Finding key biological features for cancer diagnosis from histopathology slides

Cell segmentation within histopathological slides

Naylor Peter

PSL - Mines Paristech, Institut Curie and Inserm

peter.naylor@mines-paristech.fr

Introduction

Uncovering information from histopathology image data is a difficult task and is mostly unused in cancer research. This data corresponds to thin slices of the tumor and of the surrounding tissue. Histopathology slides can thus be very informative of the cancer subtype and/or of how the patient's immune system is reacting to the cancer. Our ultimate goal would be to quantify, via the extraction of biological features, a patient histopathology data. Biological feature are features that have a true biological meaning, like proportion of cancerous cell, normal cells, etc. See [5] for more insight. I will introduce the data and the methods for segmentation.

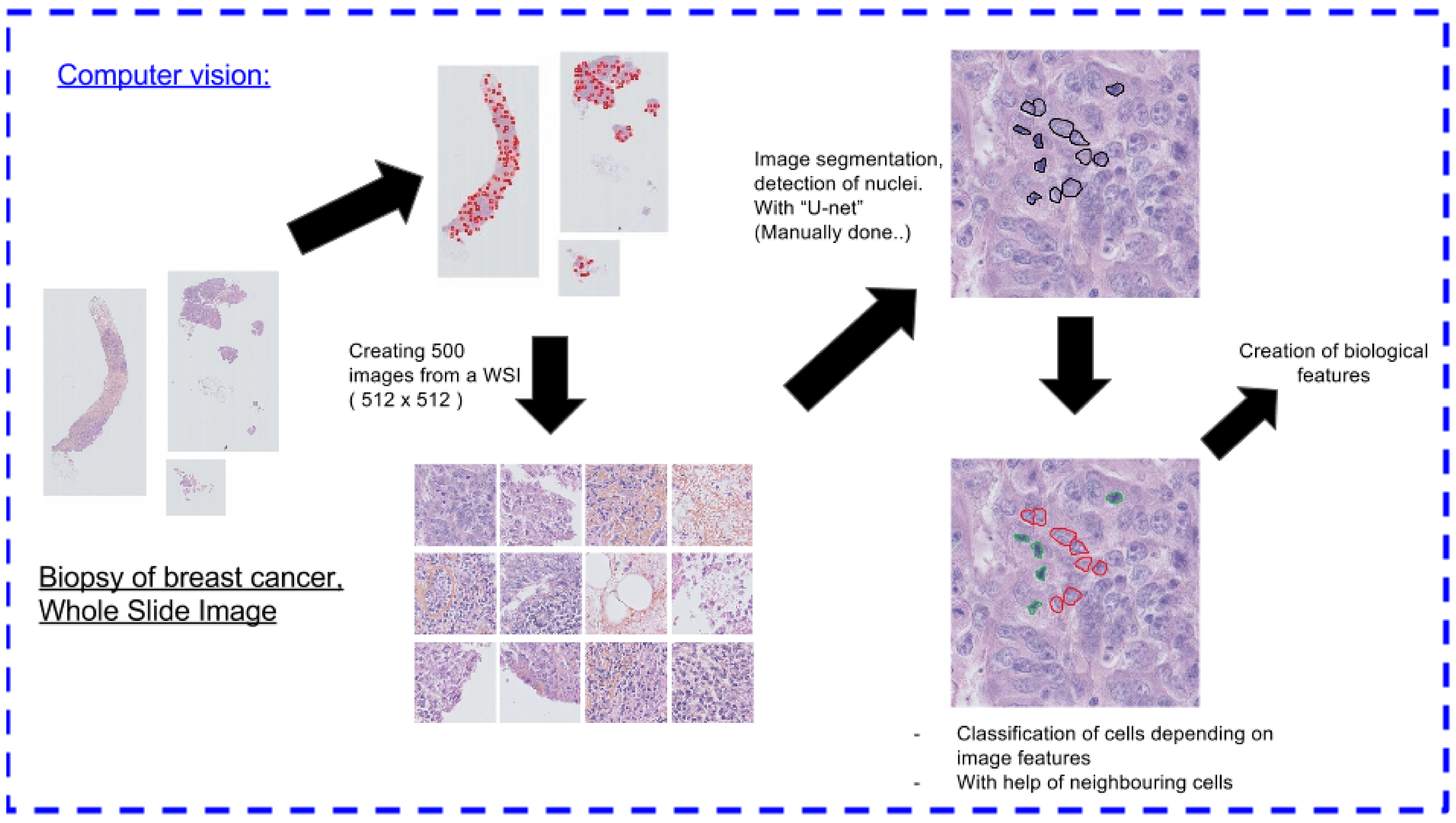
Methods

Our strategy to identify the important features is to first segment the important elements in histopathology slides (such as cells, tumor and stromal tissue, necrotic regions, etc.), second to define physiologically interpretable features for each of these elements and third to build a prediction model in order to assess the importance of each of these features. We propose a method based on fully convolutional network architectures for image segmentation that rely on standard convolutionnal networks.

Motivation

1. Unused data in cancer research.
2. Residual Cancer Burden Calculator, a tumor grading, is based on the content of histopathology slides, however pathologist only have a limited time per slide.
3. Reproducibility of the RCBC grading, it can vary between hospitals.
4. Defining biology driven features will interpretability for predicting clinical variables, such as outcome, subtype or response to treatment.
5. Allow us to investigate the link between genomic and transcriptomic features.

Pipeline



Uncovering information from histopathology data is a difficult task

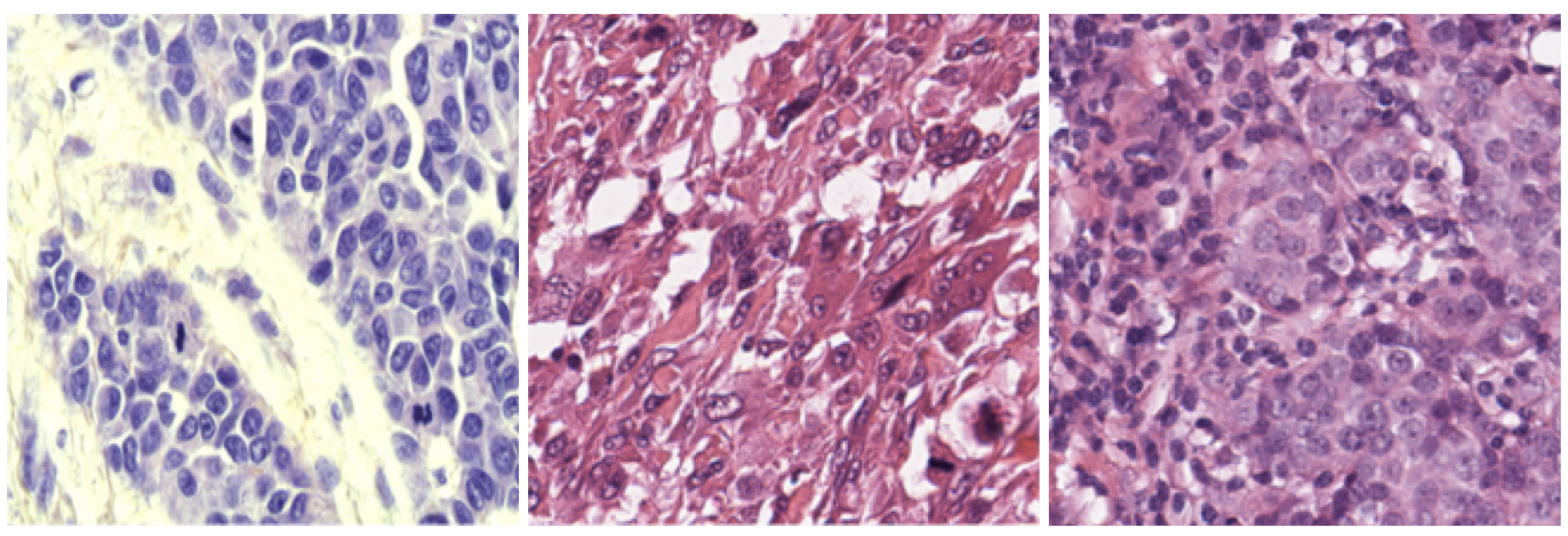


Figure 1: Histopathology data

1. *Data size*: each patient has several slides, one slide is more than $50GB$. A typical dataset: hundreds of slides.
2. *Stain variation*: Many reasons for this variability : scanner type, the stain supplier and the stain quality, differences in slide preparation and tissue type.
3. *Variability in the objects*: another variability is biological variability. Many different cells and tissue types.
4. *Projection artefacts*: a slice is actually a 3D slice, we can have overlapping cells / nuclei and other artefacts.
5. *No explicit formalization of information*: No clear descriptions of the objects we are trying to identify.

Manual annotation of histopathology slides

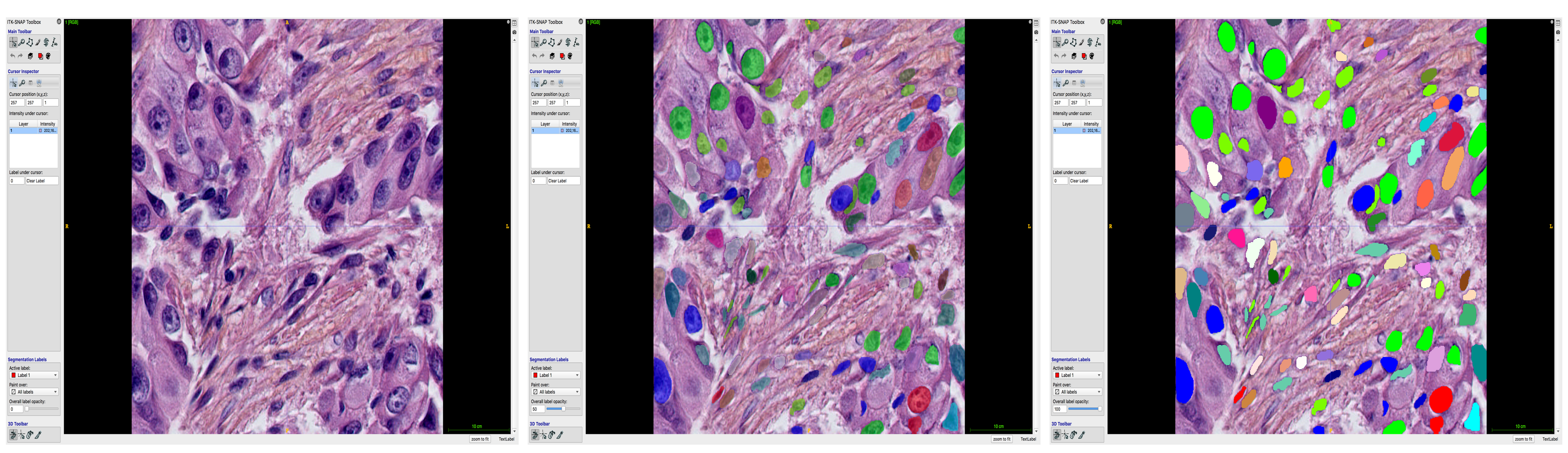


Figure 2: Histopathology data

We use ITK-snap to manually annotate our histopathology slide. We wish to detect cells vs background. We have 33 images of size 512×512 over 7 different patients.
Cells can be very tricky to annotate:

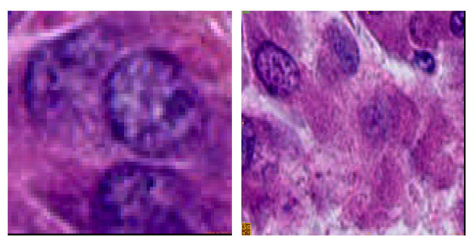


Figure 3: 3D Slice

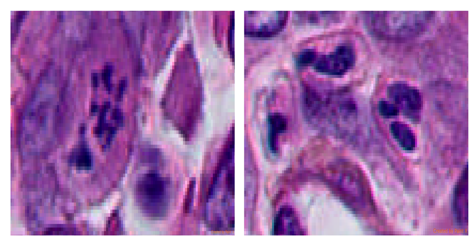


Figure 4: Wierd looking nuclei

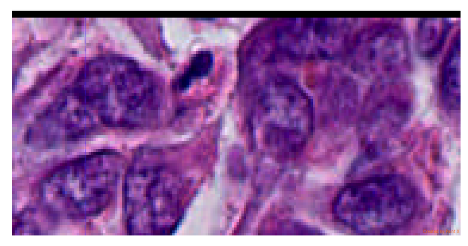


Figure 5: Dense region

Fully Convolutionnal Networks (FCN)

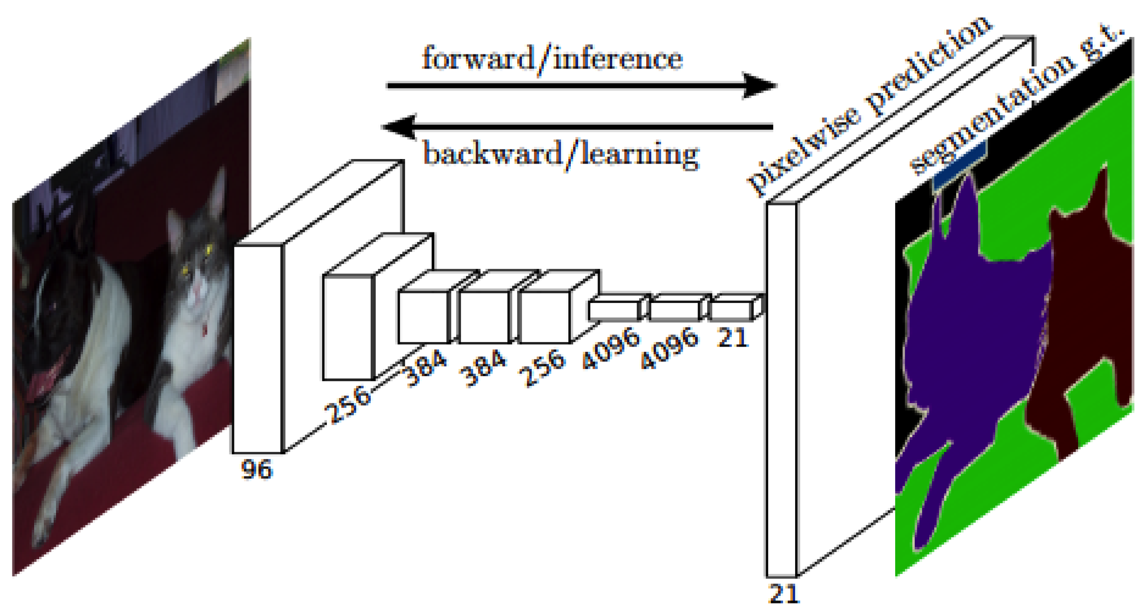


Figure 6: FCN for semantic segmentation (image taken from [1])

FCN for segmentation are an extension of standard Convolutionnal network for image recognition. These models take an end to end tuned CNN, and cast it into a FCN. Upsampling and deconvolutionnal layers are added, the standard part of the CNN provides the "what" while the added layers try to provide the "where". Finally, skip path are added between layers to provide final layers with information from the first layers.

Results

The fully convolutionnal network is fined tuned and helped with the use of data augmentation: rotation, flips, blurring and elastic deformation. The size of the input images can be of 512×512 pixels or of 256×256 pixels. Several metrics were kept, especially: mean accuracy, intersection over union, the Jaccard Index, recall and precision. Finally, training was performed on 21 images and test on 5 images across 6 patients. 7 validation images were used for reporting the validation scores, these validation images were provided by one patient.

Crop size	Network Name	MA	IU	Recall	Precision
512	FCN8	0.54	0.52	0.09	0.12
256	FCN8	0.63	0.53	0.28	0.09
256	FCN8_200	0.71	0.53	0.45	0.10
256	FCN8_2000	0.65	0.53	0.31	0.11

Table 1: First results

Forthcoming Research

- Setting up slighter different architectures: U-net [3].
- Changing the loss to make it more adapted to segmentation, [3].
- Changing the channels on the input, HE standardization [4].
- Incorporating different prior knowledge about cell segmentation directly in the architecture. [2].

References

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[2] Petter Ranefall, Sajith Kecheril Sadanandan, and Carolina Wählby. Fast adaptive local thresholding based on ellipse fit. In *International Symposium on Biomedical Imaging (ISBI'16), Prague, Czech Republic, April 13-16, 2016*, 2016.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pages 234–241. Springer, 2015.

[4] Arnout C Ruifrok and Dennis A Johnston. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology*, 23(4):291–299, 2001.

[5] Yinyin Yuan, Henrik Failmezger, Oscar M Rueda, H Raza Ali, Stefan Gräf, Suet-Feung Chin, Roland F Schwarz, Christina Curtis, Mark J Dunning, Helen Bardwell, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science translational medicine*, 4(157):157ra143–157ra143, 2012.