A HYBRID HUMAN-COMPUTER APPROACH FOR LARGE-SCALE IMAGE-BASED MEASUREMENTS USING WEB SERVICES AND MACHINE LEARNING

Raphaël Marée[1,2,3] *Loïc Rollus*[2,3] *Benjamin Stévens*[2,3] *Gilles Louppe*[2,3] *Olivier Caubo*[3]
Natacha Rocks[4] *Sandrine Bekaert*[4] *Didier Cataldo*[4] *Louis Wehenkel*[2,3]
[1] GIGA Bioinformatics Core Facility [2] Bioinformatics and Modeling (GIGA-Systems Biology and Chemical Biology, GIGA Research) [3] Systems and Modeling (Dept. of EE and CS, Montefiore Institute)
[4] Laboratory of Tumor Biology and Development (GIGA-Cancer and GIGA-I3, GIGA-Research)
University of Liège, Belgium

ABSTRACT

We present a novel methodology combining web-based software development practices, machine learning, and spatial databases for computer-aided quantification of regions of interest (ROIs) in large-scale imaging data. We describe our main methodological choices, and then illustrate the benefits of the approach (workload reduction, improved precision, scalability, and traceability) on hundreds of whole-slide images of biological tissue slices in cancer research.

***Index Terms*—** imaging informatics, machine learning, rich internet application, hybrid human-computer

1. INTRODUCTION

With recent advances in acquisition technologies, scientists generate growing amounts of biological imaging data (e.g., in anatomical pathology, neuroscience, drug discovery, or toxicology). Projects leading to terabytes of imaging data are becoming usual in various contexts, e.g. when experimental studies rely on whole-slide virtual microscopy, high-content screening, molecular imaging by mass spectrometry, or automated volume electron microscopy. As a result, better imaging informatics tools are needed [1] to ease the visualization and high-throughput analysis of such high-dimensional datasets in today's collaborative, geographically distributed, scientific context. As human interpretration of such datasets is impractical at such scale and operator-dependent, there is a strong need for computational methods to facilitate the extraction of quantitative information from these images. Despite increasing progress in machine learning, for some tasks algorithms have not yet reached reliable precision and interactive methods are still needed to proofread algorithm results.

Tissue image analysis is a very active field of research [2] including many works for tissue classification. Many of them are algorithm-oriented papers where limited imaging data is used to evaluate recognition performances of new standalone algorithms. By contrast, our work proposes a practical and scalable methodology with humans in the loop to ease the discovery of new biomedical insights. To achieve this goal, we extend Cytomine, a rich internet application for remote visualization and manual annotation [3], to enable computer-aided extraction of biologically relevant measurements from large-scale tissue imaging data. Our design choices are presented in Section 2. Results on a practical biological application that requires detection and surface measurements of tumoral regions in hundreds of large ($>$ Gigabyte) tissue slice images are given in Section 3. The potential impact of the approach is then discussed in Section 4.

2. METHODS

We propose a hybrid human-computer approach for the quantification of large sets of high-resolution bioimages by combining recent web development methodologies, spatial database concepts, machine learning techniques, and collaborative proofreading. In this work, the extraction of contours of regions of interest (ROIs) is formulated as a pixel classification problem followed by contour processing. First, manual annotations of ROIs (e.g. tumoral regions) and non-ROI (other subtypes of tissues) are used to train a pixel classification model. This model is then applied in a distributed fashion on new images and its predictions are processed and encoded in a centralized repository. Finally multiple users can proofread these predictions to derive reliable image-based measurements (e.g. surface measurements of tumoral regions). To implement this workflow, we rely on the Cytomine framework [3] which facilitates large-scale imaging data curation through a web interface. While its visualization and manual annotation modules have been described previously, in this section we first briefly recap its main design principles then describe our extensions.

2.1. General design principles

The rich internet application [3] uses recent web technologies and integrates various tools, standards and algorithms. High-resolution, two-dimensional images (with hundreds of thousands of pixels wide and tall) can be visualized at multiple resolutions in traditional web clients through caching mechanisms and distributed image tile servers supporting various

978-1-4673-1961-4/14/$31.00

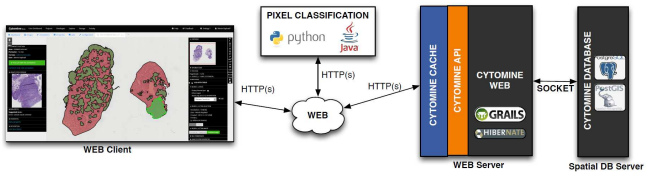

Fig. 1. An overview of the proposed architecture with the web client user interface and the tissue pixel classification application connecting through a RESTful web API.

image formats. Its underlying relational data model allows to create and manage projects which contain users, images, vocabularies with domain-specific terms, and layers of manual annotation geometries (e.g. polygons) drawn on top of original images to highlight regions of interest. All project data are stored in a spatial, relational, database and can be visualized and edited through a web interface or third-party softwares.

2.2. Spatial data model and RESTful API

The annotation is the central entity in the data model for ROI description. An annotation is a geometrical shape, located within an image, and for which a term from a user-defined, project-related, vocabulary can be associated. To enable model training (see next section) and proofreading of their results, here we propose three subtypes of annotations: user annotations (representing manual user annotations used as training sets for learning pixel classification models), job annotations (annotations generated by software instances), and reviewed annotations (for algorithm predictions proofread by users). In addition, we propose a job entity to store instances of softwares described by key-value pairs of parameter values. As we follow a REST architecture style (see Figure 1), each resource (e.g. an annotation) is referenced by a uniform resource locator (URL). The RESTful application programming interface (API) allows communication between servers and (web or third-party) clients through HTTP requests.

2.3. Generic machine learning and contour processing

For ROI detection, a supervised pixel classification model is first built using a generic algorithm based on subwindows and multiple output extremely randomized trees [4, 5]. It is trained from user manual annotations (retrieved using the web API) corresponding to ROIs and non-interesting regions. For the prediction phase, pixel classification can be performed by multiple software clients that work on small, independent, specific image areas (tiles), enabling massively distributed processing of large images. Connected components labeling then allows to extract contours from each tile pixel classifier mask. These contours need then to be translated into valid geometric shapes for spatial databases (as also pointed out by [6]). In our case we used hit-or-miss transforms to eliminate invalid geometric shapes. Point coordinates of valid geometries are then communicated through the web API to the central server that translates internally the HTTP request into spatial insert queries. Finally, once all tiles are processed, these contours are eventually merged by spatial union queries over tiles to take into account the fact that a single ROI may actually overlap several tiles.

2.4. Proofreading algorithm predictions

A polygon simplification algorithm [7] is then applied in order to slightly reduce the number of vertices of polygons hence easing manual edition of contours. To allow interactive and collaborative proofreading, we extended the image and annotation visualization web interface of [3] in order to display simplified job annotations as layers of geometrical shapes overlaid on top of the original images (see Figure 1, left). Then, for each image under review, a user is able to accept or reject individually (or all at once) predicted geometrical shapes, or edit them through drawing tools which allow to edit vertices, scale, substract or merge polygons, or fill internal holes. These user manual operations are automatically translated internally into spatial queries to update reviewed annotations.

3. EXPERIMENTAL DATA AND RESULTS

3.1. Biological application

The proposed methodology has been used to identify the impact of a pulmonary tissue composition change on lung tumor onset and progression [8]. To assess these questions, different mouse models were developed where mice were treated with components inducing a specific type of neutrophilic inflammation in lung tissues. The effects of pulmonary inflammation has to be investigated in lung hematoxylin-eosin-stained digital slides (8 tissue slices per animal) and the tumor area has to be determined and reported to the total area of lungs for different experimental conditions. A typical whole-slide scanned image has 35000×30000 pixels (with $40X$ objective and pixel resolution of $0.23\mu m$).

3.2. Evaluation of performances

From a machine learning perspective, the task could be seen as a binary pixel classification problem with tumors as positive class, and all other subtypes of tissues (including inflammatory cells, blood vessels, cartilage, bronchus, ...) as negative class (see Figure 2, bottom right). After a development phase where scientists manually annotated regions corresponding to different subtypes of tissues, several pixel classification models were trained and qualitatively evaluated on a few slides (final chosen parameter values are $T = 10$ trees, $n_{min} = 2$, and $\pm$ 100000 subwindows of fixed-size 24×24

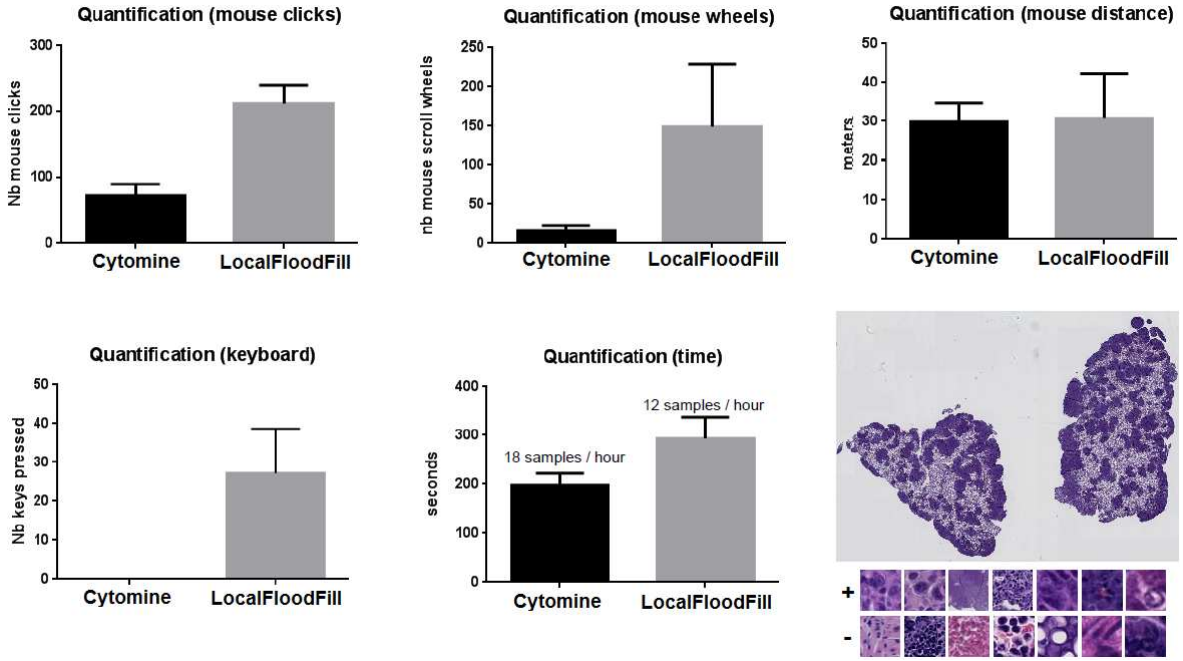

Fig. 2. Average timings and operation counts using the local "FloodFill" approach and our web-based "Cytomine" approach. Bottom right: one of the five whole-slide hematoxylin-eosin-stained tissue image used for performance evaluation (original image size: 36864×25344 pixels), and examples of patches extracted in positive (tumor) and negative regions.

pixels encoded in HSV). The approach was then used to quantify automatically about 250 whole-slide images corresponding to 5 experimental studies, which yielded the validation of a total of more than 17000 tumoral ROIs. To assess the impact of the proposed workflow on scientist daily workload, we compared the time required to proofread algorithm predictions (through a regular Wi-Fi internet connection) with respect to another semi-automatic approach (on the local computer). The latter approach, which was the standard practice in the laboratory before the introduction of our methodology, combines the use of ImageJ and Adobe Photoshop on downsampled images. It requires the user to click on tumor islets using a "magic wand" tool (with a flood fill algorithm) followed by manual corrections (e.g. to reject inflammatory cells which have similar color intensities). Timings and manual operation counts for 5 randomly choosen whole-slide images (totaling 900 tumor islets) have been recorded and averaged. Figure 2 summarizes these results which clearly show that the number of mouse clicks and total operation time is lower when using our approach. However, we observed cursor travelled distances are roughly equal due to the positioning of proofreading buttons and the lack of keyboard shortcuts in our web interface at the time of the experiment. Additional qualitative and quantitative results are given in Figure 3.

3.3. Other case study

Another preliminary experiment, related to cancer studies with tens of hematoxylin-DAB-stained immunohistochemical slides, shows that our approach with improved user interface and keyboard shortcuts reduces the required time to contour ROIs from 15 minutes (with a fully manual approach) downto 3.5 minutes on average. In that specific study, for the baseline approach it was needed to manually and systematically delineate all contour lines because usual flood fill algorithms are inoperative on regions with similar color intensities but different textured patterns.

4. DISCUSSION

4.1. Impact on workload, precision, and current practices

In practice, we expect savings of time and manual operations provided by our approach would depend on the sizes and densities of the ROIs (e.g. if only a few small ROIs have to be detected in each image, then our approach is less beneficial). Performances also depend on appearance heterogeneity of ROIs and other image regions that can affect recognition performances of the pixel classification models. Also, the reported results do not take into account the model development phase which can involve several time-consuming steps to improve recognition of difficult regions. Overall, although it was shown previously that the used algorithm works well on diverse types of imagery [4], it still requires time to tune its few parameters and to annotate a realistic training set, hence the approach will be mostly useful when one has to analyze

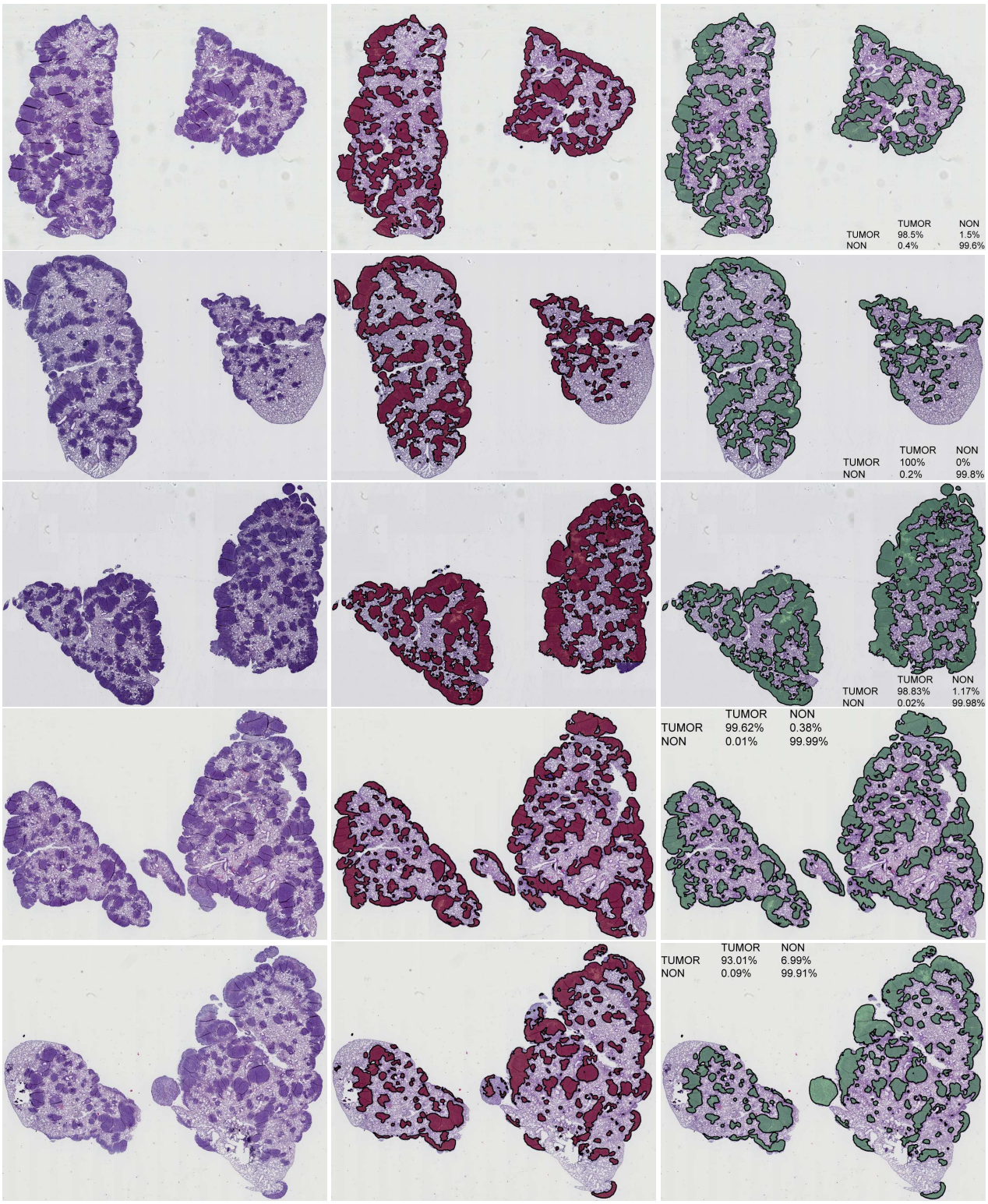

Fig. 3. Qualitative and quantitative results for the five digital slides used for performance evaluation. For each line: low magnification of the original H&E tissue image (left), tumor tissue classification automatically generated by algorithm (middle, in red), final tumor tissue classification after proofreading by expert (right, in green). Confusion matrices show algorithm recognition rates within tissues computed *a posteriori* (*i.e.* the ground-truth is the final tissue classification after proofreading).

large numbers of samples. This advantage is especially true given that the image processing can be massively distributed, and that algorithm results can be proofread remotely by multiple scientists. Regarding final quantification accuracy, we expect our hybrid human-computer approach (that combines pixel-level automated classification with careful proofreading) to produce more precise contours with respect to an approach working on downsampled images where boundaries are less accurate and small tumor islets can be missed by the human eye due to fatigue. In addition, we have implemented a blinded mode that can be activated to hide image names in order to avoid bias when proofreading algorithm results. Furthermore, the approach stores algorithm parameters and final reviewed annotations to achieve a better traceability. This also allows subsequent pixel classification model re-training with the hope to improve recognition accuracy over time. Overall we hope such an approach will contribute to more precise and reproducible results in biomedical research.

4.2. Extensibility

Other algorithms could be integrated into our approach. One could indeed leverage the existing workflow (web communication mechanisms to register algorithm parameters, retrieval of manual annotations for parameter tuning or training, distributed processing of image tiles, contour processing and simplification services), and proofread the results of such new algorithms on the web. For example, segmentation algorithms from Ilastik [9] or Fiji [10] could be integrated. In preliminary work, we have integrated Fiji routines for tile-based thresholding and color deconvolution that can be visualized on-the-fly in the web user interface.

5. CONCLUSIONS AND FUTURE WORK

In this paper we presented an approach that combines ideas from machine learning, spatial databases, and web software development to ease the quantification of regions of interest in large bioimages. We are currently extending the methodology with other user interfaces and algorithms to produce various types of quantifications for different imaging modalities (e.g. combining detection of regions of interest and positive cell counting within immunohistochemical images, cell sorting in cytology, or phenotype recognition in microscopy images).

6. ACKNOWLEDGMENTS

This work is funded by the research grants n°1017072 and n°1217606 of the Wallonia. RM is supported by GIGA research center with the help of the Wallonia and the European Regional Development Fund (ERDF). We thank P. Martinive and N. Leroi (for communication of preliminary results about the extraction of ROIs in cancer immunohistochemical slides), and F. Perin (for slide scanning).

7. REFERENCES

[1] G. Myers, "Why bioimage informatics matters," *Nature Methods*, vol. 9, pp. 659–660, July 2012.

[2] S. Kothari, J.H. Phan, T.H. Stokes, and M.D. Wang, "Pathology imaging informatics for quantitative analysis of whole-slide images," *Journal of the American Medical Informatics Association*, vol. 20, no. 6, pp. 1099–1108, November-December 2013.

[3] R. Marée, B. Stevens, L. Rollus, N. Rocks, X. Moles-Lopez, I. Salmon, D. Cataldo, and L. Wehenkel, "A rich internet application for remote visualization and collaborative annotation of digital slide images in histology and cytology," *BMC Diagnostic Pathology*, vol. 8 S1, 2013.

[4] M. Dumont, R. Marée, L. Wehenkel, and P. Geurts, "Fast multi-class image annotation with random subwindows and multiple output randomized trees," in *Proc. VISAPP*, 2009.

[5] R. Marée, L. Wehenkel, and P. Geurts, *Decision Forests in Computer Vision and Medical Image Analysis*, chapter 10, pp. 125–142, Advances in Computer Vision and Pattern Recognition. Springer, 2013.

[6] B. Katigbak, C. Vergara-Niedermayr, D. Brat, D. Adler, F. Wang, J. Gao, J. Saltz, J. Kong, L. Cooper, T. Kurc, and Z. Zhou, "A high-performance spatial database based approach for pathology imaging algorithm evaluation," *Journal of Pathology Informatics*, vol. 4, no. 1, pp. 5, 2013.

[7] D.H. Douglas and T.K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Canadian Cartographer*, pp. 112–122, 1973.

[8] N. Rocks, S. Bekaert, G. Warnock, C. Sepult, R. Marée, J. M. Foidart, A. Noël, and D. Cataldo, "Roles of polarized neutrophils on lung tumour development in an orthotopic lung tumour mouse model," September 2013, Poster P3117 presented at the European Respiratory Society Annual Congress.

[9] C. Sommer, C. Straehle, U. Koethe, and F.A. Hamprecht, ""ilastik: Interactive learning and segmentation toolkit"," in *8th IEEE International Symposium on Biomedical Imaging (ISBI 2011)*, 2011.

[10] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J. Tinevez, D.J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: an open-source platform for biological-image analysis," *Nature Methods*, vol. 9, no. 7, pp. 676–682, July 2012.