Bases de données multimédia
IX – Machine à vecteurs de support et CNN

ENSIMAG

2014-2015

Matthijs Douze & Karteek Alahari

Plan

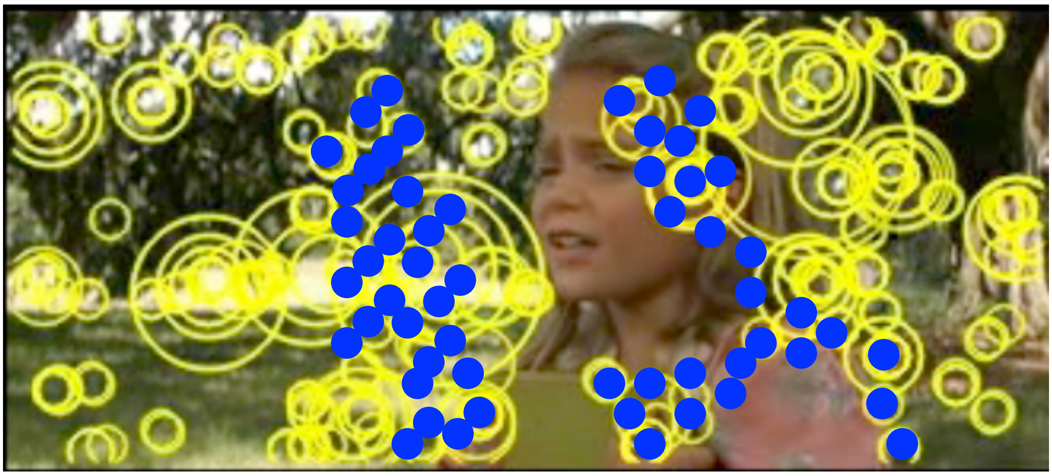
- Classification
- Machine à vecteurs de support
- Reconnaissance d'actions
- CNN
- Quelques perspectives

Classification

- What can we do, given all these features ?

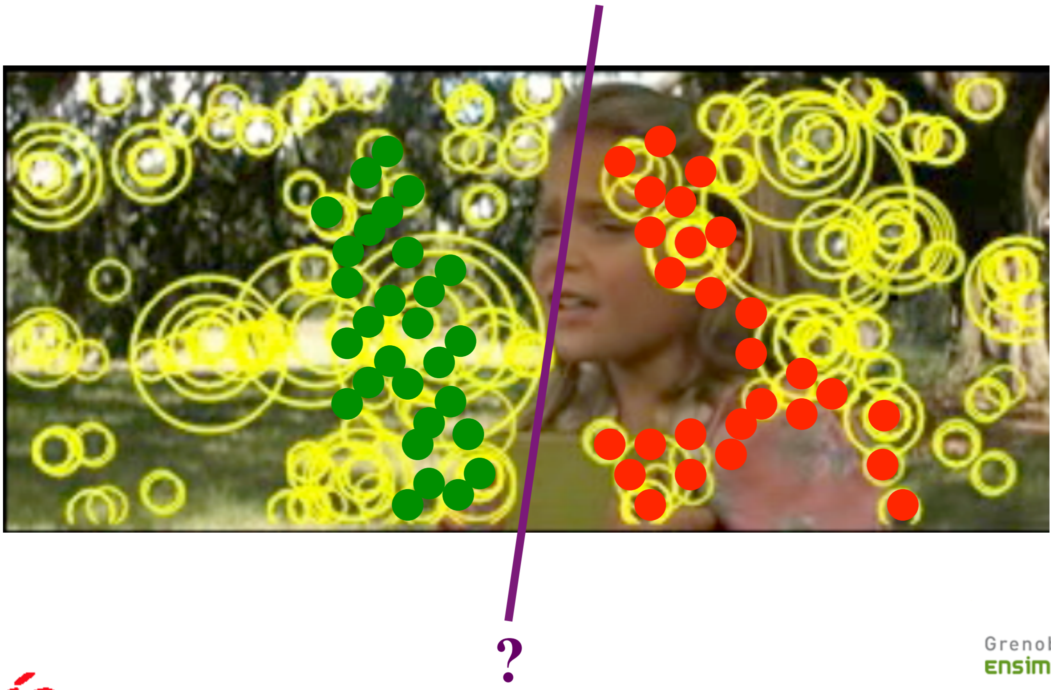
Classification

- What can we do, given all these features ?

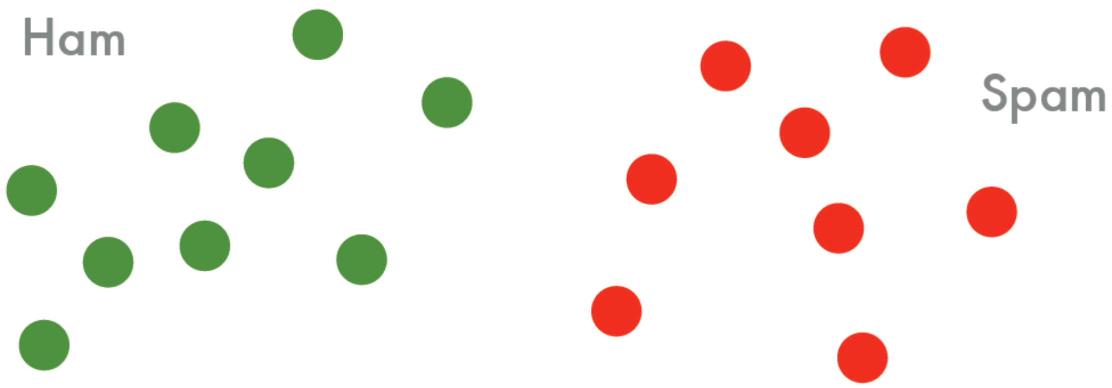


Classification

- What can we do, given all these features ?

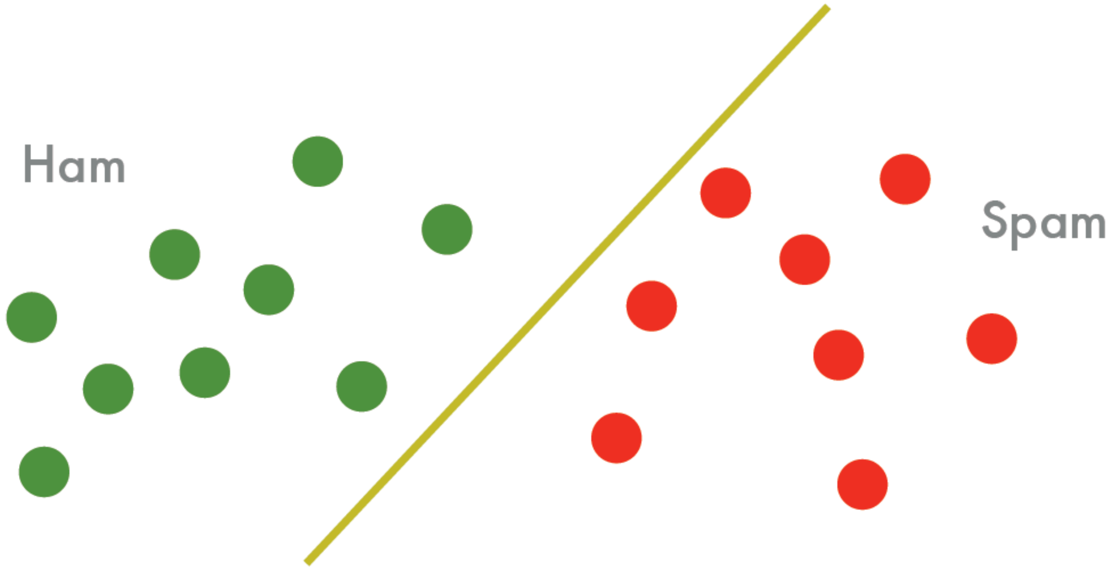


Classification

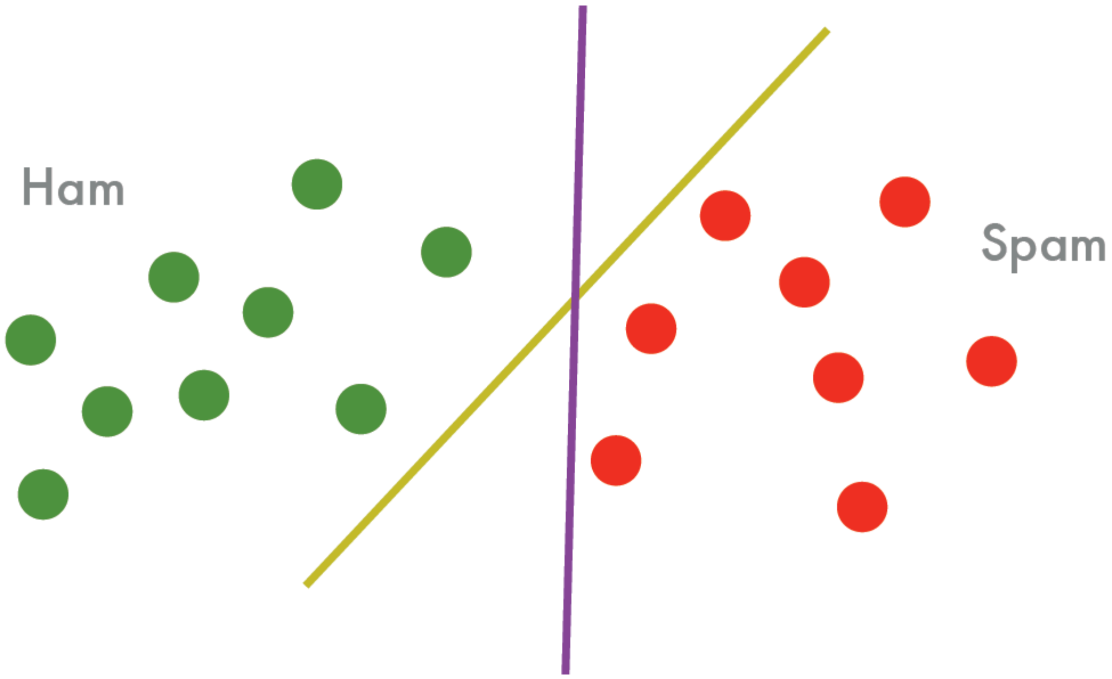


Slides courtesy Alex Smola

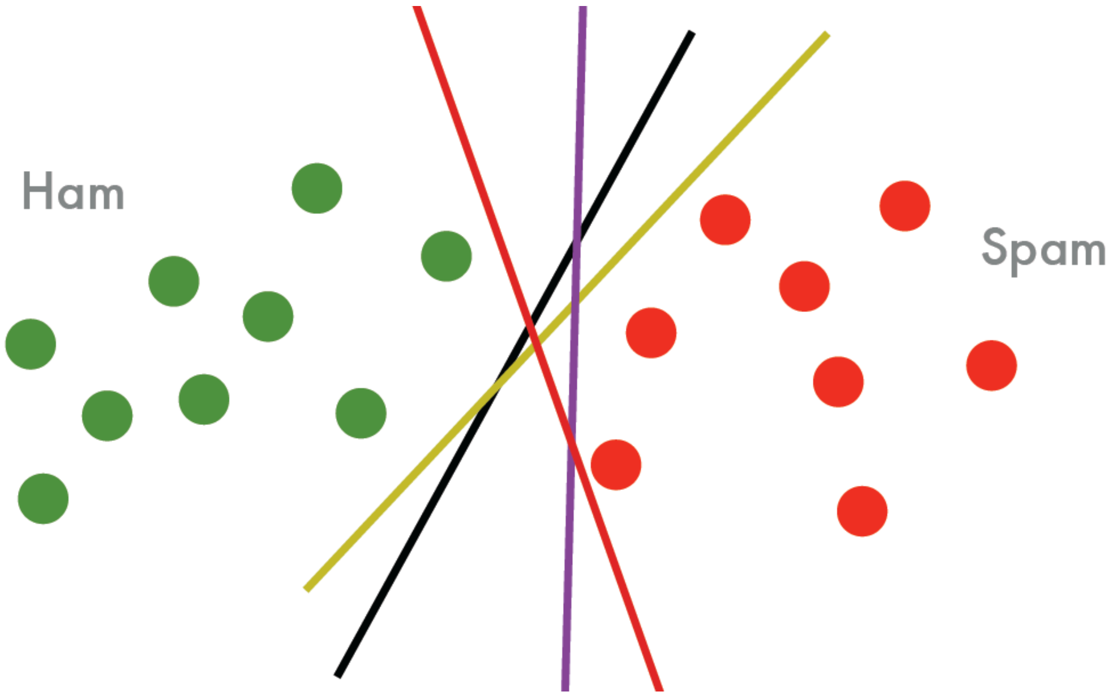
Classification



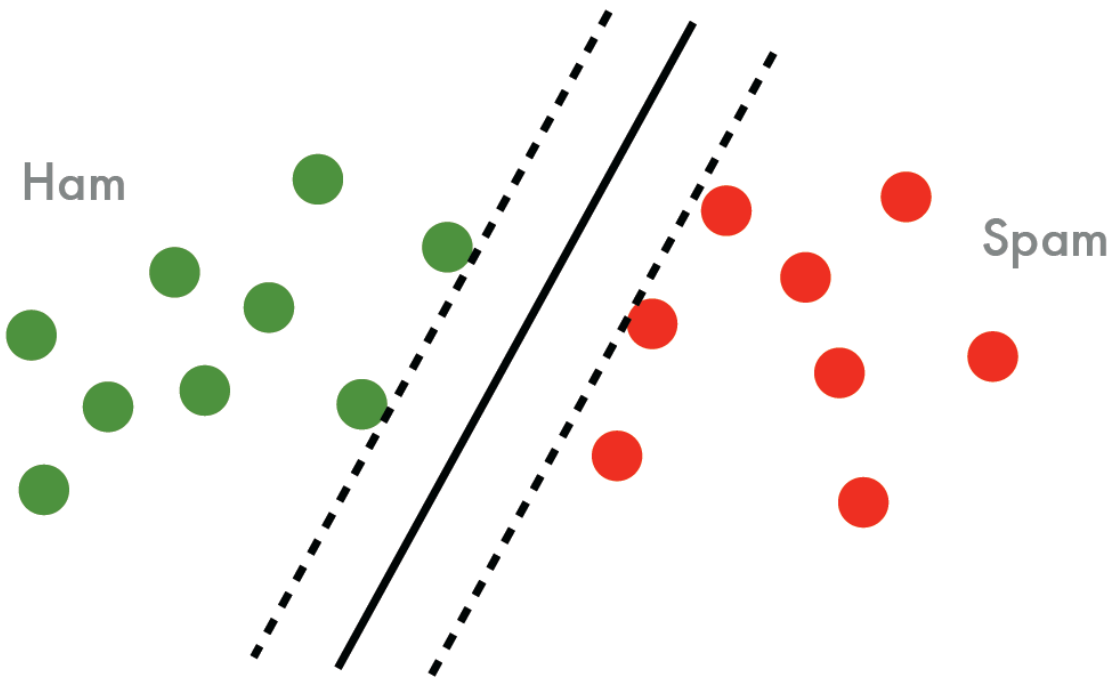
Classification



Classification

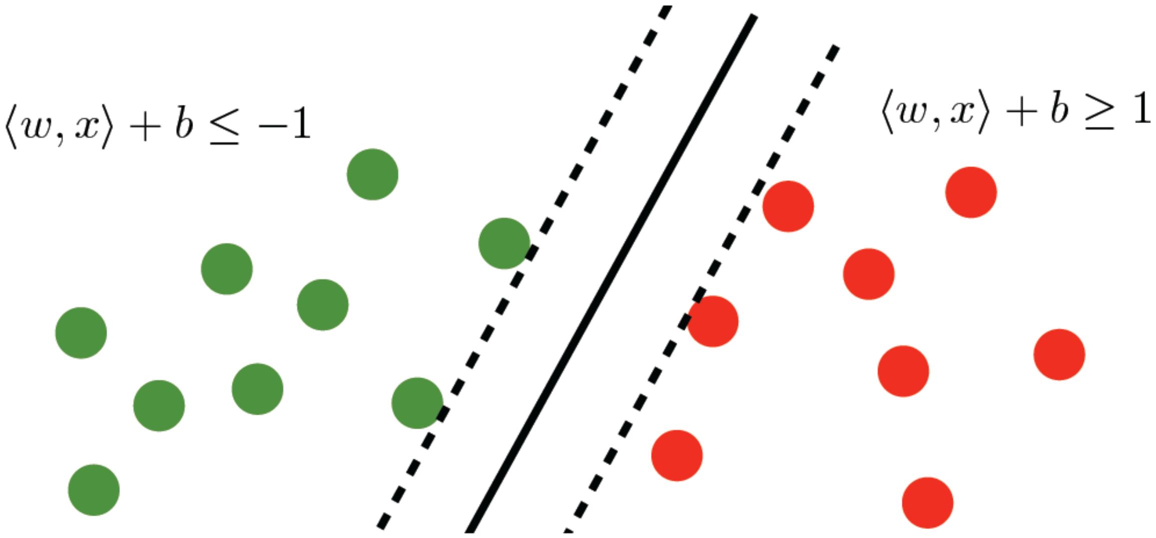


Classification



Large margin classifier

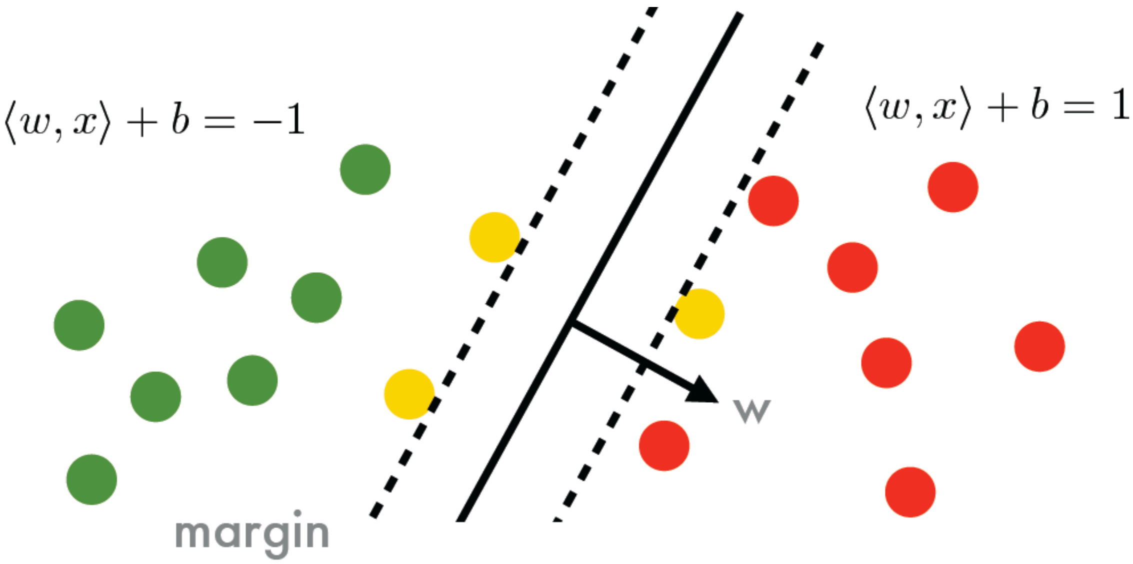
Classification



linear function

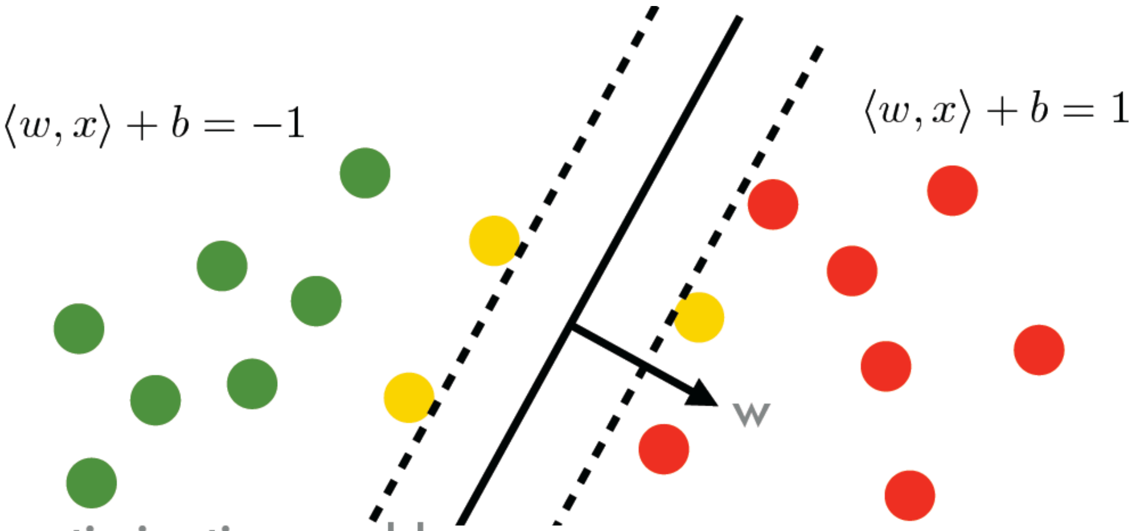
$$f(x) = \langle w, x \rangle + b$$

Classification



$$\frac{\langle x_+ - x_-, w \rangle}{2\,\|w\|} = \frac{1}{2\,\|w\|}\left[\left[\langle x_+, w \rangle + b\right] - \left[\langle x_-, w \rangle + b\right]\right] = \frac{1}{\|w\|}$$

Classification



optimization problem

$$\underset{w,b}{\text{maximize}} \frac{1}{\|w\|} \text{ subject to } y_i \left[\langle x_i, w\rangle + b\right] \geq 1$$

Support Vector Machines

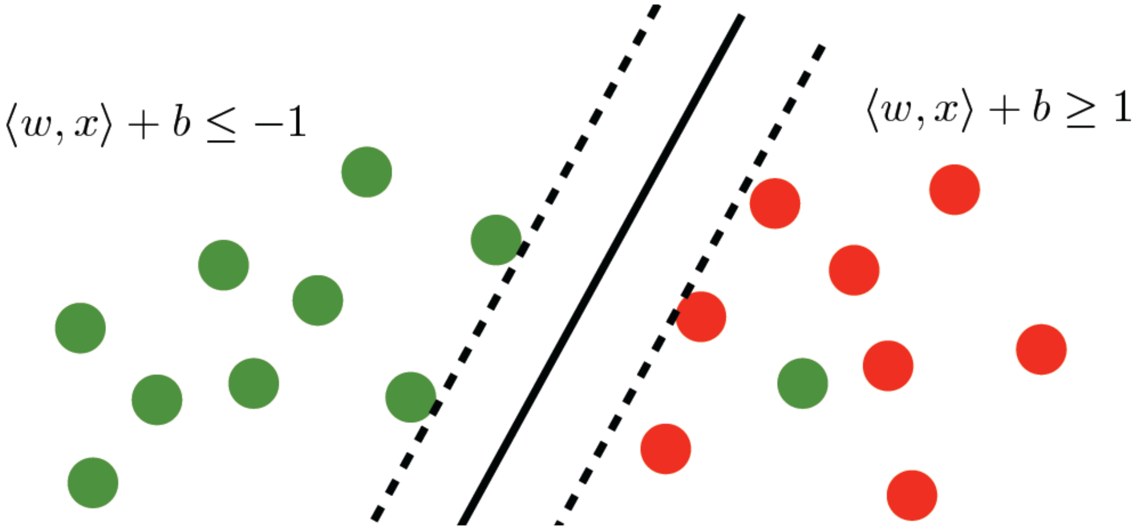
$$\underset{w,b}{\text{minimize}} \frac{1}{2} \|w\|^2 \text{ subject to } y_i \left[\langle x_i, w\rangle + b\right] \geq 1$$

- Many optimization techinques to solve this
  - e.g., Stochastic Gradient Descent (SGD)

- Implementations available
  - SVM^{light} (Thorsten Joachims)
  - SGD-SVM (Léon Bottou)

Support Vector Machines

- What about linearly inseparable cases ?

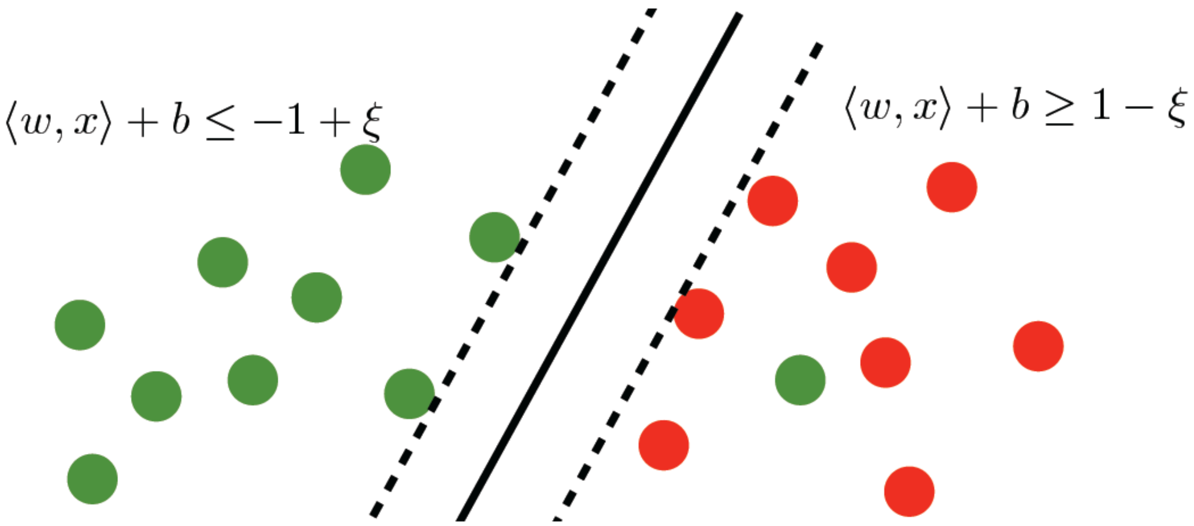


linear function

$$f(x) = \langle w, x\rangle + b$$

Support Vector Machines

- What about linearly inseparable cases ?



Action recognition in realistic videos

Challenges

- Severe camera motion
- Variation in human appearance and pose
- Cluttered background and occlusion
- Viewpoint and illumination changes

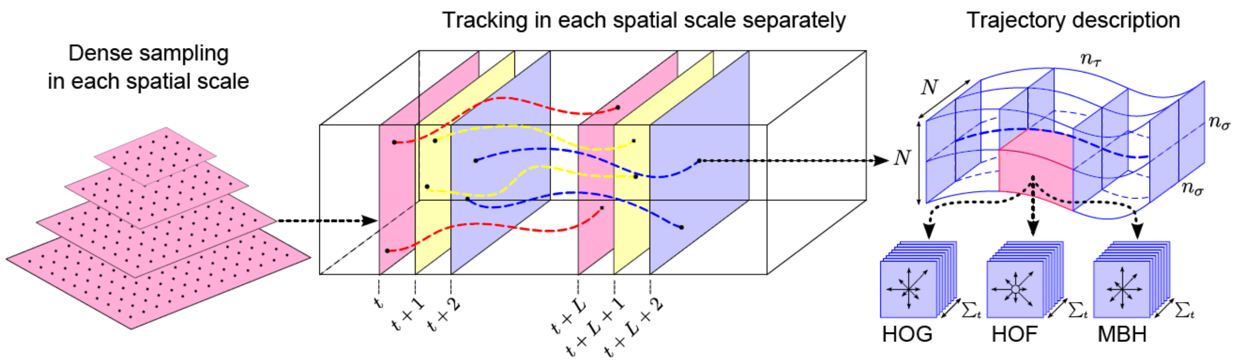
Current state of the art

- Local space-time features + bag-of-features model
- Dense trajectories performs the best on a large variety of datasets [Wang et.al. IJCV'13]

Slides courtesy H. Wang

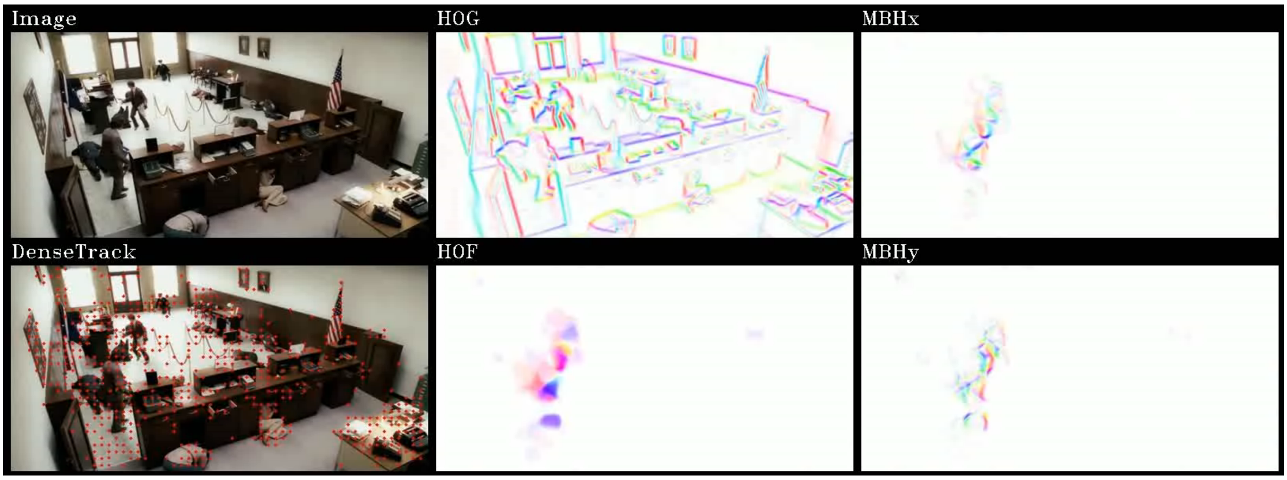
Dense trajectories

- Three major steps:
  - Dense sampling
  - Feature tracking
  - Trajectory-aligned descriptors



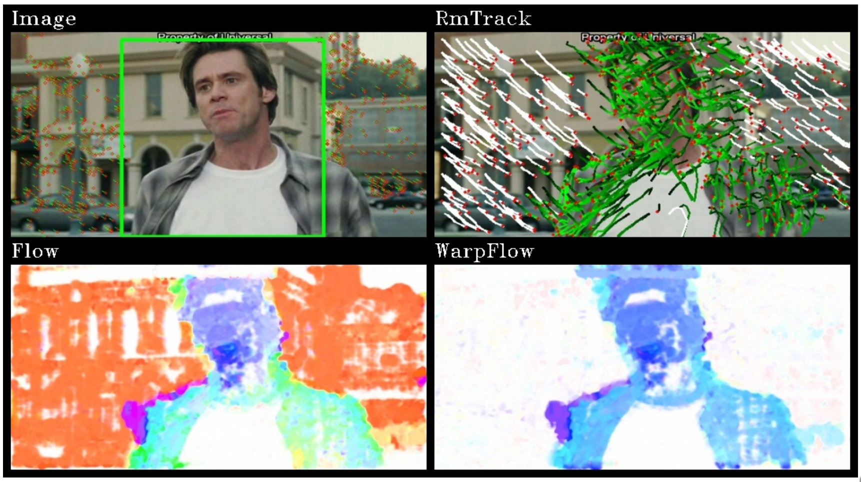
Dense trajectories

- Advantages:
  - Capture
  - MBH is robust to camera motion
- Disadvantages:
  - Generate irrelevant trajectories in background due to camera motion
  - Motion descriptors are corrupted due to camera motion, e.g., HOF, MBH



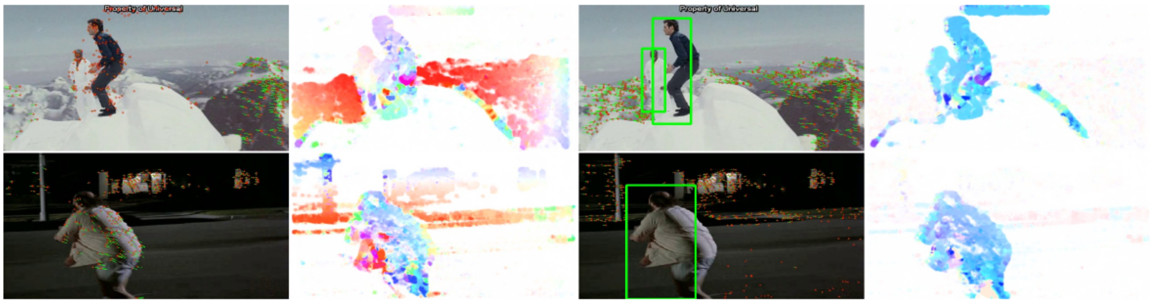
Improved dense trajectories

- Improve dense trajectories by explicit camera motion estimation
- Detect humans to remove outlier matches for homography estimation
- Stabilize optical flow to eliminate camera motion
- Remove trajectories caused by camera motion

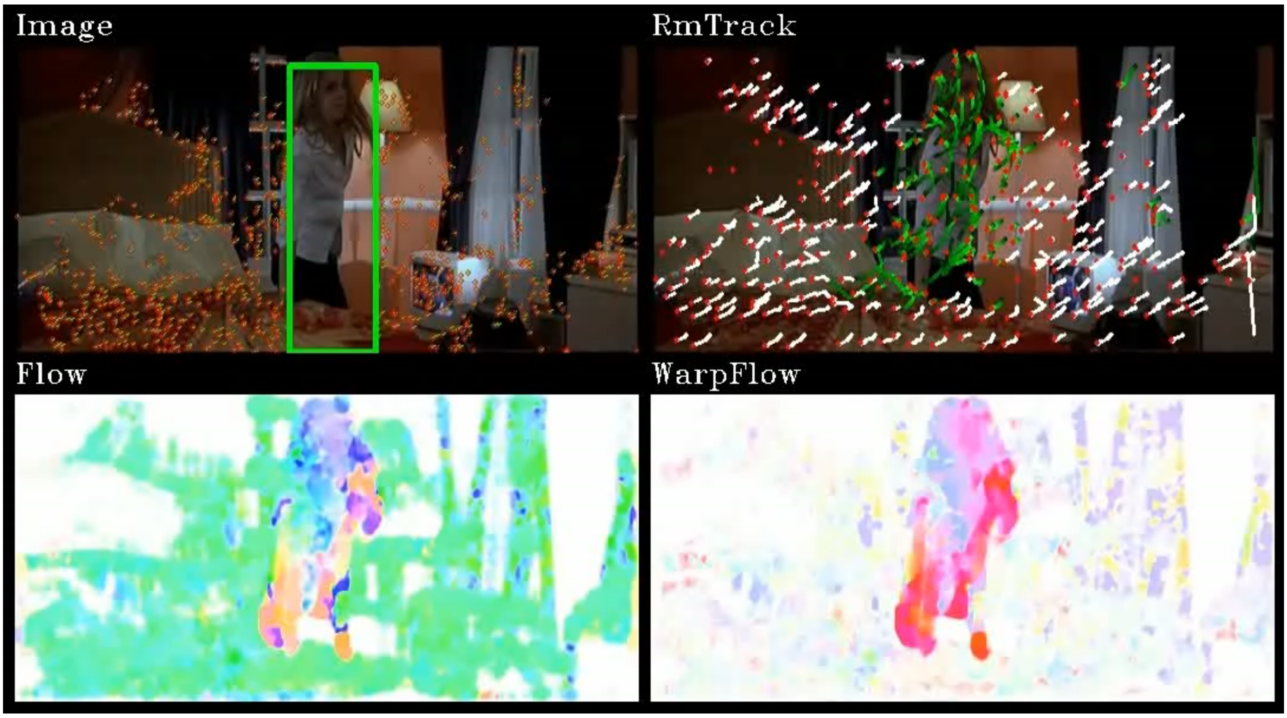


Remove inconsistent matches due to humans

- Human motion: not constrained by camera motion → generates outlier matches
- Apply a human detector in each frame, and track the human bounding box
- Remove feature matches inside the human bounding box during estimation



Example of the refined flow

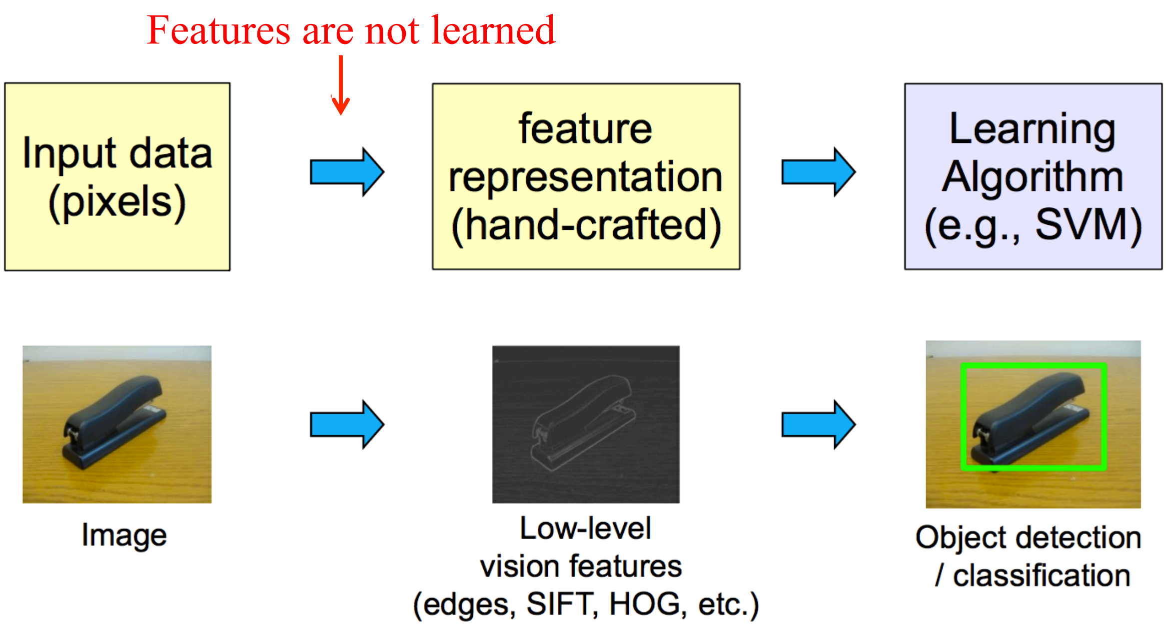


Source code: http://lear.inrialpes.fr/~wang/improved_trajectories

Plan

- Classification
- Machine à vecteurs de support
- Reconnaissance d’actions
- CNN
- Quelques perspectives

Traditional Approaches for Recognition



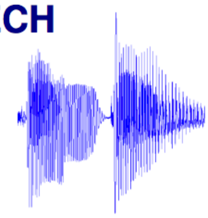
CNN Slides courtesy: M. Renzato, R. Fergus

Traditional Approaches for Recognition

VISION

→ SIFT/HOG (fixed) → K-Means/ pooling (unsupervised) → classifier (supervised) → "car"

SPEECH

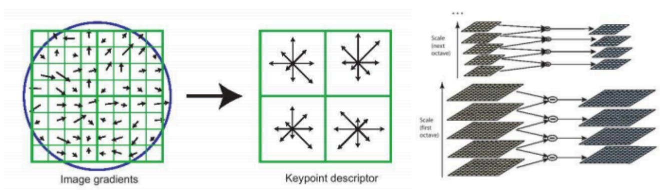


→ MFCC (fixed) → Mixture of Gaussians (unsupervised) → classifier (supervised) → \'d ē p\

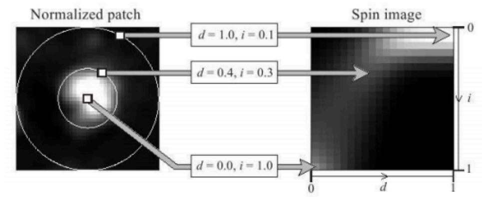
NLP

This burrito place is yummy and fun! → Parse Tree Syntactic (fixed) → n-grams (unsupervised) → classifier (supervised) → "+"

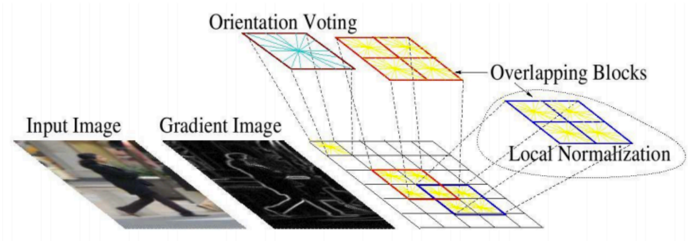
Computer Vision Features



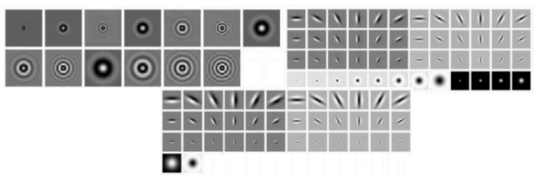
SIFT



Spin image



HoG



Textons

and many others:

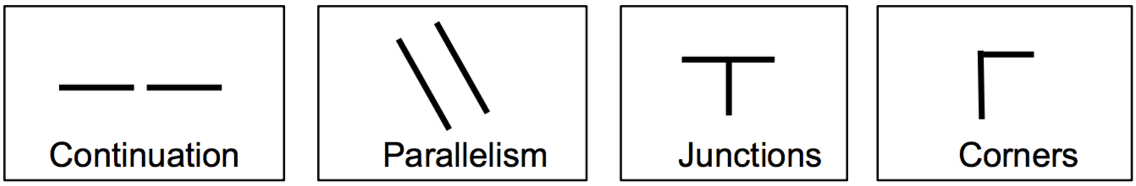
SURF, MSER, LBP, Color-SIFT, Color histogram, GLOH, …..

Computer Vision Features

- Features are key to progress
- Have led to impressive results in various competitions (e.g., PASCAL VOC)
- Where do we go from here? Better features? Better classifiers?

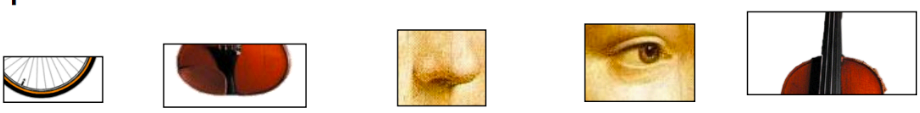
Mid-level Representations

- Mid-level cues



"Tokens" from Vision by D.Marr: 

- Object parts:



Mid-level Representations

VISION

pixels → edge → texton → motif → part → object

SPEECH

sample → spectral band → formant → motif → phone → word

NLP

character → word → NP/VP/.. → clause → sentence → story

Difficult to hand-engineer → What about learning them?

Learning Feature Hierarchy

- Learn hierarchy
- All the way from pixels → classifier
- One layer extracts features from output of previous layer

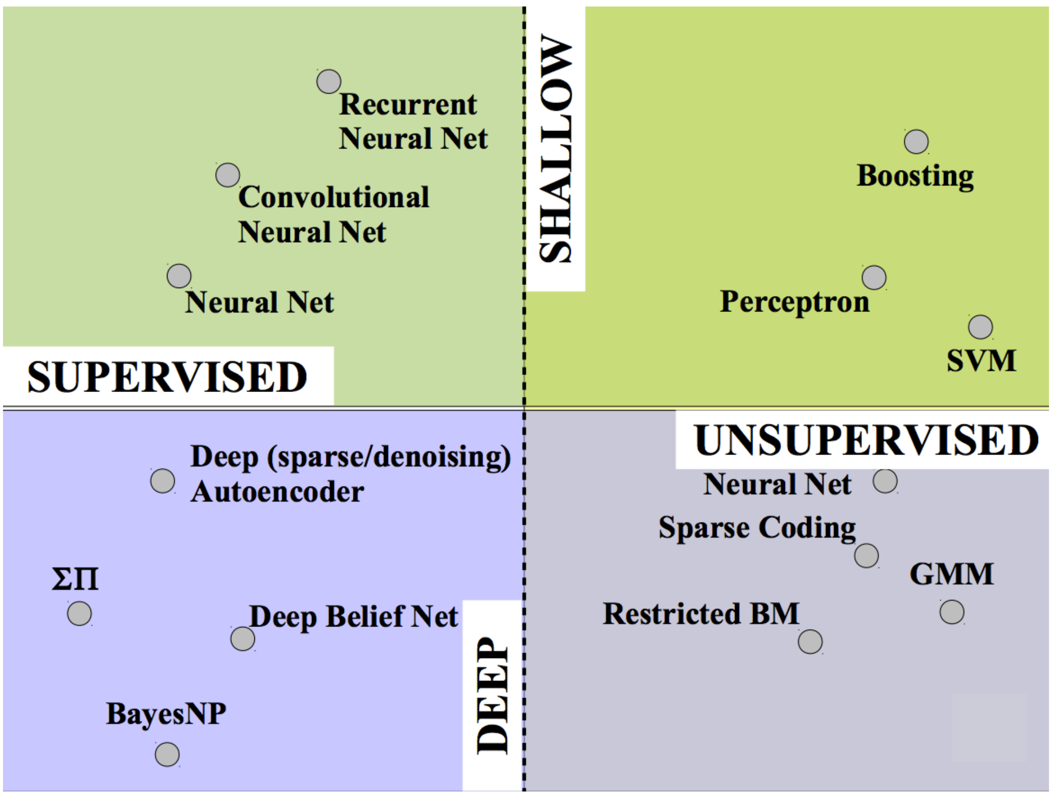
- Train all layers jointly

Learning Feature Hierarchy

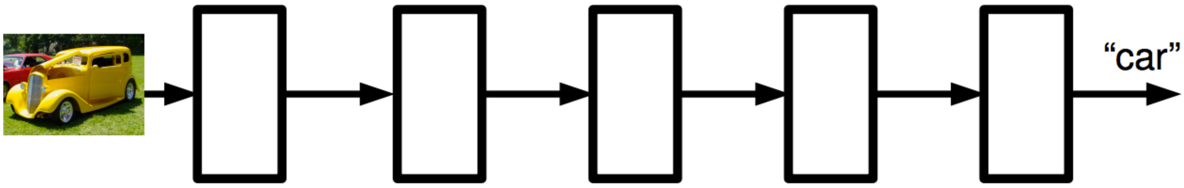
- Supervised Learning
  - End-to-end learning of deep architectures (e.g., deep neural networks) with back-propagation
  - Works well when the amounts of labels is large
  - Structure of the model is important (e.g. convolutional structure)

- Unsupervised Learning
  - Learn statistical structure or dependencies of the data from unlabeled data
  - Layer-wise training
  - Useful when the amount of labels is not large

The Space of Learning Methods



Deep Learning



What is Deep Learning

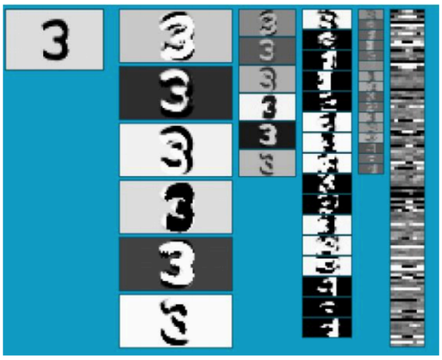
- Cascade of non-linear transformations
- End to end learning
- General framework (any hierarchical model is deep)

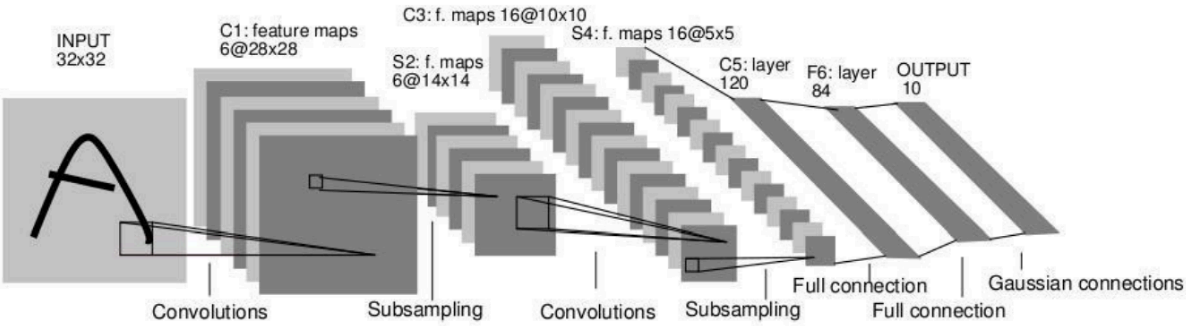
But, before that, some basics (… on the board)

- Perceptron
  - Single-layer
  - Multi-layer
- Non-linearity
  - Sigmoid
- Backpropagation algorithm
- Neural networks

Example: Convolutional Neural Networks (CNN)

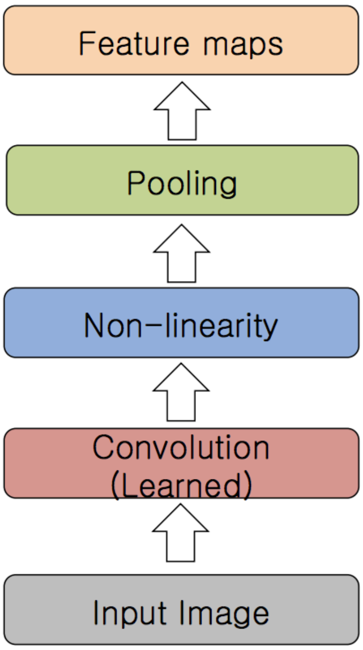
- LeCun et al. 1989
- Neural network with specialized connectivity structure



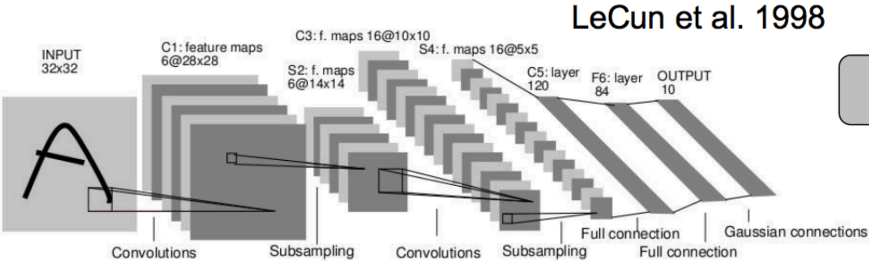


CNN

- Feed-forward:
  - Convolve input
  - Non-linearity (rectified linear)
  - Pooling (local max)
- Supervised
- Train convolutional filters by back-propagating classification error



LeCun et al. 1998

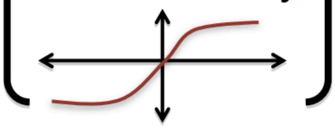


Components of each CNN Layer

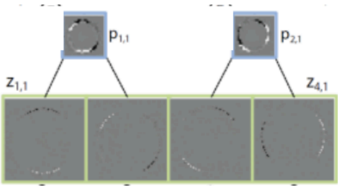
Pixels / Features

Filter with Dictionary (convolutional or tiled)

+ Non-linearity



Spatial/Feature (Sum or Max)



[Optional]

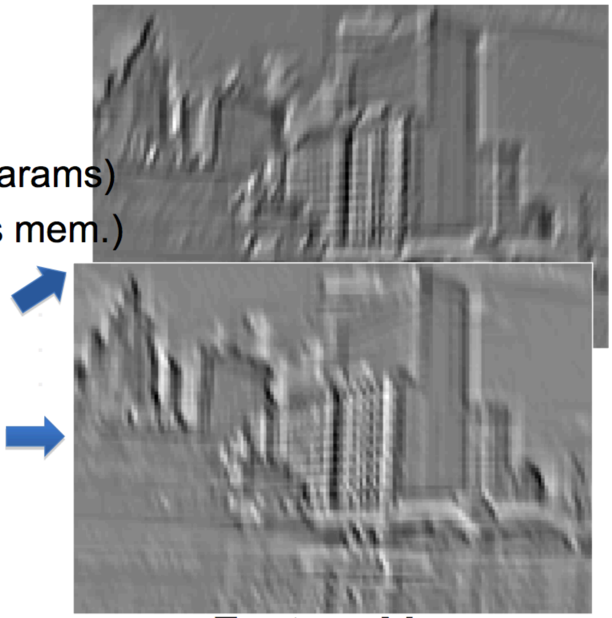
Normalization between feature responses

Output Features

Filtering: Convolutional

- Convolutional
  - Dependencies are local
  - Translation equivariance
  - Tied filter weights (few params)
  - Stride 1,2,… (faster, less mem.)

Input

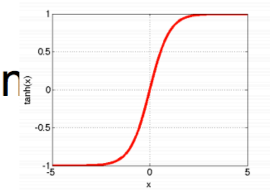


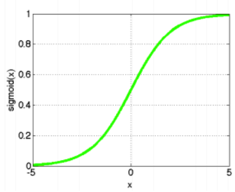
Feature Map

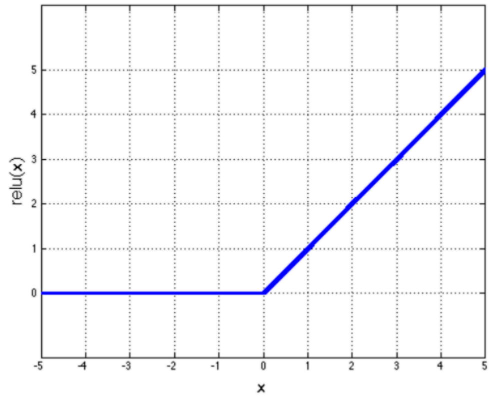
Filtering: Non-linearity

- Non-linearity
  - Per-element (independen
  - Tanh
  - Sigmoid: 1/(1+exp(-x))
  - Rectified linear
    - Simplifies backprop
    - Makes learning faster
    - Avoids saturation issues

  → Preferred option

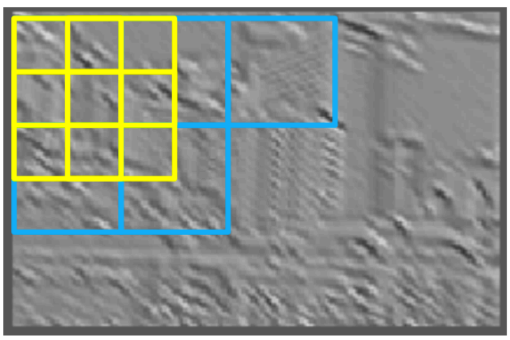


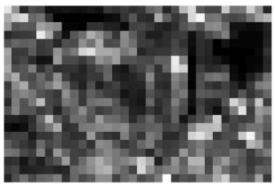


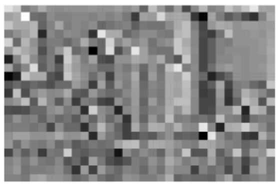


Pooling

- Spatial Pooling
  - Non-overlapping / overlapping regions
  - Sum or max
  - Boureau et al. ICML’10 for theoretical analysis

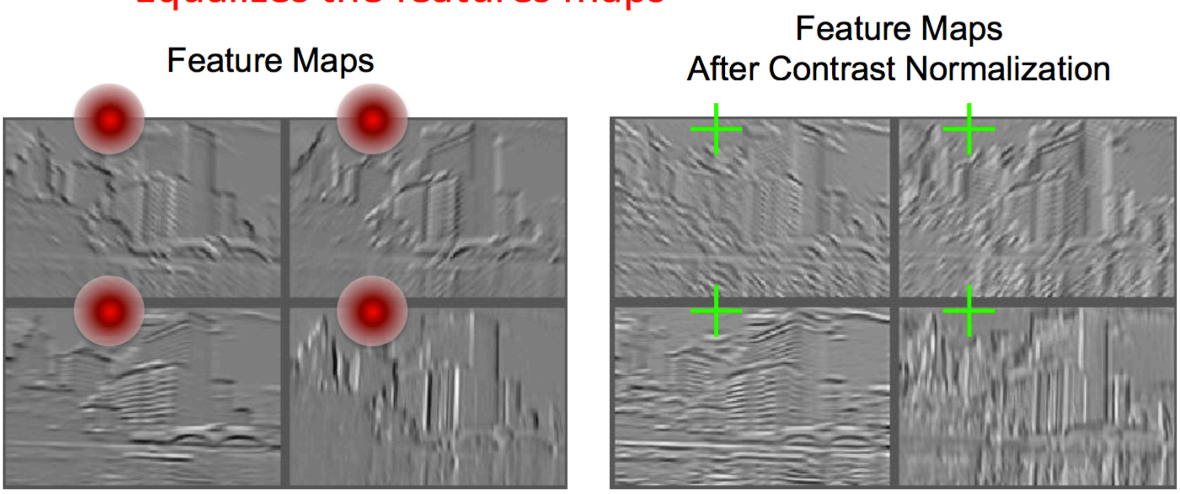


Max


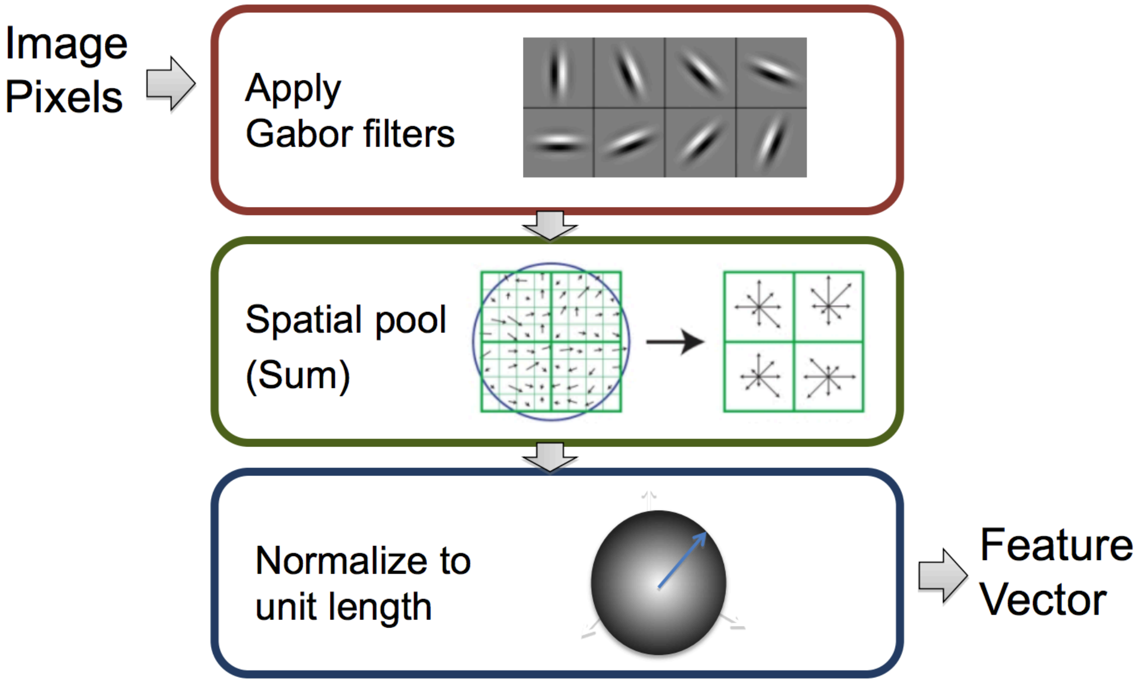
Sum


Normalization

- Contrast normalization (across feature maps)
  - Local mean = 0, local std. = 1, "Local" → 7x7 Gaussian
  - Equalizes the features maps



Comparison with SIFT

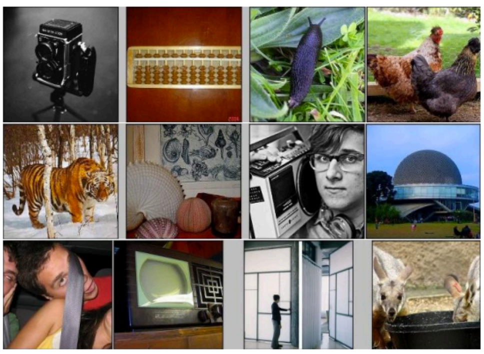


CNN: Applications

- Handwritten text/digits
  - MNIST (0.17% error [Ciresan et al. 2011])
  - Arabic & Chinese [Ciresan et al. 2012]

- Simpler recognition benchmarks
  - CIFAR-10 (9.3% error [Wan et al. 2013])
  - Traffic sign recognition
    - 0.56% error vs 1.16% for humans [Ciresan et al. 2011]

CNN: Applications

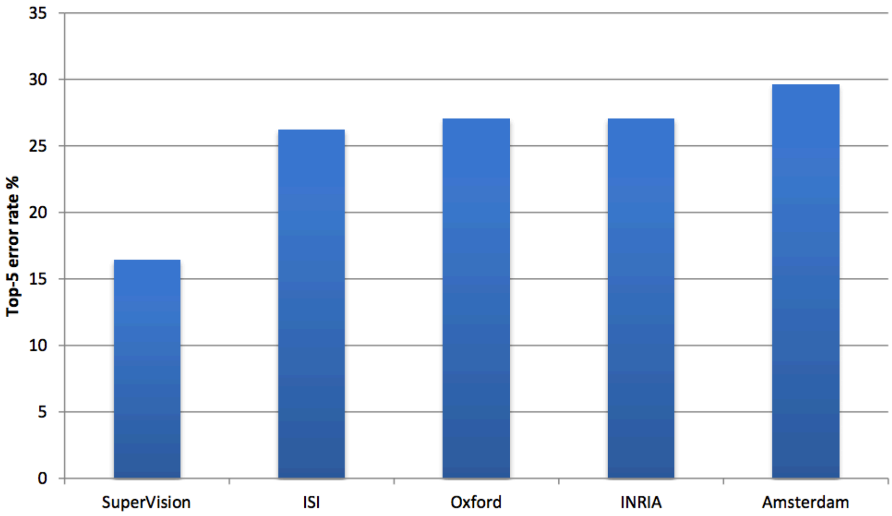


[Deng et al. CVPR 2009]

- ~14 million labeled images, 20k classes
- Images gathered from Internet
- Human labels via Amazon Turk

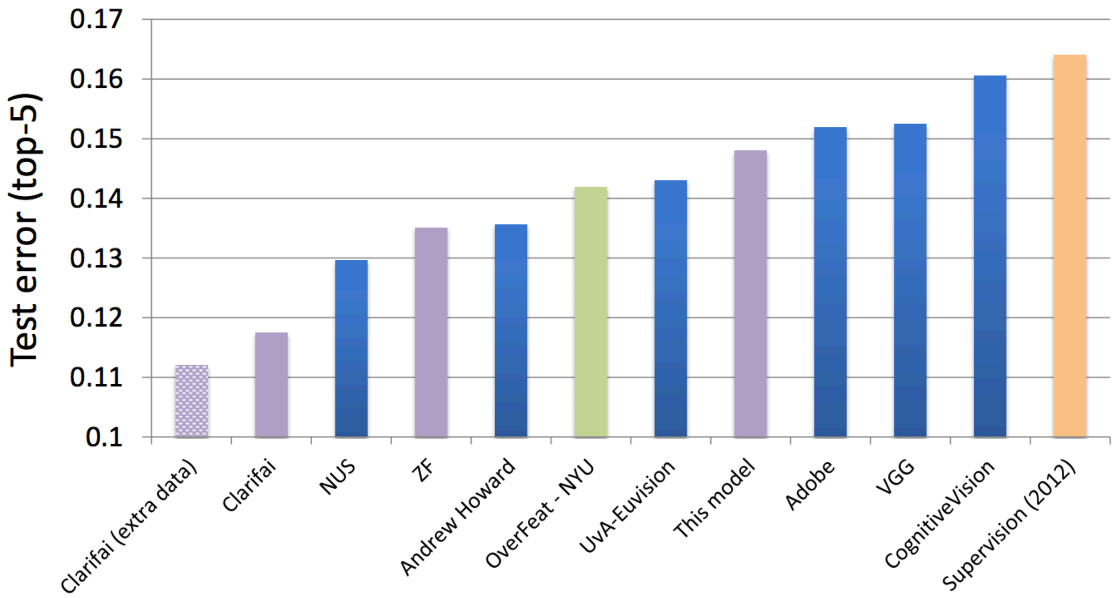
ImageNet 2012

- Krizhevsky et al. -- 16.4% error (top-5)
- Next best (non-convnet) – 26.2% error



ImageNet 2013

- http://www.image-net.org/challenges/LSVRC/2013/results.php



Plan

- Classification
- Machine à vecteurs de support
- Reconnaissance d’actions
- CNN
- Quelques perspectives

Problèmes d’intérêt et perspectives

- En conclusion
  - domaine de recherche très actif car récent

    → les capacités de calcul des ordinateurs permettent depuis très peu de temps de traiter des grands volumes de données

  - Compétitions pour comparer les méthodes : ImageNet, Trecvid, VideOlympics