STIC-AmSud First Meeting »

Santiago, June 17, 2014

Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes

Prof. Laurent Heutte

Laurent.Heutte@univ-rouen.fr

http://www.litislab.eu/Members/lheutte

University of Rouen, France

- Located in the north of Paris (100 km)
- 33000 students
- 7 Faculties (research/teaching)
  - ✓ Medicine
  - ✓ Sciences
  - ✓ Literacy
  - ✓ Law
  - ✓ Technology
  - ✓ Economic sciences
  - ✓ Psychology

Faculty of Sciences and Techniques

- 3300 students
- 66 diplomas
- 400 professors and researchers
- 200 administrative staff
- Faculty divided into:
  - ✓ 7 departments (teaching): Computer Science, Computer Engineering, Physics, Biology, Mathematics,…
  - ✓ 14 laboratories (research): LITIS, CORIA, IRCOF, … some may associated with CNRS, INRIA, INSERM,…

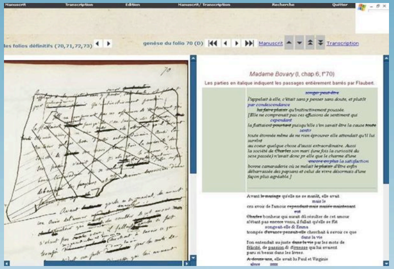
LITIS Lab. (http://www.litislab.eu)

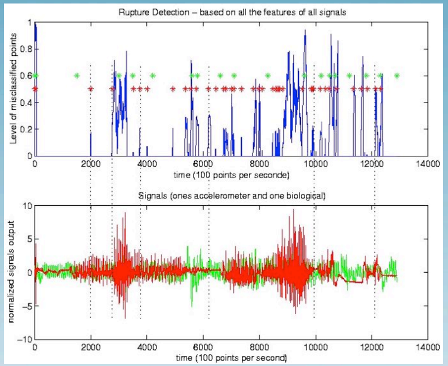
- Laboratory of Computer Science, Information Processing and Systems
- Depending on 3 organizations located in Upper Normandy: University of Rouen, University of Le Havre, INSA Rouen
- Scope: Sciences and Technology of Information and Communications
  - ✓ All formal and practical aspects of « information »
- 90 faculty members (whose 31 Prof, 5 Assoc. Prof., 54 Ass. Prof.)
- 7 research teams
- 80 PhD students and post-doc
- Head of the Lab: Prof. T. Paquet

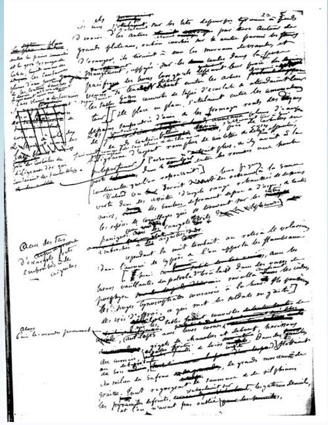
Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes

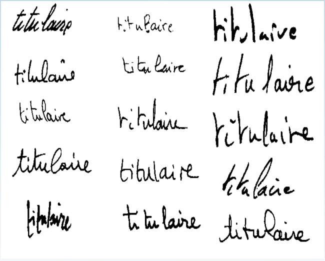
Document and Learning Team

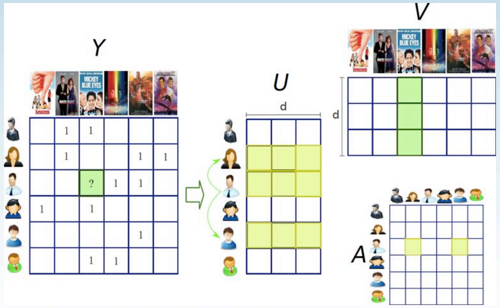
- Head: Prof. L Heutte
- 16,5 staff members – 7 PR, 9.5 Ass. Prof.
- 4 post-doc and enginers
- 16 PhD students

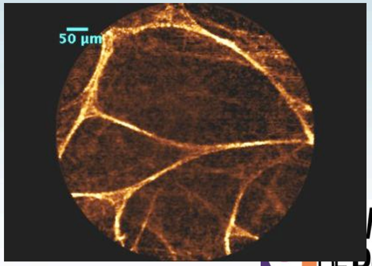


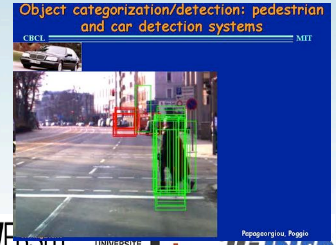












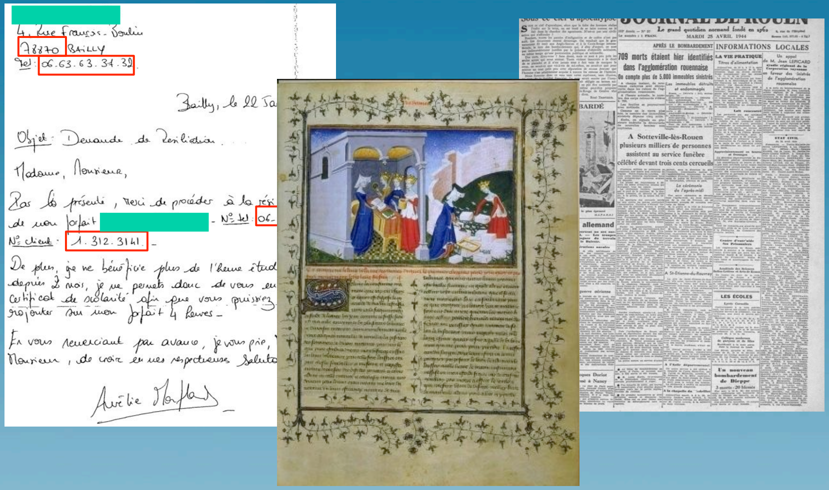
Scientific Issues

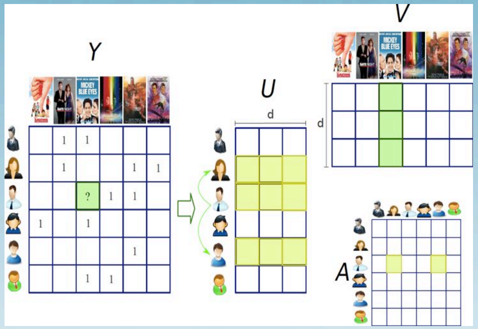
- Machine Learning and Pattern Recognition
- Joint learning of representations and decisions
  - ✓ Dictionary learning and variable selection, deep learning
  - ✓ Kernel learning (SVM, Kernel PCA, SimpleMKL, regularization path)
  - ✓ Graphs and learning (isomorphism, classification,…)
  - ✓ Model selection, bayes estimators and risks
- Model adaptability
  - ✓ Markovian models, multi-streams HMM, structure adaptation, Markov random fields and CRF
  - ✓ Learning with unknown or evolutive costs, multi-objective learning, hyper-parameters in classifier ensembles (random forests, DRF, one-class)
  - ✓ Multi-task learning

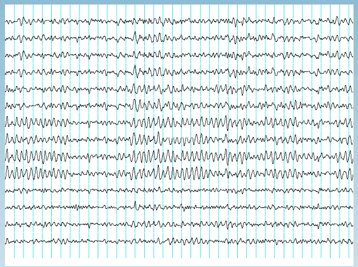
Application domains

- Access to information
  - ✓ Handwriting recognition
  - ✓ Spotting
  - ✓ Information extraction
  - ✓ Complex manuscripts
  - ✓ Digital libraries

  - ✓ Recommandation systems

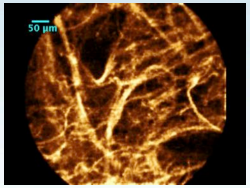


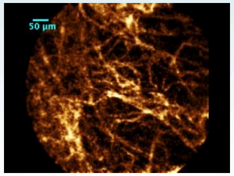


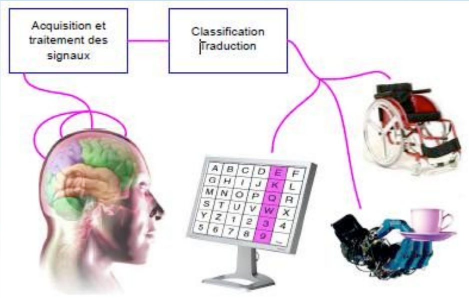


- Biomedical information processing
  - ✓ Brain Computer interface
  - ✓ Analysis of motor control data

  - ✓ Medical image classification
  - ✓ Medical image segmentation







STIC-AmSud French Team

- Prof. L. Heutte, PhD, PhD supervisor
  - ✓ Off-line and on-line andwriting analysis and recognition
  - ✓ Handwritten document analysis (bank checks, postal addresses, incoming mails, old manuscripts)
  - ✓ Information extraction and retrieval in handwritten documents
  - ✓ Classifier ensemble learning, classifier selection in ensembles
- Ass. Prof. Caroline Petitjean, PhD
  - ✓ Medical image analysis, segmentation and classification
  - ✓ Cardiac MRI image segmentation with shape prior (graph-cut)
  - ✓ Medical image modelling
- Ass. Prof. Simon Bernard, PhD
  - ✓ Classifier ensemble learning
  - ✓ Random forests

Pattern spotting in historical documents

- DocExplore project (http://www.docexplore.eu)

FIGURE : Query

- Natural image
- Scene or big enough Object

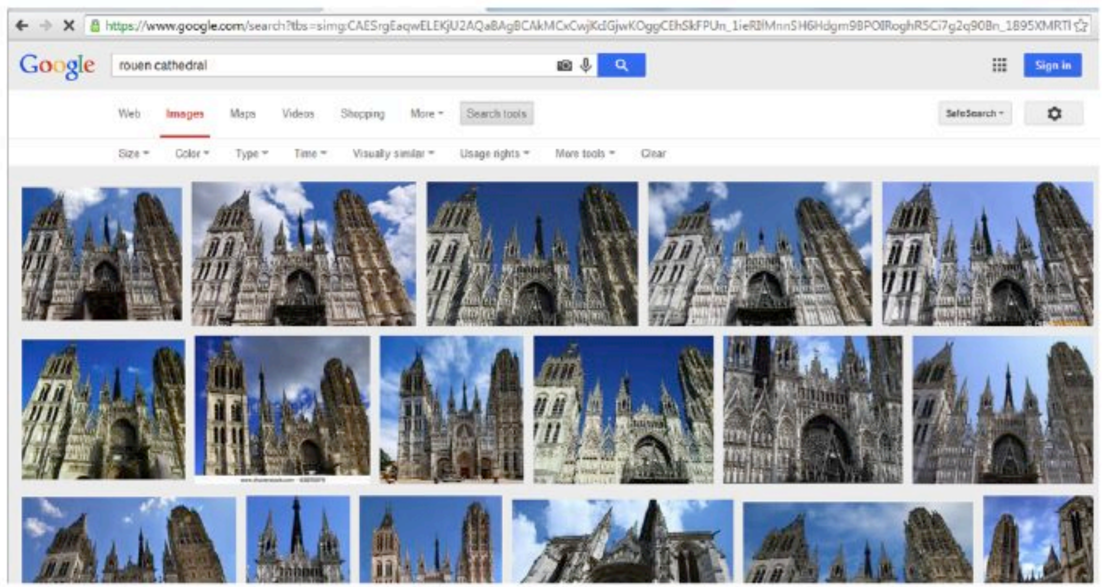


FIGURE : Results based on Google

Pattern spotting in historical documents

- Content-based sub-image retrieval

  ▸ A need expressed by historians and archivists

Pattern spotting in historical documents

- Content-based sub-image retrieval (cont'd)
  - Text and graphical objects
  - Image quality, changes in lighting, contrast,...

Pattern spotting in historical documents

- Content-based sub-image retrieval (cont'd)

Oxford dataset

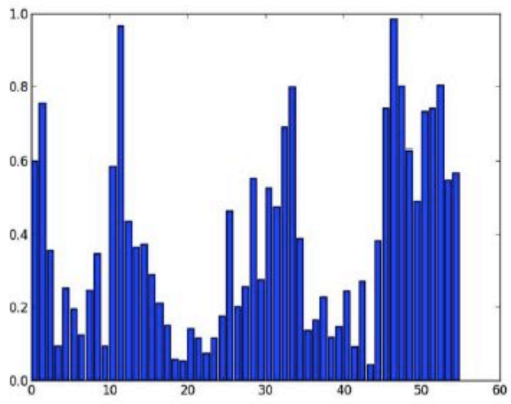


FIGURE : query average size ≈ 0.27

DocExplore dataset

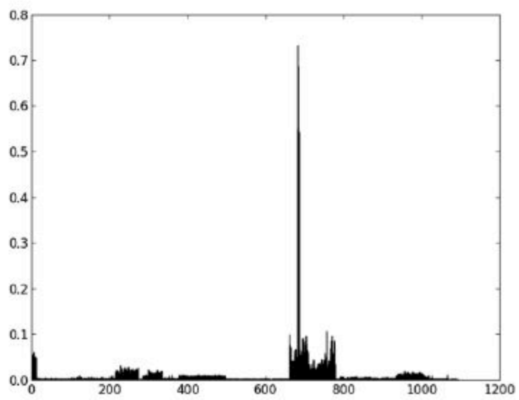


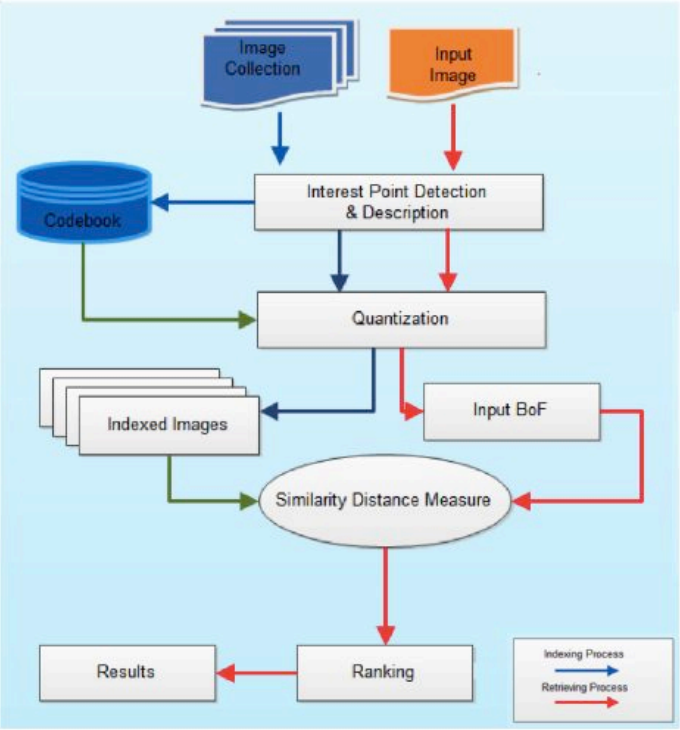
FIGURE : query average size ≈ 0.012

- Research questions :
  - Can we adapt the system developed for natural images to use it for document images ?
  - If it is not the case, what are those new challenges to be solved ?

Pattern spotting in historical documents

- Bag of visual words

  - Interest point detector
    - Hessian affine detector
  - Local descriptor
    - SIFT 128D
  - Codebook/quantization
    - HKM, 10k clusters
  - Similarity distance
    - Cosine distance

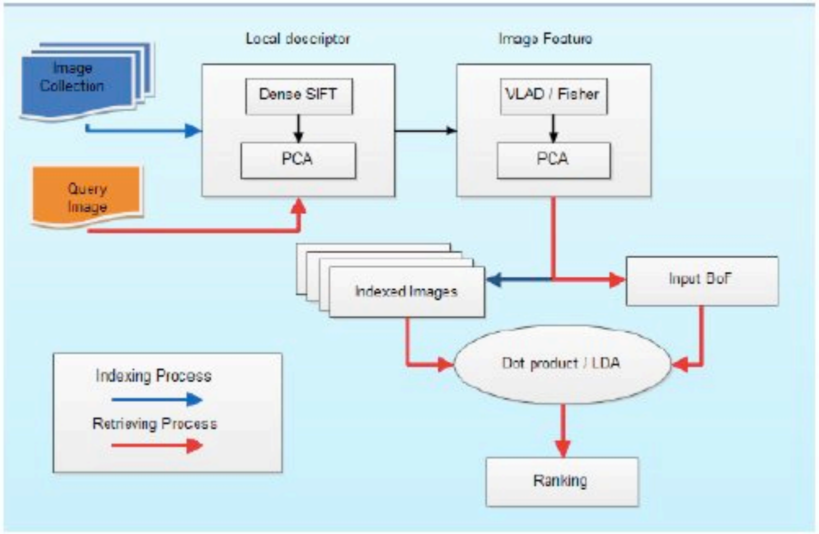


1. J. Sivic and A. Zisserman. *Video Google : A text retrieval approach to object matching in videos.* ICCV 2003

Pattern spotting in historical documents

- BoVW derivatives

- Feature representation
  - VLAD [2]
  - Fisher Vector [3]
- Similarity measure
  - Dot product
  - LDA ranking (learning on the fly a LDA model)



2. H. Jégou, et al. *Aggregating local descriptors into a compact image representation.* CVPR 2010

3. F. Perronnin, et al. *Improving the fisher kernel for large-scale image classification.* ECCV 2010

Pattern spotting in historical documents

- Experiments and results

- DocExplore dataset
  - 1591 medieval images
  - 1094 queries
  - 34 categories
    - flag
    - ornate initial letter
    - text separator
    - decorative object
    - ...

- Size of cluster = 64, PCA projection to 128, 1024

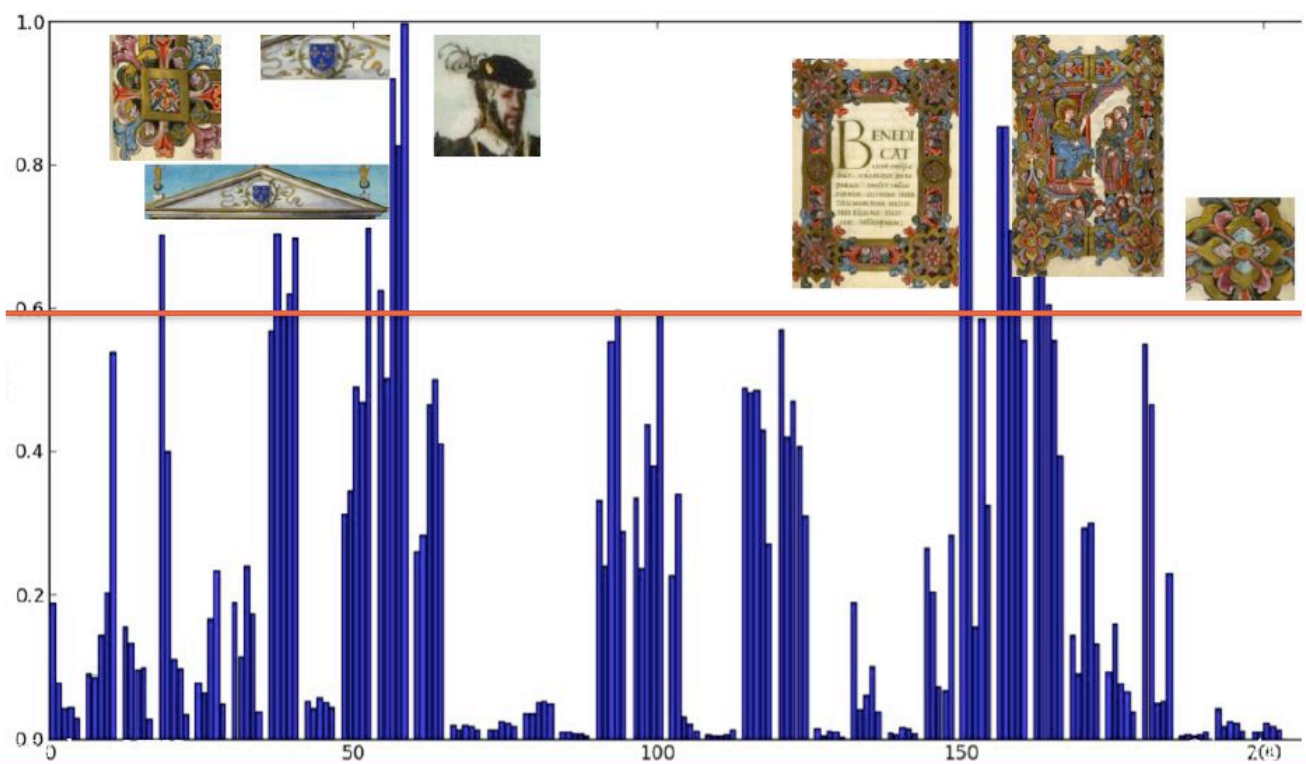
Dimension	Vlad LDA	Fisher LDA	Vlad Dot	Fisher Dot
32	0.078	0.095	0.123	0.141
64	0.116	0.111	0.137	0.150
128	0.151	0.111	0.145	0.155
1024	0.07	0.05	0.151	0.161

BoW Model	0.103

Pattern spotting in historical documents

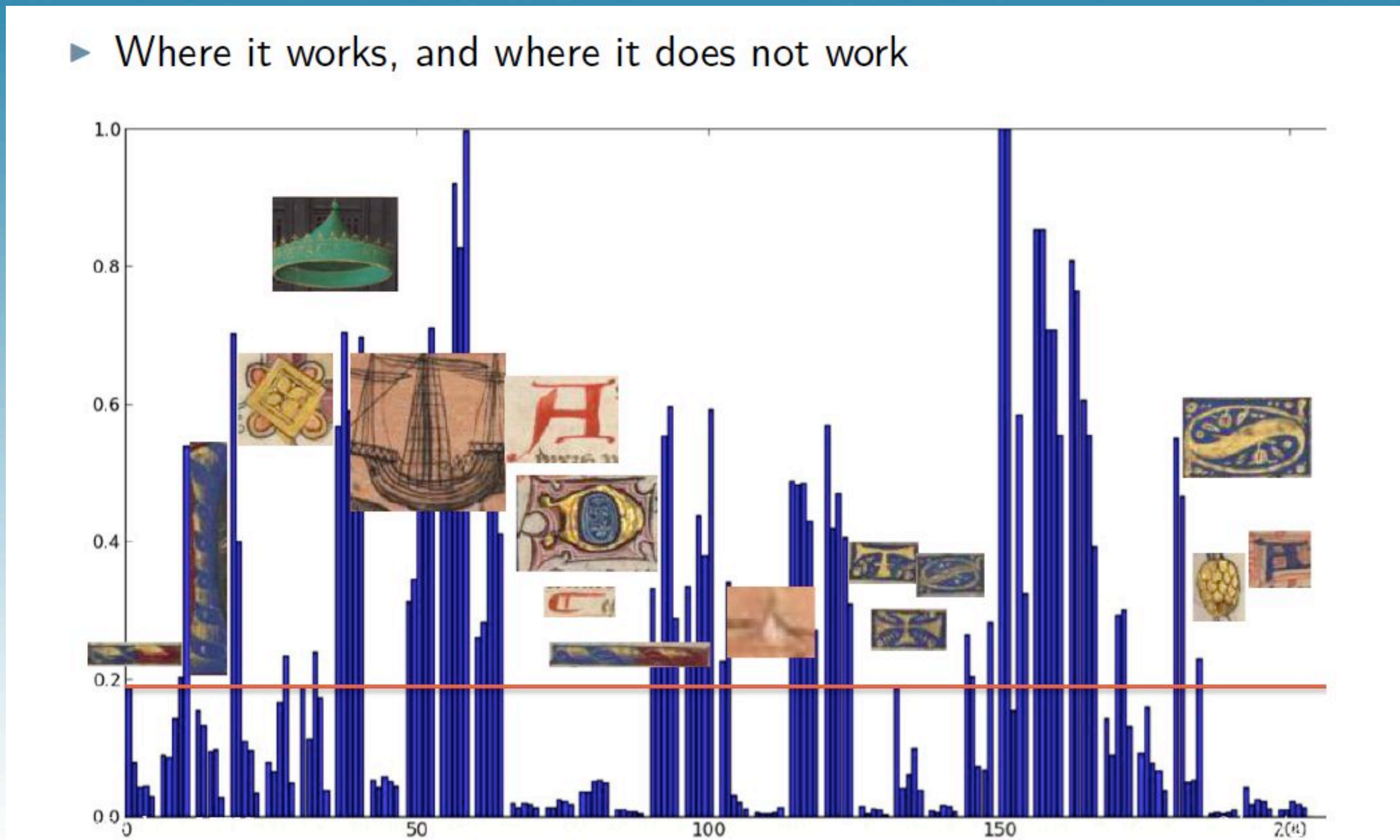
- Experiments and results (cont'd)

▶ Where it works, and where it does not work



Pattern spotting in historical documents

- Experiments and results (cont'd)



Pattern spotting in historical documents

- Experiments and results (cont'd)

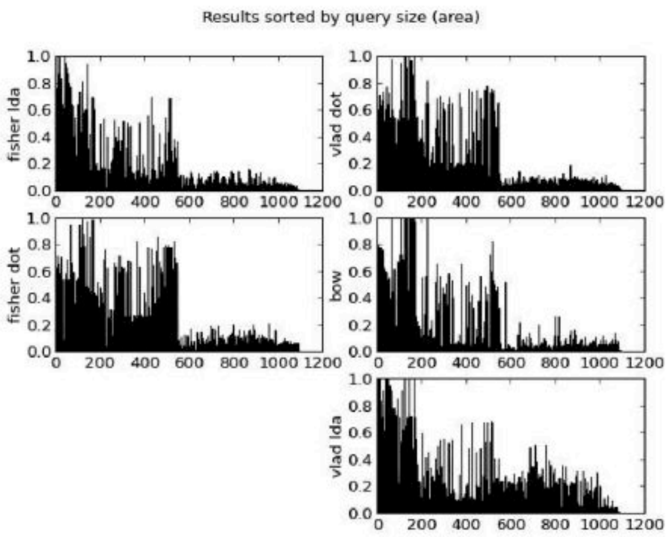


FIGURE : the query is sorted by the area (in pixels) from the biggest (left) to the smallest (right)

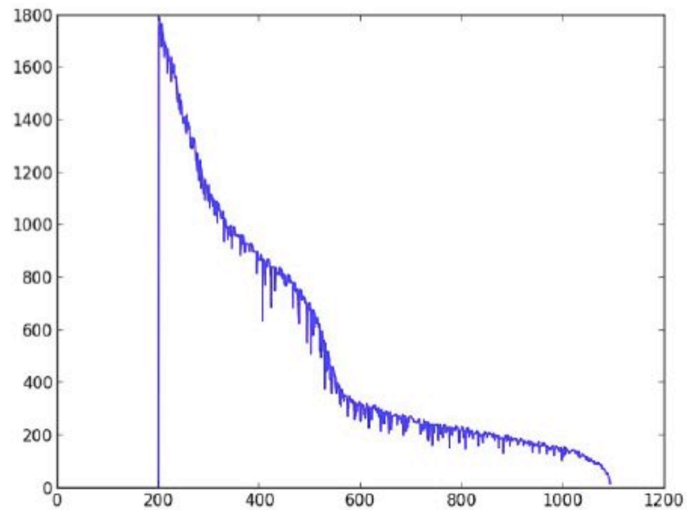
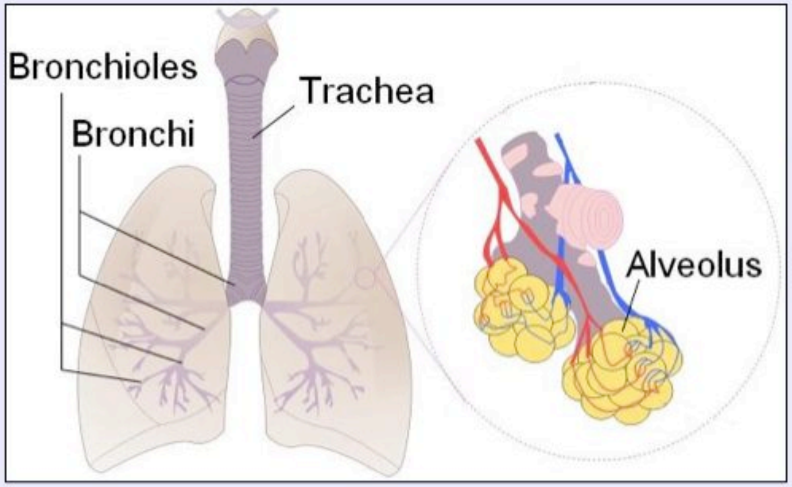


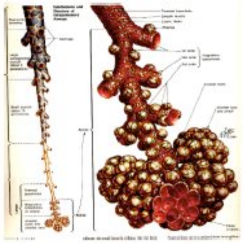
FIGURE : number of visual words extracted from each query (average number of visual word / image ≈ 10K)

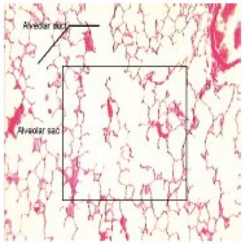
Medical Image Classification

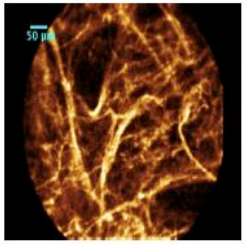
- Classification of endomicroscopic images of the lung

Background: Until recently, the alveoli were unreachable for in vivo investigation. A new endoscopic technique, called **Fibered Confocal Fluorescence Microscopy** (FCFM), has recently been developed that enables the visualization of the more distal regions of the lungs in-vivo and in real time [Thiberville09]. This promising technique could replace lung biopsy in the future.





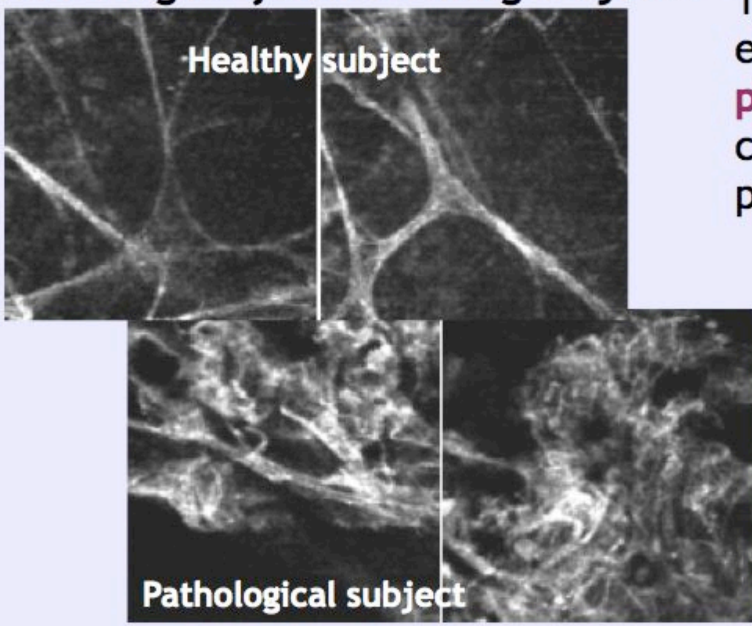




Medical Image Classification

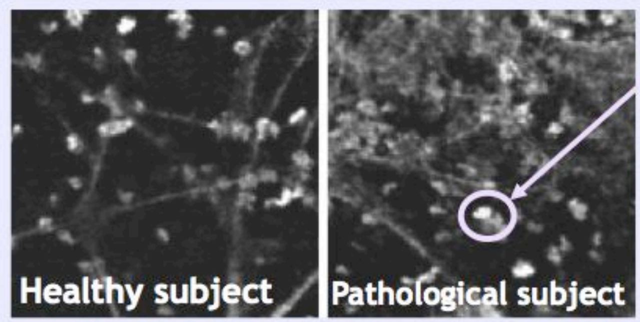
- Classification of endomicroscopic images of the lung

FCFM images of non-smoking subjects



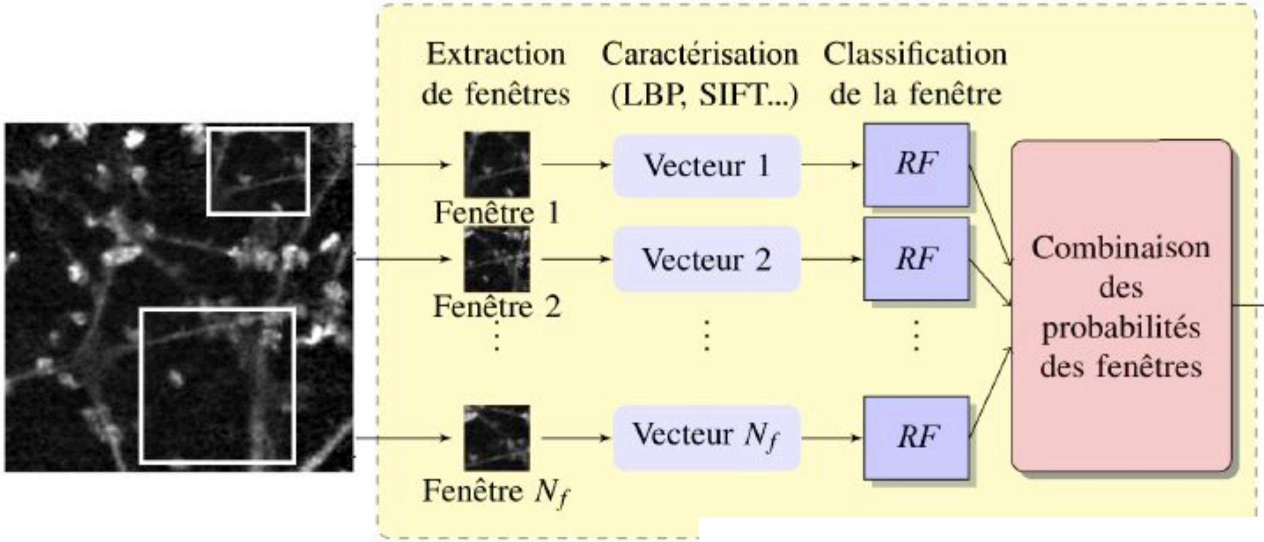
These images represent the **alveolar structure**, made of elastin fiber, with an approximate resolution of **1µm per pixel**. This structure appears as a network of (almost) continuous lines, that can be altered by distal lung pathologies (see figures).

FCFM images of smoker subjects

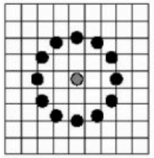


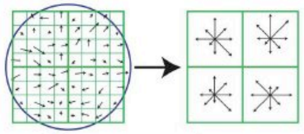
Macrophage (only in smoker)

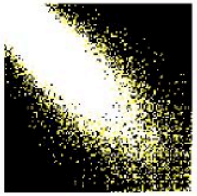
Medical Image Classification



- Local Binary Patterns (LBP) [Ojala, PAMI02]
- Scale Invariant Feature Transform (SIFT), Dense-SIFT [Lowe, IJCV04]
- Statistiques de co-occurrence [Haralick, SMC73]







	SIFT	Cooccurrence	LBP
Dimension	128	140	28

Medical Image Classification

Procédure 10-fold

- Extra-trees ($L = 30$, $K_{RFS} = \sqrt{M}$)
- SVM (noyau polynomial, $C = 10^5$)

Caractérisation locale

- 10K fenêtres en apprentissage
- 100 fenêtres par image de test

Performances pour la base non-fumeur

	Globale	Locale
ET	$91.53 \pm 8.46\%$	$\mathbf{93.84 \pm 7.06\%}$
SVM	$91.82 \pm 8.24\%$	$90.76 \pm 8.73\%$

Performances pour la base fumeur

	Globale	Locale
ET	$94.44 \pm 7.85\%$	$97.77 \pm 4.68\%$
SVM	$94.44 \pm 7.85\%$	$\mathbf{98.88 \pm 3.51\%}$

Medical Image Classification

Classification of endomicroscopic images of the lung [MLMI 2012]:

- A new technique $\rightarrow$ new images $\rightarrow$ uncertain oracle on pathological images

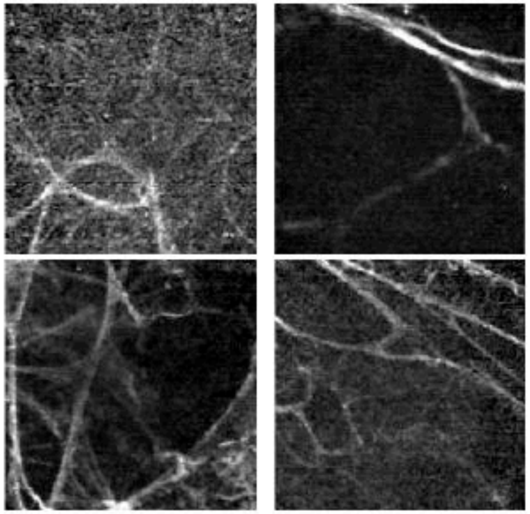


FIGURE 3 – Images alvéoscopiques présentant des difficultés de classification visuelle : sujet sain (haut), sujet pathologique (bas)

- Learning from healthy images only $\rightarrow$ One-Class Paradigm

Medical Image Classification

One-Class Random Forest

<u>Our solution</u>:

- Using classifier ensemble paradigm to break the curse of dimensionality, by subsampling the feature space and the training set.
- Generating more outlier data in sparse target regions, and less in densely populated target regions.

Our approach: One-Class Random Forest (**OCRF**), combining ensemble learning principles from traditional Random Forest algorithm with an original outlier generation method.

OCRF is composed of three main steps:

(i) extraction of **density information** from whole target data
(ii) generation of outlier data in **bootstrap** samples projected into **RSM** subspaces
(iii) induction of a **Forest-RI** on augmented dataset: bagging + RFS

Medical Image Classification

One-Class Random Forest: framework illustration

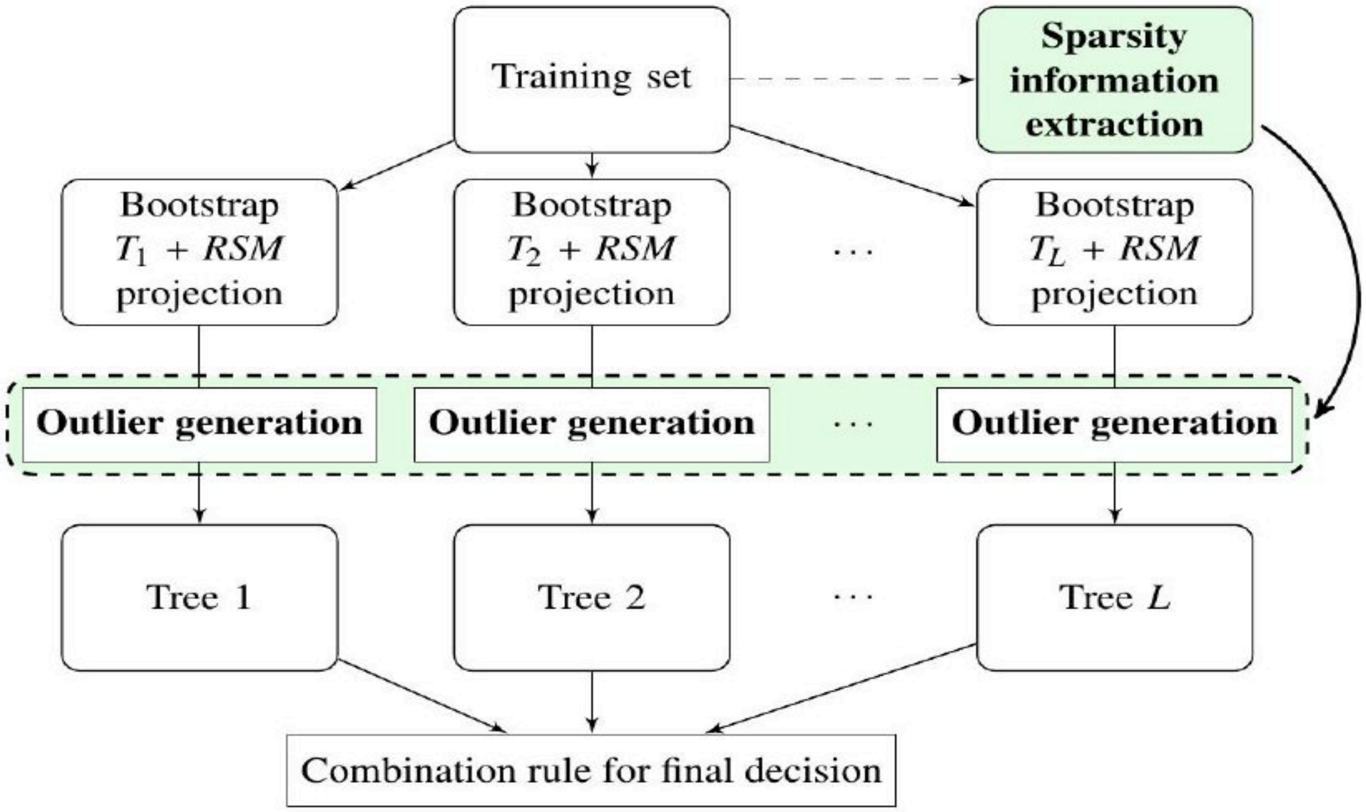


Figure: OCRF generation framework

Medical Image Classification

One-Class Random Forest: algorithm

Require: Training set T, $N_{outlier}$, $\Omega_{outlier}$, L individual trees, K_{RSM}
Ensure: A random forest classifier

(A) ***Density information extraction***
Compute H_{target}, normalized histogram of target data in T
Compute $H_{outlier}$, normalized histogram of generated outlier data, complementary of H_{target}, i.e. $H_{outlier} = 1 - H_{target}$

(B) ***Outlier generation and forest induction***
for $l = 1$ to L **do**
    (i) Draw a bootstrap sample T_l from training set T
    (ii) Project T_l onto a random subspace of dimension K_{RSM}
    (iii) Generate $N_{outlier}$ outlier data according to $H_{outlier}$ in the domain $\Omega_{outlier}$
    (iv) Train a standard decision tree on the augmented dataset
end for
return random forest model

Medical Image Classification

- Results on 78 UCI datasets

OCRF: Results (2)

Table: (a) Mean rank values; (b) Significancy results of statistical comparison with Friedman-Nemenyi test [Dem06]

	OCRF	OCSVM	Gauss	Parzen	Mog
Av. rank	2.4 ± 1.1	4.0 ± 1.3	1.9 ± 1.15	3.8 ± 1.1	2.8 ± 1.0

(a)

row $>$ col	OCRF	OCSVM	Gauss	Parzen	Mog
OCRF	-	+1	0	+1	0
OCSVM		-	-1	0	-1
Gauss			-	+1	+1
Parzen				-	-1

(b): OCRF $>$ {OCSVM,Parzen}, nothing can be said when compared to {Gauss} or {MoG}

Discussion…