Computer Vision

VISUM 2018 – Day 0

Luis F. Teixeira

What?

Computer Vision vs.

Image Processing

Computer Graphics

Pattern Recognition

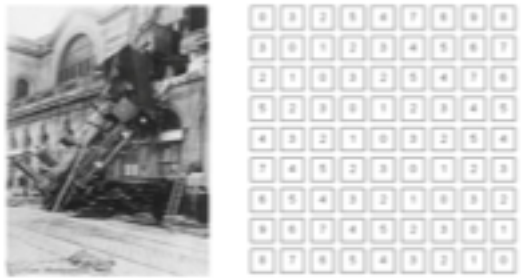
Goal of Computer Vision

To bridge the gap between pixels and "meaning"

Apple?

Goal of Computer Vision

To bridge the gap between pixels and "meaning"



Goal of Computer Vision

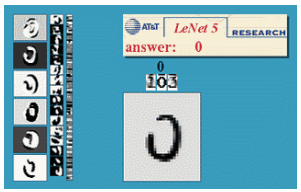
- Provide computers with human-like **perception** capabilities so that they can sense the environment, **understand** the sensed data, take appropriate actions (**make decisions**), learn from this experience in order to enhance future performance
  - **Understand visual information** with no accompanying structural, administrative or descriptive text information
- Sources of difficulties:
  - Sensory gap
  - Semantic gap

Where?

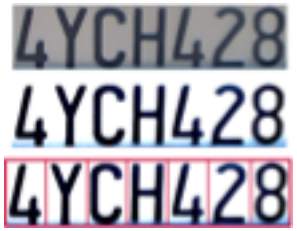
Optical character recognition (OCR)

Technology to convert scanned docs to text

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



http://www.research.att.com/~yann/

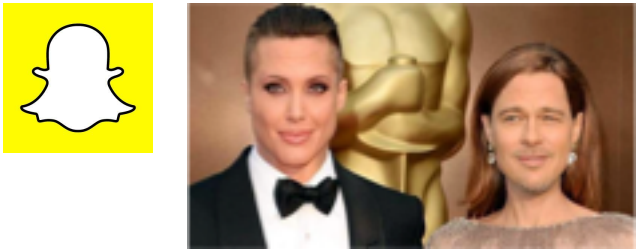


License plate readers

James Hays

Face detection

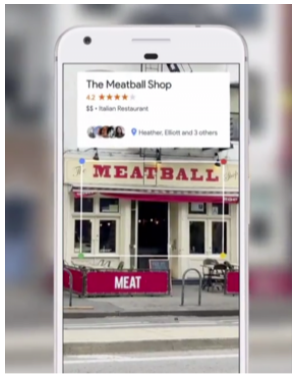
Almost all digital cameras detect faces

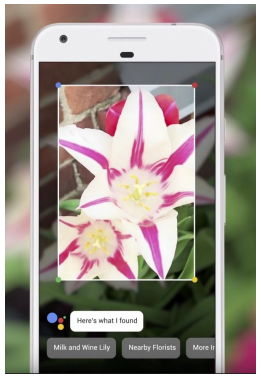


Face recognition

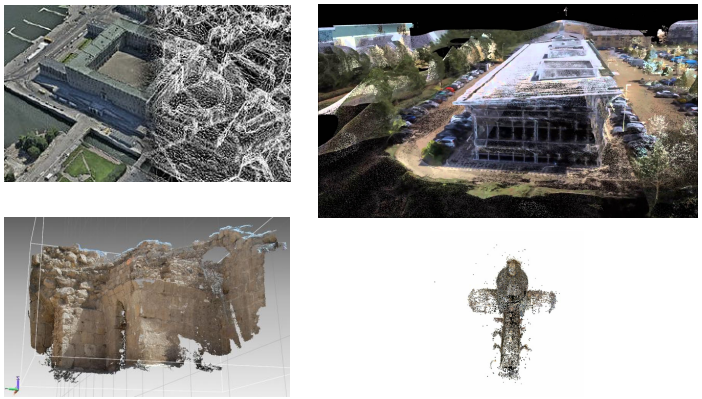
Object recognition (in mobile phones)

e.g., Google Lens





3D from images

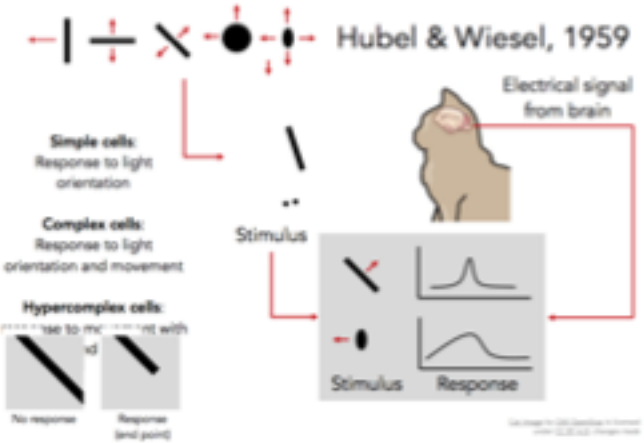


Building Rome in a Day: Agarwal et al. 2009

Autonomous cars

When?

Hubel & Wiesel, 1959



Fei-Fei Li et al.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
PROJECT MAC

Artificial Intelligence Group
Vision Memo. No. 100.

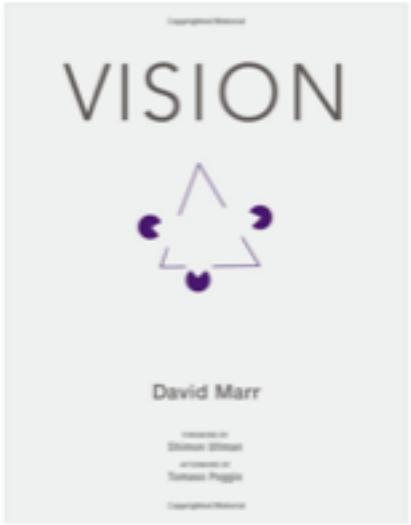
July 7, 1966

THE SUMMER VISION PROJECT

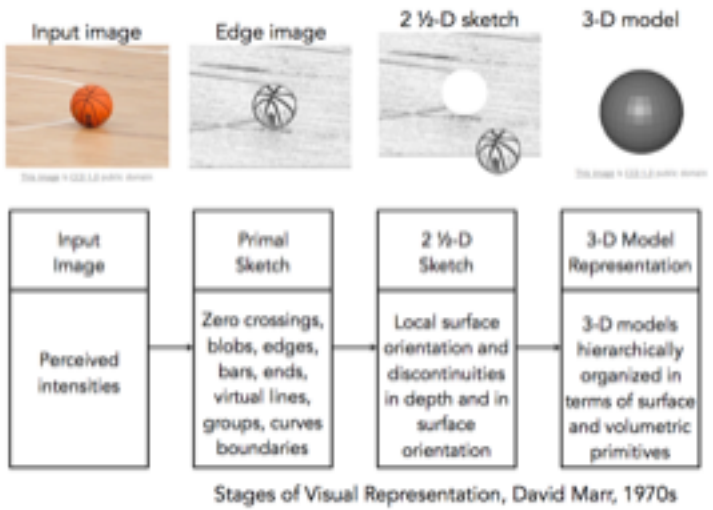
Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

Fei-Fei Li et al.



Fei-Fei Li et al.



Stages of Visual Representation, David Marr, 1970s

Fei-Fei Li et al.

How?

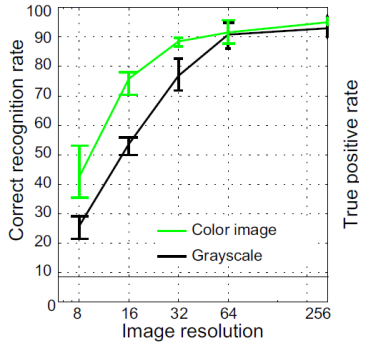
Let us focus in one task
in Computer Vision

Recognition

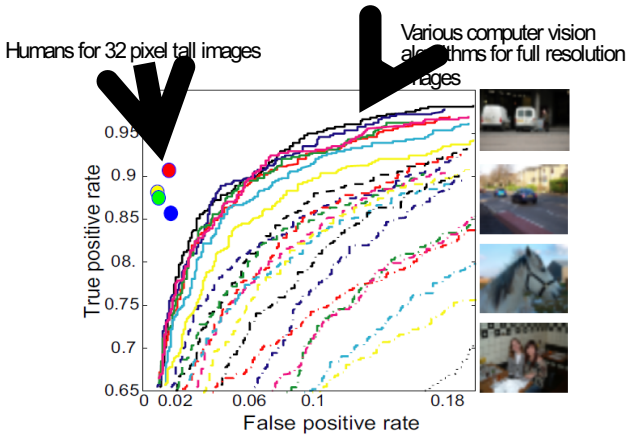
Scene recognition

Human Scene Recognition

15 scene categories

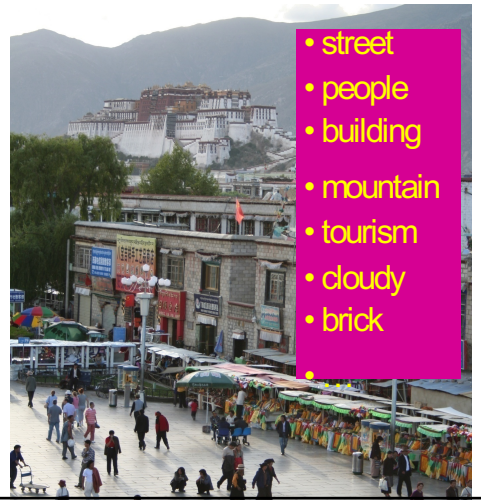


Humans vs. Computers: Car-Image Classification



things have changed since 2015…

Image annotation / tagging / attributes



Object detection

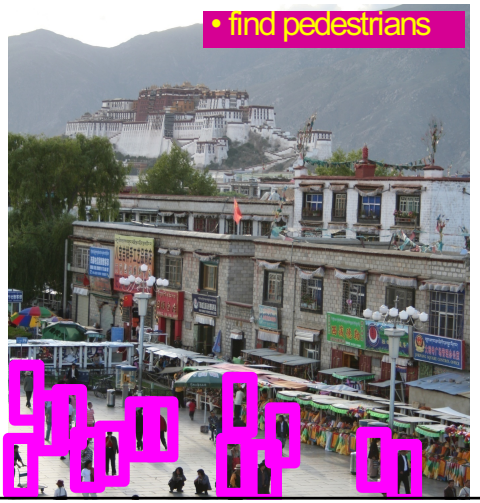


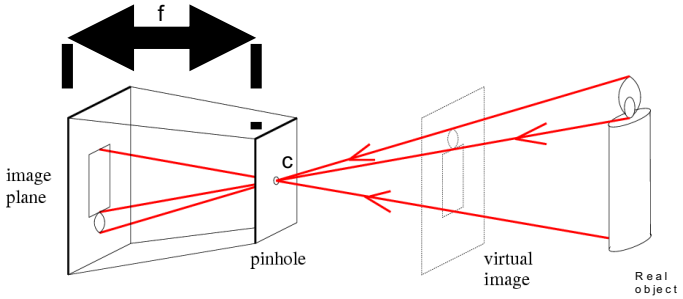
Image parsing / semantic segmentation

Typical pipeline

Cameras

- Camera obscura and pinhole camera
  - Light from a scene passes through the aperture and projects an inverted image on the opposite side of the box
- Photography in the 19th century
- Now we have cameras everywhere

Pinhole camera model



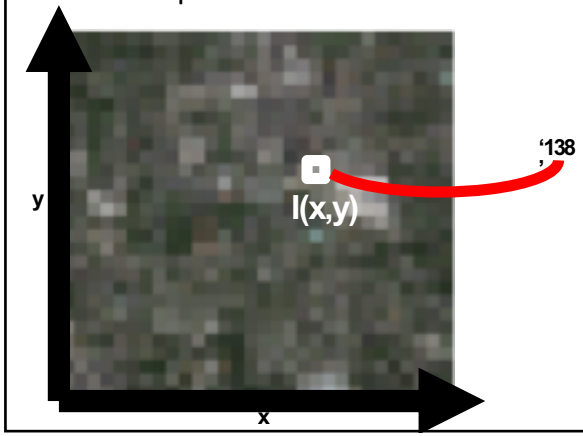
f = Focal length

c = Optical center of the camera

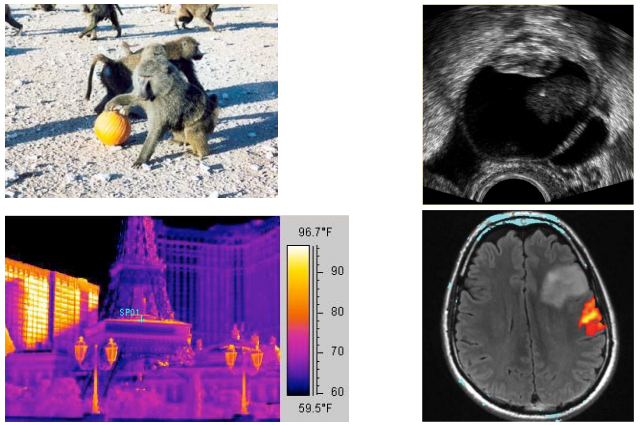
Figure from Forsyth

What is each part of an image?

- Pixel -> picture element



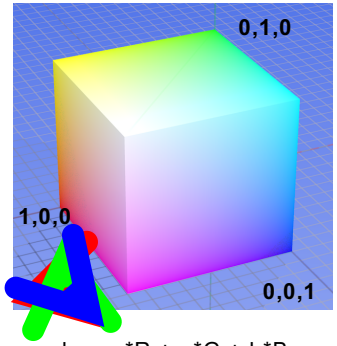
Example 2D Images



Danny Alexander

Color spaces: RGB

Default color space



R = 1 (G=0,B=0)

G = 1 (R=0,B=0)

B = 1 (R=0,G=0)

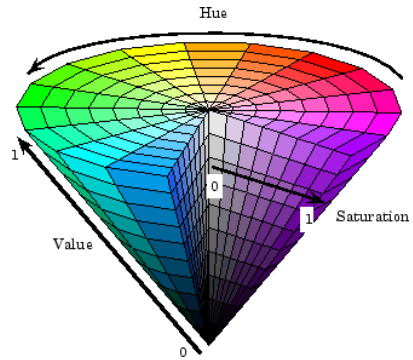
Any color = r*R + g*G + b*B

- Strongly correlated channels
- Non-perceptual

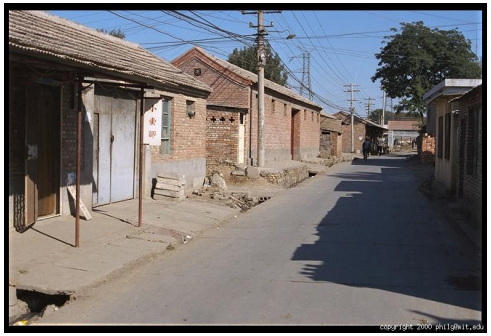
Image from: http://en.wikipedia.org/wiki/File:RGB_color_solid_cube.png

Color spaces: HSV

Intuitive color space



Most information in intensity



Original image

James Hays

Most information in intensity

Only intensity shown – constant color

James Hays

Most information in intensity

Only color shown – constant intensity

James Hays

Image filtering

Compute function of local neighborhood at each position

$$h[m,n] = \sum_{k,l} f[k,l]\, I[m+k, n+l]$$

James Hays

Image filtering

Compute function of local neighborhood at each position

h=output f=filter I=image

$$h[m,n] = \sum_{k,l} f[k,l]\, I[m+k, n+l]$$

2d coords=k,l 2d coords=m,n

[ ] [ ] [ ]

Example: box filter

$f[\cdot,\cdot]$

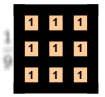
$\frac{1}{9}$

1	1	1
1	1	1
1	1	1

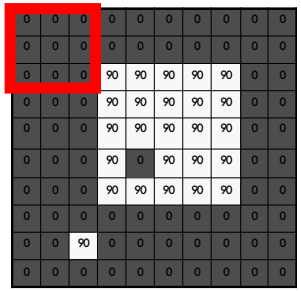
Slide credit: David Lowe (UBC)

Image filtering

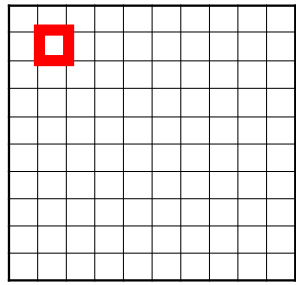
$f[\cdot,\cdot]$



$I[.,.]$



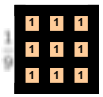
$h[.,.]$



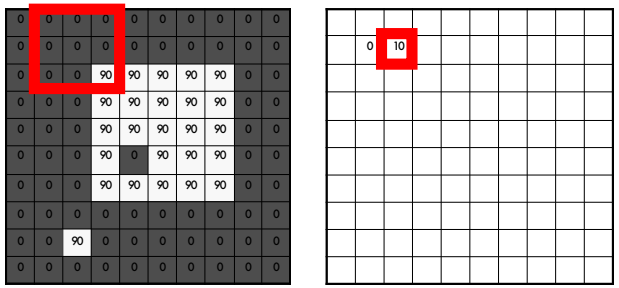
$$h[m,n] = \sum_{k,l} f[k,l]\, I[m+k, n+l]$$

Credit: S. Seitz

Image filtering

$f[\cdot,\cdot]$ 

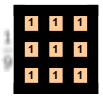
$I[.,.]$ $h[.,.]$



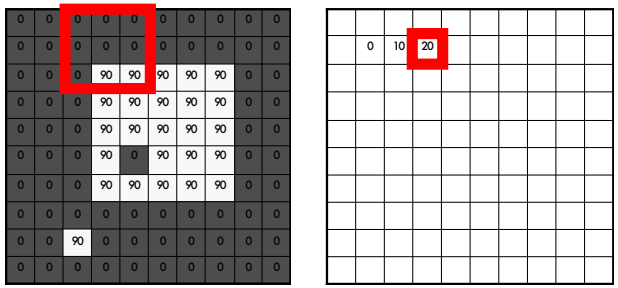
$$h[m,n] = \sum_{k,l} f[k,l]\, I[m+k, n+l]$$

Credit: S. Seitz

Image filtering

$f[\cdot,\cdot]$ 

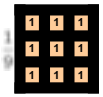
$I[.,.]$ $h[.,.]$



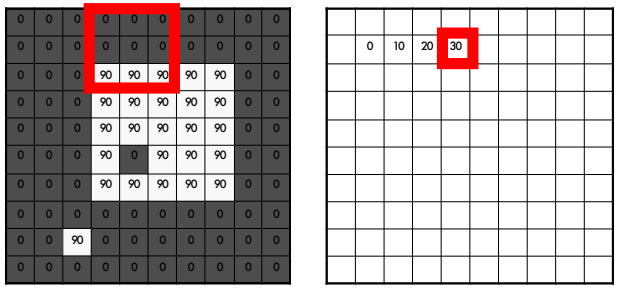
$$h[m,n] = \sum_{k,l} f[k,l]\, I[m+k, n+l]$$

Credit: S. Seitz

Image filtering

$f[\cdot,\cdot]$ 

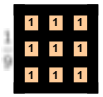
$I[.,.]$ $h[.,.]$



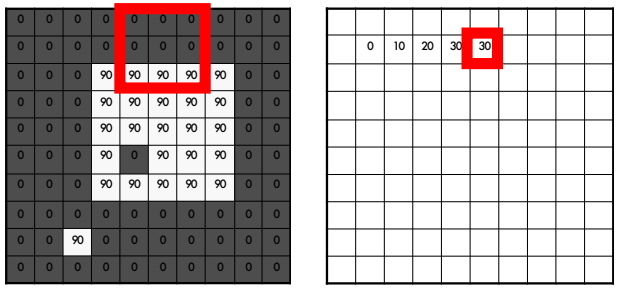
$$h[m,n] = \sum_{k,l} f[k,l]\, I[m+k, n+l]$$

Credit: S. Seitz

Image filtering

$f[\cdot,\cdot]$ 

$I[.,.]$ $h[.,.]$



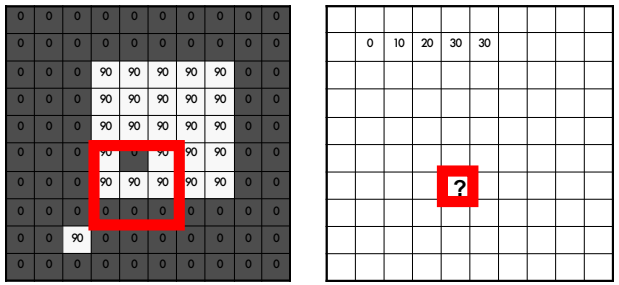
$$h[m,n] = \sum_{k,l} f[k,l]\, I[m+k, n+l]$$

Credit: S. Seitz

Image filtering

$f[\cdot,\cdot]$

$I[.,.]$ $h[.,.]$

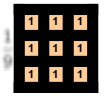


$$h[m,n] = \sum_{k,l} f[k,l]\, I[m+k, n+l]$$

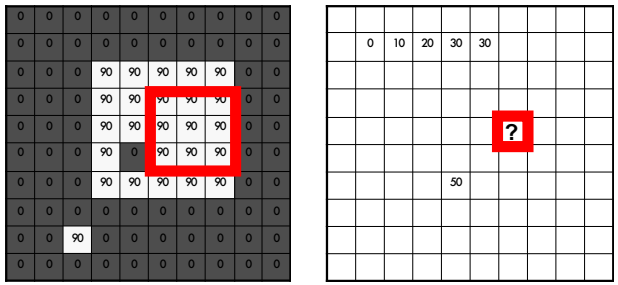
Credit: S. Seitz

Image filtering

$f[\cdot,\cdot]$



$I[.,.]$ $h[.,.]$



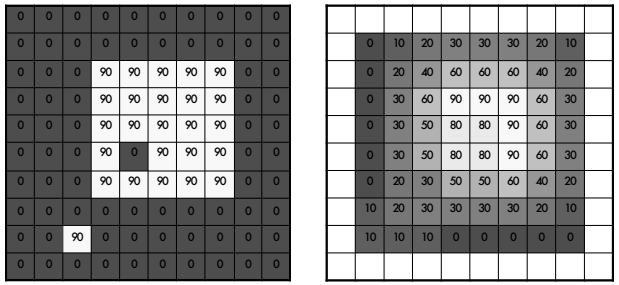
$$h[m,n] = \sum_{k,l} f[k,l]\, I[m+k, n+l]$$

Credit: S. Seitz

Image filtering

$f[\cdot,\cdot]$

$I[.,.]$ $h[.,.]$



$$h[m,n] = \sum_{k,l} f[k,l]\, I[m+k, n+l]$$

Credit: S. Seitz

Box Filter

What does it do?

- Replaces each pixel with an average of its neighborhood
- Achieve smoothing effect (remove sharp features)

$f[\cdot,\cdot]$

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

Slide credit: David Lowe (UBC)

By the way: CNNs

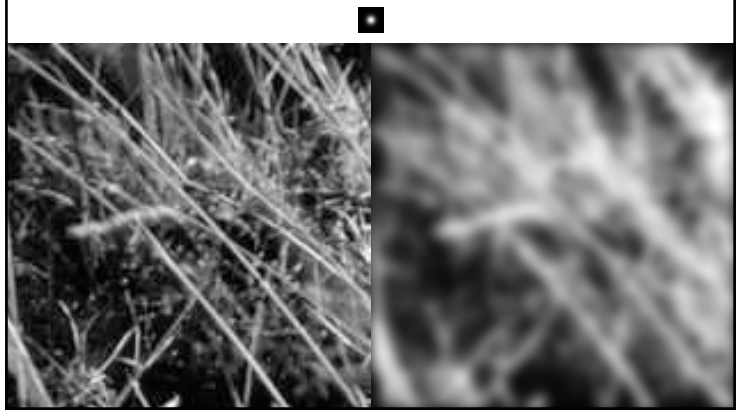
- Convolution is the basic operation in Convolutional Neural Networks
- Learning convolution kernels allows us to learn which `features' provide useful information in images.

Smoothing with box filter

James Hays

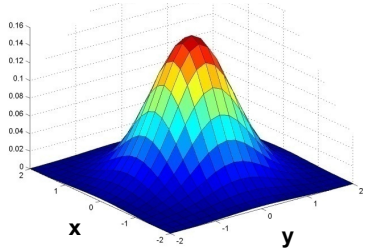
Smoothing with Gaussian filter

James Hays



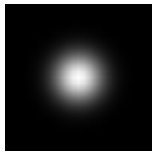
Gaussian filter

- Weight contributions of neighboring pixels by nearness



0.003	0.013	0.022	0.013	0.003
0.013	0.059	0.097	0.059	0.013
0.022	0.097	0.159	0.097	0.022
0.013	0.059	0.097	0.059	0.013
0.003	0.013	0.022	0.013	0.003

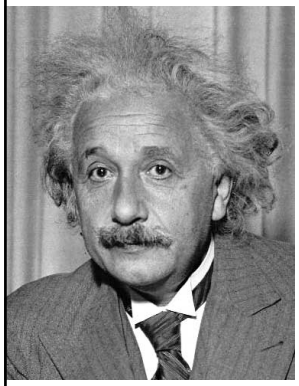
5 x 5, σ = 1



$$G_\sigma = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}}$$

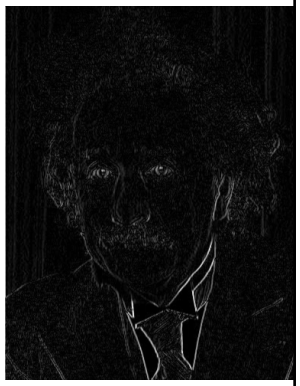
Slide credit: Christopher Rasmussen

More linear filters



1	0	-1
2	0	-2
1	0	-1

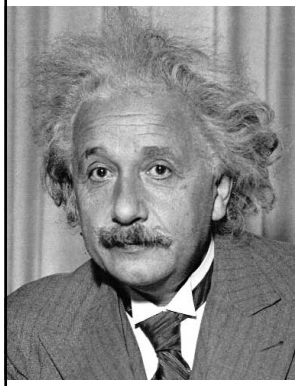
Sobel



Vertical Edge
(absolute value)

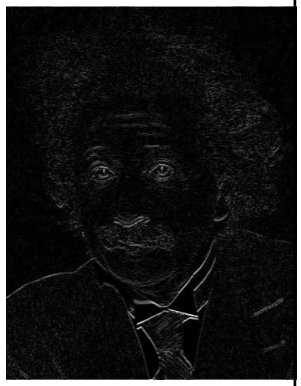
David Lowe

More linear filters



1	2	1
0	0	0
-1	-2	-1

Sobel



Horizontal Edge
(absolute value)

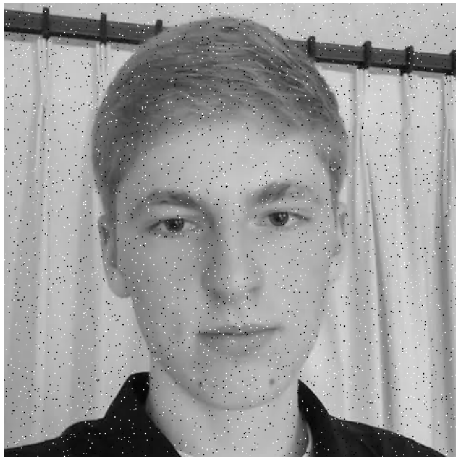
David Lowe

Median filters

- Operates over a window by selecting the median intensity in the window.
- 'Rank' filter as based on ordering of gray levels
  - E.G., min, max, range filters
- Not a convolution-based filter

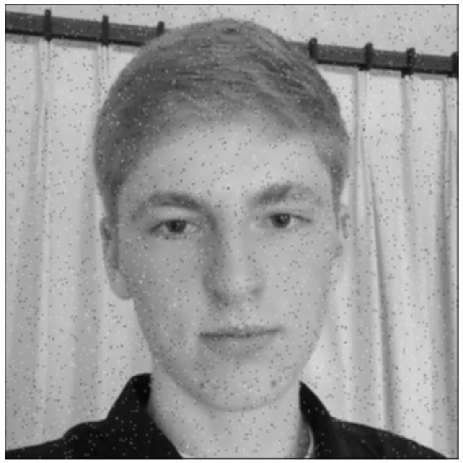
Steve Seitz, Steve Marschner

Noisy – Salt and Pepper



James Tompkin

Mean – 3 x 3 filter

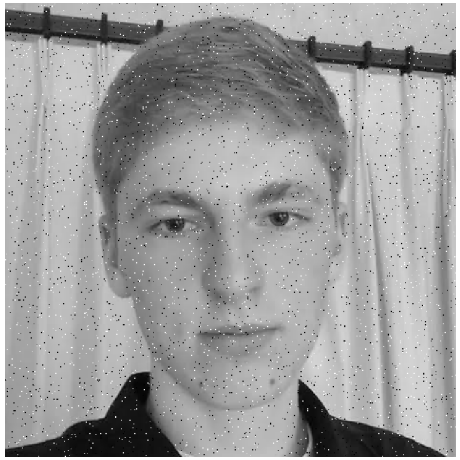


James Tompkin

Mean – 11 x 11 filter

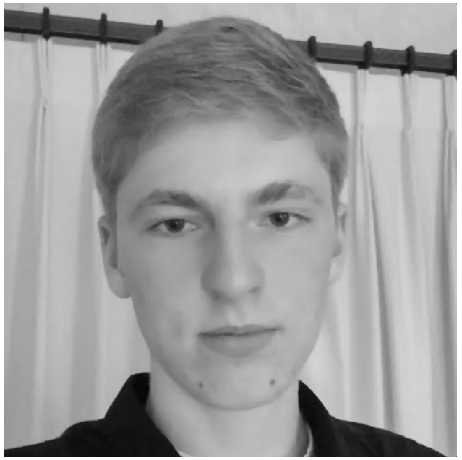
James Tompkin

Noisy – Salt and Pepper



James Tompkin

Median – 3 x 3



James Tompkin

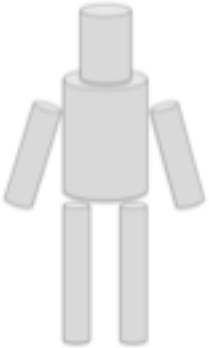
History of ideas in recognition

- 1960s – early 1990s: the geometric era

Svetlana Lazebnik

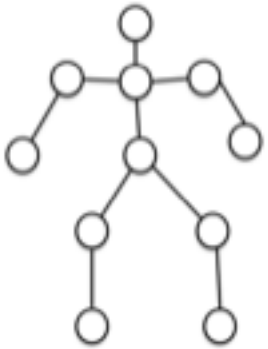
- Generalized Cylinder

Brooks & Binford, 1979



- Pictorial Structure

Fischler and Elschlager, 1973



Fei-Fei Li et al.

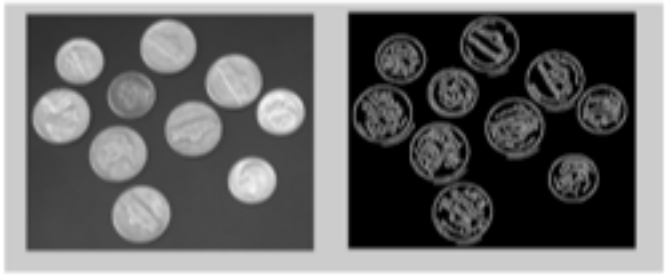
Alignment: fitting a model to a transformation between pairs of features (matches) in two images

Svetlana Lazebnik

Recognition by components

Biederman (1987)

Svetlana Lazebnik



David Lowe, 1987

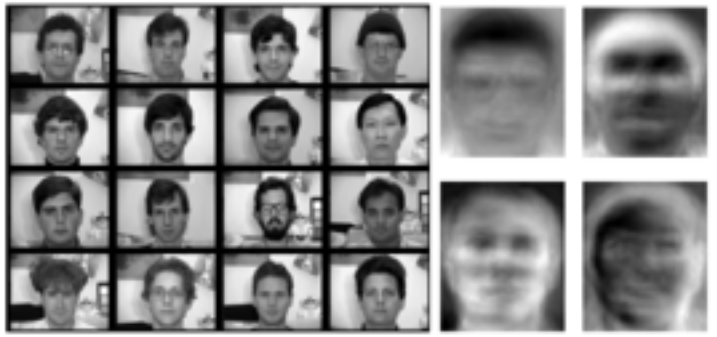
Fei-Fei Li et al.

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models

Svetlana Lazebnik

Eigenfaces (Turk & Pentland, 1991)

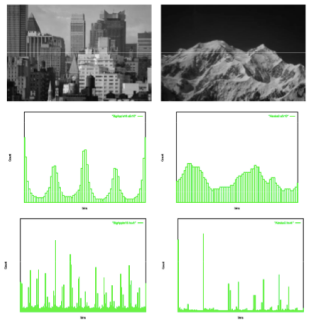


Experimental Condition	Correct/Unknown Recognition Percentage		
	Lighting	Orientation	Scale
Forced classification	96/0	85/0	64/0
Forced 100% accuracy	100/19	100/39	100/60
Forced 20% unknown rate	100/20	94/20	74/20

Svetlana Lazebnik

Scene Classification

- Example: using global features to classify city/landscape images
  - Based on colour and edge histograms
  - KNN classifier
  - Features fusion using weighted concatenation

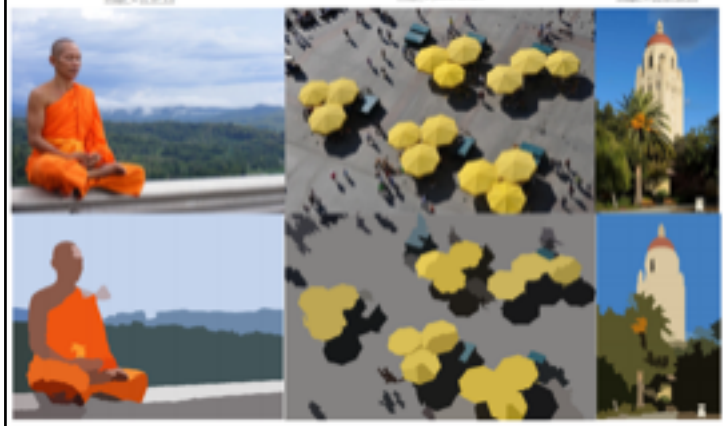


On Image Classification: City vs. Landscape. Vailaya, A. and Jain, A. and Zhang, H. IEEE Workshop on Content - Based Access of Image and Video Libraries,1998

Global Appearance Models

- Limitations of global appearance models
  - Requires global registration of patterns
  - Not robust to clutter, occlusion, geometric transformations
- If object recognition is too hard, why don't we start by performing object segmentation?
  - Group the pixels into meaningful areas and separating the objects from the background

Normalized Cut (Shi & Malik, 1997)



Fei-Fei Li et al.

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches

Svetlana Lazebnik

Sliding window approaches

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features

Svetlana Lazebnik

Fei-Fei Li et al.

"SIFT" & Object Recognition, David Lowe, 1999

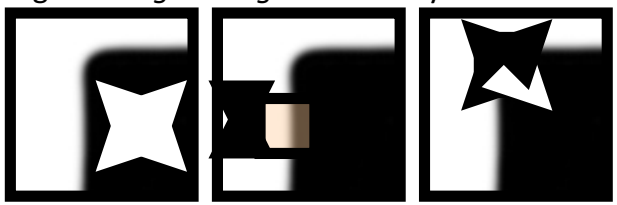
Fei-Fei Li et al.

Local features detectors

- What points would be better choices?

Corners

- We should easily recognize the point by looking through a small window
- Shifting a window in *any direction* should give *a large change* in intensity



"**flat**" region: no change in all directions

"**edge**": no change along the edge direction

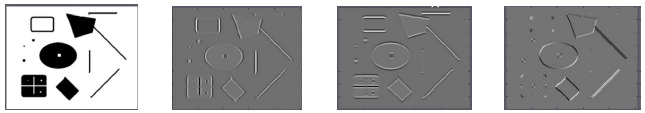
"**corner**": significant change in all directions

Alyosha Efros

Harris Detector Formulation

$$M = \sum w(x,y) \begin{bmatrix} I_x I_x & I_x I_y \\ I_x I_y & I_y I_y \end{bmatrix}$$

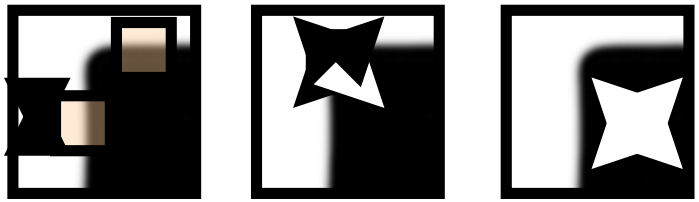
2 x 2 matrix of image derivatives (averaged in neighborhood of a point).



Notation: $I_x \Leftrightarrow \frac{\partial I}{\partial x}$ $I_y \Leftrightarrow \frac{\partial I}{\partial y}$ $I_x I_y \Leftrightarrow \frac{\partial I}{\partial x}\frac{\partial I}{\partial y}$

James Hayes

Corner response function



"edge": $\lambda_1 >> \lambda_2$, $\lambda_2 >> \lambda_1$

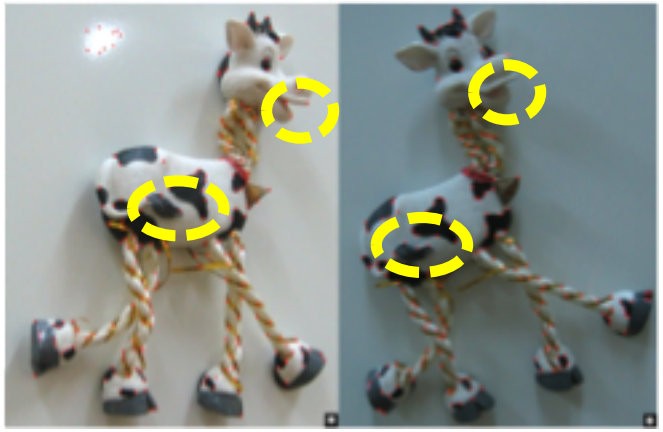
"corner": λ_1 and λ_2 are large, $\lambda_1 \sim \lambda_2$;

"flat" region λ_1 and λ_2 are small;

$$C = \lambda_1\lambda_2 - \alpha(\lambda_1 + \lambda_2)^2$$

α: constant (0.04 to 0.06)

Harris corner detector

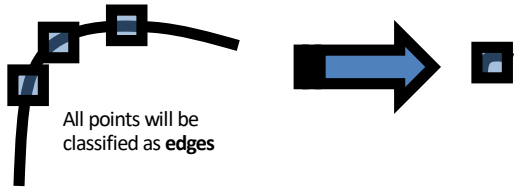


Darya Frolova, Denis Simakov

Harris corner detector

- Properties of the Harris corner detector
  - Rotation invariance
    rotates but its shape (i.e. eigenvalues) remains the **same**

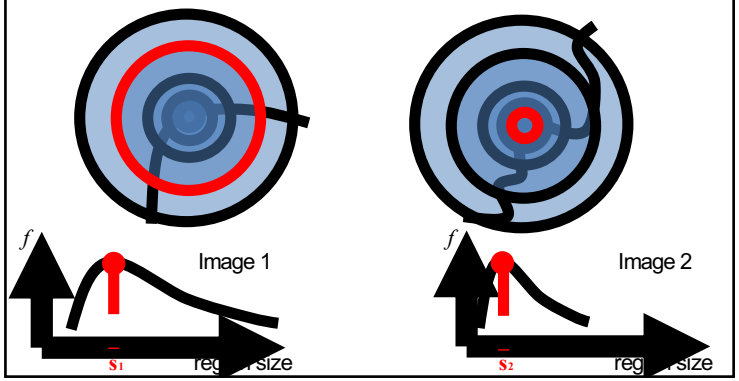
  - Not invariant to image scale



Kristen Grauman

Scale invariant detection

Intuition - Find scale that gives local maxima of some signature function f in both position and scale.



Scale invariant detection

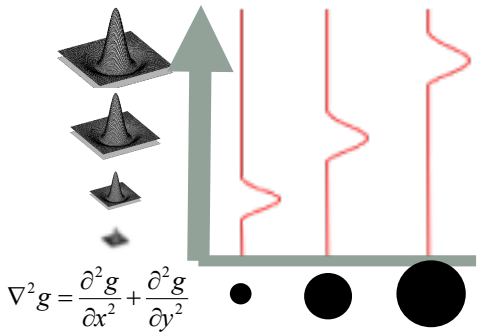
- A "good" function for scale detection:
  has one stable sharp peak

- For usual images: a good function would be one which responds to contrast (sharp local intensity change)

Juan Carlos Niebles

Scale invariance detection

- Useful signature function
  - Laplacian-of-Gaussian = "blob" detector

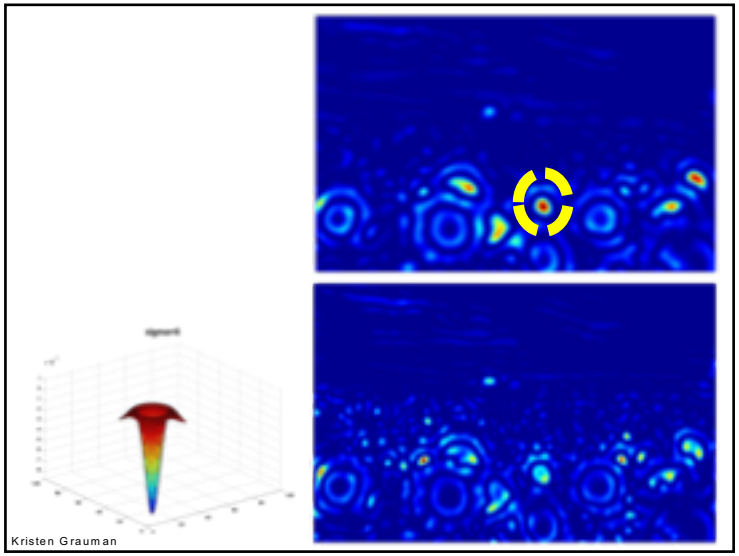


$$\nabla^2 g = \frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2}$$

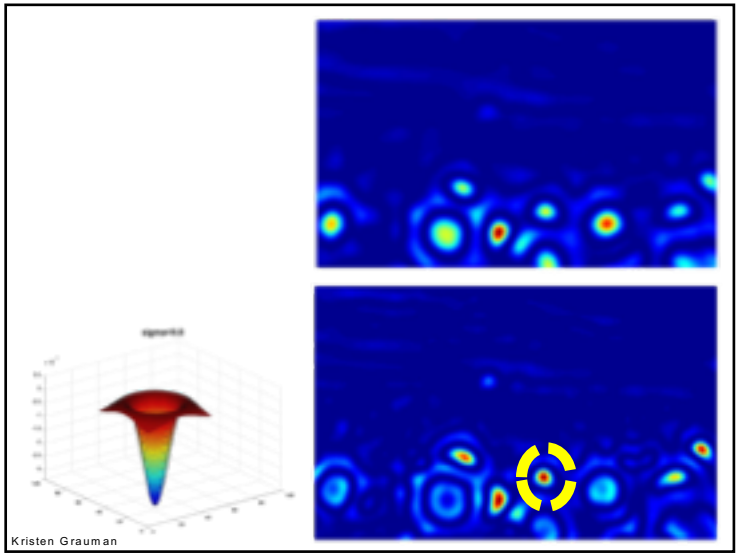
Scaled image
at ¾ the size

Original image

Kristen Grauman



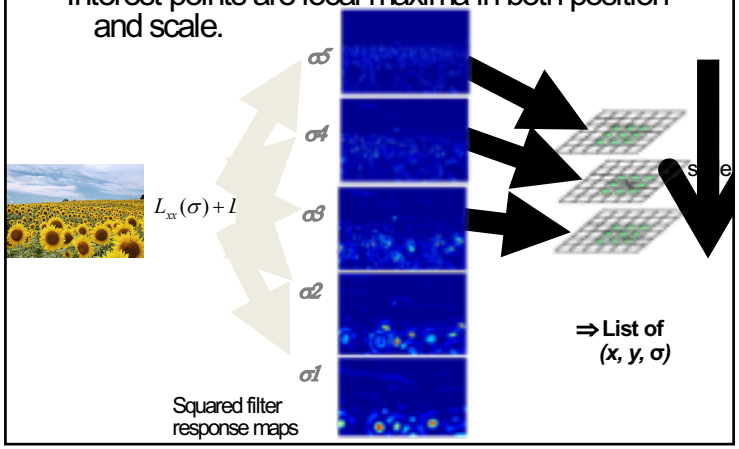
Kristen Grauman



Kristen Grauman

Scale invariant interest points

Interest points are local maxima in both position and scale.



Scale invariant interest points

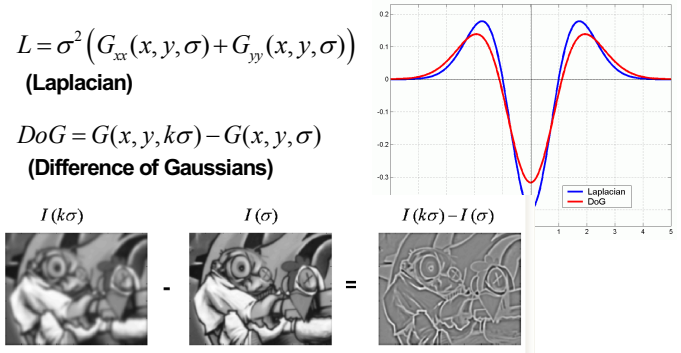
- We can approximate the Laplacian with a difference of Gaussians, which is more efficient to implement.

$$L = \sigma^2\left(G_{xx}(x,y,\sigma)+G_{yy}(x,y,\sigma)\right)$$

(Laplacian)

$$DoG = G(x,y,k\sigma)-G(x,y,\sigma)$$

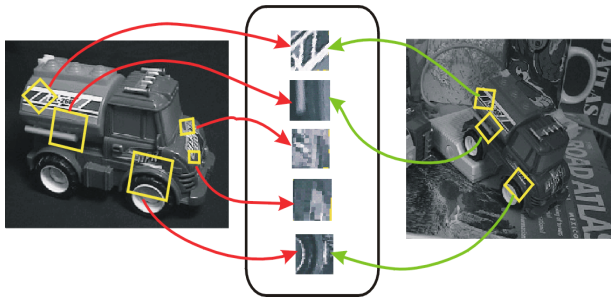
(Difference of Gaussians)



Invariant local features

Find features that are invariant to transformations

- geometric invariance: translation, rotation, scale
- photometric invariance: brightness, exposure, …



Feature Descriptors

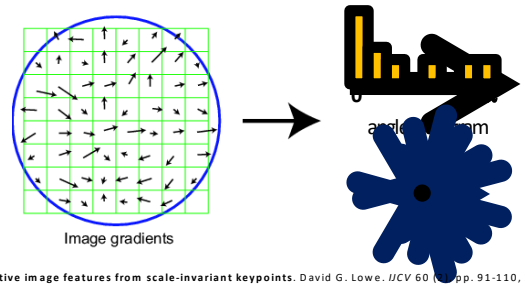
Local descriptors

- In each detected feature (point), a descriptor is then extracted
- Histogram-based descriptors
  - Based on the histogram of oriented gradient
  - SIFT, SURF, GLOH and HOG
- Compact descriptors
  - Based on binary strings obtained comparing pairs of image intensities
  - BRIEF, ORB, BRISK and FREAK

SIFT descriptor

Basic idea:

- Take 16x16 square window around detected feature
- Compute edge orientation (angle of the gradient - 90°) for each pixel
- Throw out weak edges (threshold gradient magnitude)
- Create histogram of surviving edge orientations

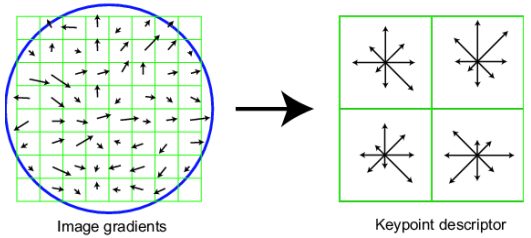


Distinctive image features from scale-invariant keypoints. David G. Lowe. *IJCV* 60 (2), pp. 91-110, 2004.

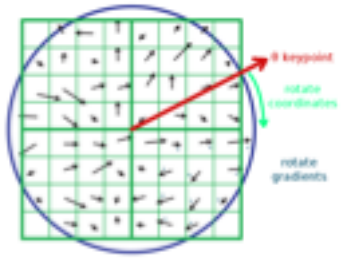
SIFT descriptor

Full version

- Divide the 16x16 window into a 4x4 grid of cells (2x2 case shown below)
- Compute an orientation histogram for each cell
- 16 cells * 8 orientations = 128 dimensional descriptor



SIFT descriptor

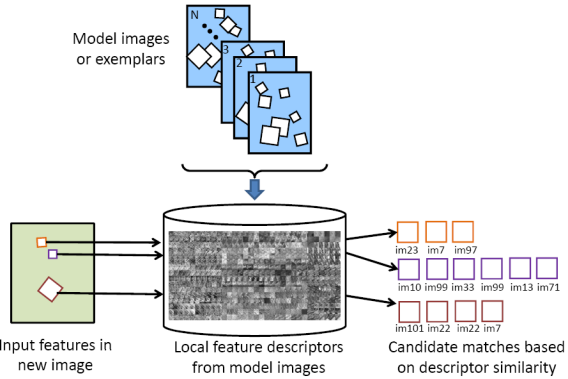


most prominent gradient

- To become rotation invariant, rotate the gradient directions AND locations by (-keypoint orientation)
  - Now we've cancelled out rotation and have gradients expressed at locations **relative** to keypoint orientation θ
  - We could also have just rotated the whole image by -θ, but that would be slower.

Large-scale image search

Combining local features, indexing, and spatial constraints



K. Grauman and B. Leibe

Spatial Verification

(for Instance Recognition)

Both image pairs have many visual words in common.

Ondrej Chum

Spatial Verification

(for Instance Recognition)

Only some of the matches are mutually consistent

We need to exclude all other matches (i.e. **outliers**)

Ondrej Chum

RANSAC - RAndom SAmple Consensus

- Idea:
  - All the inliers will agree with each other on the transformation; the (hopefully small) number of outliers will (hopefully) disagree with each other
- RANSAC only has guarantees if there are < 50% outliers

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models

Svetlana Lazebnik

Histogram of Gradients (HoG)
Dalal & Triggs, 2005

Deformable Part Model
Felzenswalb, McAllester, Ramanan, 2009

Fei-Fei Li et al.

History of ideas in recognition

- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- Mid-2000s: bags of features

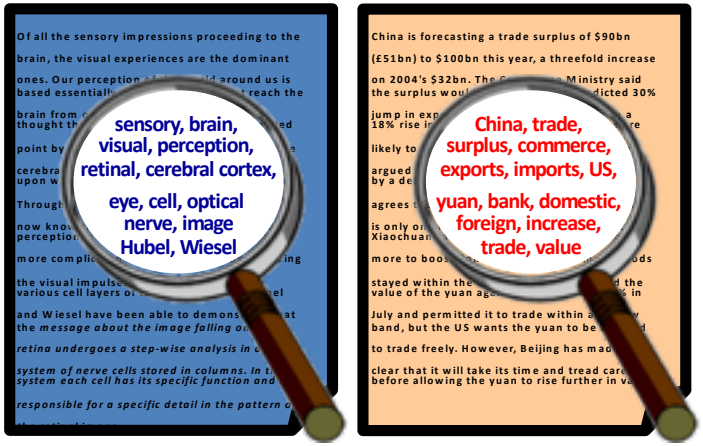
Svetlana Lazebnik

Bag-of-features models

Object → Bag of 'words'

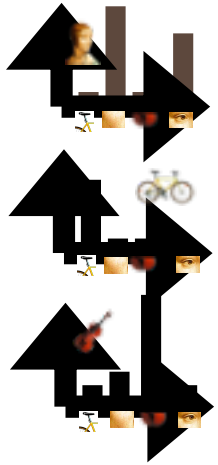
Svetlana Lazebnik

Analogy to documents



Bags of visual words

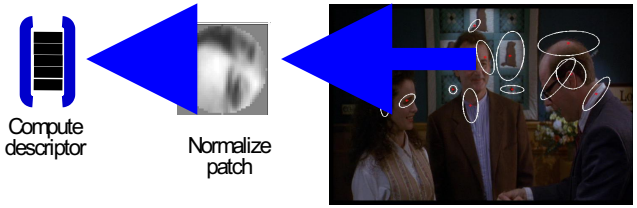
- Summarize entire image based on its distribution (histogram) of word occurrences.
- Analogous to bag of words representation commonly used for documents.



Feature extraction

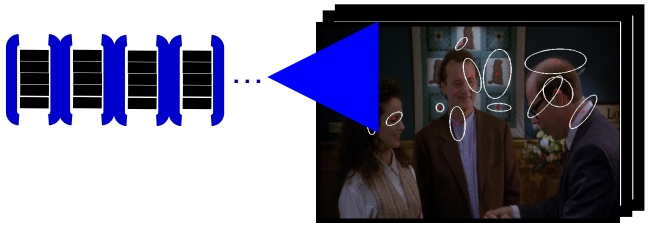
- Regular grid or interest regions

Feature extraction

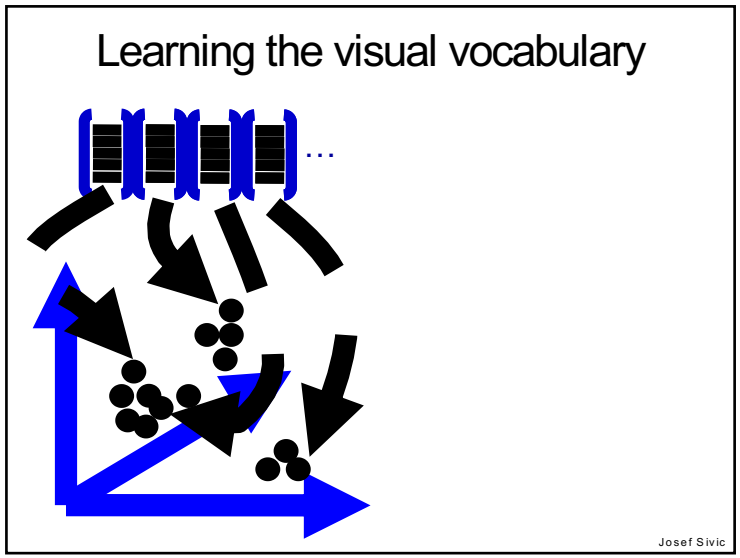


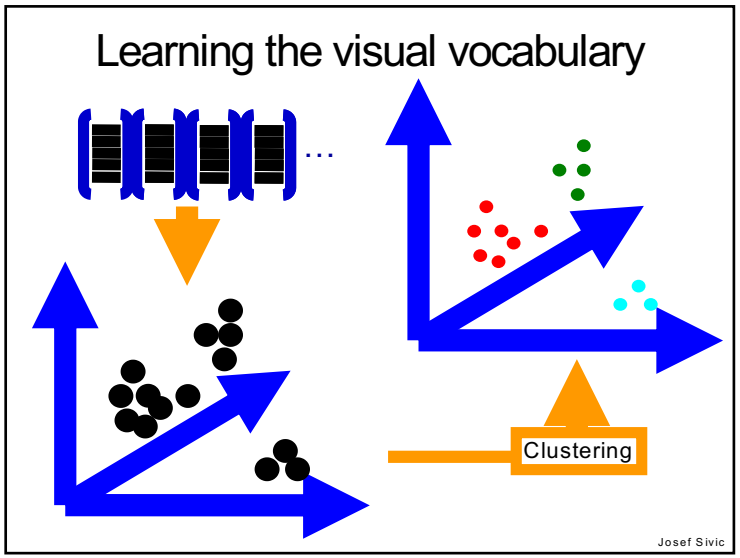
Josef Sivic

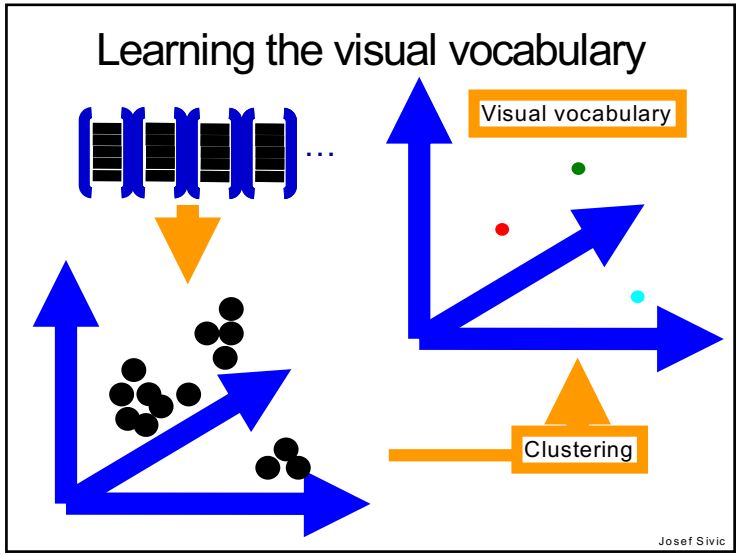
Feature extraction

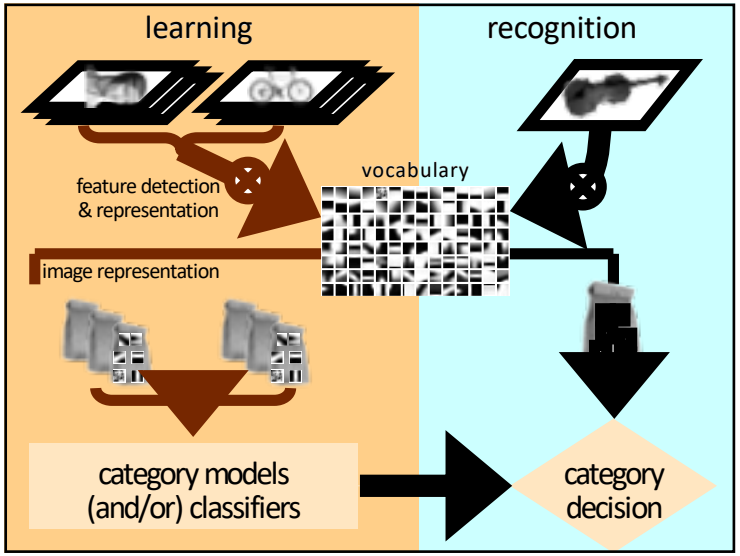


Josef Sivic


Learning the visual vocabulary
Josef Sivic


Learning the visual vocabulary
Clustering
Josef Sivic


Learning the visual vocabulary
Visual vocabulary
Clustering
Josef Sivic


learning
recognition
feature detection & representation
vocabulary
image representation
category models (and/or) classifiers
category decision

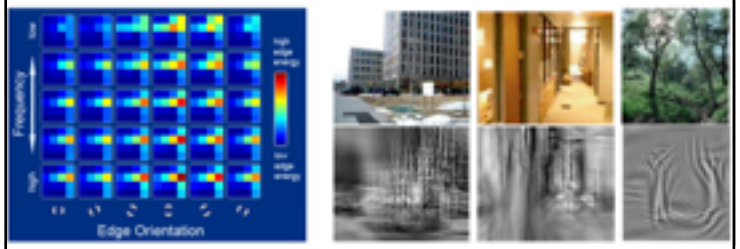
History of ideas in recognition

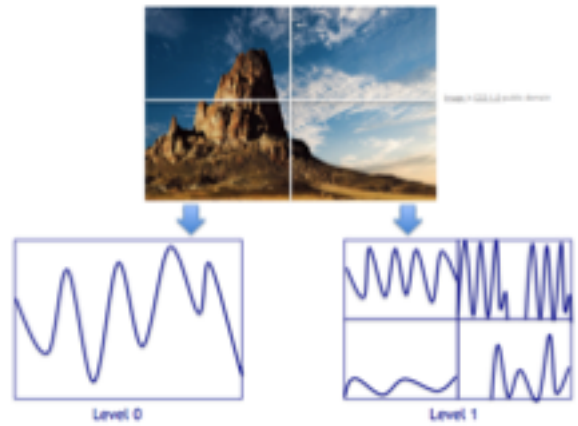
- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- Mid-2000s: bags of features
- Late 2000s: combination of local and global models

Svetlana Lazebnik

Global scene descriptors

- The "gist" of a scene: Oliva & Torralba (2001)





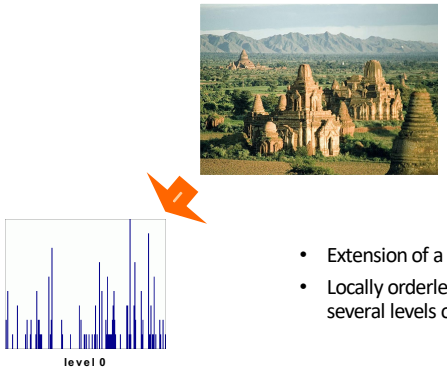
Spatial Pyramid Matching, Lazebnik, Schmid & Ponce, 2006

Fei-Fei Li et al.

Spatial pyramid

Compute histogram in each spatial bin

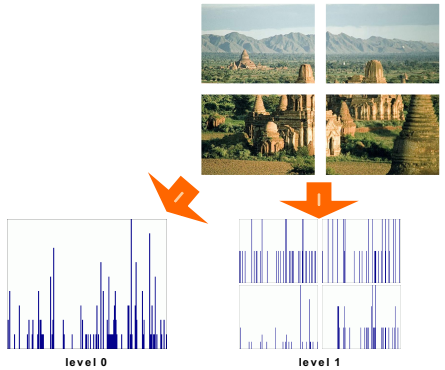
Spatial pyramid representation



- Extension of a bag of words
- Locally orderless representation at several levels of resolution

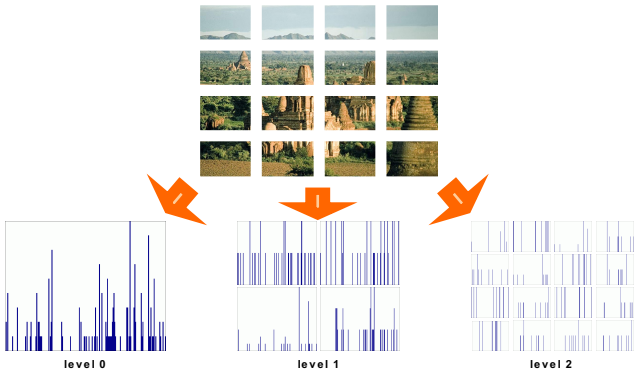
Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation



Lazebnik, Schmid & Ponce (CVPR 2006)

Spatial pyramid representation



Lazebnik, Schmid & Ponce (CVPR 2006)

History of ideas in recognition

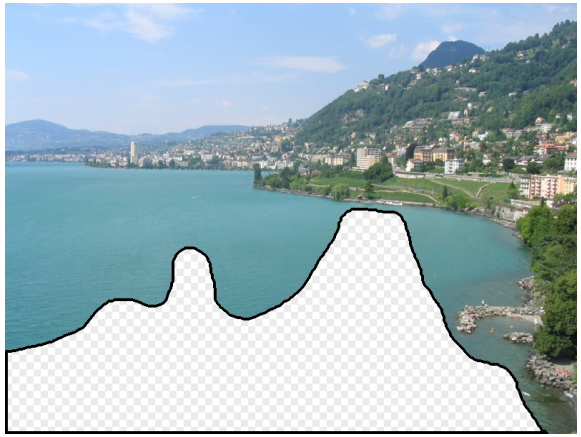
- 1960s – early 1990s: the geometric era
- 1990s: appearance-based models
- Mid-1990s: sliding window approaches
- Late 1990s: local features
- Early 2000s: parts-and-shape models
- Mid-2000s: bags of features
- Late 2000s: combination of local and global models
- Early 2010s: data-driven models

Svetlana Lazebnik

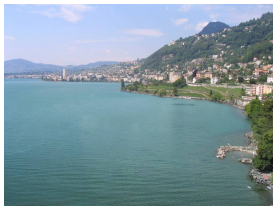
Data-driven models

- The idea
  - What if invariance / generalization isn't actually the core difficulty of computer vision?
  - What if we can perform high level reasoning with brute-force, data-driven algorithms?

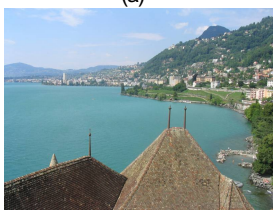
What should the missing region contain?



Which is the original?



(a)

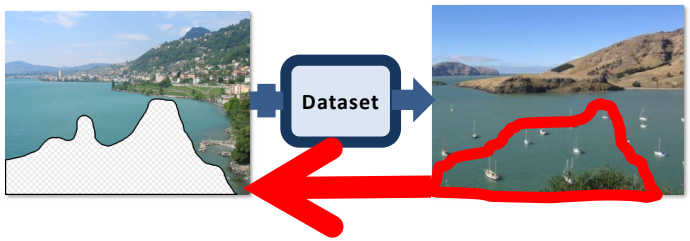


(b)

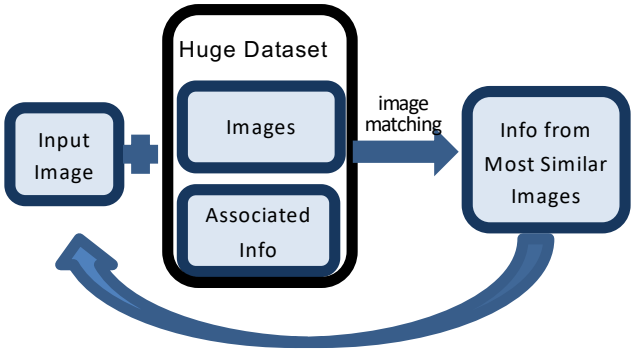
(c)

How it works

- Find a similar image from a large dataset
- Blend a region from that image into the hole



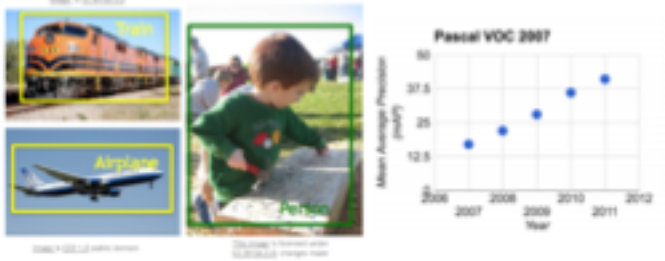
General Principal



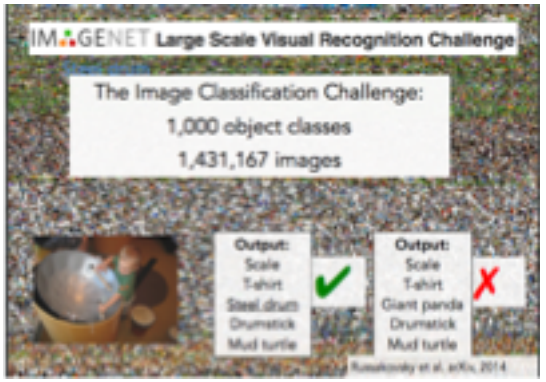
Hopefully, If you have enough images, the dataset will contain very similar images that you can find with simple matching methods.

PASCAL Visual Object Challenge (20 object categories)

[Everingham et al. 2006-2012]



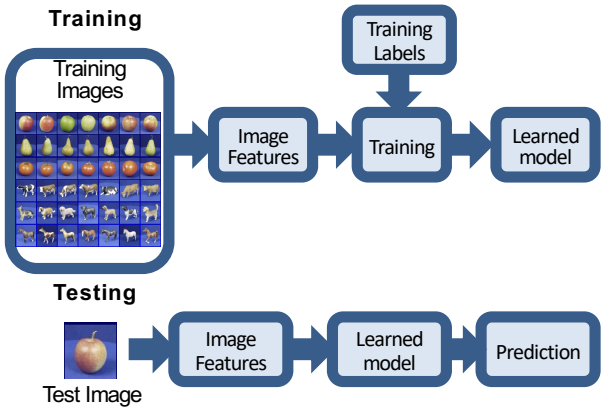
Fei-Fei Li et al.



Major breakthrough

- In 2012
  - "The Revolutionary Technique That Quietly Changed Machine Vision Forever"
  - https://www.technologyreview.com/s/530561/the-revolutionary-technique-that-quietly-changed-machine-vision-forever/

Pattern recognition in computer vision



Trends in Computer Vision

References

- Richard Szeliski, *Computer Vision: Algorithms and Applications*, 2010 - http://szeliski.org/Book/
- David Forsyth and Jean Ponce, *Computer Vision: A Modern Approach* 2nd Edition, 2012
- Simon J. D. Prince, *Computer Vision: Models, Learning, and Inference*, 2012
- Rafael C. Gonzalez and Richard E. Woods, *Digital Image Processing* 3rd Edition, 2007
- Richard Hartley and Andrew Zisserman, *Multiple View Geometry* 2nd Edition, 2004
- Ian Goodfellow , Yoshua Bengio, Aaron Courville and Francis Bach, *Deep Learning*, 2016

Online References

- Fei-Fei Li et al. (Stanford University) - CS 131 Computer Vision: Foundations and Applications
  - http://vision.stanford.edu/teaching/cs131_fall1617/index.html
- James Tompkin et al. (Brown University) - CSCI 1430: Introduction to Computer Vision
  - https://cs.brown.edu/courses/csci1430/
- Kristen Grauman et al. (University of Texas at Austin) - CS 376: Computer Vision
  - http://vision.cs.utexas.edu/376-spring2018/
- Rob Fergus et al. (New York University) - CSCI-GA.2271-001: Computer Vision
  - https://cs.nyu.edu/~fergus/teaching/vision/index.html