Biological Image Classification with Random Subwindows and Extra-Trees

Raphaël Marée, Pierre Geurts, Louis Wehenkel

GIGA Bioinformatics platform
Interdisciplinary Cluster for Applied Genoproteomics
Systems and Modeling - Montefiore Institute
Department of Electrical Engineering and Computer Science
University of Liège

Belgium

BioImage Informatics, 8th September 2006

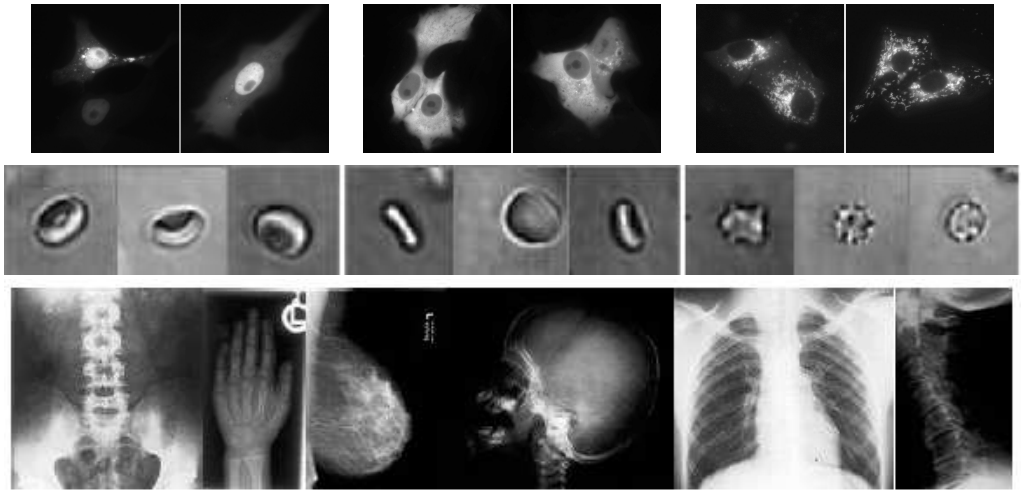
Image classification

- Goal
  - Given a training set of N labelled images (i.e. each image is associated with a class), build a model to predict the class of new images
- Challenges
  - To avoid manual adaptation to specific task
  - To be able to discriminate between a lot of classes
  - To be robust to uncontrolled conditions
    - Illumination/scale/viewpoint/orientation changes, noise, ...
  - To be fast and accurate

Our approach

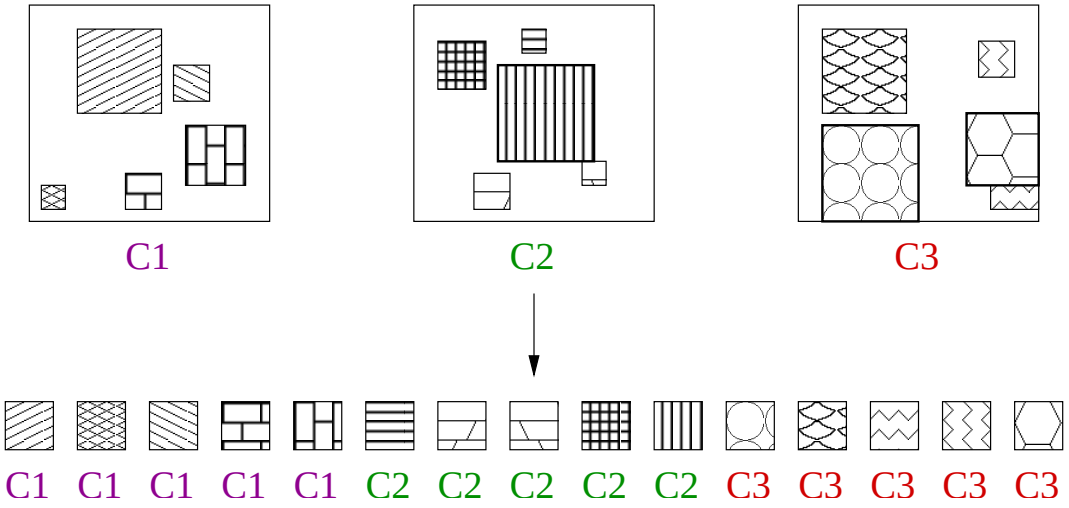
- Good results on a variety of tasks [Marée et al., CVPR 2005]

- Good baseline method for biological image classification ?



Learning: Extraction of Random Subwindows

- Extract N_{ls} ($>>N$) subwindows from N training images
  - Extract subwindows of random sizes, at random locations
  - Resize each subwindow to a fixed size (16×16)
  - Label each subwindow with the class of its parent image

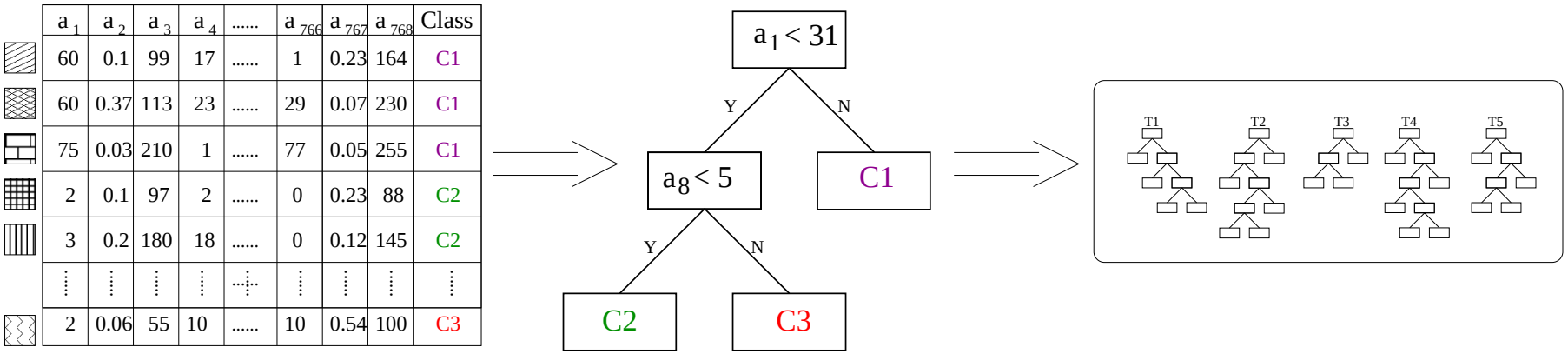


Learning: Subwindow Classification Model

- Describe each 16×16 subwindow by its raw pixel intensities (256 variables in grey, 768 if colors)
- Build a subwindow classification model from the training set of subwindows

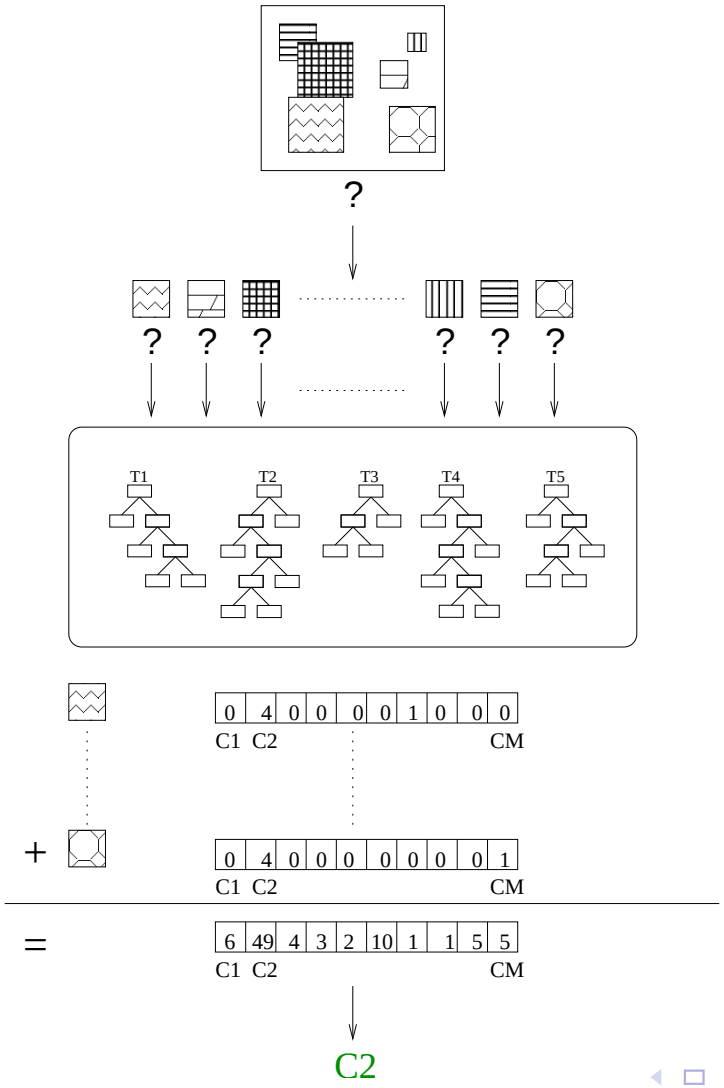
	a_1	a_2	a_3	a_4	.....	a_{766}	a_{767}	a_{768}	Class
	60	0.1	99	17	.....	1	0.23	164	C1
	60	0.37	113	23	.....	29	0.07	230	C1
	75	0.03	210	1	.....	77	0.05	255	C1
	2	0.1	97	2	.....	0	0.23	88	C2
	3	0.2	180	18	.....	0	0.12	145	C2
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	2	0.06	55	10	.....	10	0.54	100	C3

Learning: Extra-Trees [Geurts et al., Machine Learning 2006]



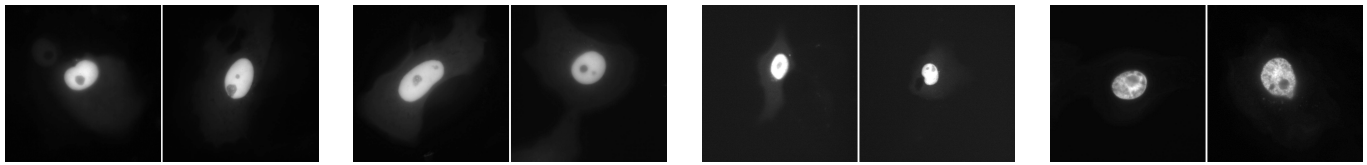
- Ensemble of T decision trees, generated independently
- Top-down growing by recursive partitioning
  - Internal *test nodes* compare a pixel-location-channel to a threshold ($a_i < v_i$), *terminal nodes* output class probability estimates
  - Choice of internal tests at random
  - Fully developed (perfect fit on training set)

Prediction: aggregation of subwindows and tree votes

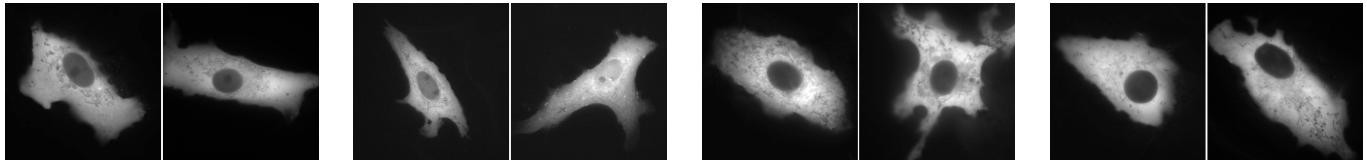


Subcellular Protein Localization (1/2)

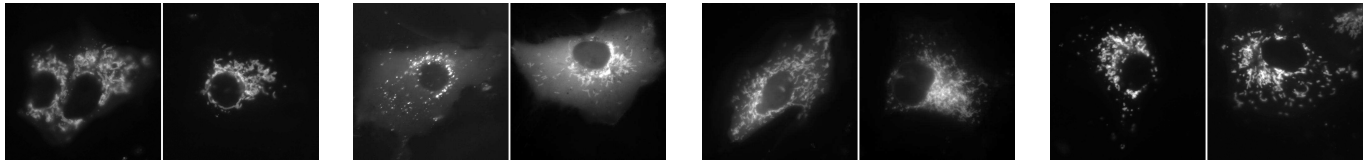
- Nucleus (32)



- Cytoplasm (32)



- Mitochondria (32)



Images from http://gfp-cdna.embl.de/

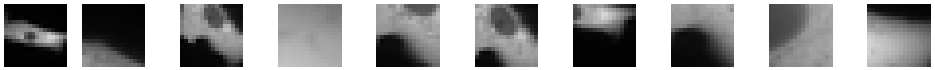
Subcellular Protein Localization (2/2)

- Protocol
  - 96 pair of images (N- and C-terminal GFP) labelled in 3 classes
  - Leave-one-pair-out prediction error rate
- Examples of random subwindows

Nucleus:

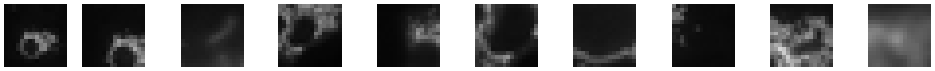
...

Cytoplasm:



...

Mitochondria:



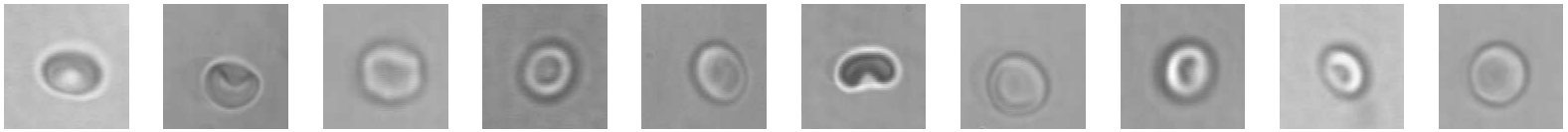
...

- Results

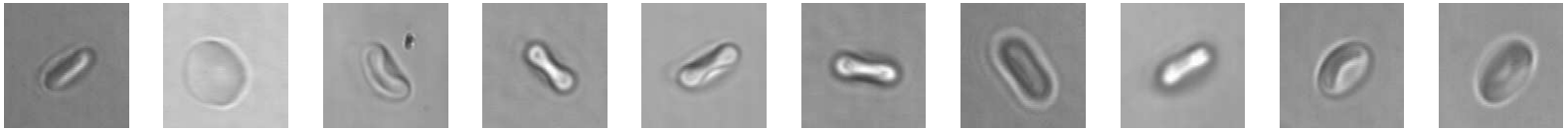
Dataset	ls/ts	nb. class	us	naïve	best
LifeDB	95/1	3	9.38%	66.67%	na

Red-Blood Cells (1/2)

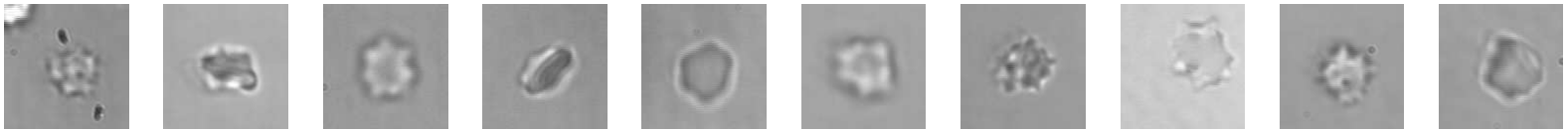
- Stomatocytes (3259)



- Echinocytes (916)



- Discocytes (887)



Images from RWTH, Aachen, Germany

Red-Blood Cells (2/2)

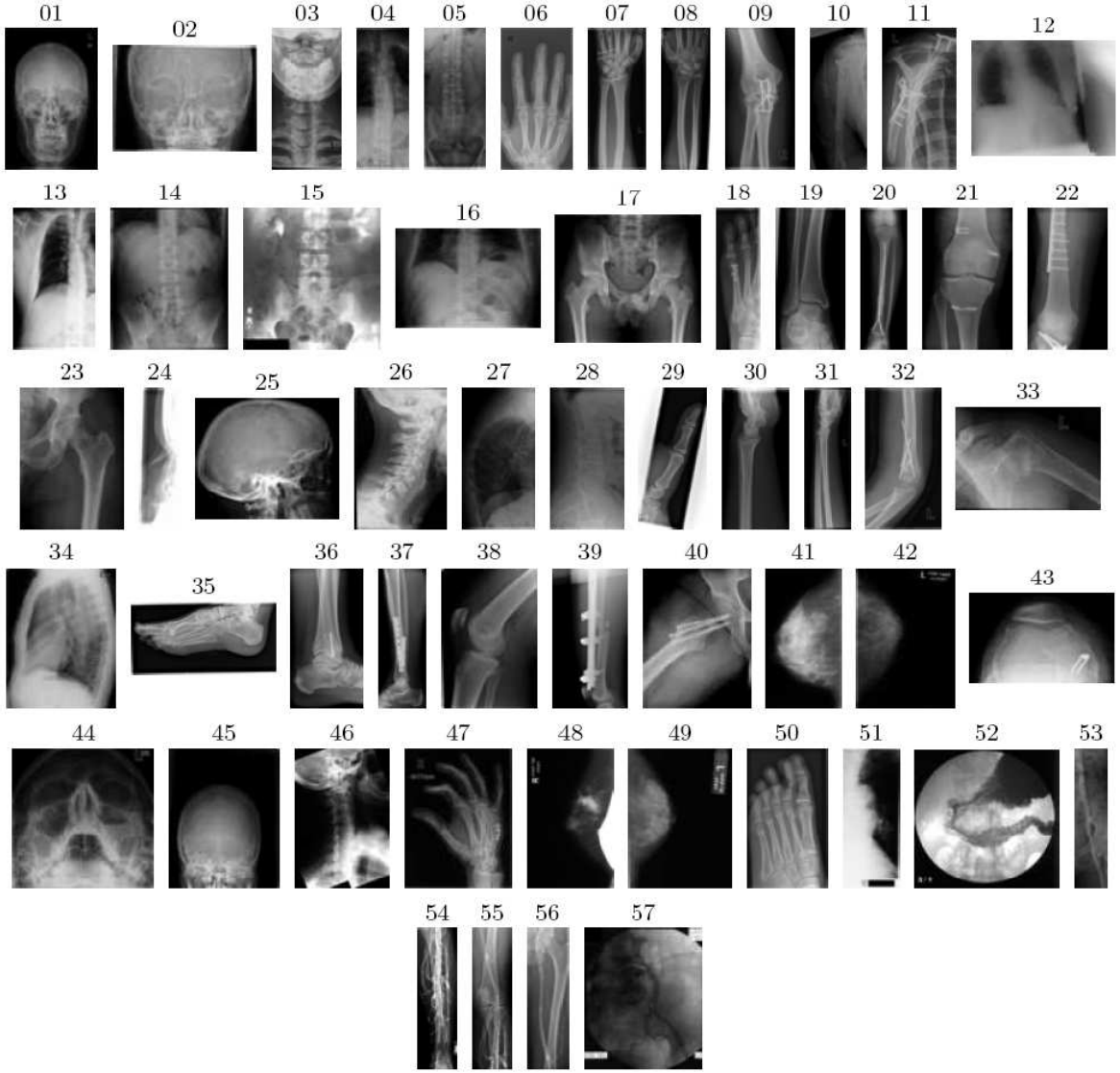
- Protocol
  - 5062 images labelled in 3 classes
  - Stratified ten-fold cross-validation error rate
- Random subwindows

  Improved result if random subwindow sizes are constrained between 80% and 100% of the image size instead of the full range of sizes
- Results

Dataset	ls/ts	nb. class	us	naïve	best
RBC	4556/506	3	20.92% ± 1.53	36%	13.5%

Nearest neighbor (euclidian distance) error rate: 24.4%

IRMA Radiographs (1/2) Images from http://irma-project.org/



IRMA Radiographs (2/2)

- Protocol
  - 10000 images labelled in 57 classes
  - Independant test set error rate
- Random subwindows

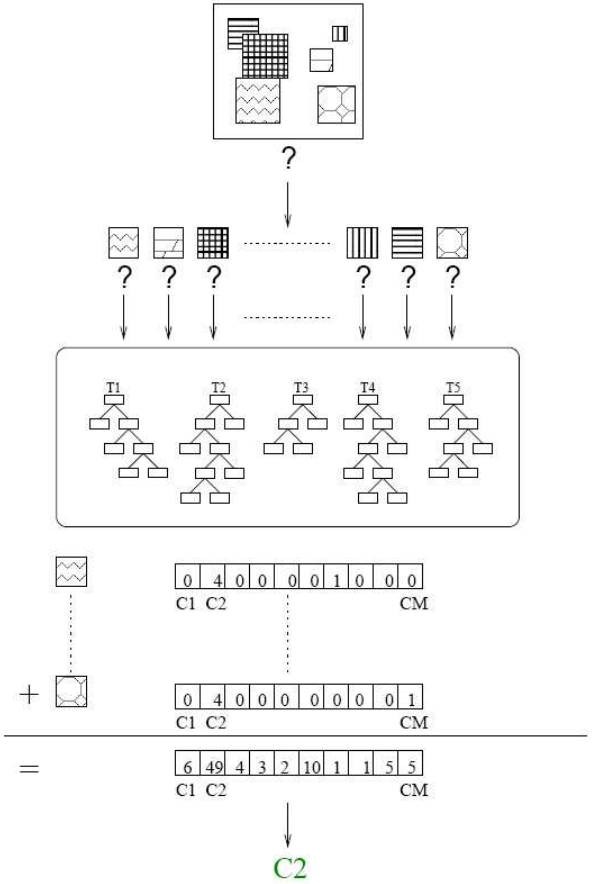
  Improved result if we add (x, y) absolute position coordinates to each subwindow feature vector
- Results

Dataset	ls/ts	nb. class	us	naïve	best
IRMA	9000/1000	57	13.1%	70.3%	9.3%

Nearest neighbor (euclidian distance) error rate: 36.8%

Interpretation

- Subwindow individual votes could help to focus on discriminative regions in the images

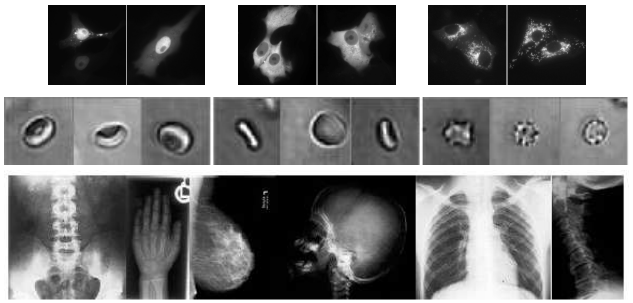


Computational complexity

- Learning: $O(nTN_{ls}logN_{ls})$
  - n: size of the feature vector
  - N_{ls}: nb. of training subwindows
  - T: nb. of trees
- Prediction: $O(N_{test}Td)$
  - N_{test}: nb. of test subwindows
  - T: nb. of trees
  - d: average tree depth $\approx logN_{ls}$
- Easy to parallelize

Conclusions

- Image classification method that...
  - combines Random Subwindows and Extra-Trees
  - yields quite good results on 3 various biological tasks



  - could be quickly evaluated on new classification problems
    - few parameters (the more trees/subwindows, the better)
    - fast learning, fast prediction
    - possible extension of the feature vectors, parameter tuning (subwindow sizes, orientation, ...)
  - is implemented in Java (executable and API available): `http://www.montefiore.ulg.ac.be/~maree/`

Extensions and Future work

- Automatic selection of informative subwindows and features
  - Active Learning, AdaBoost weights, ...
- Classification prior to segmentation, quantification, ...
  - Confidence maps

Acknowledgments

- Raphaël Marée is supported by GIGA-Interdisciplinary Cluster for Applied Genoproteomics, hosted by the University of Liège
- Pierre Geurts is a Postdoctoral Researcher at the National Fund for Scientific Research (FNRS, Belgium)
- IRMA and RBC databases courtesy of TM Lehmann and T. Deselaers, Dept. of Medical Informatics, RWTH Aachen, Germany

- PEPITe for the release of PiXiT, a Java implementation of the method, available for evaluation purpose at:
  `http://www.montefiore.ulg.ac.be/~maree/`
  `http://www.pepite.be/`

References

- "Random subwindows for robust image classification". R. Marée, P. Geurts, J. Piater and L. Wehenkel. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, June 2005, Volume 1, pages 34-40.
- "Extremely Randomized Trees". P. Geurts, D. Ernst, and L. Wehenkel. Machine Learning, Volume 36, Number 1, page 3-42, 2006.