lecture 1: introduction

deep learning for vision

Yannis Avrithis

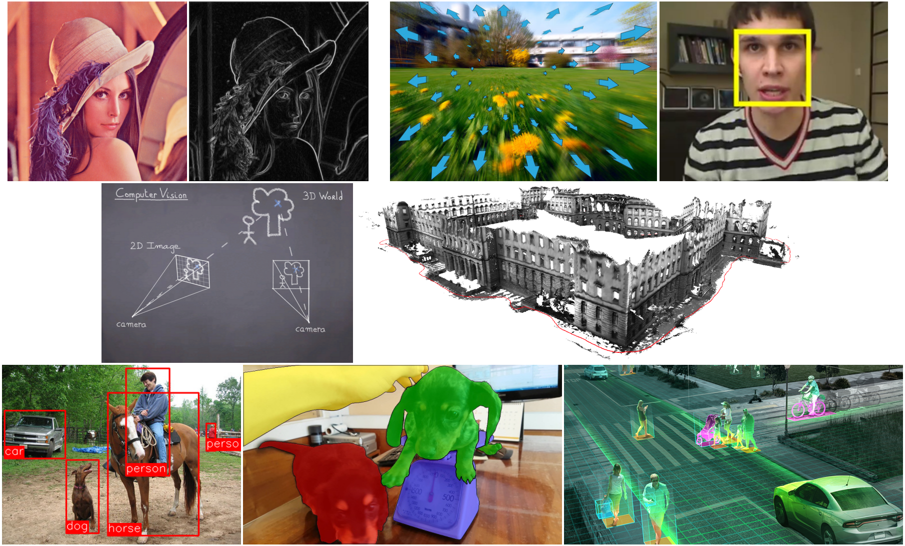
Inria Rennes-Bretagne Atlantique

Rennes, Nov. 2019 – Jan. 2020

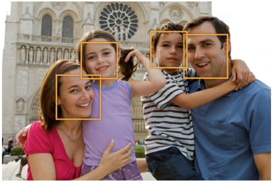
outline

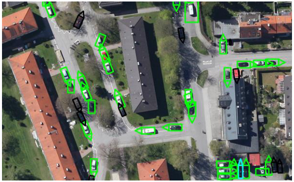
research field

computer vision in images

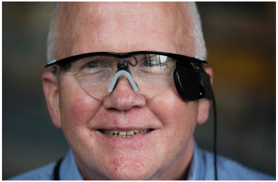
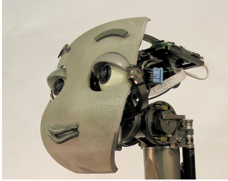


computer vision in images

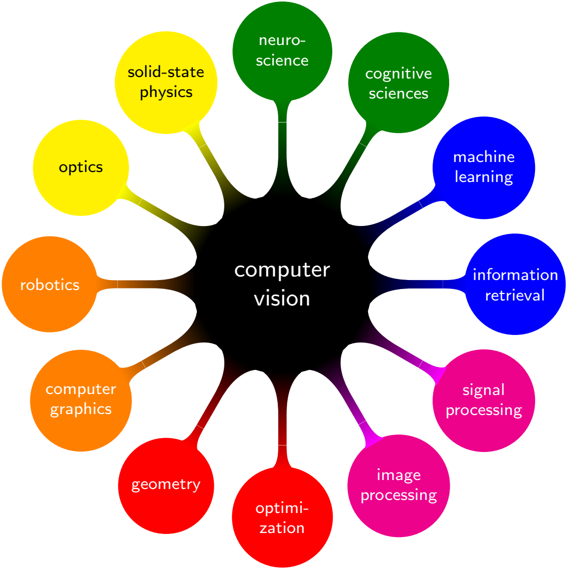




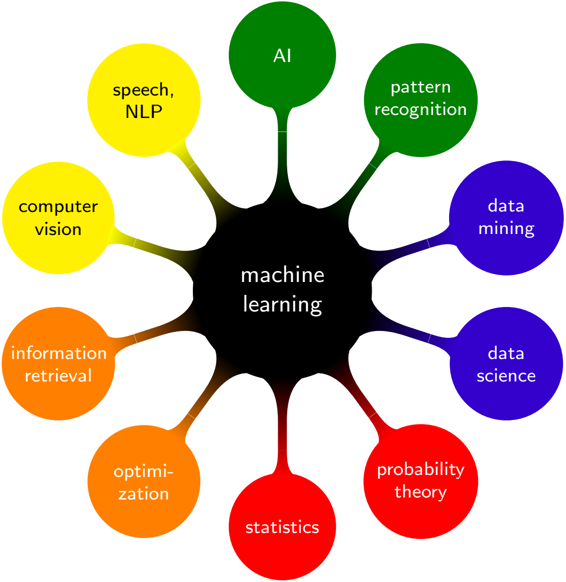


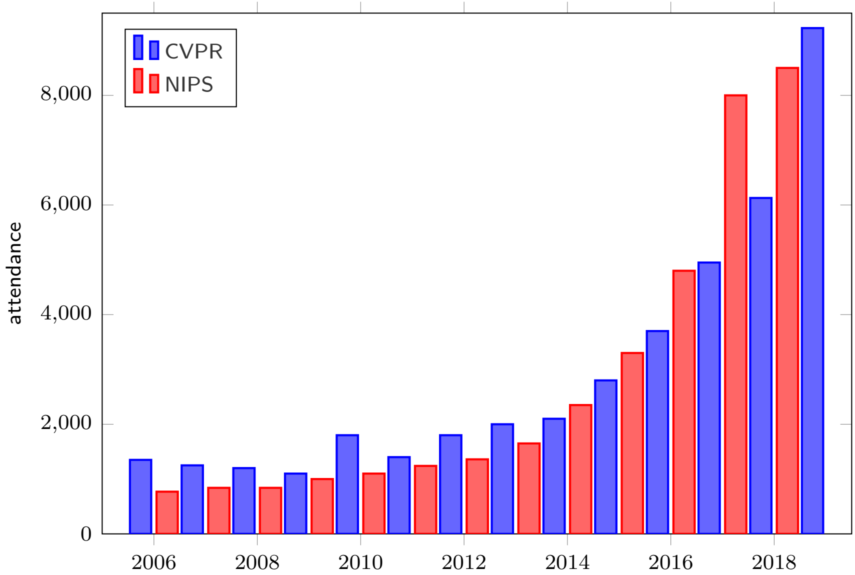
computer vision—related fields



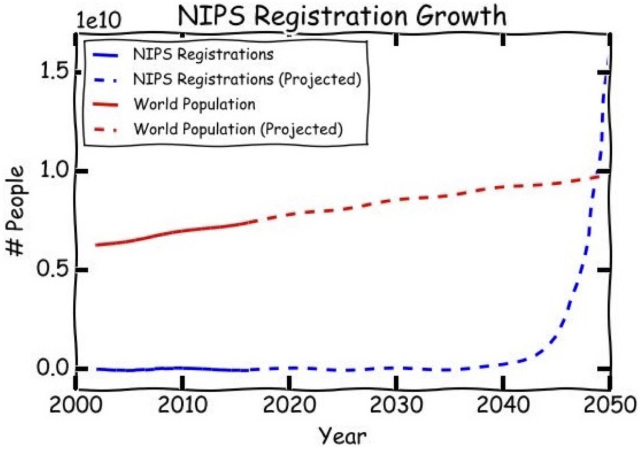
machine learning—related fields



conference attendance growth



really?



NIPS Registration Growth
1e10
NIPS Registrations
NIPS Registrations (Projected)
World Population
World Population (Projected)
People
1.5
1.0
0.5
0.0
2000
2010
2020
2030
2040
2050
Year

CVPR 2019 sponsors

amazon ANDURIL anyvision A BETTER TOMORROW ARGO AI Bai du 百度

ByteDance COGNEX cruise DeepBlue

facebook Google HUAWEI IBM Research

inspur intel AI lyft NAVER | LINE NVIDIA pony.ai

Qualcomm 快手 SAMSUNG Tencent YouTu Lab

TOYOTA RESEARCH INSTITUTE Uber WAYMO wayfair XILINX

Adobe AIBEE Alibaba Group 阿里巴巴集团 Horizon Robotics DiDi

Disney Research HRI Honda Research Institute US dji IIAI Inception Institute of Artificial Intelligence iRobot

JD.COM Lambda MALONG 码隆 Micron

Panasonic PathAI scale SUPER ANNOTATE voyage

ABEJA algolux AMAX appen AQUIFI

Aurora BAOBAB cadence CEVA Cirrascale cloudfactory

CVTE deepen 深睿医疗 deepwise figure eight gumgum GENIUS PROS HITACHI Inspire the Next

HOVER 虎牙直播 isee iMerit JUMIO KINATRAX Kitware

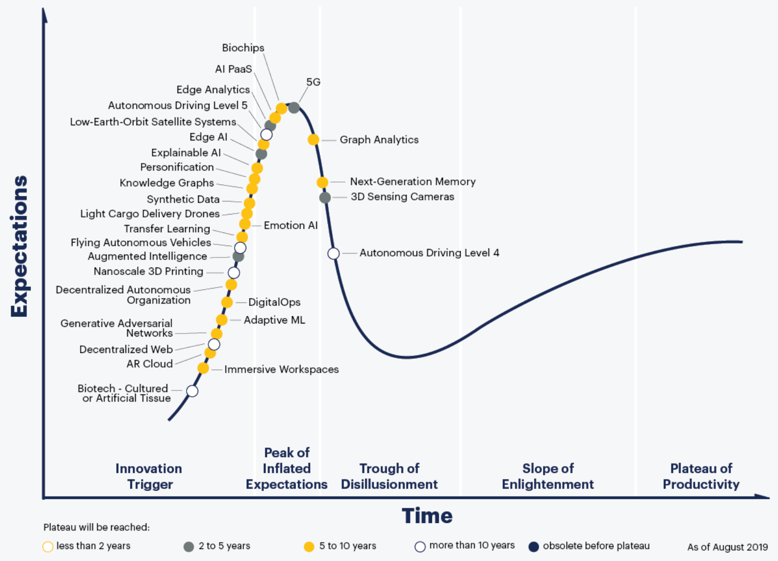
Matroid meitu MTLab MVTec nullmax 纽劢科技 ObjectVideo LABS

Orbital Insight Playment PROPHESEE SAP EMDATA SIEMENS SIGOPT

SK telecom SPORT LOGiQ Tencent AI Lab TESLA Velodyne Lidar Verisk

VinAI VOXEL51 Weights & Biases ZILLOW GROUP ZOOX

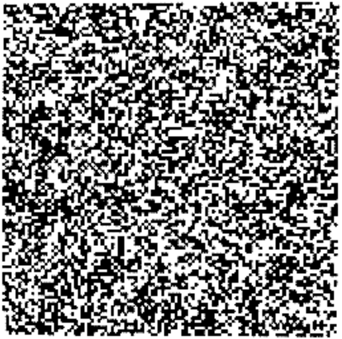
hype cycle

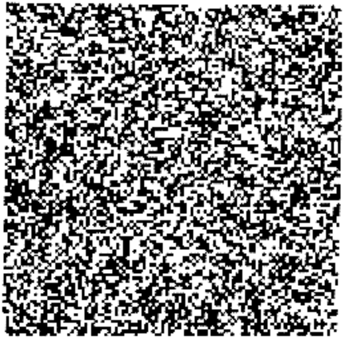


https://www.gartner.com/

psychology and neuroscience background

non-invasive: Béla Julesz

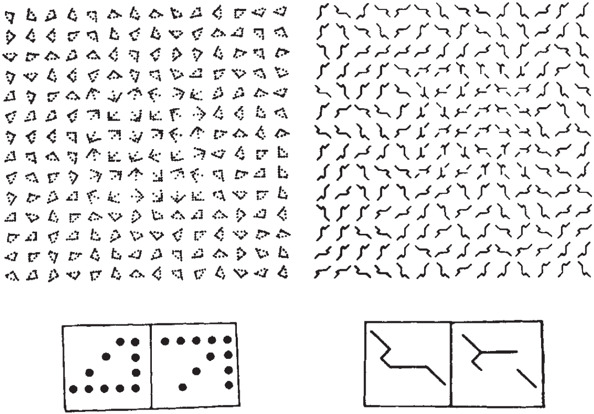




- which happens first? stereopsis or recognition?
- **random dot stereogram**: two identical images, except for a central square region that is displaced randomly in one image
- yields the impression of the square floating over the background

Julesz. BLTJ 1960. Binocular Depth Perception of Computer-Generated Patterns.

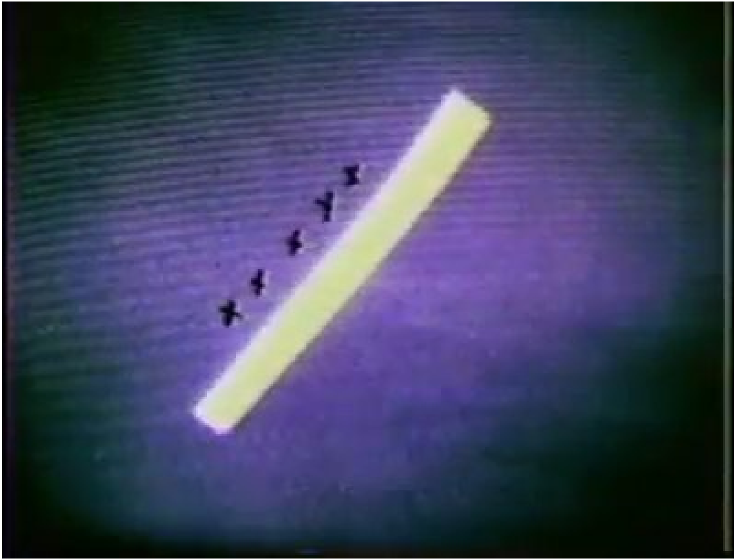
non-invasive: Béla Julesz



- study of pre-attentive (effortless, instantaneous) texture discrimination
- texture pairs with identical second order statistics
- **textons**: "basic elements of pre-attentive human texture perception"

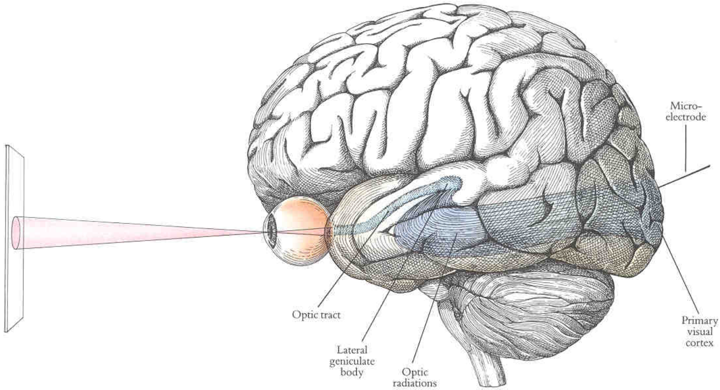
Julesz. N 1981. Textons, the Elements of Texture Perception, and Their Interactions.

invasive: Hubel & Wiesel



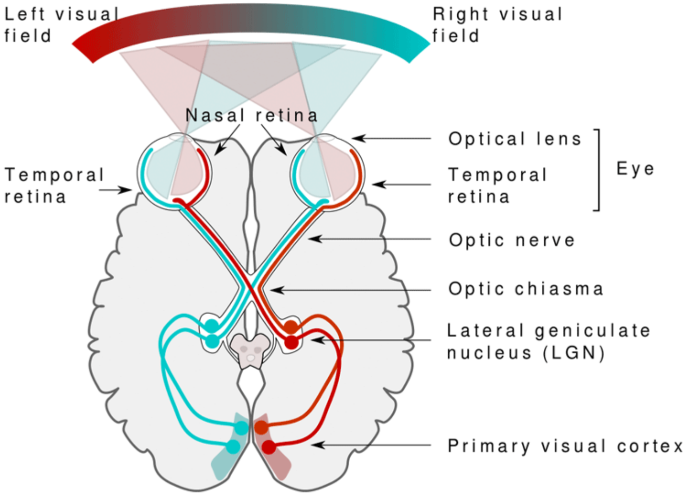
Hubel and Wiesel. JP 1959. Receptive Fields of Single Neurones in the Cat's Striate Cortex.

visual system of mammals



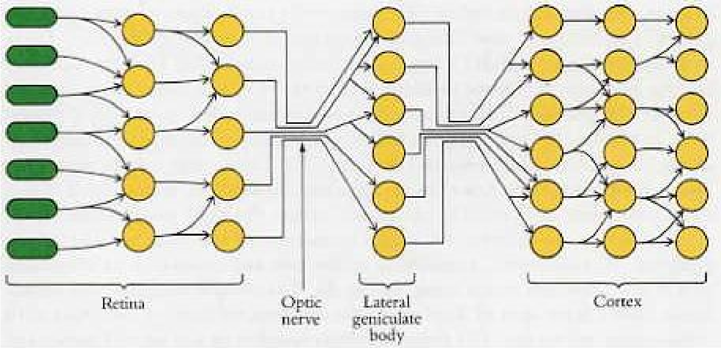
Hubel 1995. Eye, Brain, and Vision.

visual pathway



Hubel 1995. Eye, Brain, and Vision.

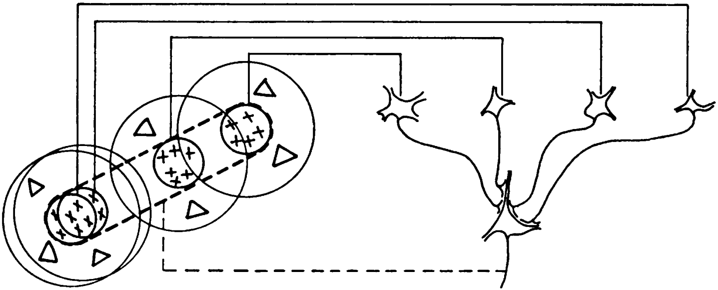
topographic representation



- as you move along the retina, the corresponding points in the cortex trace a continuous path
- each column represents a two-dimensional array of cells

Hubel 1995. Eye, Brain, and Vision.

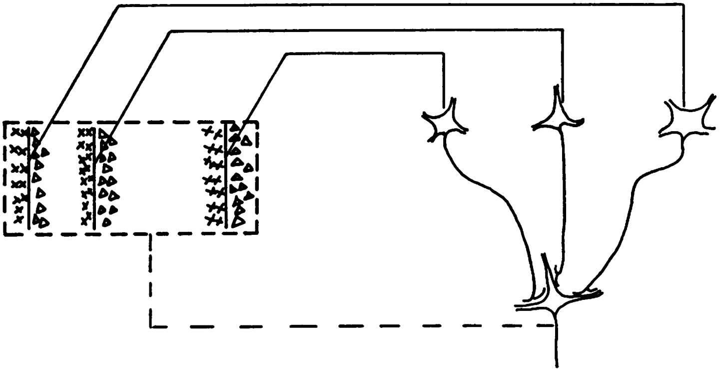
simple cells



- lower-order cells with radially symmetric receptive field with on-center and off-surround
- cells centered along a line with excitatory synaptic connections to a cell of higher order

Hubel and Wiesel. JP 1962. Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex.

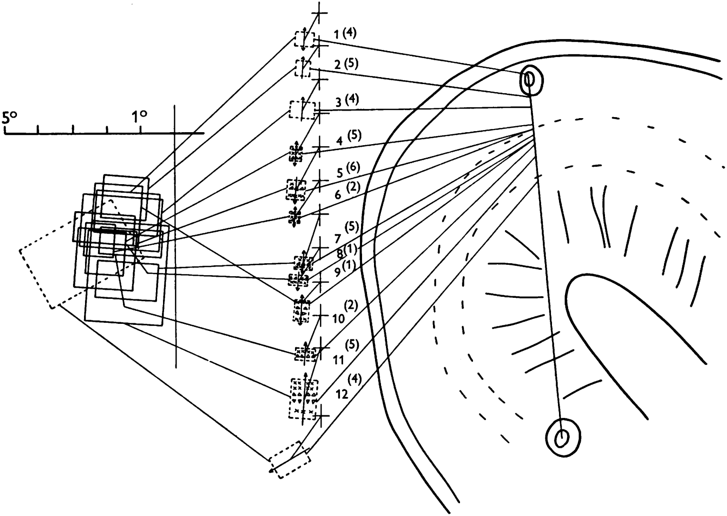
complex cells



- simple cells respond to a vertically oriented edge
- cells scattered throughout a rectangle with excitatory synaptic connections to a complex cell

Hubel and Wiesel. JP 1962. Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex.

electrode recordings



Hubel and Wiesel. JP 1962. Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex.

computer vision background

the summer vision project

[Papert 1966]

"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 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"."

general goals

FIGURE-GROUND

"divide a picture into regions such as likely objects, likely background areas and chaos"

REGION DESCRIPTION

"analysis of shape and surface properties"

OBJECT IDENTIFICATION

"name objects by matching them with a vocabulary of known objects"

Papert 1966. The Summer Vision Project.

specific goals

July

> “non-overlapping objects like balls, bricks, cylinders”
>
> “each face will be of uniform and distinct color and/or texture”
>
> “background will be homogeneous”

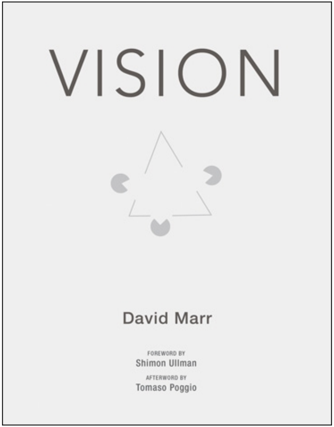
August

> “complex surfaces and background, *e.g.* cigarette pack with writing, or a cylindrical battery”
>
> “objects like tools, cups, *etc.* ”

Papert 1966. The Summer Vision Project.

David Marr, “Vision”

[Marr 1982]



- **biological plausibility**: turning psychology and neuroscience results into models of visual information processing
- **inverse graphics**: from images to surfaces through geometric and photometric models
- **philosophy**: levels of analysis, processing stages, generic principles

edge detection

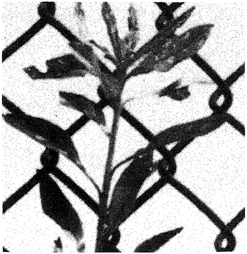
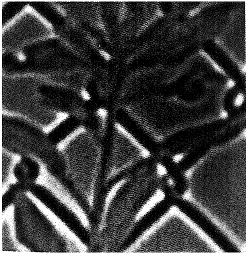
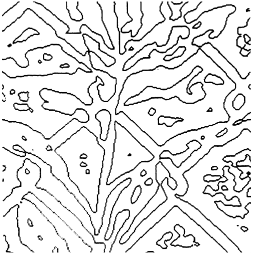


image I



$I * \nabla^2 G$

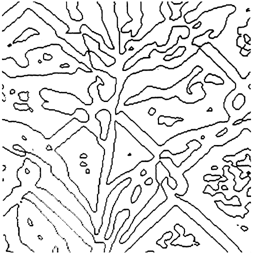
sign



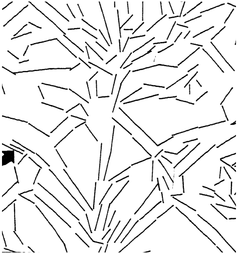
zero crossings

Marr 1982. Vision.

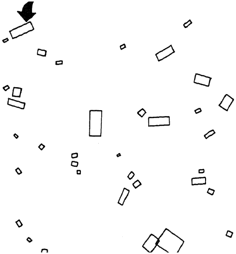
raw primal sketch



zero crossings



edge segments

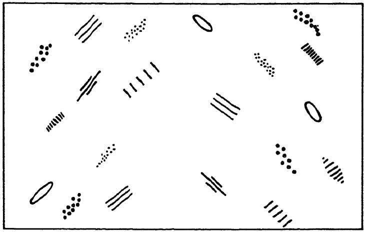


blobs

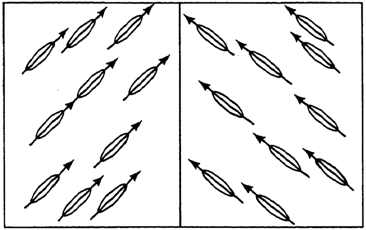
bars

Marr 1982. Vision.

full primal sketch



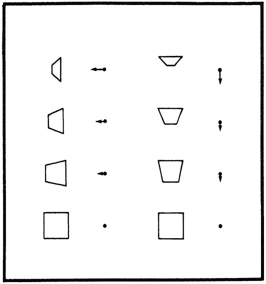
image



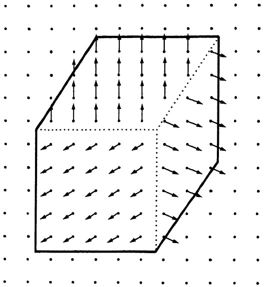
hierarchical grouping of tokens

Marr 1982. Vision.

2.5d sketch



surface orientation

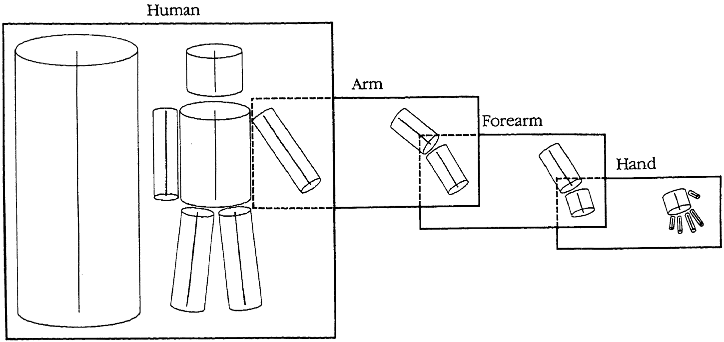


2.5d sketch

- surface orientation (vector field), surface orientation discontinuities (dotted lines), depth discontinuities (continuous lines)
- obtained via stereopsis, optical flow, motion parallax, photometric stereo

Marr 1982. Vision.

3d model representation

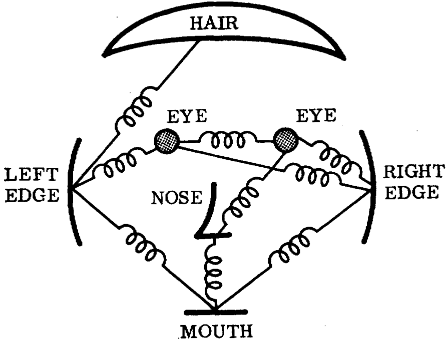


- hierarchical 3d model description
- parts of limited complexity, specified in local coordinate systems
- flexible, allowing for relative part transformation

Marr 1982. Vision.

pictorial structures

[Fischler and Elschlager 1973]

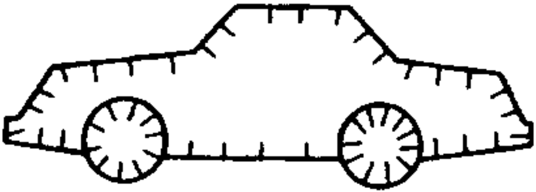


- manually specified object description
- parts-based model: part attributes and pairwise spatial relations
- efficient dynamic programming implementation

Fischler and Elschlager. TC 1973. The Representation and Matching of Pictorial Structures.

generalized Hough transform

[Ballard 1981]

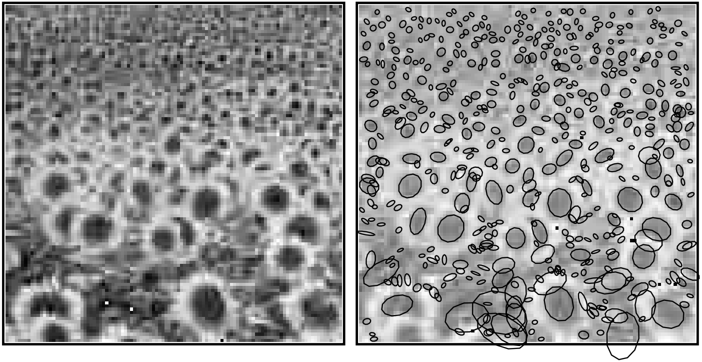


- Hough transform detects analytic curves in parameter space
- generalized version detects arbitrary non-analytic curves
- detection based on a voting process

Ballard. PR 1981. Generalizing the Hough Transform to Detect Arbitrary shapes.

scale selection

[Lindeberg 1993]

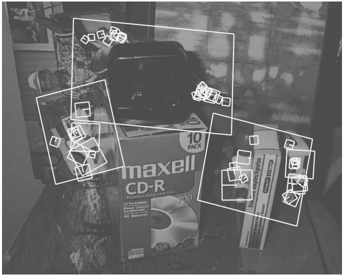


- scale-space and scale-normalized derivatives
- automatic scale selection at local maxima over scale
- applies to blobs, junctions, corners, edges or ridges

Lindeberg. SCIA 1993. On Scale Selection for Differential Operators.

scale-invariant feature transform (SIFT)

[Lowe 1999]

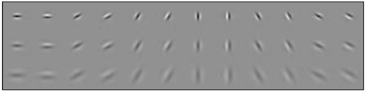


- scale selection by difference of Gaussians (DoG)
- orientation assignment, local descriptor
- Hough transform on affine space

Lowe. ICCV 1999. Object recognition from local scale-invariant features.

textons

[Malik et al. 1999]



oriented filter bank

image

texture segmentation

- textons defined as clusters of filter responses
- regions described by texton histograms

Malik, Belongie, Shi and Leung. ICCV 1999. Textons, Contours and Regions: Cue Integration in Image Segmentation.

real-time face detection

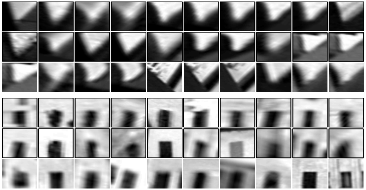
[Viola and Jones 2001]

- simple rectangle features in constant time on integral images
- learning weak classifiers by boosting
- classifier cascade provides a focus-of-attention mechanism

Viola and Jones. CVPR 2001. Rapid Object Detection Using a Boosted Cascade of Simple Features.

bag of words

[Sivic and Zisserman 2003]



visual vocabulary

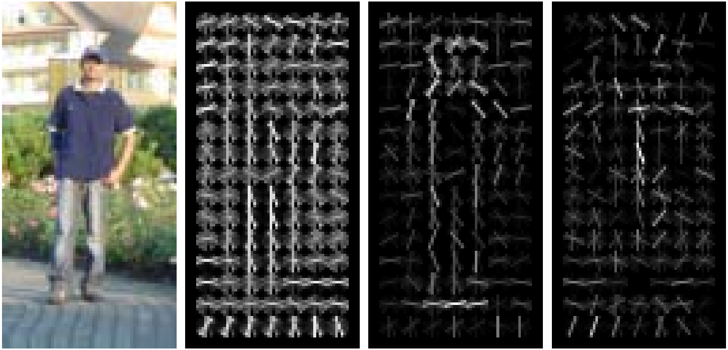
video retrieval

- “visual words” defined as clusters of SIFT descriptors
- images described by visual word histograms
- text retrieval methods applied to video retrieval

Sivic and Zisserman. ICCV 2003. Video Google: A Text Retrieval Approach to Object Matching in videos.

histogram of oriented gradients (HOG)

[Dalal and Triggs 2005]

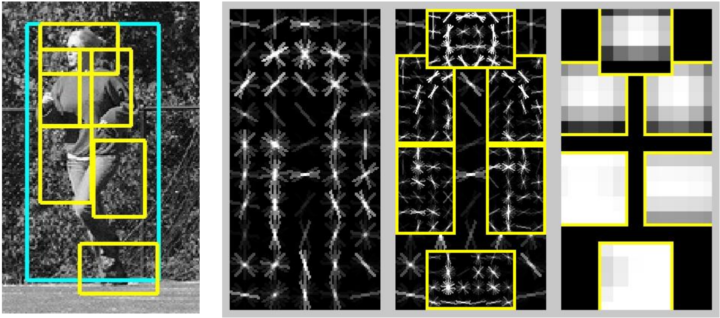


- dense, SIFT-like descriptors
- SVM classifier
- sliding window detection at all positions and scales

Dalal and Triggs. CVPR 2005. Histograms of Oriented Gradients for Human Detection.

deformable part model (DPM)

[Felzenszwalb et al. 2008]



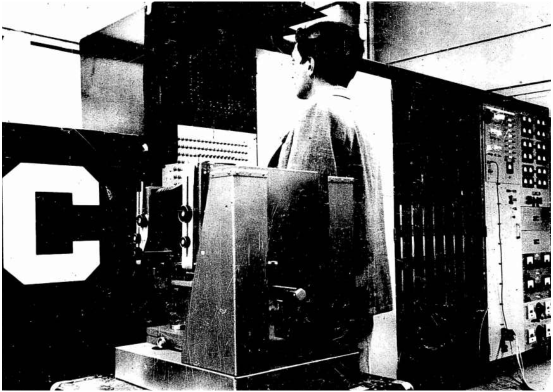
- appearance represented by HOG
- spatial configuration inspired by "pictorial structures"
- part locations treated as latent variables

Felzenszwalb, Mcallester and Ramanan. CVPR 2008. A Discriminatively Trained, Multiscale, Deformable Part Model.

machine learning background

perceptron

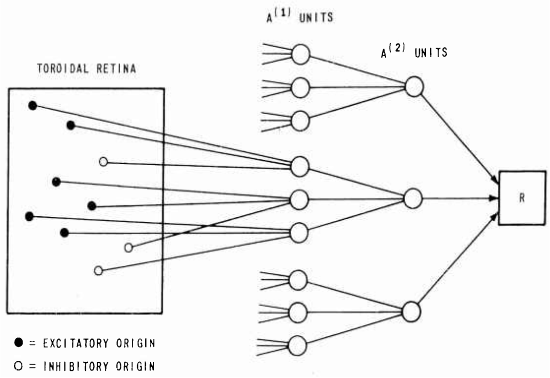
[Rosenblatt 1962]



- Mark-I perceptron
- analog circuit implementation; parameters as potentiometers

Rosenblatt 1962. Principles of Neurodynamics

perceptron



- early forms of multi-layer networks, continuous activation functions, back-propagating errors, convolution, skip connections, recurrent networks, selective attention, program learning, and multi-modality

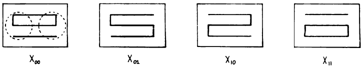
Rosenblatt 1962. Principles of Neurodynamics

perceptron

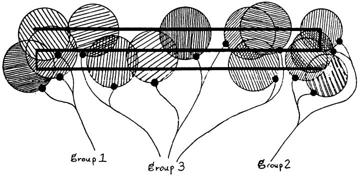
[Minsky and Papert 1969]

Theorem 0.8: No diameter-limited perceptron can determine whether or not all the parts of any geometric figure are connected to one another! That is, no such perceptron computes $\psi_{\text{CONNECTED}}$.

The proof requires us to consider just four figures



and a diameter-limited perceptron ψ whose support sets have diameters like those indicated by the circles below:



- (re-)define perceptron as a linear classifier
- then prove a series of negative results
- "AI winter" follows; misconception remains until today

Minsky and Papert 1969. Perceptrons: an Introduction to Computational Geometry.

automatic differentiation

[Werbos 1974]

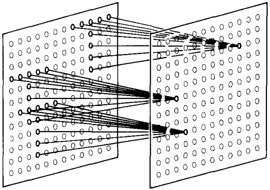
Actual Variable	Variable Number	Operation Category	Major Source	Minor Source
$(b(2))^2$	20	product	19	19
$b(2)=C(2)-k_1Y_p(2)$	19	difference	18	17
$C(2)$	18	input	-	-
$k_1Y_p(2)$	17	product	16	1
$Y_p(2)$	16	sum	15	13
$k_2Y_A(2)$	15	product	14	2
$Y_A(2)$	14	input	-	-
$(1-k_2)Y_p(1)$	13	product	12	4
$(b(1))^2$	12	product	11	11
$b(1)=C(1)-k_1Y_p(1)$	11	difference	10	9
$C(1)$	10	input	-	-
$k_1Y_p(1)$	9	product	8	1

- formulate an arbitrary function as a computational graph
- **dynamic feedback**: compute symbolic derivatives by dynamic programming

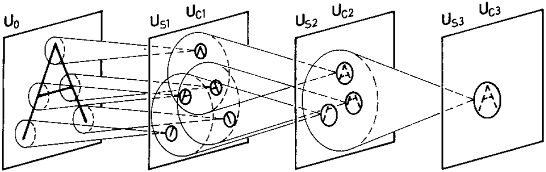
Werbos 1974. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.

neocognitron

[Fukushima 1980]



convolution



feature hierarchy

- biologically-inspired convolutional network
- unsupervised learning

Fukushima. BC 1980. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected By Shift in Position.

back-propagation

[Rumelhart et al. 1986]

The backward pass starts by computing $\partial E/\partial y$ for each of the output units. Differentiating equation (3) for a particular case, c, and suppressing the index c gives

$$\partial E/\partial y_j = y_j - d_j \tag{4}$$

We can then apply the chain rule to compute $\partial E/\partial x_j$

$$\partial E/\partial x_j = \partial E/\partial y_j \cdot \mathrm{d}y_j/\mathrm{d}x_j$$

Differentiating equation (2) to get the value of $\mathrm{d}y_j/\mathrm{d}x_j$ and substituting gives

$$\partial E/\partial x_j = \partial E/\partial y_j \cdot y_j(1 - y_j) \tag{5}$$

This means that we know how a change in the total input x to an output unit will affect the error. But this total input is just a linear function of the states of the lower level units and it is also a linear function of the weights on the connections, so it is easy to compute how the error will be affected by changing these states and weights. For a weight w_{ji}, from i to j the derivative is

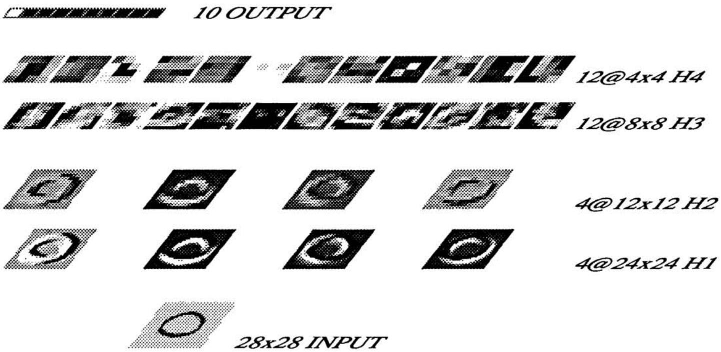
$$\begin{aligned} \partial E/\partial w_{ji} &= \partial E/\partial x_j \cdot \partial x_j/\partial w_{ji} \\ &= \partial E/\partial x_j \cdot y_i \end{aligned} \tag{6}$$

- introduce back-propagation in multi-layer networks with sigmoid nonlinearities and sum of squares loss function
- advocate batch gradient descent for supervised learning
- discuss online gradient descent, momentum and random initialization

Rumelhart, Hinton and Williams. N 1986. Learning Representations By Back-Propagating Errors.

convolutional networks

[LeCun et al. 1990]

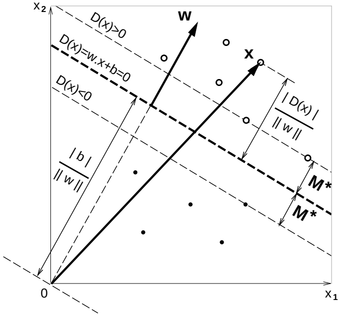


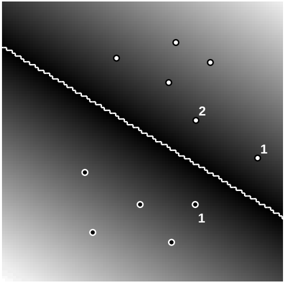
- train a convolutional network by back-propagation
- advocate end-to-end feature learning for image classification

LeCun, Boser, Denker *et al* . NIPS 1990. Handwritten Digit Recognition with a Back-Propagation Network.

support vector machines

[Boser et al. 1992]



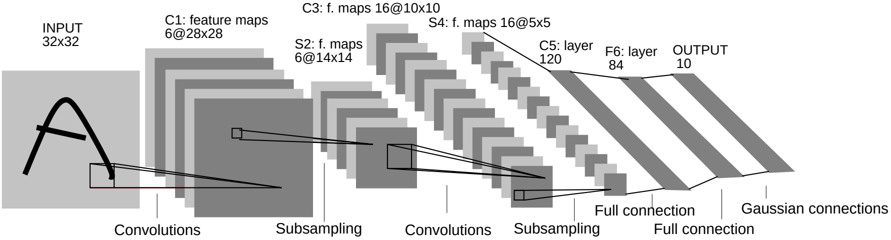


- linear classifier, made nonlinear via kernel trick
- convex optimization
- back to raw inputs; hand-crafted kernel functions
- shift focus from neural networks to kernel methods

Boser, Guyon and Vapnik. COLT 1992. A Training Algorithm for Optimal Margin Classifiers.

LeNet-5

[LeCun et al. 1998]



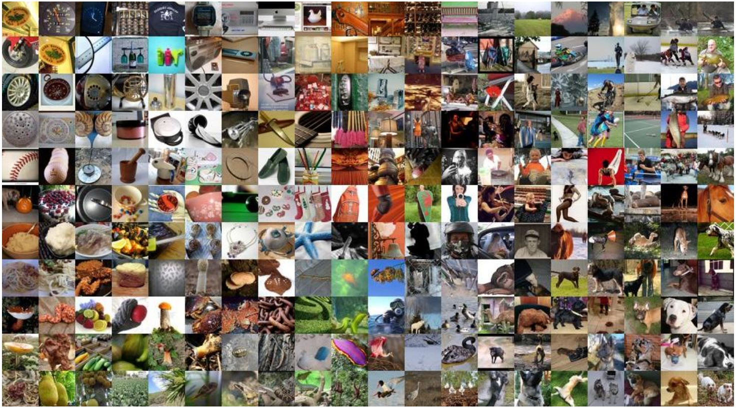
- sub-sampling gradually introduces translation, scale and distortion invariance
- non-linearity included in sub-sampling layers as feature maps are increasing in dimension

Lecun, Bottou, Bengio, Haffner. IEEE Proc. 1998. Gradient-Based Learning Applied to Document Recognition.

modern deep learning

ImageNet

[Russakovsky et al. 2014]

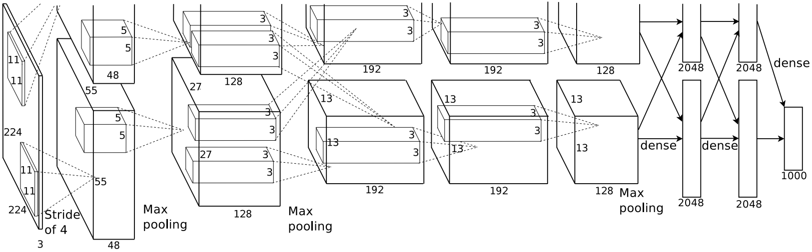


- 22k classes, 15M samples
- ImageNet Large-Scale Visual Recognition Challenge (ILSVRC): 1000 classes, 1.2M training images, 50k validation images, 150k test images

Russakovsky, Deng, Su, Krause, *et al.* 2014. Imagenet Large Scale Visual Recognition Challenge.

AlexNet

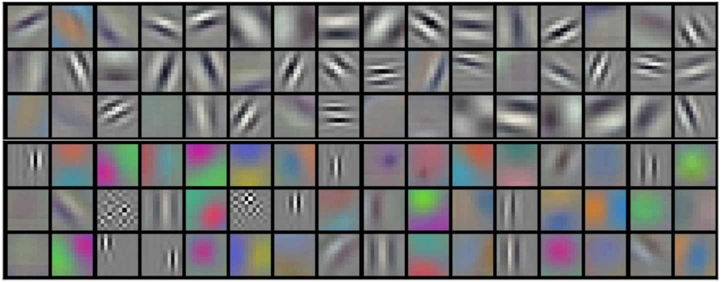
[Krizhevsky et al. 2012]



- implementation on two GPUs; connectivity between the two subnetworks is limited
- ReLU, data augmentation, local response normalization, dropout
- outperformed all previous models on ILSVRC by 10%

Krizhevsky, Sutskever, Hinton. NIPS 2012. Imagenet Classification with Deep Convolutional Neural Networks.

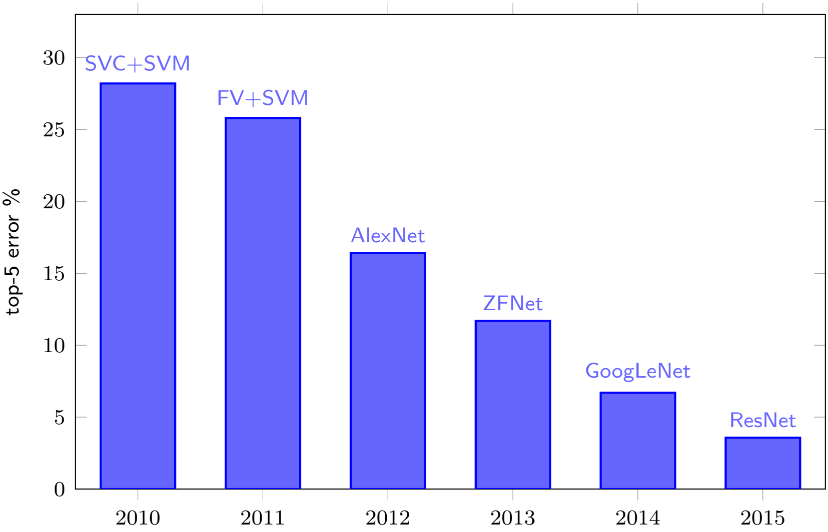
learned layer 1 kernels



- 96 kernels of size $11 \times 11 \times 3$
- top: 48 GPU 1 kernels; bottom: 48 GPU 2 kernels

Krizhevsky, Sutskever, Hinton. NIPS 2012. Imagenet Classification with Deep Convolutional Neural Networks.

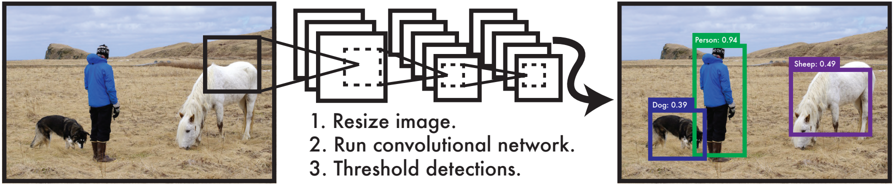
ImageNet classification performance



Russakovsky, Deng, Su, Krause, *et al.* 2014. Imagenet Large Scale Visual Recognition Challenge.

object detection

[Redmon et al. 2016]

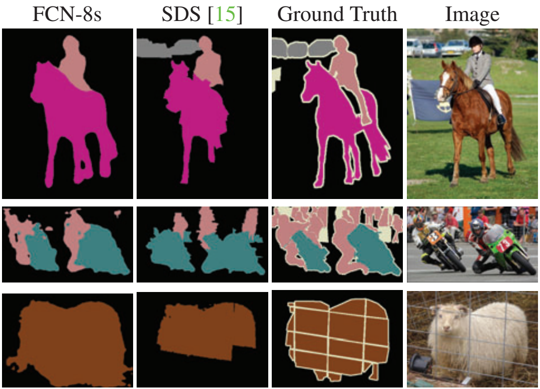


- learn to detect objects as a single classification and regression task, without scanning the image or detecting candidate regions
- first object detector to operate at 45fps

Redmon, Divvala, Girshick, Farhadi. CVPR 2016. You Only Look Once: Unified, Real-Time Object Detection.

semantic segmentation

[Long et al. 2015]



- learn to upsample
- apply to pixel-dense prediction tasks

Long, Shelhamer, Darrell. CVPR 2015. Fully Convolutional Networks for Semantic Segmentation.

instance segmentation and pose estimation

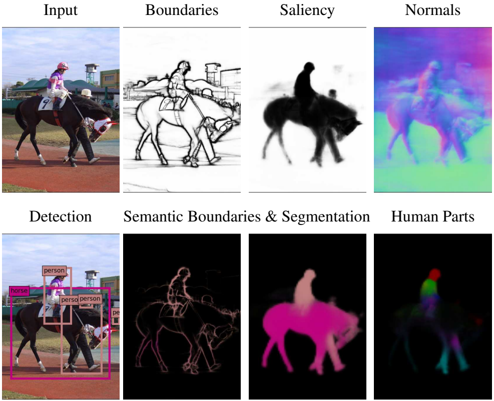
[He et al. 2017]

- semantic segmentation per detected region
- pose estimation as regression

He, Gkioxari, Dollar, Girshick. ICCV 2017. Mask R-CNN.

multi-task learning

[Kokkinos 2017]

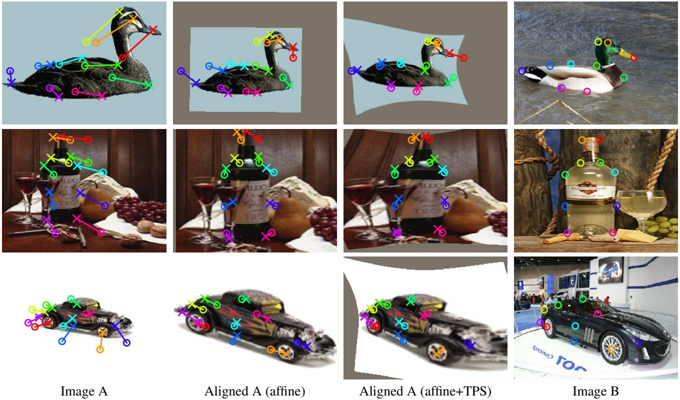


- learn several vision tasks with a joint network architecture including task-specific skip layers

Kokkinos. CVPR 2017. Ubernet: Training a Universal Convolutional Neural Network for Low-, Mid-, and High-Level Vision Using Diverse Datasets and Limited Memory.

geometric matching

[Rocco et al. 2017]

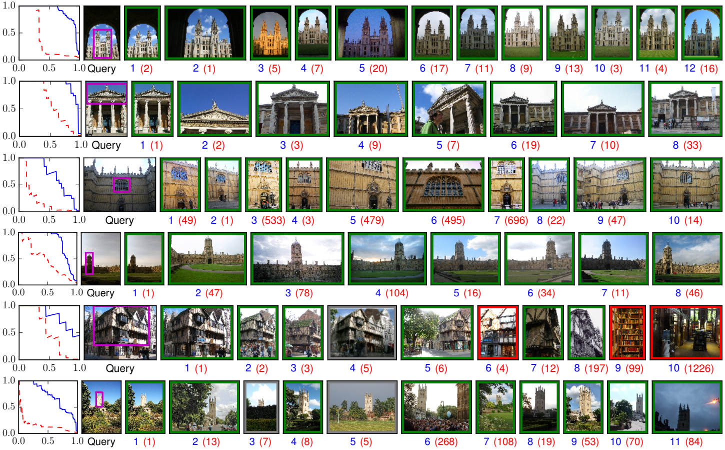


- mimic the standard steps of feature extraction, matching and simultaneous inlier detection and model parameter estimation
- still trainable end-to-end

Rocco, Arandjelovic, Sivic. CVPR 2017. Convolutional Neural Network Architecture for Geometric Matching.

image retrieval

[Gordo et al. 2016]

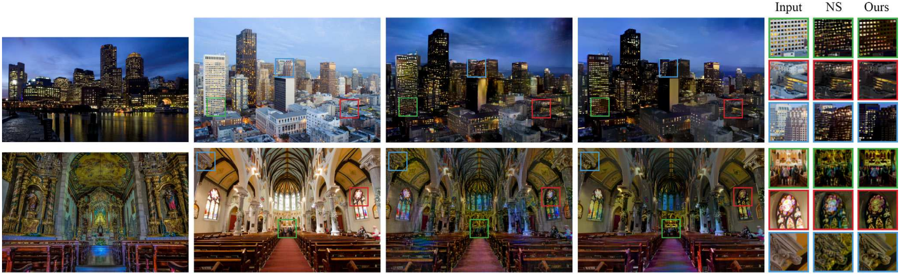


- learn to match
- apply as generic feature extractor

Gordo, Almazan, Revaud, Larlus. ECCV 2016. Deep Image Retrieval: Learning Global Representations for Image Search.

photorealistic style transfer

[Luan et al. 2017]



(a) Reference style image (b) Input image (c) Neural Style (distortions) (d) Our result (e) Insets

- generate same scene as input image
- transfer style from reference image
- photorealism regularization

Luan, Paris, Shechtman, Bala. CVPR 2017. Deep Photo Style Transfer.

image captioning

[Vinyals et al. 2017]

- image description by deep CNN
- language generation by RNN

Vinyals, Toshev, Bengio and Erhan. PAMI 2017. Show and Tell: Lessons Learned From the 2015 MSCOCO Image Captioning Challenge.

about this course

logistics

- **course website**: `https://sif-dlv.github.io/`
- **piazza**: `https://piazza.com/inria.fr/fall2019/dlv`

prerequisites

basic knowledge of

- linear algebra
- calculus
- probabilities
- signal processing
- machine learning
- python

goals

- discuss well-known methods from low-level description to intermediate representation, and their dependence on the end task
- study a data-driven approach where the entire pipeline is optimized jointly in a supervised fashion, according to a task-dependent objective
- study deep learning models in detail
- interpret them in connection to conventional models
- focus on recent, state of the art methods and large scale applications

conventional methods

- **representation**: global/local visual descriptors, dense/sparse representation, feature detectors; encoding/pooling, vocabularies, bag-of-words; VLAD*, Fisher vectors*, embeddings*, HMAX*
- **local features and spatial matching**: derivatives, scale space and scale selection; edges, blobs, corners/junctions; dense optical flow / sparse feature tracking*; wide-baseline matching; geometric models, RANSAC, Hough transform; fast spatial matching*
- **codebooks and kernels**: geometry/appearance matching; bag-of-words; k-means clustering, hierarchical*, approximate*, vocabulary tree*; soft assignment, max pooling; match kernels, hamming embedding, ASMK*; pyramid matching, spatial pyramids, Hough pyramids*.

deep learning approach (1)

- **learning**: binary classification; perceptron, support vector machines, logistic regression; gradient descent, regularization, loss functions, unified model; multi-class classification; linear regression*, basis functions; neural networks, activation functions
- **differentiation**: stochastic gradient descent; numerical gradient approximation; function decomposition, chain rule, analytical gradient computation, back-propagation; chaining, splitting and sharing; common forward and backward flow patterns; dynamic automatic differentiation*

deep learning approach (2)

- **convolution**: convolution, cross-correlation, linearity, equivariance, weight sharing; feature maps, matrix multiplication, 1×1 convolution; padded, strided, dilated convolution; pooling and invariance; convolutional networks: LeNet-5, AlexNet, ZFNet*, VGG, NiN*, GoogLeNet.
- **optimization and deeper architectures**: optimizers: momentum, RMSprop, Adam, second-order*; initialization: Gaussian matrices, unit variance, orthogonal*, data-dependent*; normalization: input, batch, layer*, weight*; deeper networks: residual, identity mappings*, stochastic depth*, densely connected

applications

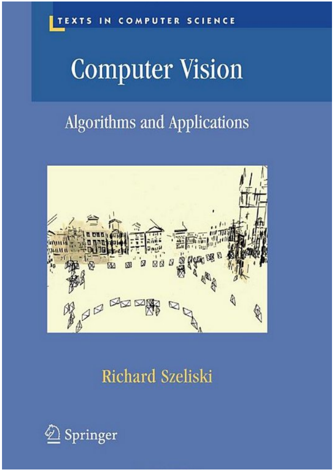
- **object detection**: background: Viola and Jones, DPM, ISM, ESS, object proposals, non-maximum suppression; two-stage: R-CNN, SPP, fast/faster R-CNN, RPN; bounding box regression; part-based: R-FCN, spatial transformers*, deformable convolution; upsampling*: FCN, feature pyramids; one-stage: OverFeat*, YOLO, SSD*, RetinaNet*, focal loss
- **retrieval**: local/global descriptors; pooling from CNN representations: MAC, R-MAC, SPoC*, CroW*; manifold learning, siamese and triplet architectures; fine-tuning: constrastive/triplet loss, learning to rank; graph-based methods, diffusion, unsupervised fine-tuning

related courses at sif

- ADM Advanced Probabilistic Data Analysis and Modeling (Guillaume Gravier)
- BSI Big Data Storage and Processing Infrastructures (Gabriel Antoniu)
- CG Computer Graphics: Rendering and Modeling 3D Scenes (Rémi Cozot)
- CV Computer Vision (Eric Marchand)
- DMV Data Mining and Visualization (Alexandre Termier)
- GDP Graph Data Processing (Pierre Vandergheynst)
- HDL High-Dimensional Statistical Learning (Rémi Gribonval)
- REP Image Representation, Editing and Perception (Olivier Le Meur)
- SML Supervised Machine Learning (François Coste)

computer vision: algorithms and applications

http://szeliski.org/Book/



1 introduction

3 image processing

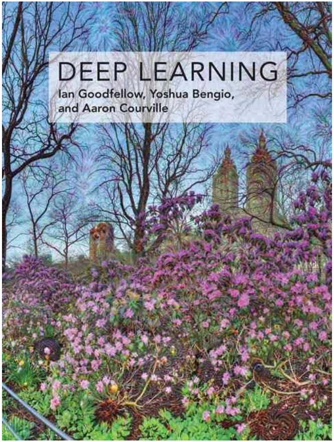
4 feature detection and matching

6 feature-based alignment

14 recognition

deep learning book

http://www.deeplearningbook.org/



1 introduction

5 machine learning basics

6 deep feedforward networks

7 regularizaton for deep learning

8 optimization for training deep models

9 convolutional networks

11 practical methodology

evaluation

- oral presentation: 50%
- written exam: 50%

oral presentations

- teams of two
- instructions, paper list: `https://sif-dlv.github.io/oral`
- choose 2-5 papers, report your choice by mid-December
- should be interesting but not too hard
- study and find more related work; find connections
- present on second half of January
- focus presentation on ideas; not too detailed
- 8 min/talk, 4 min questions: total 20 min/team
- the class is your audience
- ask questions!

good luck!