Ontology-driven Image Analysis for Histopathological Images

Ahlem Othmani, Carole Meziat, Nicolas Loménie

Ontology-driven Image Analysis for Histopathological Images

Ahlem Othmani, Carole Meziat, and Nicolas Loménie*

CNRS - French National Center for Scientific Research
IPAL joint lab - UMI CNRS
Institute for Infocomm Research, A*STAR
http://ipal.i2r.a-star.edu.sg/

Abstract. Ontology-based software and image processing engine must cooperate in new fields of computer vision like microscopy acquisition wherein the amount of data, concepts and processing to be handled must be properly controlled. Within our own platform, we need to extract biological objects of interest in huge size and high-content microscopy images. In addition to specific low-level image analysis procedures, we used knowledge formalization tools and high-level reasoning ability of ontology-based software. This methodology made it possible to improve the expressiveness of the clinical models, the usability of the platform for the pathologist and the sensitivity or sensibility of the low-level image analysis algorithms.

1 Introduction

Usually in medical imaging, after the acquisition step, computer vision researchers propose new algorithms dedicated to a specific task like the segmentation of the liver out of MRI images or the counting of cells over stained images. For macroscopic natural images also, dedicated softwares for face recognition for instance have already been successfully delivered on the market so far. But, when considering new devices from satellite to microscopy imaging systems, the resolution and size at which images are acquired provide huge amount of biological and natural data to process in a parallel way, within a limited frame time and more or less on a pervasive mode in the near future [1].

For instance, the digitalization of biopsy images is raising new issues due to the exploration of what is called Whole Slide Images (WSI). For one patient, the amount of visual data to process over this WSI is about eight gigabyte. Various biological objects must be detected and segmented in order to infer any aid to the pathologist for the diagnosis. The spatial relationships between these different objects must be used as well to improve the efficiency of the automatic analysis of the data. As a matter of fact, if a WSI in histopathology is about a 50 000 by 40 000 pixels size image, it is now common to produce satellite images

* Corresponding author: nicolas.lomenie@mi.parisdescartes.fr

at very high resolution of about 30 000 by 30 000 pixels size (see the Pleiades satellite resolution).

It is not reasonable to systematically plan the design of image analysis algorithm on the fly as new needs are required. In any case, it will take time and money to improve the capacity of automatic annotation of these images. Ontology-driven interface and processing can be an alternative to this engineering constraints. First, systematically involving ontological descriptions on a platform improves the interaction standards with the novice end-user and also within the software designer team by modeling the knowledge and the objects in a formal way. Second, high-level reasoning based on the formalized concepts can provide an alternative way to detect biological objects. Last, it provides the end-user (like the pathologist) a semantic way to specify a query based on the results of the image analysis modules currently available in the system. From a pattern recognition point of view, it can help to lower the false alarm rate by adding high-level constraint rules or to improve the correct detection rate within a fixed time frame constraint by triggering the image analysis algorithms only on specific areas in the image defined by high-level spatial relationships rules for instance.

The ontology contribution is particularly relevant in the field of pathology and clinical imaging where a mental database is constantly used by the physician either coming from books or from his/her acquired experience over years of visual inspection of clinical data. This is the reason why our research work -even though aimed at being quite generic however- is however dedicated to a specific application and a platform we designed to automatically grade breast cancer out of histopathological images [2].

This work proposes to leverage the high-level reasoning and knowledge formalization ability of ontology-based softwares to make annotation of high-content images more efficient and interactive. Few works have operationally explored this kind of idea among which we can mention [3, 4].

Section 2 focuses on the low-level image analysis modules currently available in our system. Section 3 elaborates on the ontology part of the system and illustrates the use of the reasoning capability to infer new results or control the low-level image engine. Section 4 gives and discusses elements of quantitative assessment of the ontology-driven strategy before drawing a conclusion in the last section.

2 Low-level image annotation

The low-level image processing aims at outlining and describing general biological objects in the histopathological image. The current platform uses images from breast cancer biopsies. Three characteristics are used in breast cancer grading according to the Nottingham Grading System [5]:

- tubular formation of cells around lumina: the better formed the tubular formations are, the lower the cancer grade is;

- nuclear pleomorphism, that comes from nuclei features (area, mean and standard deviation intensity and circularity or roundness features): the bigger the nuclei are, the less regular their shape is and the less uniform their texture is, the higher the cancer grade is.
- mitosis number: the more mitoses are present in the image, the higher the cancer grade is.

Currently in our application, three different kinds of biological objects are detected to be able to provide an image with a cancer grade: the nuclei, the lumina and the invasive areas.

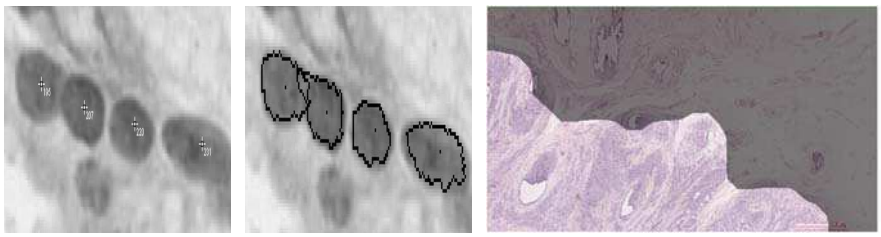
Nuclei segmentation. The nuclei detection module proceeds in two steps. First nuclei seeds are identified and then each detected nucleus is automatically segmented to extract geometric and radiometric features about it. The nuclei seeds extraction follows two processing steps: the regions of interest detection and then the nuclei identification (see Figures 1(b) and (c)).

(a) (b) (c)

Fig. 1: Nuclei identification - (a) Regions of interest detection, (b) nuclei identification (coarse nuclei separation), (c) nuclei identification based on a distance map.

The region of interest detection step locates the part of the images that contains the nuclei. Usually, cells are grouped together all around lumina and form what is called tubules. This step creates a mask to locate nuclei clusters that contain useful information. The following processing chain is performed: (1) Automatic image thresholding in order to distinguish the nuclei from the image background; (2) Morphological closure in order to group close nuclei together; (3) Removal of small objects not useful for ulterior processing or studies (see Figure 1(a)). The nuclei identification step proceeds by similar morphological filtering operators before drawing a distance map over which points within the nuclei area being the furthest from the boundaries are identified as the nuclei seeds (see Figure 2(a)). The nuclear boundaries are extracted using a snake-based method described in [6]. Patches of images that contain nuclei are extracted and are subjected to a polar transform of the coordinate system. After a first processing that constructs the first lace close to the real nuclei boundaries, the

iterative snake algorithm outlines the nuclei boundary (see Figure 2(b)). Then geometric and radiometric features can be extracted over each detected nucleus.



(a) (b) (c)

Fig. 2: An example of (a) seeds detection and (b) nuclei segmentation at high magnification x40 (c) invasive area detection at low magnification x1.2

Lumina and Invasive Area segmentation The low-level detection of the lumina uses mathematical morphology tools . The invasive ROI detection is currently casted as a classification problem whereby we exploited the relationship between human vision and neurosciences [7]. As the low-level processing part is not the core of this paper, we just give an illustration of the obtained results in our platform for the low-level detection of the invasive areas (see Figure 2(c)). The idea is now to exploit these biological landmarks to perform reasoning and knowledge management over the microscopy slide, as the extraction of all the biological concepts in an exhaustive way is not possible.

3 Ontology-driven image analysis

The algorithms briefly described in the previous section are actually standard low-level ones based on signal analysis: they allow a "black-box" detection of biological structures useful to draw a diagnosis based on a medical protocol. Yet, medical protocols or knowledge are constantly evolving or refining and can be very specific to an expert mental database as well, so that developing a complete image analysis platform in a complex field as histopathology is very costly and versatile in terms of engineering.

One way to overcome this issue and to facilitate the design of such complex, evolving platform as well is to work at a higher semantic level. For that purpose, the ontology framework constitutes a powerful tool for formalizing knowledge and for reasoning. In fact, the ontology is used to make the engineering of our heterogeneous knowledge database: medical knowledge but also image processing results coming from the image engine.

With this in mind, a few applications have been developed so far on our grading platform to experiment the benefits of articulating ontology capabilities

with image processing outcomes. They correspond to two generic objectives of our research work:

- Consistency-checking annotation: to improve the specificity rate;
- Image Analysis Engine Triggering Control: to improve the sensibility rate within a limited response time.

Before illustrating these two concepts, the next subsection draws a brief technical description of the core anatomical ontology for breast cancer grading we already built up.

3.1 Anatomical ontology: OWL

An ontology is a system of knowledge representation of a domain in the form of a structured set of concepts and relationships between these concepts. An ontology is expressed in the form of a XML graph and produces reasoning through a rule language. Our Breast Cancer Ontology (BCO) is based on two languages: OWL-DL (Web Ontology Language Description Logics) to describe the ontology and SWRL (Semantic Web Rule Language) to write and manage rules for the reasoning part. Technically, OWL and SWRL are specifications of the W3C[1], OWL is an extension of RDF (Resource Description Framework) used in the description of classes and types of properties, SWRL combines OWL and RuleML (Rule Markup Language) to produce the rules for the reasoning. The annotated images are described with the Wide Field Markup Language (WFML) [2] specific to the histopathology field (see Figure 4). Finally, the query language SPARQL (Simple Protocol And RDF Query Language) is used for querying in Java. SPARQL has been chosen for its ease of use and the very good integration of the API in Java. A thorough description of this ontology-based platform can be found in [8, 9].

3.2 Rules and reasoning

Once the anatomical and medical core concepts are formalized, we can feed our WFML database with new annotations based on a reasoning process.

Consistency Checking Annotation. Usually in the bio-medical field, the objects of interest are described by the biologists with high-level descriptions. However, the image analyzers use signal-based definition of these concepts. Subsequently, it is not uncommon to have the opportunity to cross both ways of defining a biological structure, like for the mitosis for instance. On an ideal platform, we will get two ways for defining mitoses:

- a low-level - in a sense implicit - signal-based extraction providing a set of results $\mathcal{R}^{signal}$, usually by statistical learning ;

[1] World Wide Web Consortium

[2] a XML language produced by the company TRIBVN for its platform ICS Framework

– an explicit high-level description corresponding to a SWRL rule like the one expressed in the Protégé[3] platform in Figure 3 and potentially providing a set of results $\mathcal{R}^{knowledge}$, and where Circularity and Roundness are the standard shape features.

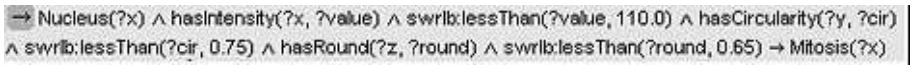


Fig. 3: A SWRL rule for mitosis description in our BCO (Breast Cancer Ontology) within the Protégé platform

In the case that $\mathcal{R}^{signal}$ is currently available on the platform, we can check the consistency of this result set by the semantic rule expressed in Figure 3 in the way Mechouche et al. proceeded for brain annotation issues [3]. This semantic checking will provide a set of results $\mathcal{R}^{signal \times knowledge}$ lowering the false alarm rate and subsequently improving the specificity rate of the image engine.

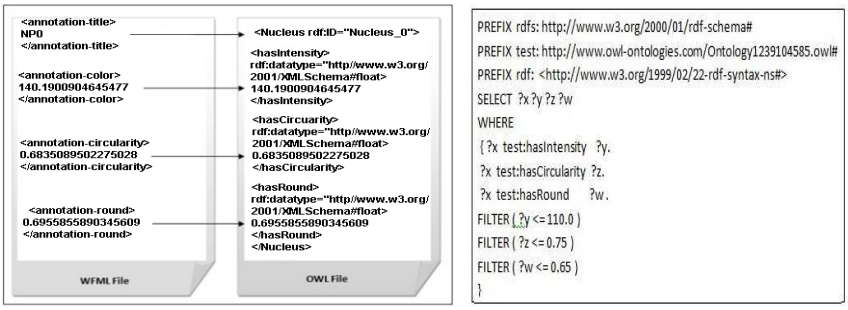
In the case that an image analysis module detecting the nuclei is currently available but not a mitosis detector, the platform can use the semantic rules defined by the pathologist on the fly to enrich the annotation WFML file by reasoning and providing a knowledge-based result set $\mathcal{R}^{knowledge}$ to this kind of semantic query. The basic principle of this case study is the following: from an original WFML file containing annotations about nuclei, we seek those corresponding to mitoses based on the semantic rule in Figure 3 and enrich the annotation file whenever it detects a mitosis. Technically, the WFML is parsed to retrieve information for each nucleus. The OWL file in the Protégé platform is powered by the list of nuclei in order to use the logic reasoning engine. This is the reason why we need a matching procedure between the WFML Files (specific to our application) and generic OWL Files (to benefit from the reasoning capability) as described in Figure 4(a). A reasoning procedure is then performed with the SPARQL query language which defines the syntax and semantics necessary to express queries on RDF type Database (Figure 4(b)).

Then the nuclei that are recognized as mitoses are modified in the WFML file by changing the annotation from a nucleus NP (standing for Nuclear Pleomorphism) into Mitosis (see Figure 5 for the global overview of the annotation updating process and Figure 6 for the WFML-based annotated resulting images [4]).

Image Analysis Engine Triggering Control. Another issue of high-content image annotation is the limited response time we must fit in. The ability to

[3] `http://protege.stanford.edu/`

[4] In the ICS^{TM} Technology interface from TRIBVN S.A., the image format is a SVS format involving both a pyramidal TIFF multiscale description and WFML description file for the annotations currently available in the database (`http://www.tribvn.com/`.



(a) (b)

Fig. 4: (a) Matching between WFML and OWL files (b) SPARQL query sample.

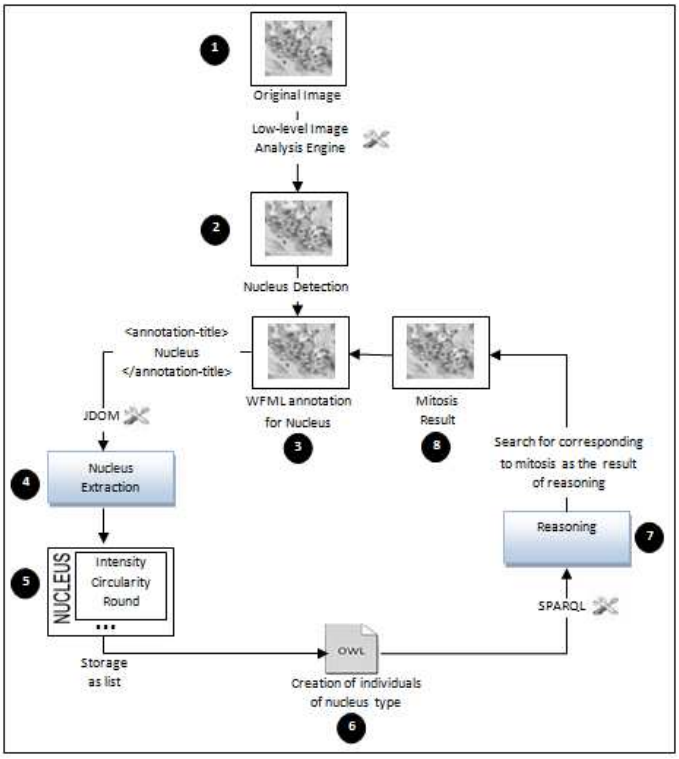
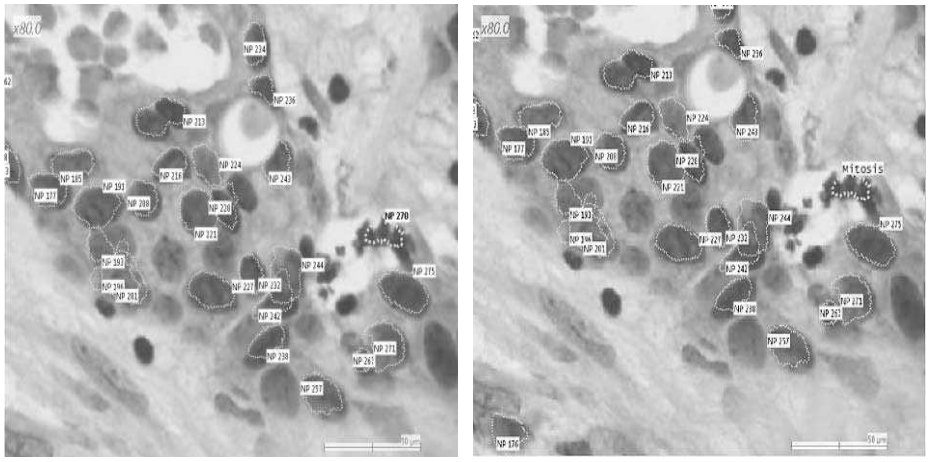


Fig. 5: Mitosis detection process

control the image analysis triggering over the While Slide Image can help to improve the sensibility rate of the platform under the time constraint.

From rule R1 for instance defined in first order logic in Equation 1, we can trigger the image analysis algorithms to detect the neoplasm as an invasive area,



(a) (b)

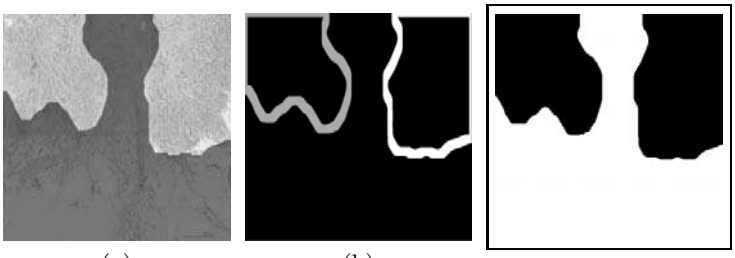
Fig. 6: (a) WFML before Mitosis Detection corresponding to step 3 in Figure 5 (b) WFML after Rule-based Mitosis Detection corresponding to step 8 in Figure 5

then locate formally the border of the neoplasm using a formalization of the spatial relation *Border* (see Figures 7 and 8).

$$R1 : Mitosis(X) \rightarrow \exists NeoplasmN/X \in Border(N) \quad (1)$$

Fig. 7: A SWRL rule for the expression of the spatial relationship constraint of Eq. 1 within the Protégé platform

For the location of the border, we model the spatial relationship " Around" as a landscape resulting from mathematical morphology operators such as dilations [4, 10] (see Figure 8). If a request is sent to the system to detect mitoses, we can scan the WSI image by image or does trigger that detector where this is relevant regarding the spatial relations constraints we are currently formalizing in the knowledge base according to the pathologist experience. The processing of the request is the result of a reasoning step that can evolve as new rules are added. For example, it triggers the rule R1 which is linked to mitoses. By doing this, we save between five and ten-fold increase in processing time which is of dramatic importance for WSI exploration. In addition, this kind of spatial relationship rules can help to check the consistency of $\mathcal{R}^{signal}$ results.



(a) (b) (c)

Fig. 8: An example of (a) invasive area detection at low magnification and (b) of its instantiated "Around" region after various morphological based filtering like (c) dilation and erosion of the invasive area

4 Results and discussion

The preliminary assessment of the ontology-driven annotation was achieved by quantifying the improvement of the grading platform specificity and sensibility rate related to the mitosis detection based on the previous ideas. The database is made of several histopathological samples as listed in Figure **??**.

Query with intensity. As quantified in Figure 9(a), all mitoses (TP=9) are detected but the detection is not very specific, there are a lot of false alarms. For 253 cells on average per frame, the algorithm returns 5 mitoses on average, sensitivity is equal to 1, all true mitoses were detected among the 5 mitoses returned. This could be a good diagnosis aid for histopathologists who can focus on the study of detected mitoses instead of having to analyze all the cells. This will reduce the workload and save time, which is of utmost importance in this field regarding the size of the images. The pathologist's task is complex and requires a lot of experience and we noticed an important point: real mitoses were detected by the semantic procedure but were not identified by the pathologist.

Query with intensity and geometrical constraints. A normal cell has a regular shape almost round or oval while a mitosis has an irregular shape and tends to divide. The test consists in reducing the false alarm rate and increasing the specificity by adding geometric constraints (see Figure 9(b)). In half the cases the algorithm becomes more specific, false alarms are reduced but the sensitivity may decrease in some cases. Of course, this result is very dependent of the low-level image processing algorithms. If the outlines of cells are correctly detected specificity gets better and sensitivity is maintained, the algorithm is efficient. In addition the algorithm works better on images with a grade NGS (Nottingham Grading System) equals to 1 than image with a NGS equals to 3. The reason is that cells are less deformed for grade 1, the algorithm is less dependent on the quality of segmentation and just need to locate each cell.

Query with intensity										
Image	Frame	Nb of cells	Correct Nb of mitosis	Nb of mitosis detected	TP	TN	FP	FN	Specificity	Sensitivity
IMG001	f001	257	3	6	3	251	3	0	0,988	1,000
IMG001	f003	314	2	4	2	310	2	0	0,994	1,000
IMG001	f008	296	2	3	2	293	1	0	0,997	1,000
IMG002	f001	244	0	8	0	236	8	0	0,967	1,000
IMG002	f004	242	0	18	0	224	18	0	0,926	1,000
IMG003	f002	297	2	4	2	293	2	0	0,993	1,000
IMG004	f002	261	0	6	0	255	6	0	0,977	1,000
IMG004	f003	214	0	1	0	213	1	0	0,995	1,000
IMG004	f011	229	0	5	0	224	5	0	0,978	1,000
IMG004	f016	234	0	5	0	229	5	0	0,979	1,000
		2588	9	60	9	2528	51	0	0,980	1,000

(a) Ontological intensity constraints

Query with intensity and geometrical constraints										
Image	Frame	Nb of cells	Correct Nb of mitosis	Nb of mitosis detected	TP	TN	FP	FN	Specificity	Sensitivity
IMG001	f001	257	3	2	2	255	0	1	1,000	0,667
IMG001	f003	314	2	3	2	311	1	0	0,997	1,000
IMG001	f008	296	2	3	2	293	1	0	0,997	1,000
IMG002	f001	244	0	0	0	244	0	0	1,000	1,000
IMG002	f004	242	0	0	0	242	0	0	1,000	1,000
IMG003	f002	297	2	3	2	294	1	0	0,997	1,000
IMG004	f002	261	0	5	0	256	5	0	0,981	1,000
IMG004	f003	214	0	1	0	213	1	0	0,995	1,000
IMG004	f011	229	0	3	0	226	3	0	0,987	1,000
IMG004	f016	234	0	0	0	234	0	0	1,000	1,000
		2588	9	20	8	2568	12	1	0,995	0,889

(b) Ontological intensity and geometry constraints

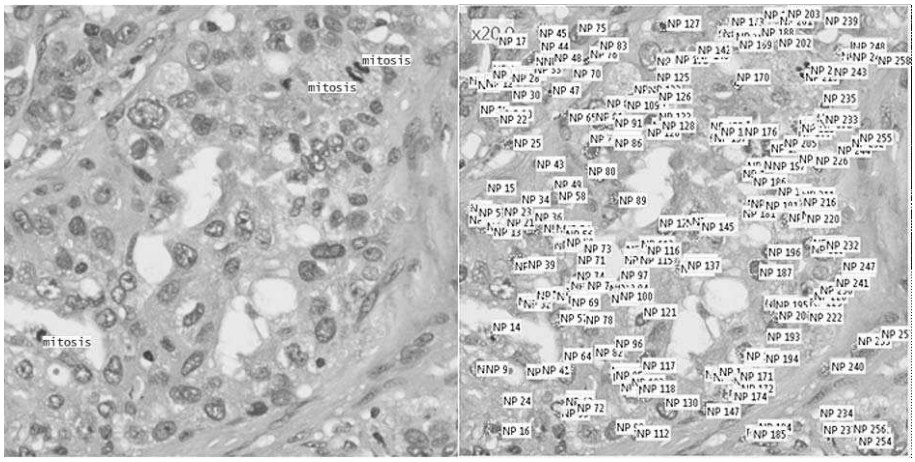
Fig. 9: Query results

A detailed test. The image in Figure 10 contains 258 cells and 3 mitoses. A pathologist detects three mitoses. By testing with the intensity without geometric constraints only six mitoses (N194, N170, N14, N82, N210, N224) are detected among which three true mitoses (N210, N224, N14). By testing with geometrical constraints like Circularity ≤ 0.75 and Roundness ≤ 0.65, only 2 mitosis are detected, one of the three true mitosis is not detected. The algorithm detects the mitosis N224 and N210 but not the mitosis N14. These results show that the algorithm with geometric constraints is more specific but that it decreases sensibility.

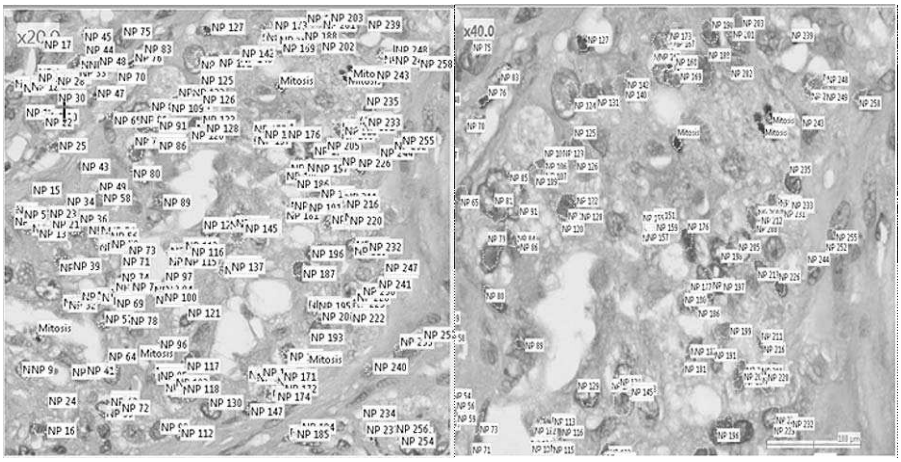
Furthermore, a qualitative assessment about usability by the novice end-user (that is the pathologist in our case) remains to be drawn. However, the formalization of the knowledge is a definite asset that the clinical world requires (see the European-based virtual physiological human project) for sharing and reusing tools that are developed worldwide [11]. In addition, for internal development requirement, the need for knowledge engineering in clinical and medical imaging fields is gaining momentum in order to be able to share issues and experience between the various key players of the platform design, from the imaging researcher to the clinician expert.

5 Conclusion

We experimented and made preliminary assessment of the articulation between ontology-based platforms and image analysis engine in a field where images contain a lot of complex, documented information, partly in the form of a mental database acquired by experience over years of practice. We showed that formalizing the knowledge can lead to improvement of image analysis results both in terms of specificity and sensibility of the pattern recognition system by involving semantic reasoning procedures. The new amount of visual data available in fields like satellite or bio-clinical imaging definitely calls for new paradigms whereupon



(a) Manual annotation (b) Low-level image annotation



(c) High-level image annotation (d) A zoomed-in area of the (c) image

Fig. 10: Detailed test image

knowledge engineering and computer vision issues must cooperate for the end-user benefits. In particular, digitized pathology is a growing market like digitized radiology has been over the previous couple of decades but with more than ever increased requirement for interoperability and expressiveness in the modeling [12]. Ontology-driven and reasoning based software engineering should play a key role for these new issues within the visual computing paradigm. In the next phase of our research program, we will consolidate the spatial relation reasoning part for the image analysis engine control.

References

1. Wu, C., Aghajan, H.: Using context with statistical relational models - object recognition from observing user activity in home environment. In: Workshop on Use of Context in Vision Processing (UCVP), ICMI-MLMI. (2009)
2. Dalle, J.R., Leow, W.K., Racoceanu, D., Tutac, A.E., Putti, T.C.: Automatic breast cancer grading of histopathological images. In: Proceedings of the 30th International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, Canada (2008) 3052–3055
3. Mechouche, A., Morandi, X., Golbreich, C., Gibaud, B.: A hybrid system using symbolic and numeric knowledge for the semantic annotation of sulco-gyral anatomy in brain MRI images. IEEE Transactions on Medical Imaging **28** (2009) 1165–1178
4. Hudelot, C., Atif, J., Bloch, I.: A spatial relation ontology using mathematical morphology and description logics for spatial reasoning. In: ECAI Workshop on Spatial and Temporal Reasoning. (2008) 21–25
5. Frkovic-Grazio, S., Bracko, M.: Long term prognostic value of nottingham histological grade and its components in early (pT1N0M0) breast carcinoma. Journal of Clinical Pathology **55** (2002) 88–92
6. Dalle, J.R., Li, H., Huang, C.H., Leow, W.K., Racoceanu, D., Putti, T.C.: Nuclear pleomorphism scoring by selective cell nuclei detection. In: IEEE Workshop on Applications of Computer Vision, Snowbird, Utah, USA (2009)
7. Miikkulainen, R., Bednar, J.A., Choe, Y., Sirosh, J.: Computational Maps in the Visual Cortex. Springer, Berlin (2005)
8. Tutac, A.E., Racoceanu, D., Leow, W.K., Dalle, J.R., Putti, T., Xiong, W., Cretu, V.: Translational approach for semi-automatic breast cancer grading using a knowledge-guided semantic indexing of histopathology images. In: Microscopic Image Analysis with Application in Biology, MIAAB, third MICCAI Workshop, 11th International Conference on Medical Image Computing and Computer Assisted Intervention, New-York, USA (2008)
9. Roux, L., Tutac, A., Lomenie, N., Balensi, D., Racoceanu, D., Leow, W.K., Veillard, A., Klossa, J., Putti, T.: A cognitive virtual microscopic framework for knowlege-based exploration of large microscopic images in breast cancer histopathology. In: IEEE Engineering in Medicine and Biology Society - Engineering the Future of Biomedecine, Minneapolis, Minnesota, USA (2009)
10. Lomenie, N.: Reasoning with spatial relations over high-content images. In: IEEE World Congress on Computational Intelligence - International Joint Conference on Neural Networks, Barcelona, Spain (2010)
11. Gianni, D., McKeever, S., Yu, T., Britten, R., Delingette, H.: Sharing and reusing cardiovascular anatomical models over the web: a step towards the implementation of the virtual physiological human project. Philosophical Transactions of The Royal Society A **368** (2010) 3039–56
12. Klipp, J., Kaufman, J.: Adoption trends in digital pathology. Laboratory Economics **5** (2010) 3–10