Satellite Image Mining: Indexing and Retrieval

Master AIC (Apprentissage, Information et Contenu) and D&K (Data & Knowledge) – Université Paris Saclay

Henri Maître - Télécom ParisTech
henri.maitre@telecom-paristech.fr
# Content

## I - Remote Sensing (RS) and RS Images
- Why remote sensing? ............................................. 3
- Preparing a RS program ....................................... 4
- Image parameters (resolution, spectral bands, repetition …) 8
- Image diversity (12 not presented slides) .................. 14

## II - RS Image mining
- RS archiving problems ........................................ 23
  - RS image mining IS NOT multimedia image mining .... 25
  - RS image mining specificity ................................. 28
- Hand-crafted features and classification .................... 35
  - Expert in the loop, relevance feedback ................. 45
- Deep Neural Networks ........................................... 57
  - DNN toolbox ................................................ 60
  - Some instances ............................................. 64
- From Lo to Hi level of semantics ............................ 67

09/01/2019  Master AIC
Part I - Remote Sensing and Remote sensing images

Why? How? For Whom?
Why do we need Remote Sensing

Environment:

- Meteorology: short-term weather prediction
- Climate: long-term monitoring
  - volcanos
  - earthquake, tsunamis, floods
  - Industrial hazards
  - Marine pollution
Why do we need Remote Sensing

- **Agriculture:**
  - Survey and evaluation of crop & farming production
  - Fish & Aquaculture resources management
  - Forestry resources planning
  - Water management, dams, watering
  - Desertification & urban pressure
Why do we need Remote Sensing

- **Town & country planning:**
  - Mapping and inventories
  - Constructions & public work: railways, airports, harbours, dams, …
  - Cities and Mega-cities management
  - Management of moving populations, displacements, installation
  - Climatic impact management
  - Crisis management: fires, floods, …
Why do we need Remote Sensing

- Defence & Security applications:
  - Military deployment preparation
  - Military mission debriefing
  - Intelligence and survey of national/foreign territory
How to prepare a remote sensing program
How is prepared a remote sensing program

- Conceive the sensor: application, customers
- Determine which satellite / which launcher
- Conceive the ground-station and the data management process: economical, social and technical issues

→ 15 to 20 years
Satellite links with the Earth

- **S Band Station Emitter/Receiver**
  - S-Band Antenna
  - Telemetry link 1.6 Mbps
  - Telecontrol link 4 kbps

- **X Band receiving Station**
  - X-Band Antenna
  - Image down link 250 Mbps

- **Operating and Control Decision Center**
  - Acquisition Programming

- **Processing center**
  - Request for an image
  - Images

- **User**
  - Images
Satellite: orbit choice

- Mecanics laws:
  - Newton = centripetal force
  - Satellite speed = driving force
  ➜ elliptical or circular trajectory (Kepler)

\[
\vec{F} = -\frac{\mu m \vec{r}}{r^2}
\]
Orbit choice

1) Geostationary
   - Always in the Equator Plane
   - Always at vertical of the same point on the Equator
   - Altitude ~ 36 700 km
   - Field of view: ~1/3 Earth: always the same

   - Applications: meteo, survey of catastrophies, telecoms, TV
2) Processing satellite (low orbit)

- Altitude ~ 800 km (down to 250 km)
- Circular ~ N/S
- Trajectory: ± polar
- ~ 15 revolutions / day
- Helio-synchronous
Choice of resolution

- Pixel size = smallest measured terrain on the ground
  - from 30 cm to 10 km

SPOT 5
\[ \Delta x = 2.5 \text{m} \]
On Ground resolution

- **Depends on:**
  - **Sensor:**
    - Photosites size: \( \delta x \)
    - \( G = \frac{f}{D} \) = enlargement
    - \( \Delta x = \frac{\delta x}{G} \) = smallest detail
  - **The camera lens**
    - \( \delta' x = \frac{\lambda f}{d} \) = diffraction limited resolution

\[
\Delta x_{\text{min}} = \frac{\lambda f}{G d} = \frac{\lambda D}{d} \quad \Rightarrow \text{Smallest detail}
\]

\[D = \text{satellite-Earth distance} \sim 1\ 000 \ \text{km} = 10^6 \ \text{m}\]

\[\lambda = \text{wave length} = 0,5 \cdot 10^{-6} \ \text{m}\]

\[d = \text{lens diameter} \sim 0,5 \ \text{m}\]

\[\Delta x_{\text{min}} = 1 \ \text{m}\]

Possible with: \( f = 1 \ \text{m} \)

if \( \delta' x = \frac{\lambda f}{d} = 1 \mu\text{m} \)

the photosite measures \( 10^{-6} \ \text{m} \)
Often *push-broom sensor*

- **Sensor size along track:**
  - On line sensor
  - \( \text{speed} \times \text{aperture time} \)

- **In the other direction**
  - Number of sensors on a line
  - From 6,000 to 40,000

- **Resolution:**
  - Depends on the lens
Swath choice

- **Swath = image width**
  - from 10 km to 10 000 km
  - = from 3 000 to 40 000 pixels / line
  - Given by the sensor size
  - Limited by the communication link with Earth

- **Revisit delay**
  - **15 min for geostationnary sat. (to dump the memory)**
    - from 1h30 (min) to 1 month for processing satellites
    - But … sensor agility!
Video possibility

- **Angle of view ~ + or – 50 degrees:**
  - MN ~ 2000 km
  - 1 rotation around the Earth = 90 min ~ 40 000 km
  - Time to go from M to N
    \[ = \frac{90 \times 2000}{40000} = 4 \text{ min 30 s} \]
Which wave length?

1 – Passive sensors: measure the energy sent back from Sun by Earth or the energy radiated by Earth

- Emitted from the Sun (Wien’s law) \times \text{Atmosphere transparency} \times \text{Ground Reflexion}
- Black and White (Panchromatic)
- Visible = Blue - Green - Red
- Visible and Near Infra-Red: G - R - IR = false colors
- Multispectral : 7 \rightarrow 20 channels
- Hyperspectral : 64 \rightarrow 512 channels
False colors:  \( \text{NIR-R-G} \rightarrow \text{R-G-B} \)

vegetation = red

False colors:  

True colors:
Multispectral image visualisation: pseudo colors

Landsat = 7 channels

321
(a) combination 321

432
(b) combination 432

542
(c) combination 542

435
(d) combination 754

754
(e) combination 435

1 2 3 4 5 6 7
R G + B

41(7+5)

© UVED
Which wave length?

2 – Active sensors: EM emitter + receiver

radar = Micro waves: $\lambda = 1$ cm to 10 m

- But low resolution: $\Delta x = \frac{\lambda f}{Gd}$
- With complex processing: Synthetic Aperture Radar $\Rightarrow$ hi resolution
Real antenna is too small, it covers a very large field
When moving, superposition of the seen areas

One point is seen from several antenna positions
From computation we obtain an accurate information = synthetic antenna
Satellite images = big data!

- Television HD: 1,280 x 720 pixels
- Television 4k: 4,000 x 2,000 pixels
- PC display screen: 1,600 x 1,200 pixels
- Photo camera: 5,000 x 4,000 pixels
- Spot 1 … 4: 6,000 x 6,000 pixels
- SPOT 5: 24,000 x 24,000 pixels
- Quickbird: 40,000 x 40,000 pixels

1,600,000,000 pixels = 1.6 Gpixels = 800 PC display screens

1 SPOT 5 image = 10 s of satellite run
Diversity of Remote Sensing Images (slides are not presented in the lecture notes)
Part II – Remote Sensing Image Mining
Remote Sensing Imaging: Archiving Problems and Issues
Satellite Image archives

- How can we store millions of images?
- How can we ensure durability of storage?
- How knowing that information exists?
- How retrieving information?
- How exploiting information?

➡ Data Mining directly on image files
When searching in a small set of images

➡ Indexing images when received
➡ data mining on index
When searching in large sets
RS Image mining IS NOT MultiMedia Image Mining
Mining in Multimedia Image databases

**Multimedia information retrieval:**
- Either from **semantic information**: name, description, caption, text (90% of Google-like retrieval)
- Or from **instance** (i.e. with a reference image) (Face or fingerprint recognition)

**I – Classical Machine Learning techniques (2000-2012)**
- Hand-crafted feature detection and/or salient point detection
- Classification in p-dimensional space
  - \( \rightarrow \) few parameters
  - \( \rightarrow \) few learning images (groundtruth) \(~ 1000\)

**II – Deep neural networks (2012 - …)**
- Directly with images as input and/or with extracted features
- Several +/- linear classifiers in cascade
  - \( \rightarrow \) thousands of parameters
  - \( \rightarrow \) hundred of thousands of images as groundtruth
Multimedia image mining: handcrafted features + classification

- Multimedia information retrieval from exemple:
  - Choices: to be robust to possible differences
    - scale, lighting, orientation, color, ...  \(\Rightarrow\) invariance
  - Strategy: detect invariant features
    - Histograms, color distribution, area-based segmentation, graph description, ...
    - Textures
    - Salient point detection: Harris, SIFT, SURF, ...
  - Represent the image as a vector in a \(p\) dimensional space \(\mathbb{R}^p\)
  - Classification: Bayès, k-NN, dynamic clustering, SVM (Support Vector Machine), Graph tree, ...
### Ambiguous semantics: Venus

<table>
<thead>
<tr>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
<th>Image 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Venus" /></td>
<td><img src="image2" alt="Venus" /></td>
<td><img src="image3" alt="Venus" /></td>
<td><img src="image4" alt="Venus" /></td>
</tr>
<tr>
<td><img src="image5" alt="Venus" /></td>
<td><img src="image6" alt="Venus" /></td>
<td><img src="image7" alt="Venus" /></td>
<td><img src="image8" alt="Venus" /></td>
</tr>
<tr>
<td><img src="image9" alt="Venus" /></td>
<td><img src="image10" alt="Venus" /></td>
<td><img src="image11" alt="Venus" /></td>
<td><img src="image12" alt="Venus" /></td>
</tr>
<tr>
<td><img src="image13" alt="Venus" /></td>
<td><img src="image14" alt="Venus" /></td>
<td><img src="image15" alt="Venus" /></td>
<td><img src="image16" alt="Venus" /></td>
</tr>
</tbody>
</table>

---

*Institut Mines-Télécom*

*09/01/2019 Master AIC*
Textual categorisation
invariance
Salient points: SIFT

\[ L(x, y, \sigma) = G(x, y, \sigma) \ast I(x, y), \]
\[ \frac{\partial G}{\partial \sigma} = \sigma \nabla^2 G. \]

\[ D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) \ast I(x, y) \]
\[ = L(x, y, k\sigma) - L(x, y, \sigma). \]

\[ D(x) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x \]

\[ \hat{x} = -\frac{\partial^2 D^{-1}}{\partial x^2} \frac{\partial D}{\partial x}. \]

\[ \frac{\text{Tr}(H)^2}{\text{Det}(H)} < \frac{(r + 1)^2}{r} \]
Specificities of RS Image mining
Category-based retrieval in specific data-bases

- **Examples:**
  - Biomedical
  - Biology
  - Astronomy
  - Remote sensing and satellite images

- **Goal:** to retrieve images « looking the same » as a given sample in very specialized data-bases

- Different from: retrieving the exact object in a very broad data-base
Satellite images

- A very specific content
A same region, different signals

From: Tong et al. arXiv 1807.05713 - 2018
The role of scale

High-Badakchan, Tadjikistan - Ikonos

15 m

1 m
Main scales

- **<1 meter = Very high resolution**: fine details in urban context, cars, pedestrians, containers, fences, small boats, ...

- **1 m < ... < 5 m = High resolution**: urban fine structures, houses, streets, gardens, individual trees, railway & road networks, ...

- **5 m < ... < 30 m = Middle resolution**: fine landcover, coarse urban structure: dense urban, residential or commercial areas,

- **> 30 m = low resolution**: global landcover
Available information on satellite images (semantic information) = Ancillary data

- **Accurate positionning in universal geographical references**: UTM, Mercator, Lambert, etc…
- **Precise time referencing**: seasonal variations (vegetation, insolation, agricultural production, …), sun positionning (shadows), tide effects (precise coast-line, harbours and fishing activities), meteorological conditions (snow, floods, …)
- **Satellite parameters**: resolution, spectral sensitivity, noise
- **Often**: Image quality, cloud cover
Satellite images

- **What are we looking for?**

  It is not clear!

  - **Precise objects:**
    - A boat
    - A building
    - A road-crossing
    - An airplane landing area
  
  - **Generic objects:**
    - A marina
    - Greenhouse cultures
    - Oil pipeline
    - A forest fire
    - Refugee camps
    - Typhoon hazards
  
  - **Specific terrain configurations:**
    - Conducive to: ... floods, ... desertification, ... urban pollution, ...
    - Conducive to: ... build a factory, ... plan a bombing, ... cultivate marijuana
Spatial scale vs. Semantic complexity

Semantic Complexity

pixels regions zones

greenhouses flower culture intensive farming middle-age city marina

greenhouses flower culture intensive farming middle-age city marina

greenhouses flower culture intensive farming middle-age city marina

greenhouses flower culture intensive farming middle-age city marina

greenhouses flower culture intensive farming middle-age city marina

greenhouses flower culture intensive farming middle-age city marina

greenhouses flower culture intensive farming middle-age city marina
Hierarchical representation

- **Pixel**
  - spectral properties (R,G,B,IR)
  - contrast / texture
  - edges, contours

- **Region**
  - form / shape

- **Objects**
- **Scene**
  - Increasing semantics
  - sea
  - warehouse
  - house
  - network
  - wharf
  - fields
  - Master AIC
RS image processing & hand-crafted feature detection
Mining in RS Image databases

- Semantic information retrieval:
  - From ancillary data

I – Classical Machine Learning techniques (2000-2012)
- Image Processing
- Hand-crafted feature detection and/or salient point detection
- Classification in p-dimensional space
  - \( \Rightarrow \) few parameters
  - \( \Rightarrow \) few learning images (groundtruth) \( \sim 1000 \)

II – Deep neural networks (2012 - …)
- Directly with images as input and/or with extracted features
- Several +/- linear classifiers in cascade
  - \( \Rightarrow \) thousands of parameters
  - \( \Rightarrow \) hundred of thousands of images as groundtruth
Probabilistic evaluation

Image space $I$  
Class space $\omega$  
Semantic label space $L$

$p(\omega_i | I)$  
$p(L_i | w)$
Hand crafted features

- **Radiometry**
  - Multispectral: channels
  - Specific combinations for remote sensing: NDVI + IB + ISU

- **Textures**
  - Gabor Filters
  - Haralick cooccurrence matrices and their descriptors
  - Quadratic Mirror Filters (wavelets)
  - Contourlet decomposition
  - Steerable wavelets
  - Markov random fields parameters (Gaussian, Laplacian, Log-laplacian …)

- **Structures**
  - Contours, regions, lakes, forests, deserts
  - Objects: roads, buildings, rivers, lakes
  - Roads, Train or River networks
Some efficient choices

- **Indexing:** small subimages: (~ 64 x 64 pixels) = 320 m x 320 m on the ground for SPOT 5 images

- **Mixed features:**
  - Radiometry (Panchro only)
  - Structure (contours)
  - Wavelets: 2 directions, 4 scales

- **Automatic feature selection (supervised):** ReliefF, Fisher FS, SVM-RFE or Unsupervised: MIC (*Max Information Compression*), k-means FS
  ~ 100 features with redundancy or ➔ 10 to 20 features without redundancy

- **Give names to classes** (*from label to name*)
  - Waste fields
  - Cultures
  - Housing
  - Road and river networks
Classification

Many different classifiers:

- MAP & Bayes decision
- K-nearest neighbours
- Graph tree
- Kernel methods (SVM = Support Vector Machine)
- Hierarchical clustering

label = 24
or
Semantic labelling
name = « Corn field »

Supervised
or
Unsupervised
# Supervised classes

<table>
<thead>
<tr>
<th>Supervised classes</th>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
<th>Image 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residential areas</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>Planes</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>Industrial tanks &amp; cisterns</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>Railway marshalling yard</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
</tbody>
</table>
## Supervised classes

<table>
<thead>
<tr>
<th>Class</th>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
<th>Image 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>factories</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>Dense urban area</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>villages</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>Urban parks</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
</tbody>
</table>
## Supervised classes

<table>
<thead>
<tr>
<th></th>
<th><img src="image1.png" alt="Image" /></th>
<th><img src="image2.png" alt="Image" /></th>
<th><img src="image3.png" alt="Image" /></th>
<th><img src="image4.png" alt="Image" /></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Graveyards</strong></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>Road interchange</strong></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>Castle parks</strong></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>Parking lots</strong></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
</tr>
</tbody>
</table>
How to express results?

- **Classification rate**  97.3 %  (or **error rate**: 2.7 %)
- **Confusion matrix**

<table>
<thead>
<tr>
<th></th>
<th>Present object</th>
<th>Absent object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive detection</td>
<td>True positive (TP)</td>
<td>False positive (FP)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(type I error)</td>
</tr>
<tr>
<td>Negative detection</td>
<td>False negative</td>
<td>True Negative</td>
</tr>
<tr>
<td></td>
<td>(type II error)</td>
<td></td>
</tr>
</tbody>
</table>

- **Receiver Operating Characteristic (ROC Curve)**

Convert TP and FP into FPR and TPR $\in [0,1]$

Plot $TPR = f(FPR)$ for many different parameters

Without specific instruction, take the closest point from $A = (0,1)$ as working condition
Sub image classification (128 x 128):

- city, wood, fields, sea, desert & clouds

600 images for each class

Results: Gaussian SVM, Mean error 1.4% ± 0.4%
(147 features, cross validated)

| True
| Found (%) | city | clouds | desert | fields | woods | sea |
|----------|-----------------|-------|--------|--------|--------|------|-----|
| city     | 98.8            | 0     | 0      | 0.5    | 0      | 0    | 0   |
| cloud    | 0               | 99.3  | 0.2    | 0      | 0      | 0    | 0   |
| desert   | 0               | 0     | 99.0   | 0.3    | 0      | 0    | 0   |
| fields   | 0.5             | 0.2   | 0.8    | 98.1   | 0.3    | 0.4  |     |
| woods    | 0               | 0.2   | 0      | 0      | 98.0   | 1.4  |     |
| sea      | 0.7             | 0.3   | 0      | 1.0    | 1.7    | 98.2 |     |
How many features?

Automatic feature selection
• Wrappers
• Filters (mutual information)
• Embedded (Lasso)
Using a human expert to improve learning
Learning with Relevance feedback

- Man Machine dialog

- Subjective

- Objective
- Database composed of 600 SPOT5 images divided in 6 classes
- Used features: Gabor, Haralick, QMF and GMRF
- Gaussian Kernel
- System evaluation: Precision-Recall graphs
Deep Neural Networks
Mining in RS Image databases

- Semantic information retrieval:
  - From ancillary data

I – Classical Machine Learning techniques (2000-2012)
  - Image processing
  - Hand-crafted feature detection and/or salient point detection
  - Classification in p-dimensional space
    - few parameters
    - few learning images (groundtruth) ~ 1000

II – Deep neural networks (2012 - …)
  - Directly with images as input and/or with extracted features
  - Several +/- linear classifiers in cascade
    - thousands of parameters
    - hundred of thousands of images as groundtruth
Some references (dated 01/01/2019)

- Penatti, O. A., Nogueira, K., & dos Santos, J. A. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 44-51).
Deep Neural Network

**Which input?**
- Raw image
- Processed image (filtered, segmented …)
- Feature detected image (classified, edge detected, …)
- Features

**Which architecture?**
- # layers,
- type of layers

**Which protocol?**
- Feature learning
- Fine tuning

**Which output?**
- Densely classified image
- Detected targets
- List of targets
- List of Features

From: I. Bloch, AIC
CNN basic components

- **Convolutional layer**: with \( r 	imes r \) kernel – down scaling

- **Nonlinearity**: sigmoïd or RELU (rectified linear unit)

- **Pooling layer**: single value taken from a set of values - ex: \( \text{max} \) on a \( r 	imes r \) patch

- **Autoencoder**: symetrical NN to reduce the model dimensionality
CNN basic components

- **Fully convolutional layer**: to perform a large distance context dependance

- **Transfer coding**: to learn from a database and use for another one

- **Fine Tuning**: to specify a network to a given task after training on a general purpose data base

- **Yoyo architecture**: downsampling for feature extraction then upsampling for fine positioning of targets
Most used components for RS-CNN (2019)

- CNN from the Pattern Recognition community
  - AlexNet
  - GoogleNet
  - VGGNet
  - ResNet
  - Inception

- Training sets
  - ImageNet (General purpose image library for pattern recognition)
  - UC Merced DataSet (Aerial images / 21 classes)
  - OSM - OpenStreetMap (Aerial Image Database)
  - Google Street Map (hi level semantic)
  - NLCD - USGS data Base (Geological survey)
  - Corinne Landcover (Agriculture & vegetation)
  - Gaofen Image Dataset (GID) (Hi Resolution Satellite)
  - ...
Instance # 1: Basic CNN (DLR)

- With UC Merced Land database (aerial / 21 classes)
- With pre-trained CNN (Imagenet)
- Fine-tuned full convolutional layers with enhanced data

---

**Table II: Classification Components and Algorithm Comparison**

<table>
<thead>
<tr>
<th>Method &amp; Algorithm</th>
<th>Test-set Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest with RGB feature</td>
<td>44%</td>
</tr>
<tr>
<td>CNN with RGB feature</td>
<td>44.5%</td>
</tr>
<tr>
<td>Random Forest with Overfit features</td>
<td>86.9%</td>
</tr>
<tr>
<td>CNN with Overfit feature</td>
<td>92.4%</td>
</tr>
</tbody>
</table>

---

**Table III: Method Comparison Over the UCML Benchmark**

<table>
<thead>
<tr>
<th>Method &amp; Algorithm</th>
<th>Test-set Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOVW [2]</td>
<td>71.8%</td>
</tr>
<tr>
<td>SPMK [1]</td>
<td>74%</td>
</tr>
<tr>
<td>SPCk+[2]</td>
<td>76%</td>
</tr>
<tr>
<td>Sparse Coding [4]</td>
<td>81.7%</td>
</tr>
<tr>
<td>Salient Unsupervised Learning [6]</td>
<td>82.8% ± 1.18%</td>
</tr>
<tr>
<td>MinTree + KD Tree [3]</td>
<td>83.1% ± 1.2%</td>
</tr>
<tr>
<td>CNN with Overfit feature</td>
<td>92.4%</td>
</tr>
</tbody>
</table>
Instance # 2: fully CNN (Inria)  

Maggiori et al. IEEE TGRS, Feb 2017

**Patch-based CNN**

- 3@64x64
- 64@14x14
- 112@11x11
- 16@9x9
- 12x12
- Stride: 4
- 4x4
- 3x3

**Fully convolutional Patch-based CNN**

- 3@80x80
- 64@18x18
- 112@15x15
- 80@13x13
- 1@5x5
- 12x12
- Stride: 4
- 4x4
- 3x3
- 9x9
- Deconv. (4x upsampling)

**Graph:**

- True Positive rate vs False Positive rate
- Patch-based
- Fully convolutional

**Images:**

- (a) Image
- (b) Ground truth
- (c) Patch-based
- (d) Fully convolutional
- (e) SVM
Instance # 3 : RS CNN (Liemars/Wuhan)

Figure 1: Sample selection for model fine-tuning.
Instance # 3: RS CNN (Liemars/Wuhan)

Cooperation between classifying (sparse) and segmenting (dense)

From: Tong et al. arXiv 1807.05713 - 2018
Figure 8: Land-use classification maps of the GF-2 image obtained in Dongguan, Guangdong Province on January 23, 2015. (a) Ground truth. (b)-(f) Results of eCognition, RF+Fusion, SVM+Fusion, PT-GID, and FT-\(U_{tg}\).

From: Tong et al. arXiv 1807.05713 - 2018
From Low to High Level - Changing the scale
Complexity of images

Analysis window: real size
128 x 128 pixels

Analysis window: enlarged
Hierarchical representation

Two goals:
- Enlarge the field of view
- Increase the semantic level

Grouping strategy:
- Sliding window
- Pyramid
- Growing and Merging

Decision strategy:
- Bag of Visual Words (BOVW)
Increasing the semantics

Park = \{trees + fields + tracks\}

Waste area = \{waste + lawns + trees + roads\}

Residential area = \{houses + lawns + pools + roads\}

Commercial area = \{buildings + houses + parking lots + waste\}
Probabilistic evaluation

\[ p(L_i | w) \]

Image space I \quad \text{Class space} \ \omega \quad \text{Semantic label space} L

\[ I_1 \quad \omega_{k,1} \quad \omega_{l,1} \quad L_1 \]
\[ I_2 \quad \omega_{k,1} \quad \omega_{l,2} \quad L_2 \]
\[ I_3 \quad \omega_{k,1} \quad \omega_{l,3} \quad L_3 \]
\[ I_4 \quad \omega_{k,2} \quad \omega_{l,3} \quad L_4 \]
\[ I_5 \quad \omega_{k,3} \quad \omega_{l,4} \quad \vdots \]
\[ I_6 \quad \vdots \quad \vdots \quad \vdots \]
\[ \vdots \quad p(\omega_i | I) \quad \vdots \quad \vdots \]

\[ p(L_i | w) \]
Decision making: Bag of Words

- 2 levels → **H=high** (unknown)  **L = low** (known)
- List of *N* classes at **H** = \{*c*<sub>1</sub>,*c*<sub>2</sub>,… *c*<sub>*N*</sub>\}
- At **H** : 1 super-region with *n* objects, each ∈ 1 class = *n* labels described by the ordered list of the probability (or the occurrence) of each class:
  \[R_k = \{p_1,p_2, \ldots p_n\}\]
- **Classify H according to the** \(R_k\)
  - Naïve Bayes : \(c^* = \text{argmax } p(c|x) = \text{argmax } p(c) \prod_{k=1}^{n} p(x_k|c)\)
  - Improving Naïve Bayes:
    - pLSA = Probabilistic Latent Semantic Analysis
    - LDA = Latent Dirichlet Analysis