Learning Approaches for Remote Sensing Image Classification

Yuliya Tarabalka

Inria Sophia Antipolis-Méditerranée - TITANE team

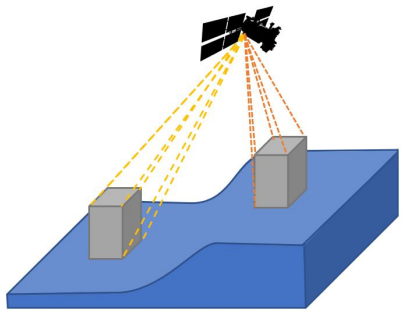
Université Côte d'Azur - École Doctoral STIC

HdR committee:

Mihai Datcu	Reviewer
Florence Tupin	Reviewer
Lorenzo Bruzzone	Reviewer
Paul Scheunders	Examiner
Gabriele Moser	Examiner
Xavier Descombes	Examiner

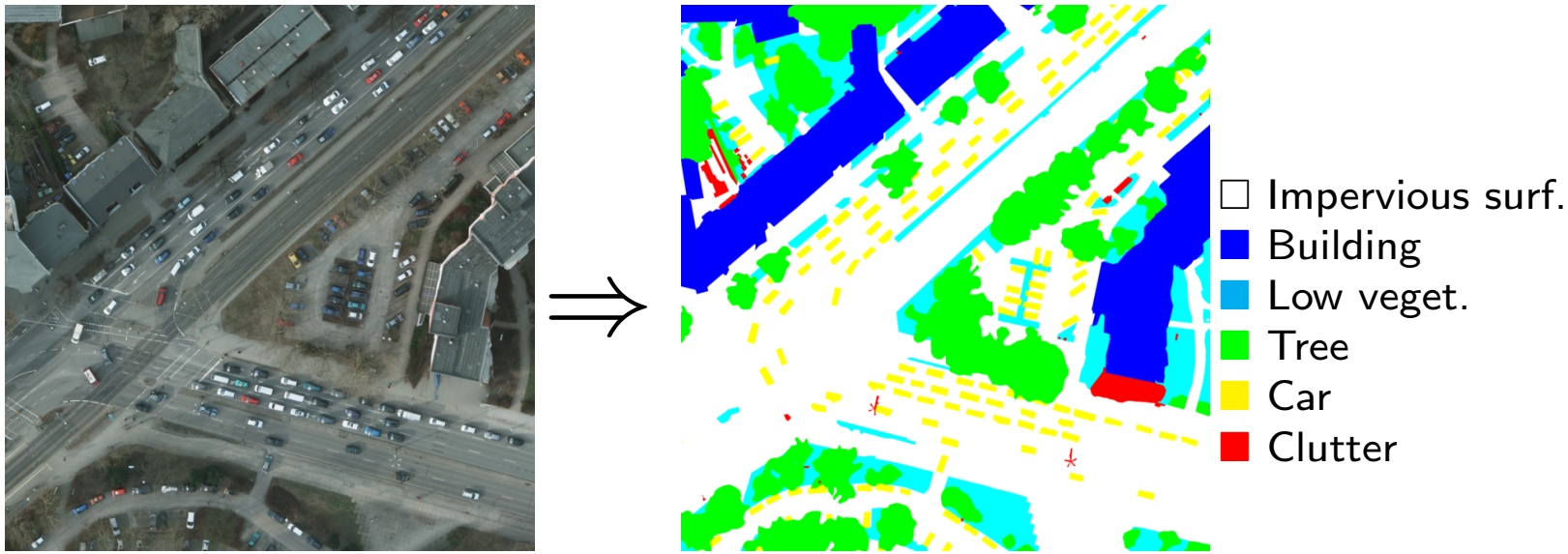
Context

- Continuous proliferation & improvement of remote data sensors
- Huge volume of unstructured satellite images



Context

- **Crucial need:** automatize the analysis of remote sensing data
- **Remote sensing image classification:** assign a semantic class to every pixel



Input Output

Classification approaches

Pixelwise

- SVMs [Camps-Valls 2006]
- Random forests [Ham 2005]
- Neural networks [Goel2003, Ratle 2010]

Graph-based

- Minimum spanning forest [Bernard 2012]
- Partition trees [Valero 2010]
- Graph cut [Tarabalka 2014]

Feature engineering

- Morphological, attribute & extinction profiles [Fauvel 2008, Dalla Mura 2010]
- Texture + SVMs [Huang 2008]
- Strong shape prior [Lacoste 2005, Wang 2010, Jeong 2015]

Deep learned features

- Convolutional neural networks [Mnih 2012]
- Deep features + SVMs [Chen 2014]
- Fully convolutional networks [Marmanis 2016, Volpi 2017]

Spectral-spatial methods

Challenges: Large-scale data sources

- Increasing amount & openness of data, e.g.:
  - Pléiades: entire earth every day ($<$ 1 m resolution)
  - Free Copernicus satellite data

  $\Rightarrow$ Scalability: temporal/space complexity

- Intra-class variability:

Chicago

Austin

Vienna

- Interest in semantic classes (e.g., *building*, *road*, *lane*)

  $\Rightarrow$ Need for high-level contextual reasoning (shape, patterns,...)

  $\Rightarrow$ Generalization to different locations

Outline

1. Introduction
2. Classification with shape constraints
3. Video segmentation with shape growth/shrinkage constraint
4. Deep learning for large-scale image classification
5. Conclusions and future work

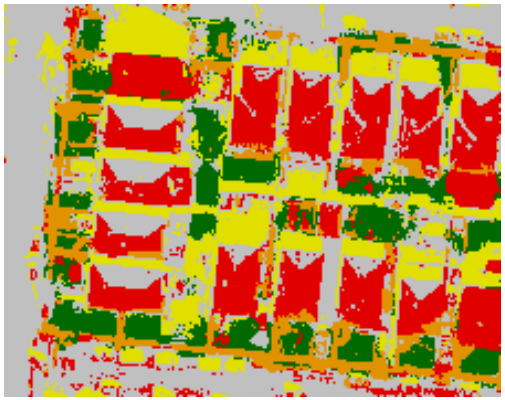
Outline

Optimization with shape constraints

Common graph-based classification approach

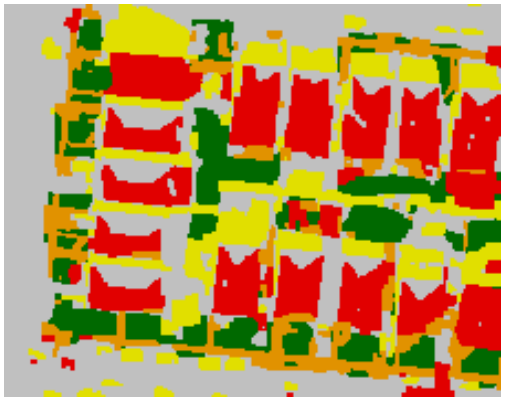
Input

$\Rightarrow$



Pixelwise classif.
(e.g., SVM)

$\Rightarrow$



Spatial regularization
(e.g., graph cut)

Introducing shape constraints

- Shape information may improve classification [Tarabalka 2010]
  - E.g., rectangularity, area, compactness priors
- Difficult to optimize energies with high-level "regional" priors

Optimization with shape constraints

State of the art

- Template matching
  - Graph-based [Fredman 2005]
  - Active contours [Cremers 2006]
  - Marked point processes [Ortner 2008, Jeong 2015]
- Soft priors: specific optimizer
  - Express in MRF's pairwise interaction term (e.g., compactness [Das 2009], star-shaped [Veksler 2008])
  - Linear approx. + trust regions opt. framework (e.g., shape moments, convexity [Gorelick 2013 & 2014])

- **Goal:** Incorporate soft shape priors into the classification process
  - Multiple objects & classes
  - Multiple constraints

Problem formulation: Multi-label classification

Image I, set of regions $\mathcal{R} = (R_i)$, class labels $\mathcal{L} = (L_i)$, $L_i \in \Omega$. Minimize:

$$E(\mathcal{R}, \mathcal{L}) = \underbrace{E_C(I, \mathcal{R}, \mathcal{L})}_{\text{Color prior}} + \underbrace{\sum_{R_i \in \mathcal{R}} E_S(R_i, L_i)}_{\text{Shape prior}} .$$

Color prior

$$E_C(I, \mathcal{R}, \mathcal{L}) = \sum_{R_i \in \mathcal{R}} \sum_{I_j \in R_i} -\log \underbrace{P(L_i|I_j)}_{\text{Classifier}}$$

- How well the labeling fits a classifier's prediction, based on color/spectrum
- *E.g.*, SVM-derived probabilities

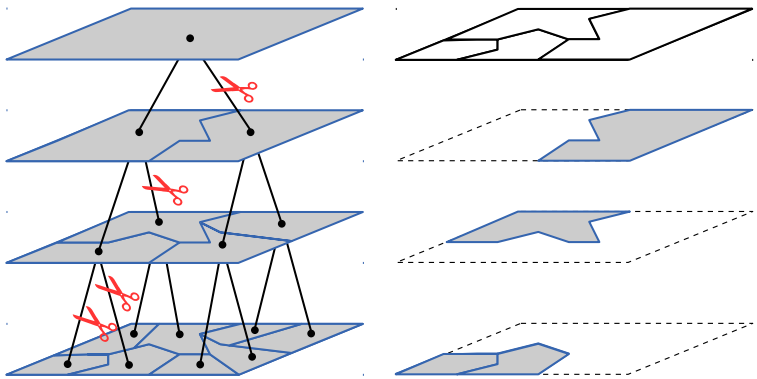
Shape prior

$$E_S(R_i, L_i) = -|R_i| \log P(L_i|\overbrace{\mathbf{S}_i}^{\text{Shape feat. vector}})$$

- How plausible are the objects' shapes
- Train shape feature distributions by kernel density estimation to derive $P(L_i|\mathbf{S}_i)$

Proposed optimization method

- Progressively update a solution to reduce $E(\mathcal{R}, \mathcal{L})$
- Not feasible to explore entire search space to update regions
  - $\rightarrow$ Maintain a hierarchical tree of superpixels (BPT)
  - $\rightarrow$ The current solution is obtained by selecting the optimal combination of nodes in the tree (dynamic prog.)
  - $\rightarrow$ The solution is improved by local transformations on the tree

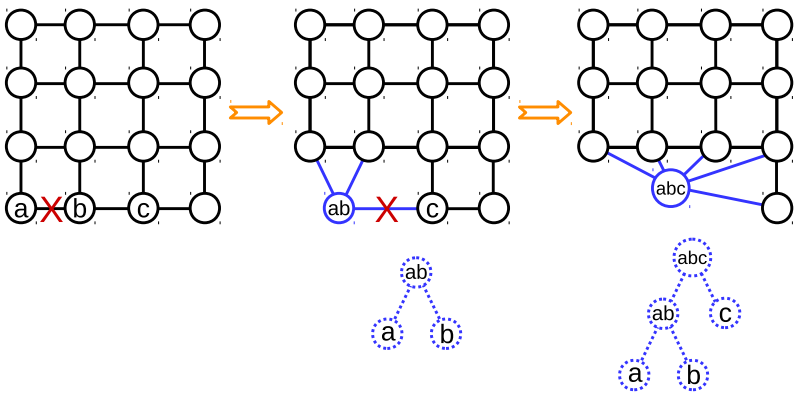


Binary partition tree (BPT)

Constructing an initial BPT

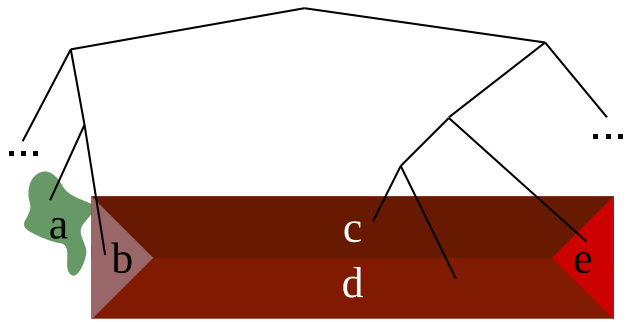
Greedy construction algorithm

Iteratively merge the *most similar* pair of regions:



- Color histograms to represent regions
- Earth Mover Distance to measure histogram dissimilarity

Issue with unoptimized BPTs



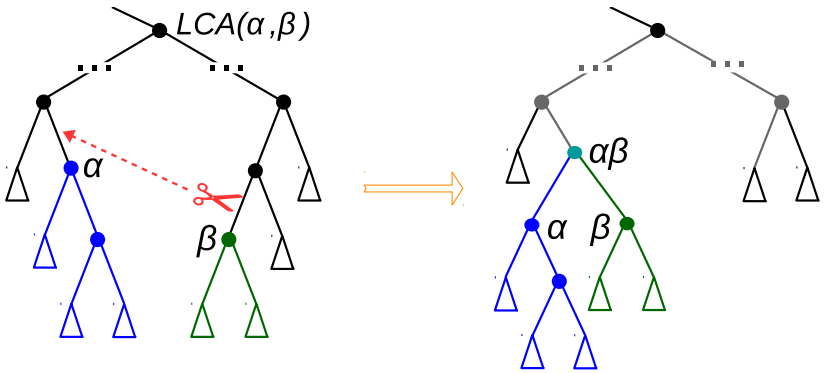
- Shape info cannot be used during BPT construction
- Objects often not represented by single node

E. Maggiori, Y. Tarabalka, G. Charpiat. "Improved partition trees for multi-class segmentation of remote sensing images". IGARSS 2015.

Optimization algorithm

Optimization operator

"Prune and paste" move:



α and β spatial neighbors

Algorithm

Iterate:

1. Construct a heap of **all moves** according to the energy gain.
2. Apply **the best k moves**

until no moves that decrease the energy left.

→ Theoretical properties reduce search space of possible moves

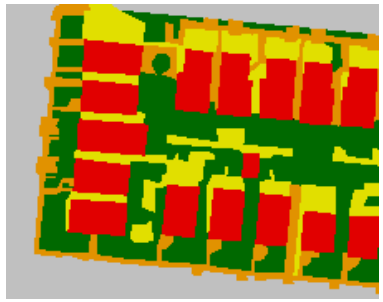
- For every prune place (β) we only try a constant number of paste places (α)

E. Maggiori, Y. Tarabalka, G. Charpiat. "Optimizing Partition Trees for Multi-Object Segmentation with Shape Prior". BMVC 2015.

Experiments

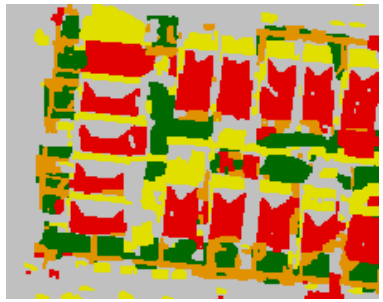
- Input: Google Maps image over Long Island
- SVM and shape features (area, rectangularity, elongatedness) trained on adjacent image

Input (225×180)



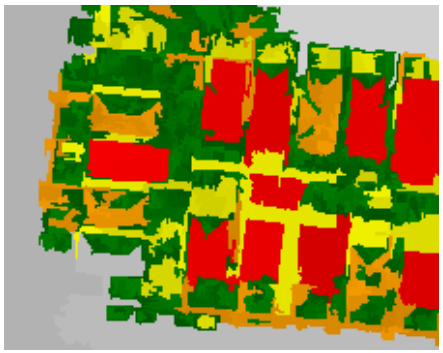
Reference

Tiles, roads, internal roads, veget., shadow



SVM + Graph Cut

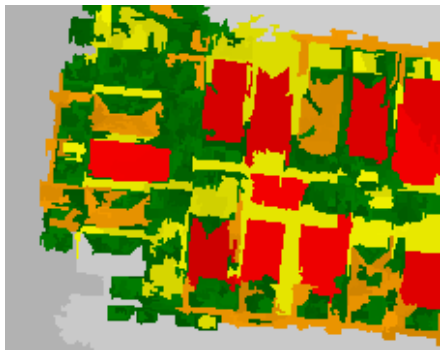
Acc. = 68%



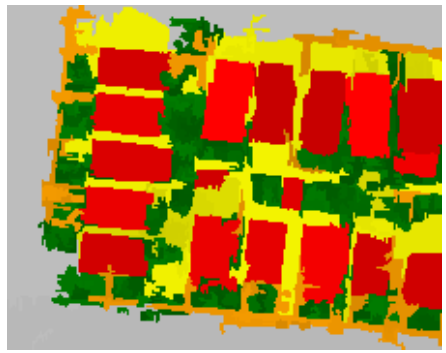
SVM + BPT

Acc. = 65%

→



→

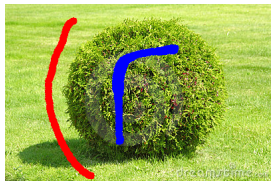


SVM + BPT opt.

Acc. = 79%

Experiments

Convexity prior

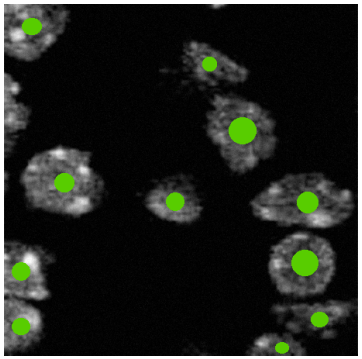


Input

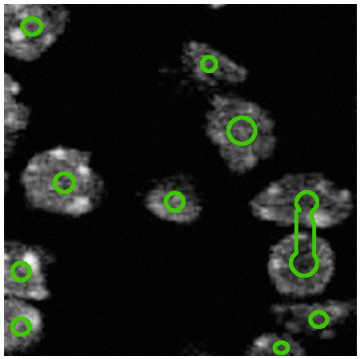
Gorelick et al. (ECCV 2014)

BPT Opt.

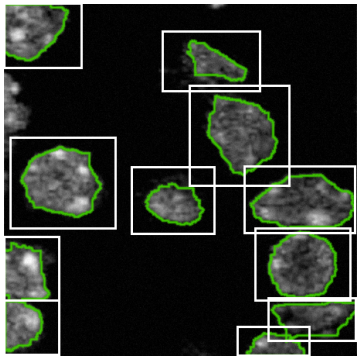
Cell nuclei:



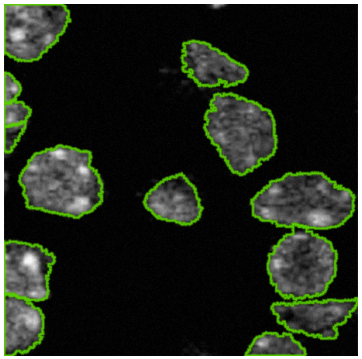
Input (457×454)



Gorelick et al.



Gorelick et al., object by object



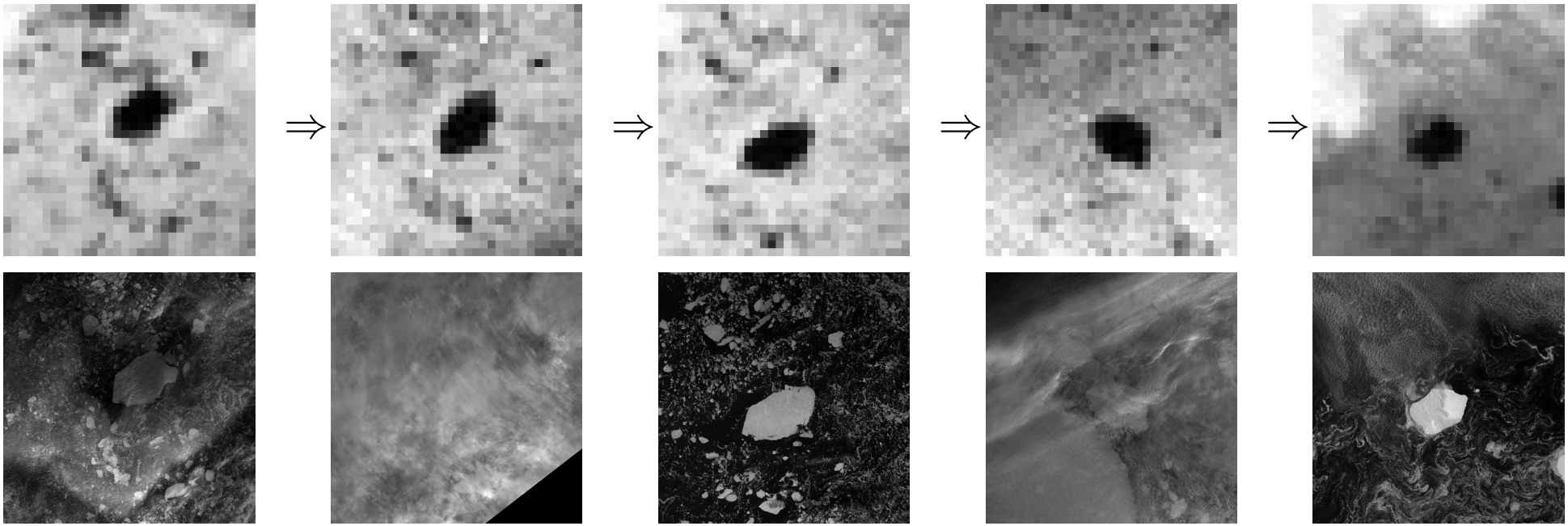
BPT opt.

Concluding remarks

- Shape-aware classification method has been proposed
  - Multiple soft shape constraints
  - Multi-class
  - Multi-object
- Issue: scalability
  - Computational complexity
  - Features

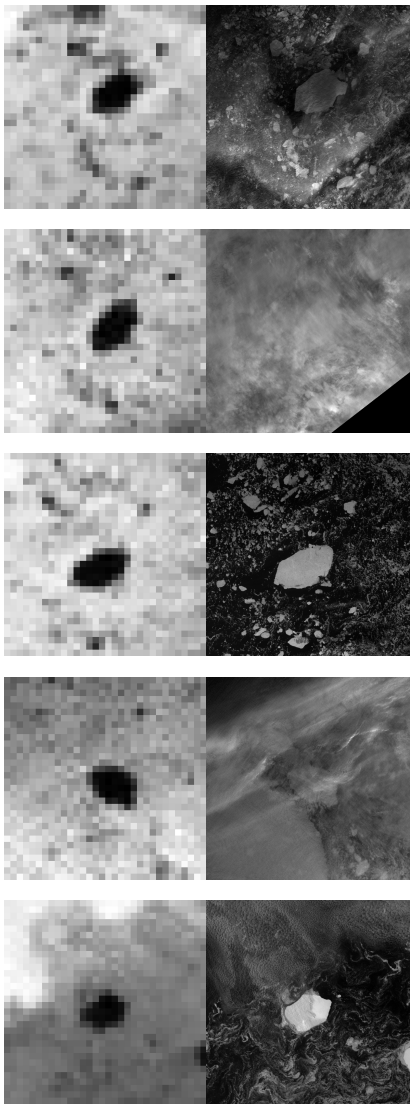
Outline

Motivation: segment a melting flow in time series



- Segment ice floes from time series of AMSR-E + MODIS images
- **Main difficulties:**
  - Low signal-to-noise ratio
  - Foreground & background intensity distributions vary significantly
  - Data for some pixels can be missing
- **Solution:** exploit temporal coherence

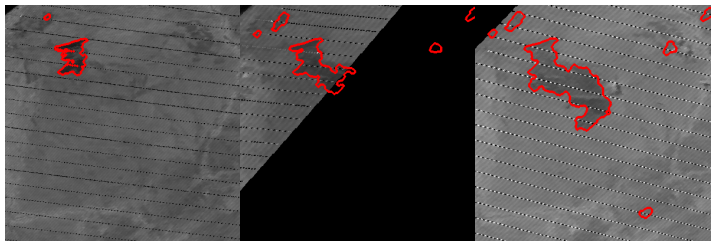
How to exploit temporal coherence?



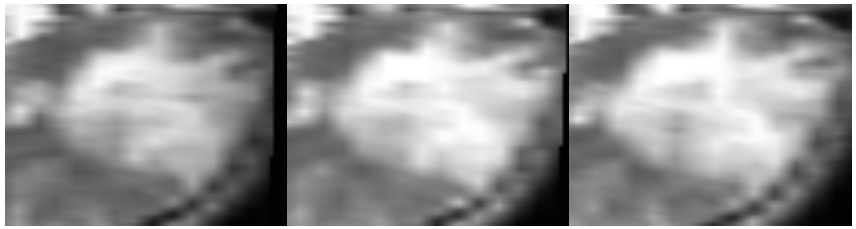
Previous works:	**Our problem:**
• Rely on coherence of foreground/background intensity distributions over time [Shi'98, Grundmann'10]	• Foreground/background intensity distributions vary significantly over time
• Introduce shape priors into image segmentation [Cremers'02, Schoenemann'07]	• Shape prior is unknown • Shape is changing over time
• Smooth 2D+T spatio-temporal volume [Riklin-Raviv'10, Wolz'10]	• Rapid shrinkage events will be underestimated

Solution: introduce shape shrinkage or growth constraint

- **Objective:**
  - Segment **monotonously growing or shrinking** shapes,
  - From time sequences of extremely noisy images,
  - In a low computational time
- **Method:**
  - Formulate video segmentation as an optimization problem,
  - Using the spatio-temporal graph of pixels,
  - With shape **growth or shrinkage constraint** expressed with directed infinite links.
  - Globally-optimal solution is computed with a **graph cut**
- **Examples of growing shapes:**



Savanna fires, 2D satellite data



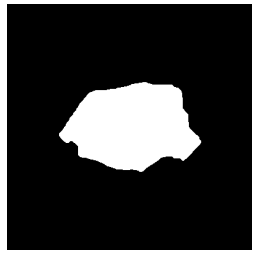
Brain tumor, 3D medical MRI volumes

Y. Tarabalka, G. Charpiat, L. Brucker, B. H. Menze. "Spatio-temporal video segmentation with shape growth or shrinkage constraint". IEEE Trans. on Image Processing, 2014.

Graph cut for image segmentation

- **Goal:** Compute $T(t \in [1, T])$ segmentation maps $L^t = \{L^t_{(x,y)} \in [0,1], x = [1..H], y = [1..W]\}$,

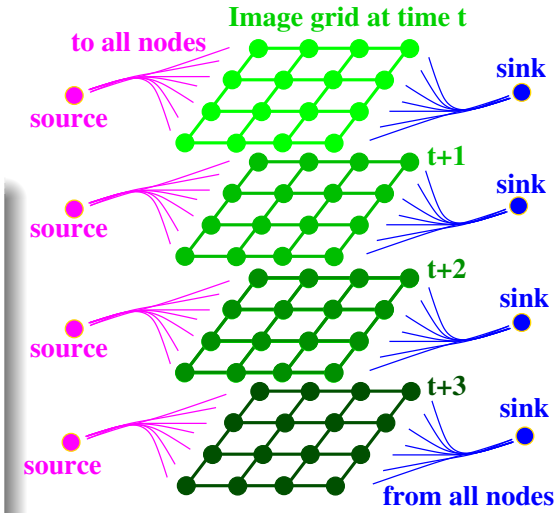
$$L^t_{(x,y)} = \begin{cases} 1, & \text{if } (x,y) \in \text{foreground at time t;} \\ 0, & \text{otherwise.} \end{cases}$$



- **Graph-cut segmentation:**
  1. map each image $I(t)$ onto a graph
  2. minimize a submodular energy of the form:

$$E^t(L) = \sum_{\text{pixels } i} V^t_i(L^t_i) + \sum_{i \sim j} W^t_{i,j}(L^t_i, L^t_j)$$

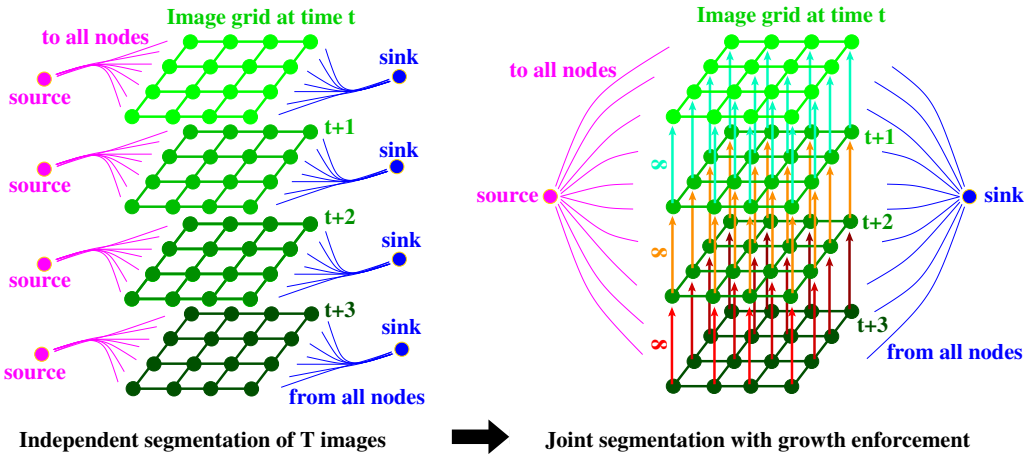
- L^t_i = label of pixel i at time t
- $V^t_i(L^t_i)$ = penalty for a pixel i to have a label L^t_i
- $W^t_{i,j}(L^t_i, L^t_j)$ = interaction term between adjacent pixels i and j



Independent segmentation of T images

Enforcing shape growth

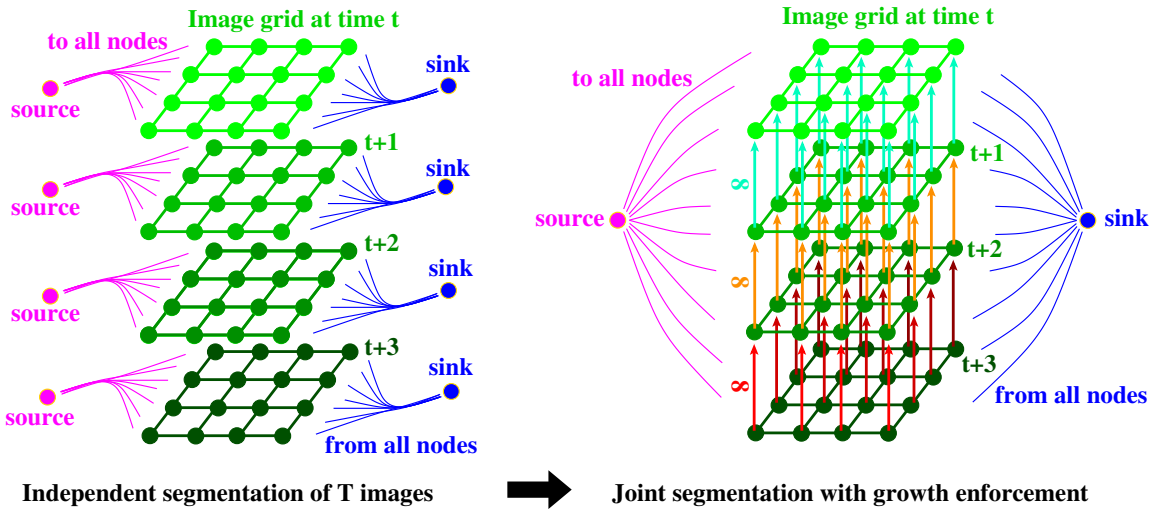
- **Shape growth** = property that the foreground cannot lose any pixel when time advances
- **Enforcing shape growth** (label 1 = foreground, label 0 = background)
  - $\Leftrightarrow$ if $L_i^{t_1} = 1$, then $L_i^{t_2} = 1 \; \forall t_2 > t_1$
  - $\Leftrightarrow$ pair of pixels $((x, y, t), (x, y, t+1))$ cannot have the pair of labels $(1, 0)$
  - $\Leftrightarrow$ **directed infinite link** from each pixel to its predecessor in time



Graph cut with shape growth constraint

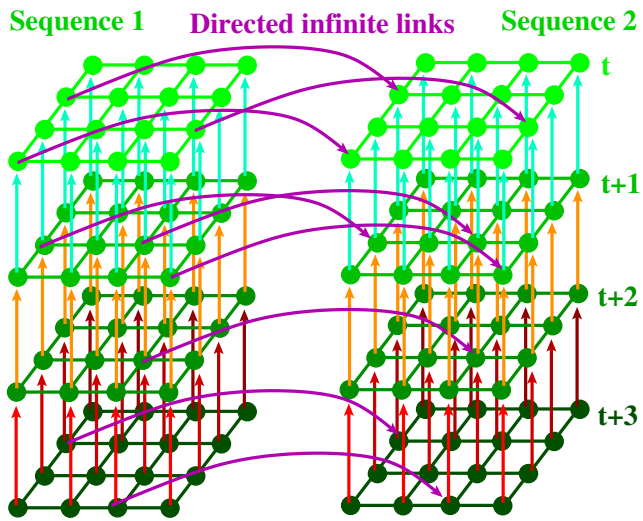
- Segment **jointly** all T images together
  - apply graph cut to the 3D grid $W \times H \times T$
  - with directed infinite links in time
- Criterion minimized: $E = \sum_t E^t$ under the **constraint of shape growth**:

$$E = \sum_{\text{pixels } i} V_i(L_i) + \sum_{i \sim j} W_{i,j}(L_i, L_j) + \infty \sum_t \delta_{L_i^t > L_i^{t+1}}$$



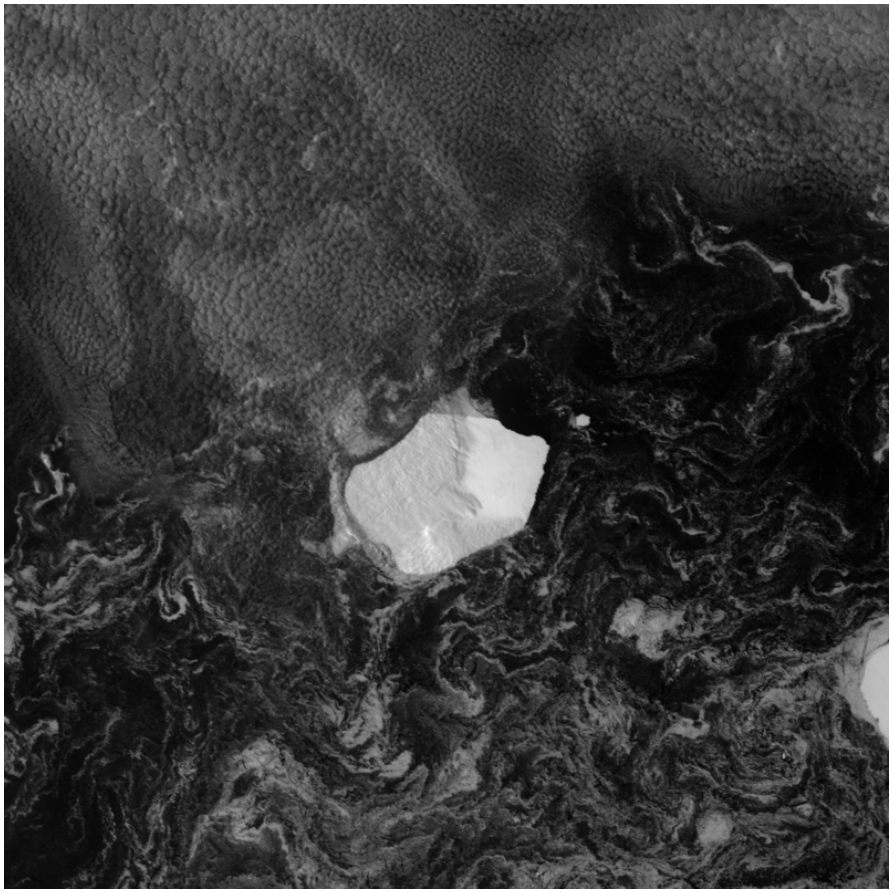
Extensions

- **Shape shrinkage:** reverse the direction of infinite links
  - from each pixel to its successor in time
- **3D shape:** set directed infinite links for all voxel pairs $((x, y, z, t), (x, y, z, t-1))$
- **Encourage, but not impose shape growth:** replace directed infinite links by directed finite links
- **Inter-sequences inclusion constraint:** foreground in one sequence has to be included in foreground of another sequence
  - see figure
- **Weighting frames by reliability**
  - strong level of noise at time $t \rightarrow$ multiply E^t by a small reliability factor < 1

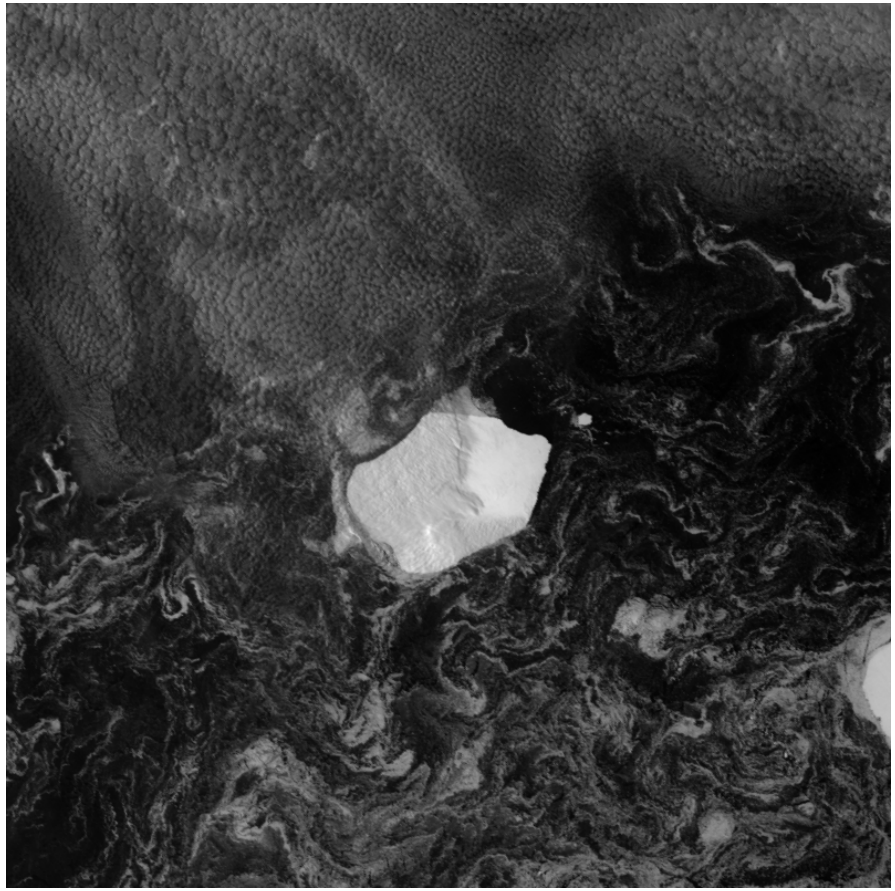


Segmenting jointly two sequences S1 and S2, by enforcing the foreground of S1 to contain the foreground of S2, with directed infinite links between all pixels of coordinates (x, y, t), from S1 towards S2

Experiments: shrinking ice floe segmentation



Original MODIS data



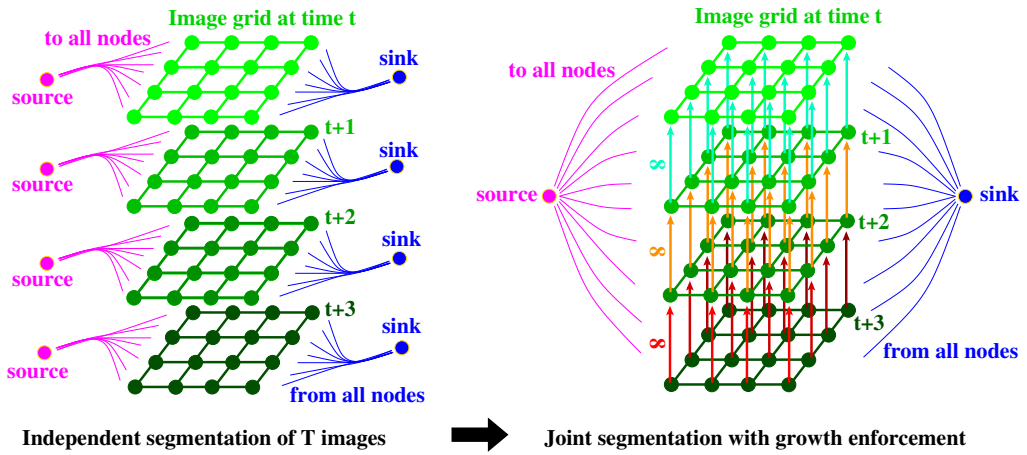
Graph-cut with directed infinite links

Manual segmentation

Dice score (DC) $= 0.980 \pm 0.007$

Comparison with other graph-cut-based methods

- *[w/o]* No temporal links = independent segmentation of each frame
- *[Feedforward]* Foreground pixels of the frame t are marked as seeds with infinite unary costs in the frame $(t+1)$
- *[Bi=const]* Bidirectional temporal links with a constant weight
- *[Bi=variable]* Bidirectional temporal links are computed based on intensity differences between pixels in successive frames [Wolz'10]

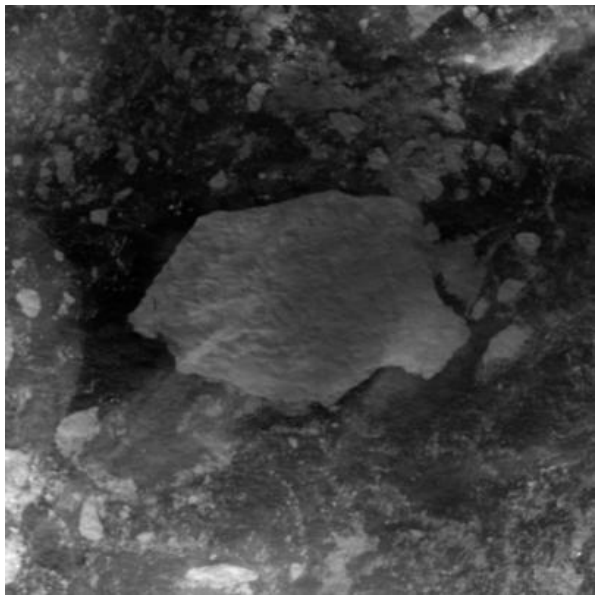


Comparison with other graph-cut-based methods

[w/o]

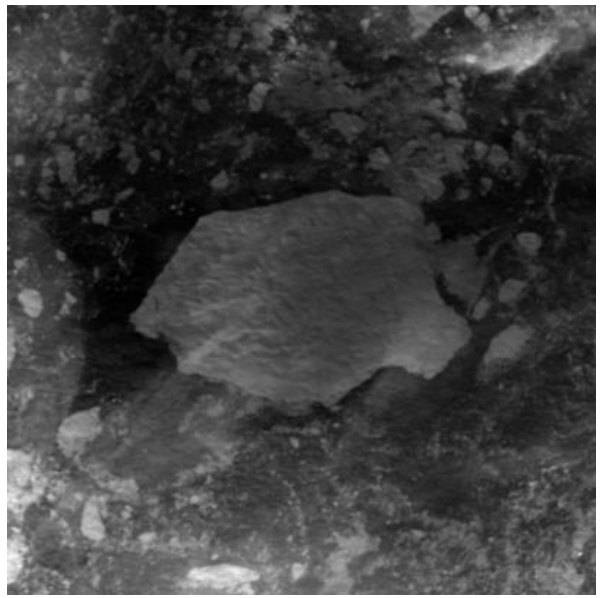
DC = .933 ± .099

[Feedforward]



DC = .554 ± .128

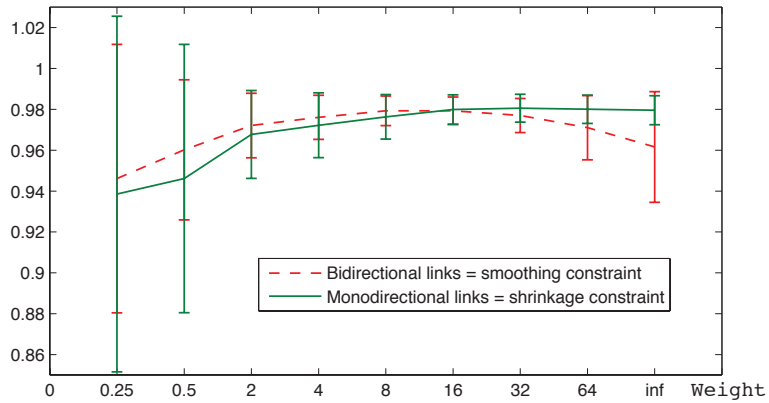
[Bi=variable]



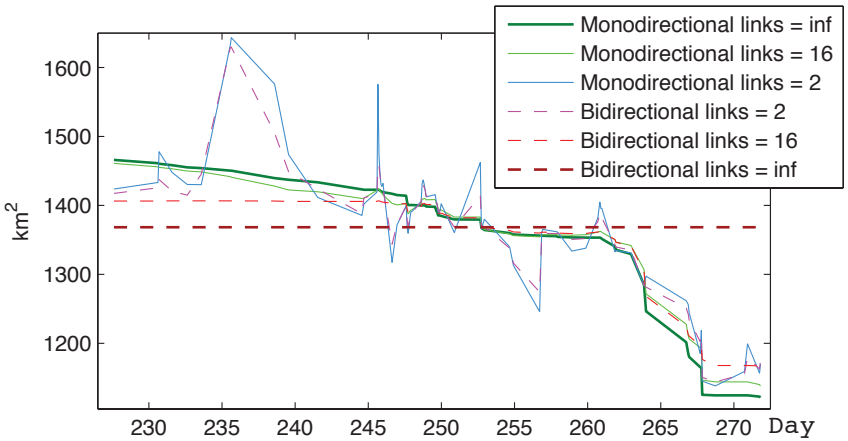
DC = .958 ± .048

- **Conclusion:** These methods are very sensitive to:
  - noise
  - variations of foreground/background intensities

Using temporal links with constant weights



Mean and standard deviation for the dice score as a function of the temporal link's weight, when using mono- and bidirectional temporal links



Area of a multiyear ice floe as a function of time, computed by using mono- and bidirectional links with different weights

Concluding remarks

- The main contribution:
  1. framework for segmentation of 2D/3D image time series with the constraint of shape growth/shrinkage,
  2. in order to be able to segment very noisy/low-contrast/incomplete data,
  3. in a very low computational time.

- Limitations:
  - Designed for specific shape priors
  - Limited scalability

} Convolutional neural networks?

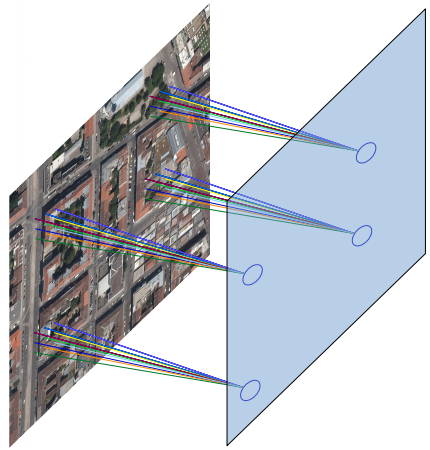
Outline

Convolutional neural networks (CNNs) [LeCun 1998]

- Jointly learn to extract contextual features & conduct classification

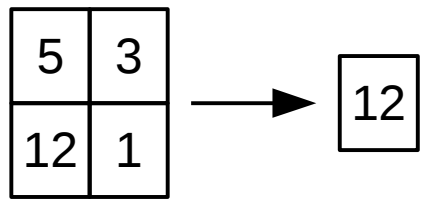
Convolutional layers:

- Only local spatial connections
- Location invariance
- Learned convolution filters $\rightarrow$ feature maps



Pooling layers: subsample feature maps

- Increase *receptive field* ☺
- Downgrade resolution
  - Robustness to spatial variation ☺
  - Not good for *pixelwise* labeling ☹



Max pooling

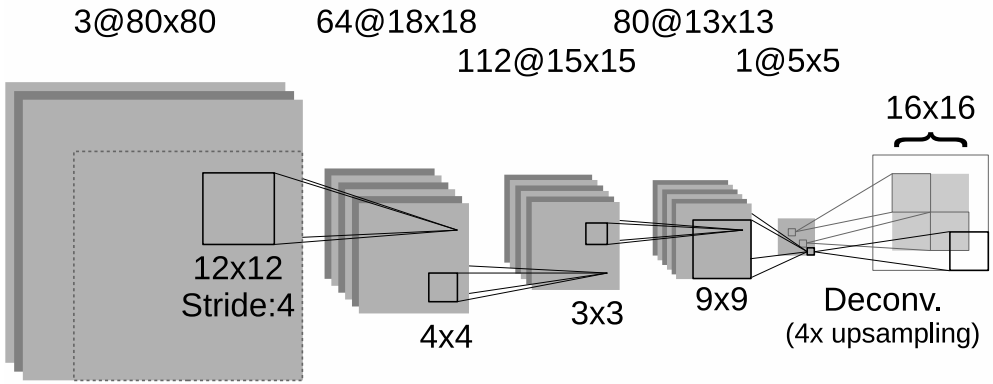
Remote sensing: *dense* classification with CNNs?

Fully convolutional networks (FCNs)

[Long et al., CVPR 2015]

- Interpolation with a learned kernel ("deconvolutional" layer)
- Lost resolution is upsampled

Proposed FCN for remote sensing

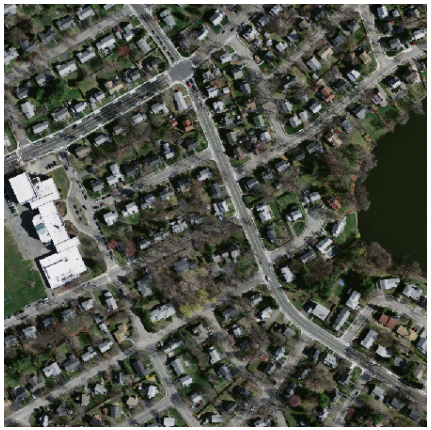


- Adapted from previous work (Mnih, 2013) and made it fully conv.
- 10x faster and more accurate

Classification with FCNs: some results

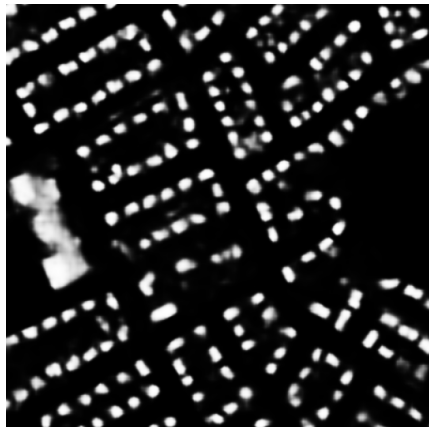
Massachusetts dataset

[Dataset: Mnih, 2013]



Color input

Reference



FCN

Pixelwise SVM

- Classification of 22.5 km^2 (1 m resolution): 8.5 seconds (2.7 GHz 8-core, Quadro K3100M GPU)

E. Maggiori, Y. Tarabalka, G. Charpiat, P. Alliez. "Convolutional neural networks for large-scale remote sensing image classification", IEEE TGRS 2017.

Yielding high-resolution outputs

Recognition/localization (RL) trade-off

Subsampling:

- increases the receptive field (improving recognition)
- reduces resolution (hampering localization)

Input Ref. FCN

Recent work

Three families of architectures:

- *Dilation* (Chen et al., 2015; Dubrovina et al., 2016,...)
- *Unpooling/deconv.* (Noh et al., 2015; Volpi and Tuia, 2016,...)
- *Skip networks* (Long et al., 2015; Badrinarayanan et al., 2015,...)

Goal: architecture that addresses RL trade-off

Premise

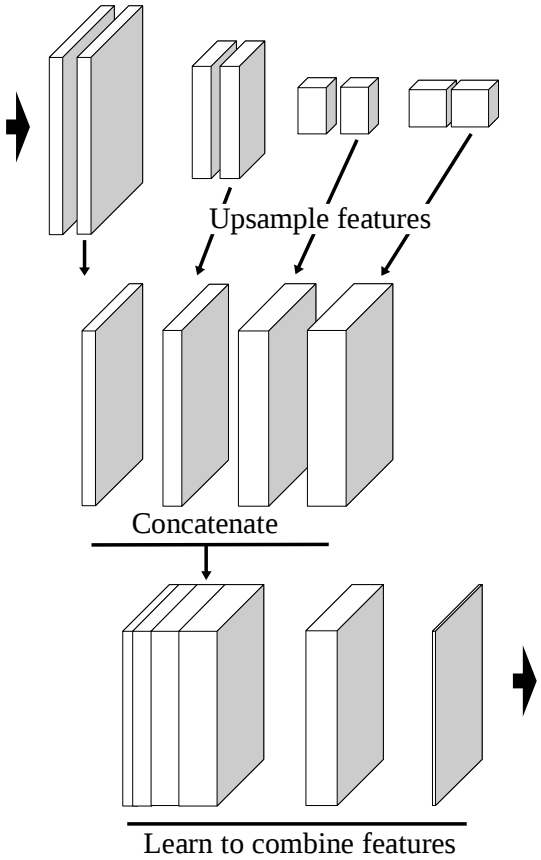
- CNNs do not need to "see" everywhere at the same resolution
- E.g., to classify central pixel:

Full resolution context

Full resolution only near center

$\Rightarrow$ Combine resolutions in a flexible way to address trade-off

Proposed method: MLP network



1. Base FCN

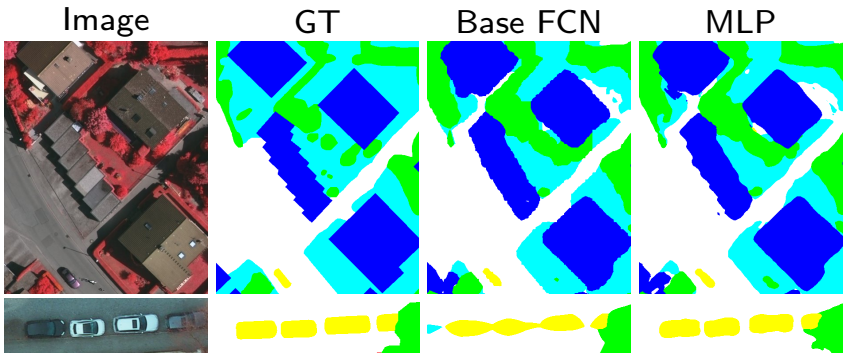
2. Extract intermediate features
   $\Rightarrow$ Pool of features

3. Multi-layer perceptron (1 hidden layer) learns how to combine those features
   $\Rightarrow$ Output classification map

E. Maggiori, Y. Tarabalka, G. Charpiat, P. Alliez. "High-Resolution Aerial Image Labeling with Convolutional Neural Networks", IEEE TGRS 2017.

Experiments

Vaihingen & Postdam ISPRS datasets:



Impervious surface (white), Building (blue), Low veget. (cyan), Tree (green), Car (yellow)

Vaihingen	Imp. surf.	Build.	Low veg.	Tree	Car	Acc.
CNN+RF	88.58	94.23	76.58	86.29	67.58	86.52
CNN+RF+CRF	89.10	94.30	77.36	86.25	71.91	86.89
Deconvolution						87.83
Dilation	90.19	94.49	77.69	87.24	76.77	87.70
Dilation + CRF	90.41	94.73	78.25	87.25	75.57	87.90
MLP	**91.69**	**95.24**	**79.44**	**88.12**	**78.42**	**88.92**

Submission to ISPRS server

- Overall accuracy: 89.5%
- Second place (out of 29) at the time of submission
- Significantly simpler and faster than other methods

Classifying cities over the earth: can CNNs generalize?

Inria Aerial Image Labeling Dataset (810 km^2):

Bellingham Innsburck San Francisco Tyrol

- Images over US and Austria with open images and building footprints
- Different cities in training and test sets

⇒ project.inria.fr/aerialimagelabeling

E. Maggiori, Y. Tarabalka, G. Charpiat, P. Alliez. "Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark". IGARSS 2017.

Inria Aerial Image Labeling Benchmark

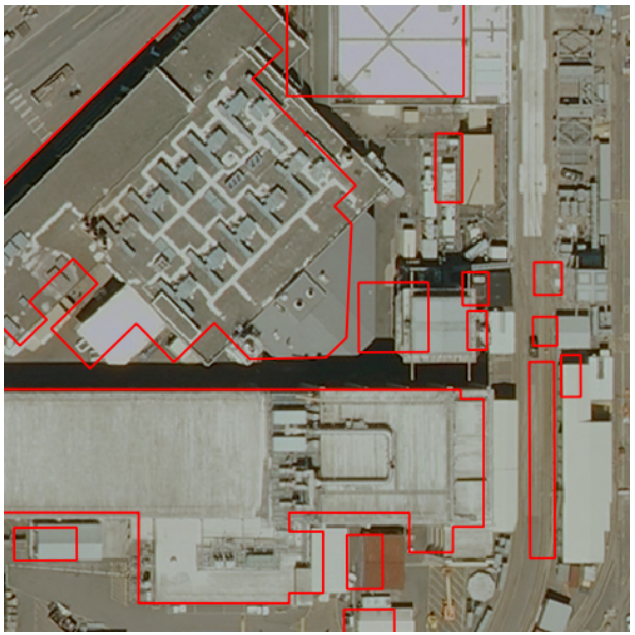
- Dec. 2016 - Nov. 2017: $>$ 500 downloads

Leaderboard

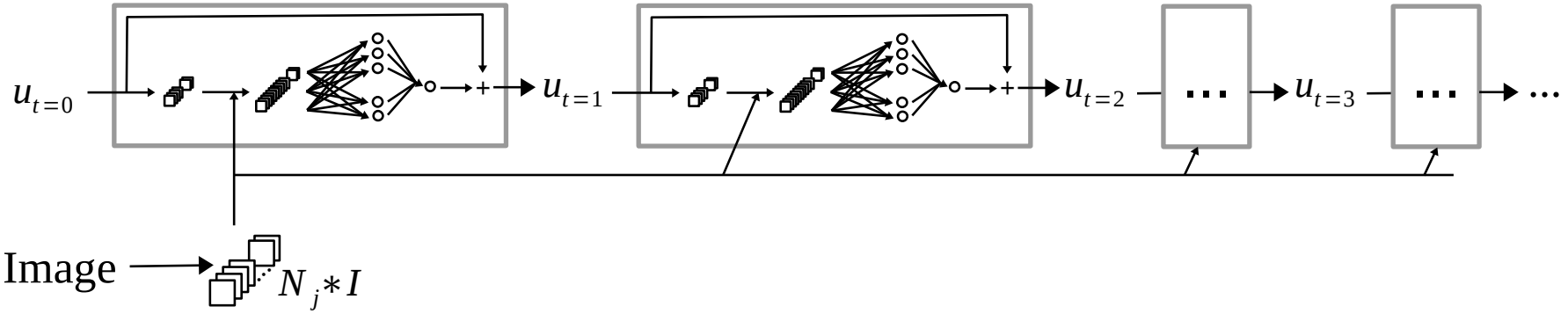
Method	Date	Bellingham		Bloomington		Innsbruck		San Francisco		East Tyrol		Overall	
		IoU	Acc.	IoU	Acc.	IoU	Acc.	IoU	Acc.	IoU	Acc.	IoU	Acc.
Inria1	3-Jan-17	52.91	95.14	46.08	94.95	58.12	95.16	57.84	86.05	59.03	96.40	55.82	93.54
Inria2	3-Jan-17	56.11	95.37	50.40	95.27	61.03	95.37	61.38	87.00	62.51	96.61	59.31	93.93
TeraDeep	5-May-17	58.08	95.88	53.38	95.61	59.47	95.26	64.34	88.71	62.00	96.57	60.95	94.41
RMIT	16-July-17	57.30	95.97	51.78	95.60	60.70	95.69	66.71	89.23	59.73	96.59	61.73	94.62
Raisa Energy	8-Sep-17	64.46	96.04	56.63	95.38	66.99	95.97	67.74	87.48	69.21	96.92	65.94	94.36
Duke University	6-Nov-17	66.90	96.69	58.48	96.15	69.92	96.37	75.54	91.87	72.34	97.42	70.91	95.70
NUS	13-Nov-17	65.36	96.34	58.50	95.95	68.45	96.21	71.17	90.08	71.58	97.32	68.36	95.18
Onera 1	14-Nov-17	63.42	96.11	62.74	96.20	63.77	95.44	66.53	89.18	65.90	96.76	65.04	94.74
Onera 2	14-Nov-17	68.92	96.94	68.12	97.00	71.87	96.72	71.17	89.74	74.75	97.78	71.02	95.63

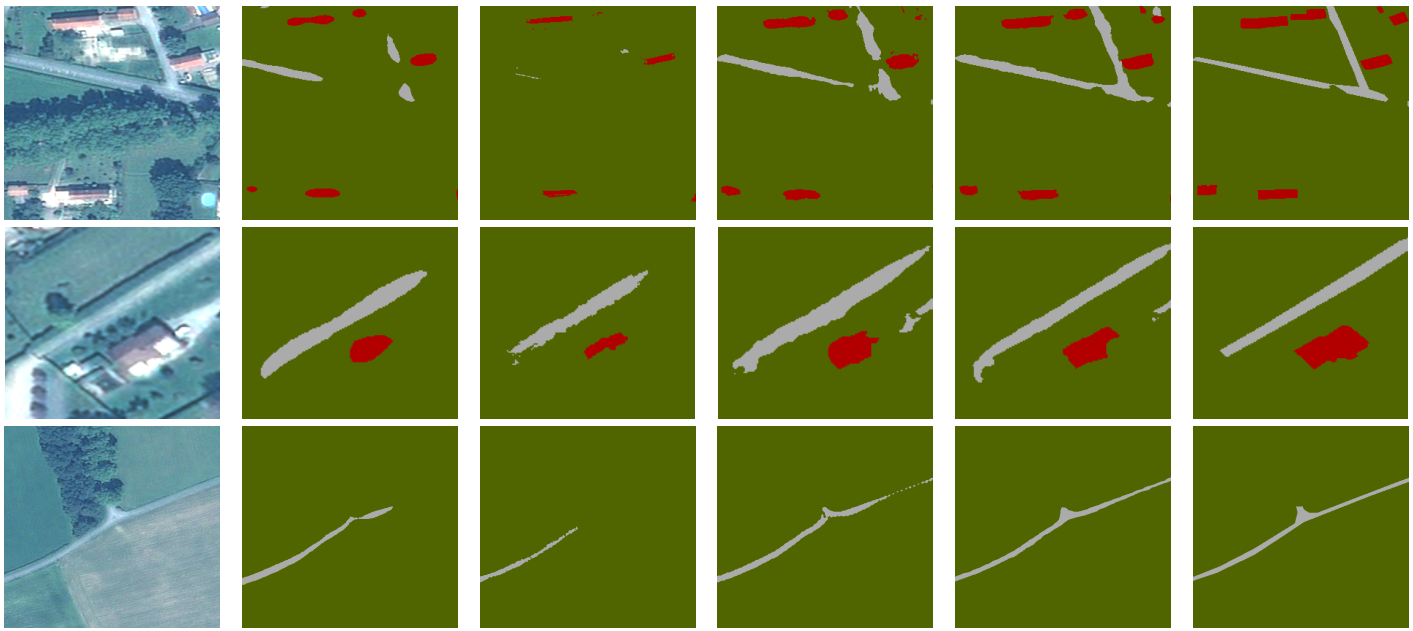
Dealing with imperfect training data

- Frequent misregistration/omission in large-scale data sources
  - Example: OpenStreetMap data are mostly misaligned with satellite data



Recurrent neural networks to enhance classification

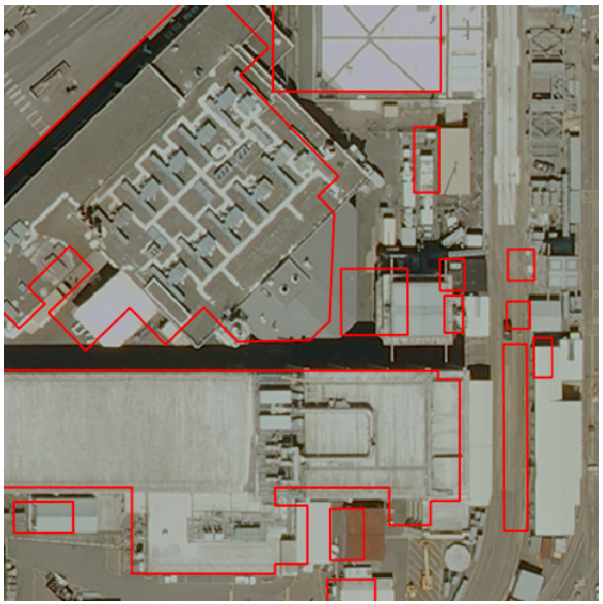


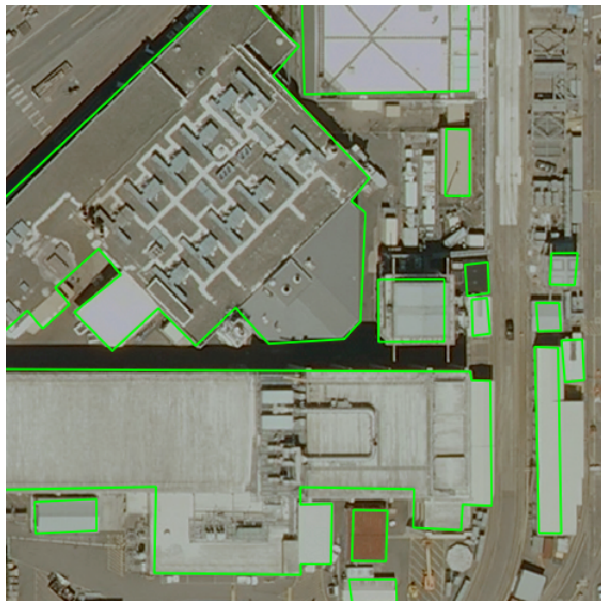


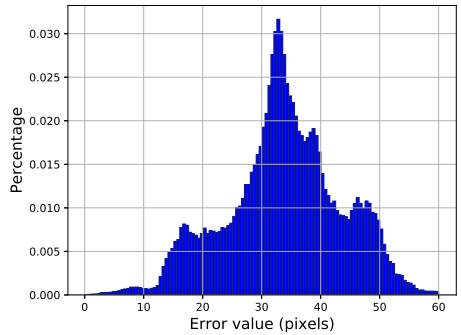
Color image | Coarse CNN IoU = 48% | CNN+CRF IoU = 44% | Class-agnostic CNN+RNN IoU = 65% | CNN+RNN IoU = 73% | Reference

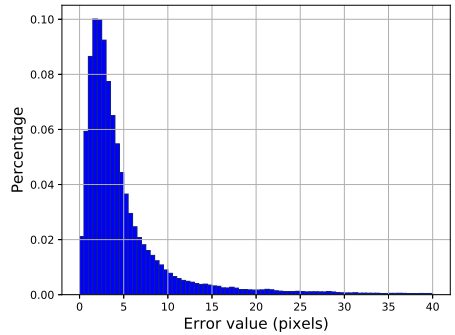
Fully-convolutional net for multimodal image registration

- Chain of scale-specific neural networks to solve alignment problem









Misalignment distribution before and after processing

A. Zampieri, G. Charpiat, Y. Tarabalka. "Coarse to fine non-rigid registration: a chain of scale-specific neural networks for multimodal image alignment with application to remote sensing", submitted to CVPR 2018.

Concluding remarks

- CNN-based architectures for image classification and alignment have been developed
- CNNs exploit the properties of images particularly well
- Shifting efforts from feature engineering to network engineering
- Good *payoff* of the efforts,
  e.g., learning better features than handmade ones,
  convolutions $\rightarrow$ GPUs, borrowing pretrained network

Outline

Conclusions

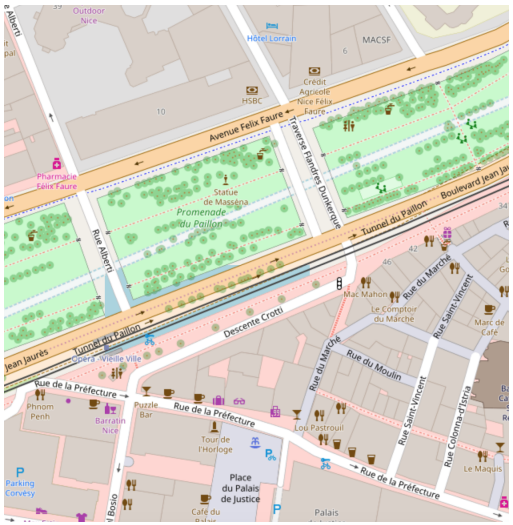
There is no such thing as a universally better classifier

- For specific applications:
  - Manually designed features and priors may work very well
- To classify remote sensing images on a world-scale:
  - Learning methods must be **generic** and **highly scalable**
  - CNNs have shown a remarkable computational performance
    - Capable to learn expressive multi-scale contextual features
    - Succeed in classifying new unseen earth areas
  - Still significant work to be done to design automatic mapping systems

Future work

2017-2021: **ANR JCJC project EPITOME**

- **Epitome** = summary, an instance that represents a larger reality
- **Objective:** devise novel epitome-like, or **summary, representations** for large-scale satellite images:
  - **Generic** = applicable for diverse images & applications
  - **Structure-preserving** = represent meaningful objects within images
  - We opt for **multi-resolution vector-based** representations



ANR project EPITOME - Approach

- Learn about both geometric and semantic structures & their accurate scale alignment
  - Use crowd-sourced maps to derive weakly labeled training data
  - Solve multimodal alignment problem
  - CNN segmentation to update/correct maps
- Devise vector-based epitome representation + algos for generation & manipulation
  1. Take inspiration from recent multi-resolution image vectorization techniques
     - Preserve geometric & semantic structures discovered by learning
  2. Learn in a space of vector primitives

Future work

- ANR PRC project "Properties of faults, a key to realistic generic earthquake modeling and hazard simulation" (PI: Geoazur)
  - VRH DEM construction from multiple Pléiades images to recover complex topology
  - CNN-based learning to identify and measure properties of the damage faults
- FAPESP project with INPE, Brasil
  - Registration of UAV images with satellite images in tropical forests
- IRT project with Thales Alenia Space
  - Onboard image processing for autonomous space systems

Yuliya Tarabalka - Research scientist at Inria

- Publications: 65 scientific articles (>3000 citations, h-index = 18)
- Teaching: $\sim$ 96 h/year
  - Discrete inference & learning (ENS Paris-Saclay), Optimization & Math. methods (CentralSupelec), Advanced algos (IUT Nice)
- Supervision:
  - 2010-2017: 2 PhD and 8 MSc students
  - 2017-pres: 3 PhD students
    - L. Matteo "Study of faults from VHR satellite images"
    - O. Tasar "Learning approaches for efficient image representations"
    - N. Girard "Methods to structure large-scale satellite data"
- Professional service:
  - PhD thesis committees, expert panels, editor & reviewer for internat. journals, program & organizing committees for internat. conferences

Thank you for your attention!

Questions?