SLDC: an open-source workflow for object detection in multi-gigapixel images

http://www.cytomine.be/

Romain Mormont R.MORMONT@STUDENT.ULG.AC.BE
Jean-Michel Begon JM.BEGON@ULG.AC.BE
Renaud Hoyoux RENAUD.HOYOUX@ULG.AC.BE
Raphaël Marée RAPHAEL.MAREE@ULG.AC.BE
Montefiore Institute, University of Liege, Belgium

1. Introduction and Motivation

In several fields of application, multi-gigapixel images must be analysed to gather information and take decision. This analysis is often performed manually, which is a tedious task given the volume of data to process. For instance, in cytology, a branch of medical sciences which studies cells, cytopathologists analyse cell samples from microscope slides in order to diagnose diseases such as cancers. Typically, malignancy is assessed by the presence or absence of cells with given characteristics. In geology, climate variations can be analysed by studying the concentration of micro-organisms in core samples. The concentration is usually evaluated by smearing the samples onto microscope glass slides and counting those micro-organisms. In those situations, computer science and, especially, machine learning and image processing provide a great alternative to a pure-human approach as they can be used to extract relevant information automatically. Specifically, those kind of problems can be expressed as object detection and classification problems.

2. Methods

This work proposes SLDC (Segment-Locate-Dispatch-Classify), a generic framework for solving problems of object detection and classification in multi-gigapixel images. Particularly, given as input a two-dimensional (large) image, it is expected to output the information about the objects of interest contained in this image. Those information include the shape of each object, its location as well as a classification label and the probability estimates.

Our implementation provides algorithm developers with a structure to define problem-dependent components of the workflow (i.e. segmentation and classification) in a concise way, as illustrated by Figure 1. Every other concerns such as parallelization and large image handling are encapsulated by the framework. For instance, in order to avoid loading the full image into memory, it splits this image in tiles which are processed independently. Parallelism is also encapsulated by the framework which applies it to accelerate tiles processing. In addition, it provides a way to execute several processing workflows one after another on the same image as well as a powerful and customizable logging system. Moreover, we have integrated it with the open-source, web-based, Cytomine software for collaborative analysis and algorithm proofreading (Maree et al., 2016b).

3. Results

The performances of the framework have been assessed on a real-world problem: thyroid nodule malignancy (illustrated in Figure 2), in collaboration with the Department of Pathology, ULB-Erasme hospital (Prof. I.Salmon).

Especially, a specific instance of the workflow was built to detect malignant proliferative clusters of cells and individual cells with nuclear inclusions in whole-slide images of thyroid cell samples. In this case, the segmentation step involves color deconvolution (Ruifrok & Johnston, 2001) and mathematical morphology operations.

A binary classification (i.e. malignant or healthy) is then performed on the detected cells and patterns crop images. Especially, this operation relies on random subwindows and extremely randomized trees (Maree et al., 2016a). This generic image classification algorithm consists in extracting randomly-sized subwindows from the image. Then, the classifier predicts the

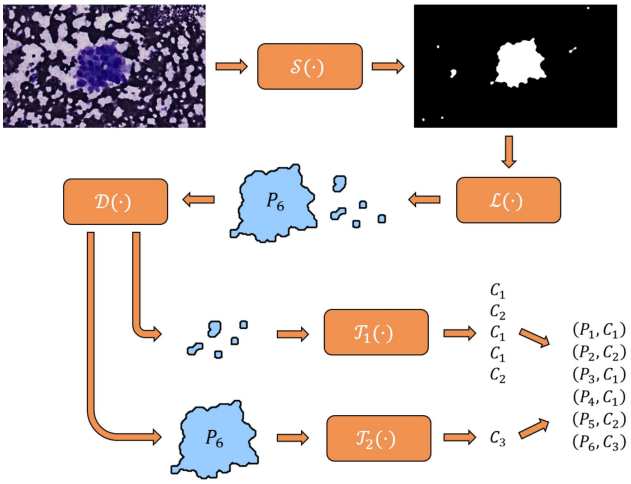

Figure 1. Illustration of the SLDC workflow. The Segmentation step ($\mathcal{S}$) produces a binary mask. The Location ($\mathcal{L}$) procedure extracts polygons representing the geometrical contours of the objects of interest from a binary mask. The Dispatch ($\mathcal{D}$) step is applied to each polygon to produces an integer which identifies the most appropriate classifier for processing this polygon. The Classify step ($\mathcal{T}$) produces a classification label (and probability estimates) for each polygon.

actual class by a majority vote of the predictions produced by the extremely randomized trees classifier on each window's raw pixel values.

Results are promising: the effective processing time for an image containing 8 gigapixels is less than 10 minutes (executed on 32 processes). As far as the thyroid application is concerned, our specific workflow has not reached our expectation yet as it sometimes fails at detecting objects of interest and produces an important number of false positives. However, the whole framework is production-ready and can be found on GitHub (`https://github.com/waliens/sldc`) as a Python library.

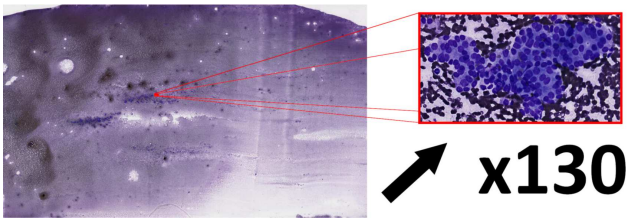

Figure 2. Illustration of the thyroid nodule malignancy detection problem. Left: original whole-slide image (163840× 95744 pixels). Right: a proliferative clusters of cells to be detected.

4. Conclusions

We developed SLDC, a generic workflow to foster the application of machine learning and image analysis techniques on large-scale imaging data. Future work will evaluate the workflow more extensively and improve its recognition performances for computer-aided cytology.

Acknowledgments

We acknowledge financial support from the Wallonia (DGO6) through the Cytomine and Histoweb research grants (1017072 and 1318185).

References

Maree, R., Geurts, P., & Wehenkel, L. (2016a). Towards generic image classification using tree-based learning: an extensive empirical study. *Pattern Recognition Letters*, *74*, 17–23.

Maree, R., Rollus, L., Stevens, B., Hoyoux, R., Louppe, G., Vandaele, R., Begon, J.-M., Kainz, P., Geurts, P., & Wehenkel, L. (2016b). Collaborative analysis of multi-gigapixel imaging data using cytomine. *Bioinformatics*, *32*, 1395–1401.

Ruifrok, A., & Johnston, D. (2001). Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.*, 292–299.