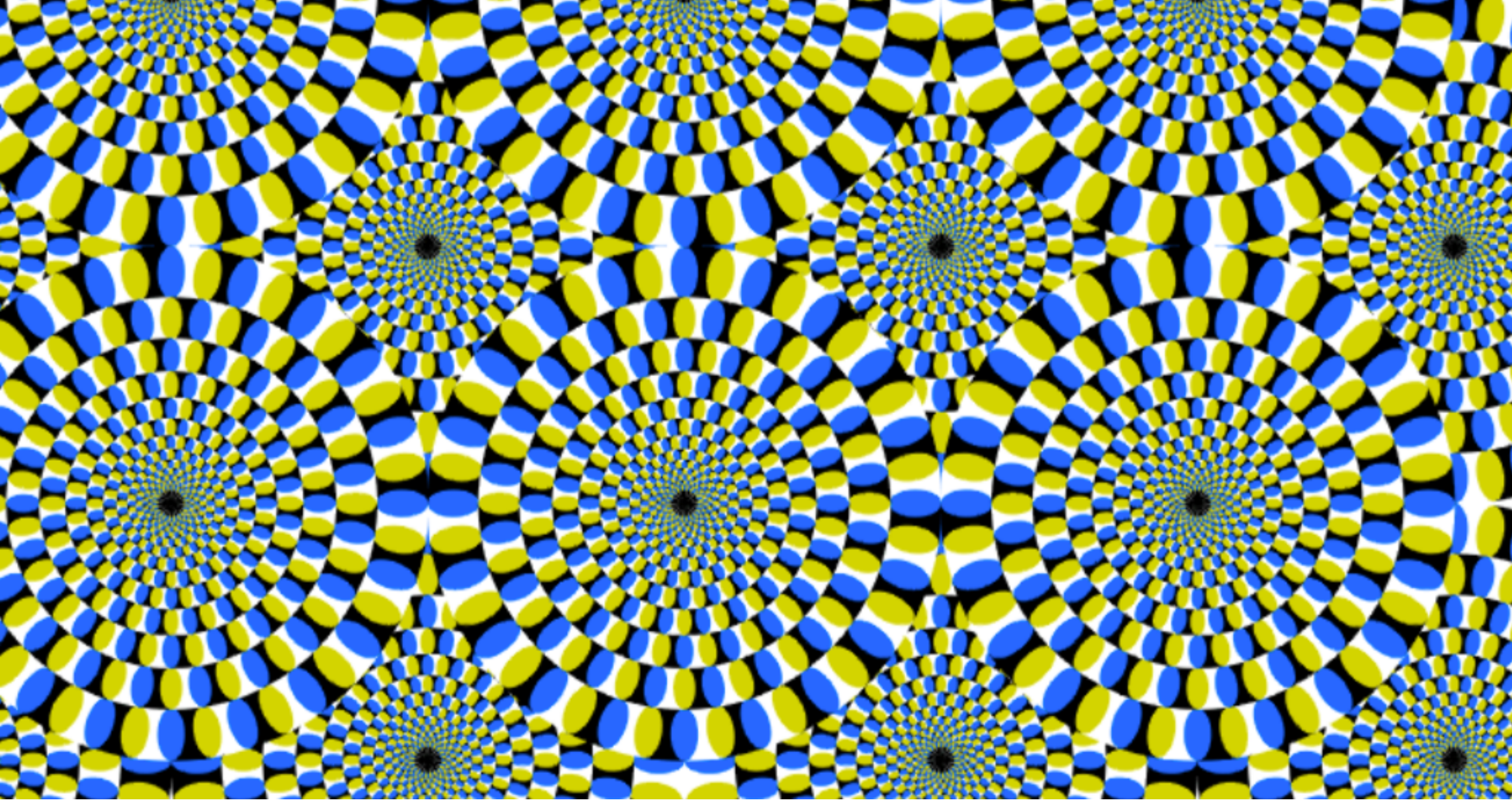


Computer Vision

EE382V Activity Sensing and Recognition

UT Austin • Dept. Electrical and Computer Engineering • Fall 2016

Today: Computer Vision and Recognition

Definition, Applications and Challenges

Images, Filters and Edges

Recognition and Bag-of-Features

Paper

Resources for learning more

What is Computer Vision?

Computer Vision

Make computers understand images and video.

What kind of scene?

Where are the cars?

How far is the building?

…

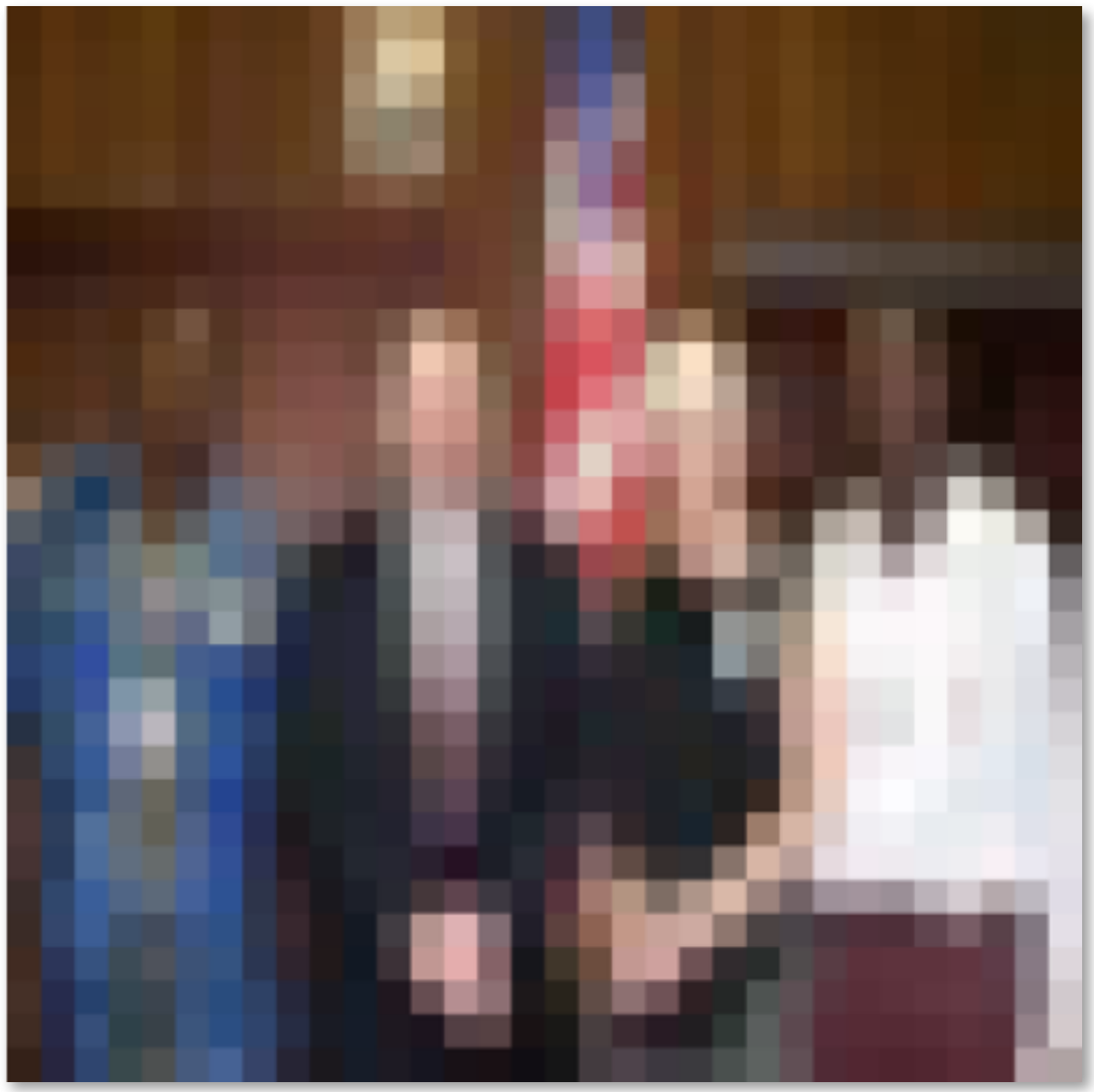
Vision is really hard

- Vision is an amazing feat of natural intelligence
  - Visual cortex occupies about 50% of Macaque brain
  - More human brain devoted to vision than anything else

Human perception has its shortcomings

Sinha and Poggio, *Nature*, 1996

But humans can tell a lot about a scene from a little information…



Source: “80 million tiny images” by Torralba, et al.

Observation

- We can recognize familiar faces even in low-resolution images

Observation

Jim Carrey

Kevin Costner

- High frequency information is not enough

Observation

- Image Warping is OK

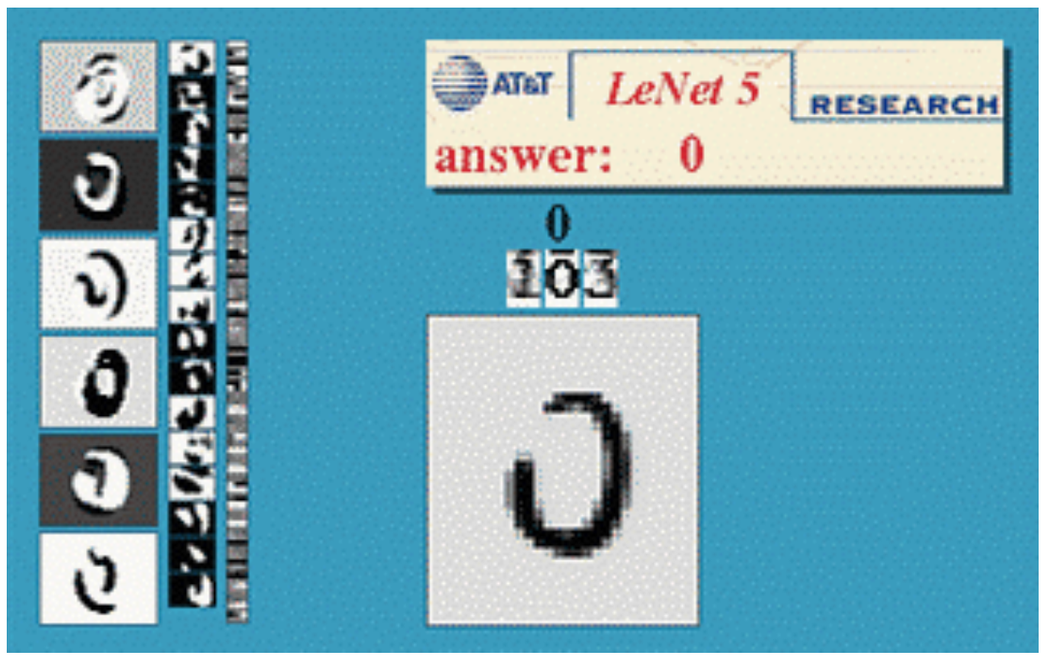
Can the computer match human perception?

- Yes and no (mainly no)
  - computers can be better at “easy” things
  - humans are much better at “hard” things
- But huge progress has been made
  - Especially in the last 10 years
  - What is considered “hard” keeps changing

Optical character recognition (OCR)

Technology to convert scanned docs to text

- If you have a scanner, it probably came with OCR software



Digit recognition, AT&T labs
http://www.research.att.com/~yann/

License plate readers
http://en.wikipedia.org/wiki/Automatic_number_plate_recognition

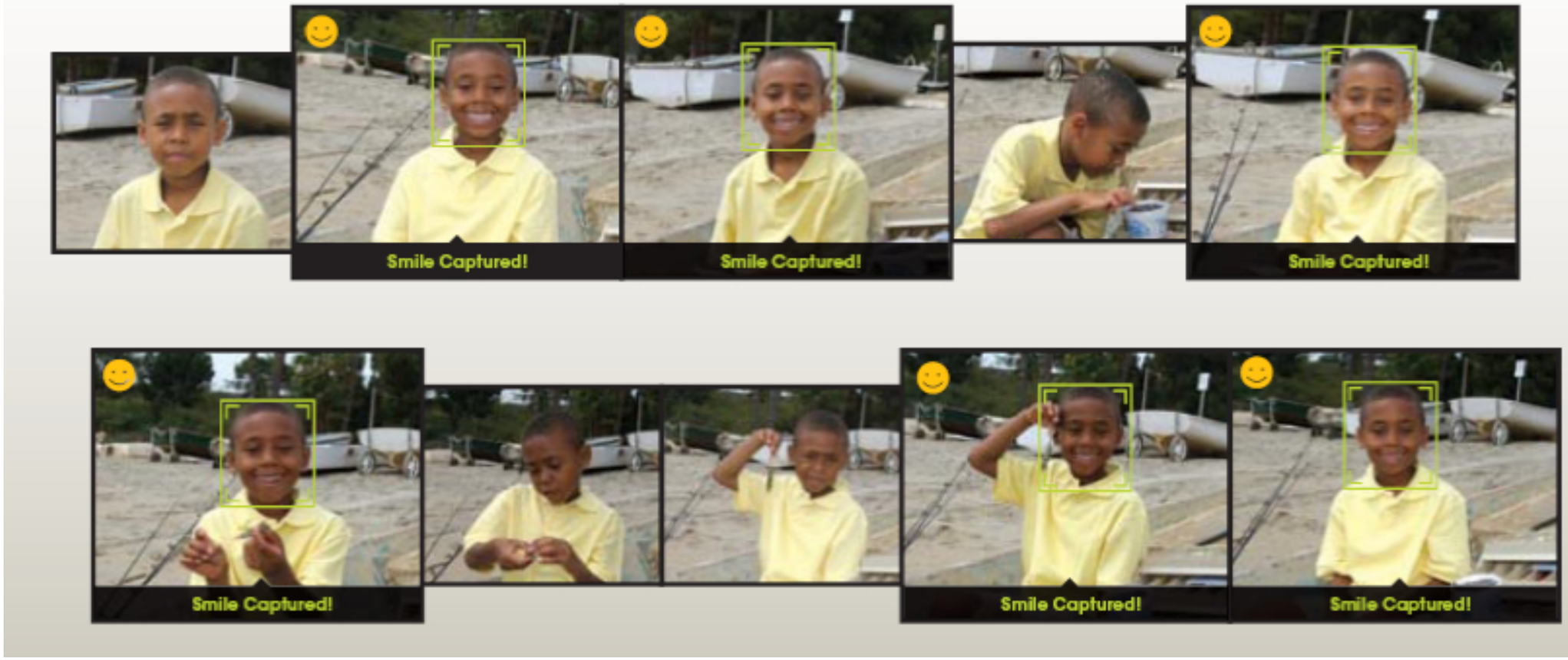
Face detection

- Many new digital cameras now detect faces
  - Canon, Sony, Fuji, …

Smile detection

The Smile Shutter flow

Imagine a camera smart enough to catch every smile! In Smile Shutter Mode, your Cyber-shot® camera can automatically trip the shutter at just the right instant to catch the perfect expression.



[Sony Cyber-shot® T70 Digital Still Camera]

3D from thousands of images

Building Rome in a Day: Agarwal et al. 2009

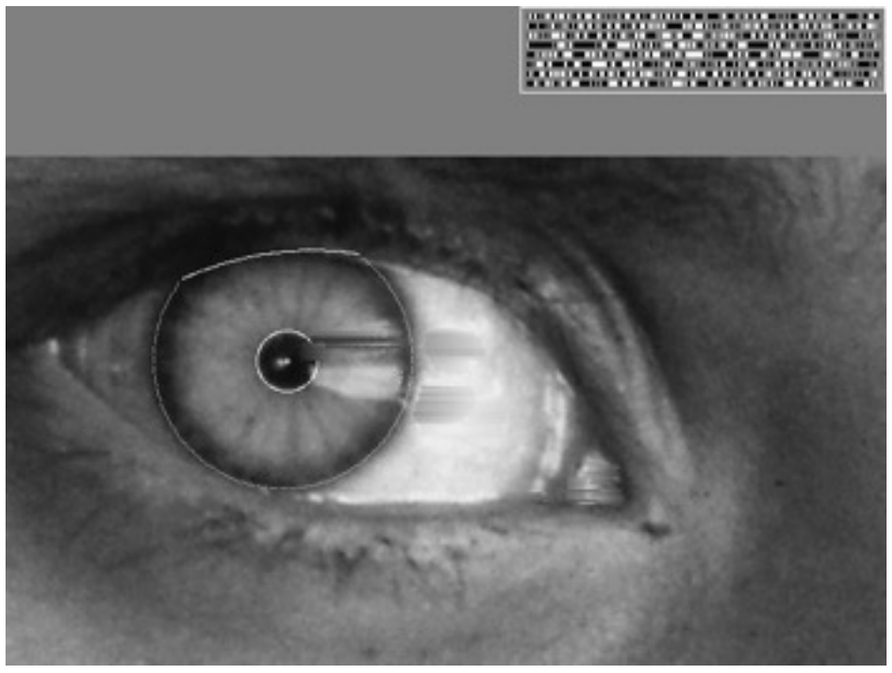
Object recognition (in supermarkets)

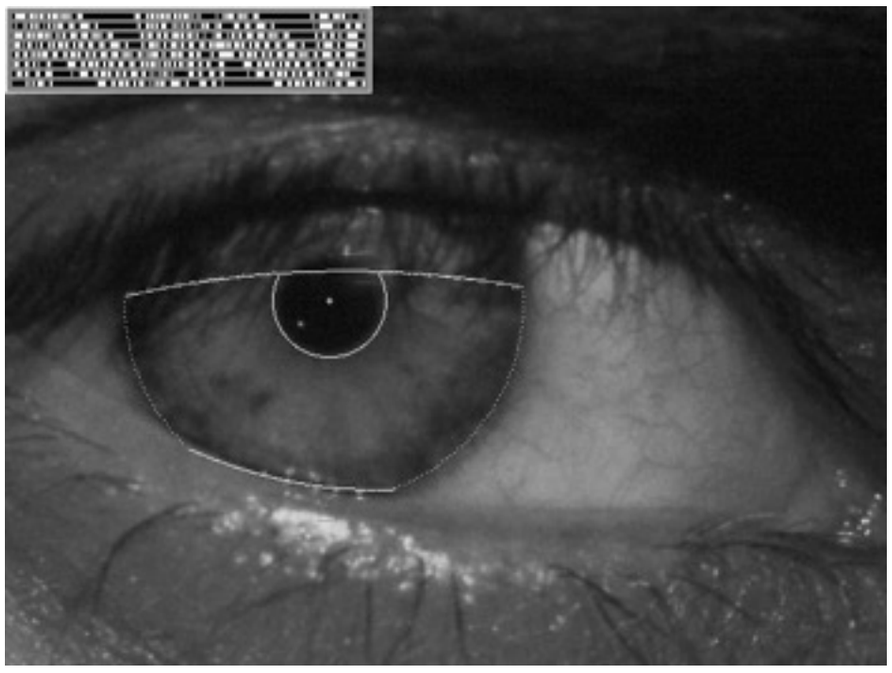
LaneHawk by EvolutionRobotics

“A smart camera is flush-mounted in the checkout lane, continuously watching for items. When an item is detected and recognized, the cashier verifies the quantity of items that were found under the basket, and continues to close the transaction. The item can remain under the basket, and with LaneHawk,you are assured to get paid for it… “

Vision-based biometrics

"*How the Afghan Girl was Identified by Her Iris Patterns*" Read the story wikipedia





Forensics

Source: Nayar and Nishino, “Eyes for R

Login without a password…

Fingerprint scanners on many new laptops, other devices

Face recognition systems now beginning to appear more widely
http://www.sensiblevision.com/

Object recognition (in mobile phones)

[Point & Find](), [Nokia]()

[Google Goggles]()

Special effects: shape capture

The Matrix movies, ESC Entertainment, XYZRGB, NRC

Special effects: motion capture

Pirates of the Carribean, Industrial Light and Magic

Sports

Sportvision first down line
Nice explanation on www.howstuffworks.com

http://www.sportvision.com/video.html

Smart cars

Slide content courtesy of Amnon Shashua

- Mobileye
  - Vision systems currently in high-end BMW, GM, Volvo models
  - By 2010: 70% of car manufacturers.

Google cars

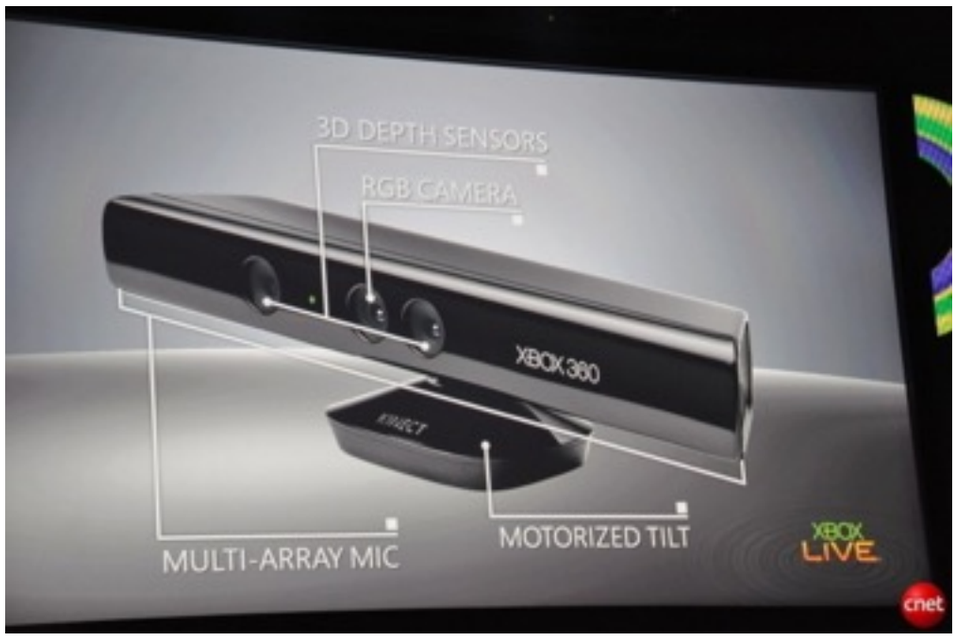
Oct 9, 2010. "Google Cars Drive Themselves, in Traffic". *The New York Times*. John Markoff

June 24, 2011. "Nevada state law paves the way for driverless cars". *Financial Post*. Christine Dobby

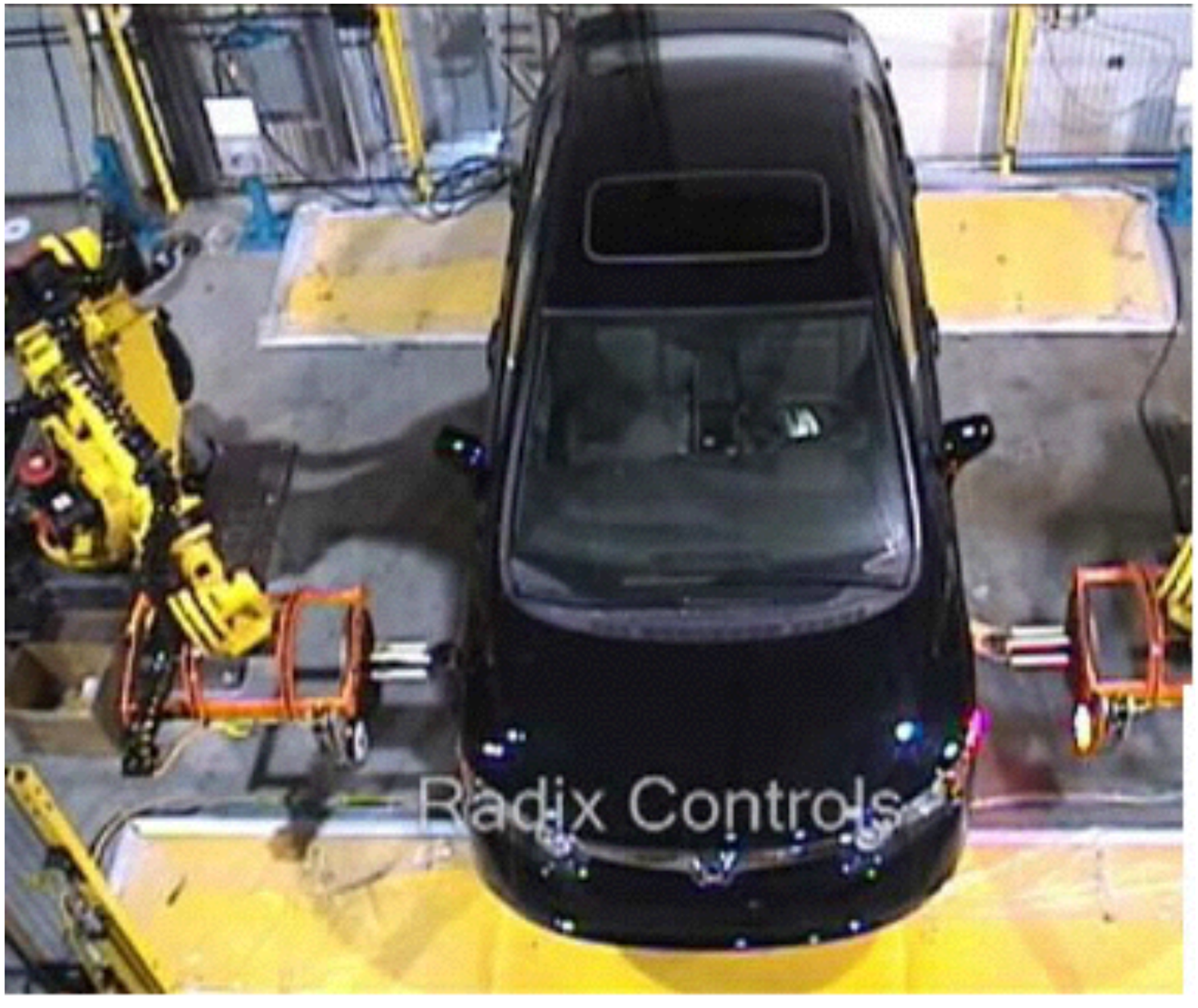
Aug 9, 2011, "Human error blamed after Google's driverless car sparks five-vehicle crash". *The Star* (Toronto)

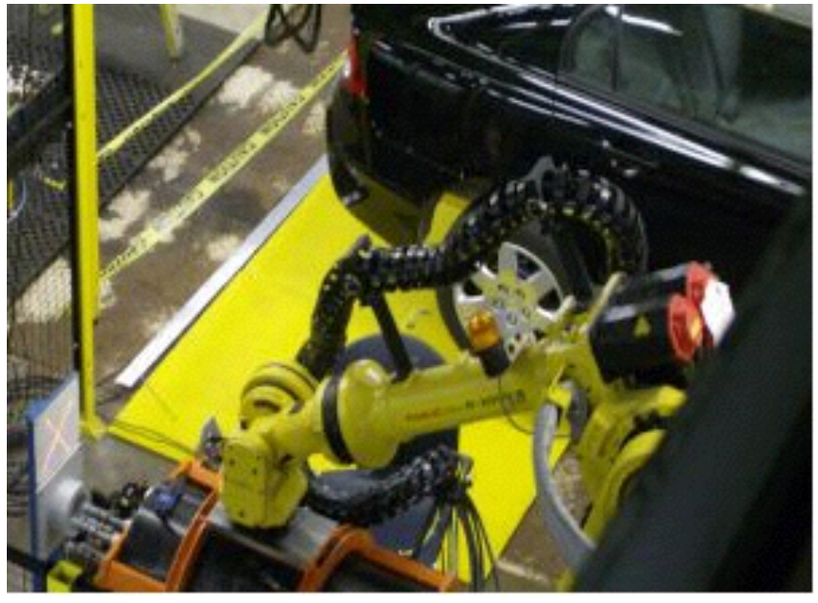
Interactive Games: Kinect

- Object Recognition: http://www.youtube.com/watch?feature=iv&v=fQ59dXOo63o
- Mario: http://www.youtube.com/watch?v=8CTJL5lUjHg
- 3D: http://www.youtube.com/watch?v=7QrnwoO1-8A
- Robot: http://www.youtube.com/watch?v=w8BmgtMKFbY



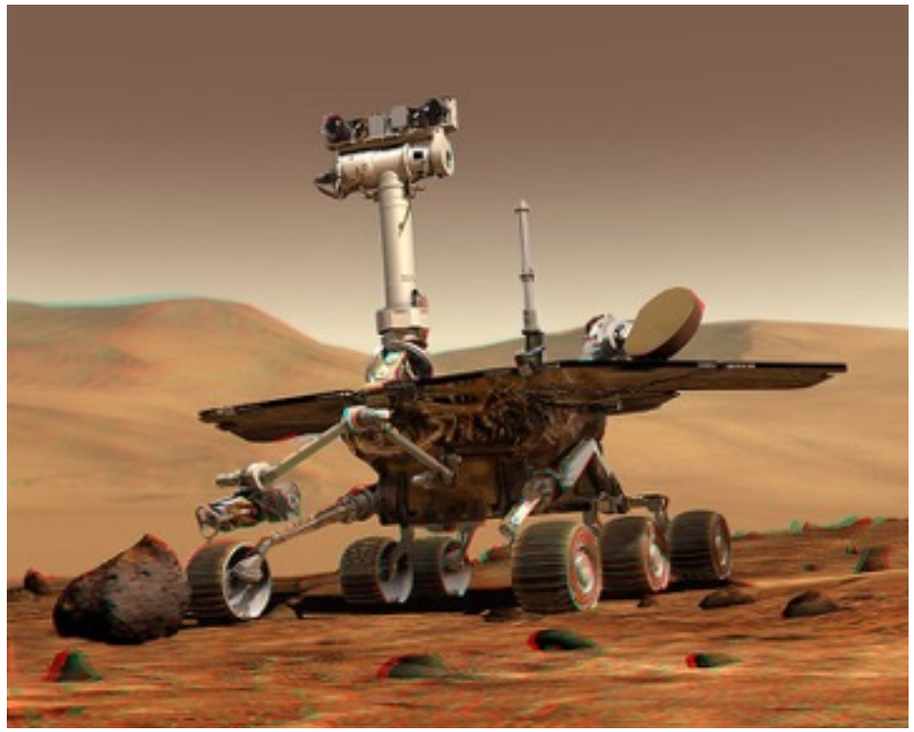
Industrial robots



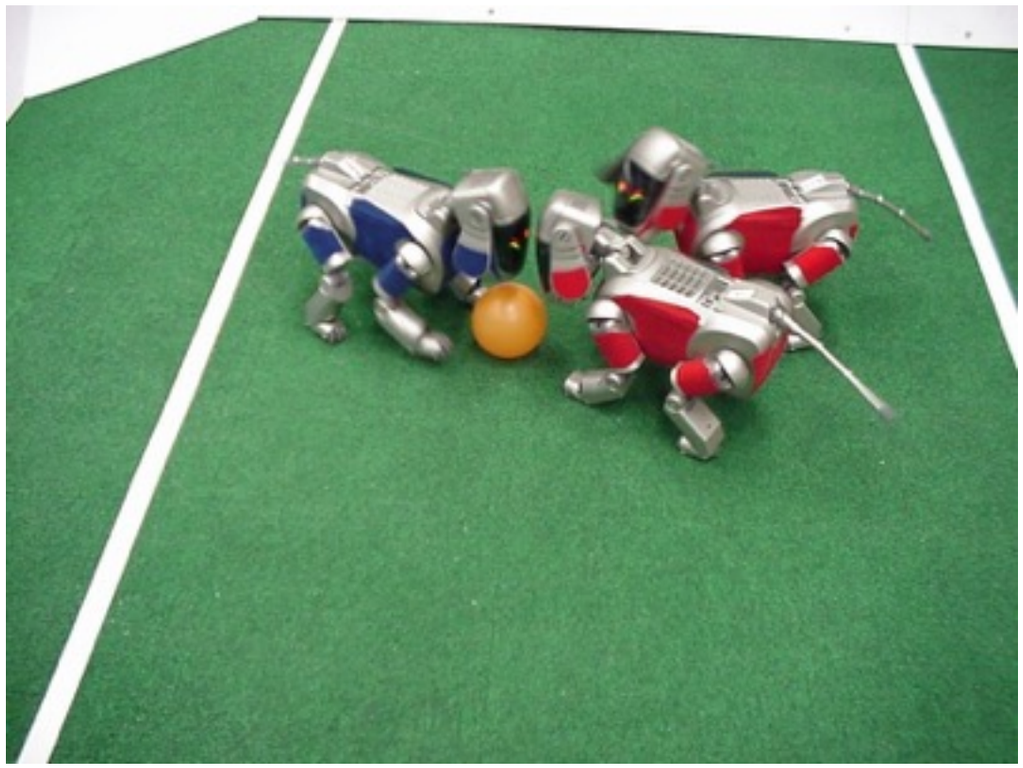


Vision-guided robots position nut runners on wheels

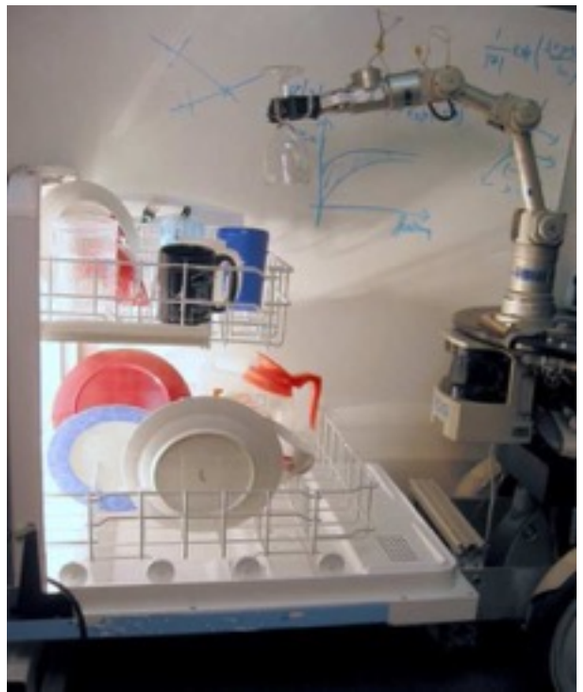
Mobile robots



NASA's Mars Spirit Rover
http://en.wikipedia.org/wiki/Spirit_rover

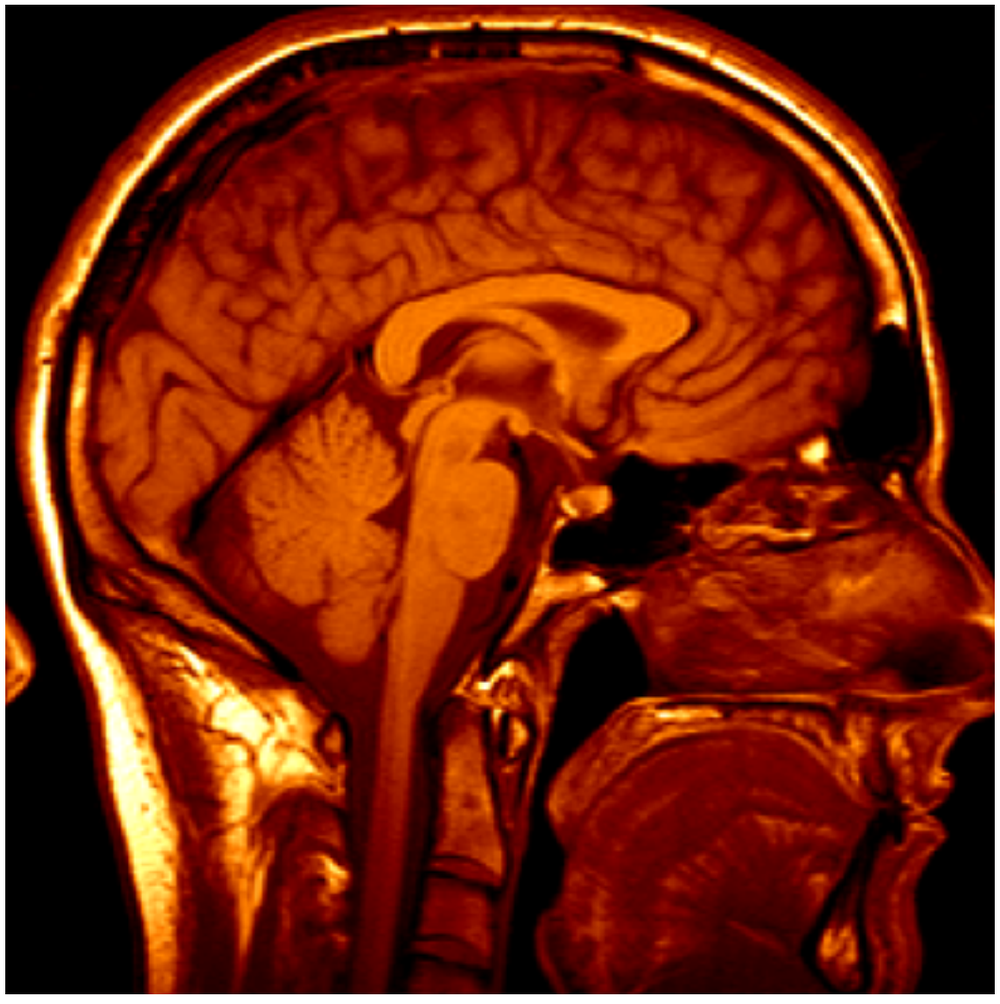


http://www.robocup.org/



Saxena et al. 2008
STAIR at Stanford

Medical imaging



3D imaging
MRI, CT

Image guided surgery
Grimson et al., MIT

Why is computer vision difficult?

Viewpoint variation

Illumination

Scale

Why is computer vision difficult?

Intra-class variation

Motion (Source: S. Lazebnik)

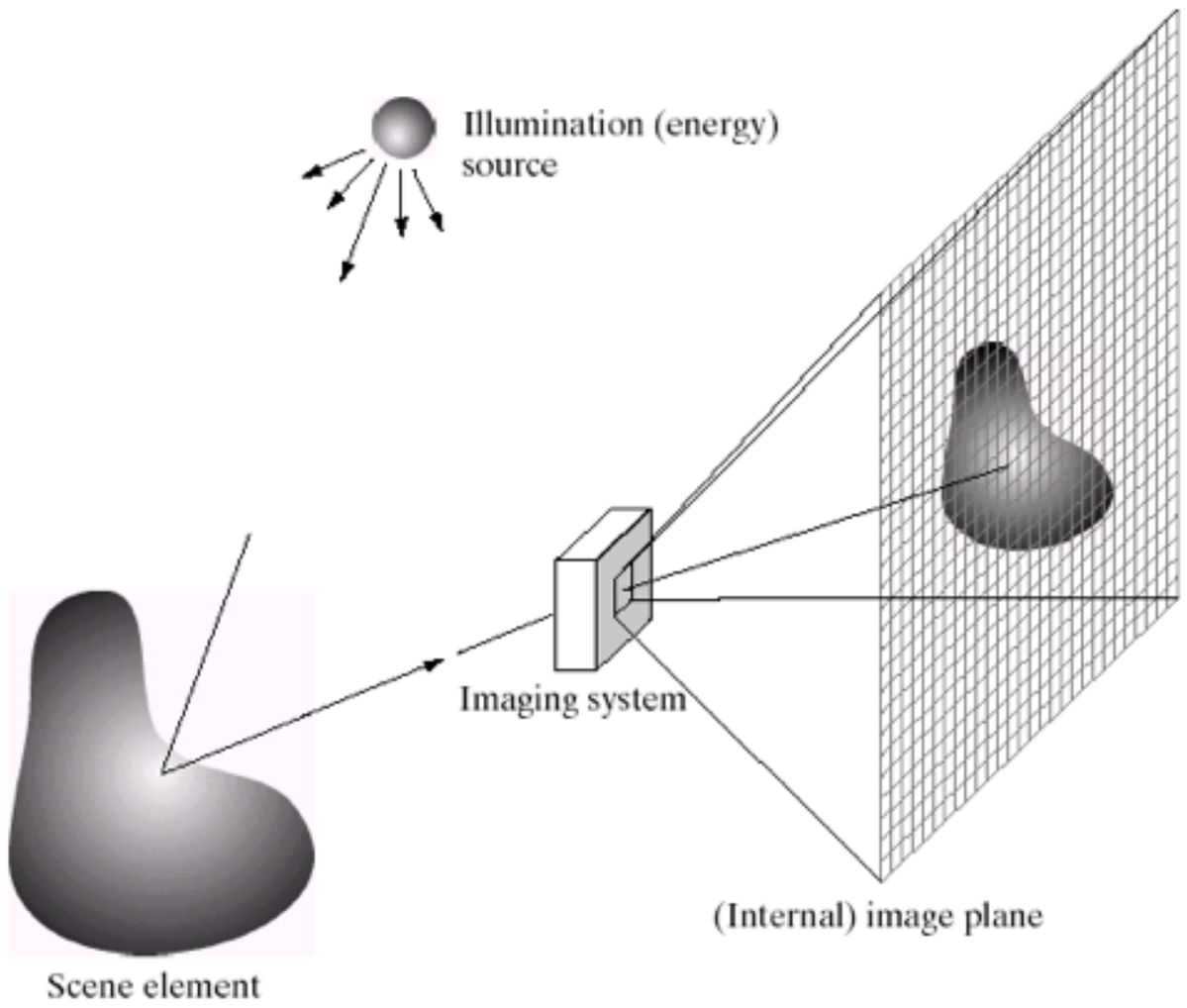
Background clutter

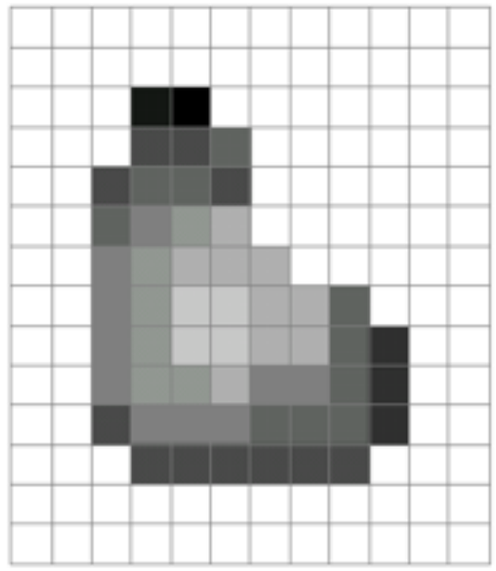
Occlusion

Fundamentals

What is an image?

What is an image?



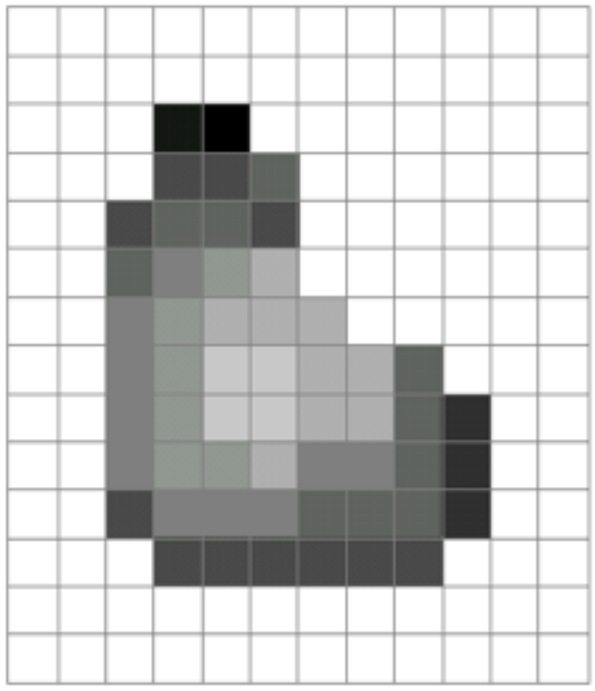


Digital Camera

Source: A. Efros

What is an image?

- A grid (matrix) of intensity values



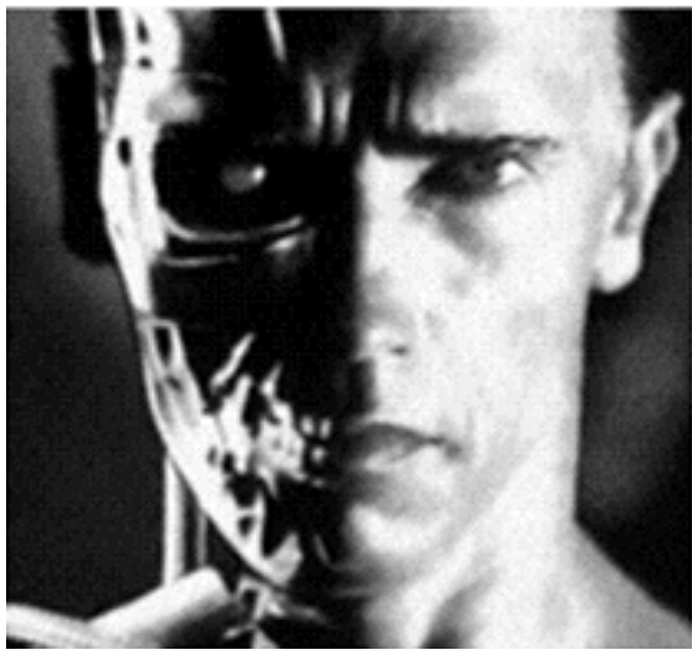
=

255	255	255	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	255	255
255	255	255	20	0	255	255	255	255	255	255	255
255	255	255	75	75	75	255	255	255	255	255	255
255	255	75	95	95	75	255	255	255	255	255	255
255	255	96	127	145	175	255	255	255	255	255	255
255	255	127	145	175	175	175	255	255	255	255	255
255	255	127	145	200	200	175	175	95	255	255	255
255	255	127	145	200	200	175	175	95	47	255	255
255	255	127	145	145	175	127	127	95	47	255	255
255	255	74	127	127	127	95	95	95	47	255	255
255	255	255	74	74	74	74	74	74	255	255	255
255	255	255	255	255	255	255	255	255	255	255	255
255	255	255	255	255	255	255	255	255	255	255	255

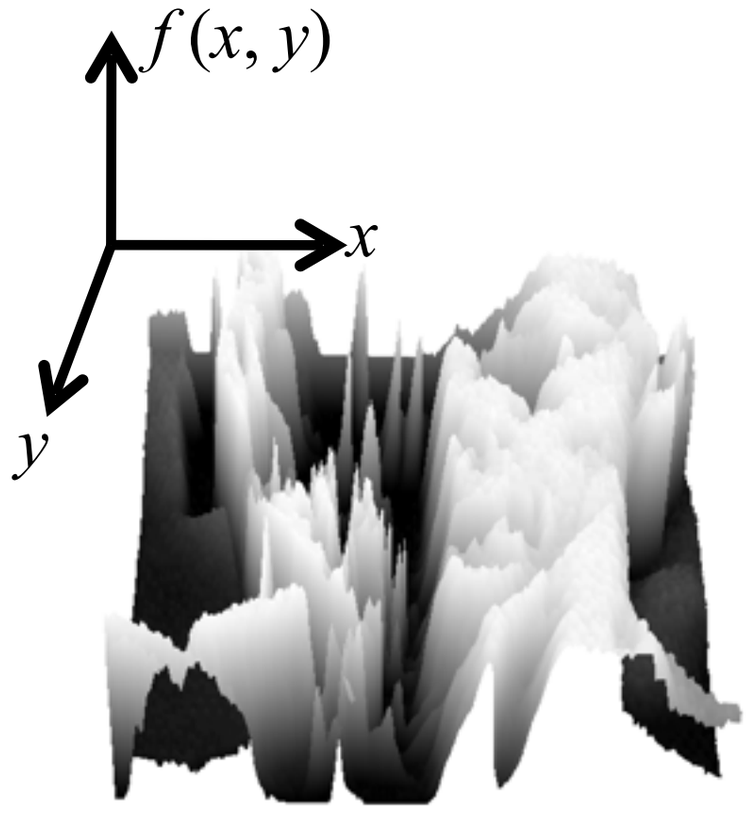
(common to use one byte per value: 0 = black, 255 = white)

What is an image?

- We can think of a (grayscale) image as a **function**, f, from R^2 to R:
  - $f(x,y)$ gives the **intensity** at position (x,y)



snoop

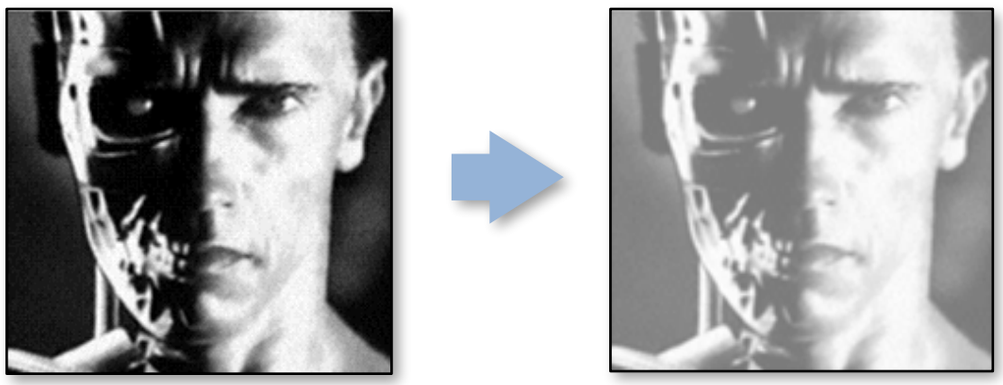


3D view

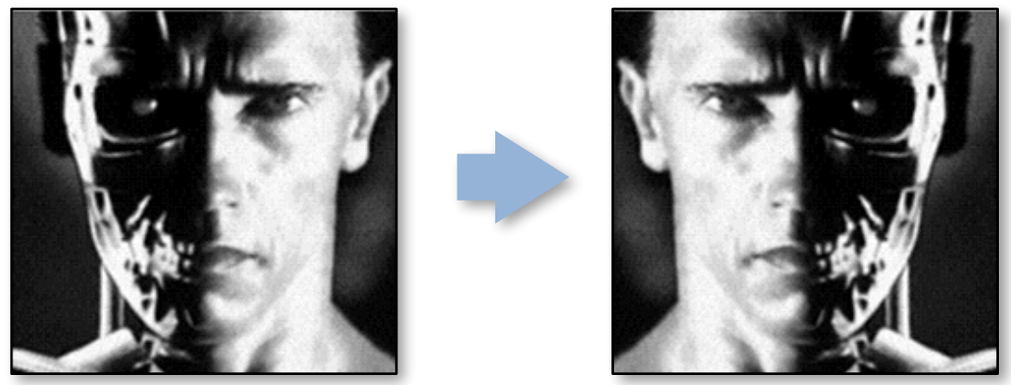
  - A **digital** image is a discrete (**sampled**, **quantized**) version of this function

Image transformations

- As with any function, we can apply operators to an image



$g(x,y) = f(x,y) + 20$



$g(x,y) = f(-x,y)$

- We'll talk about a special kind of operator, *convolution* (linear filtering)

Image filtering

- Modify the pixels in an image based on some function of a local neighborhood of each pixel

10	5	3
4	5	1
1	1	7

Local image data

Some function

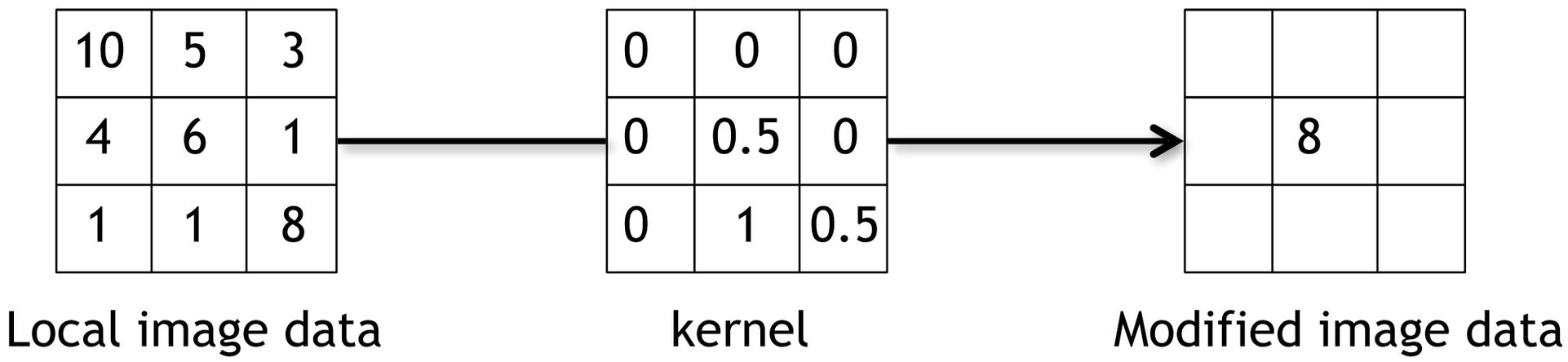
	7	

Modified image data

Source: L. Zhang

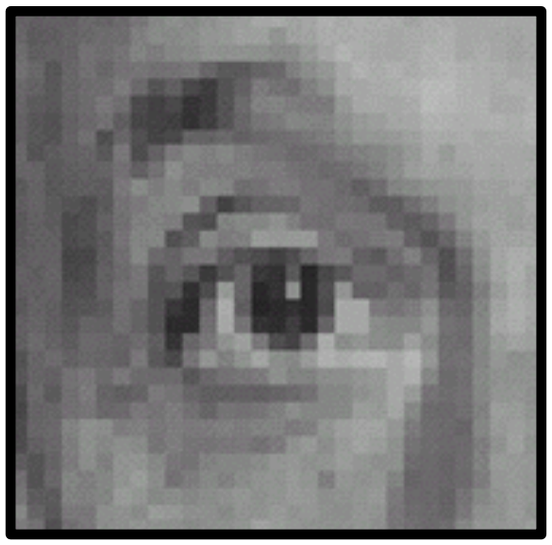
Linear filtering

- One simple version: linear filtering (cross-correlation, convolution)
  - Replace each pixel by a linear combination (a weighted sum) of its neighbors
- The prescription for the linear combination is called the "kernel" (or "mask", "filter")



Source: L. Zhang

Linear filters: examples

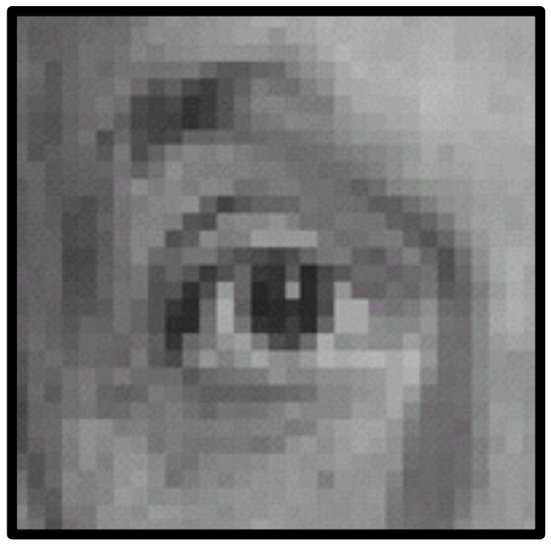


Original

$*$

0	0	0
0	1	0
0	0	0

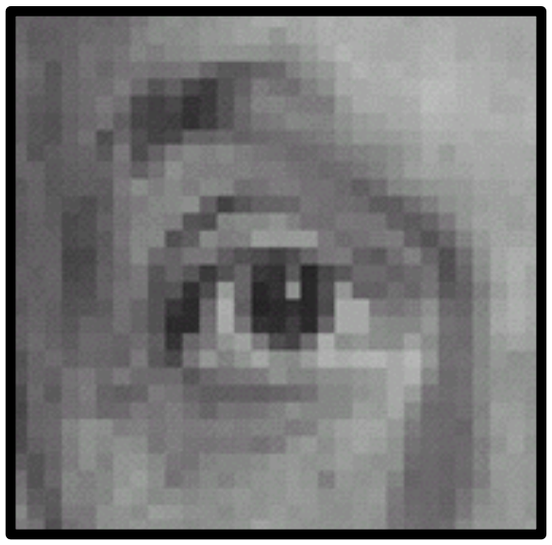
$=$



Identical image

Source: D. Lowe

Linear filters: examples

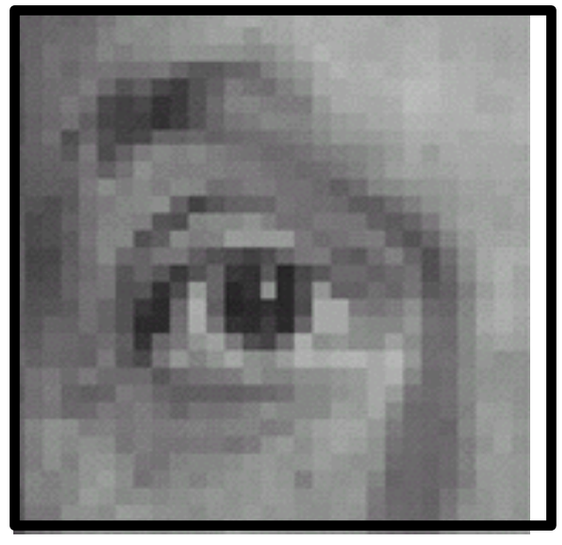


Original

$*$

0	0	0
1	0	0
0	0	0

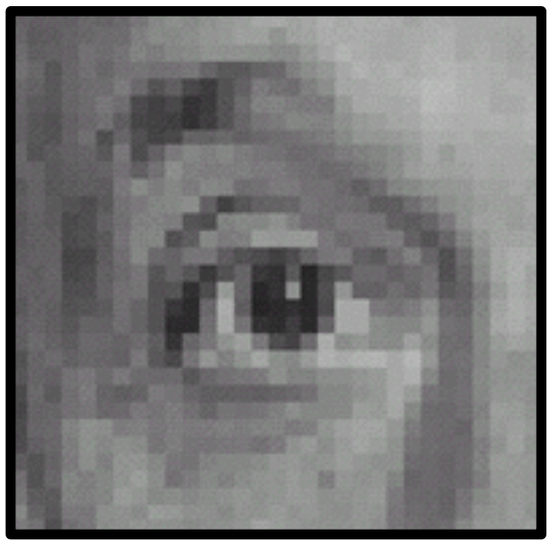
$=$



Shifted left
By 1 pixel

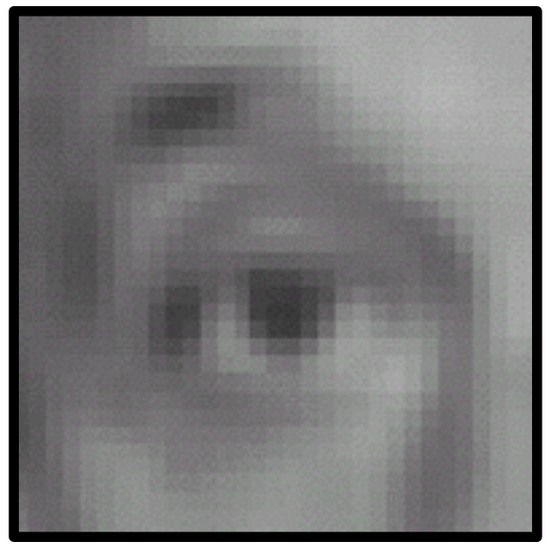
Source: D. Lowe

Linear filters: examples



Original

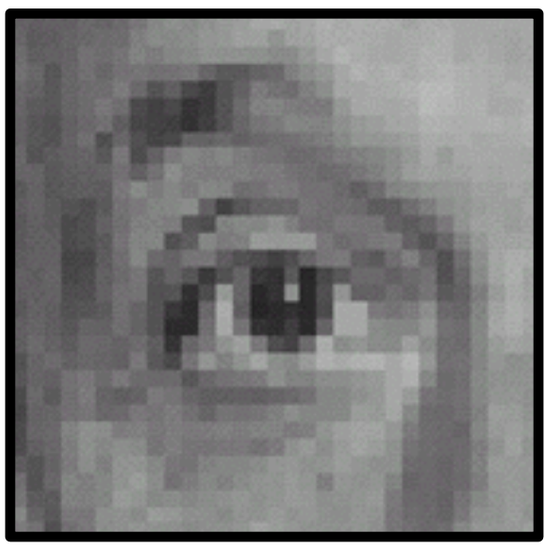
$$*\ \frac{1}{9}\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix} =$$



Blur (with a mean filter)

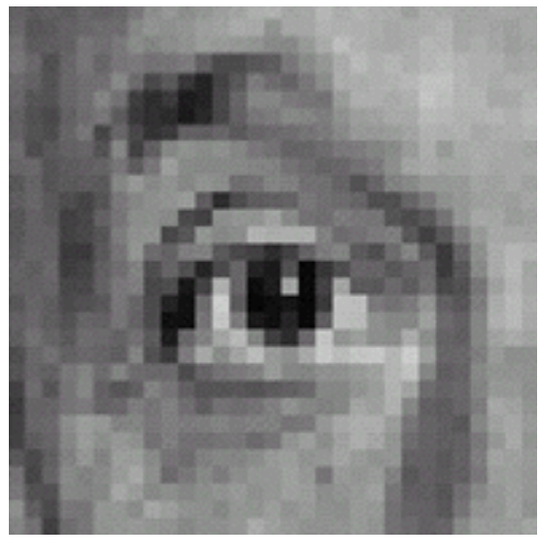
Source: D. Lowe

Linear filters: examples



Original

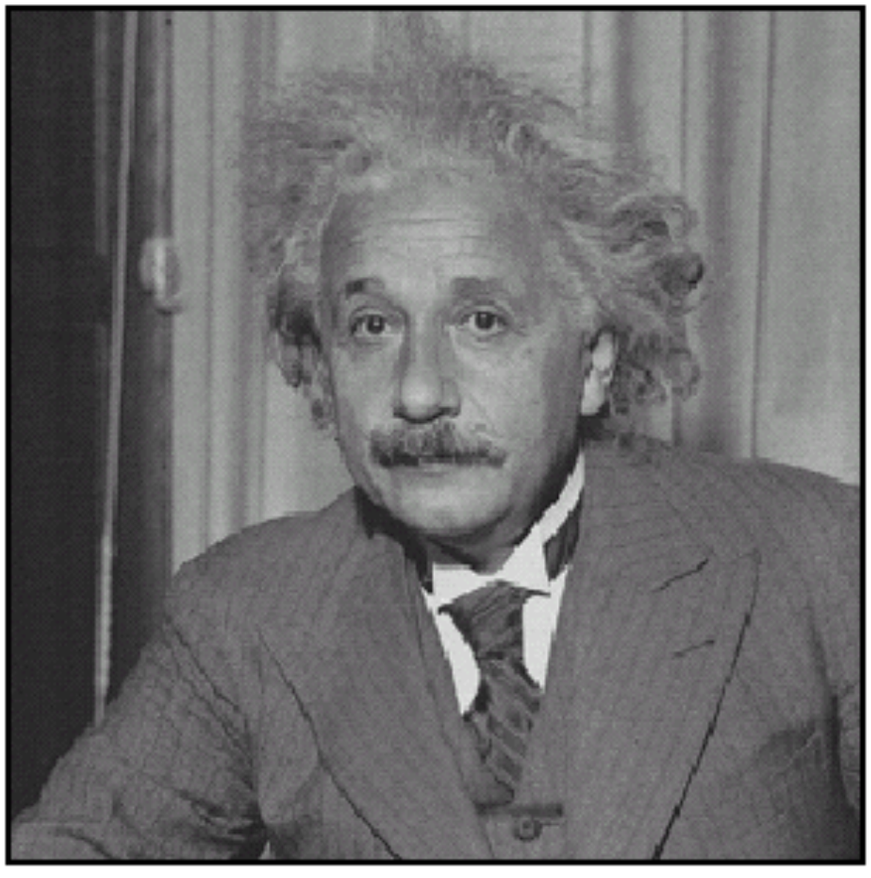
$$*\left(\begin{bmatrix}0&0&0\\0&2&0\\0&0&0\end{bmatrix}-\frac{1}{9}\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}\right)=$$



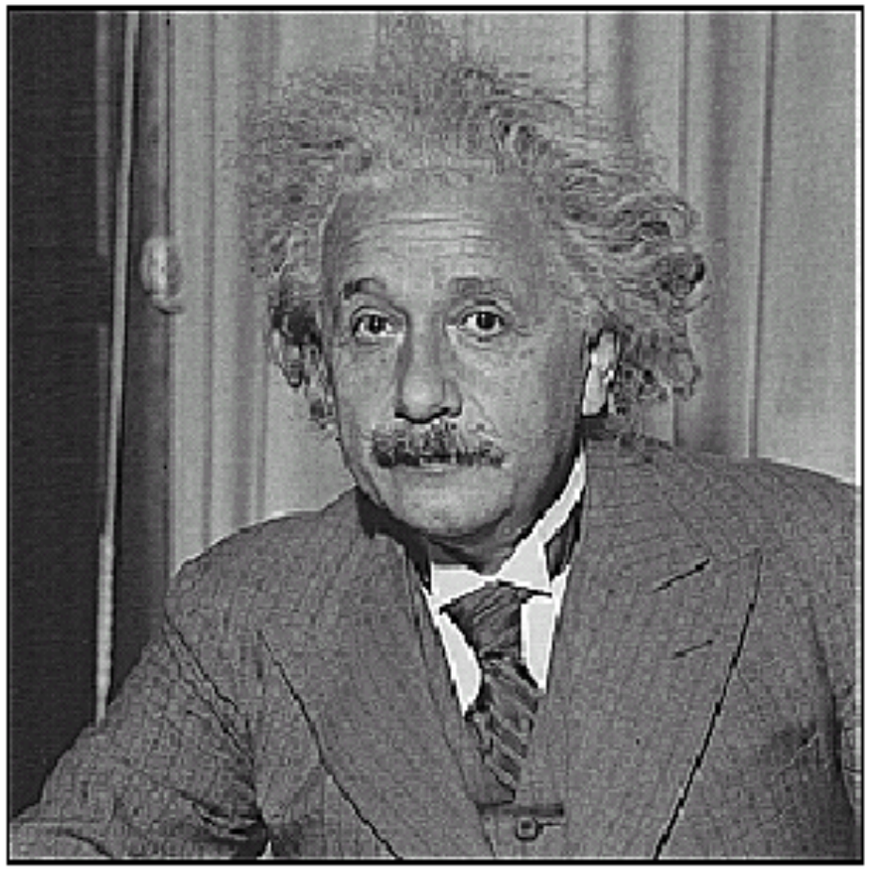
Sharpening filter
(accentuates edges)

Source: D. Lowe

Sharpening



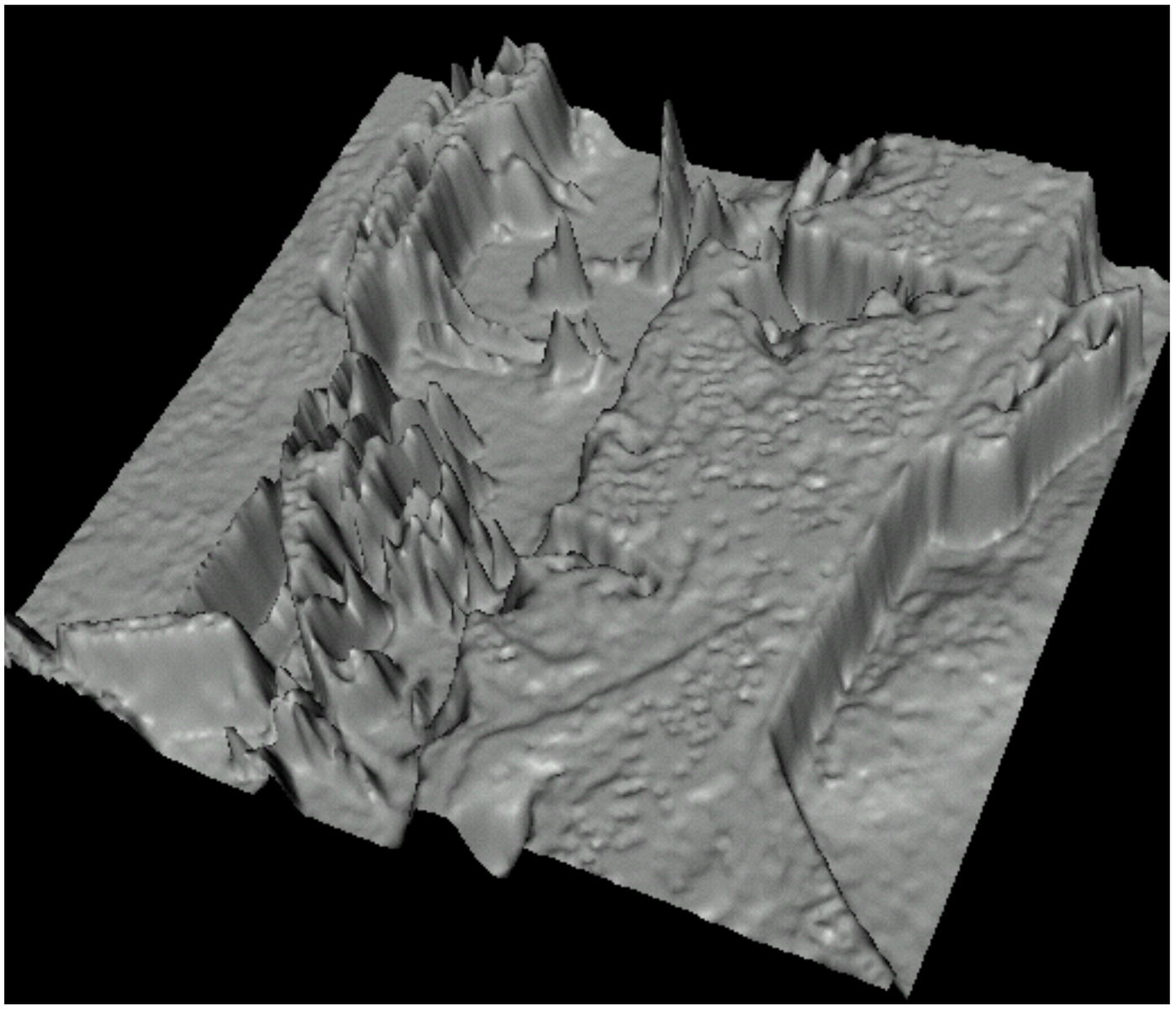
before



after

Source: D. Lowe

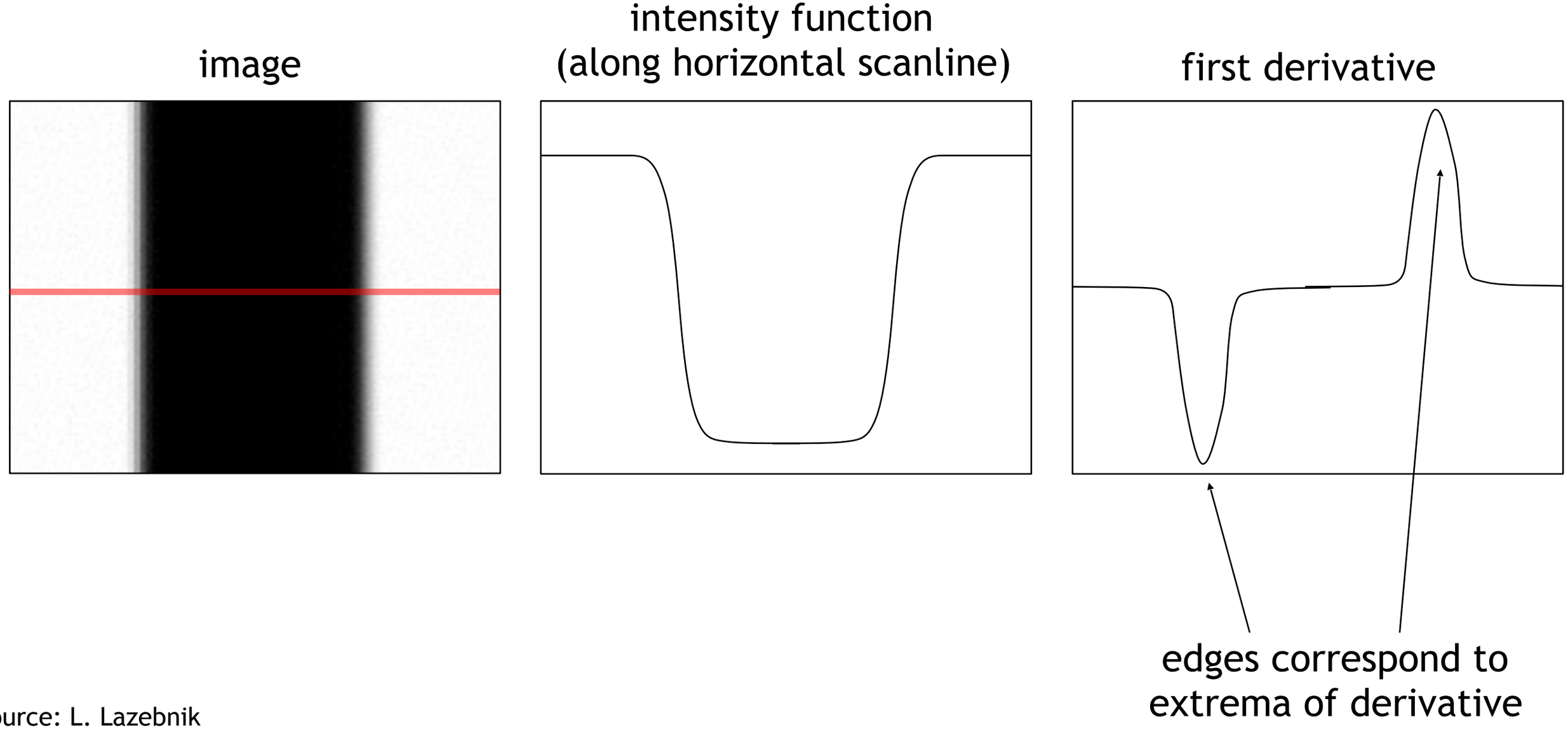
Images as functions…



- Edges look like steep cliffs

Characterizing edges

- An edge is a place of *rapid change* in the image intensity function



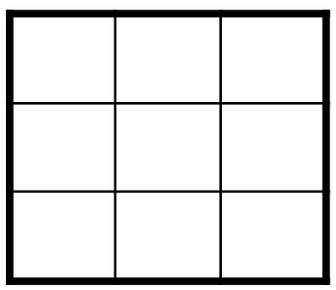
Source: L. Lazebnik

Image derivatives

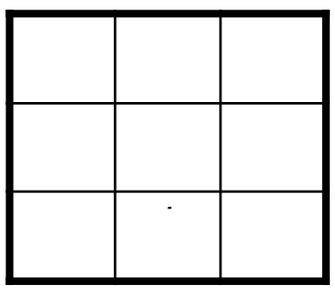
- How can we differentiate a *digital* image F[x,y]?
  - Option 1: reconstruct a continuous image, *f*, then compute the derivative
  - Option 2: take discrete derivative (finite difference)

$$\frac{\partial f}{\partial x}[x, y] \approx F[x+1, y] - F[x, y]$$

How would you implement this as a linear filter?

$\frac{\partial f}{\partial x}$: 

H_x

$\frac{\partial f}{\partial y}$: 

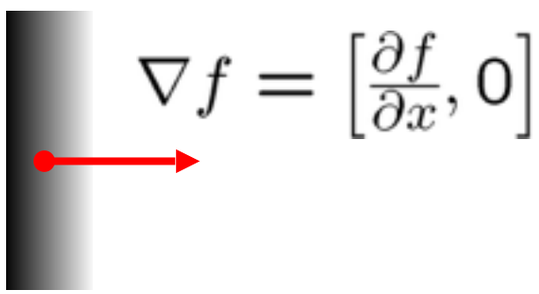
H_y

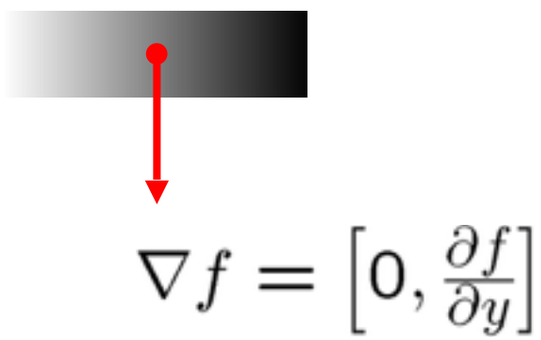
Source: S. Seitz

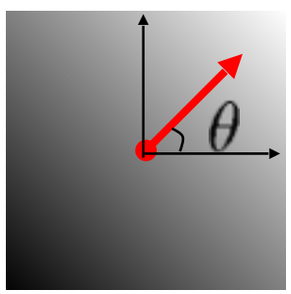
Image gradient

- The *gradient* of an image: $\nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right]$

The gradient points in the direction of most rapid increase in intensity







The *edge strength* is given by the gradient magnitude:

$$\|\nabla f\| = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}$$

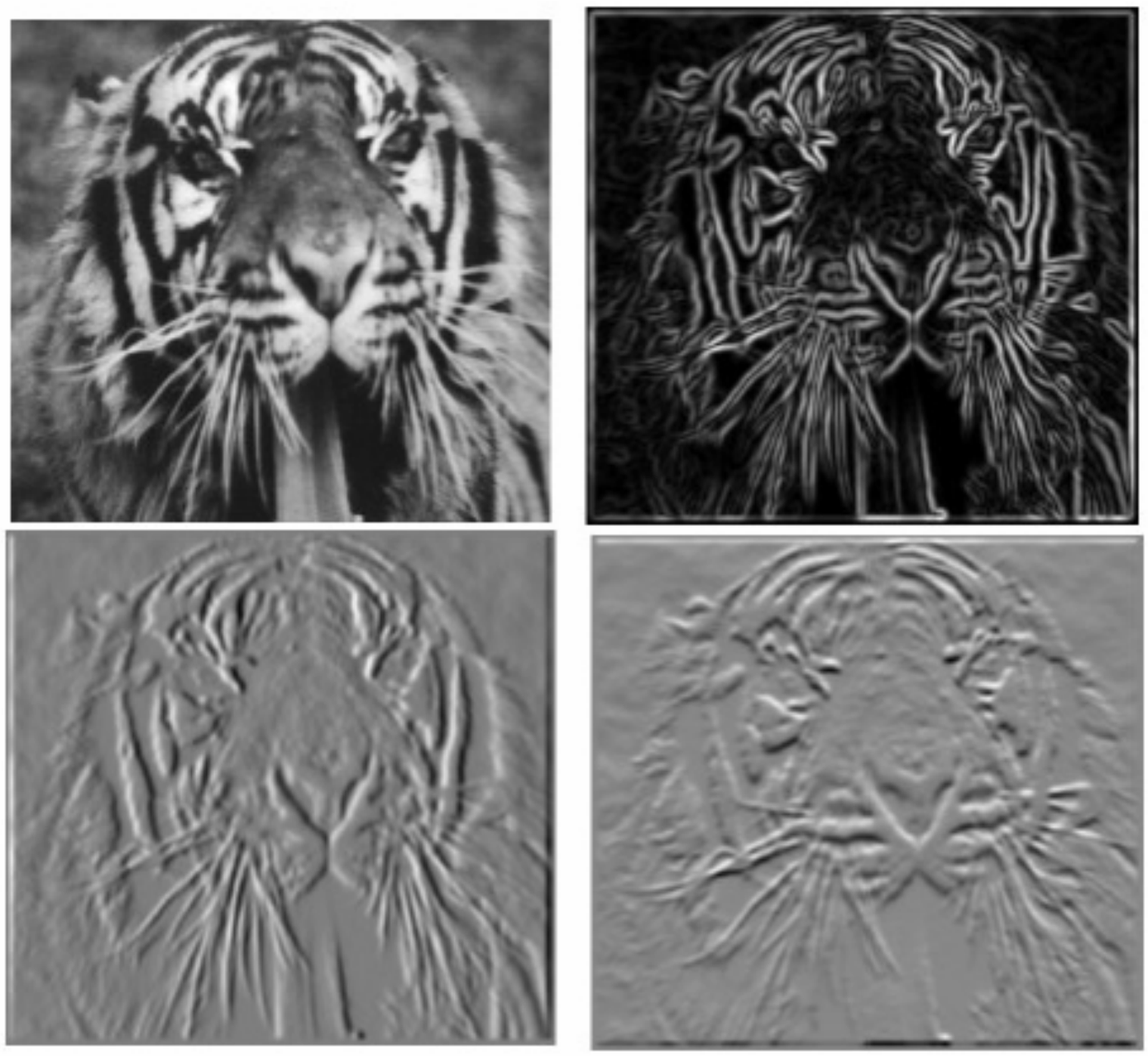
The gradient direction is given by:

$$\theta = \tan^{-1}\left(\frac{\partial f}{\partial y} / \frac{\partial f}{\partial x}\right)$$

- how does this relate to the direction of the edge?

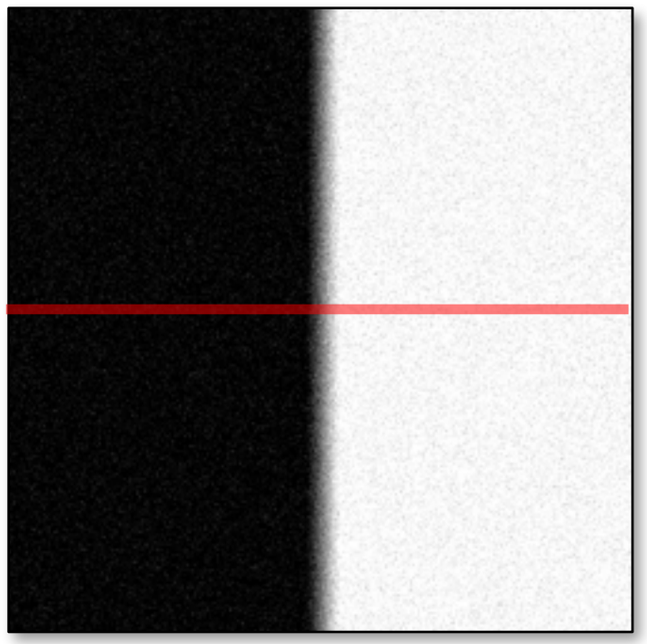
Source: Steve Seitz

Image gradient



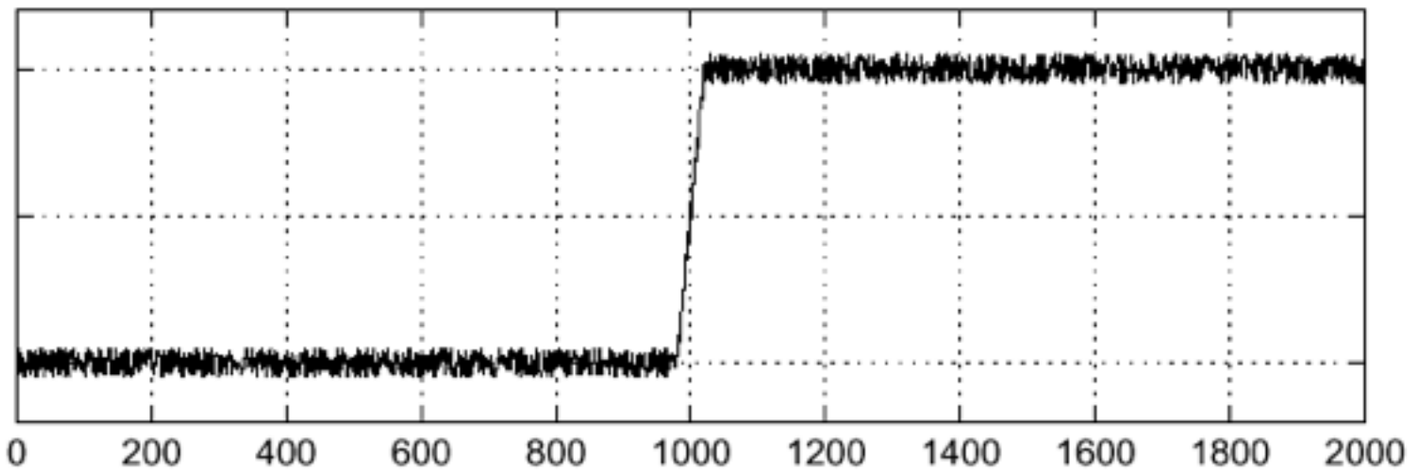
Source: L. Lazebnik

Effects of noise

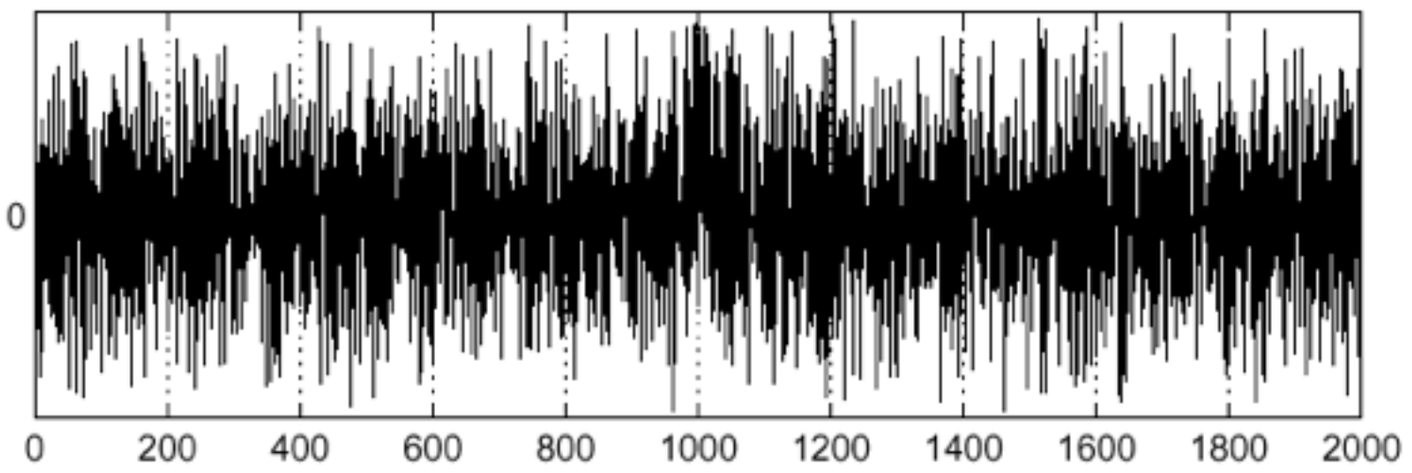


Noisy input image

$f(x)$



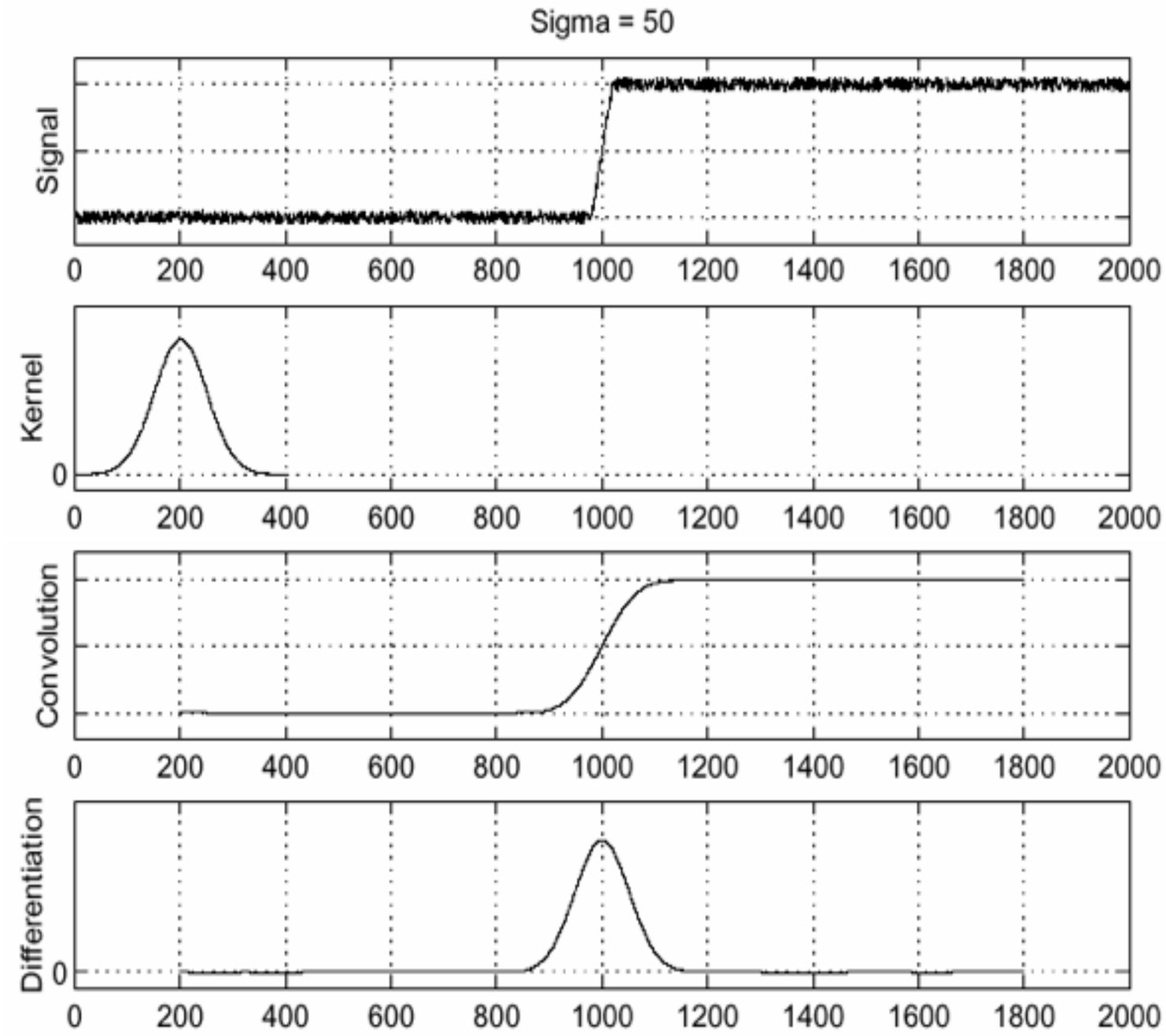
$\frac{d}{dx}f(x)$



Where is the edge?

Source: S. Seitz

Solution: smooth first



f

h

$f * h$

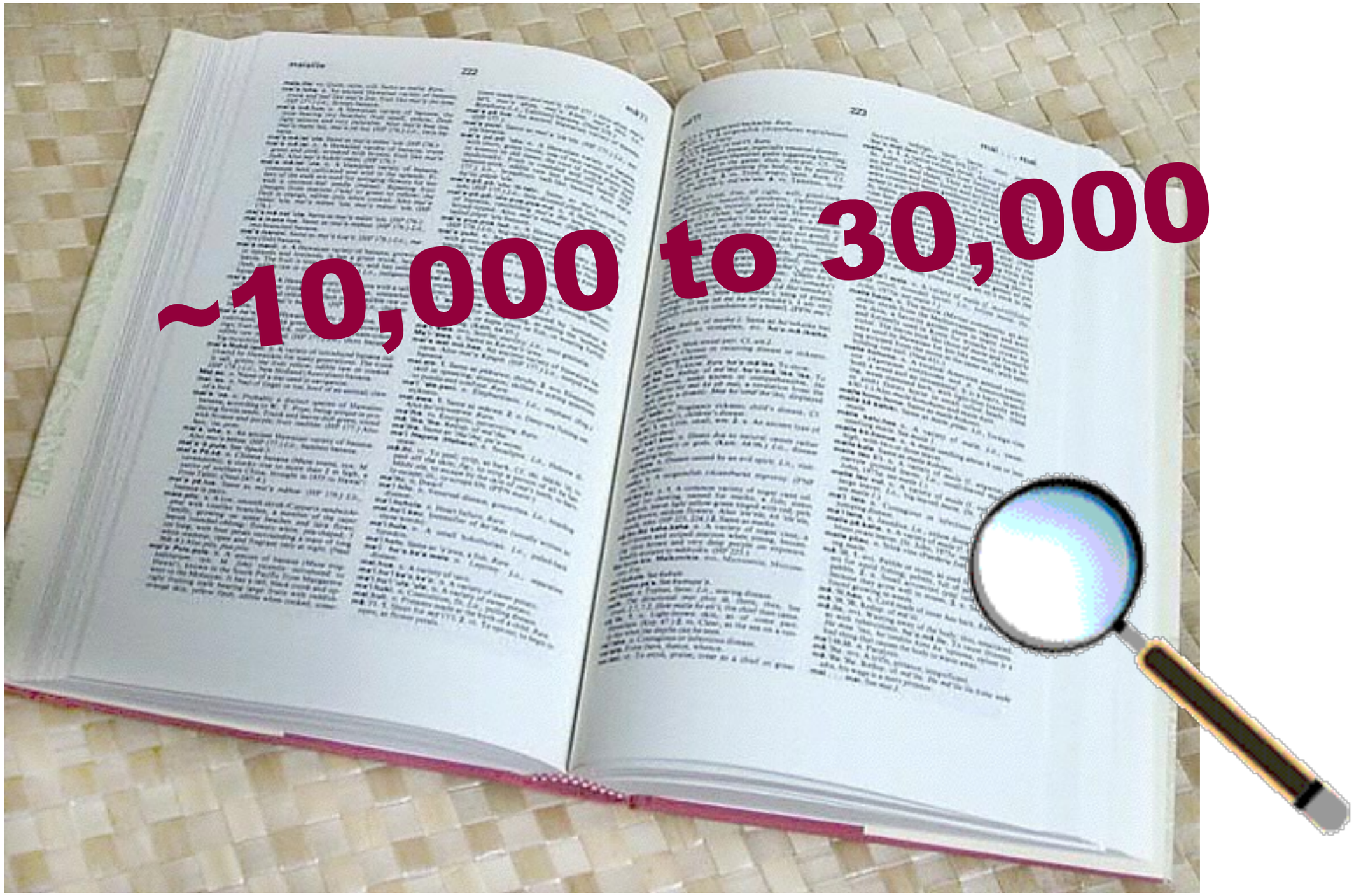
$$\frac{d}{dx}(f * h)$$

To find edges, look for peaks i $\frac{d}{dx}(f * h)$

Source: S. Seitz

Recognition

How many visual object categories are there?



Biederman 1987

~10,000 to 30,000

OBJECTS

ANIMALS

PLANTS

INANIMATE

.....

VERTEBRATE

NATURAL

MAN-MADE

MAMMALS

BIRDS

TAPIR

BOAR

GROUSE

CAMERA

Specific recognition tasks

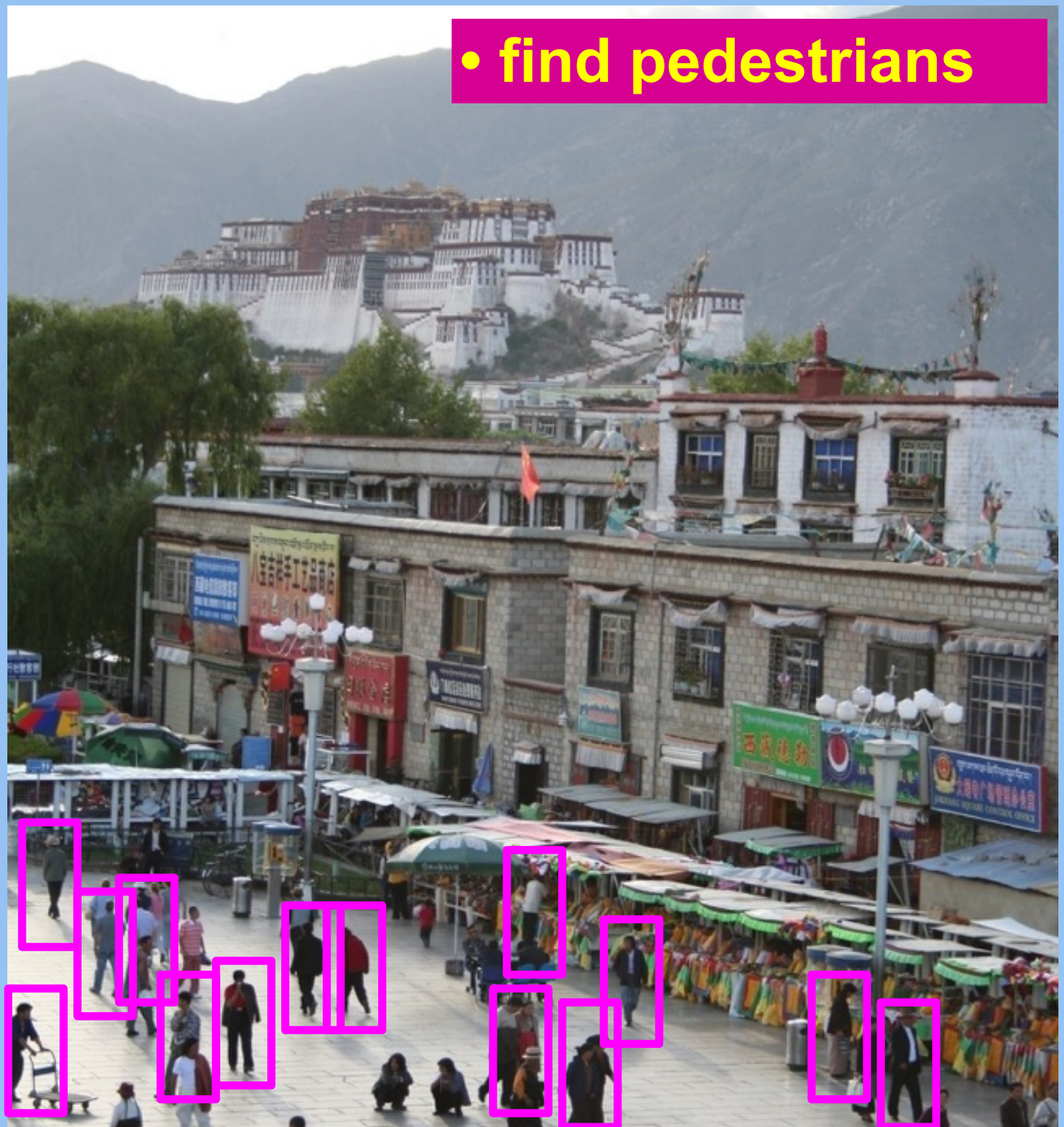
Svetlana Lazebnik

Scene categorization or classification

- **outdoor/indoor**
- **city/forest/factory/etc.**

Svetlana Lazebnik

Object detection

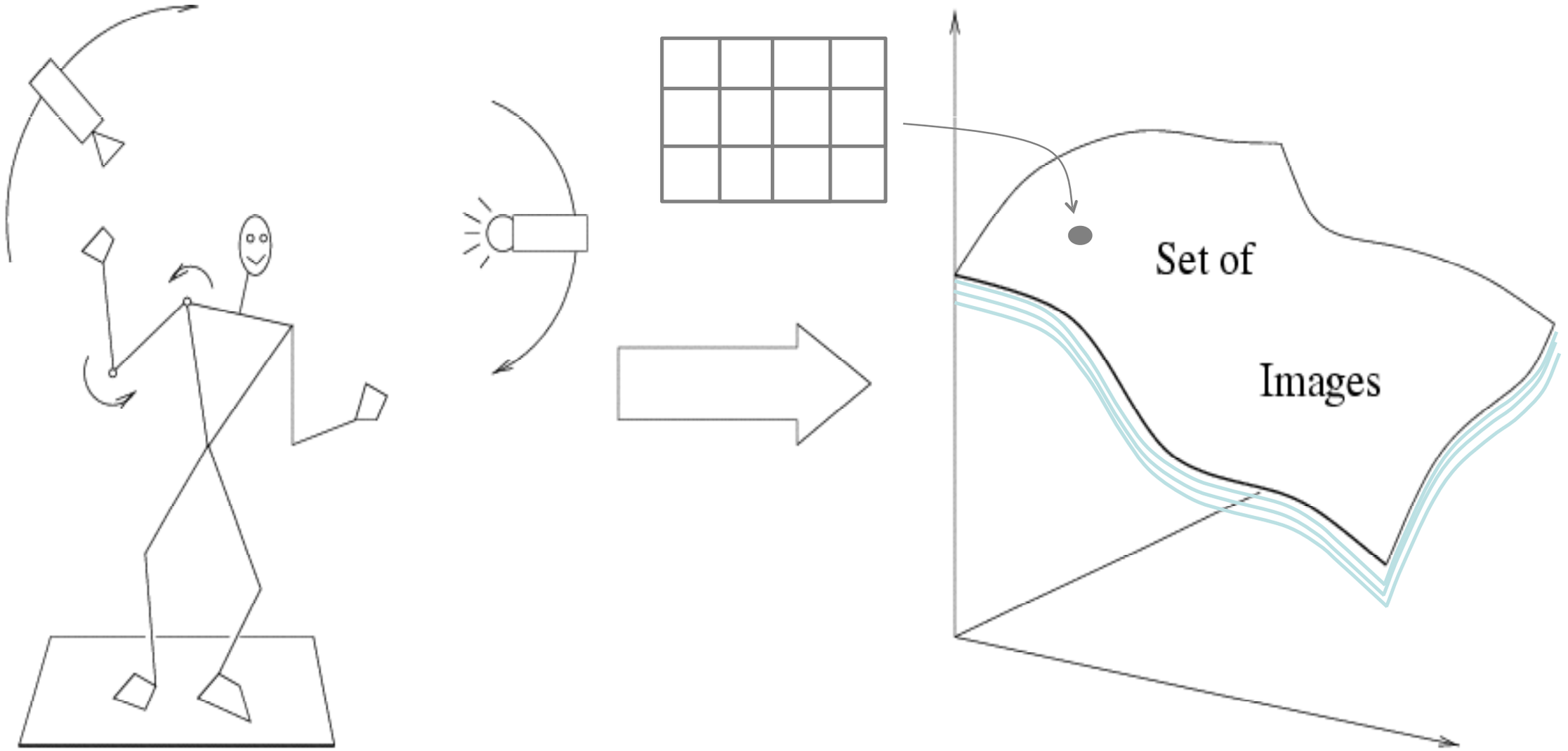


Svetlana Lazebnik

Image parsing / semantic segmentation

Svetlana Lazebnik

Recognition is all about modeling variability



Variability: Camera position
Illumination
Shape parameters

Within-class variations?

Train

Test


Coca-Cola

Train


1.25 LITRES
Coca-Cola
real taste. uplifting refreshment

Test

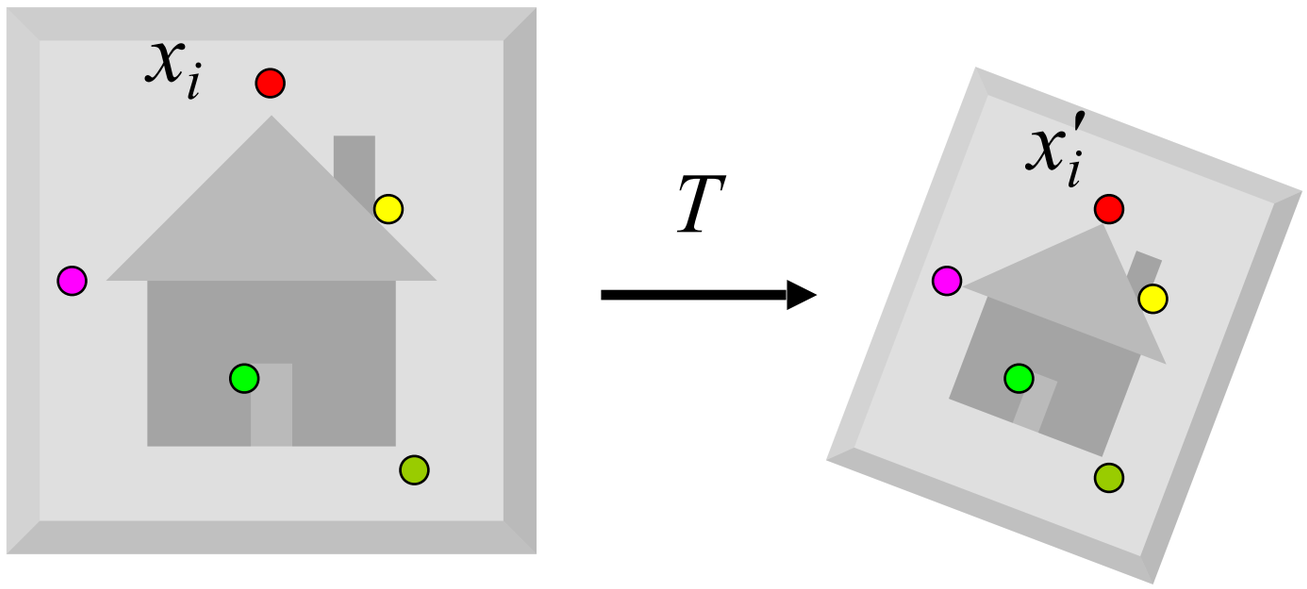
Train

Test

Within-class variations

Alignment

- Fitting a model to a transformation between pairs of features (*matches*) in two images

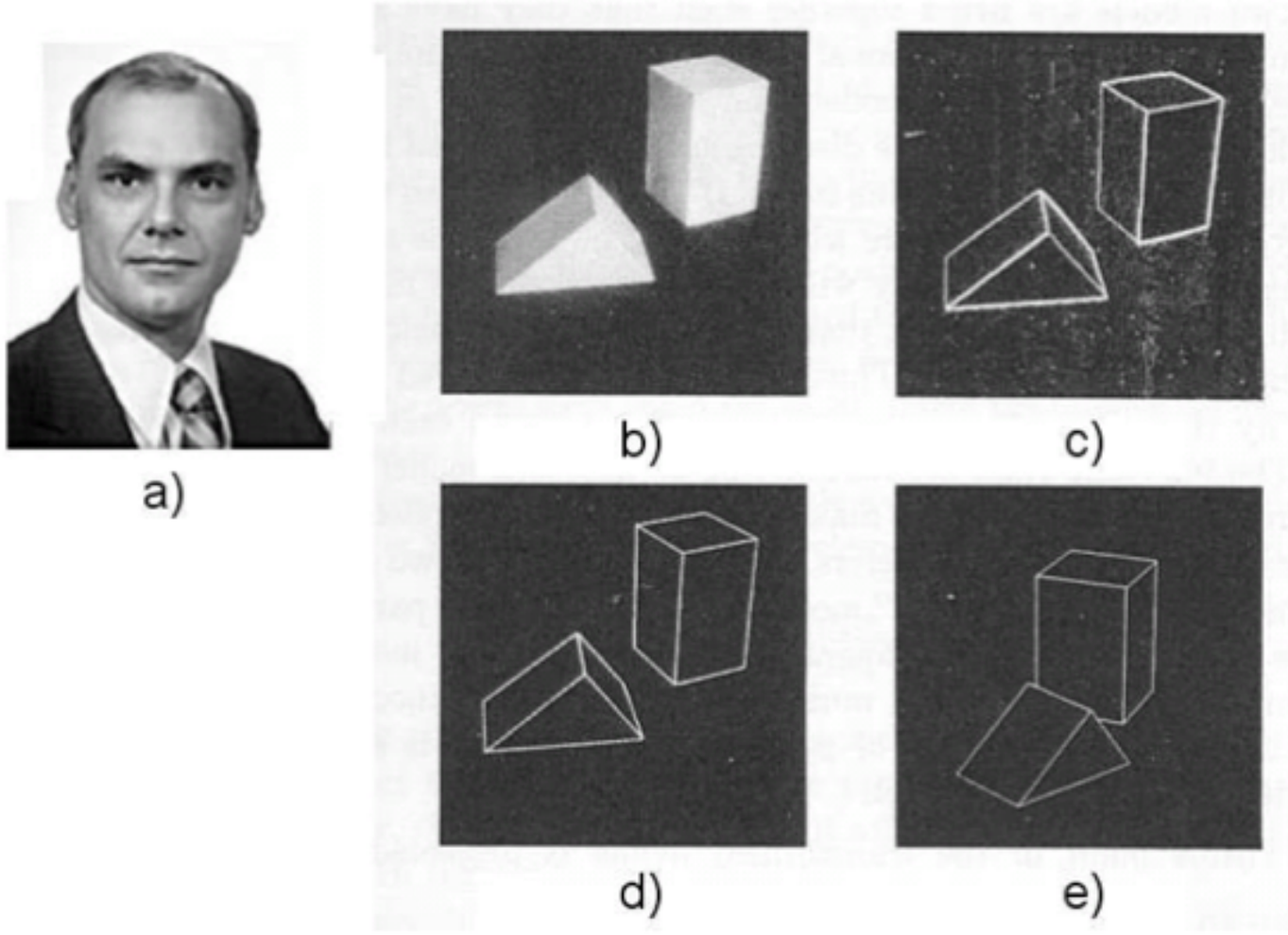


Find transformation T that minimizes

$$\sum_i \text{residual}(T(x_i), x_i')$$

Svetlana Lazebnik

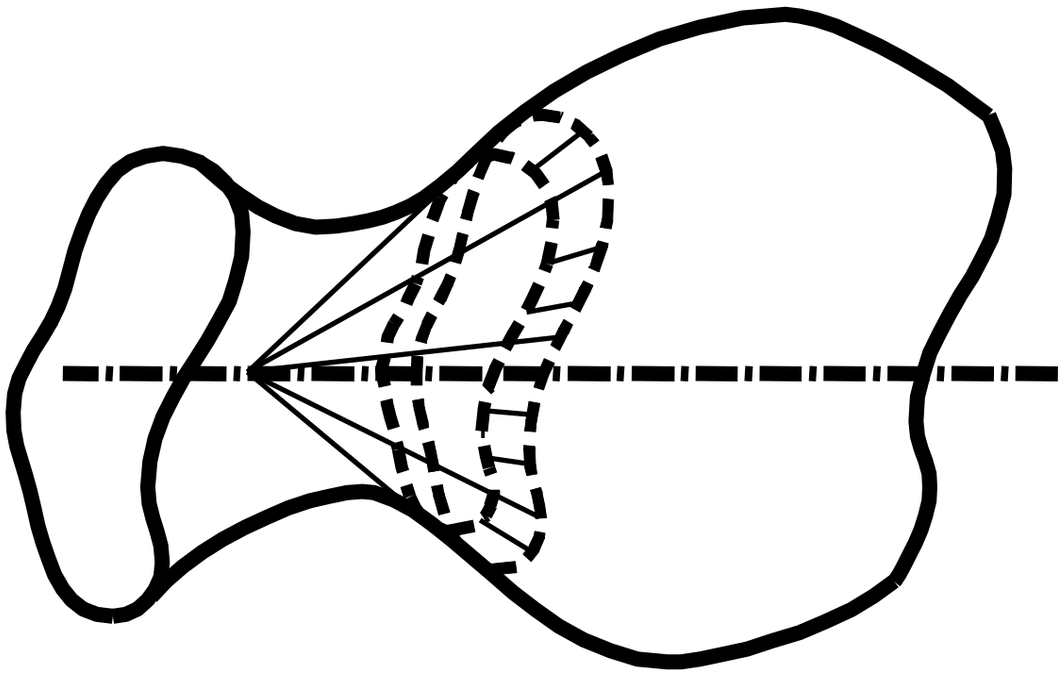
Recognition as an alignment problem: Block world



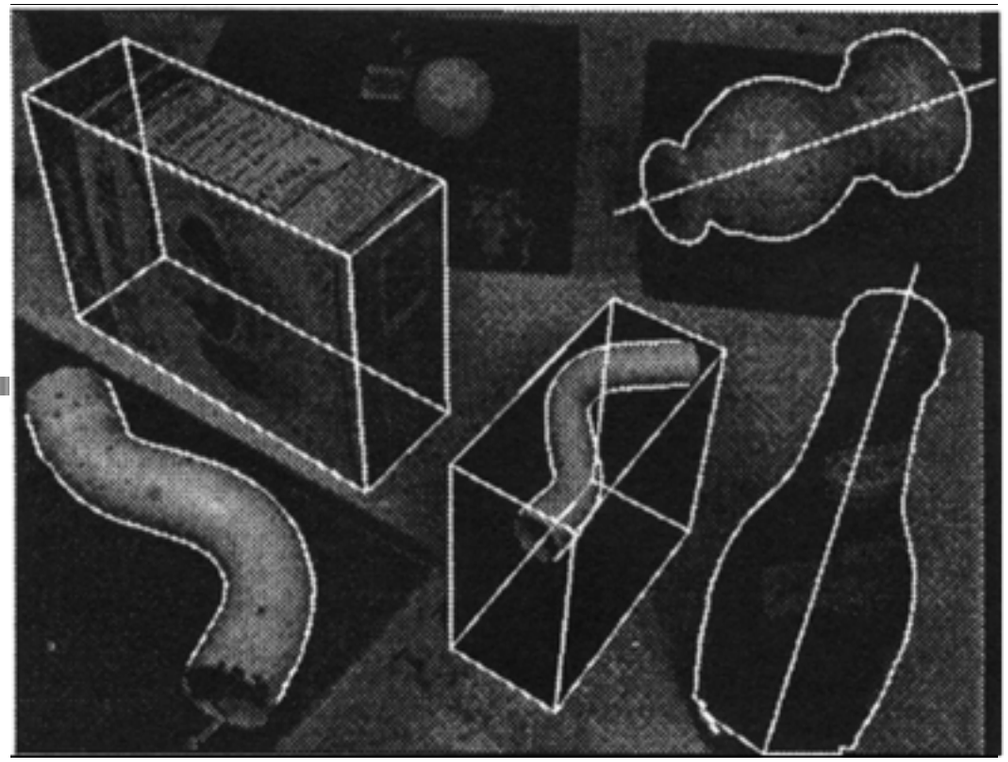
L. G. Roberts, *Machine Perception of Three Dimensional Solids*, Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

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.)

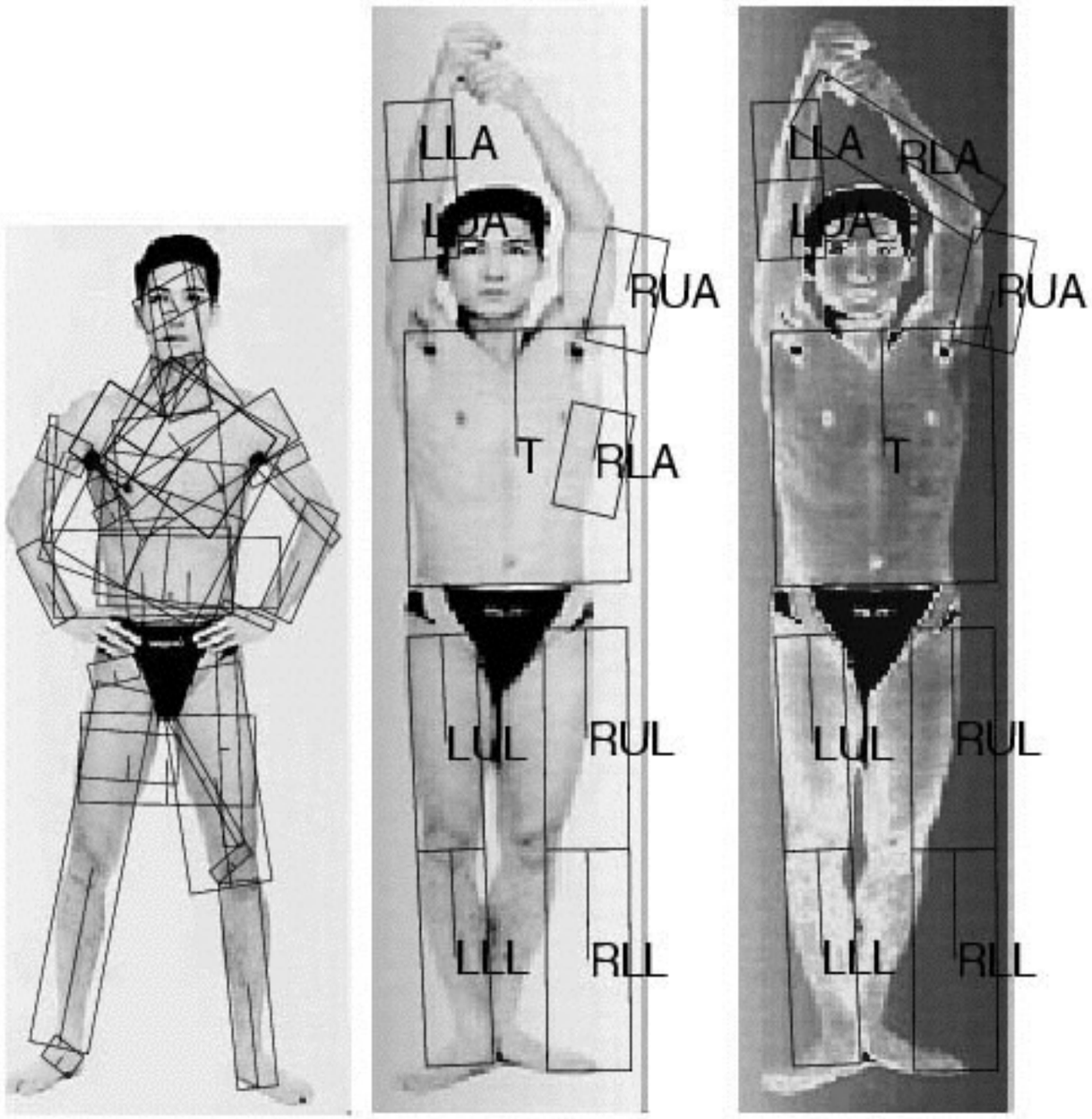
J. Mundy, Object Recognition in the Geometric Era: a Retrospective, 2006



Generalized cylinders
Ponce et al. (1989)

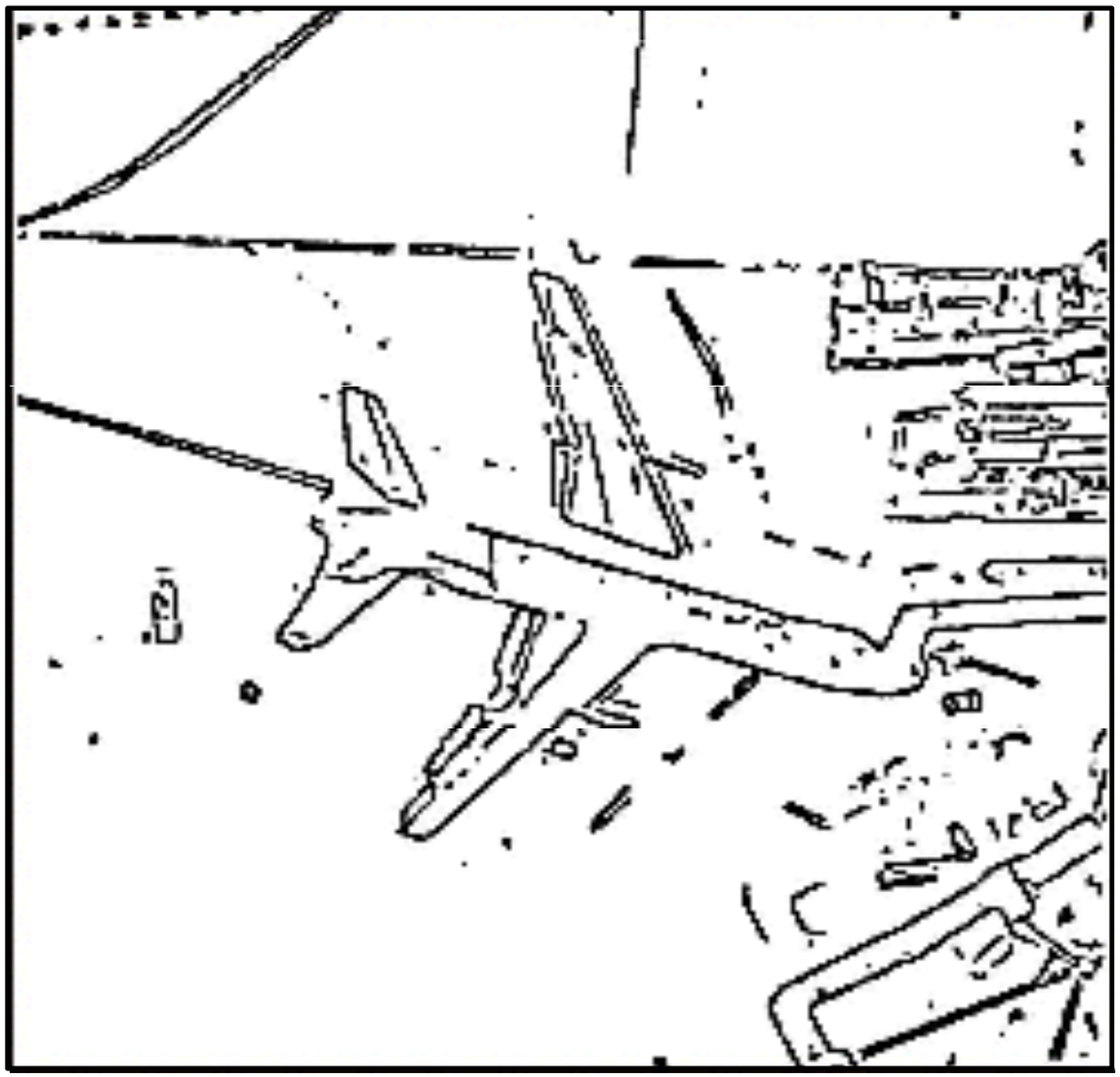


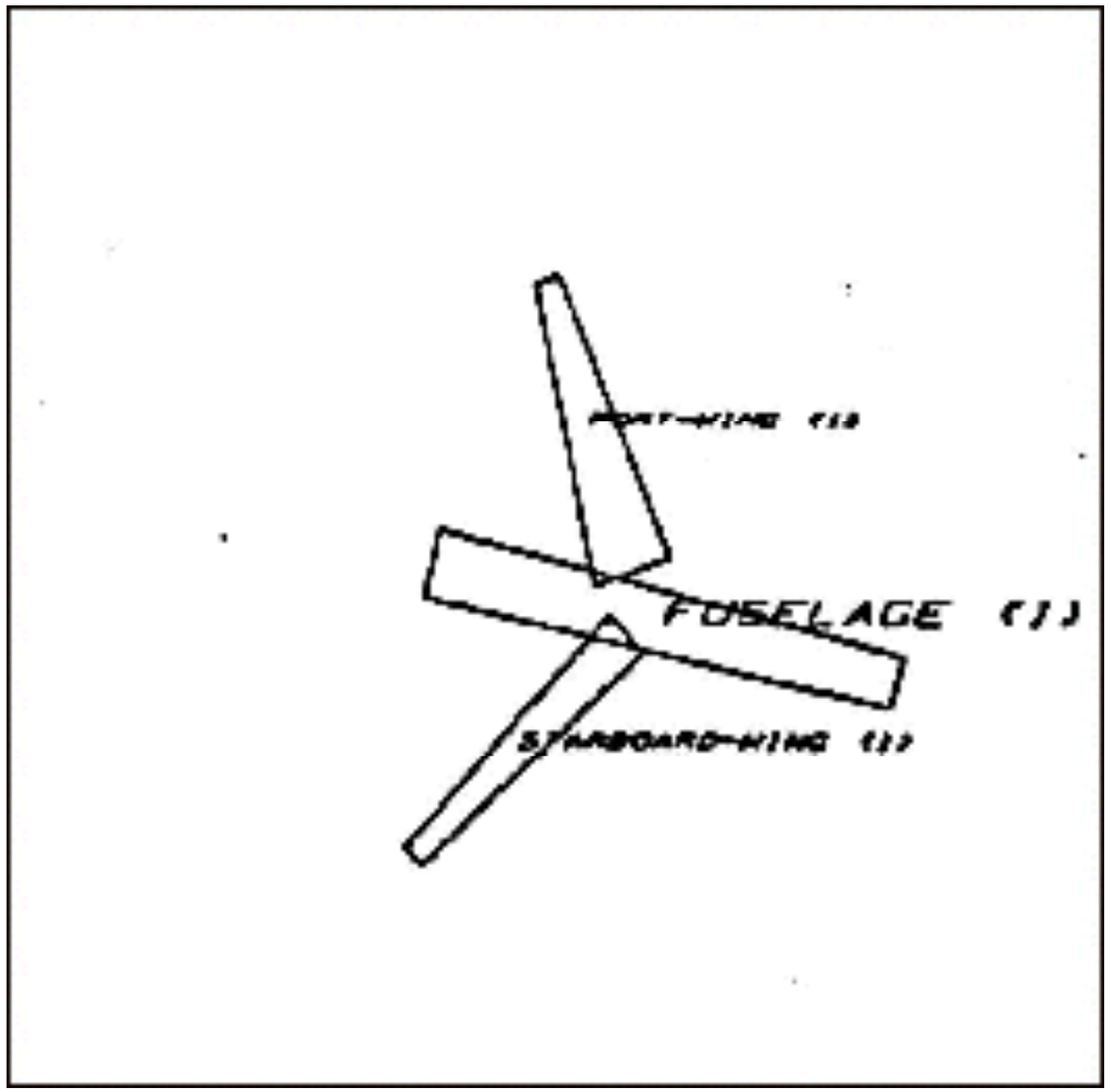
Zisserman et al. (1995)



Forsyth (2000)

Representing and recognizing object categories is harder...





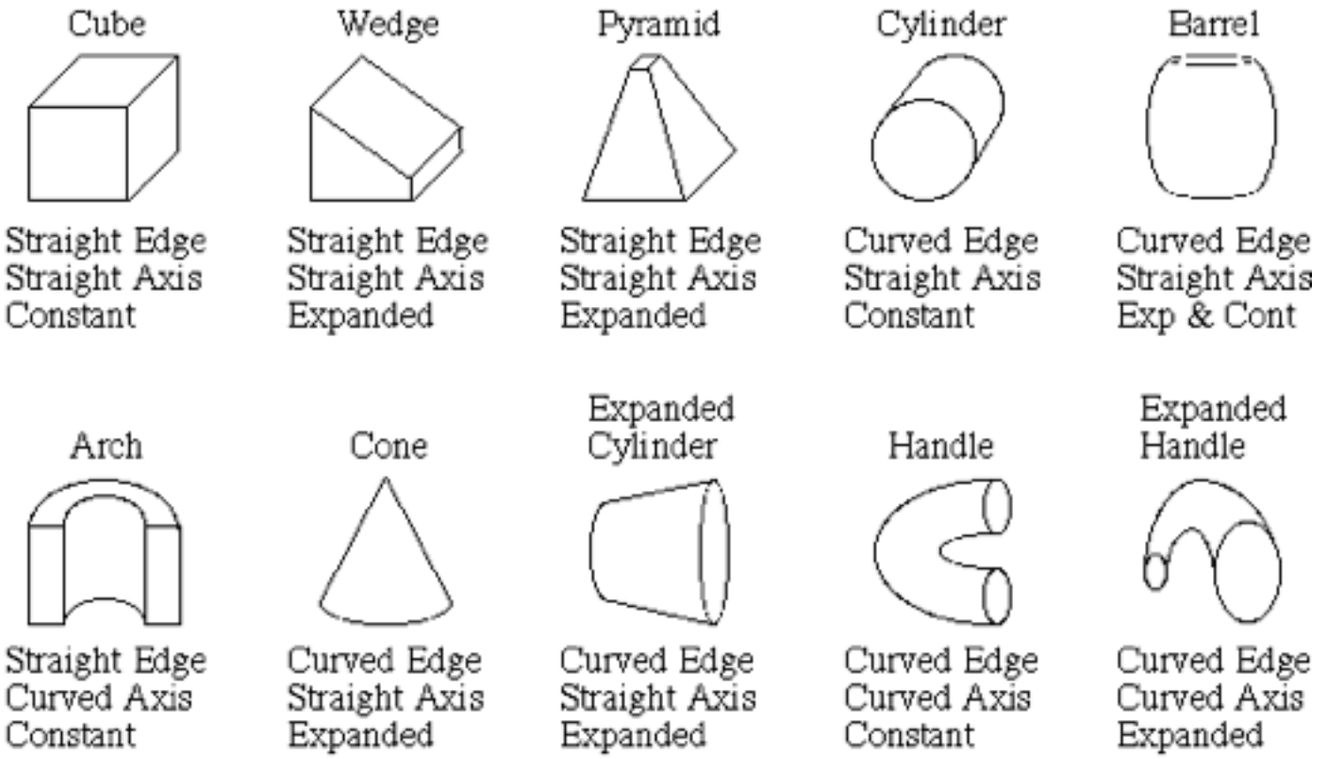
ACRONYM (Brooks and Binford, 1981)

Binford (1971), Nevatia & Binford (1972), Marr & Nishihara (1978)

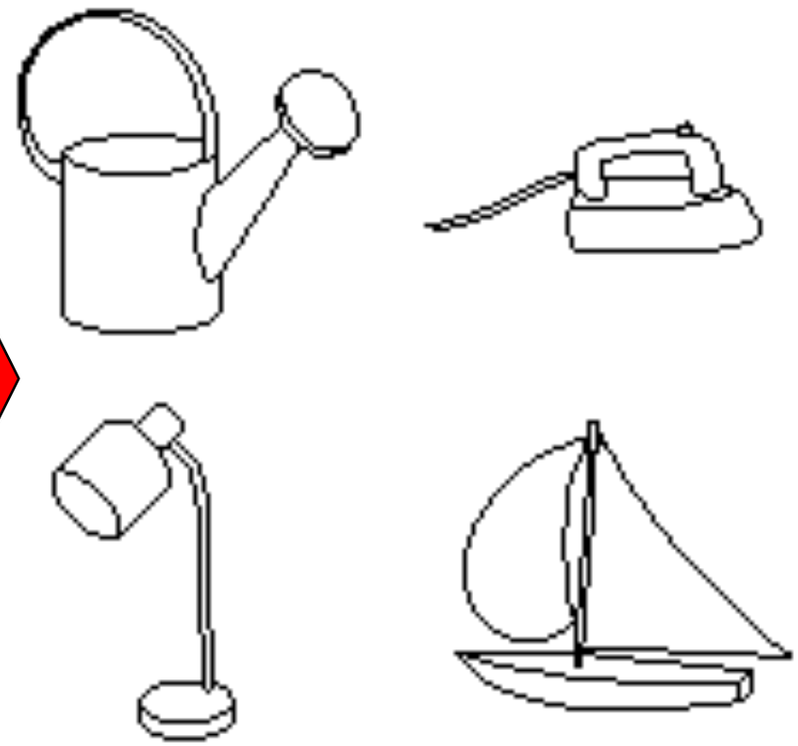
Recognition by components

Biederman (1987)

Primitives (geons)



Objects

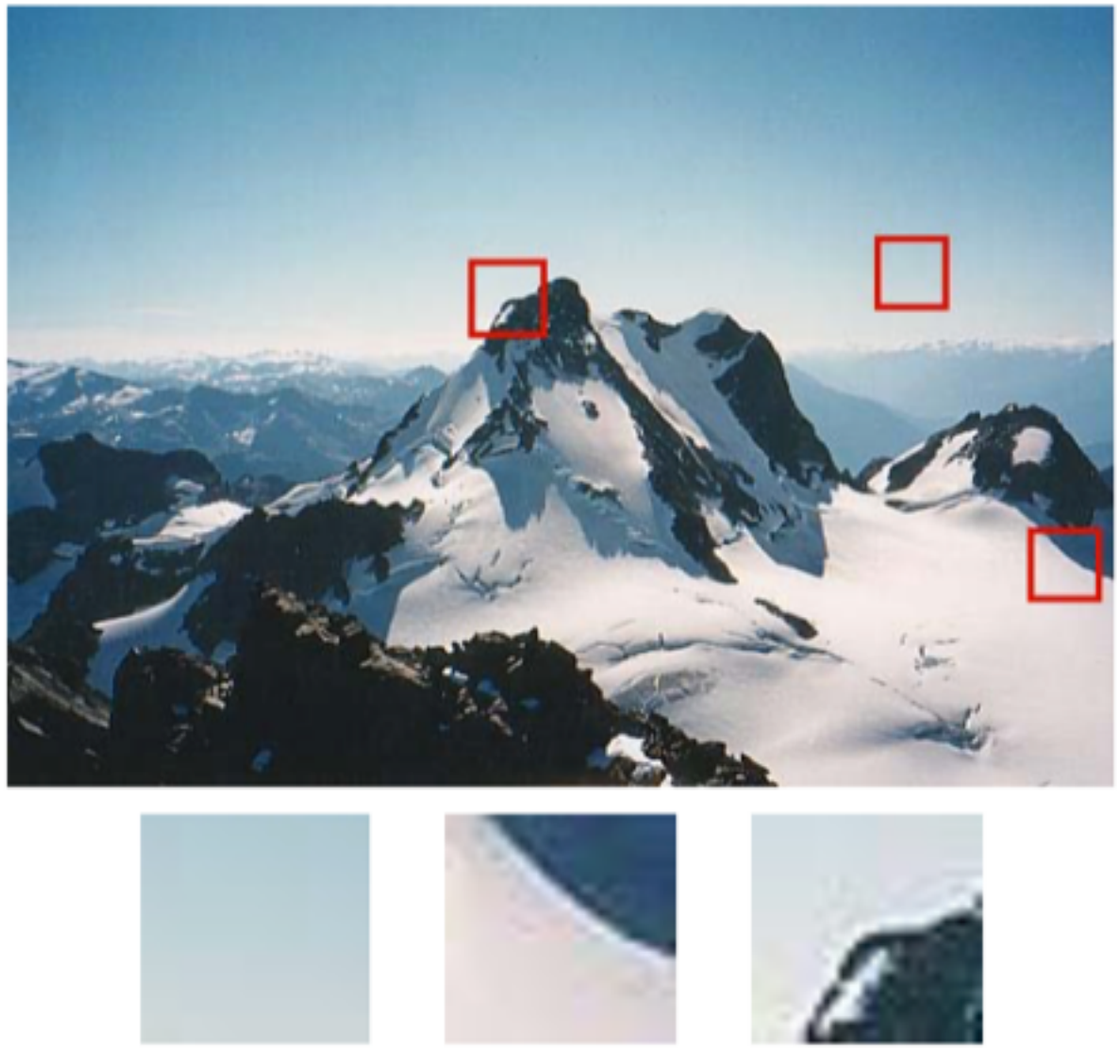


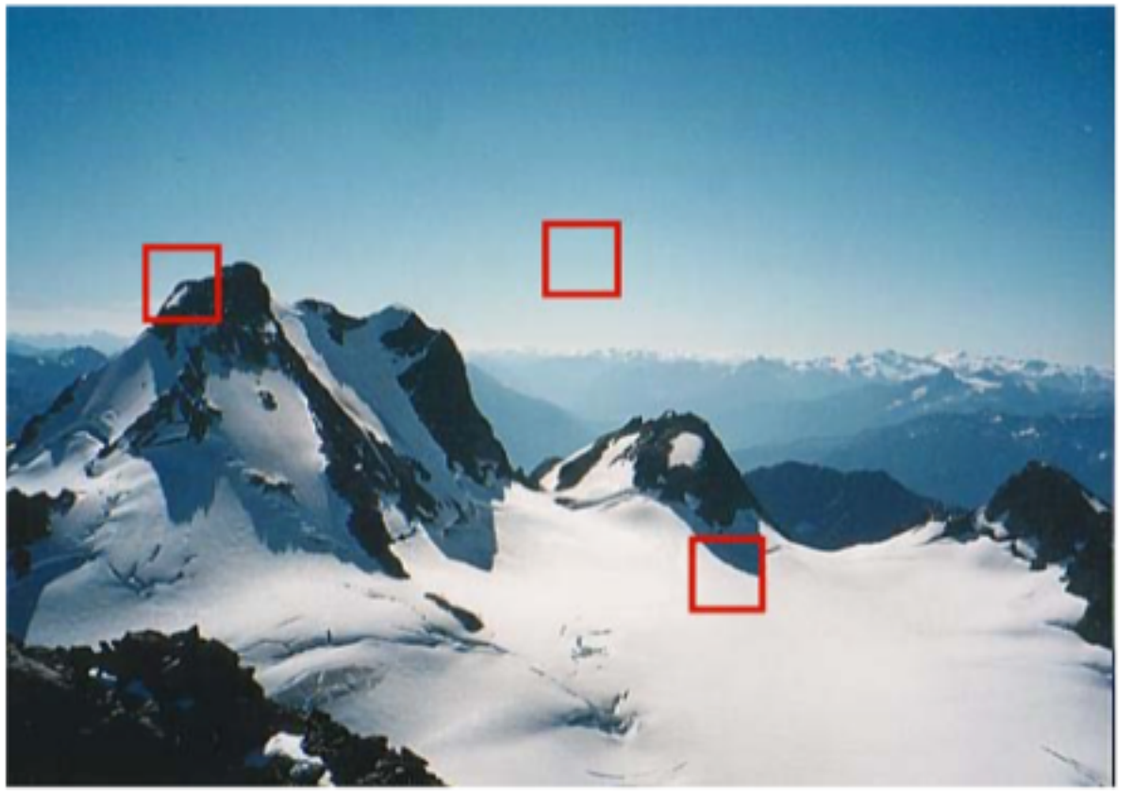
http://en.wikipedia.org/wiki/Recognition_by_Components_Theory

Local features for object instance recognition

D. Lowe (1999, 2004)

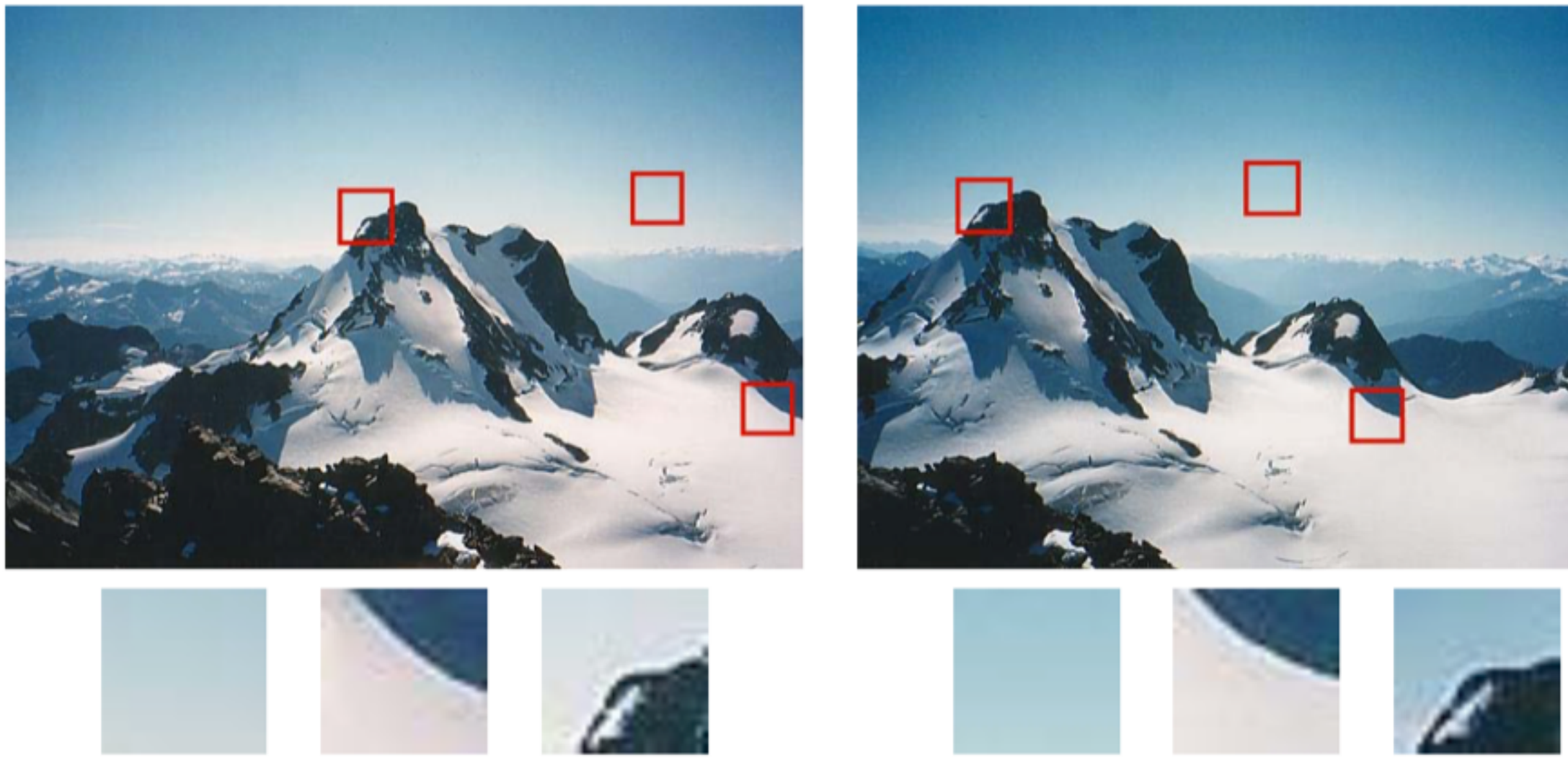
Identifying Keypoint Features: Peaks, Corners, Edges





Textured patches
Large contrast changes

Calculating Feature Descriptors



Regions around detected key points are converted to a compact, stable (invariant) representation: descriptor
e.g., SIFT, MOPS

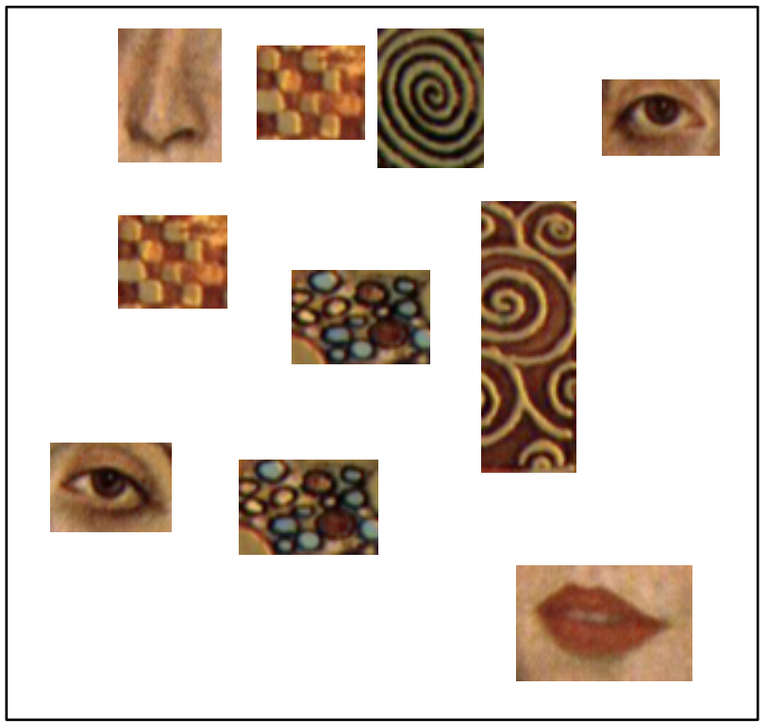
Bag-of-features models

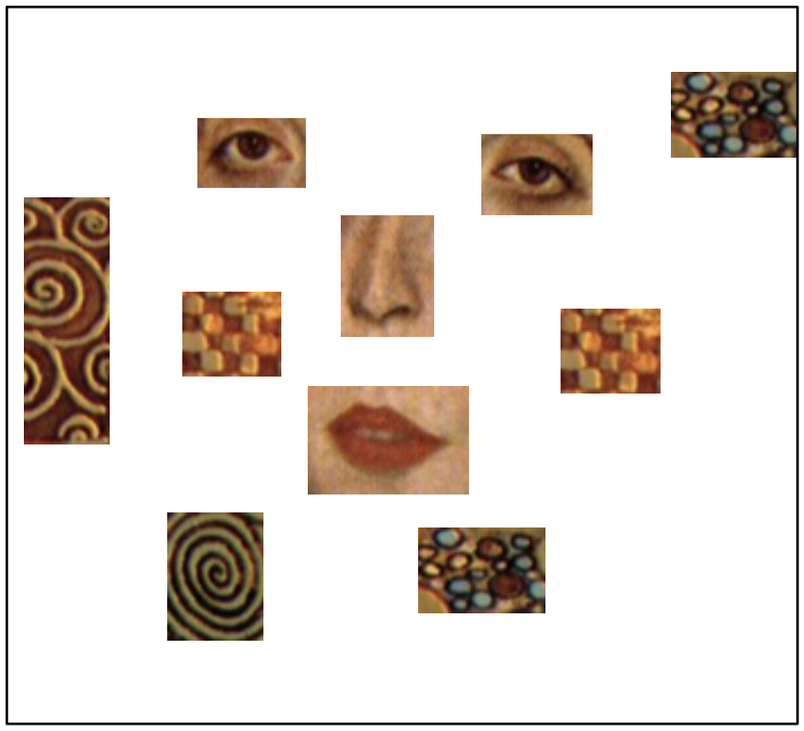
Bag-of-features models

Object → **Bag of ‘words’**

Objects as texture

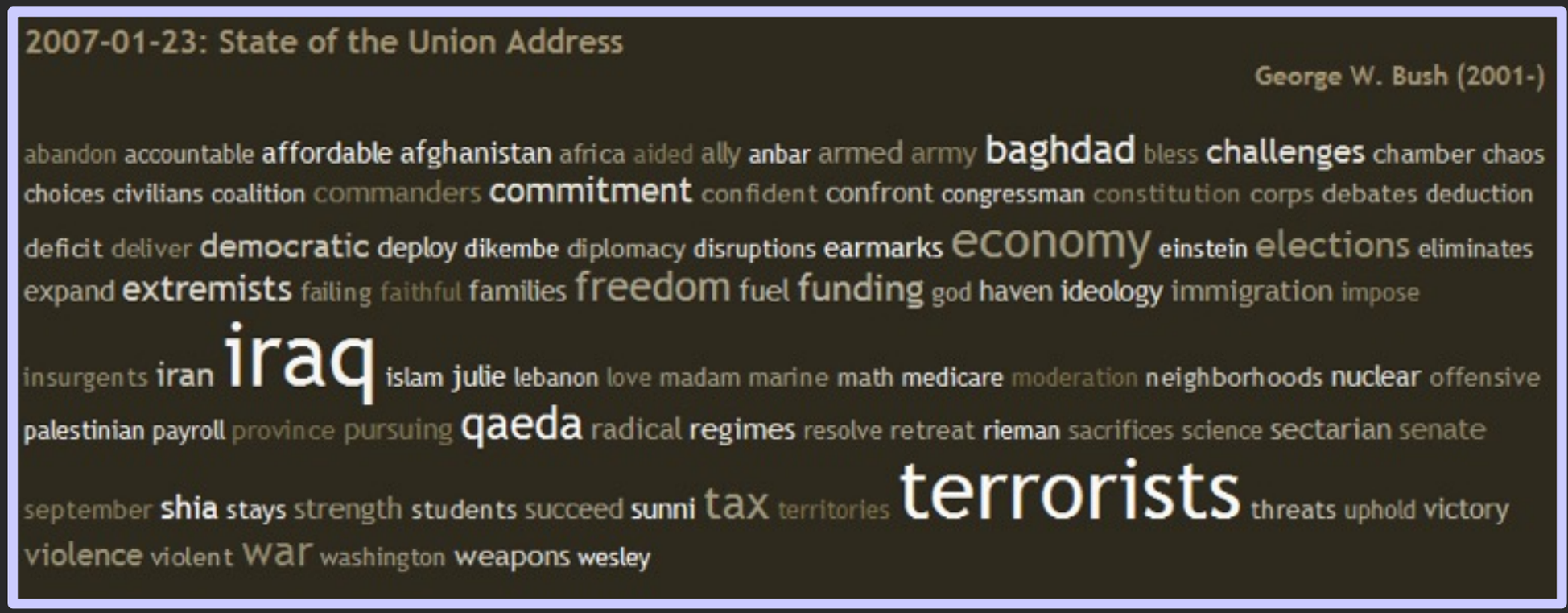
- All of these are treated as being the same





Origin 2: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary Salton & McGill (1983)



US Presidential Speeches Tag Cloud
http://chir.ag/phernalia/preztags/

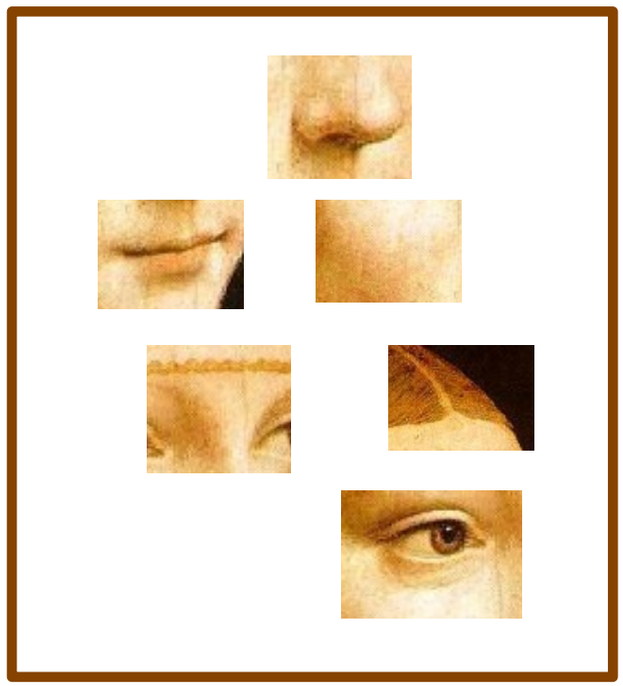
Bag of features

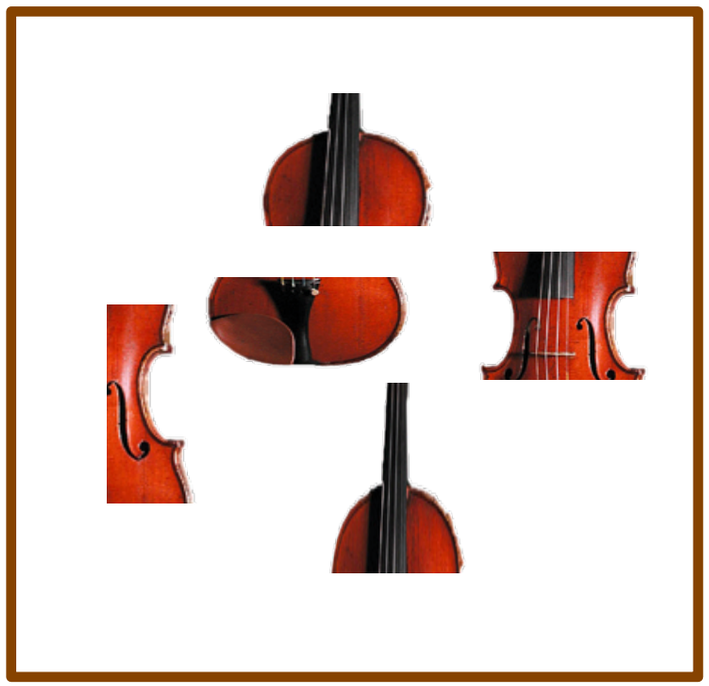
- First, take a bunch of images, extract features, and build up a “dictionary” or “visual vocabulary” – a list of common features

- Given a new image, extract features and build a histogram – for each feature, find the closest visual word in the dictionary

Bag of features: outline

1. Extract features





Bag of features: outline

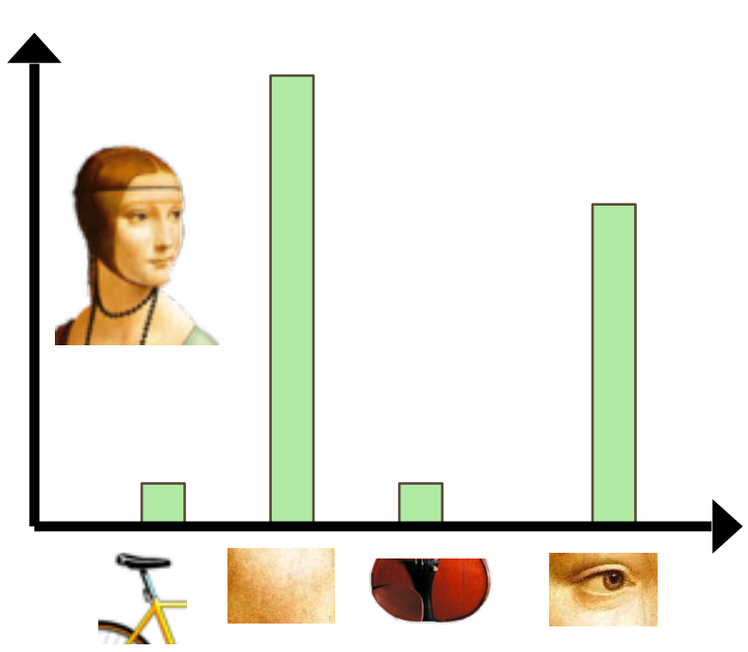
1. Extract features
2. Learn “visual vocabulary”

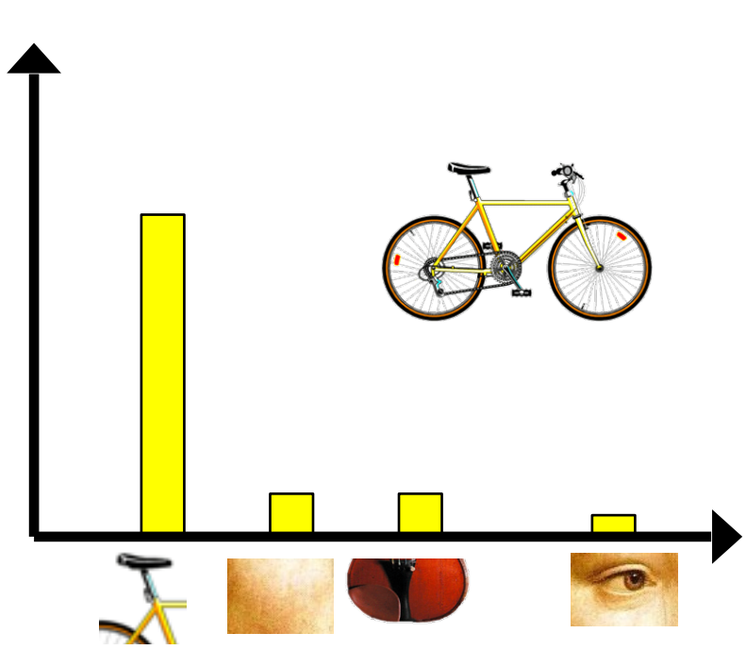
Bag of features: outline

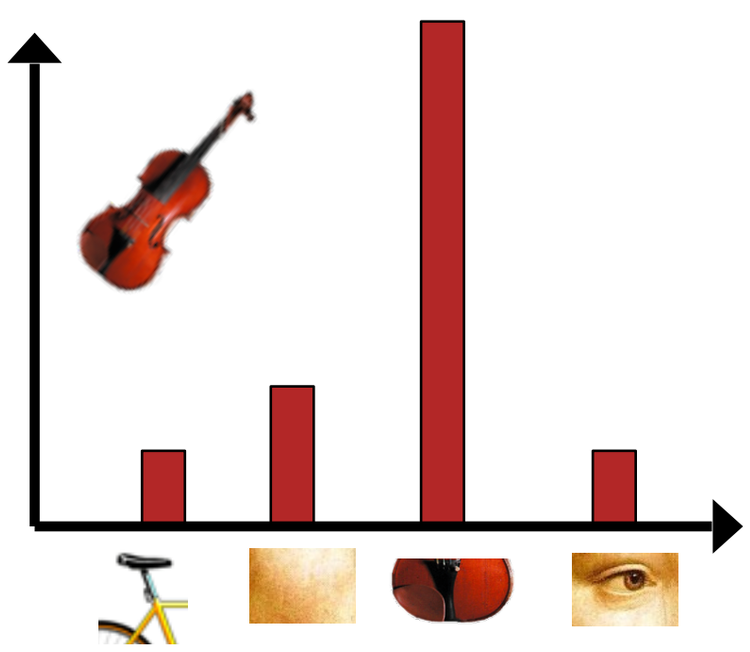
1. Extract features
2. Learn “visual vocabulary”
3. Quantize features using visual vocabulary

Bag of features: outline

1. Extract features
2. Learn “visual vocabulary”
3. Quantize features using visual vocabulary
4. Represent images by frequencies of “visual words”







1. Feature extraction

Regular grid

- Vogel & Schiele, 2003
- Fei-Fei & Perona, 2005

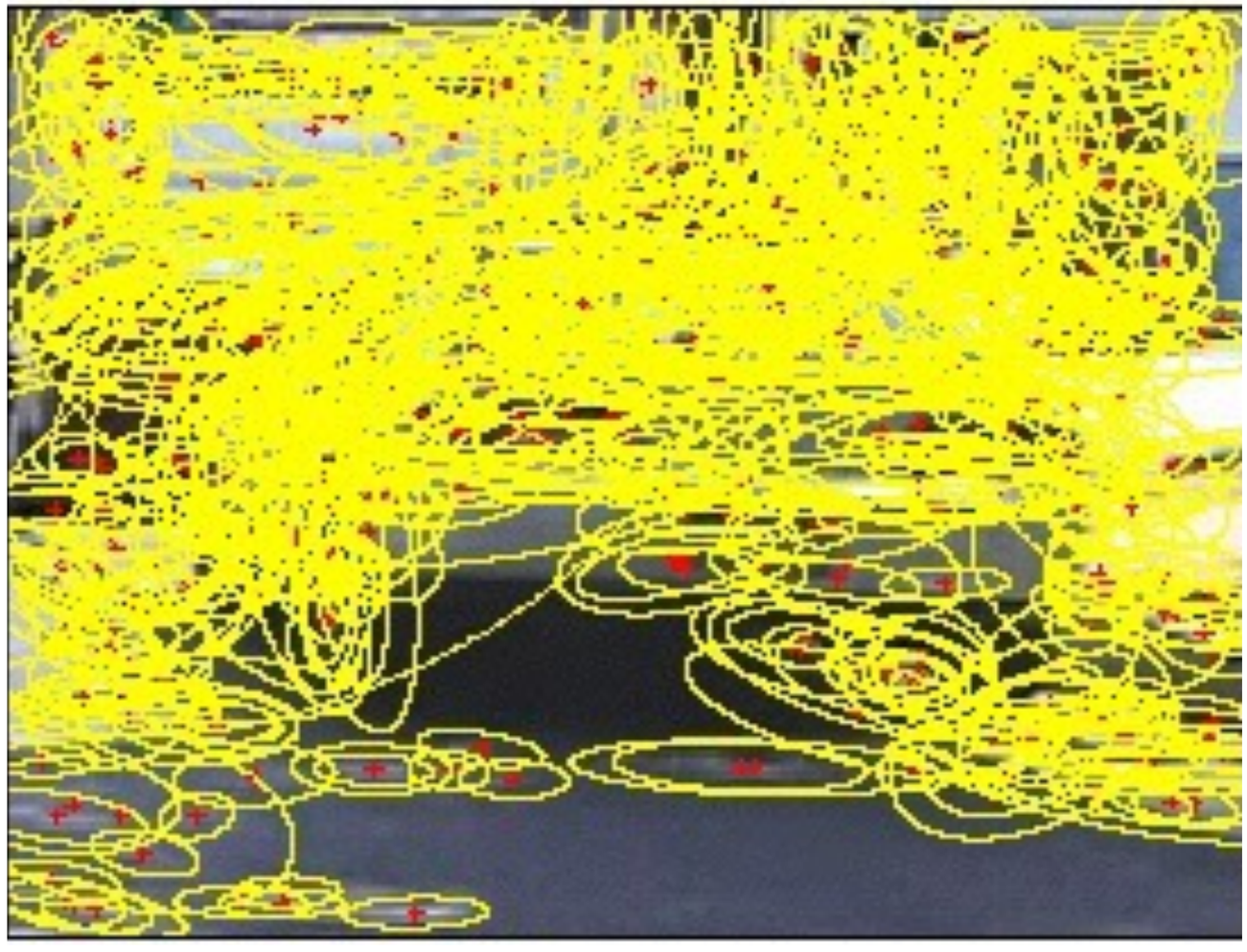
1. Feature extraction

Regular grid

- Vogel & Schiele, 2003
- Fei-Fei & Perona, 2005

Interest point detector

- Csurka et al. 2004
- Fei-Fei & Perona, 2005
- Sivic et al. 2005



1. Feature extraction

Regular grid

- Vogel & Schiele, 2003
- Fei-Fei & Perona, 2005

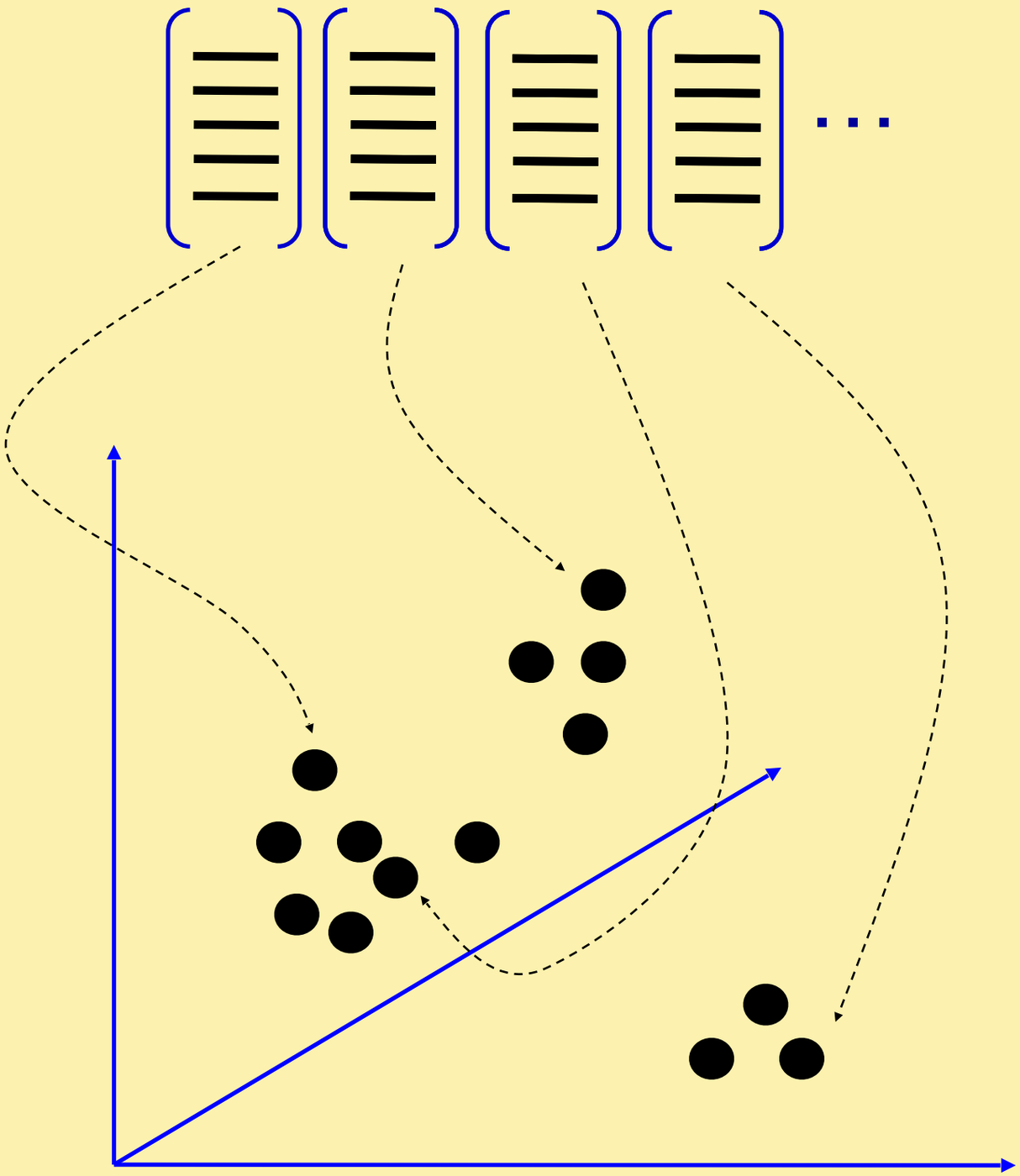
Interest point detector

- Csurka et al. 2004
- Fei-Fei & Perona, 2005
- Sivic et al. 2005

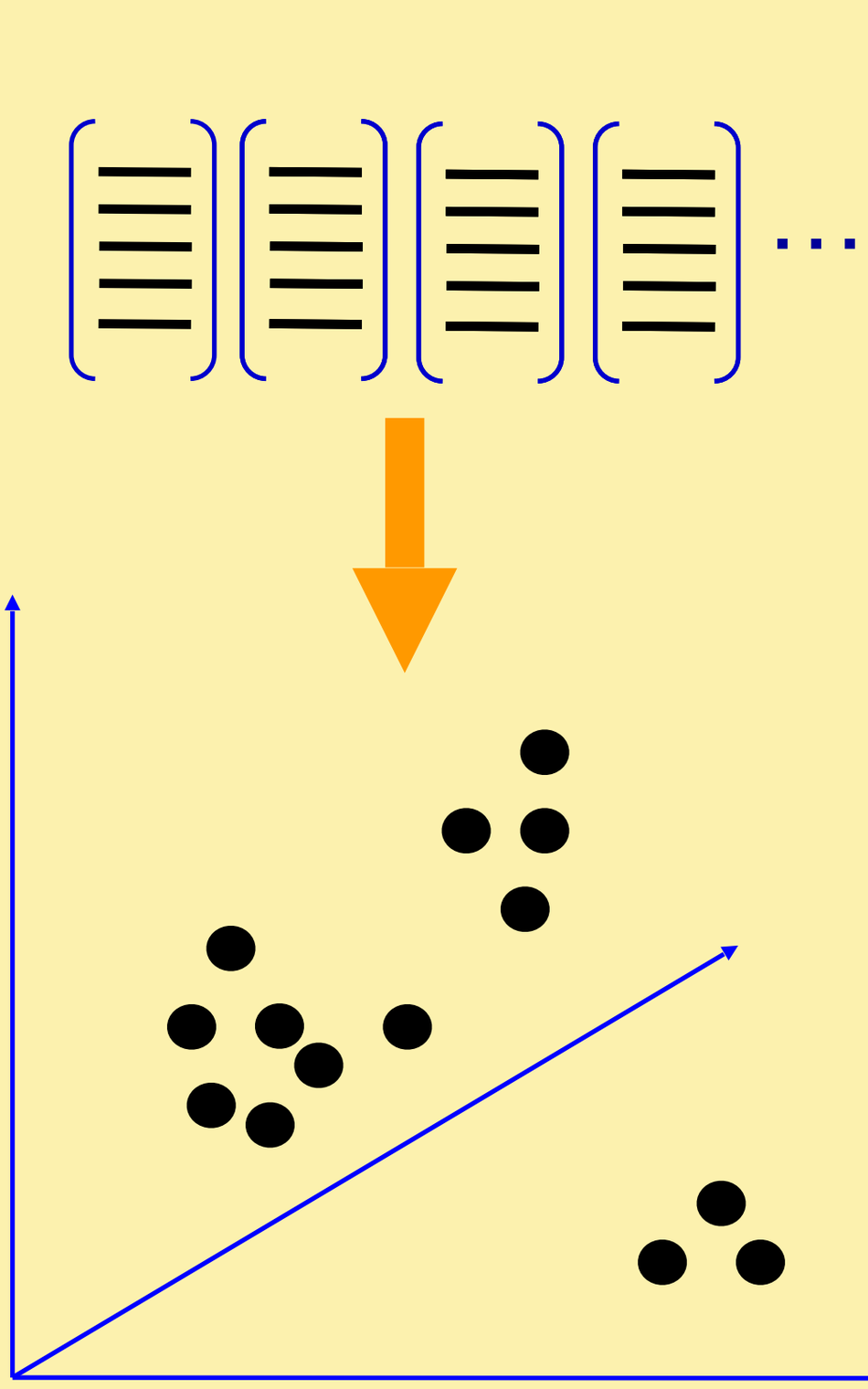
Other methods

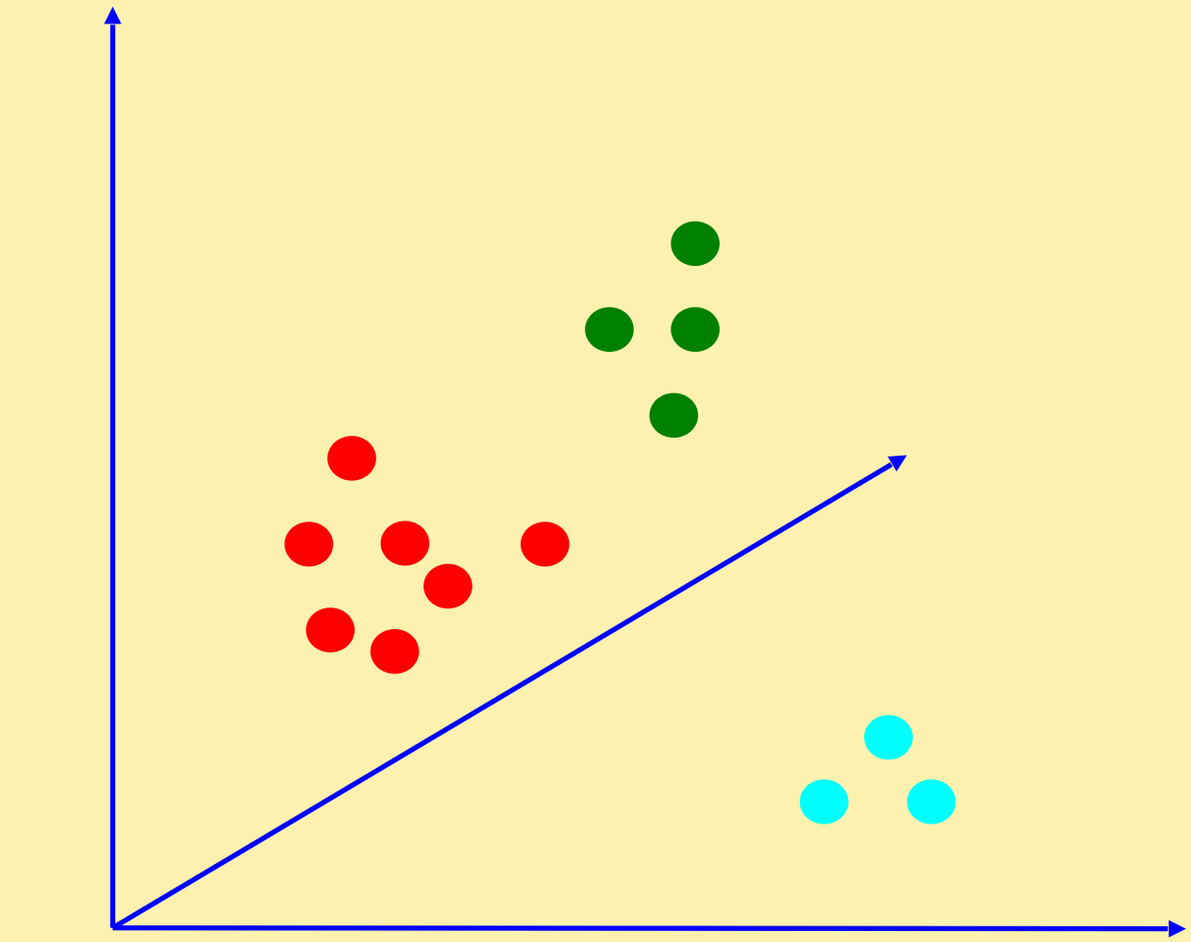
- Random sampling (Vidal-Naquet & Ullman, 2002)
- Segmentation-based patches (Barnard et al. 2003)

2. Learning the visual vocabulary



2. Learning the visual vocabulary

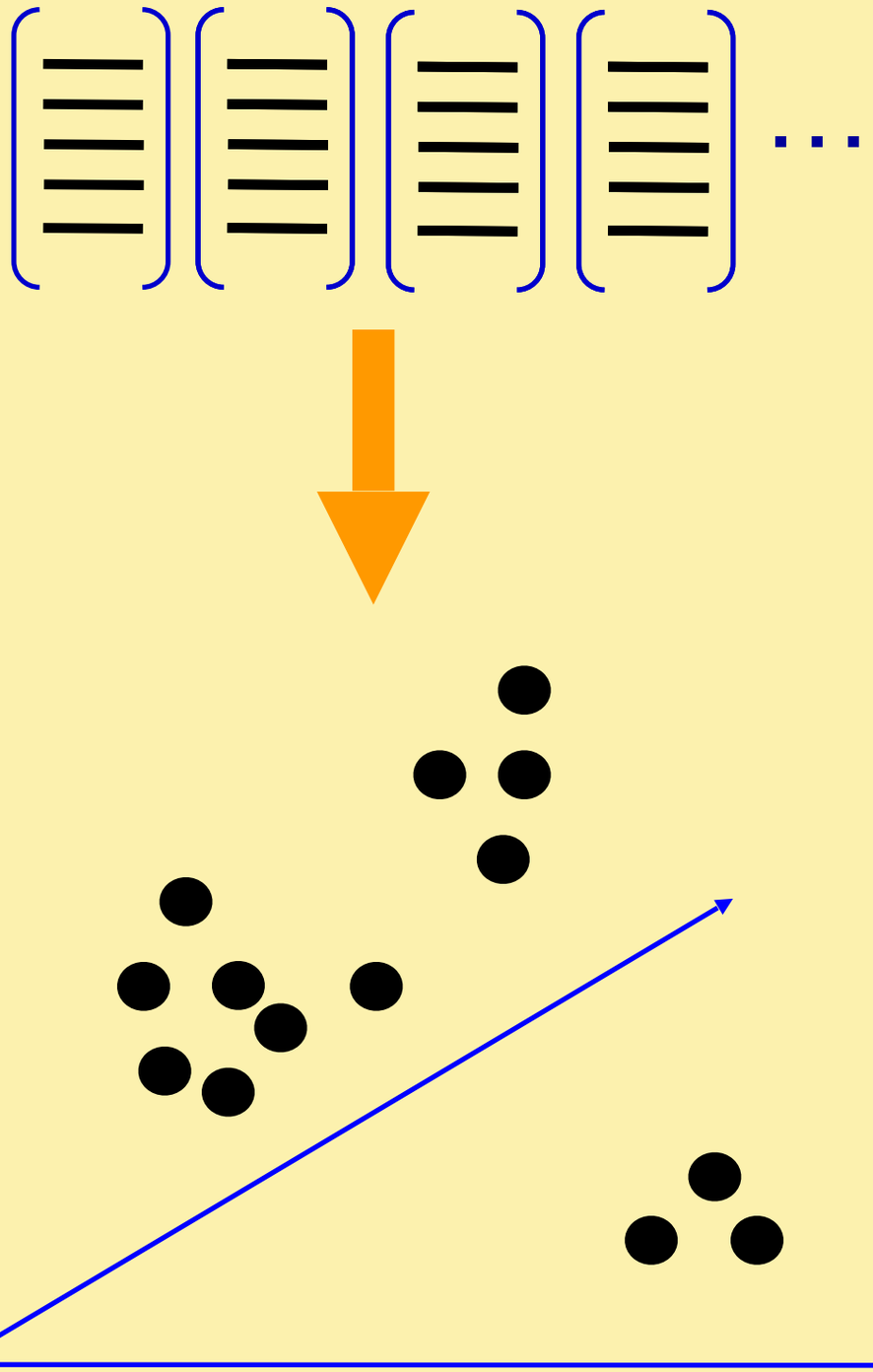


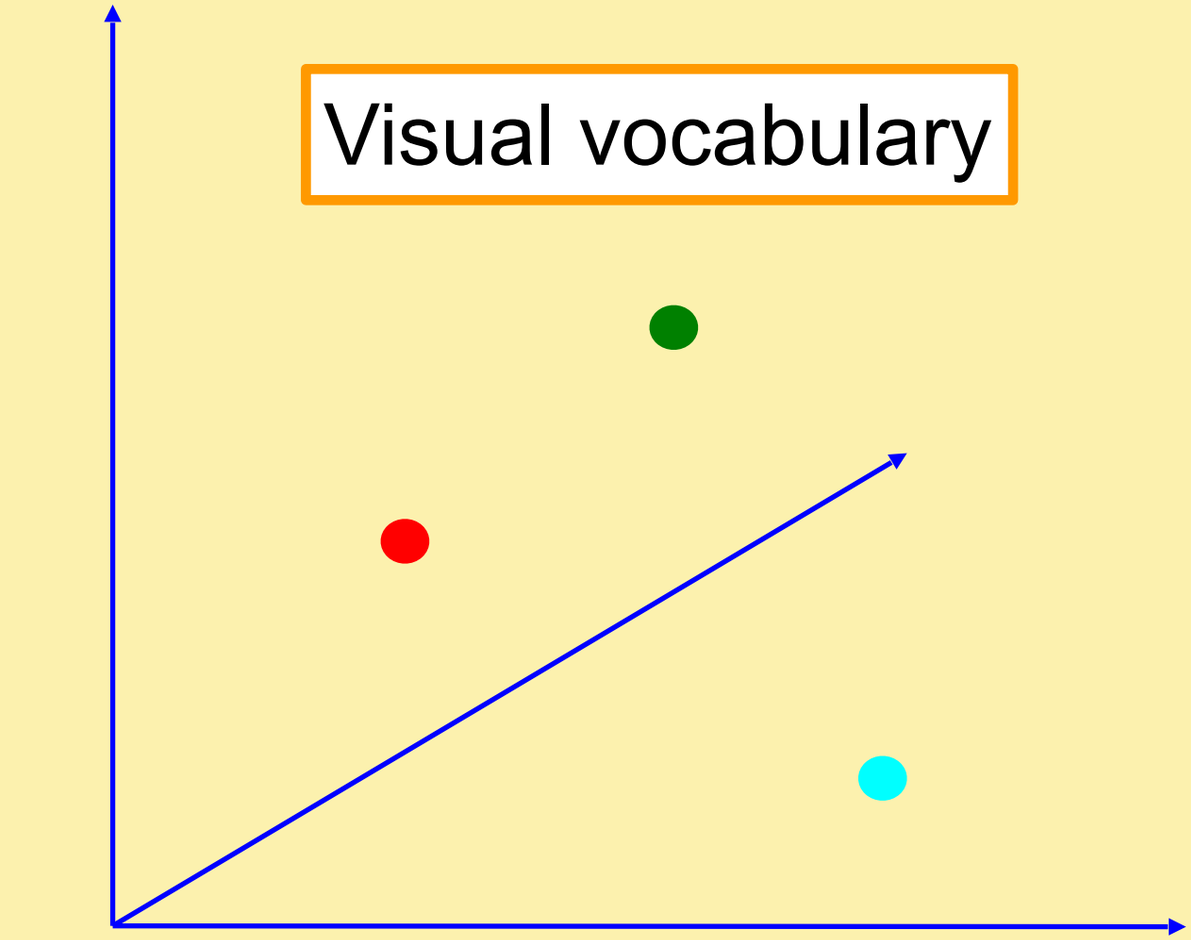


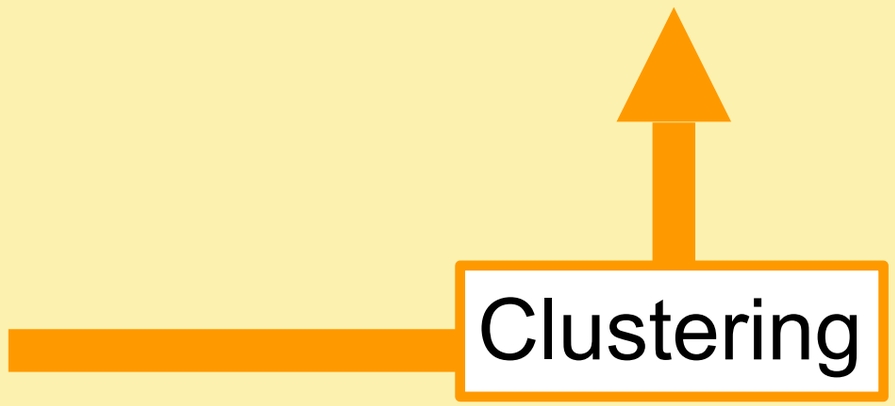
Clustering

Slide credit: Josef Sivic

2. Learning the visual vocabulary







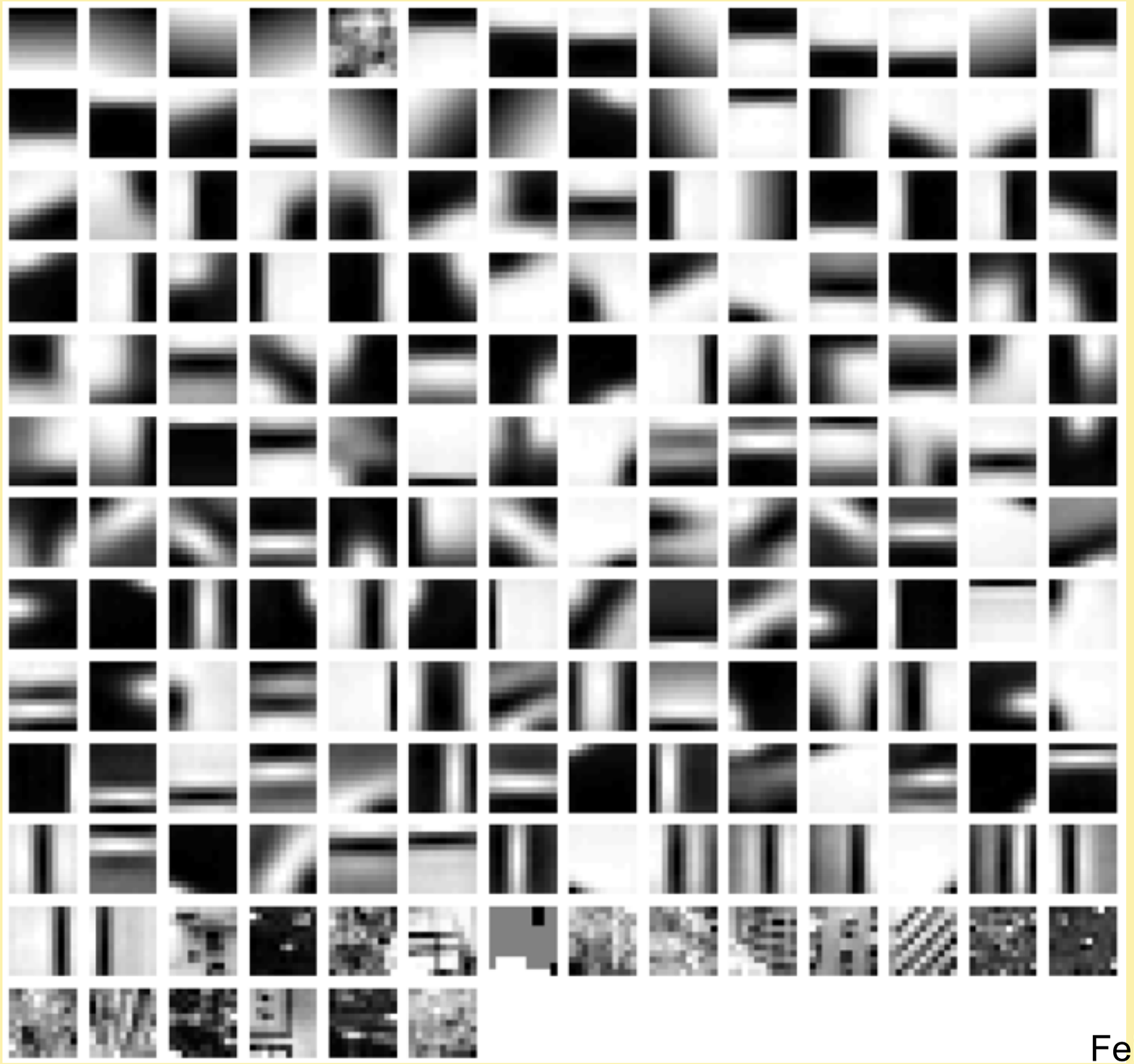
Slide credit: Josef Sivic

From clustering to vector quantization

- Clustering is a common method for learning a visual vocabulary or codebook
  - Unsupervised learning process
  - Each cluster center produced by k-means becomes a codevector
  - Codebook can be learned on separate training set
  - Provided the training set is sufficiently representative, the codebook will be “universal”

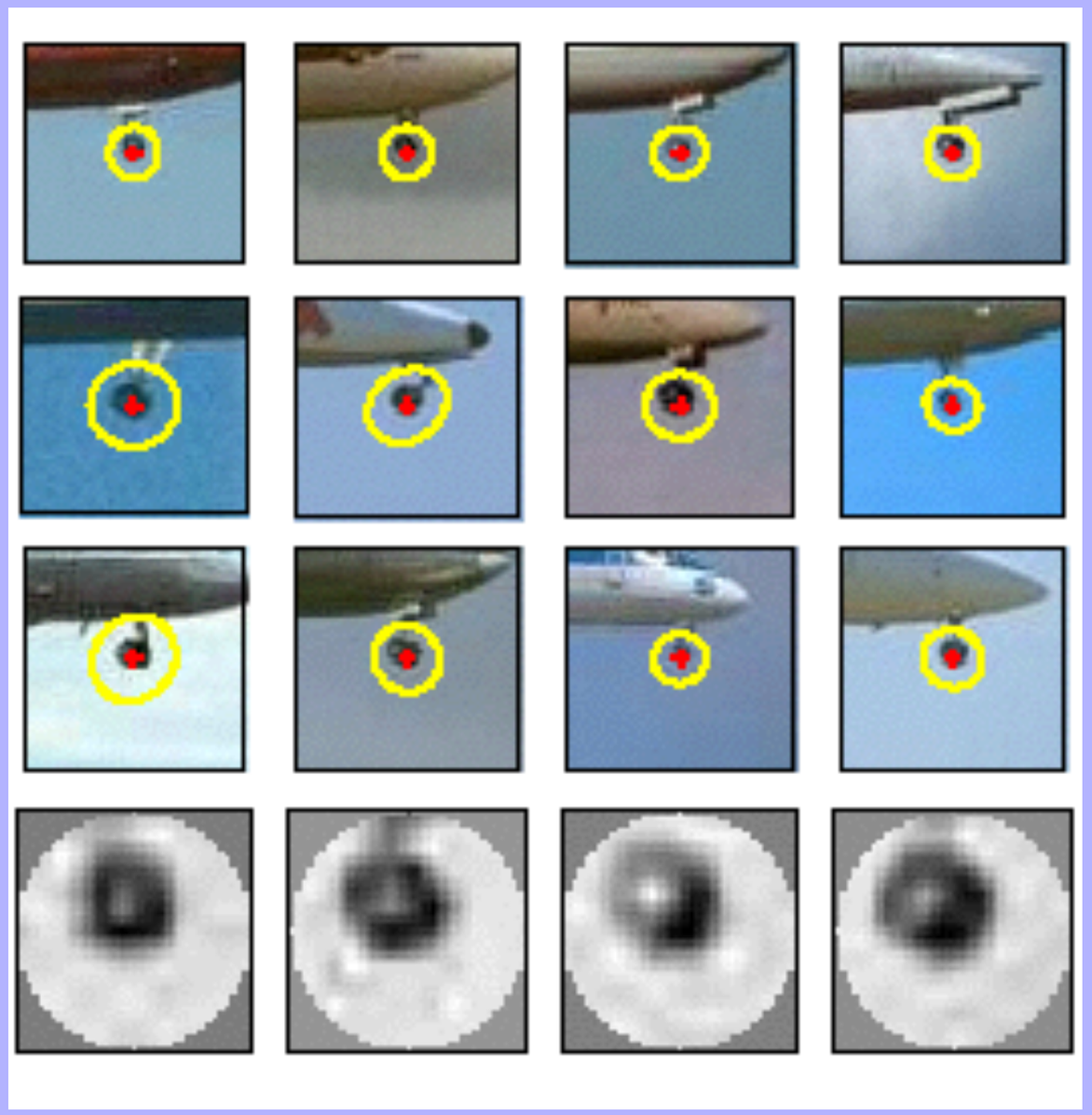
- The codebook is used for quantizing features
  - A *vector quantizer* takes a feature vector and maps it to the index of the nearest codevector in a codebook
  - Codebook = visual vocabulary
  - Codevector = visual word

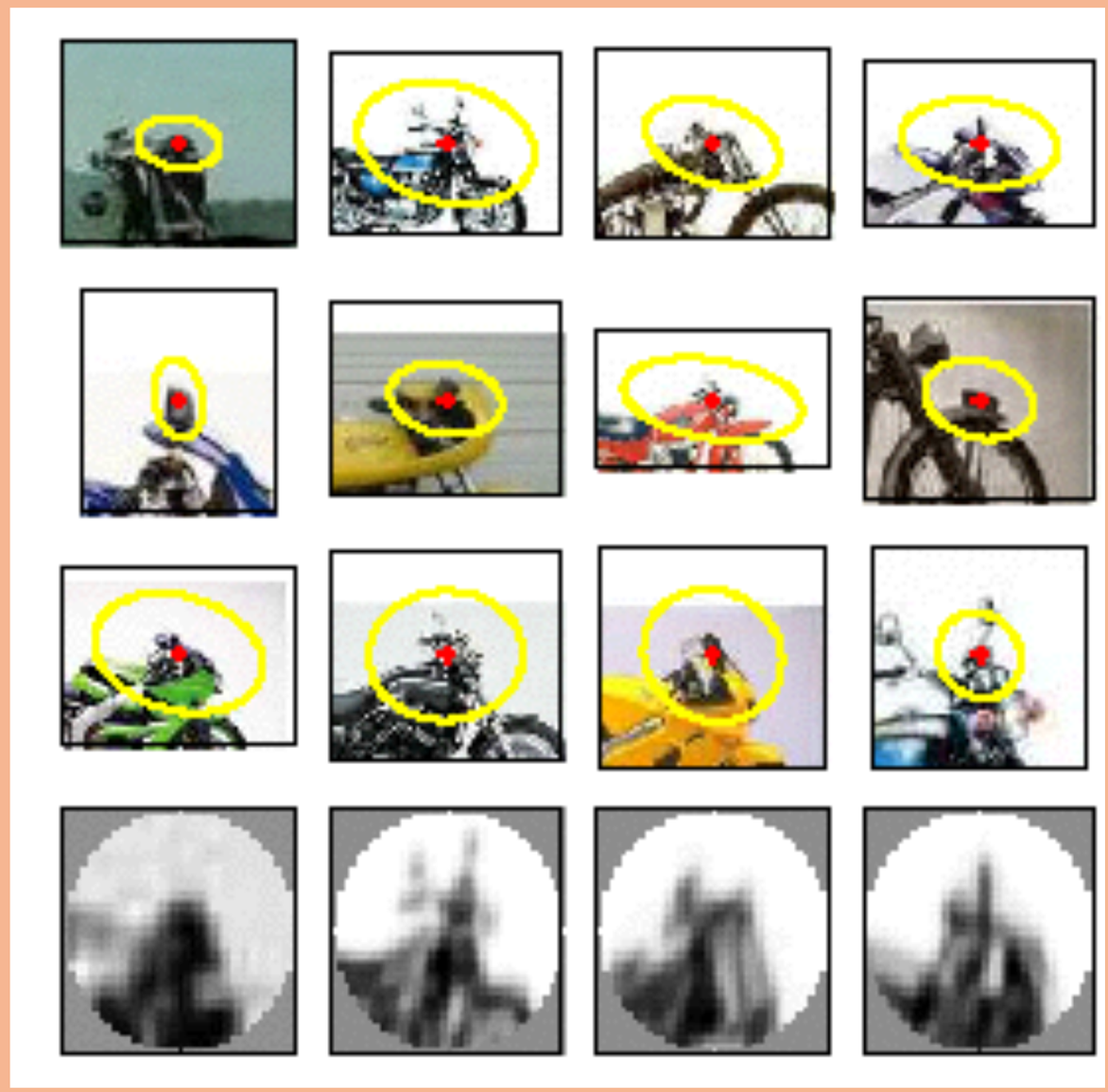
Example visual vocabulary



Fei-Fei et al. 2005

Image patch examples of visual words

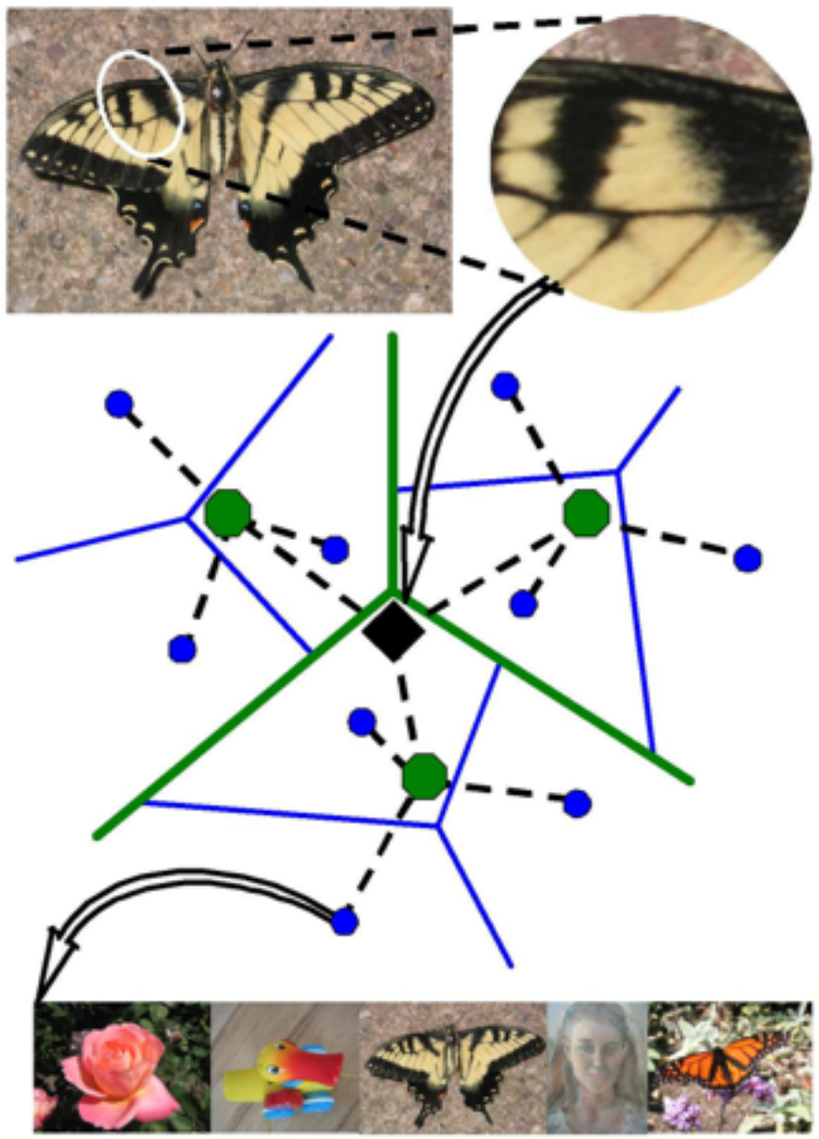




Sivic et al. 2005

Visual vocabularies: Issues

- How to choose vocabulary size?
  - Too small: visual words not representative of all patches
  - Too large: quantization artifacts, overfitting
- Computational efficiency
  - Vocabulary trees
    (Nister & Stewenius, 2006)



3. Image representation

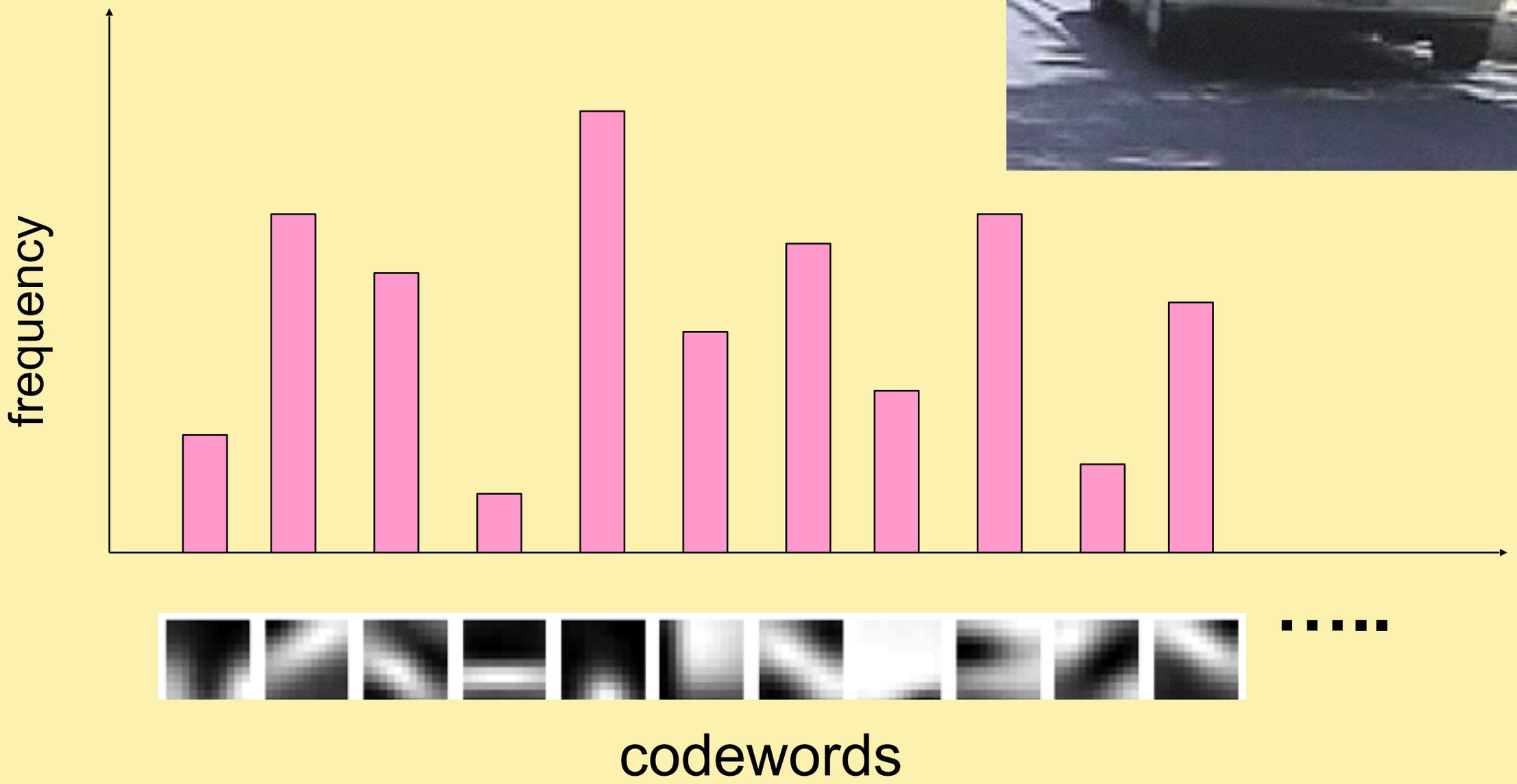
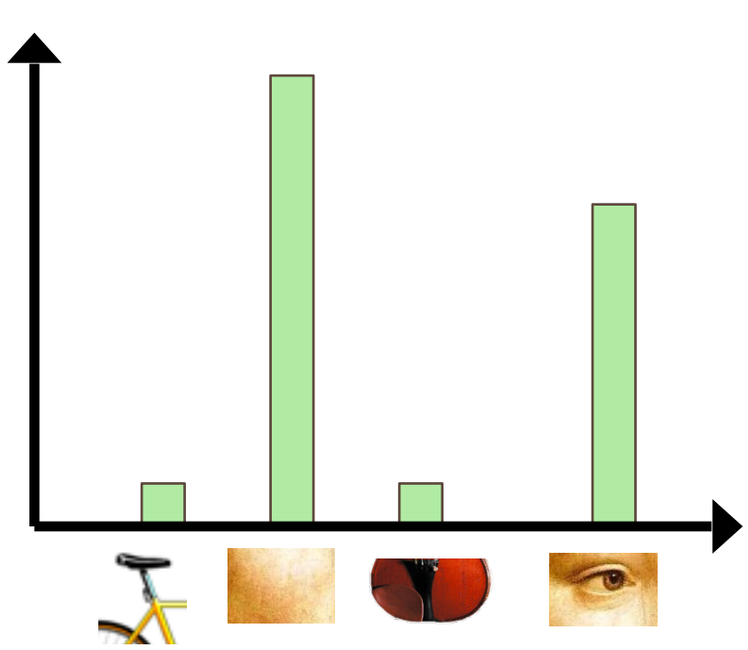
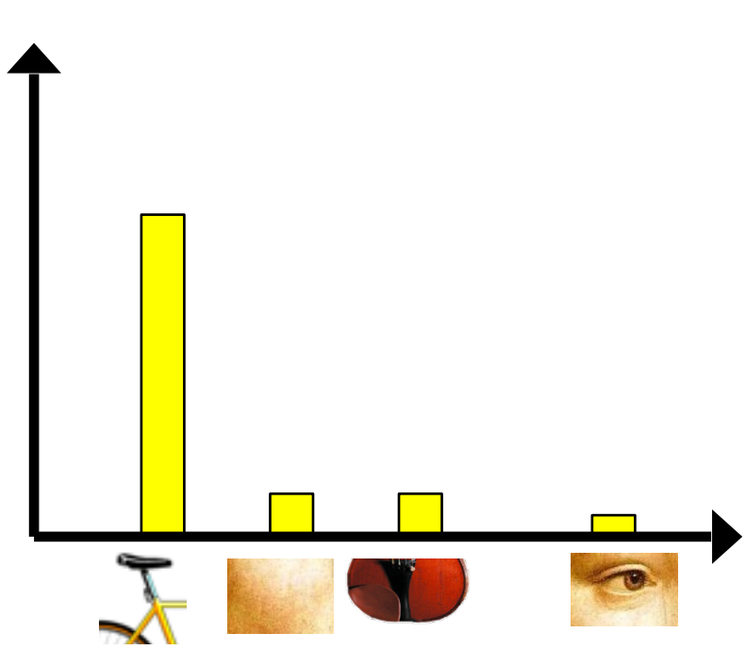
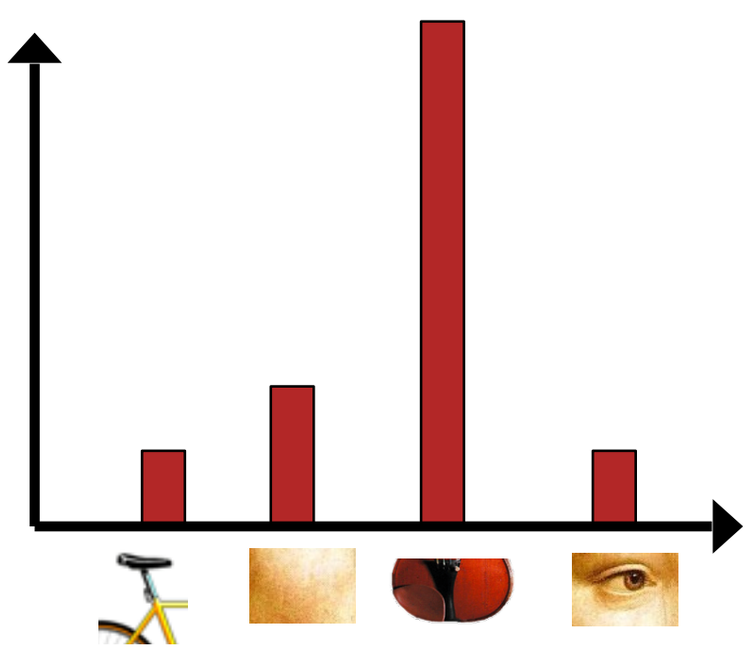


Image classification

- Given the bag-of-features representations of images from different classes, how do we learn a model for distinguishing them?







Weighting the words

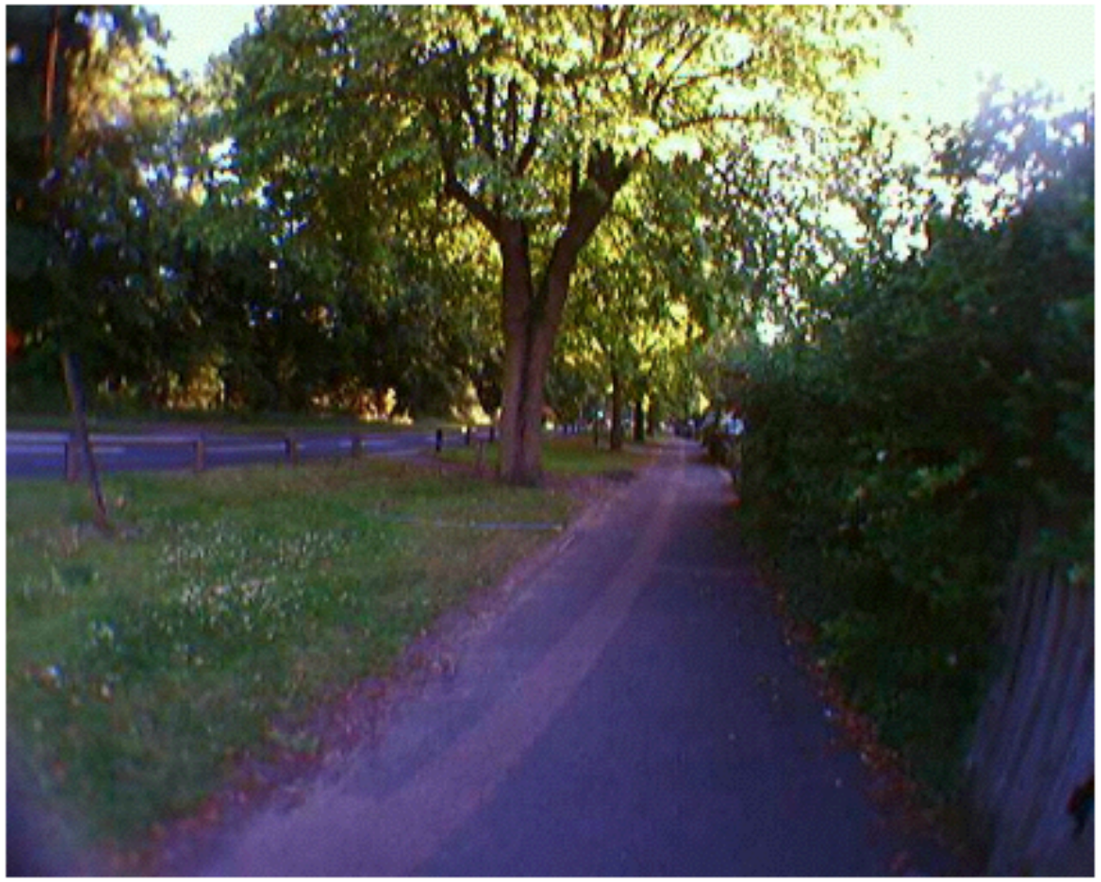
- Just as with text, some visual words are more discriminative than others

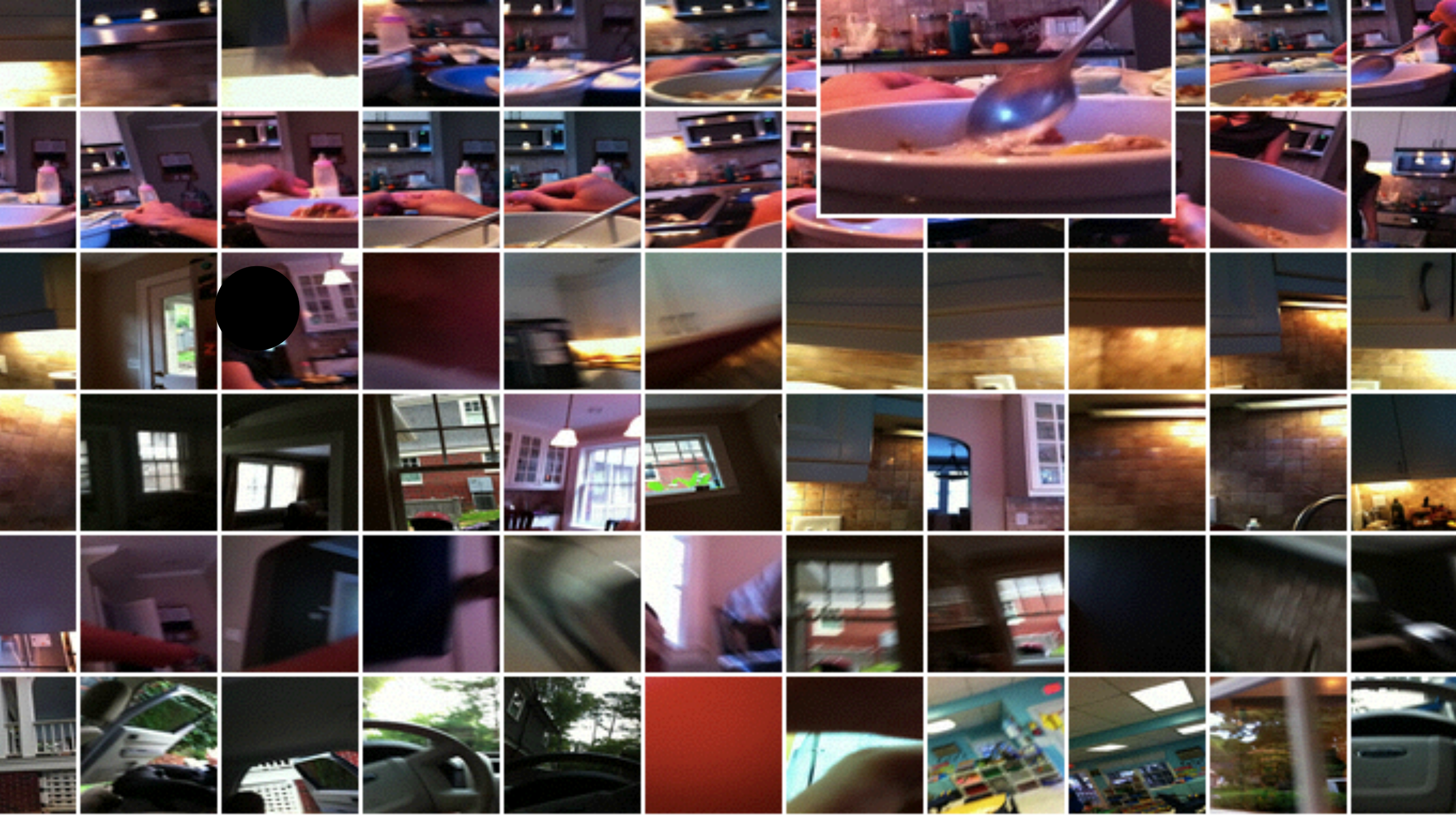
the, and, or vs. ***cow, AT&T, Cher***

- the bigger fraction of the documents a word appears in, the less useful it is for matching
  - e.g., a word that appears in *all* documents is not helping us

Paper

Can we devise a computational approach to automatically identify everyday activities with first-person images taken with wearable cameras *in-the-wild*?





iPhone as camera

6 months (26 Weeks)

Photo every 30s

40,103 Photos Total

Working

Family

Eating

TV

Reading

Meeting

Hygiene

Dogs

Driving

Socializing

Presentation

Cooking

Chores

Biking

Cleaning

Shopping

Exercising

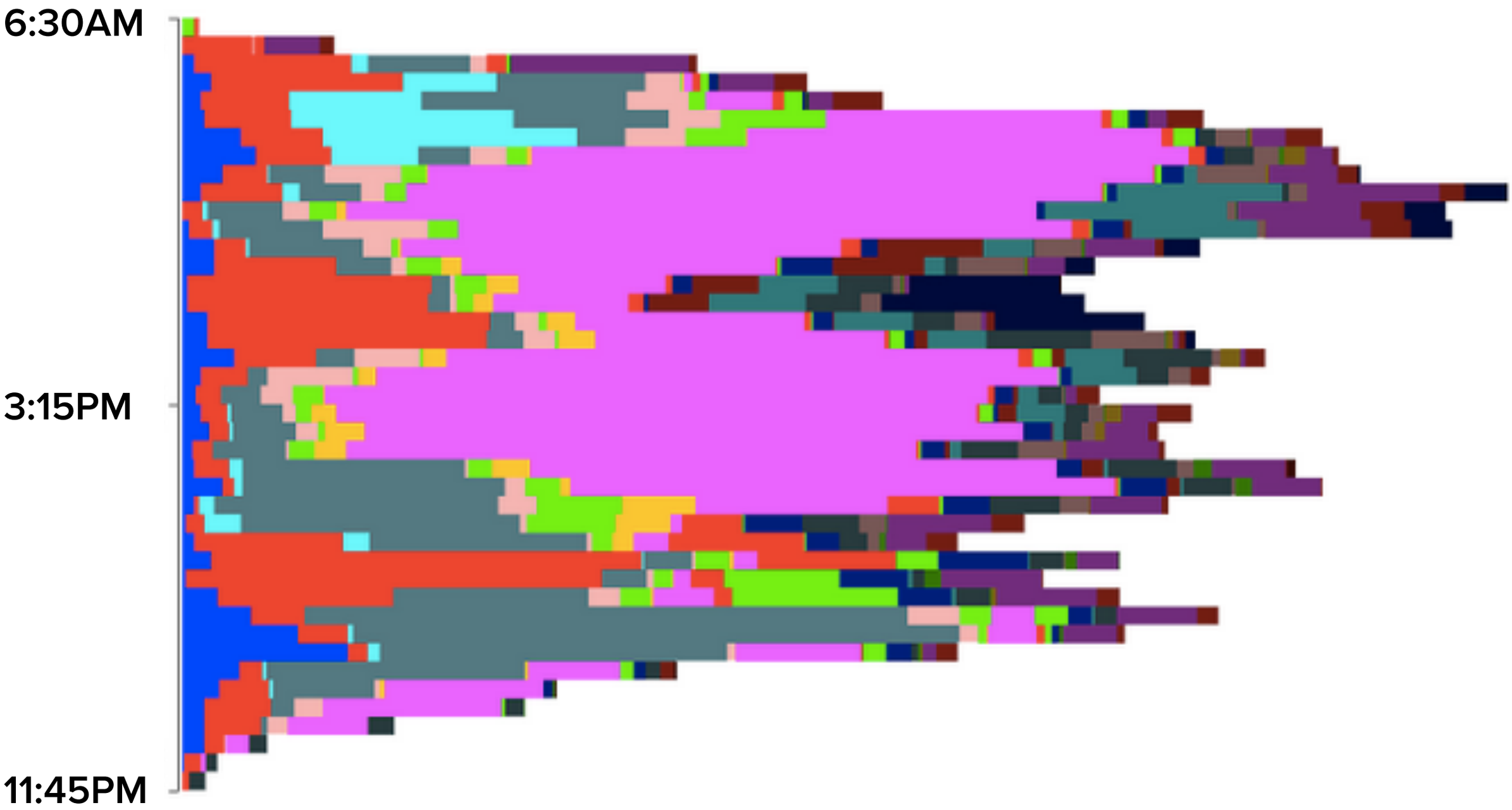
Chatting

Resting

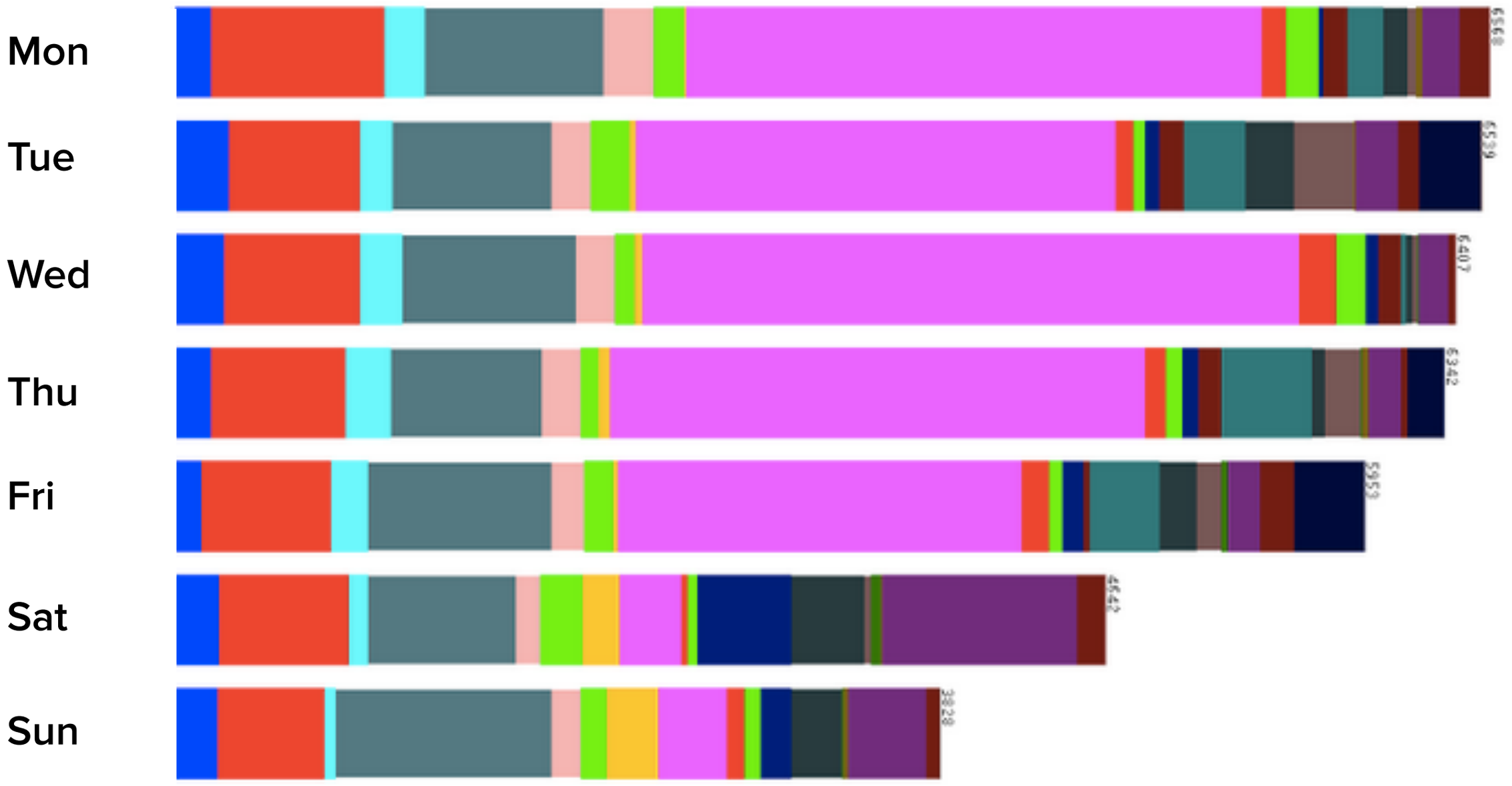
	# Images	% Dataset
Working	13895	34.24
Family	8267	20.37
Eating	4699	11.58
TV	1584	3.90
Reading	1414	3.48
Meeting	1312	3.23
Hygiene	1266	3.12
Dogs	1149	2.83
Driving	1031	2.54
Socializing	970	2.39
Presentation	848	2.09
Cooking	759	1.87
Chores	725	1.79
Biking	696	1.71
Cleaning	642	1.59
Shopping	606	1.49
Exercising	502	1.24
Chatting	113	0.28
Resting	106	0.26

	# Images	% Dataset
Working	**13895**	**34.24**
Family	**8267**	**20.37**
Eating	**4699**	**11.58**
TV	1584	3.90
Reading	1414	3.48
Meeting	1312	3.23
Hygiene	1266	3.12
Dogs	1149	2.83
Driving	1031	2.54
Socializing	970	2.39
Presentation	848	2.09
Cooking	759	1.87
Chores	725	1.79
Biking	696	1.71
Cleaning	642	1.59
Shopping	606	1.49
Exercising	502	1.24
Chatting	113	0.28
Resting	106	0.26

Activities Distribution: Day



Activities Distribution: Week



Convolutional Neural Network (CNN)

Good results in image recognition

Trained with ImageNet dataset (1 million images)

Last layer of CNN trained with collected images

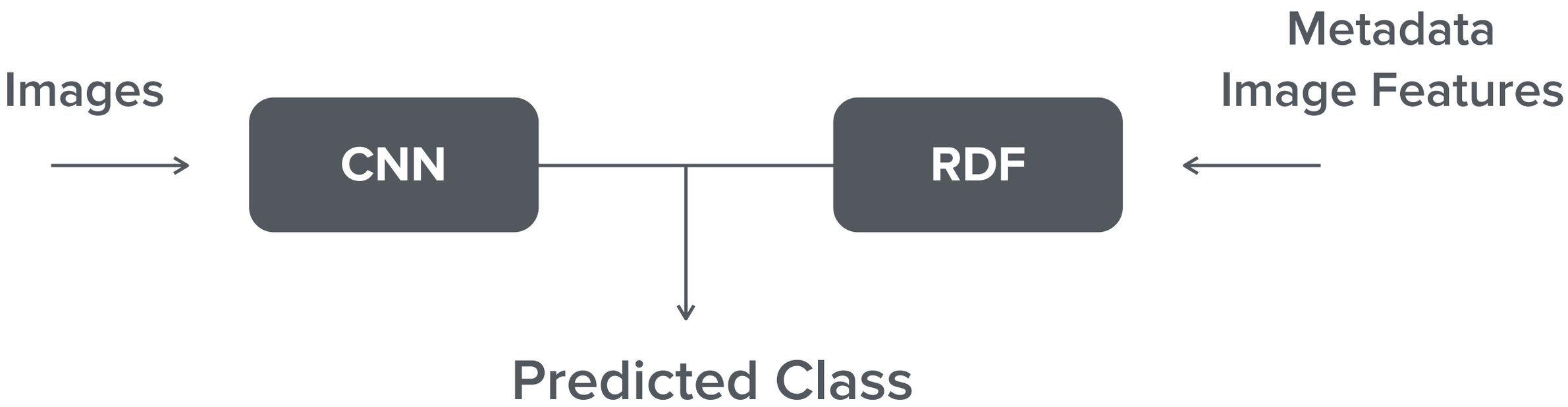
Additional Features: day of week, hour of day

Classic Ensemble

Late Fusion Ensemble

Krizhevsky, A. *et al.*. Imagenet classification with deep convolutional neural networks. NIPS 2012.

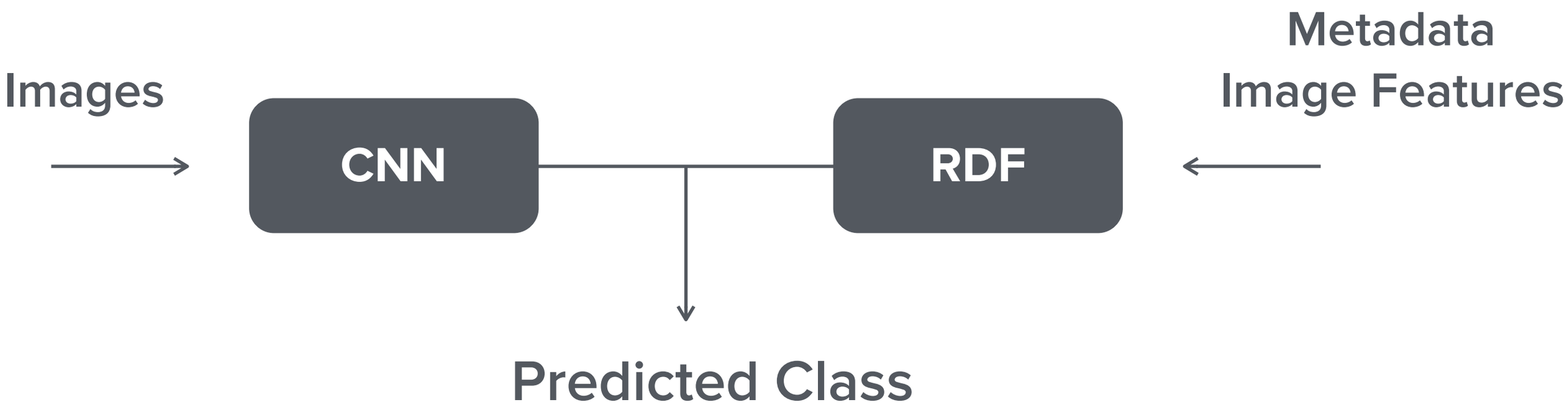
Classic Ensemble



78.56%

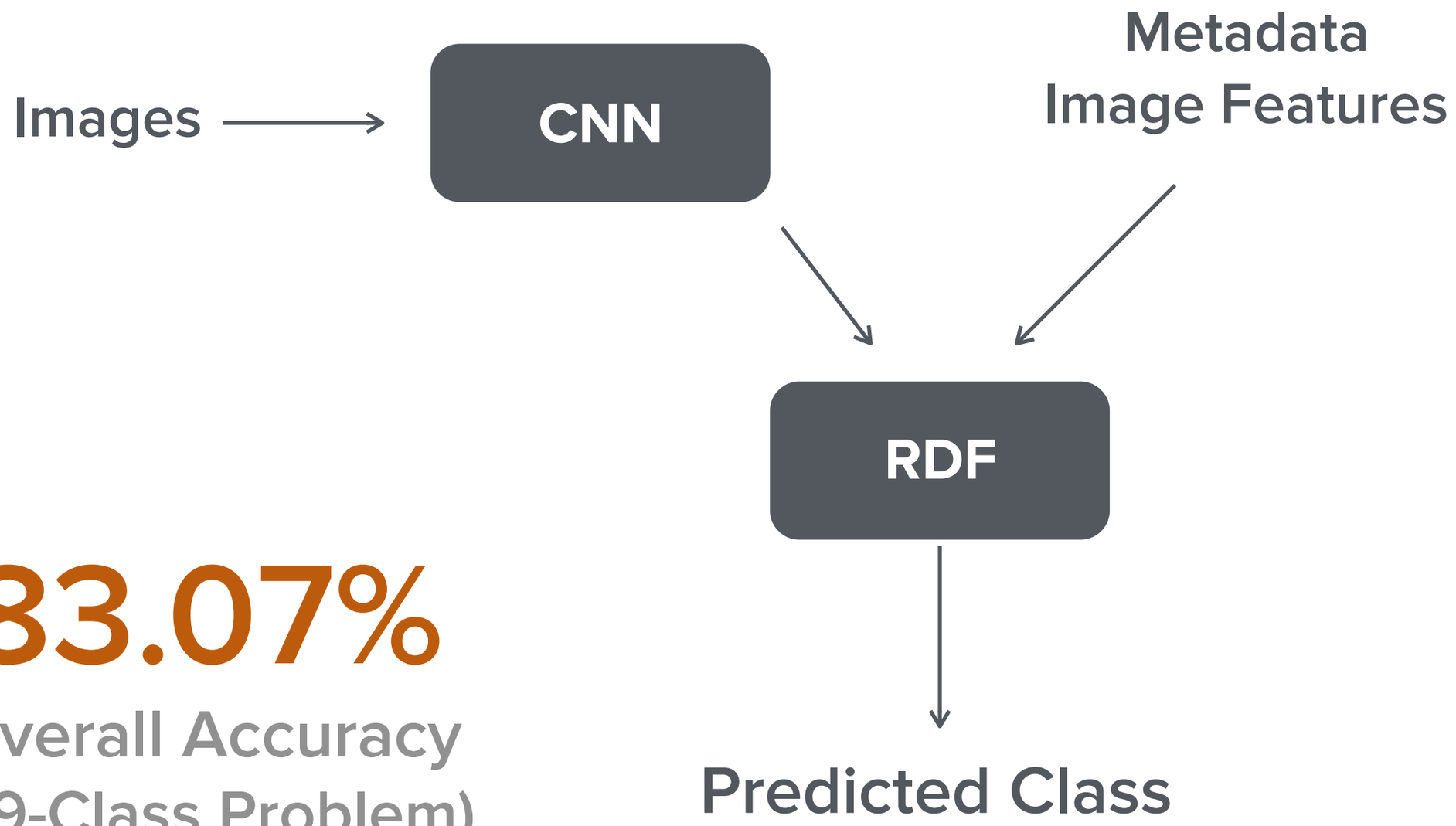
Overall Accuracy
(19-Class Problem)

Classic Ensemble



Not Learning All Feature Relationships

Late Fusion Ensemble



83.07%

Overall Accuracy
(19-Class Problem)

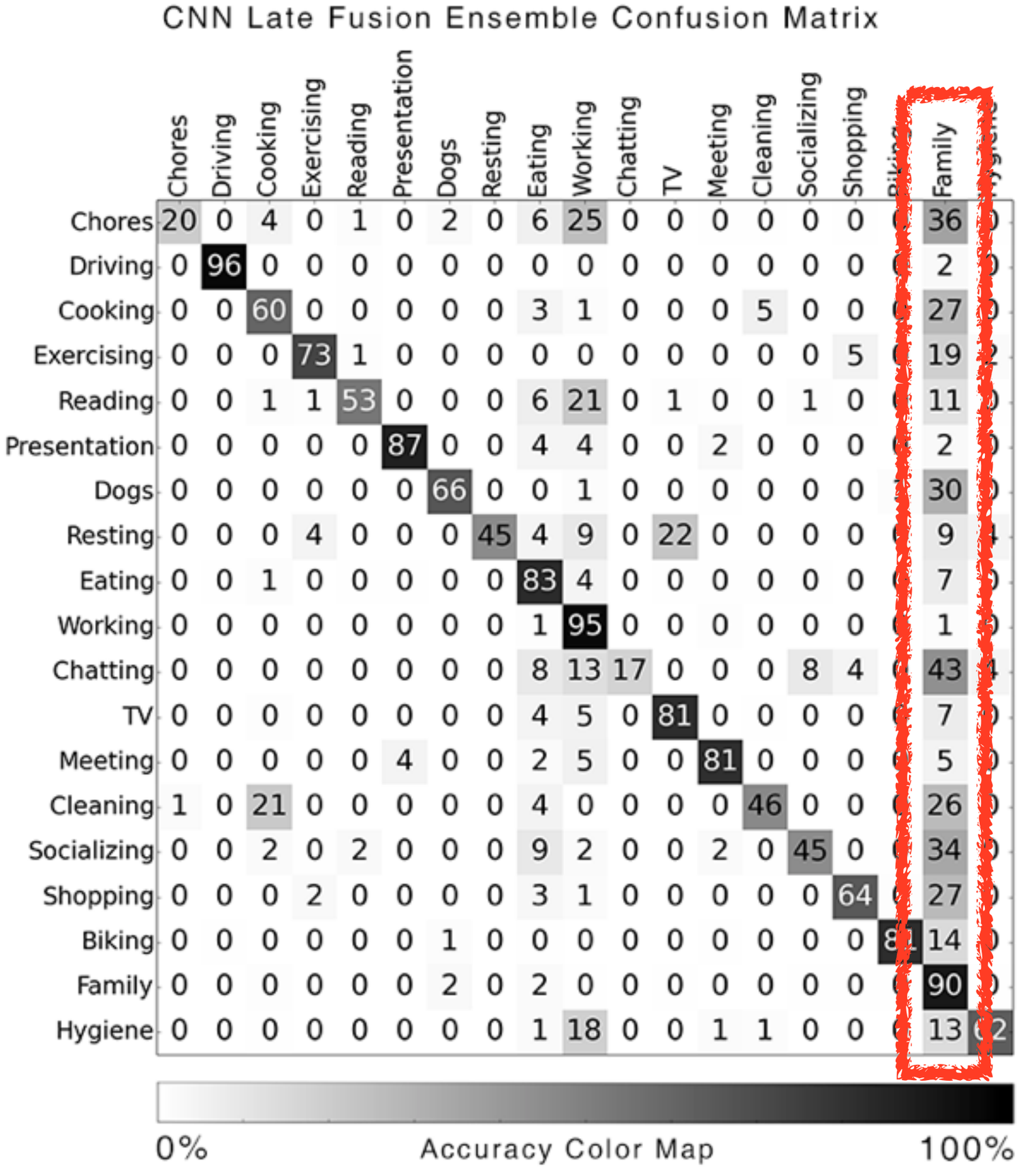
CNN Late Fusion Ensemble Confusion Matrix

	Chores	Driving	Cooking	Exercising	Reading	Presentation	Dogs	Resting	Eating	Working	Chatting	TV	Meeting	Cleaning	Socializing	Shopping	Biking	Family	Hygiene
Chores	20	0	4	0	1	0	2	0	6	25	0	0	0	0	0	0	0	36	0
Driving	0	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
Cooking	0	0	60	0	0	0	0	0	3	1	0	0	0	5	0	0	0	27	0
Exercising	0	0	0	73	1	0	0	0	0	0	0	0	0	0	0	5	0	19	2
Reading	0	0	1	1	53	0	0	0	6	21	0	1	0	0	1	0	0	11	0
Presentation	0	0	0	0	0	87	0	0	4	4	0	0	2	0	0	0	0	2	0
Dogs	0	0	0	0	0	0	66	0	0	1	0	0	0	0	0	0	1	30	0
Resting	0	0	0	4	0	0	0	45	4	9	0	22	0	0	0	0	0	9	4
Eating	0	0	1	0	0	0	0	0	83	4	0	0	0	0	0	0	0	7	0
Working	0	0	0	0	0	0	0	0	1	95	0	0	0	0	0	0	0	1	0
Chatting	0	0	0	0	0	0	0	0	8	13	17	0	0	0	8	4	0	43	4
TV	0	0	0	0	0	0	0	0	4	5	0	81	0	0	0	0	0	7	0
Meeting	0	0	0	0	0	4	0	0	2	5	0	0	81	0	0	0	0	5	0
Cleaning	1	0	21	0	0	0	0	0	4	0	0	0	0	46	0	0	0	26	0
Socializing	0	0	2	0	2	0	0	0	9	2	0	0	2	0	45	0	0	34	0
Shopping	0	0	0	2	0	0	0	0	3	1	0	0	0	0	0	64	0	27	0
Biking	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	81	14	0
Family	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0	90	0
Hygiene	0	0	0	0	0	0	0	0	1	18	0	0	1	1	0	0	0	13	62

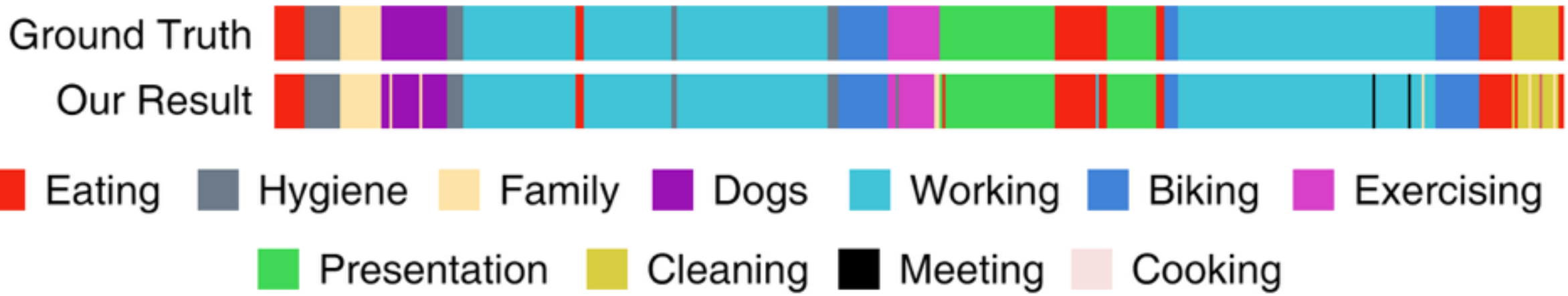
CNN Late Fusion Ensemble Confusion Matrix

	Chores	Driving	Cooking	Exercising	Reading	Presentation	Dogs	Resting	Eating	Working	Chatting	TV	Meeting	Cleaning	Socializing	Shopping	Biking	Family	Hygiene
Chores	20	0	4	0	1	0	2	0	6	25	0	0	0	0	0	0	0	36	0
Driving	0	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
Cooking	0	0	30	0	0	0	0	0	3	1	0	0	0	5	0	0	0	27	0
Exercising	0	0	0	73	1	0	0	0	0	0	0	0	0	0	0	5	0	19	2
Reading	0	0	1	1	53	0	0	0	6	21	0	1	0	0	1	0	0	11	0
Presentation	0	0	0	0	0	87	0	0	4	4	0	0	2	0	0	0	0	2	0
Dogs	0	0	0	0	0	0	66	0	0	1	0	0	0	0	0	0	1	30	0
Resting	0	0	0	4	0	0	0	45	4	9	0	22	0	0	0	0	0	9	4
Eating	0	0	1	0	0	0	0	0	83	4	0	0	0	0	0	0	0	7	0
Working	0	0	0	0	0	0	0	0	1	95	0	0	0	0	0	0	0	1	0
Chatting	0	0	0	0	0	0	0	0	8	13	17	0	0	0	8	4	0	43	4
TV	0	0	0	0	0	0	0	0	4	5	0	81	0	0	0	0	0	7	0
Meeting	0	0	0	0	0	4	0	0	2	5	0	0	81	0	0	0	0	5	0
Cleaning	1	0	21	0	0	0	0	0	4	0	0	0	0	46	0	0	0	26	0
Socializing	0	0	2	0	2	0	0	0	9	2	0	0	2	0	45	0	0	34	0
Shopping	0	0	0	2	0	0	0	0	3	1	0	0	0	0	0	64	0	27	0
Biking	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	81	14	0
Family	0	0	0	0	0	0	2	0	2	0	0	0	0	0	0	0	0	90	0
Hygiene	0	0	0	0	0	0	0	0	1	18	0	0	1	1	0	0	0	13	62

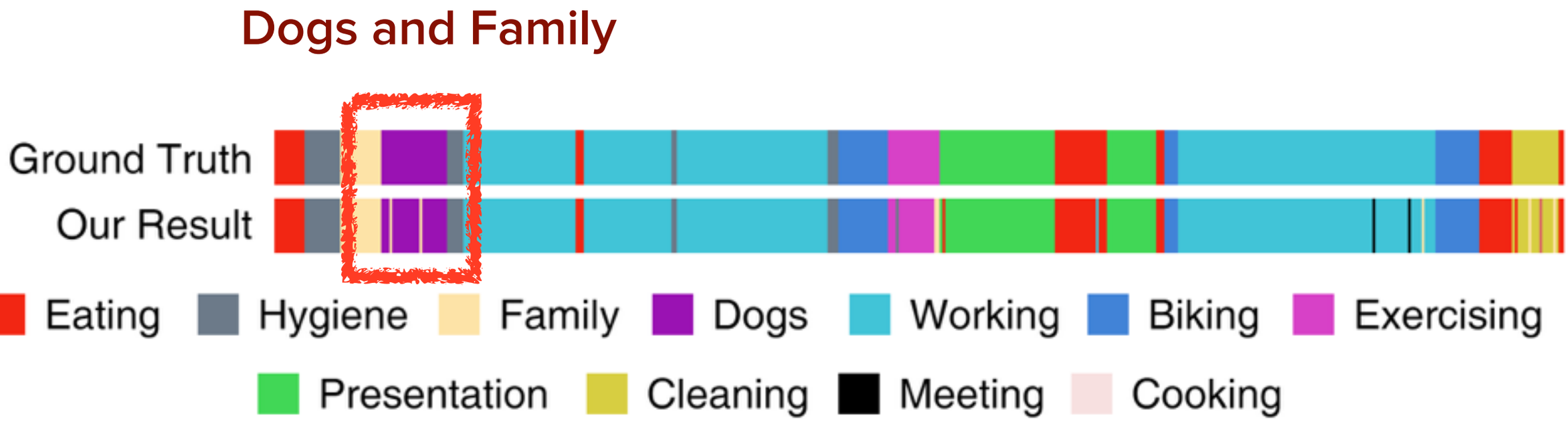
0% Accuracy Color Map 100%


CNN Late Fusion Ensemble Confusion Matrix
Chores
Driving
Cooking
Exercising
Reading
Presentation
Dogs
Resting
Eating
Working
Chatting
TV
Meeting
Cleaning
Socializing
Shopping
Biking
Family
Hygiene
Chores 20 0 4 0 1 0 2 0 6 25 0 0 0 0 0 0 0 36 0
Driving 0 96 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0
Cooking 0 0 60 0 0 0 0 0 3 1 0 0 0 5 0 0 0 27 0
Exercising 0 0 0 73 1 0 0 0 0 0 0 0 0 0 0 5 0 19 2
Reading 0 0 1 1 53 0 0 0 6 21 0 1 0 0 1 0 0 11 0
Presentation 0 0 0 0 0 87 0 0 4 4 0 0 2 0 0 0 0 2 0
Dogs 0 0 0 0 0 0 66 0 0 1 0 0 0 0 0 0 0 30 0
Resting 0 0 0 4 0 0 0 45 4 9 0 22 0 0 0 0 0 9 4
Eating 0 0 1 0 0 0 0 0 83 4 0 0 0 0 0 0 0 7 0
Working 0 0 0 0 0 0 0 0 1 95 0 0 0 0 0 0 0 1 0
Chatting 0 0 0 0 0 0 0 0 8 13 17 0 0 0 8 4 0 43 4
TV 0 0 0 0 0 0 0 0 4 5 0 81 0 0 0 0 0 7 0
Meeting 0 0 0 0 0 4 0 0 2 5 0 0 81 0 0 0 0 5 0
Cleaning 1 0 21 0 0 0 0 0 4 0 0 0 0 46 0 0 0 26 0
Socializing 0 0 2 0 2 0 0 0 9 2 0 0 2 0 45 0 0 34 0
Shopping 0 0 0 2 0 0 0 0 3 1 0 0 0 0 0 64 0 27 0
Biking 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 81 14 0
Family 0 0 0 0 0 0 2 0 2 0 0 0 0 0 0 0 0 90 0
Hygiene 0 0 0 0 0 0 0 0 1 18 0 0 1 1 0 0 0 13 62
0%
Accuracy Color Map
100%

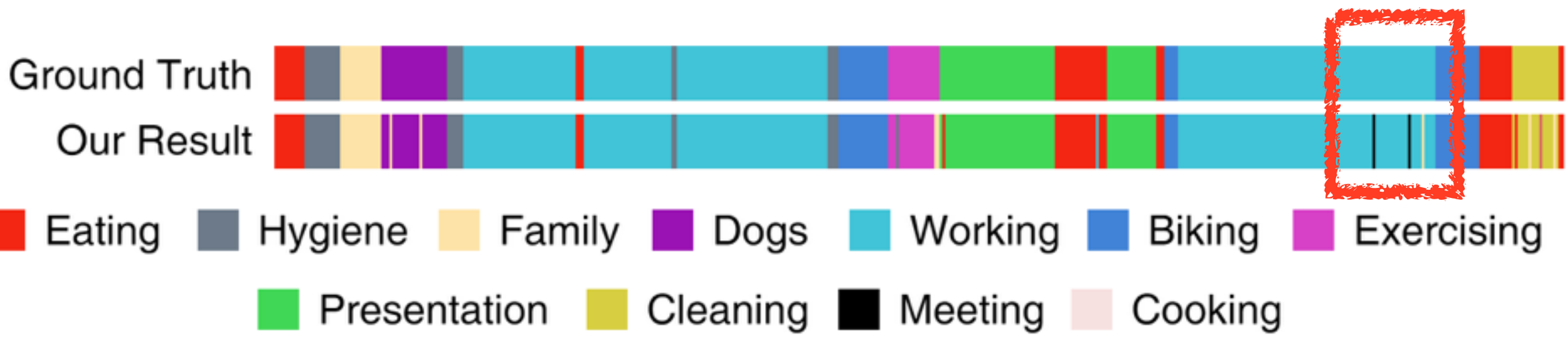
Actual vs. Predicted Activities



Actual vs. Predicted Activities



Actual vs. Predicted Activities



40,000+
Images

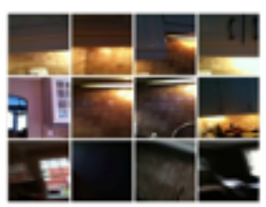
Classifier

Split Train + Test

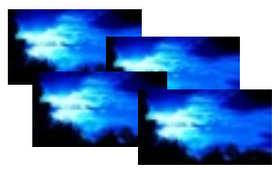
83.12%

How does it perform with images from other people?

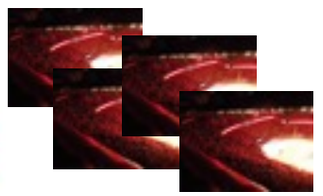
Generalizability



40,000+
Images



2 Days



2 Days

Classifier

Classifier

Classifier

Split Train + Test

83.12%

Prediction

11.48%

Prediction

30.68%

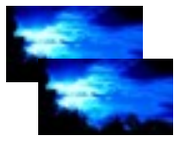
Did not generalize well

40,000+
Images

↓

Classifier

↑



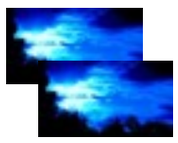
1 Day

What if we train the model with a bit of data from another person?

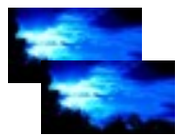
Personalization

40,000+
Images

Classifier



1 Day



1 Day

Classifier

Prediction

76.92%

Average Improvement of **65.44%**
with Personalization

Computer Vision and Nearby Fields

- Computer Graphics: Models to Images
- Comp. Photography: Images to Images
- Computer Vision: Images to Models

Very broad and exciting discipline!

Lots of Resources Online

Kristen Grauman

Associate Professor

Department of Computer Science
University of Texas at Austin

Computer Vision: Algorithms and Applications

© 2010 Richard Szeliski, Microsoft Research