Semantic scene labeling using feature learning

Camille Couprie,

IFP Energies Nouvelles, Rueil Malmaison, France

Work achieved while at New York University with

Clément Farabet, Laurent Najman

and Yann LeCun

4 March 2013

Extract
information
from the data

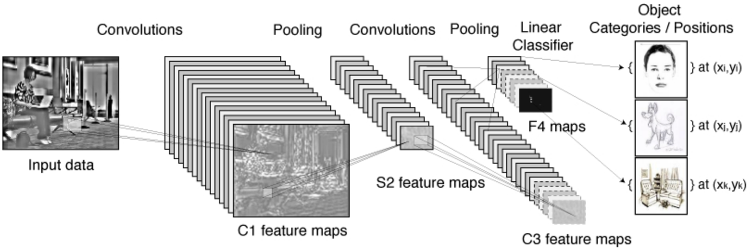
- Ultimate goal of vision : semanticaly label everything

$\rightarrow$
?

- Classical way : hand crafted features, probabilistic approaches defining graphical models
- Here : *reproducible* semantic scene labeling *in real time*

Feature Learning with Convolutional Networks

- Feature learning : not new [LeCun *et al.* 89, 98]

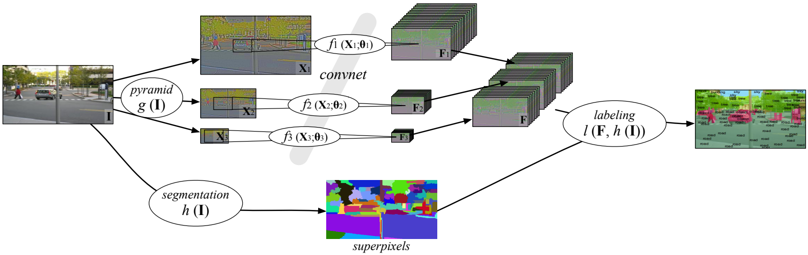


- Visual cortex organizes recognition in a hierarchical way
- Convnet : Layered convolutions and downsampling steps
- Applications : classification, object recognition. Semantic segmentation ?

Multiscale feature learning for scene labeling

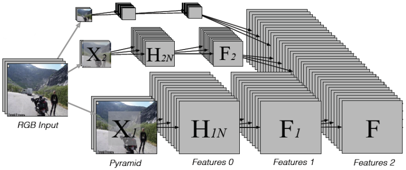
- Full image labeling implies joint

$$\left\{\begin{array}{l}\text{Recognition}\\ \text{Localization}\\ \text{Delineation}\end{array}\right. \quad \text{of objects}$$



[Clément Farabet, C. Couprie, L. Najman, Y. LeCun ICML,PAMI 2012]

Multiscale feature learning for scene labeling



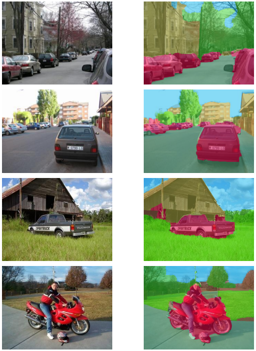
- How to compute the feature maps F_s at each scale s :
  Input image $H_0 = X_s$
  Hidden layers $H_{lp} = \text{maxp}\ tanh(b_{lp} + \sum_{q \in \text{parents}(p)} W_{lpq} * H_{l-1,q})$
  Output feature map $F = [F_1, u(F_2), ...u(F_N)]$ (u : upsampling)
- Pixelwise predictions $\hat{\mathbf{c}}_{i,a} = \frac{e^{\mathbf{w}_a{}^T \mathbf{F}_i}}{\sum_{b \in \text{classes}} e^{\mathbf{w}_b{}^T \mathbf{F}_i}}$
- Learning the parameters (b, W, w) by minimization of the multiclass cross entropy loss function
  $L = -\sum_{i \in \text{pixels}} \sum_{a \in \text{classes}} \mathbf{c}_{i,a} \ln(\hat{\mathbf{c}}_{i,a})$, using stoch. gradient descent

Results on the Stanford Background dataset (8 classes)

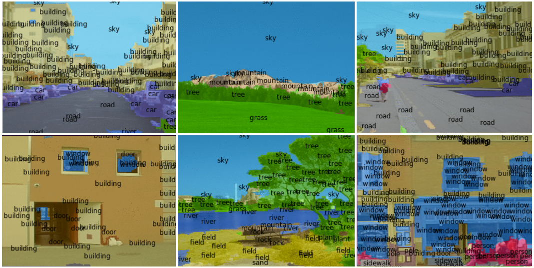
Performance of our system : Per-pixel / Average per-class accuracy.

	Per pixel	Per Class	time
Gould *et al.*, 2009	76.4%	-	10-600s
Munoz *et al.*, 2010	76.9%	66.2%	12s
Tighe *et al.*, 2010	77.5%	-	10-300s
Socher *et al.*, 2011	78.1%	-	?
Kumar *et al.*, 2010	79.4%	-	< 600s
Lempitsky *et al.*, 2011	**81.9%**	72.4	> 60s
singlescale convnet	66.0 %	56.5 %	0.35s
multiscale convnet	78.8 %	72.4%	0.6s
multiscale net + sup. pix.	80.4%	74.6%	**0.7s**
multiscale net + cover	80.4%	**75.2%**	61s
multiscale net + CRF	81.4%	**76.0%**	61s

(computation times of our system measured on a 4-core Intel i7)



Results on the SIFT-Flow dataset (33 classes)



	Pixel acc.	Class accuracy
Liu *et al.*, 2009	74.75 %	-
Tighe *et al.*, 2010	76.9 %	29.4 %
multiscale net + cover[1]	**78.5 %**	**29.6 %**
multiscale net + cover[2]	74.2 %	**46.0 %**

[1] respecting natural frequencies of classes,

[2] balancing them so that an equal amount of each class is shown to the network.

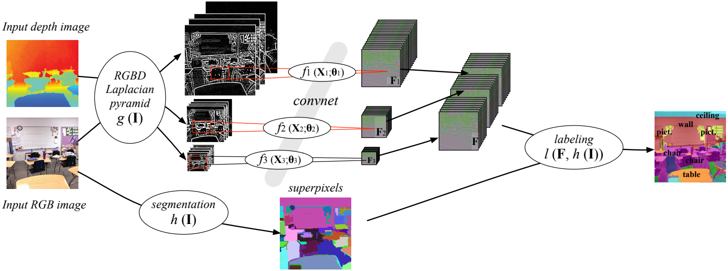
Does it work for real? Near Broadway

Play

Does it work for real ? In Washington Square Park

Play

Multiscale feature learning using depth information



ICLR 2013, submitted to JMLR

Results on NYU depth dataset v2

Ground truths

Our results

wall	books	chair	furniture	sofa	TV screen
bed	ceiling	floor	picture or deco	table	window

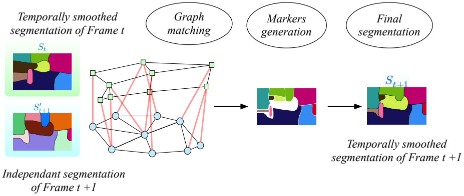
Comparison with Silberman *et al.* on NYU depth dataset

Structure Class Feature Descriptions	**Dims**
Color	**36**
C1: Color histograms: 10-bin histograms for the values of each channel. [1]	30
C2: Mean and standard deviation of color channels	6
Shape	**1086**
A1: Sparse coded SIFT descriptor histograms	1000
A2: 2D Bounding box dimensions	2
A3: 3D Bounding box dimensions	3
A4: Pyramid of Surface normal histograms	78
A5: Mean, median, max of planar errors	3
Scene	**6**
N1: Distance to closest wall: absolute and normalized by room size	2
N2: Relative Depth: mean and variance relative depth over the region [2]	2
N3: Height: minimum and maximum heights above the ground	2

Table 3. Structure Class Features. Used to classify each region of the image into one of four structure classes: Ground, Furniture, Prop and Structure.

	Ground	Furnit.	Props	Structure	Class	Pixel
Silberman *et al.* ECCV'12	68	**70**	**42**	59	59.6	58.6
Multiscale convnet	68.1	51.1	29.9	**87.8**	59.2	63.0
Multiscale+depth convnet	**87.3**	45.3	35.5	86.1	**63.5**	**64.5**

Temporal smoothing superpixels



1. Independent segmentation
2. Graph matching to identify corresponding regions
3. The corresponding regions are mined to create markers
4. The final segmentation is the solution to a global optimization procedure given the markers as constraints

Semantic segmentation

Our temporal semantic segmentation - Our temporal super-pixels

Play

Independent segmentation - Our temporal semantic segmentation

Play

Legend :
- awning
- balcony
- building
- car
- door
- person
- road
- sidewalk
- sun
- tree
- window

Comparison with the approach of Miksik et al.

(a) Independent segmentations with no temporal smoothing. Accuracy : 71.1

(b) Result using the temporal smoothing method [Miksik et al. 2012]. Accuracy : 75.3, Computation time :0.8s

(c) Our temporally consistent segmentation. Accuracy : 76.3, Computation time :0.1s

Legend :	balcony	car	person	sidewalk	tree
awning	building	door	road	sun	window

Power to the data ...
To make Power

Current and future work

Problems encountered at IFPEN

- Process Genomic data to build gene regulatory networks for biofuel production improvements
- Semantic segmentation of materials images
- Chemical signal analysis
- Seismic data restoration

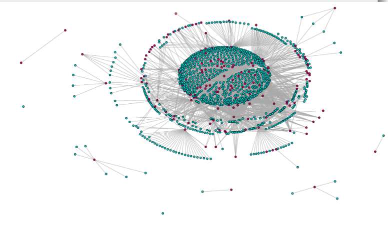


Image from A. Pirayre

Current and future work

Problems encountered at IFPEN

- Process Genomic data to build gene regulatory networks for biofuel production improvements
- Semantic segmentation of materials images
- Chemical signal analysis
- Seismic data restoration

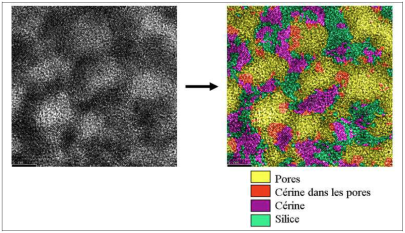


Image from M. Moreaud

Current and future work

Problems encountered at IFPEN

- Process Genomic data to build gene regulatory networks for biofuel production improvements
- Semantic segmentation of materials images
- Chemical signal analysis
- Seismic data restoration

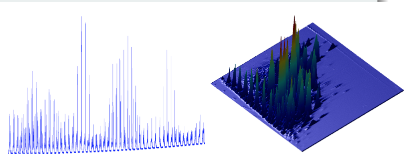


Image from L. Duval

Current and future work

Problems encountered at IFPEN

- Process Genomic data to build gene regulatory networks for biofuel production improvements
- Semantic segmentation of materials images
- Chemical signal analysis
- Seismic data restoration

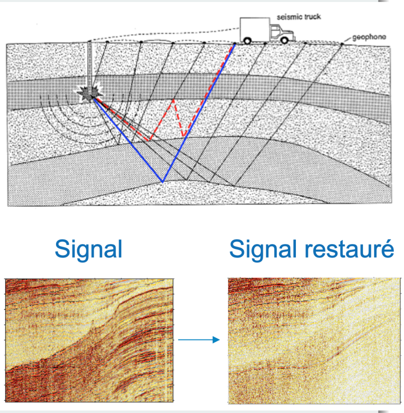
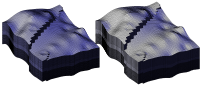


Image from L. Duval

Postdoc position



Very large data management in Geosciences

- Propose new data compression techniques for volumetric meshes able to manage seismic data values attached to geometry elements (billions of nodes or cells) with adaptive decompression for post-processing functionalities (visualization).
- Applications : geoscience fluid-flow simulation or transport combustion simulation on very large meshes.
- Propose new software solutions for the storage, the transfer and the processing (exploration, visualization) of these large data sets.

Questions