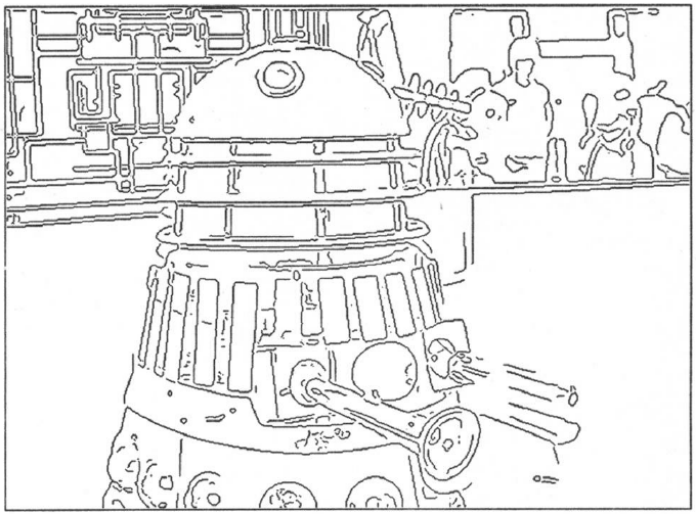
1973


HAIR WAS LOCATED AT (13, 23)
L/EDGE WAS LOCATED AT (25, 13)
R/EDGE WAS LOCATED AT (25, 28)
L/EYE WAS LOCATED AT (22, 16)
R/EYE WAS LOCATED AT (22, 23)
NOSE WAS LOCATED AT (27, 20)
MOUTH WAS LOCATED AT (29, 19)

1980's

AI winter… …back to basics

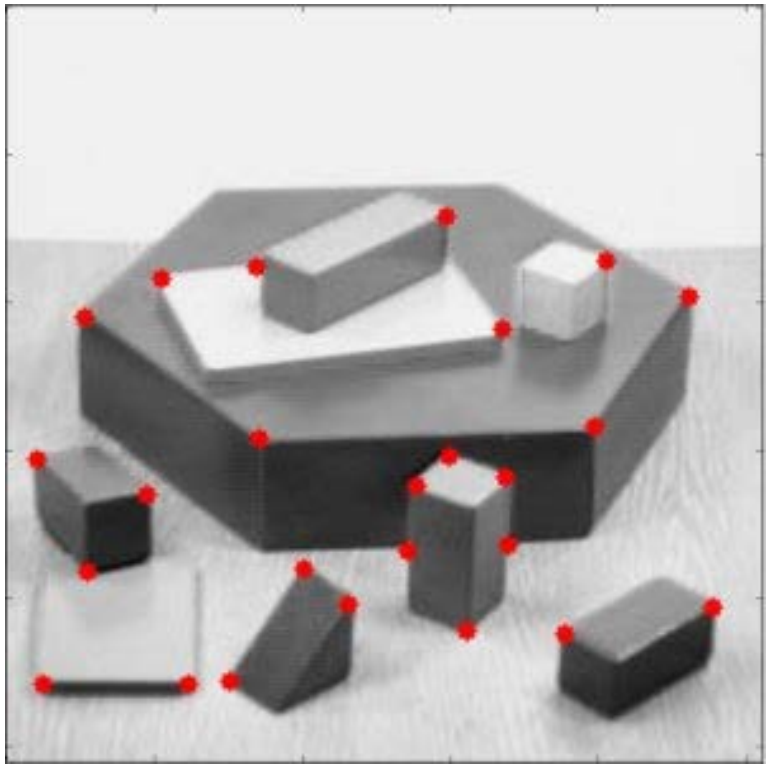




A Computational Approach to Edge Detection, Canny 1986

1980's

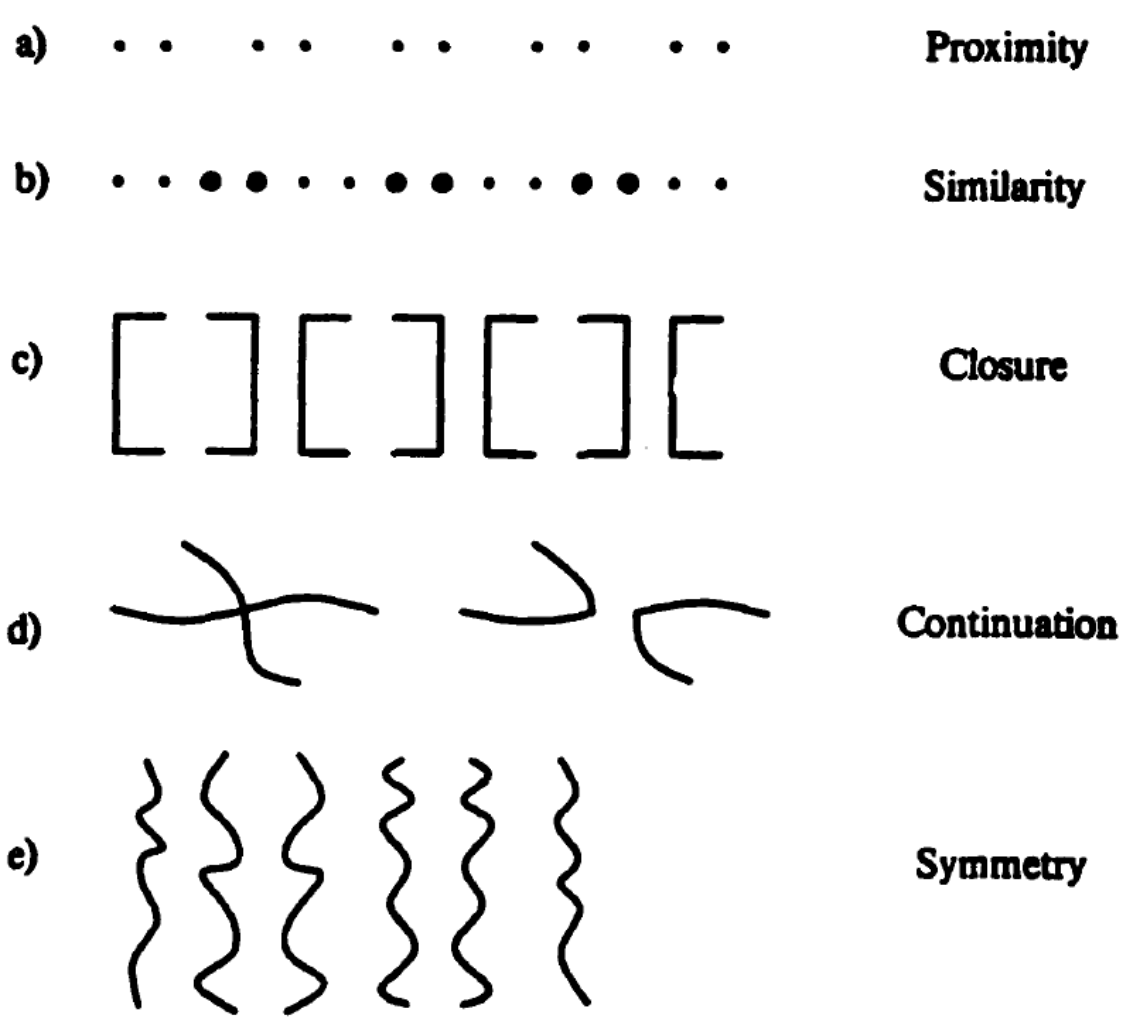
AI winter… …back to basics



Harris Corner Detection

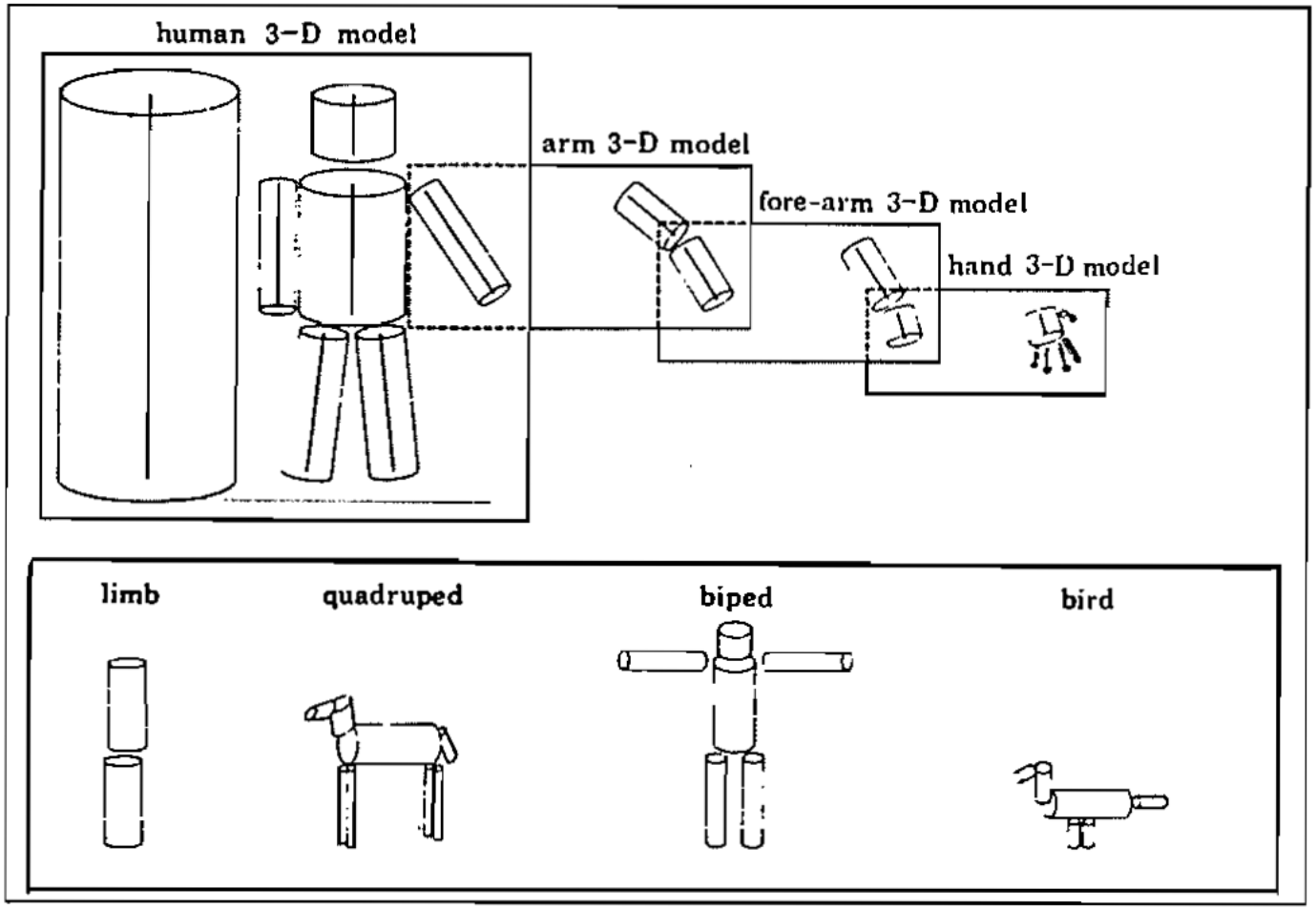
Harris and Stephens, 1988

1984



Perceptual Organization and Visual Recognition,
David Lowe, 1984

1986

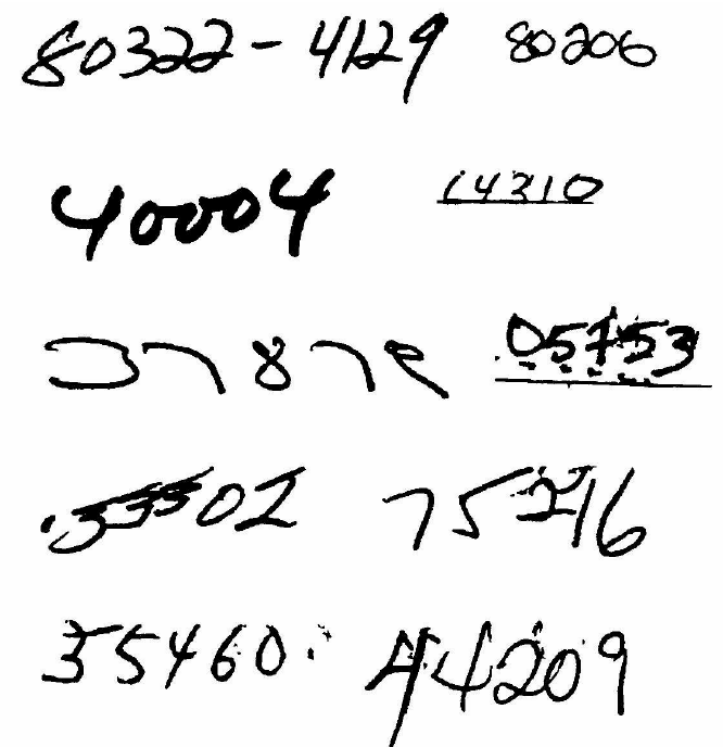


Perceptual organization and the representation of natural form, Alex Pentland, 1986

Goodbye science

1989

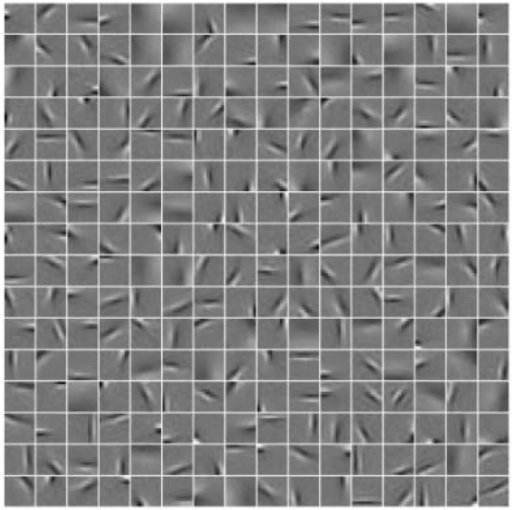
MNIST

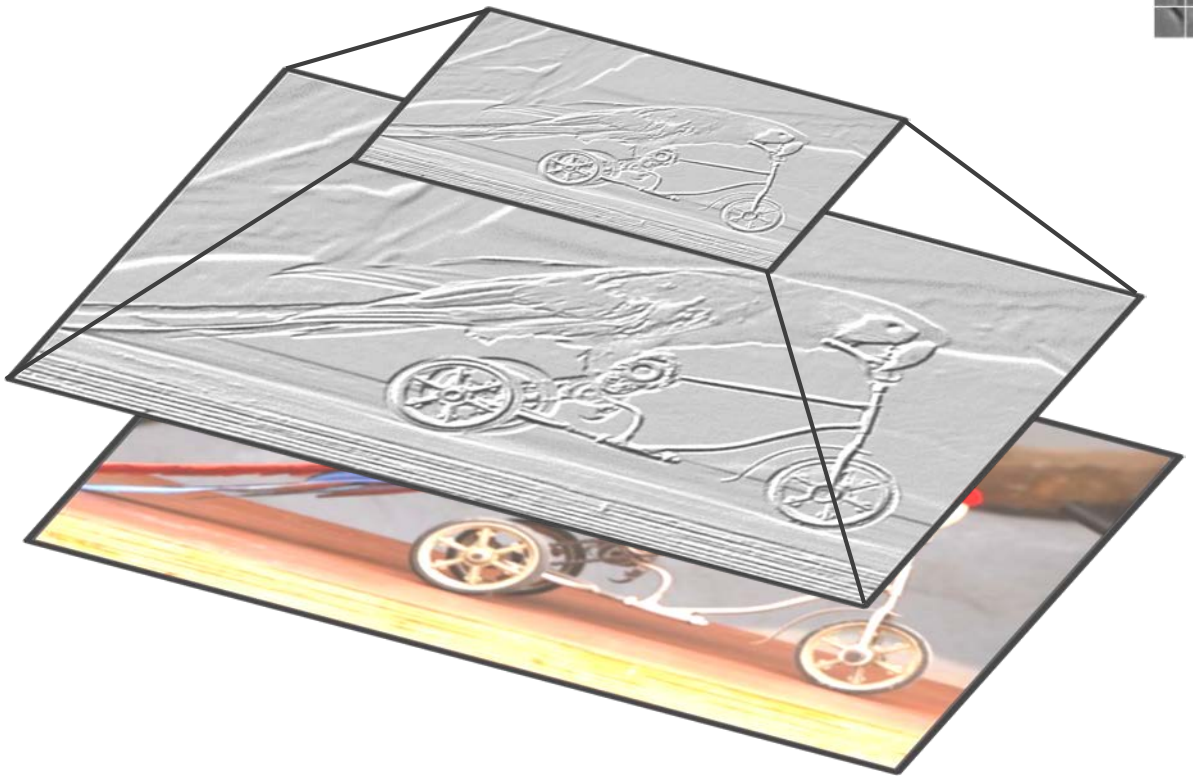


Zip codes

Backpropagation applied to handwritten zip code recognition,
Lecun et al., 1989

1989



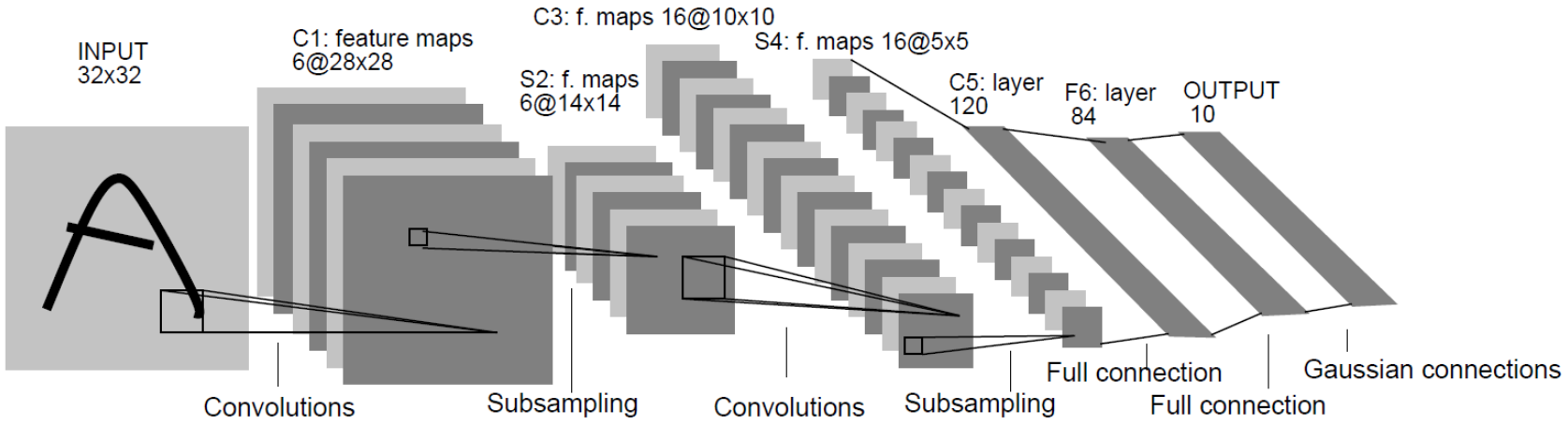


Pooling

Convolution

Image

1989



Backpropagation applied to handwritten zip code recognition,
Lecun et al., 1989

1973

Picture Processing System By Computer Complex And Recognition Of Human Faces

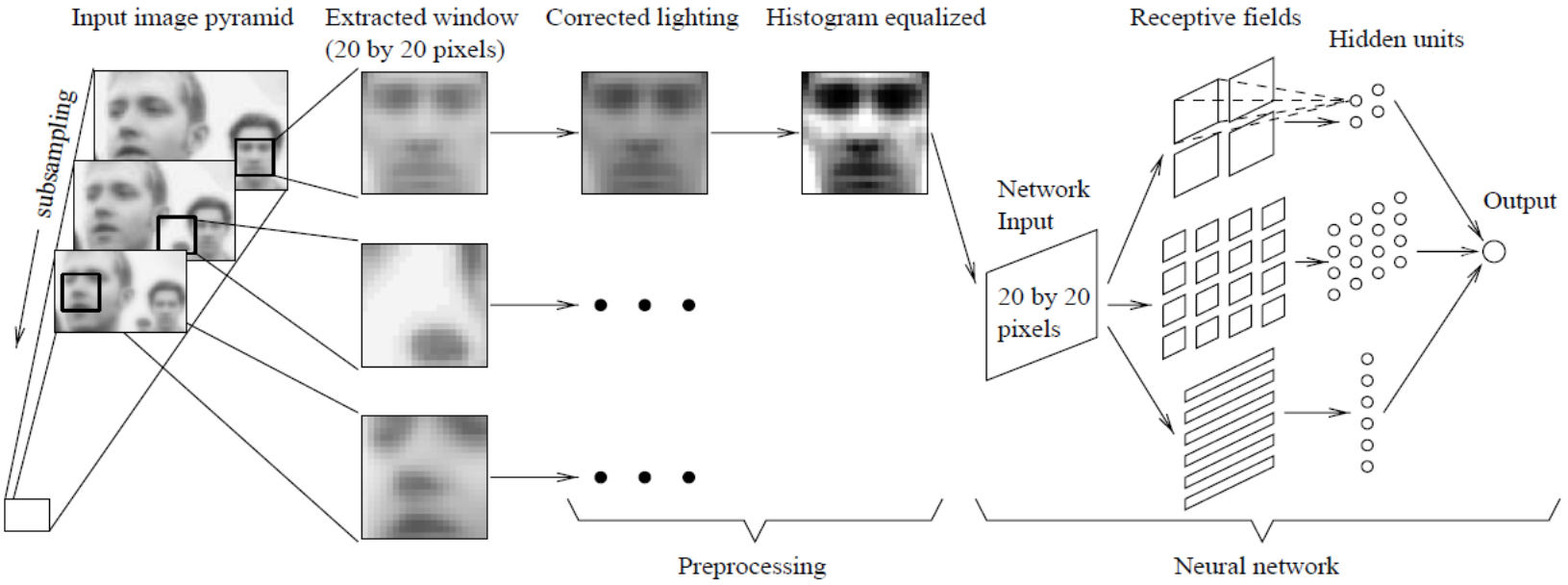
Takeo Kanade

It might be considered that recognition of faces or other natural scenes would be only a little more complex than characters, but actually, completely new aspects appear.

Figure 1-1 Pictures of human face.

1998

Faces



Neural Network-Based Face Detection, Rowley at al., PAMI 1998

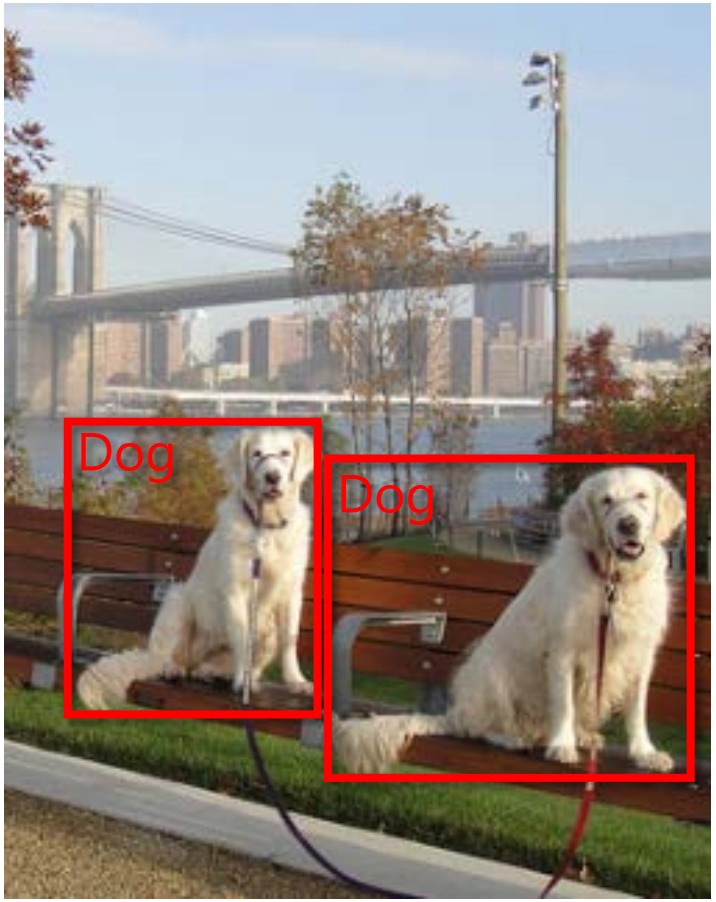
Limits

Numbers worked great, and so did faces…

Why did other categories fail?

3 main reasons… (2012)

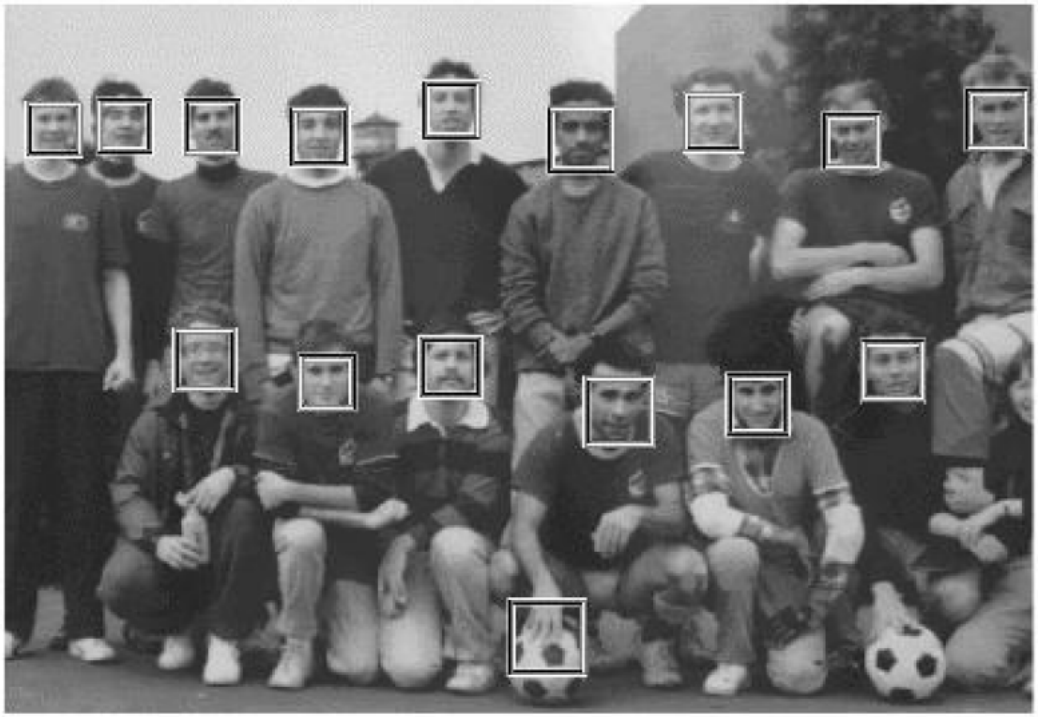
Classification vs. Detection



2001

Sliding window in real time!

Boosting + Cascade = Speed



Rapid Object Detection using a Boosted Cascade of Simple Features, Viola and Jones, CVPR 2001

Why did it work?

- Simple features (Haar wavelets)

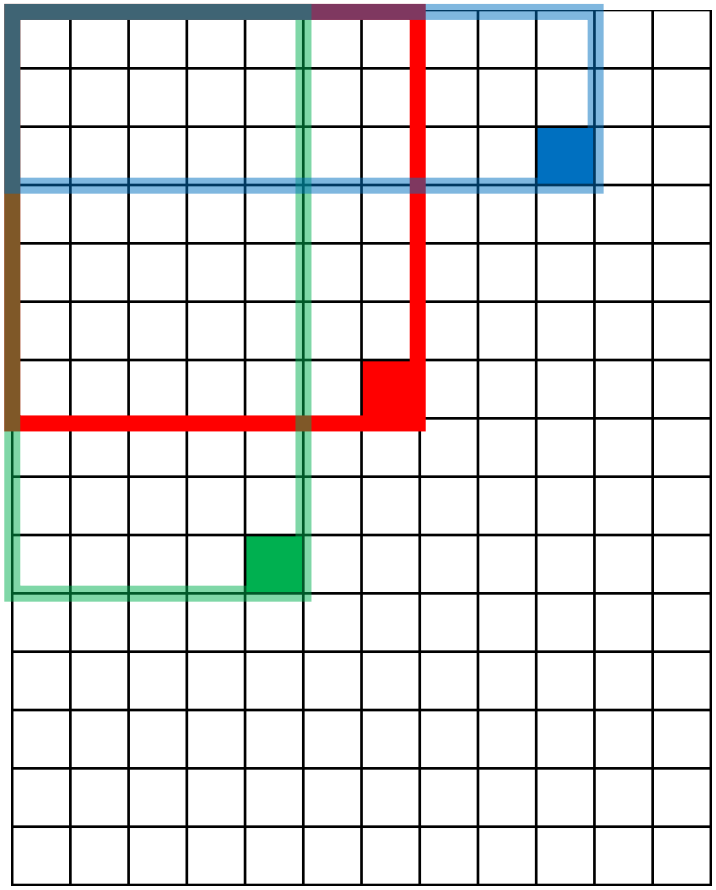
Integral images + Haar wavelets = fast

How do we compute the sum of the pixels in the red box?

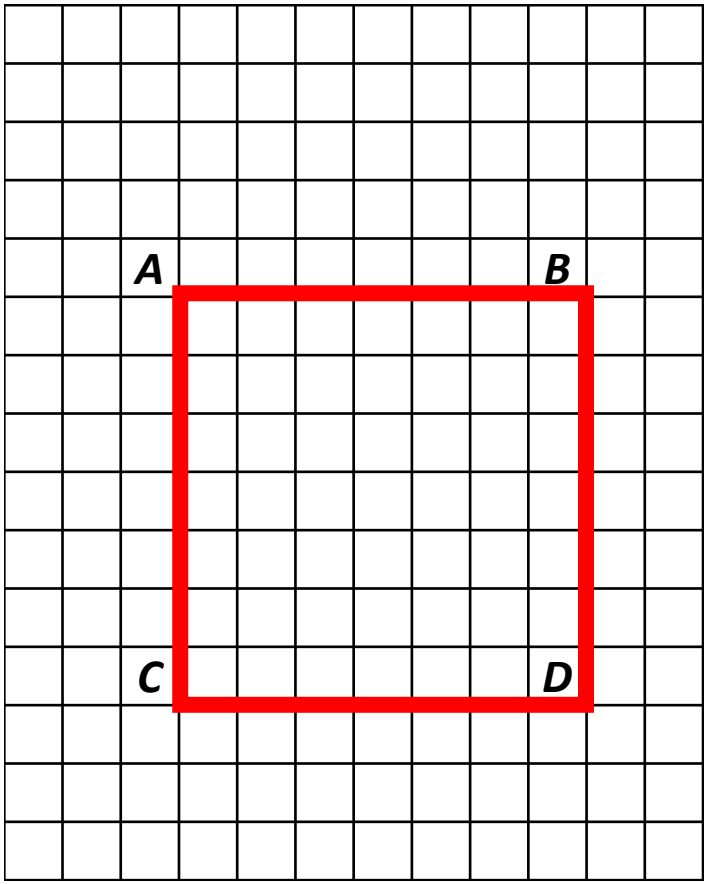
After some pre-computation, this can be done in constant time for any box.

243	239	240	225	206	185	188	218	211	206	216	225
242	239	218	110	67	31	34	152	213	206	208	221
243	242	123	58	94	82	132	77	108	208	208	215
235	217	115	212	243	236	247	139	91	209	208	211
233	208	131	222	219	226	196	114	74	208	213	214
232	217	131	116	77	150	69	56	52	201	228	223
232	232	182	186	184	179	159	123	93	232	235	235
232	236	201	154	216	133	129	81	175	252	241	240
235	238	230	128	172	138	65	63	234	249	241	245
237	236	247	143	59	78	10	94	255	248	247	251
234	237	245	193	55	33	115	144	213	255	253	251
248	245	161	128	149	109	138	65	47	156	239	255
190	107	39	102	94	73	114	58	17	7	51	137
23	32	33	148	168	203	179	43	27	17	12	8
17	26	12	160	255	255	109	22	26	19	35	24

The trick is to compute an "integral image." Every pixel is the sum of its neighbors to the upper left.



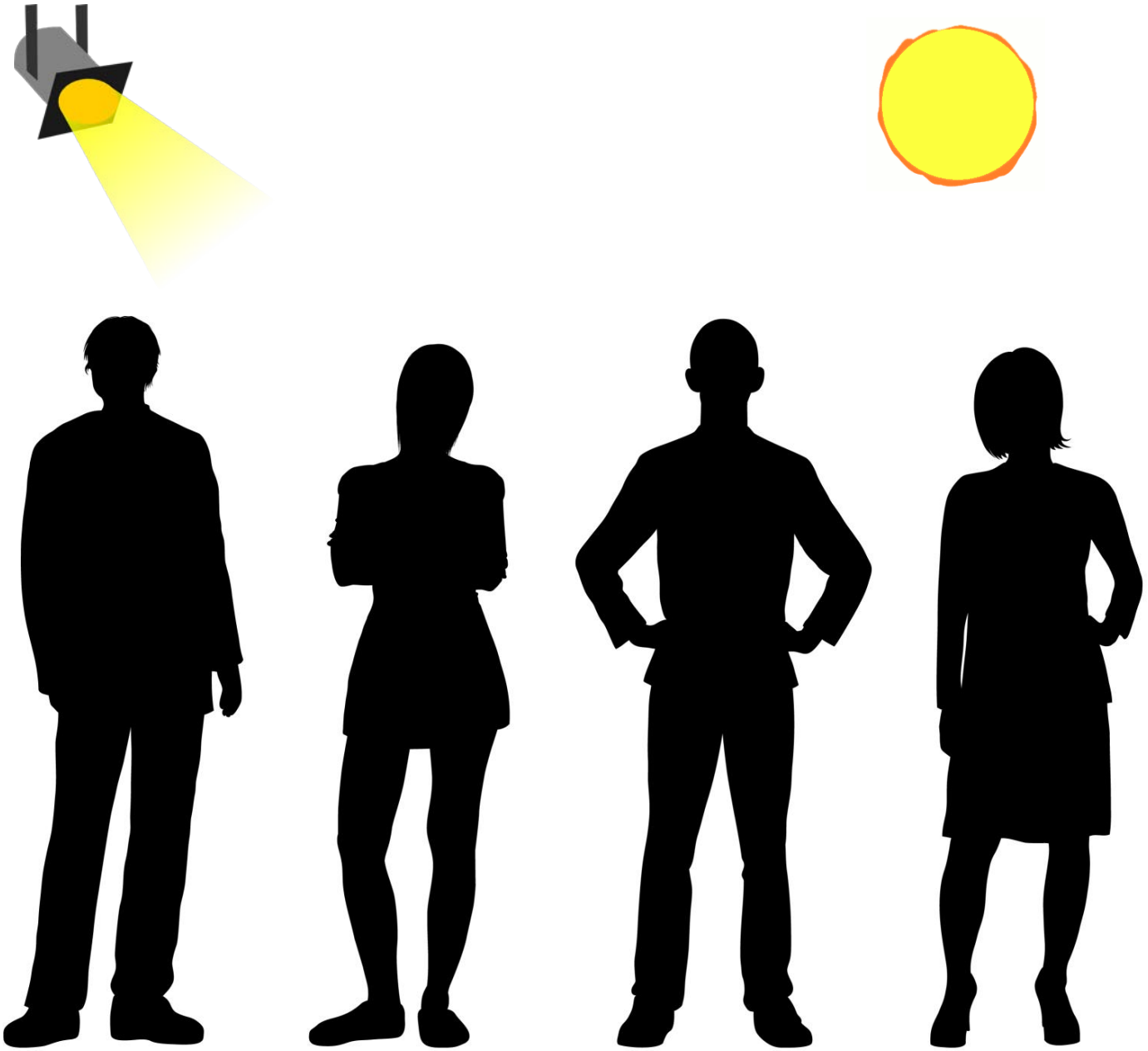
Solution is found using:



Why did it work?

- Simple features (Haar wavelets)

Integral images + Haar wavelets = fast

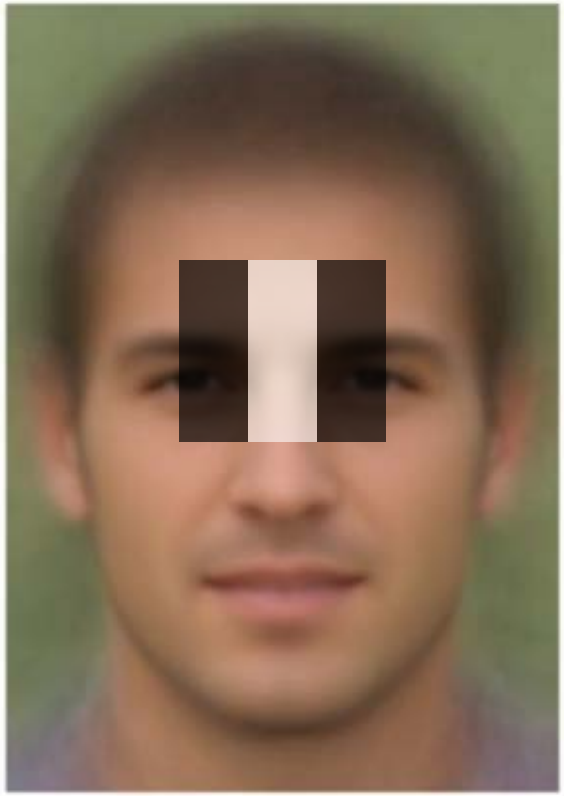


Northern Italian
(*N* = 227)

Central Italian
(*N* = 116)

Southern Italian
(*N* = 152)

Why did it work?



Why did it fail?

1999

In order to recognize objects we need better features…

…but better features are expensive to compute.

Where do humans fixate?

Called “fixation points”, and a “saccade” is the process of moving between fixation points.

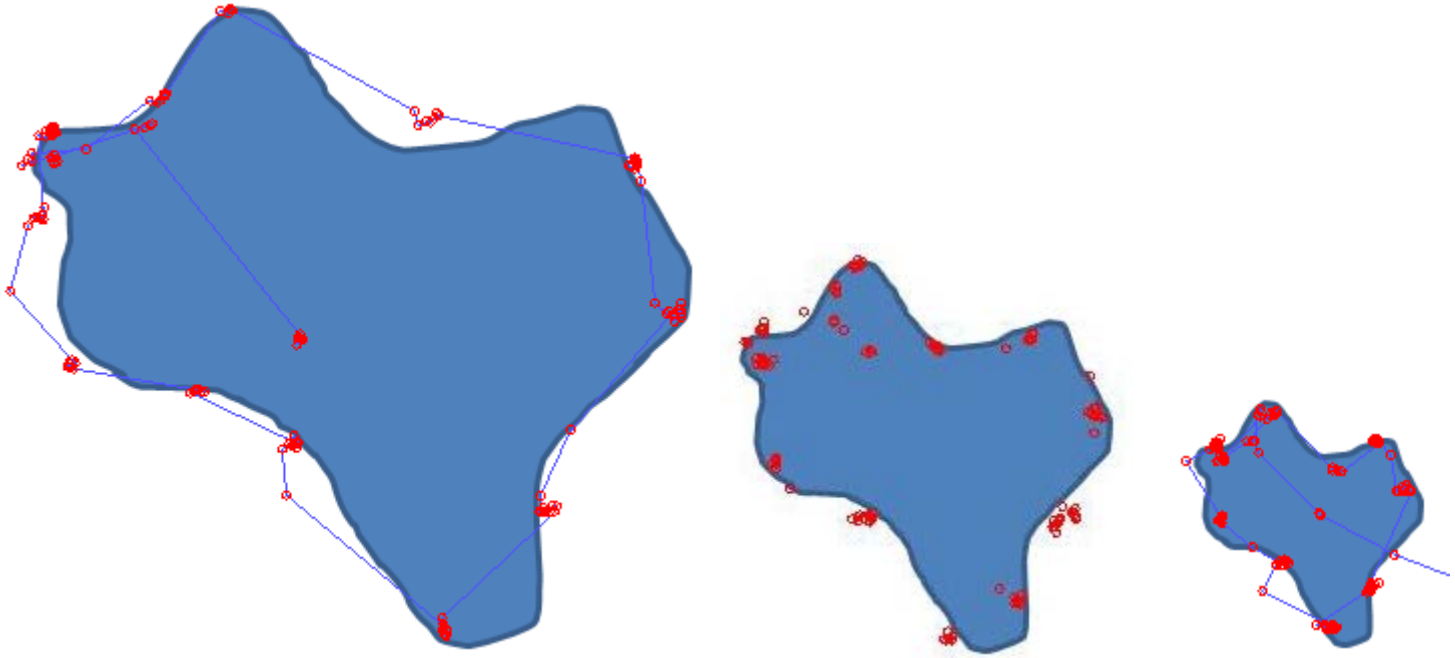
Slowly trace the outline of the above object.

Where do humans fixate?

Called “fixation points”, and a “saccade” is the process of moving between fixation points.

Slowly trace the outline of the above object.

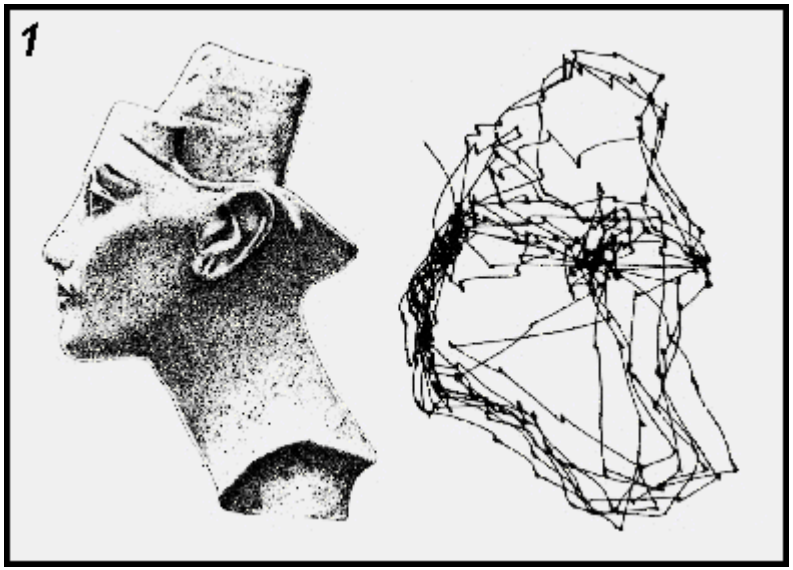
Where do humans fixate?



Result of one subject “me”.

Where do humans fixate?

Top down or bottom up?



"Eye Movements and Vision" by A. L. Yarbus; Plenum Press, New York; 1967

1999* SIFT (scale invariant feature transform)

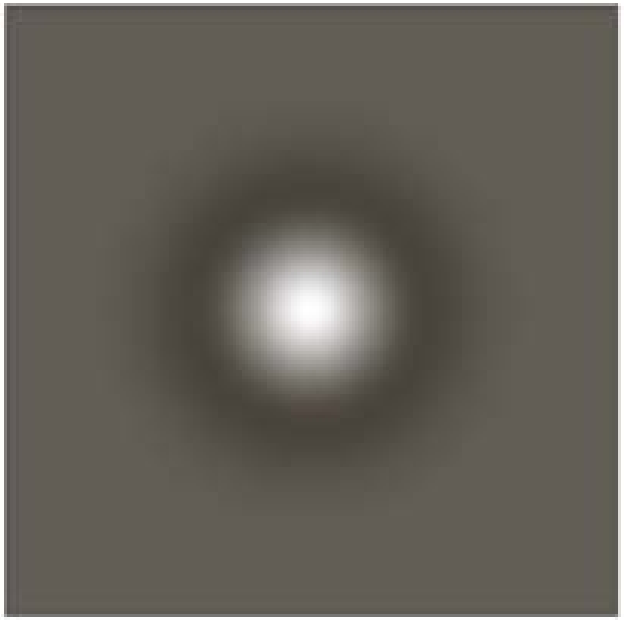
No more sliding windows (interest points)

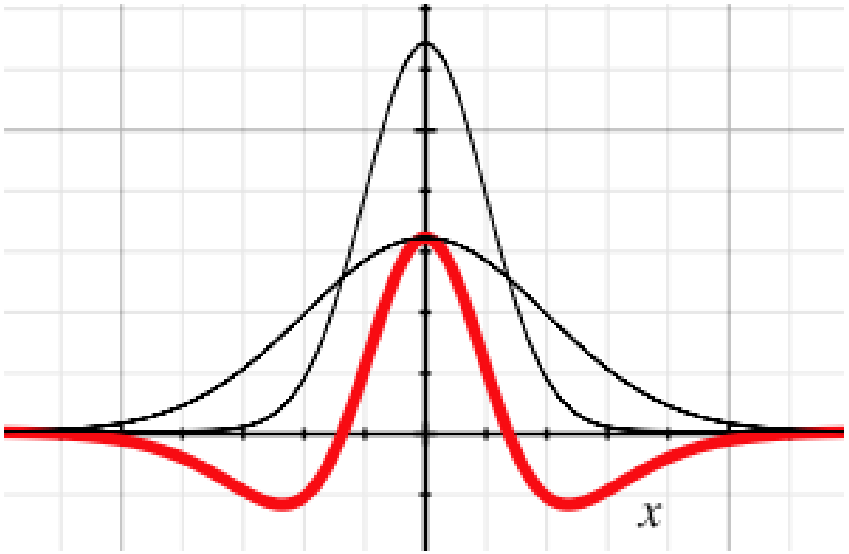
Better features (use more computation)

Object Recognition from Local Scale-Invariant Features,
Lowe, ICCV 1999.

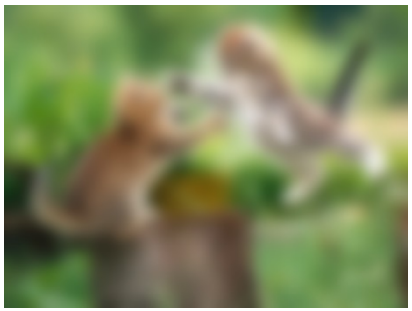
Difference of Gaussians

Look for blobs of dark or light color.

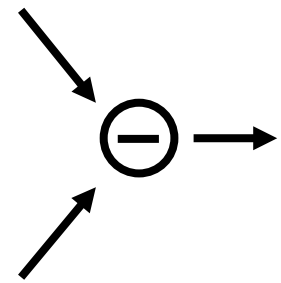


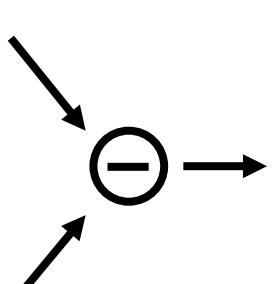


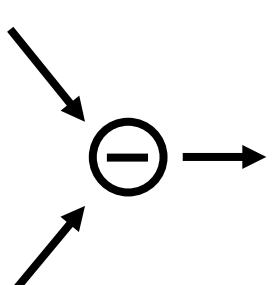
Less blur

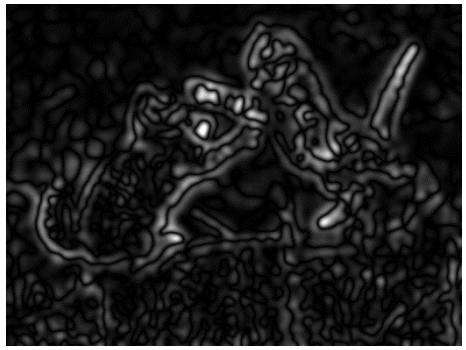


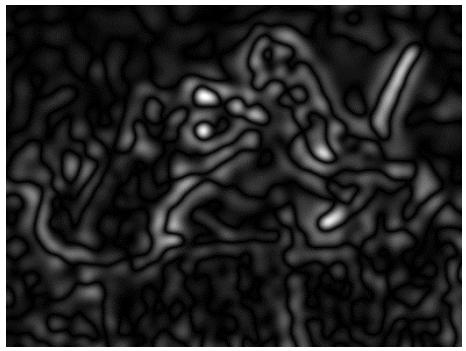
More blur





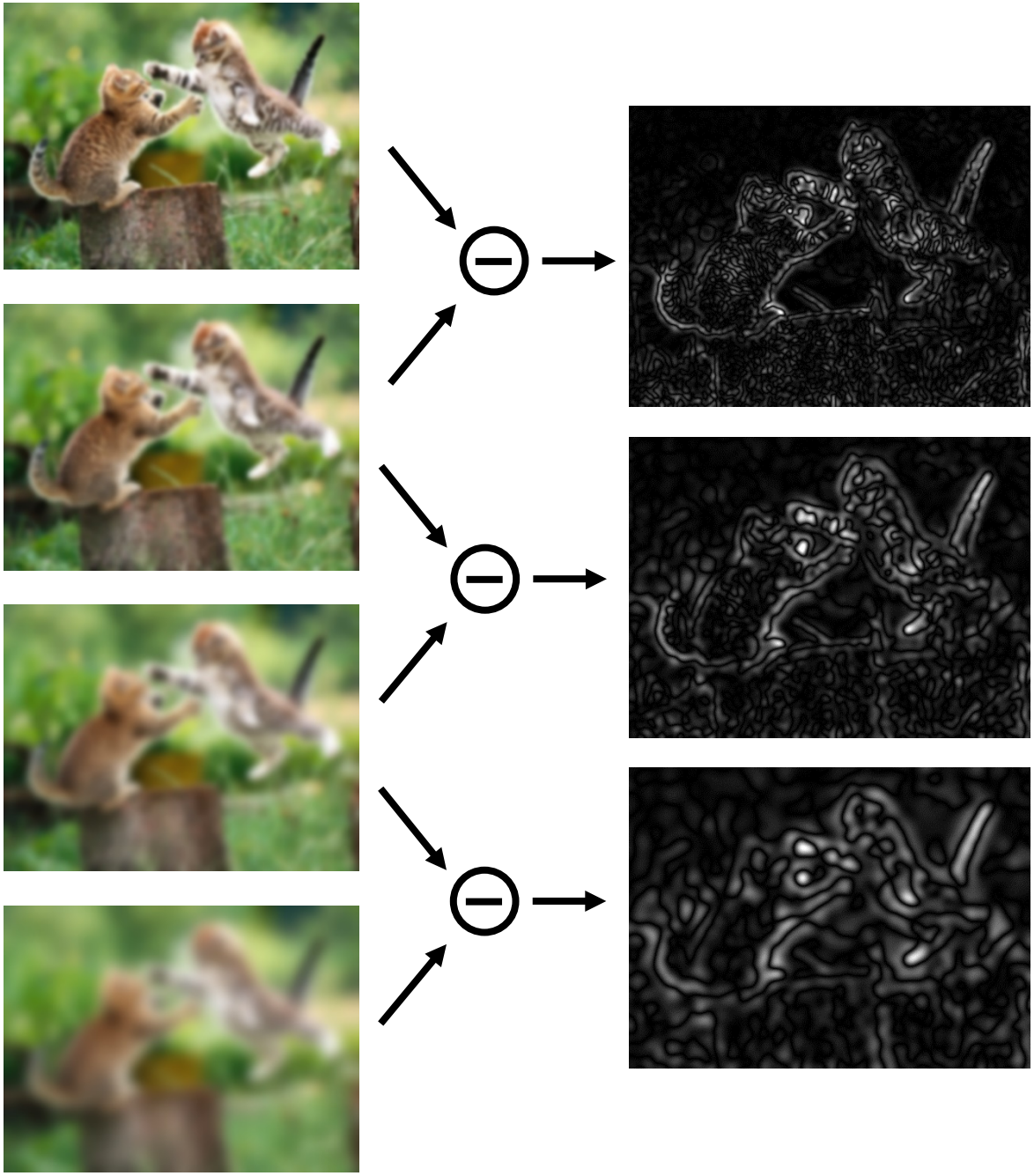




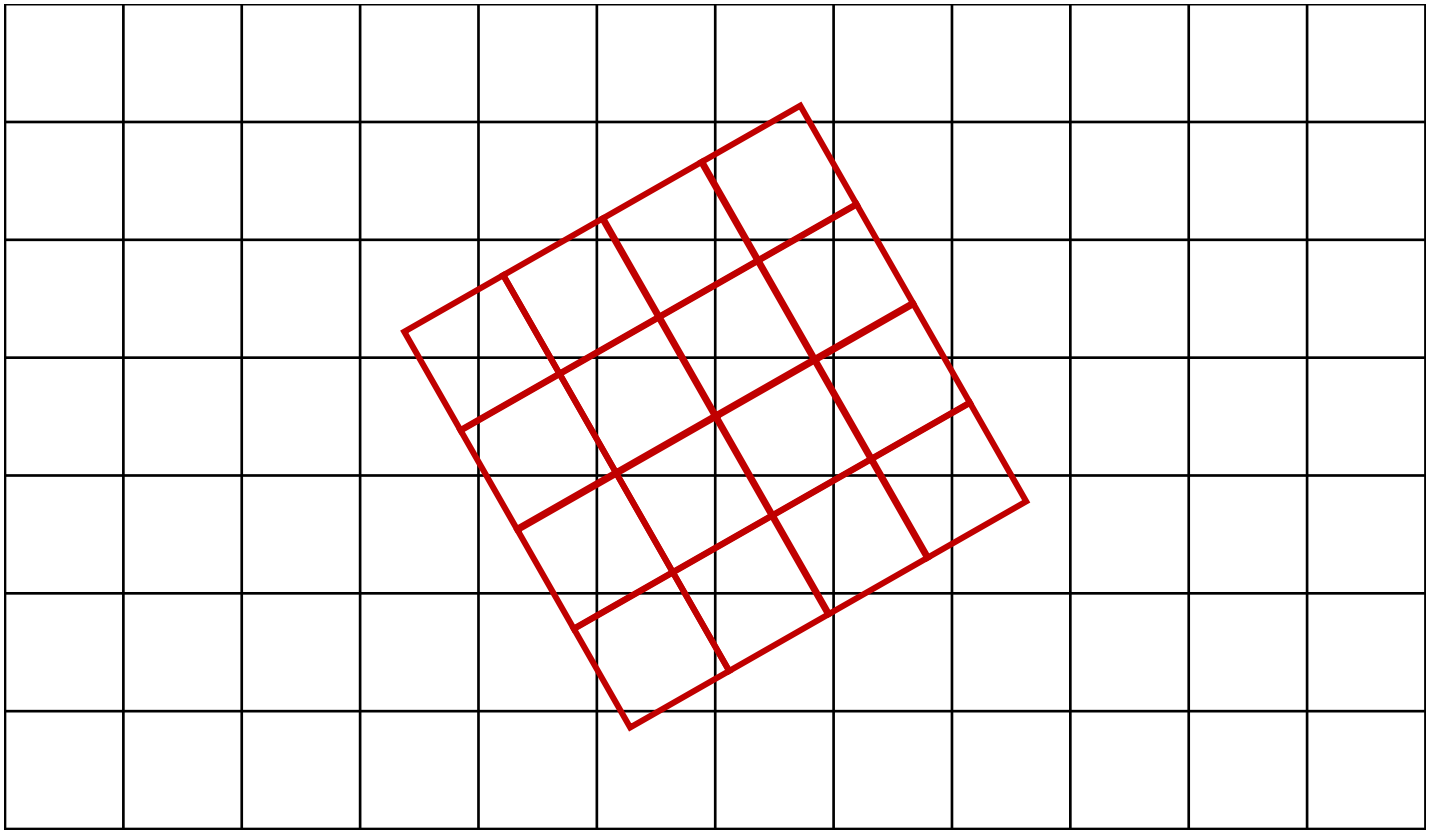


Look for peaks!

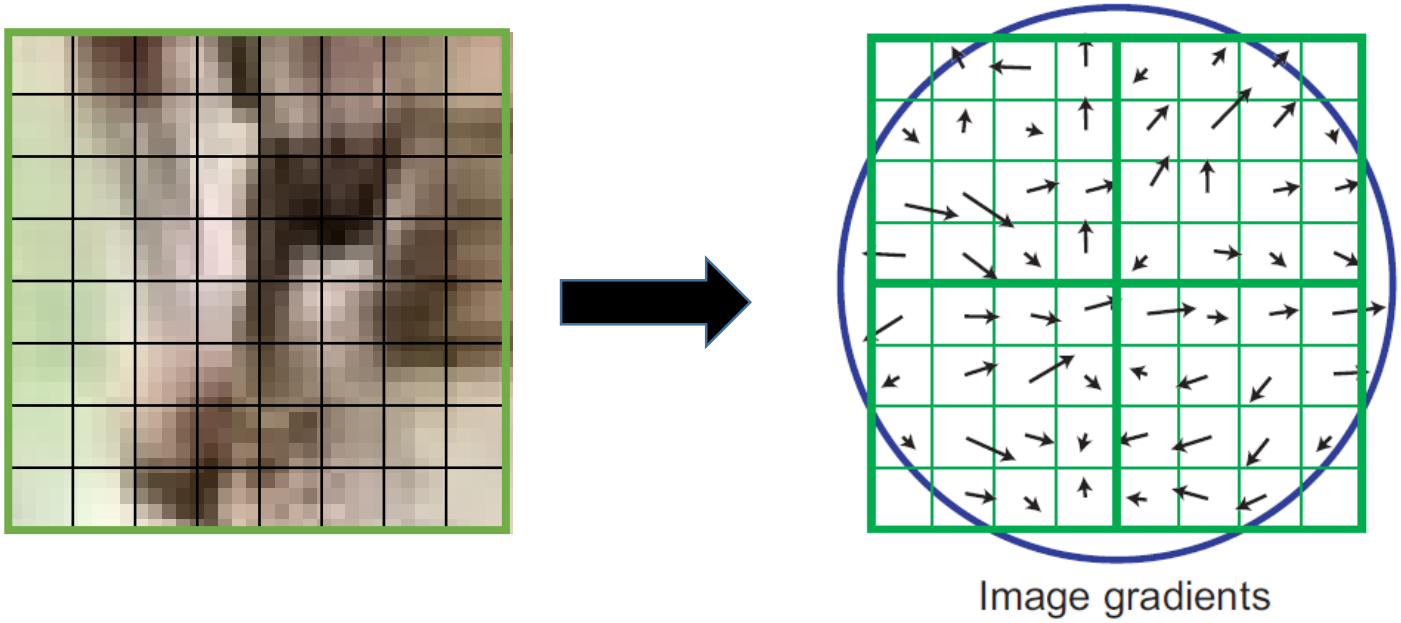
Less blur



More blur

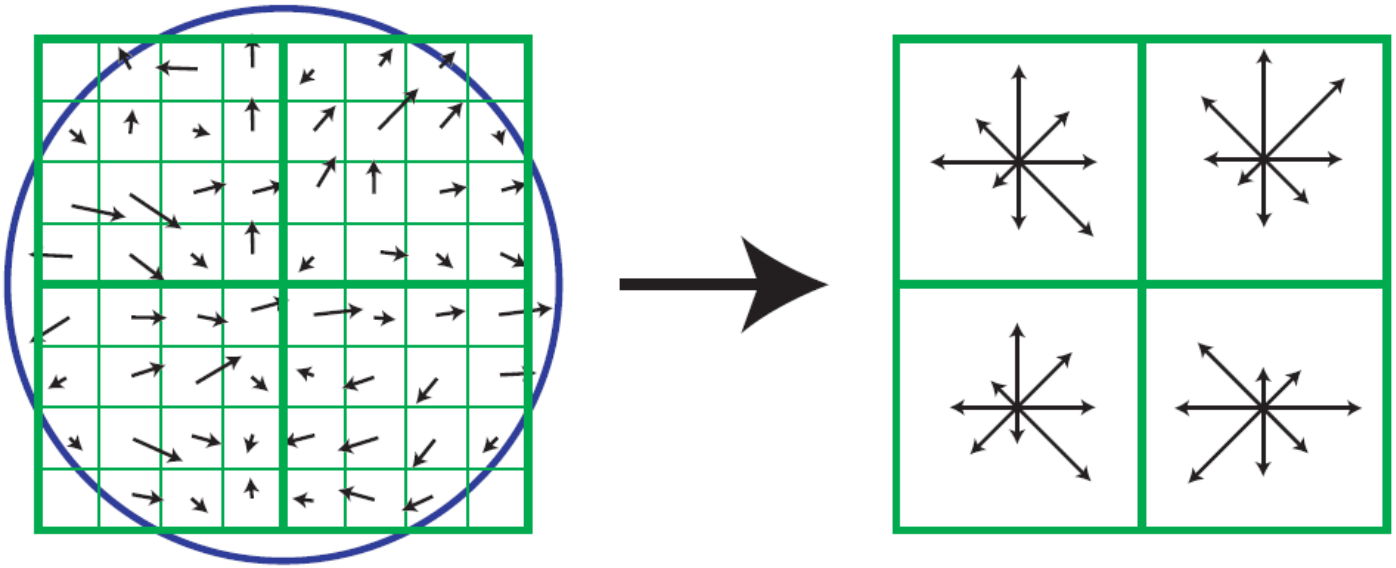


Better descriptor:



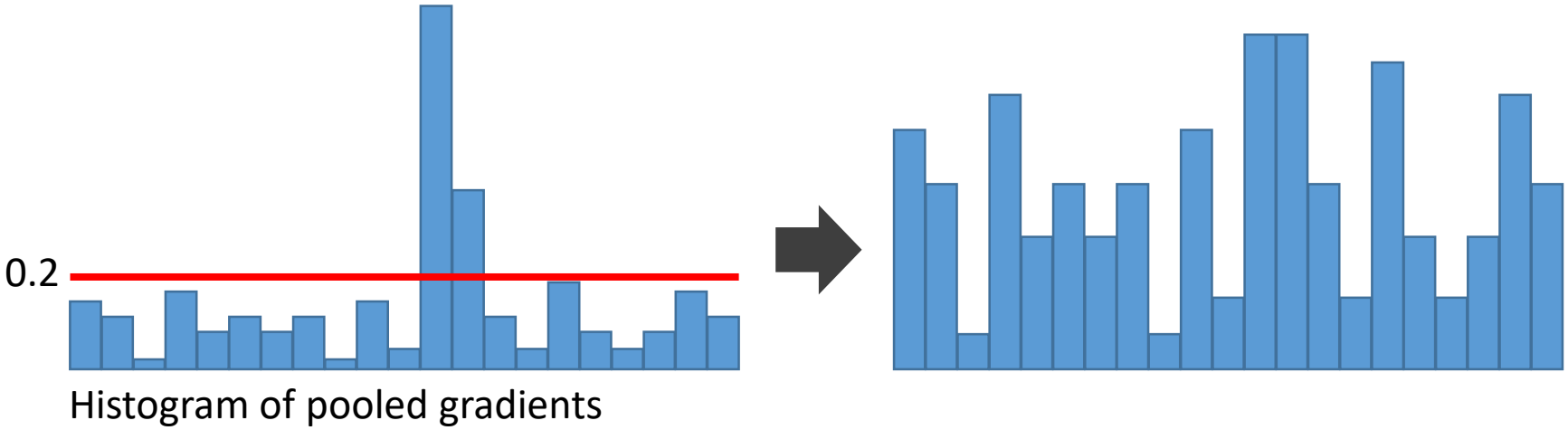
Distinctive image features from scale-invariant keypoints, Lowe, *IJCV* 2004

Better descriptor:



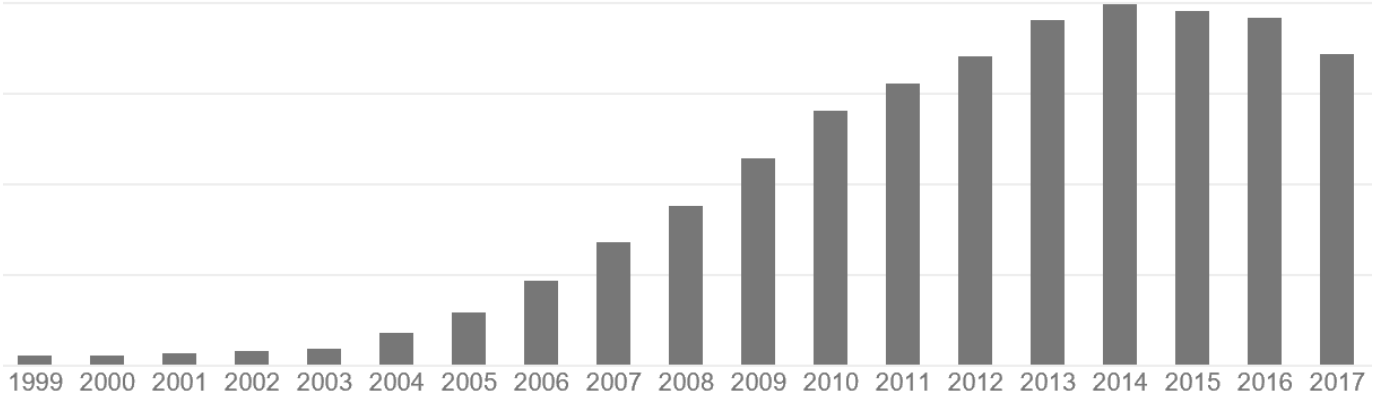
Distinctive image features from scale-invariant keypoints, Lowe, *IJCV* 2004

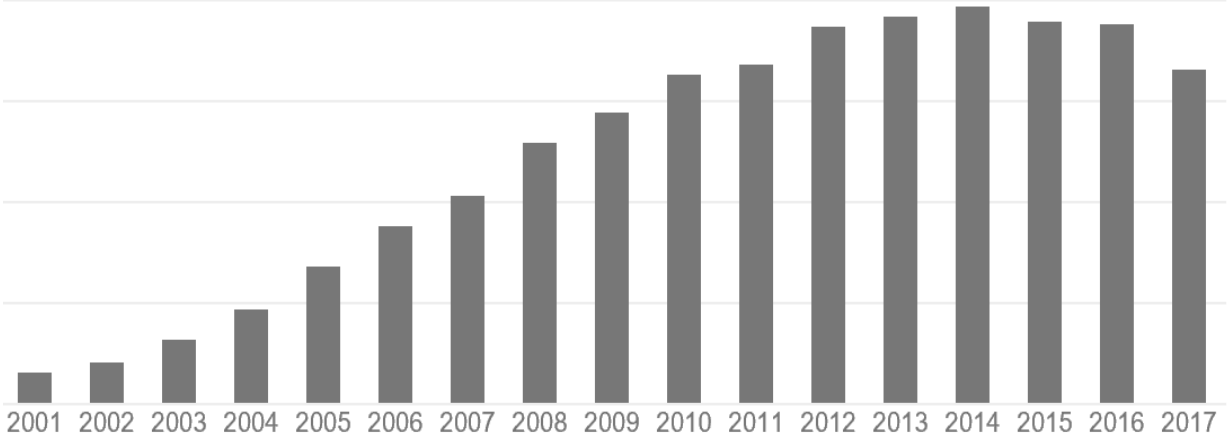
Truncated normalization



Distinctive image features from scale-invariant keypoints, Lowe, *IJCV* 2004

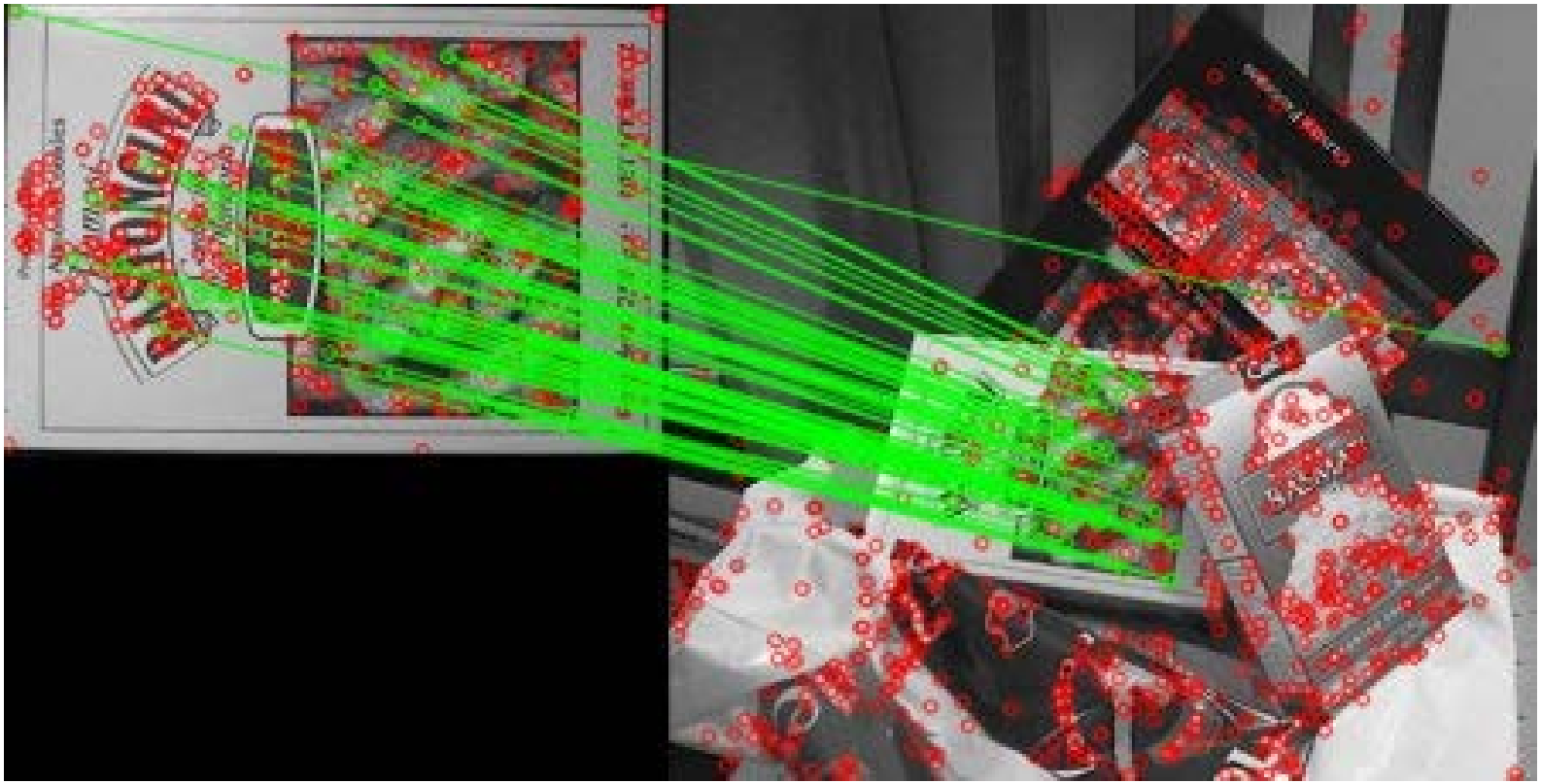
1999* SIFT


1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017


2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017

What worked!

“Object instance” recognition (flat, textured objects) CDs!



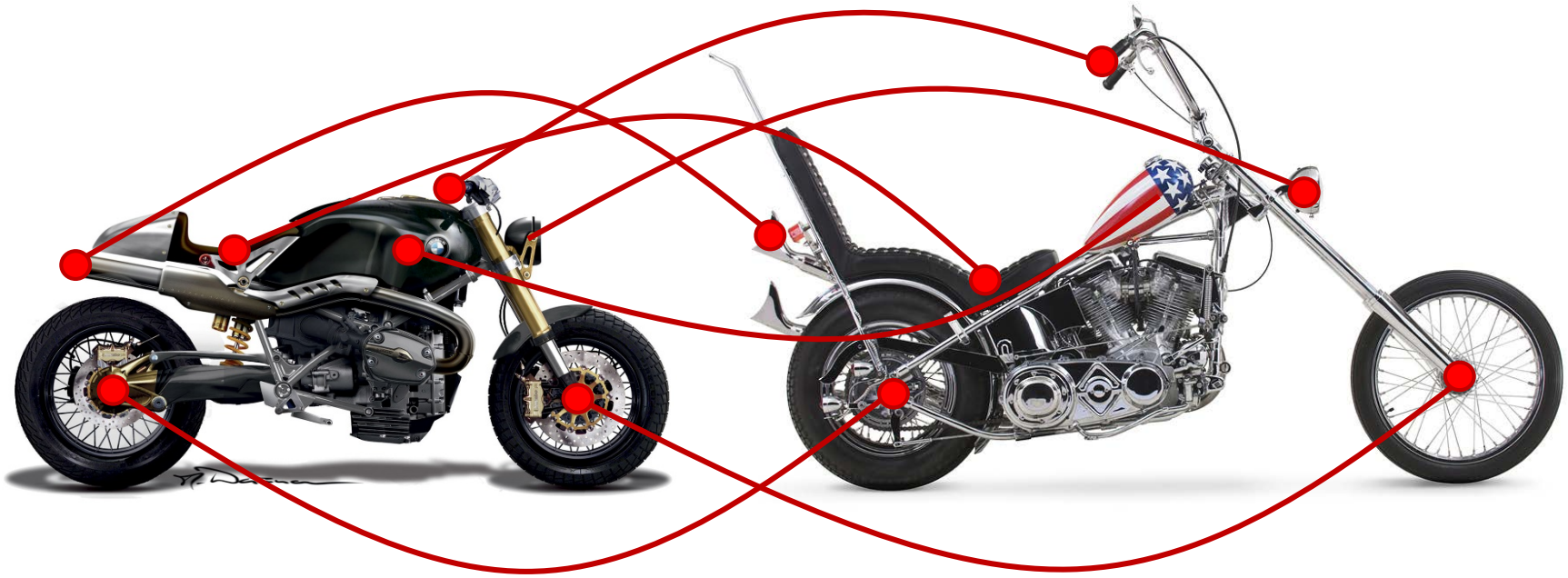
What worked

Panorama stitching

Recognizing panoramas, Brown and Lowe, *ICCV* 2003

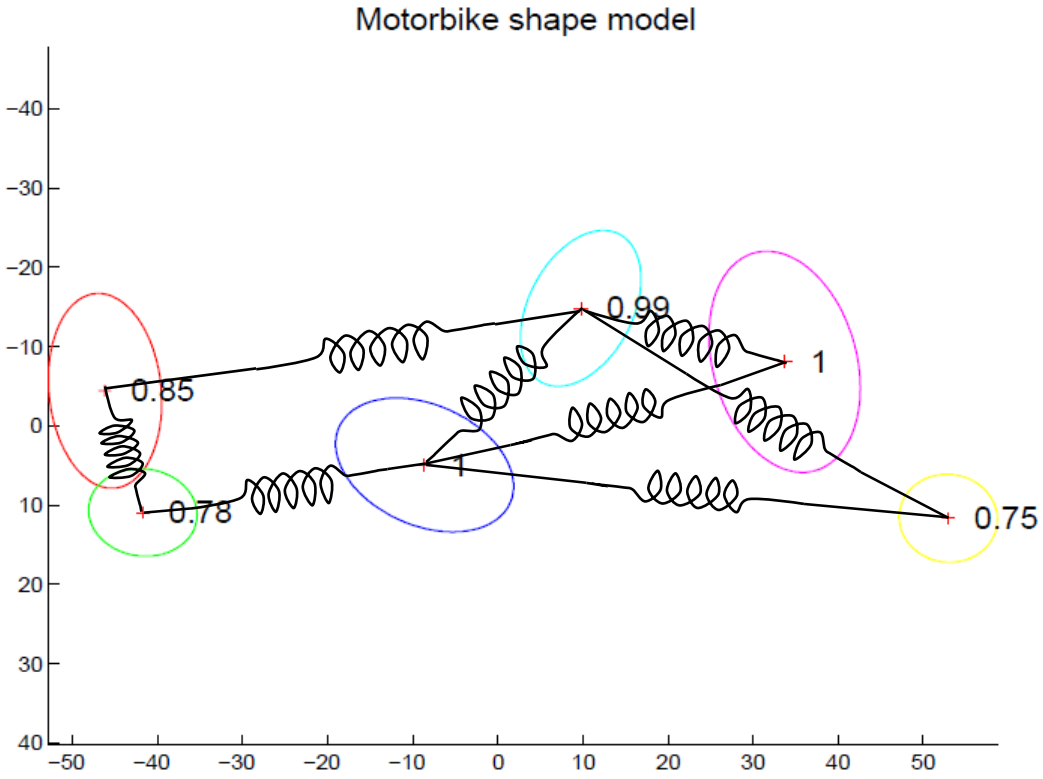
What failed…

2003 Constellation model (redux)

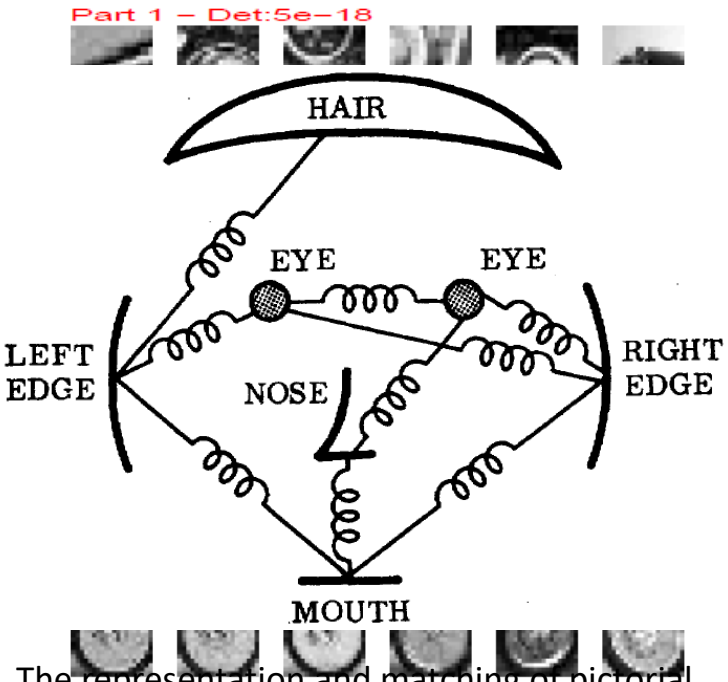


Object Class Recognition by Unsupervised Scale-Invariant Learning, Fergus et al., *CVPR* 2003.

2003 Constellation model (redux)

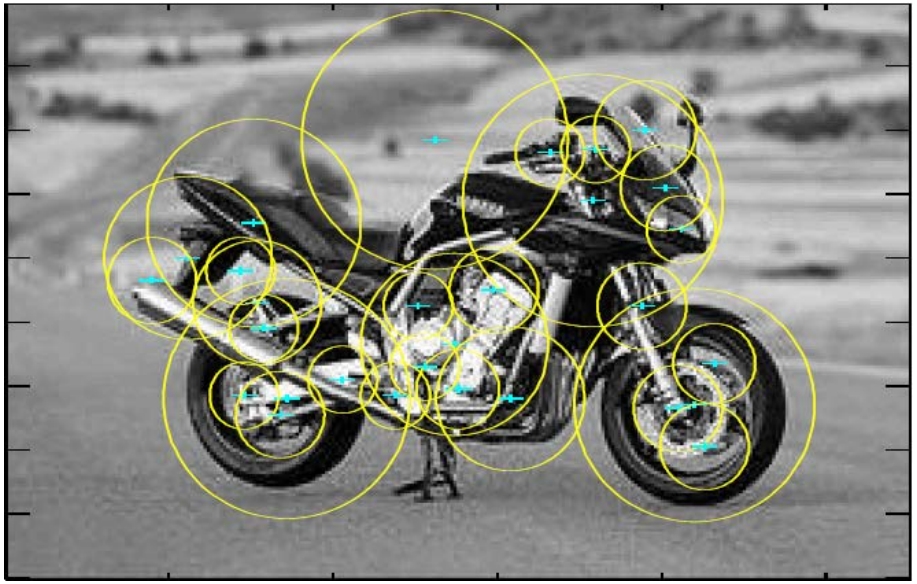


Joint Gaussian density

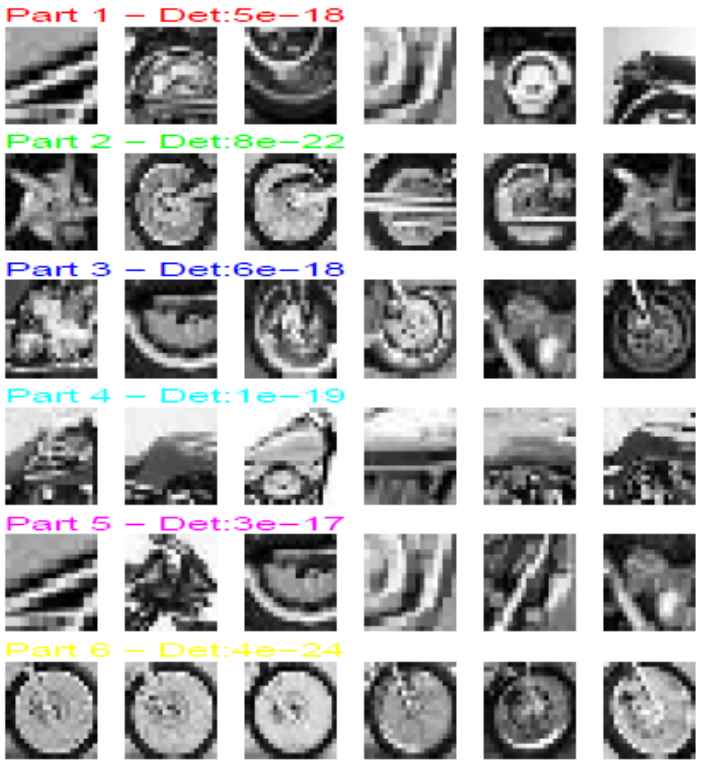


The representation and matching of pictorial structures, Fischler and Elschlager, 1973

Interest points used to find parts:



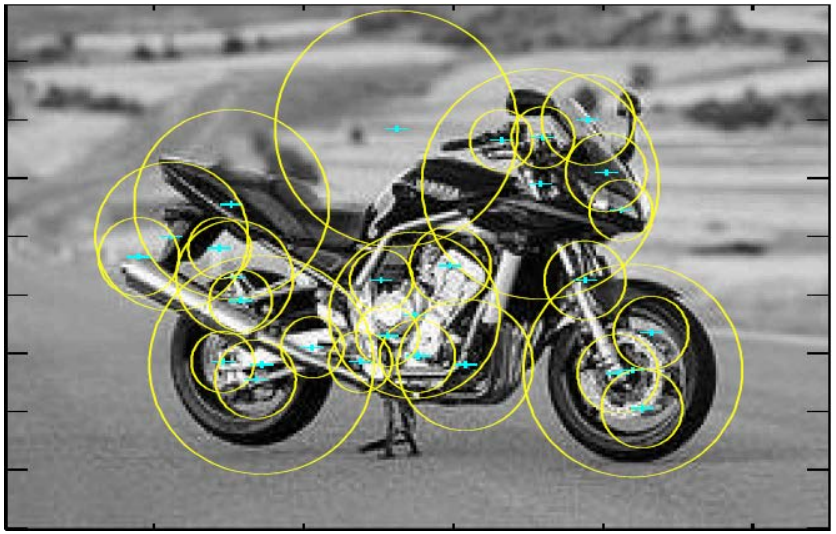
Smaller number of candidate parts allows for more complex spatial models.

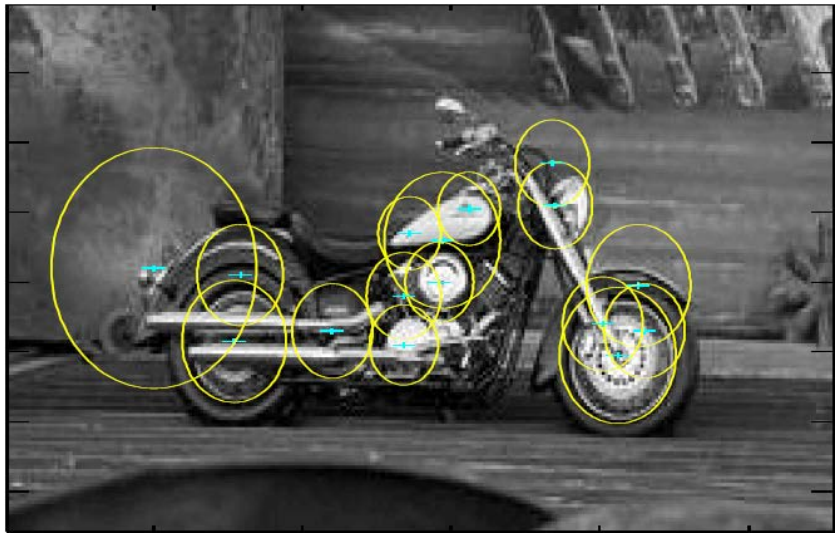


Why does it fail?

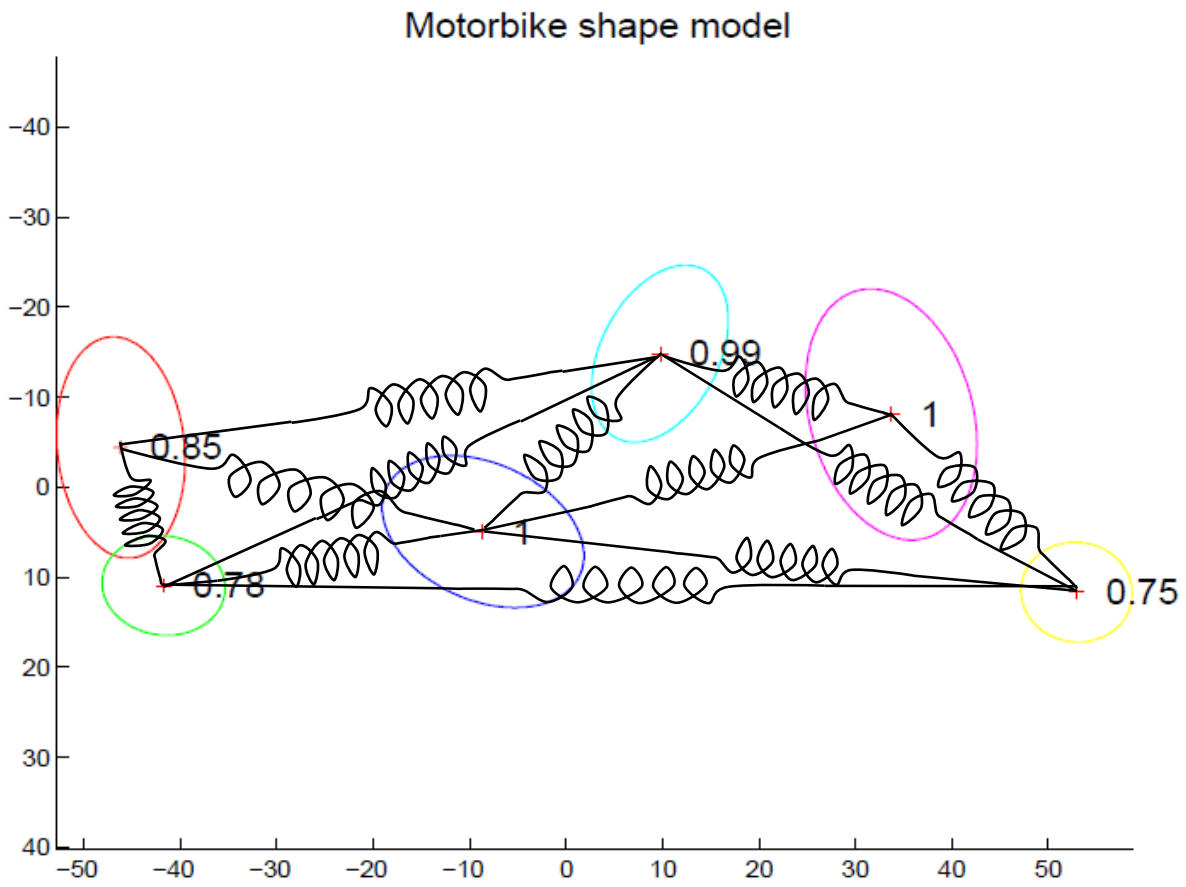
Why it fails…

- Interest points don’t work for category recognition





Too many springs…



Why it worked…







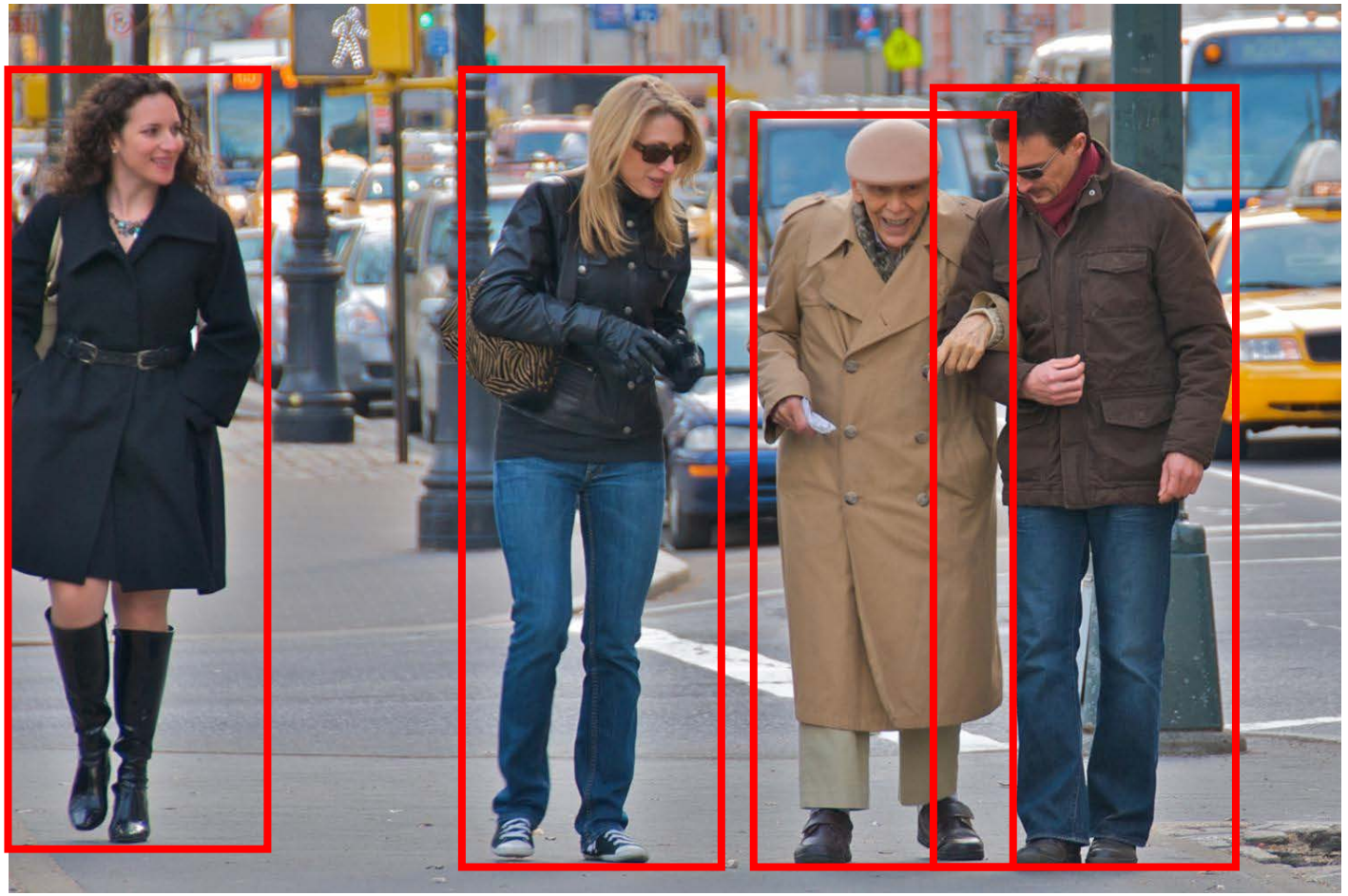









2005 HOG (histograms of oriented gradients)



Histograms of oriented gradients for human detection, Dalal and Triggs, CVPR 2005.

Pedestrians

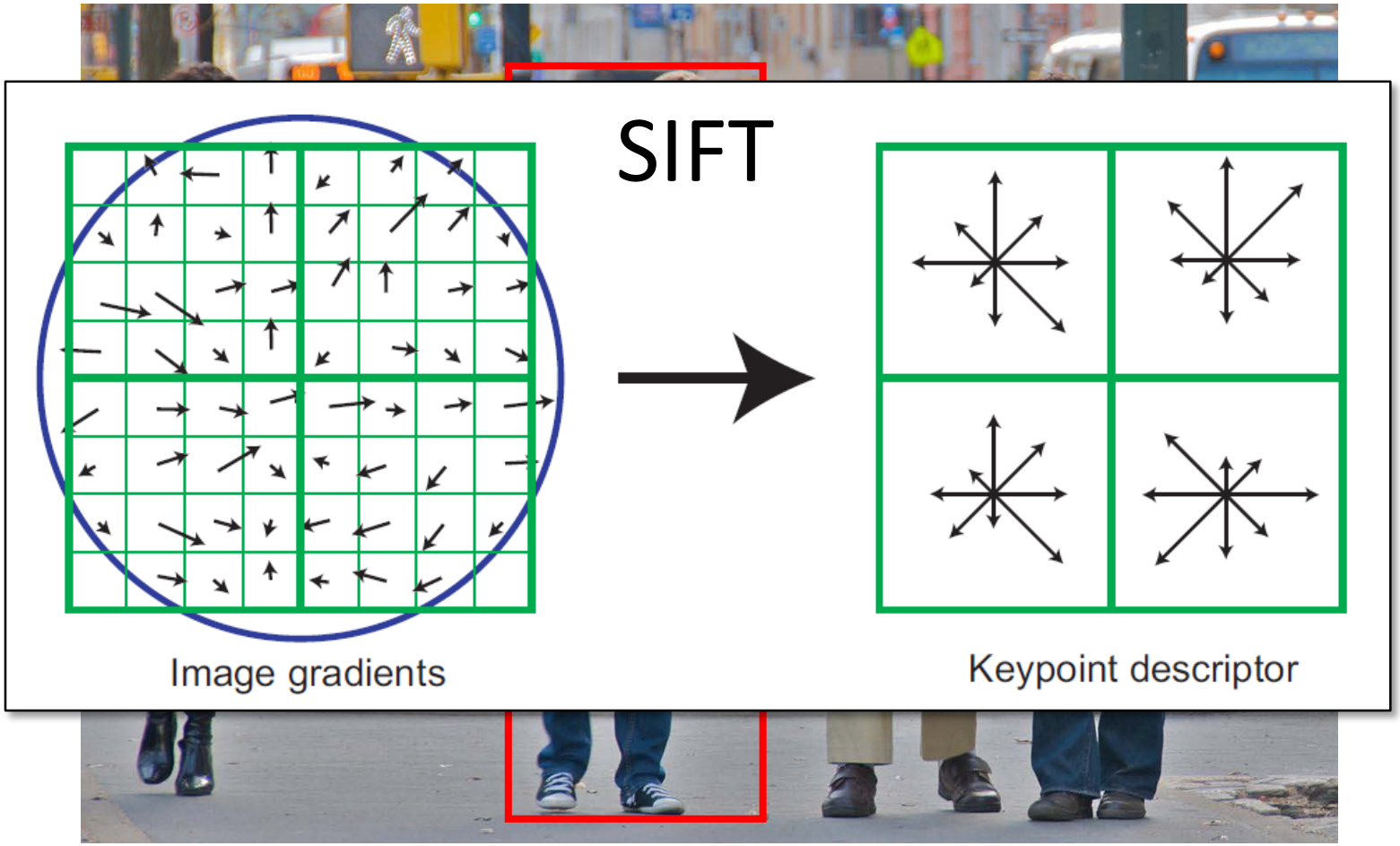
- Defined by their contours
- Cluttered backgrounds
- Significant variance in texture

Interest points won't work…

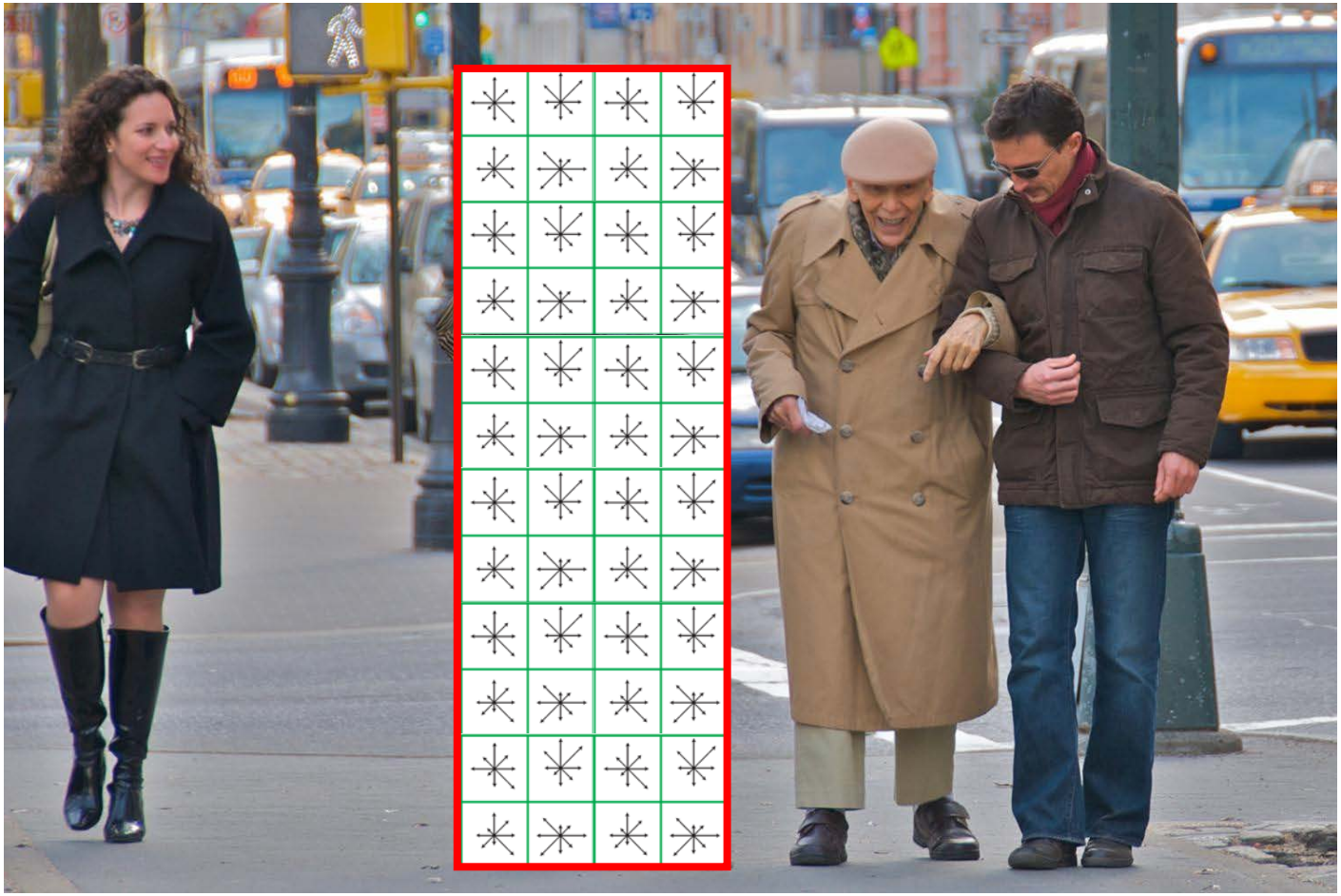
…back to sliding window.

2005 HOG (histograms of oriented gradients)

2005 HOG (histograms of oriented gradients)



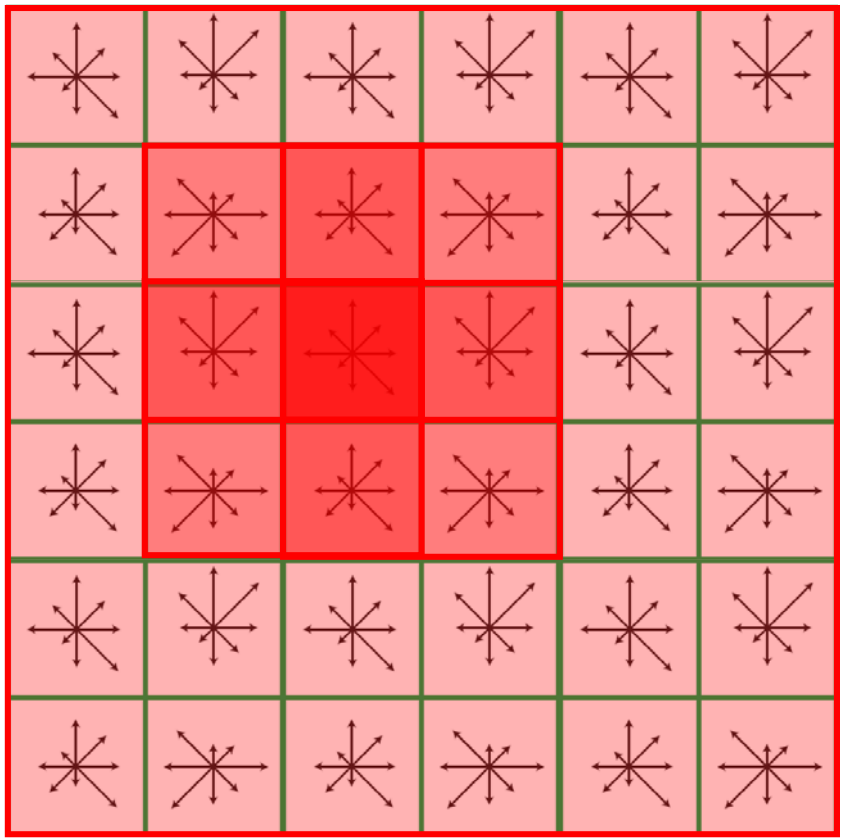
2005 HOG (histograms of oriented gradients)



Histograms of oriented gradients for human detection, Dalal and Triggs, *CVPR* 2005.

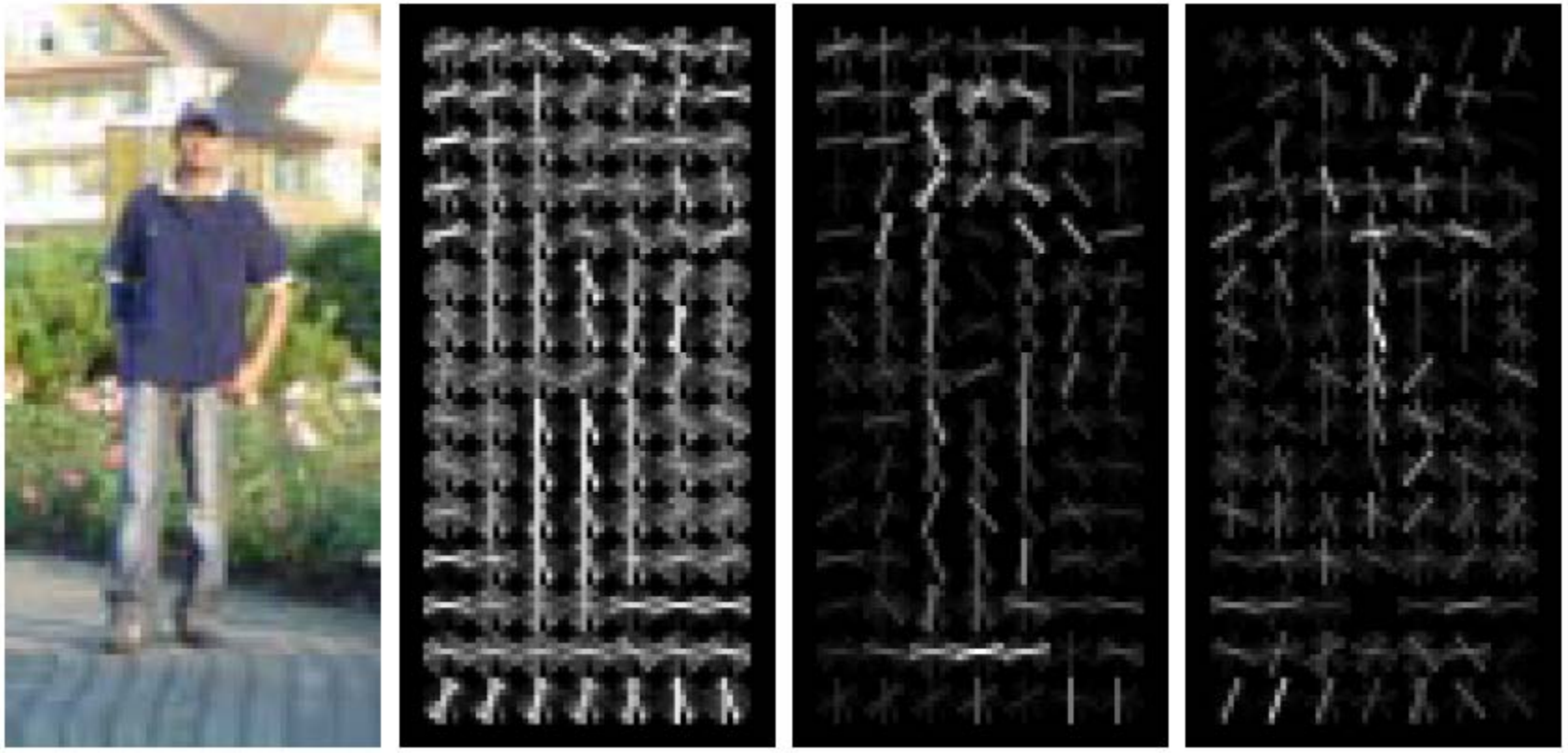
2005 HOG (histograms of oriented gradients)

Normalize locally not globally

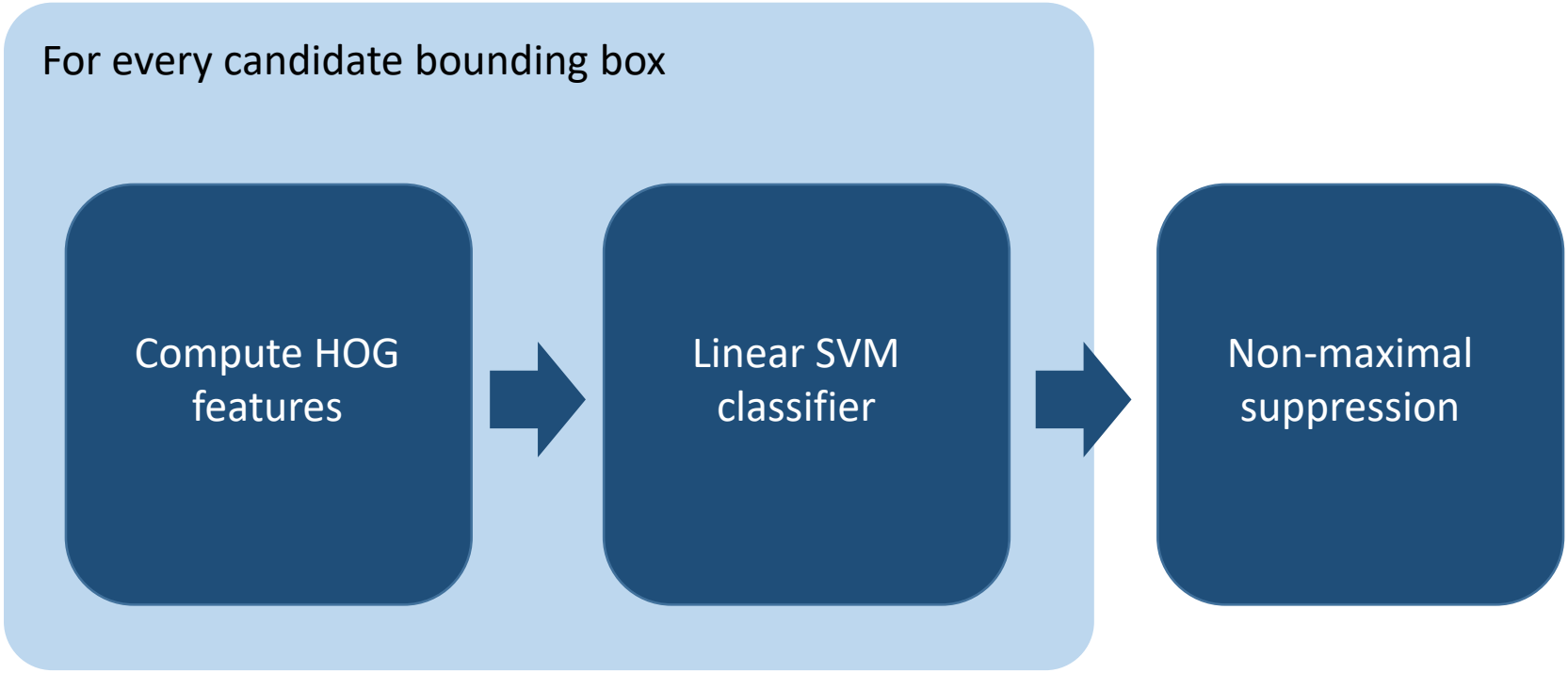


2005 HOG (histograms of oriented gradients)

Presence > Magnitude



2005 HOG (histograms of oriented gradients)



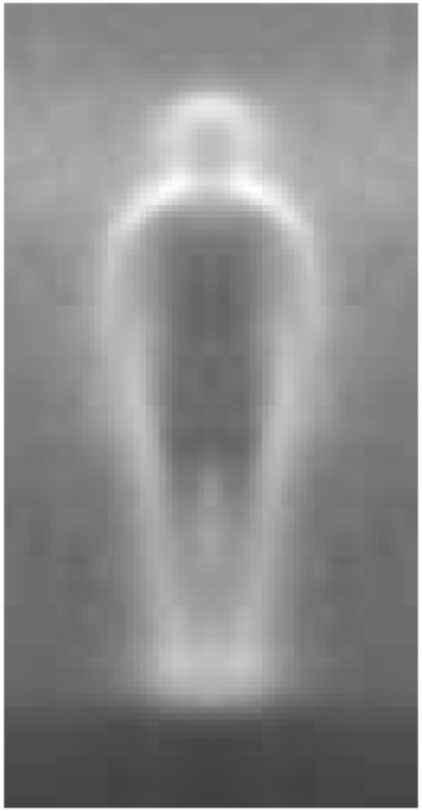
Why it worked

We can finally detect object boundaries in a reliable manner!

Hard negative mining

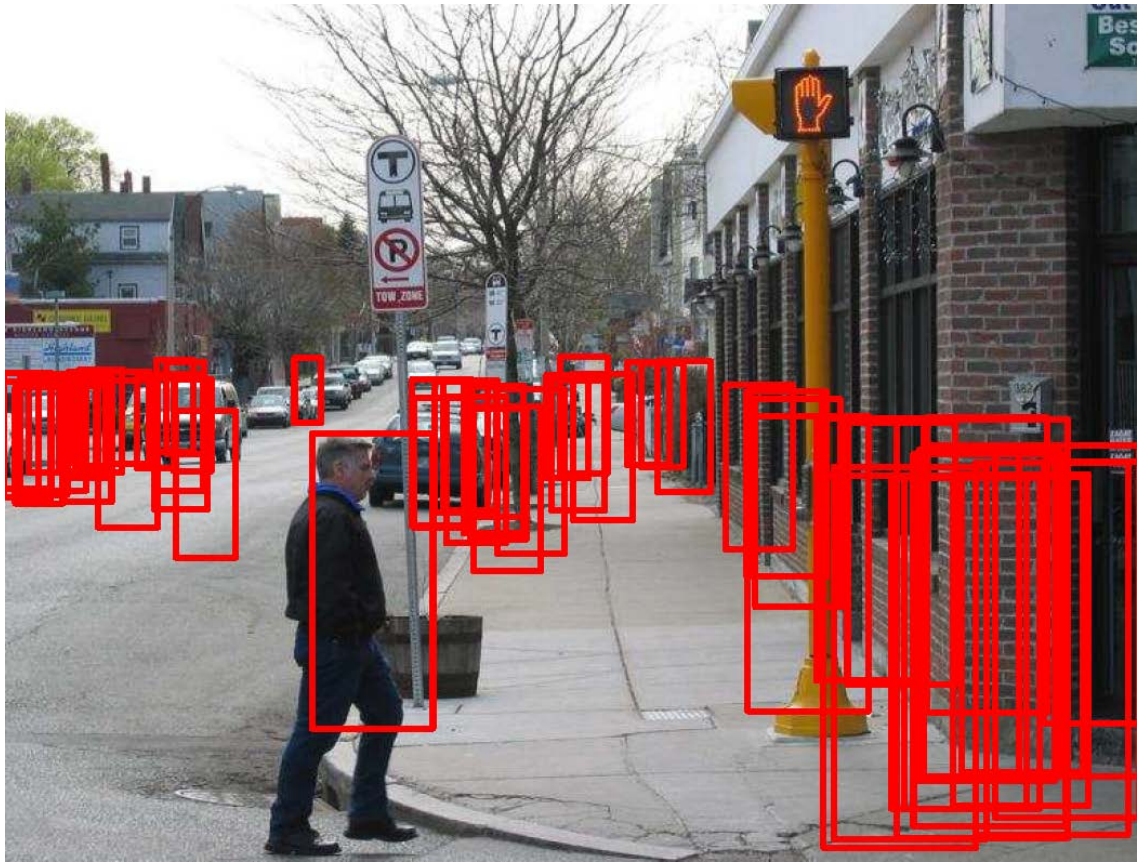
Computers are fast enough.

Why it failed



2006 Context

People don’t walk in the sky…



Putting Objects in Perspective, Hoiem, Efros, and Hebert, *CVPR* 2006.

Why it worked

Our world is not random

Some contextual cues may be reliably found

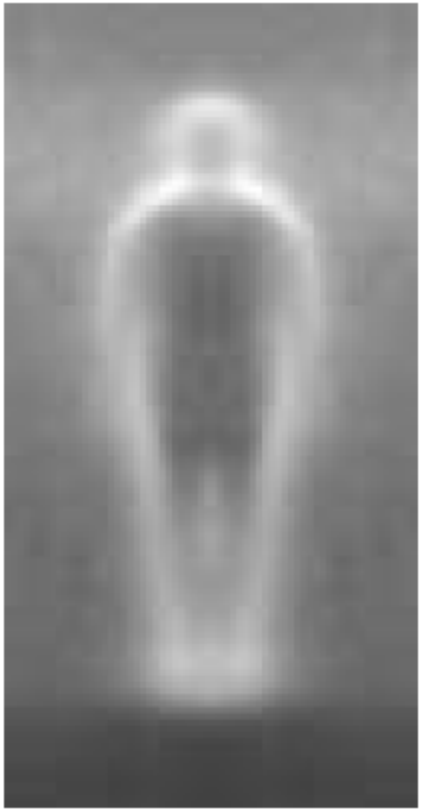
Why it failed

Many contextual cues are hard to detect

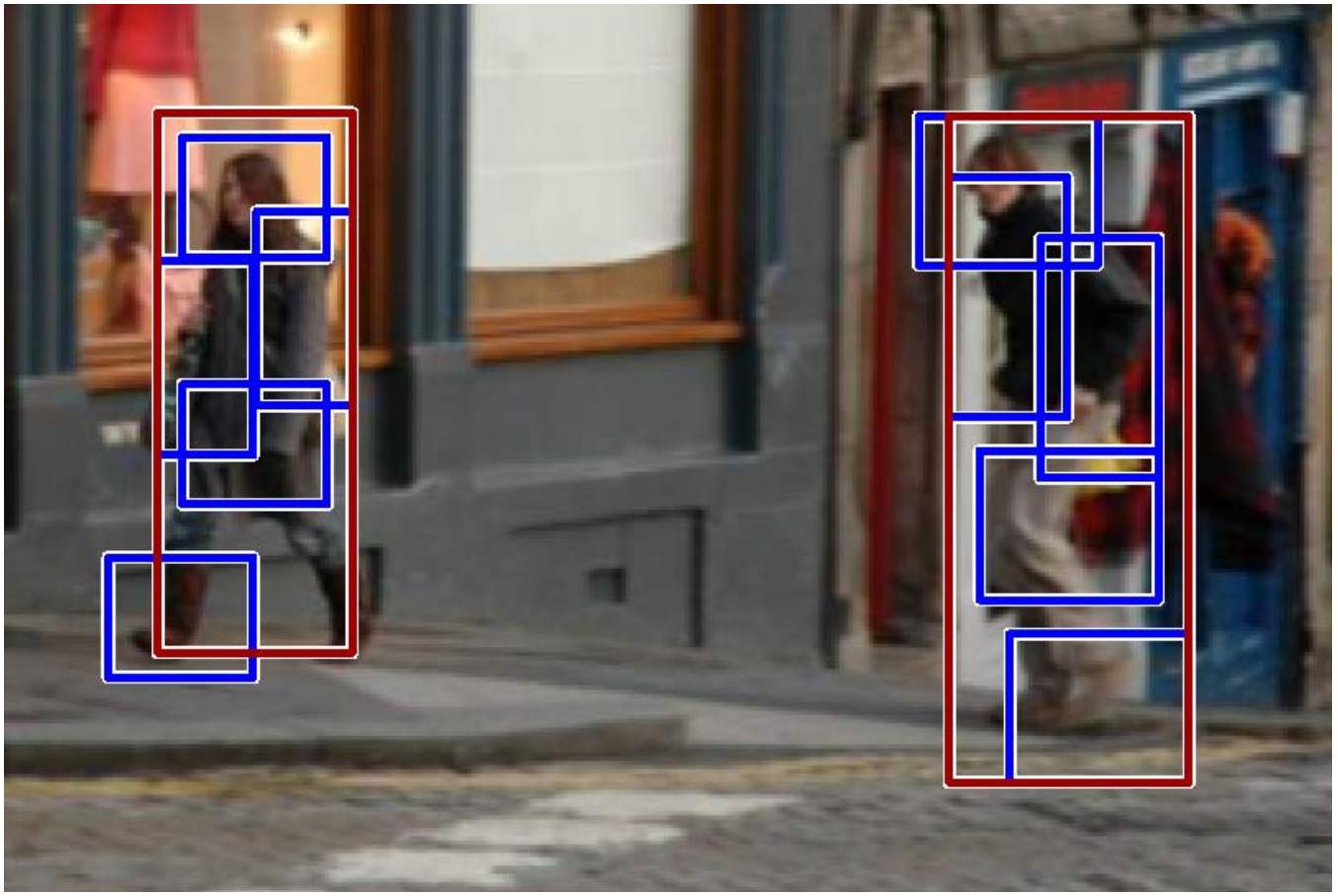
We need more data

Doesn’t help with false negatives

Why it failed

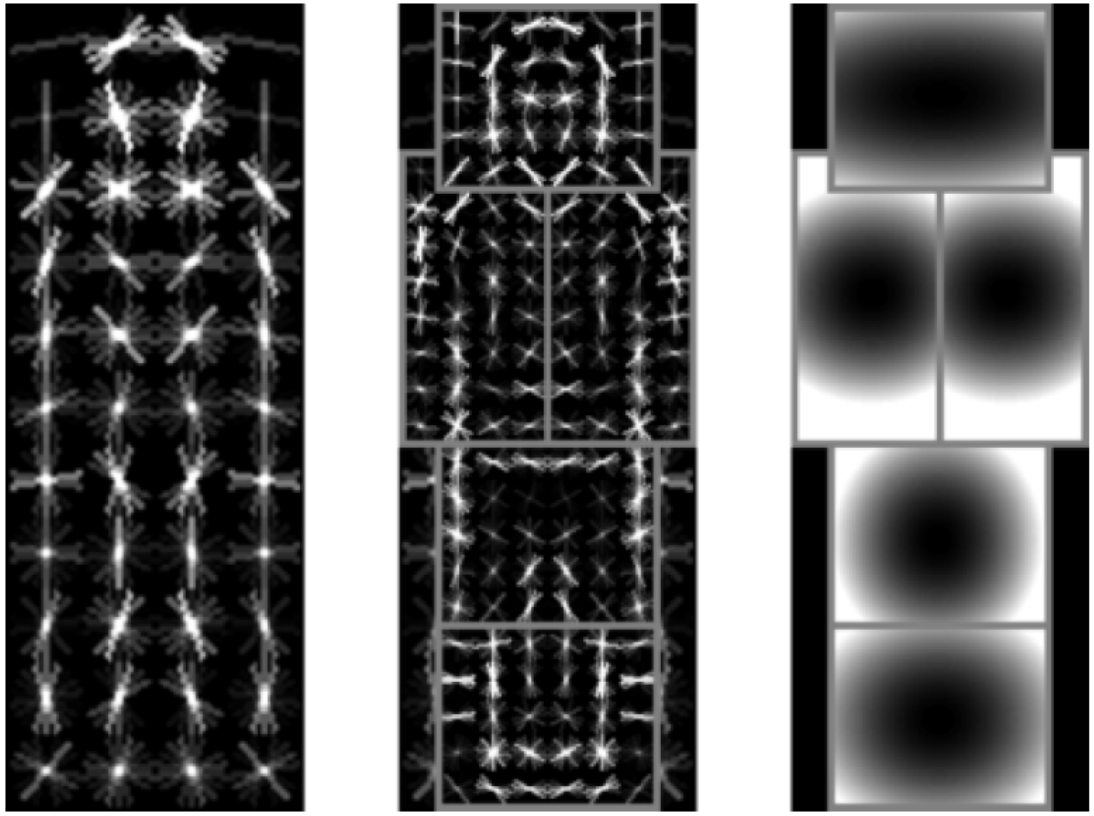


2008 DPM (Deformable parts model)



Object Detection with Discriminatively Trained Part Based Model,
Felzenszwalb, Girshick, McAllester and Ramanan, *PAMI*, 2010

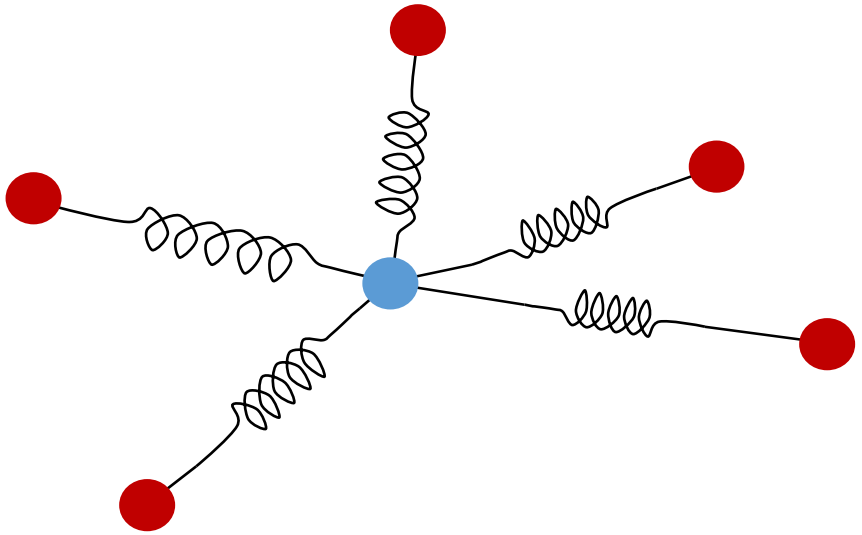
2008 DPM (Deformable parts model)



Object Detection with Discriminatively Trained Part Based Model,
Felzenszwalb, Girshick, McAllester and Ramanan, *PAMI*, 2010

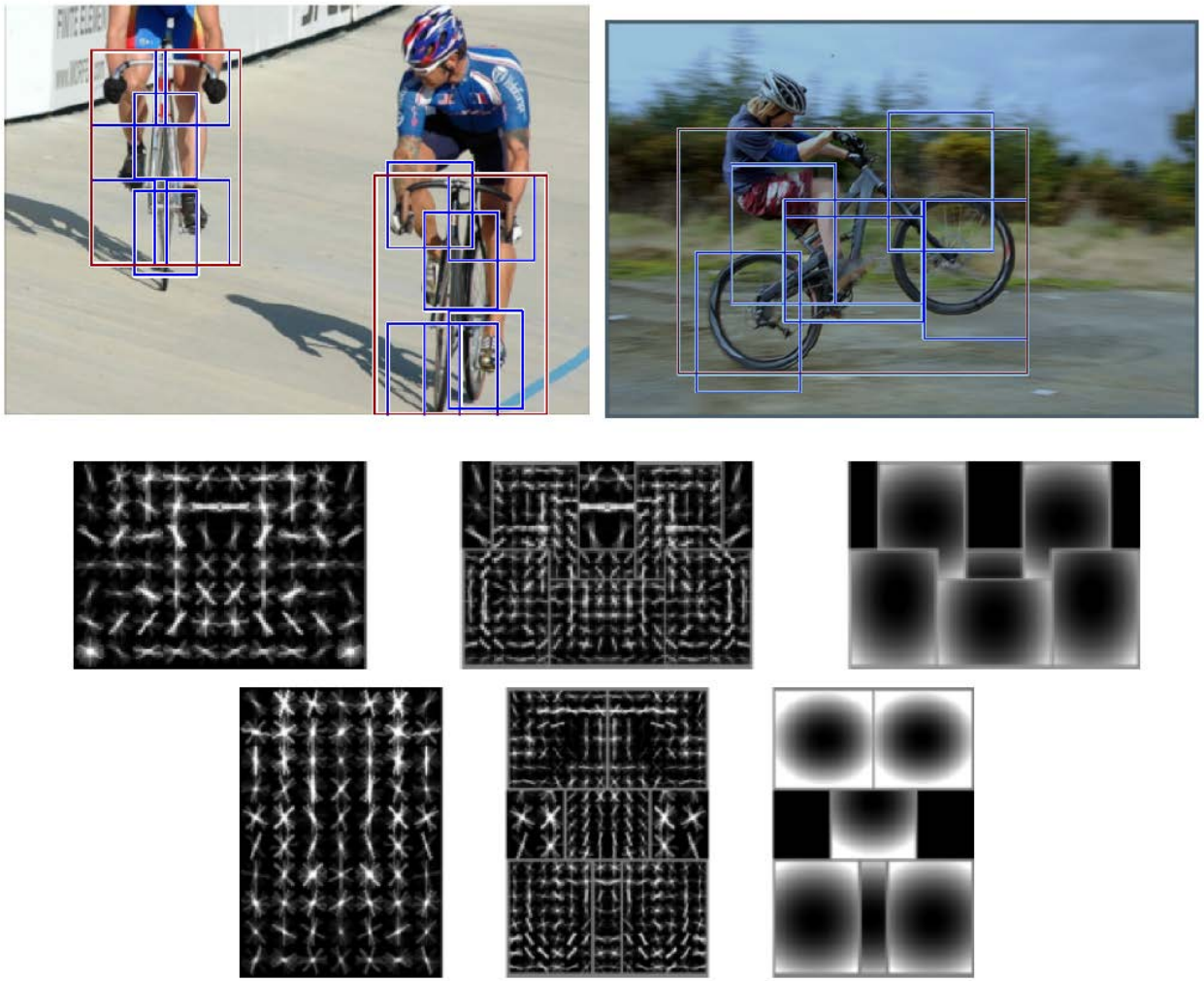
Star-structure

- Computationally efficient (distance transform)



Distance transforms of sampled functions, Felzenszwalb and Huttenlocher, Cornell University CIS, Tech. Rep. 2004.

Multiple components



Why it worked

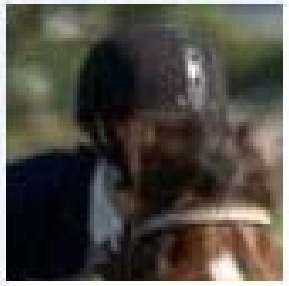
- Multiple components
- Deformable parts?
- Hard negative mining
- Good balance

"How important are 'Deformable Parts' in the Deformable Parts Model?“,
Divvala, Efros, and Hebert, *Parts and Attributes Workshop, ECCV,* 2012

Why it failed…

Human debugging

Is it a head, torso, arm, leg, foot, hand, or nothing?



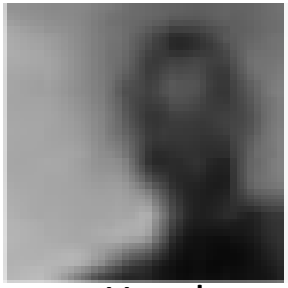
Head

Leg

Nothing

Low resolution 20x20 pixels

Feet

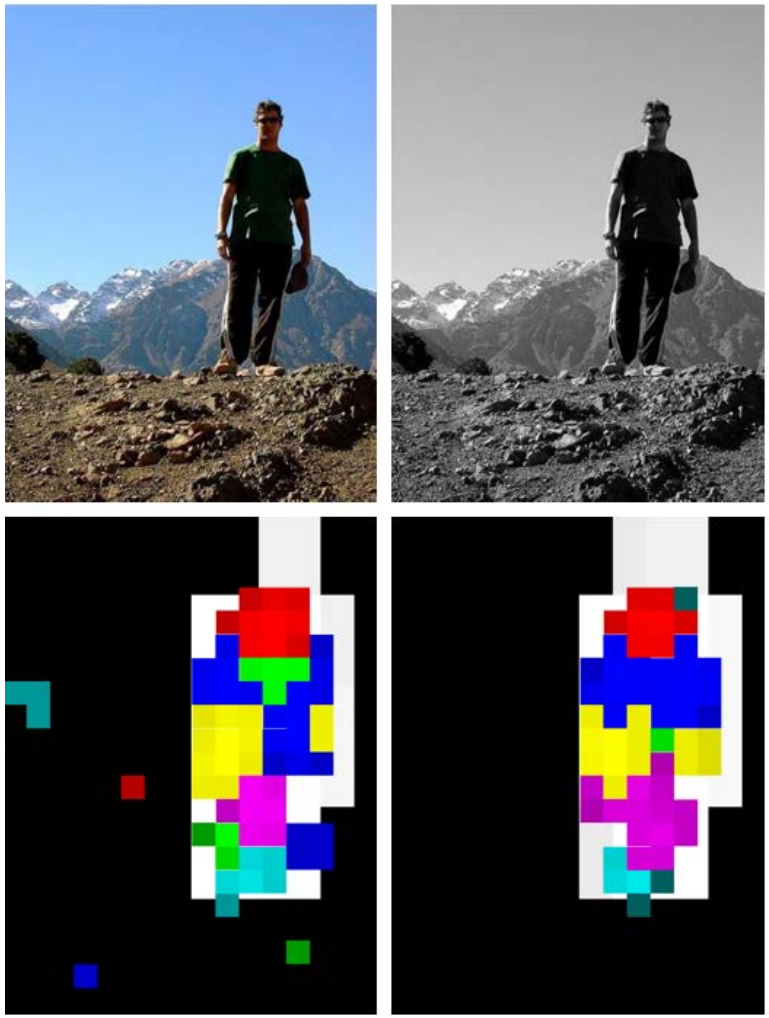


Head

?

Finding the Weakest Link in Person Detectors, Parikh and Zitnick, *CVPR* 2011.

Part detections

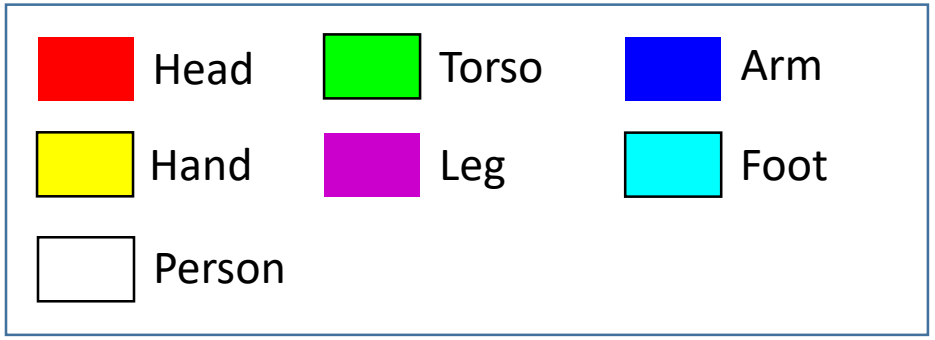


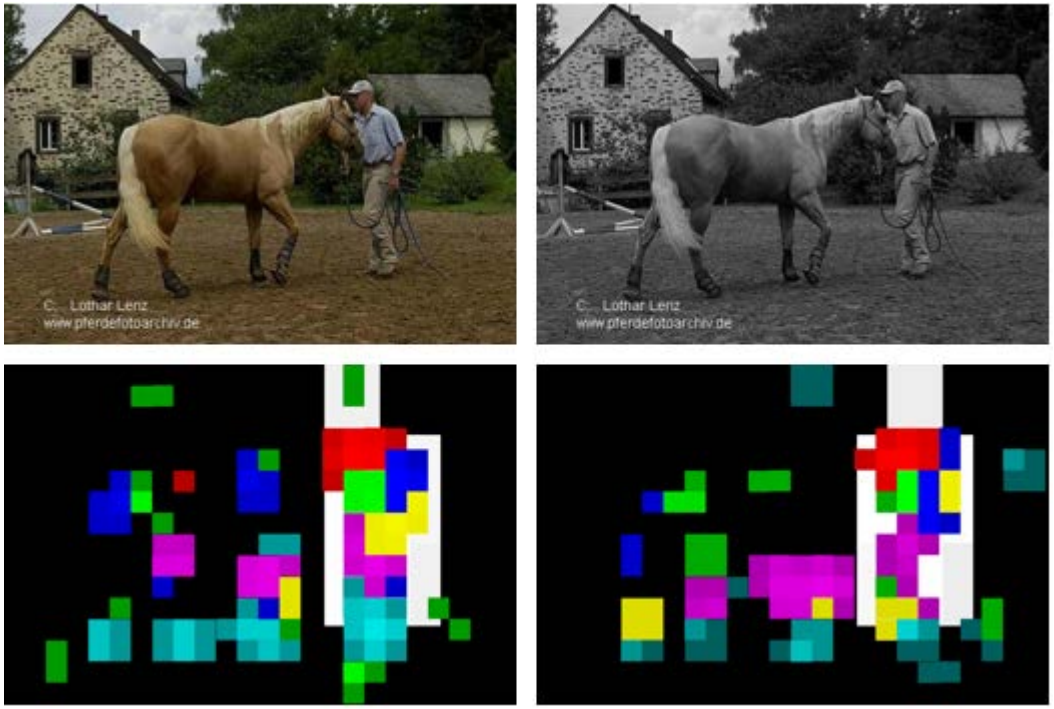
Humans

Head
Torso
Arm
Hand
Leg
Foot
Person

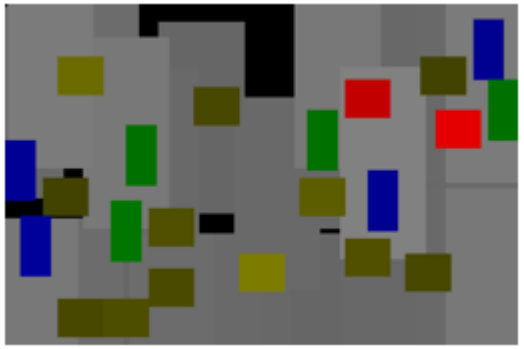
Machine

Part detections



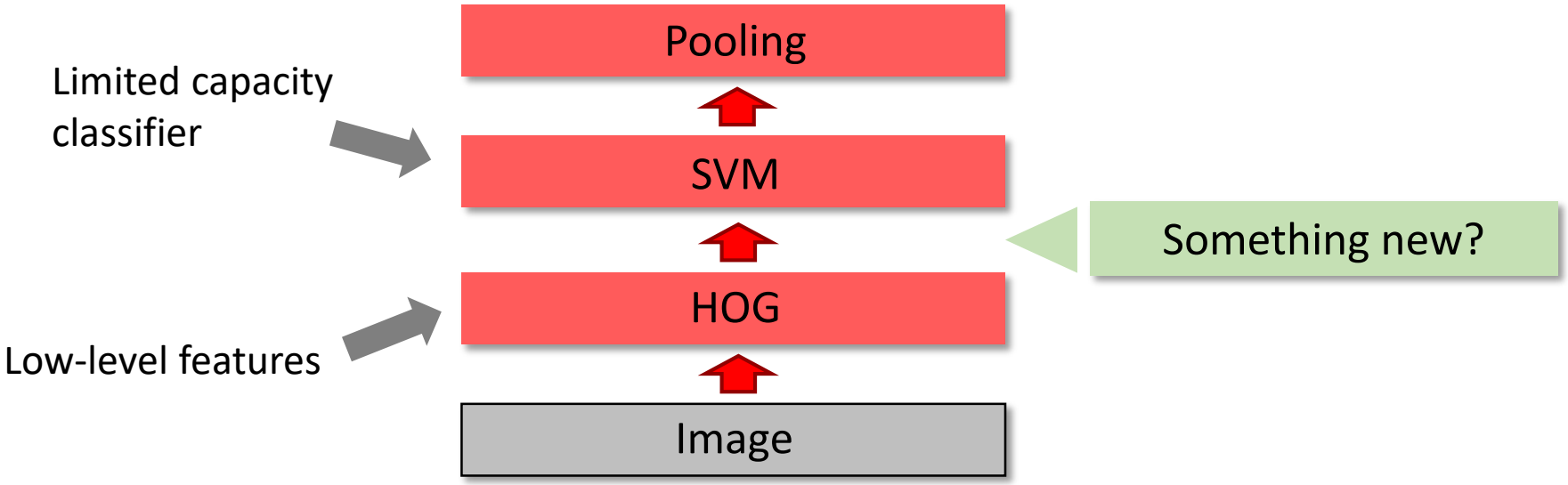


Humans



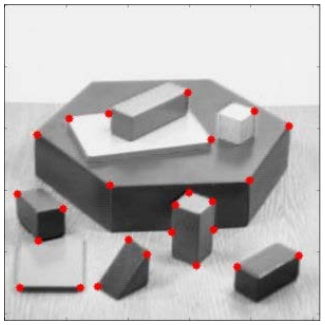
Machine

DPM


Pooling
Limited capacity classifier
SVM
Something new?
HOG
Low-level features
Image

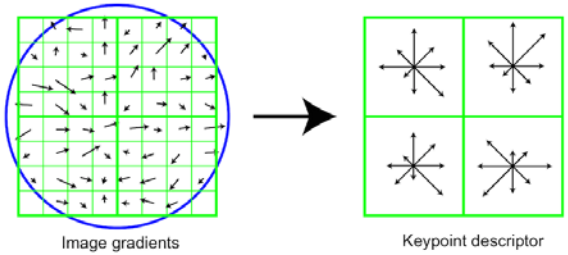
25 years of feature designing…

Canny Edge Detection
Canny, 1986

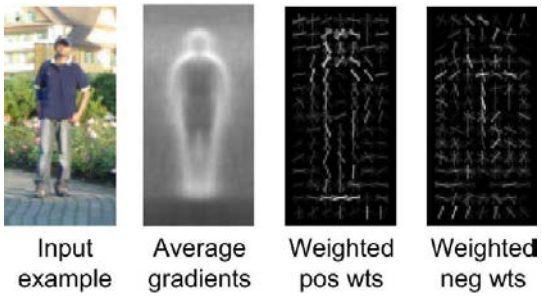


Harris Corner Detection
Harris and Stephens, 1988

Harr Wavelets,
Viola and Jones, CVPR 2001

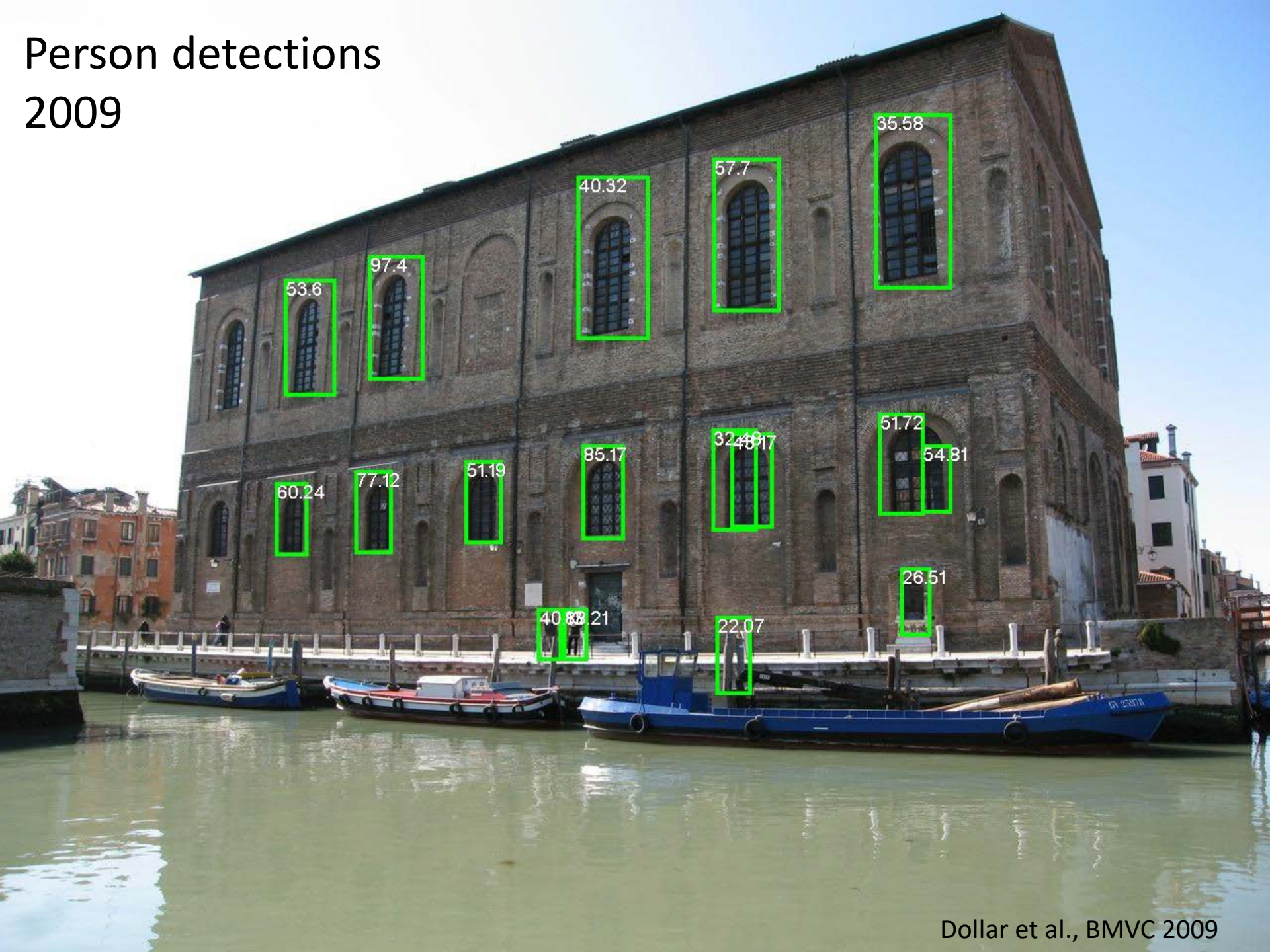


SIFT,
Lowe, 2004



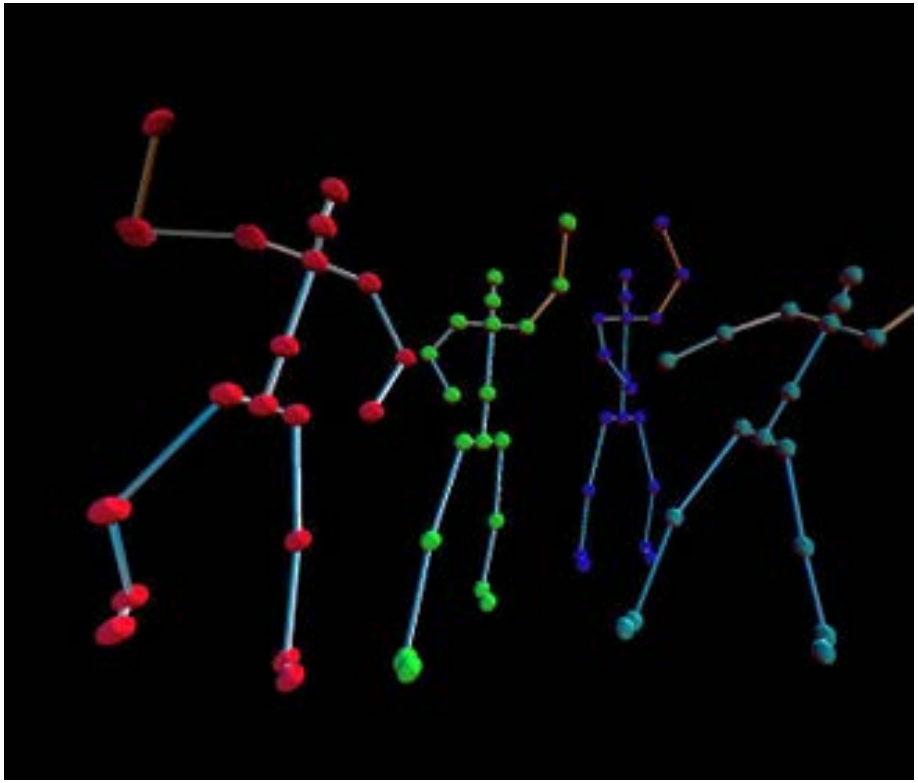
HOG,
Dalal and Triggs, 2005

Person detections 2009



Dollar et al., BMVC 2009


XBOX 360
KINECT



Real-time human pose recognition in parts from single depth images, Shotton et al., CVPR 2011

2011 What works?

Tomorrow