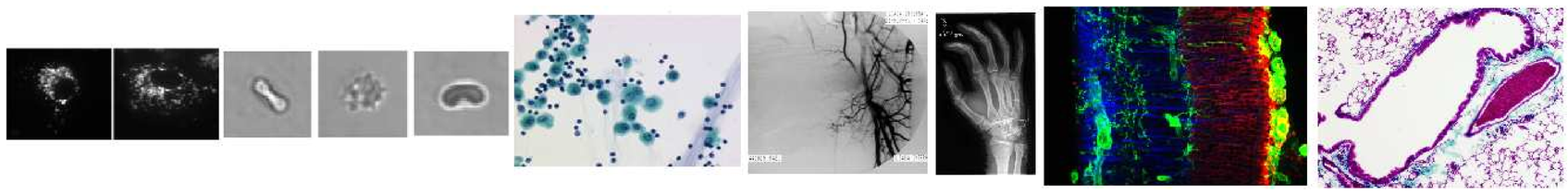
Random Subwindows and Randomized Trees for Image Retrieval, Classification, and Annotation

Raphaël Marée[1], Marie Dumont[2], Pierre Geurts[2], and Louis Wehenkel[2]

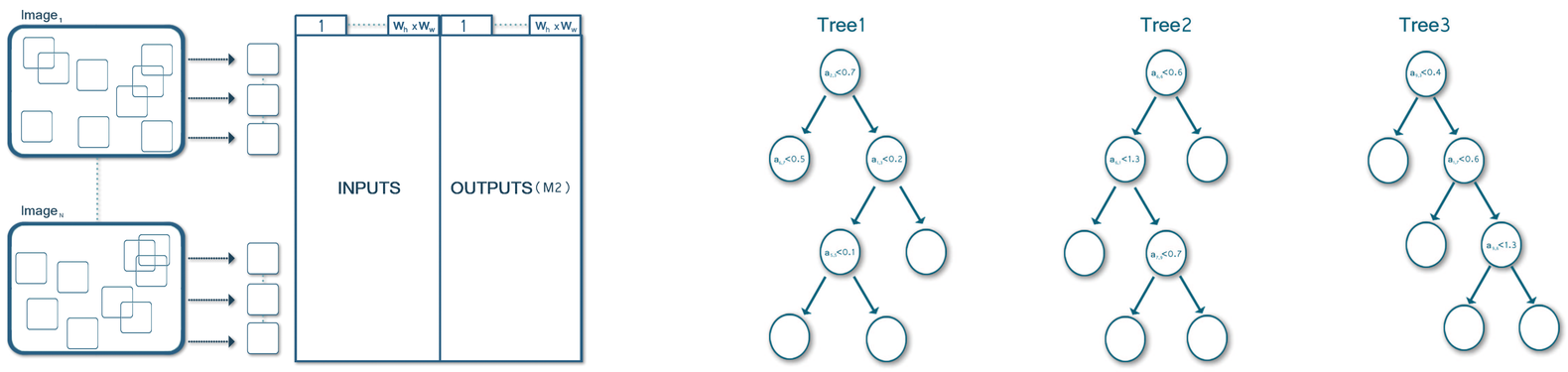
[1]GIGA Bioinformatics Platform, [2]Bioinformatics and Modeling, Department of EE & CS, University of Liège, Belgium

Abstract

Background: With the improvements in biosensors and high-throughput image acquisition technologies, life science laboratories are able to perform an increasing number of experiments that involve the generation of a large amount of images at different imaging modalities/scales. This stresses the need for computer vision methods that automate image retrieval, classification, and annotation tasks.



Method: We propose a unified framework involving the extraction of random subwindows (square patches) within images and the induction of ensemble of randomized trees [GEW06]. For image retrieval, we exploit the similarity measure and indexing structure of totally randomized tree ensembles induced from the set of random subwindows. For image classification, extremely randomized trees are used to build a subwindow classification model. For image annotation, we use extremely randomized trees with multiple outputs so as to predict the class of every subwindow pixels. To retrieve similar images, or to predict the class or the annotation of a new image, the method extracts random subwindows from this image, propagates these through the trees and aggregates output predictions.



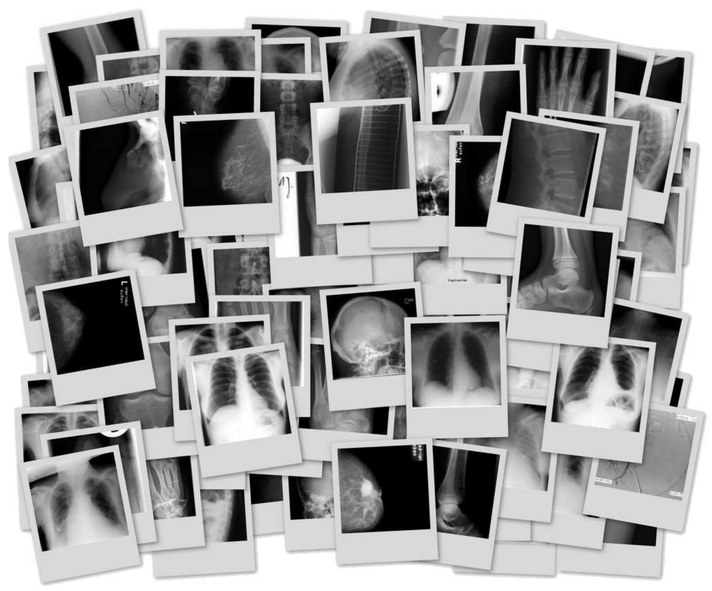
Results: We illustrate the potential of our generic method by providing results on datasets of biomedical images related to protein distributions or subcellular localizations, cell phenotypes, and radiographs. Accuracy results are compared to the state-of-the-art and to manual annotation. Beyond quantitative results, the method can also provide qualitative information such as the highlight of discriminative subwindows between classes of images, hence it can be used as an exploratory tool for further biological interpretation. We foresee the use of this automatic approach as a baseline method and first try in various biological studies that can be formulated as image retrieval, classification or segmentation problems. PiXiT (Java software) is available upon request for research studies [pix].

Content-Based Image Retrieval

▷ **Goal**

Given a reference database of images without any labeling nor any text description, the goal is to retrieve images similar to a new query image based on visual content.

▷ **IRMA 2005 RWTH**



Reference database of 9000 images.

▷ **Examples of queries and first retrieval results**

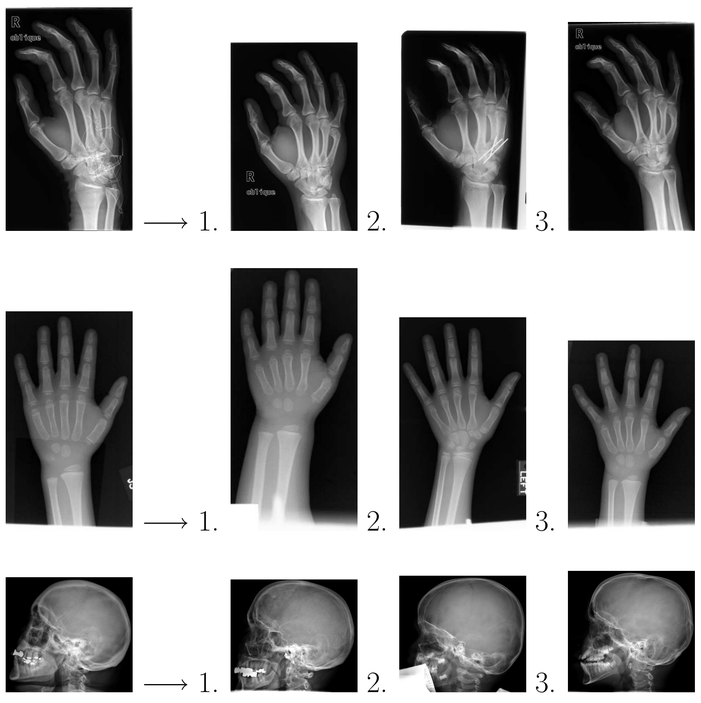
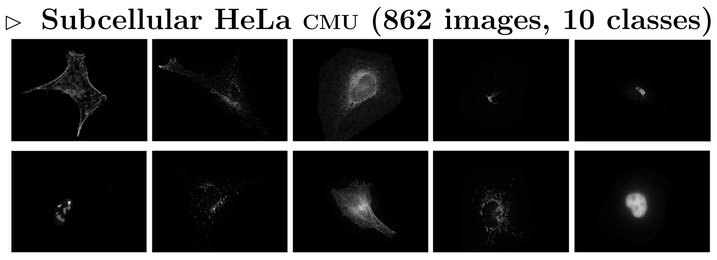


Image Classification

▷ **Goal**

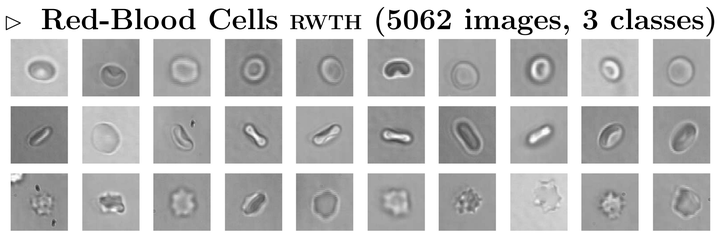
Given a training set of images where each image is labeled with one class among a finite number of predefined classes, the goal is to build a model that will be able to predict accurately the class of new, unseen images. [MGW07]

▷ **Subcellular HeLa CMU (862 images, 10 classes)**



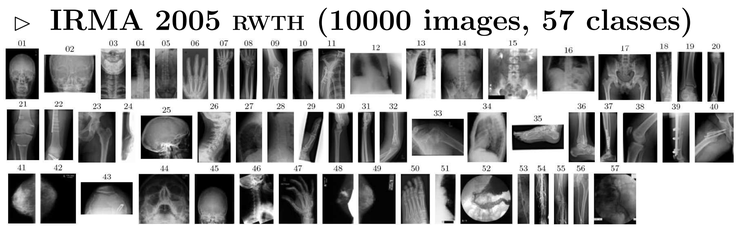
Goal: Identification of subcellular localization of proteins.
Classes: *ActinFilaments, Endosome, ER, Golgi Gia, Golgi Gpp, Nucleolus, Lysosome, Microtubules, Mitochondria, Nucleus.*
Results: 16.63%±2.75 misclassification error rate (at least as good as estimated human error rate).

▷ **Red-Blood Cells RWTH (5062 images, 3 classes)**



Goal: Inspection of individual cell shape changes following a drug treatment.
Classes: *Stomatocytes, Discocytes, Echinocytes.*
Results: 20.92%±1.53 misclassification error rate (at least as good as estimated human error rate).

▷ **IRMA 2005 RWTH (10000 images, 57 classes)**



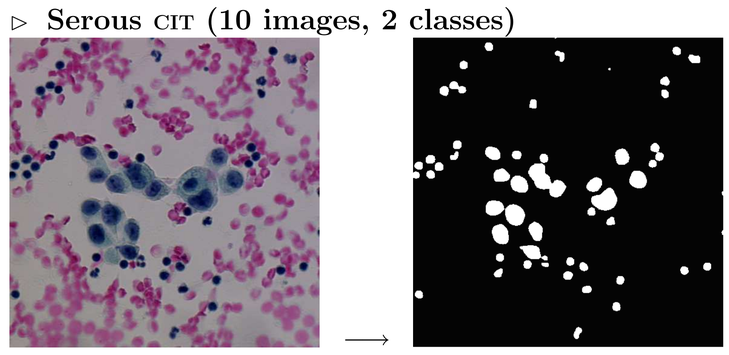
Goal: Categorization of clinical radiographs.
Classes: Codes describing imaging modalities, body orientations, body regions, biological systems.
Results: 11.8% misclassification error rate.

Image Annotation

▷ **Goal**

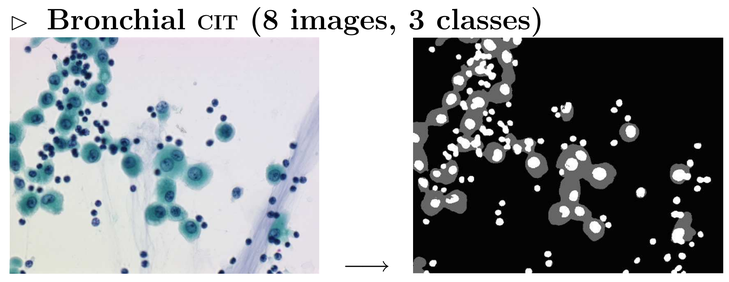
Given a training set of images with pixel-wise labelling (ie. every pixel is labeled with one class among a finite set of predefined classes), the goal is to build a model that will be able to predict accurately the class of every pixel of any new, unseen image.

▷ **Serous CIT (10 images, 2 classes)**



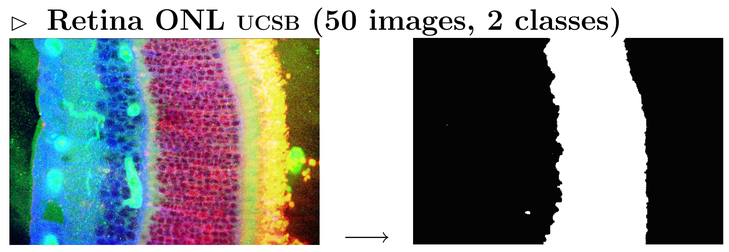
Results: 3.28% pixel misclassification error rate.

▷ **Bronchial CIT (8 images, 3 classes)**



Results: 3.11% pixel misclassification error rate.

▷ **Retina ONL UCSB (50 images, 2 classes)**



Results: 5.14% pixel misclassification error rate.

References

[GEW06] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 36(1):3–42, 2006.

[MGW07] Raphaël Marée, Pierre Geurts, and Louis Wehenkel. Random subwindows and extremely randomized trees for image classification in cell biology. *BMC Cell Biology*, 8(S1), 2007.

[pix] Pixit software. http://www.montefiore.ulg.ac.be/bioinformatics/ , http://www.pepite.be/.

Acknowledgments

RM is supported by the GIGA interdisciplinary cluster of Genoproteomics of the University of Liège with the help of the Walloon Region and the European Regional Development Fund. PG is a research associate of the FNRS, Belgium. Bronchial and Serous cytology images courtesy of O. Lezoray, Cherbourg Institute of Technology, University of Caen. RBC and IRMA databases courtesy of T. Deselaers and T. Lehmann, RWTH Aachen. Retina images courtesy of Retinal Cell Biology Laboratory & Center for BioImage Informatics, UCSB.