Translational Approach for Semi-Automatic Breast Cancer Grading Using a Knowledge-Guided Semantic Indexing of Histopathology Images

Adina Eunice Tutac[1,4], Daniel Racoceanu[1,6], Wee-Kheng Leow[1,3], Jean Romain Dalle[1,3], Thomas Putti[2], Wei Xiong[5], Vladimir Cretu[4]

[1]Image Perception, Access & Language IPAL (UMI CNRS 2955,UJF,NUS,I2R) Singapore,
[2]National University Hospital, Singapore, [3]National University of Singapore,
[4]Politehnica University of Timisoara Romania, [5]Institute for Infocomm Research, Singapore
[6]University of Besançon France

Abstract – **Within the last decade, histological grading has become widely accepted as a powerful indicator of prognosis in breast cancer. Currently, Breast Cancer Grading (BCG) is achieved by pathologists using tedious subjective visual examinations of hundred of slices day. In order to eliminate these drawbacks, we propose a semi-automatic grading system in a structured semantic Content-Based Image Retrieval (CBIR) framework. Although considered as an encouraging technology to enhance the intrinsic functionality of managing medical images, CBIR faces various issues with respect to clinical applications. One of these problems is the content gap, conceptually consisting of two major gaps: the semantic gap, defined as the discrepancy between the low-level visual features and high-level semantic concepts- and the context gap, which refers to the limitation of CBIR usage to a specific context. To tackle with these issues, this paper introduces two approaches, related to the semi-automatic breast cancer grading challenge: on one hand, a medical knowledge guided paradigm for semantic indexing of histopathology images, to overcome the semantic gap, and on the other hand, we propose a semi-automatic BCG approach, in order to improve pathologists' current manual procedures biased by subjectivity and tedious factors. The key idea is to build a Web Ontology Language standards compliant semi-automatic translation framework, from the medical concepts/rules related to the BCG, to computer vision (CV) concepts/symbolic rules. The application is related to a generic framework for BCG which narrows the context gap. This approach was tested over six breast cancer cases consisting of 7000 frames with domain knowledge from experts from the Pathology Department of Singapore National University Hospital. Our method provides pathologists a consistent approach for BCG and opens interesting perspectives for multi-scale image processing and analysis, semantic retrieval and bona-fide diagnosis/prognosis assistance.**

1 INTRODUCTION

Worldwide, breast cancer is the second most common type of cancer after lung cancer and the fifth most common cause of cancer death. Breast cancer is by far the most common cancer amongst women, with an increasing incidence rate.

Within the last decade, histological grading has become widely accepted as a powerful indicator of the prognosis in breast cancer. Currently, Breast Cancer Grading (BCG) is achieved by visual examinations (hundreds of slides per day) by the pathologists. Such a manual work is time-consuming and subjective. Thus, developing a semi-automatic grading system in a structured medical imaging framework represents an important medical requirement.

This study aims to introduce a medical knowledge guided paradigm for semantic indexing of histopathology images, applied to BCG. Our method proposes to improve pathologists' current manual procedures consistency of the diagnosis, by employing a semantic indexing technique, using a case/image based reasoning approach related to Nottingham BCG.

The paper is organized as follows. Section 2 introduces the concept of CBIR, describing the main characteristics, challenges, emphasizing our approach to overcome the semantic gap. Section 3 provides a generic description of BCG focusing on our solution to narrow the context gap and initiate a semi-automatic BCG. Section 4 introduces medical domain knowledge analysis and modeling by describing a synthesis of the breast cancer grading standard system and showing the importance of grading in breast cancer prognosis, followed by a breast cancer grading ontology model inspired from the medical concepts and rules and the specific rule modeling adapted to a translational approach between computer vision and pathologic rules. The semantic indexing of image features to give the local and the global grading is presented in section 6. Section 7 contains experiments and results leading to understanding semantic breast cancer image analysis, thus, to achieve the grading. Finally, the results and approaches are analyzed and research and clinical conclusion/perspectives are indicated.

Corresponding author : Adina Eunice Tutac, email: newlifeadit@gmail.com

2 CBIR IN MEDICAL APPLICATIONS

Content Based Image Retrieval is generally seen as a technology to access pictures from image database by visual content according to the users' interest [1] [2]. Principally, CBIR consists of three main phases: the indexing, the retrieval and the relevance feedback, typically based on visual similarity. The advantages of having CBIR systems oriented on medical axis are illustrated in Table I, along with some drawbacks related to specific techniques not yet used in medical.

Table I. Advantages & Drawbacks of Medical CBIR

	Advantages	*Drawbacks*
Medical CBIR	- increasing rate of everyday image production in hospitals - applications in diagnosis teaching & research	- usual relevance feedback doesn't allow capitalizing the contextual information (the process needs to restart from scratch for every new query) - user interfaces - performance - gaps

However, despite its promising characteristic to innovatively exploit actual huge amount of digital data, the clinical usage of CBIR is almost inexistent nowadays. One of the reasons is the complexity of a medical application. Another facet is related to the CBIR gaps. A comprehensive overview of these gaps in medical CBIR is provided by [3]. In particular, in [4] the spotlight is set on semantic and sensory gaps. An excellent review of Content-Based Medical Image Retrieval (CBMIR), showing that the semantic and sensory gaps inherently account for CBIR lack of significant clinical usage is given in [5]. A compilation of all gaps is presented in the Table II, with our own emphasis on *perception gap* instead of *aesthetic gap* proposed by [2].

Table II. CBMIR gaps

CBMIR gaps	*Characteristics*
Content	modeling & understanding image/information vs. real image/information
Features	computational numerical features vs. real image/information
Performance	application, integration, indexing, evaluation
Usability	query, feedback, refinement
Perception	Visual information perception vs. real image/information perception
Sensory	Information description vs. real image/information

Content-based image indexing [6] has been a subject of significant research in the context of medical imaging domain [7], [8]. Bridging the semantic gap [9] between low-level features and high-level semantic concepts [10] represents cutting-edge research [11], [12], since it is influenced by the versatility of image content and the lack of knowledge.

The research trend is to model ontologies and medical diagnosis rules that can capture the essence of domain knowledge and structure the information at the semantic level.

Our work is focused on finding a solution to bridge the semantic gap, by proposing a *knowledge-guided semantic indexing approach* based on a novel Breast Cancer Grading (BCG) ontology and rules base build in Protégé [13], a free, open source ontology editor and knowledge-base framework [14].

3 BREAST CANCER GRADING

Most of histological breast cancer grading systems [15] combine criteria in nuclear pleomorphism, tubule formation and mitotic counts. In general, each grading criteria is evaluated by a score of 1 to 3 (the grade 3 being associated to the most serious condition) and the score of all three components are added together to give the global grade. Breast Cancer Grading requires time and attention, dealing with hundreds of cases by day, each of them having around 2000 frames. Currently, BCG is achieved by visual examinations by pathologists. Such a manual work is time-consuming and inconsistent, according even to the pathologists' opinion. The diagnosis made on the same slide of the same patient by different pathologists or at different time during the week can differ. This is mainly related to the subjective manual scanning and evaluation of the mitosis, tubule formations and the cells nuclei. Considering these drawbacks, developing a semi-automatic grading system could considerably improve the consistency of the diagnosis.

Several approaches have been developed considering only individual parts of the BCG. Automated nuclear pleomorphism score was proposed by [16], [17], [18], while tubule formation score was addressed by [19] and mitosis count by [20]. Yet, no attempt has been done to combine all criteria in order to provide a complete automated BCG. Therefore, we propose a solution to meet pathologist's needs for a novel semi-automatic BCG, thus alleviating the shortcomings of the manual grading procedure.

Such a semi-automatic grading system should naturally be able to semantically index the images according to their content, in line with the medical domain knowledge. Beyond this, we further model the BCG-related medical knowledge (MK) as reasoning rules. These rules are embedded in the semantic indexing approach.

The proposed method provides pathologists a robust and consistent tool, as a second opinion for breast cancer grading, using the Nottingham grading system [15]. The actual precision of the proposed approach has been evaluated considering six breast cancer cases consisting of 7000 frames with domain knowledge from pathologist experts.

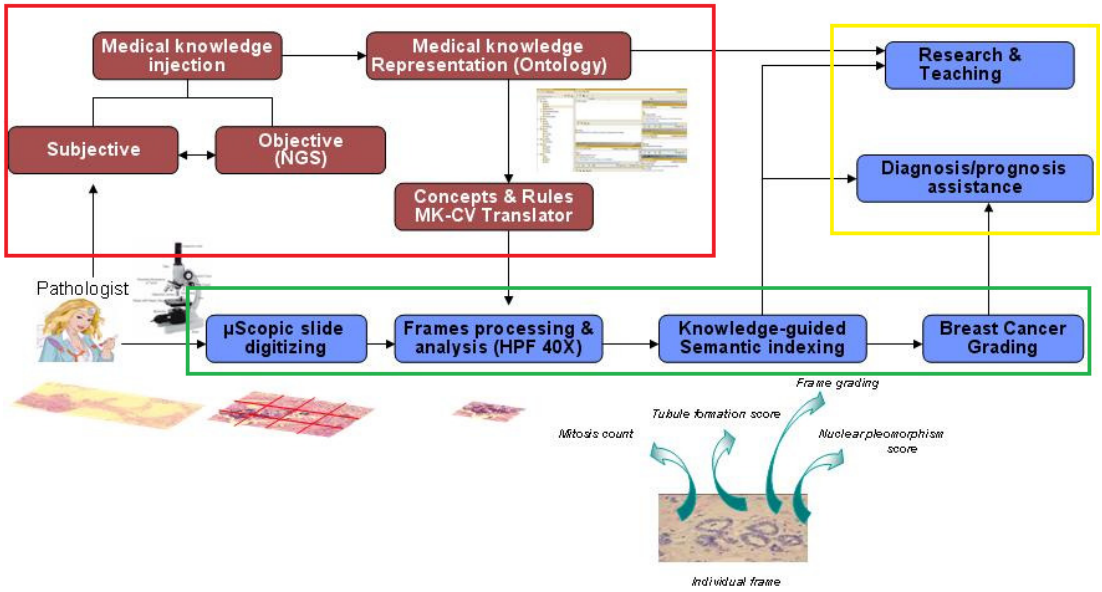


Figure 1. Knowledge-guided semantic indexing workflow (one time processing/training – red/upper left frame, each case testing- green/bottom frame, services provided upon request – yellow/upper right frame)

Our paradigm workflow depicted in Figure 1 respects the following steps. The first step consists of digitizing the histopathology slides analyzed under the microscope by the pathologist. The slide consists of 2000 to 3000 frames that will be processed and analyzed in the next step, by integrating the medical knowledge. The subjective knowledge coming from the pathologist as well as the objective knowledge coming from the Nottingham Standard Grading System are structured in a formal representation based on Breast Cancer Ontology. Medical knowledge concepts and rules are then translated into computer vision (CV) concepts and rules (required for the image processing and analysis step) thus providing the means for the semantic indexing step. The output will be the Breast Cancer Grading.

Breast Cancer ontology has its implications in research and teaching, knowledge-guided semantic indexing is of high interest in research & teaching as well as in diagnosis/prognosis assistance while BCG is nowadays the most used procedure for prognosis of breast cancer.

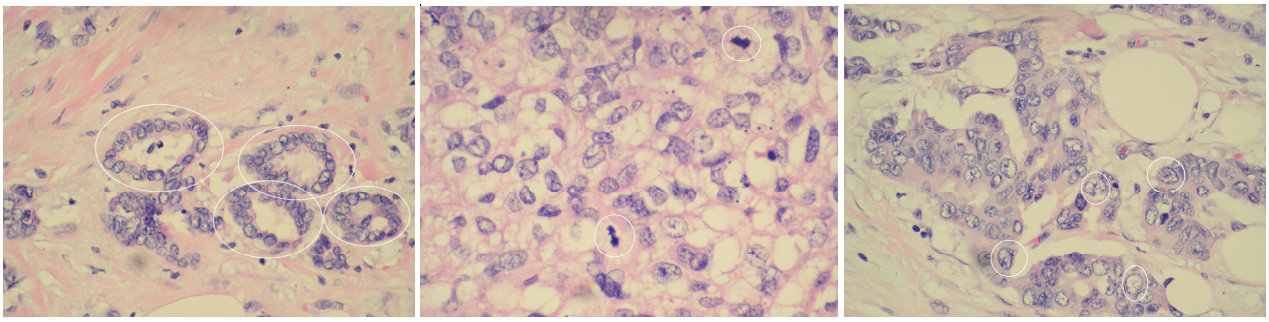
4 Medical Domain Knowledge

4.1 Domain Knowledge analysis

To have a complete domain knowledge analysis, *duality of objective knowledge and subjective knowledge* is required. In our case, objective knowledge is provided by the Nottingham Grading System (NGS) gold standard. Identification of the regions of interest (ground truth) with specific medical knowledge is given by the pathologists.

Among the standard grading systems used all over the world, NGS is preferred for the reason of providing more objective criteria for the three components of grading and specifically addresses mitosis counting in a more rigorous fashion. The three components of NGS criteria are briefly described below (see Figure 2):

- Tubule Formation score (TF) - are referred as the density of the Tubule Formations - white blobs (lumina) surrounded by a continuous string of cell nuclei.



a *b* *c*

Figure 2. NGS components : a) Tubule formation : lumina surrounded by string of cell nuclei , b) Mitosis: dividing cells nuclei, c) Big size / irregular shape nuclei-- NPS grade 3

- Mitosis Count (MC) score represent the number of Mitoses - diving cells nuclei. MC is assessed in the peripheral areas of the neoplasm and it's based on the number of mitoses per 10 High Power Field's (HPF's) – high resolution (usually 400X) frames obtained using microscopic acquisition.
- Nuclear Pleomorphism Score (NPS) - categorizes cells nuclei based on two features: size and shape.

The scores for the three separate parameters (tubules, nuclei and mitoses) are summated and the overall grade of the neoplasm is determined [1].

4.2 Breast Cancer Grading Ontology Design

We propose to narrow the semantic gap using a top-down approach. Our rationale is to associate meaning to features extracted from the image, thus indexing images by semantic means. Without any doubt, this has to be done according to the domain knowledge, a vital point in our approach. Therefore we propose a BCG knowledge modeling starting with the BCG ontology.

The modeling follows the Ontology Web Language (OWL) framework issued from the Semantic Web framework of Protégé [21], [14]. The key idea is to create definition rules for the concepts, to define the relationships between concepts in terms of classes, properties and instances. Various instances for different classes are created as individuals, where specific values are assigned for classes and properties. Figure 3 gives an insight of Breast Cancer Ontology development.

Table III illustrates the representation of breast cancer grading knowledge (medical concepts and rules) into computer vision ontology concepts and definition rules in Protégé. To give a glimpse, lumina medical concept is defined in the ontology as a class inherited from the WhiteBlobs (compact segments of white parts) with associated properties. They are following the medical rules (semantic meaning) and using visual features (low-level) of the histopathology images. The property hasIntensity White (which is an instance of Intensity class) is correlated with color feature (intensity-based), hasSize is defined as an instance of Size class, related to the dimension and Included_In (an instantiation of Localization class) DarkCellsCluster shows where the lumina appears in the microscopic image, surrounded by the dark cells cluster (inherited from Cells), meaning the detection of tubule Formations.

Table III. CONCEPTS and Rules correspondence (Medical –CV-Protégé)

Medical Concepts	Protégé concepts type
Slide	Super class
Grading	Super class
Cells	Class inherited from Image – *relations*
CellsCluster	Class inherited from Cells - *relations*
DarkCellsCluster/ VeryDarkCellsCluster	Class inherited from Cells- relations hasIntensity – *attributes (property)* Dark/VeryDark (*instances* of Intensity class)
Lumina	Class inherited WhiteBlobs hasIntensity (*property*) White (*instance* of Intensity class), hasSize (*property*) Small (*instance* of Size class), hasLocalization (*property*) Included_In (*instance* of Localization class) DarkCellsCluster (*instance* of Cells)
TF/Mitosis/NP	Classes inherited from Grading- *relations*
Local Grading/Global Grading (10 HPFs)	Class inherited from TubuleFormation/ MC/ NPS- *relations*

4.3 Breast Cancer Grading Rule-Base System Modeling

The scope of this section is to introduce the approach proposed to translate medical concepts and rules related to the breast cancer grading, to the computer vision (CV) concepts and symbolic rules. The aim is to move towards a future generic frame for an assisted semi-automatic generation of CV rules and (in future) computer programs, starting from specific medical queries. Therefore, we define

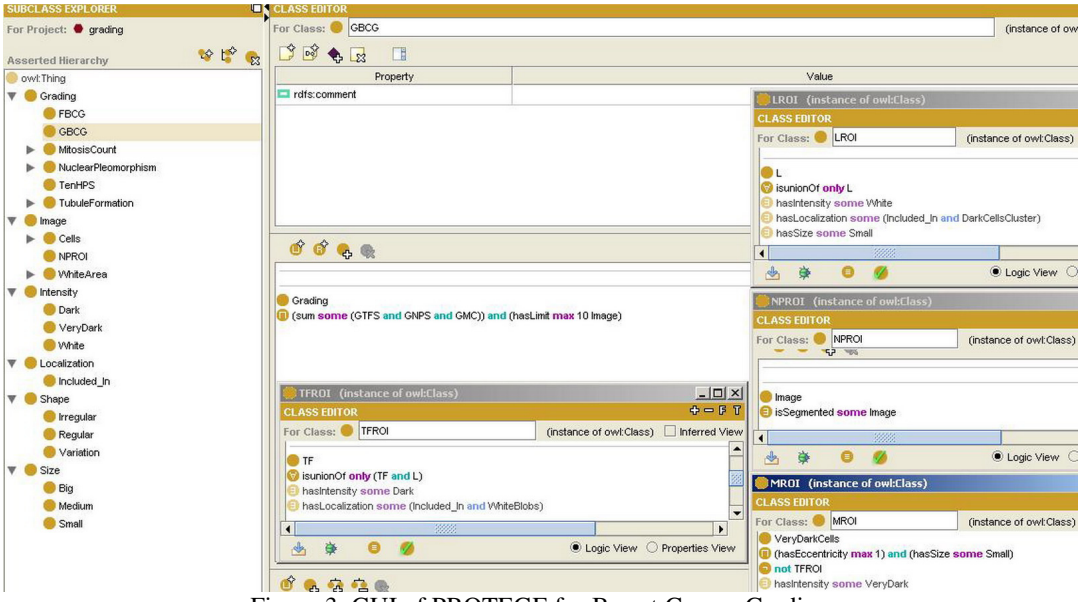


Figure 3. GUI of PROTEGE for Breast Cancer Grading

a Generic Translation Framework (GTF) to provide the translation of medical concepts and rules into CV concepts and rules.

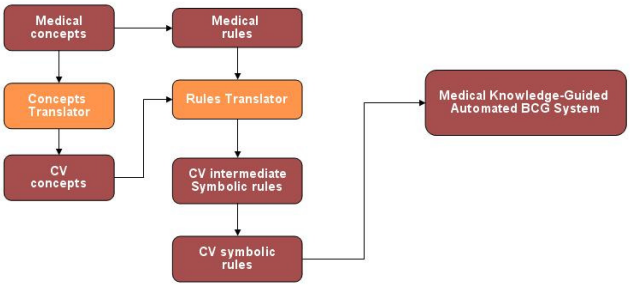


Figure 4. Generic Translation Framework

As illustrated by Figure 4, the section is structured in three parts, according to the main steps of the proposed approach: development of the correspondence between medical concepts and computer vision concepts (with respect to the OWL standard); definition of intermediate CV rules and generation of the final Symbolic Rules by fusion of the CV concepts and intermediate CV symbolic rules. Note that CV concepts are used as an input for the Rules Translator together with the medical rules to generate the CV preliminary symbolic rules.

4.3.1 *Correspondance between the Medical concepts and adequate Computer Vision concepts*

According to the NGS synthesis, we proceed to the TF extraction as the NPS and MC computation in order to create the rule-based method able to automatically generate the grading (see Table IV). Therefore, to clearly define the rules, medical concepts are transformed into computer vision concepts.

The MK-CV concept translator is based on a classification of elements that need to be taken into consideration to give the final grading.

- objects : Image, Cells, CellsCluster, Lumina, Tubule, Mitosis, Nuclei Pleomorphism
- attributes: size, shape, intensity, localization
- values : small, medium, big, regular, variated, irregular, white, dark, very dark, ecc
- operators : $\exists, \cup, \subset, \supset, \not\subset, \sum, Area, count$

An illustration of objects translation is given by Table IV.

Table IV. MK-CV objects of concept translator

Medical Objects	CV objects
Slide	Image (digitized)
Cells	Cells
CellsCluster	Union of Cells
DarkCellsCluster/ VeryDarkCellsCluster	Union of Cells
Lumina	White compact segments of the Image included in the union of dark cells
TF/Mitosis/NP	Union of Cells/Diving Cells nuclei/ dimension & shape features of the nuclei
Local Grading/Global Grading	Grading Computation for TF.MC, NP single frame/10 frames

4.3.2 *Intermediate rules*

To obtain the symbolic tubule formation rule, we create intermediate rules for each domain concept used for this criterion.

- *DarkCellsCluster* is defined as containing group of adjacent cells with intensity property value setup between *VeryDark* and *White* limits.

$$DarkCellsClusters = \{\bigcup morphology(adjacent(Cell_i)) \mid VeryDark < intensity(Cell_i) < White\} = \{DarkCellsCluster_c\}_{c=\overline{1,C}}$$

In terms of Protégé, this rule is defined as: Cells with *hasIntensity* (property) some Dark, which is an instance of Intensity class.

WhiteBlobs intermediate rule composes the Lumina (L) rule as white blobs included in the existing DarkCellsCluster.

$$WhiteBlobs = \{morph(WhiteArea)\} = \{WhiteBlob_k\}_{k=\overline{1,b}}$$

$$L_k = \{WhiteBlob_k \mid \exists c \in \overline{1,C}, DarkCellsCluster_c \supset WhiteBlob_k\}$$

- L_{ROI} intermediate CV rule is a union of all lumen detected in the image.

$$L_{ROI} = \bigcup_{k=\overline{1,l}} L_k$$

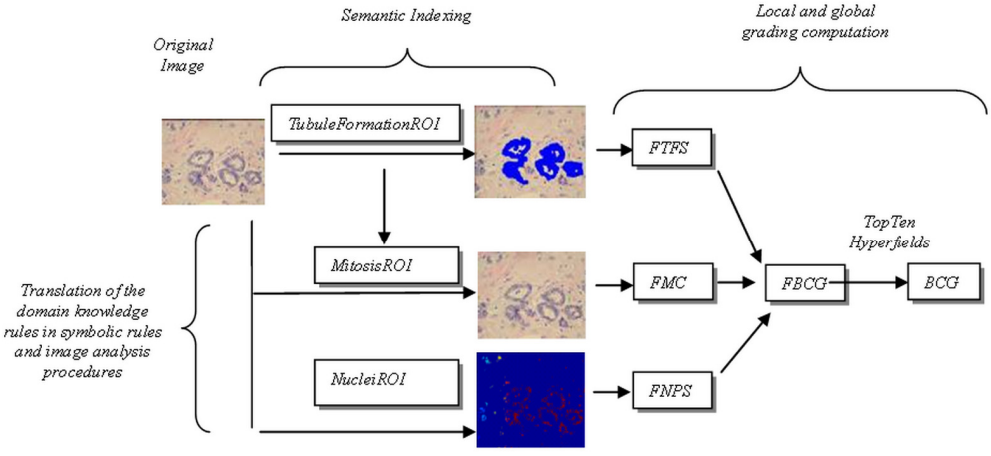


Figure 6. SEMANTIC Indexing in BCG Context

Following the same idea, intermediate rules are defined for the mitosis count symbolic rule.

$$VeryDarkCells = \{\bigcup Cells_j \mid intensity(Cells_j) \le VeryDark\} = \\ = \{VeryDarkCell_j\}_{j=\overline{1,J}}$$

$ecc(VeryDarkCell_j)$ rule represents an eccentricity deterministic operation computed for the *VeryDarkCells.*

- $size(VeryDarkCell_j)$ rule applies a size detection threshold onto the *VeryDarkCells*. In practical image processing/analysis, the detection of *DarkCellsCluster, VeryDarkCells* or *WhiteBlobs* becomes a simple intensity-based segmentation method.

For the nuclear pleomorphism rule definition, image segmentation methods are performed to detect

$$size(DarkCell_i), shape(DarkCell_i).$$

4.3.3 *Generation of the final symbolic computer vision rules*

Considering the tubule formation criteria given by the pathologist:

- Pathologist rule for Tubule = white lumina blobs surrounded by string of dark cells nuclei."
- Symbolic rule (used in our algorithm):

TF symbolic rule specifies that if there are *WhiteBlobs* included in the *DarkCellsCluster*, the pathologic criterion is satisfied.

$$TF_c = \{DarkCellsCluster_c \mid \exists WhiteBlob_k \subset DarkCellsCluster_c\}$$

where: TF_{ROI}*(TF region of interest)* is defined by:

$$TF_{ROI} = \{\bigcup DarkCellsCluster_c\}$$

The TF_{ROI} symbolic rule creates the union of all *DarkCellsCluster* – with intensity and localization dependence and *L*. The $\bigcup DarkCellsCluster$ detection is performed using morphologic operators.

The result of this operation is to index the medical image by the TF_{ROI}. This is an important point of our approach, since we are able to associate to each frame precise ROI structure corresponding to the detected tubule formations.

- Pathologist Rule: Mitosis = very dark dividing cells nuclei from the peripheral area of neoplasm
- Symbolic Rule:

MitosisROI:

$$M_{ROI} = \{\bigcup VeryDarkCell_j \mid ecc(VeryDarkCell_j), \\ size(VeryDarkCell_j), VeryDarkCell_j \not\subset TF_{ROI})\}$$

VeryDarkCells structures must not be contained in the tubule formation area (TF_{ROI}), specified in the rule by the $\not\subset$ operator. Thus, M_{ROI} rule is defined as a union of all *VeryDarkCells* dependent of particular *ecc*, *size* and *localization* values.

- Pathologist Rule for Nuclear Pleomorphism: *Size and Shape* features of nuclei
- Symbolic Rule :

Nuclei$_{ROI}$: $NP_{ROI} = \{segment(\mathrm{Im})\}$ where $segment(\mathrm{Im}) = \{size(DarkCellsCluster_i), shape(DarkCellsCluster_i)\}$ A n example of MK-CV rules translation, in the case of mitosis detection, is given by:

- Mitosis Pathologist Rule :

  very dark diving cells nuclei from the peripheral area of neoplasm

- Mitosis CV intermediate symbolic Rule :

$$VeryDarkCells = \{\bigcup Cells_j \mid intensity(Cells_j) \le VeryDark\} = \\ = \{VeryDarkCell_j\}_{j=\overline{1,J}}$$

- Mitosis CV symbolic Rule :

$$M_{ROI} = \{\bigcup VeryDarkCell_j \mid ecc(VeryDarkCell_j), \\ size(VeryDarkCell_j), VeryDarkCell_j \not\subset TF_{ROI})\}$$

5 Semantic Indexing Approach. Breast Cancer Grading Computation

This approach intends to overcome the drawbacks of classical indexing methods. The conceptual annotations are rule-based defined in the grading model for every particular frame and globally transmitted in a structure for the entire case. Some prerequisites need to be considered. An important role is played by the scale. The images have been scanned at high-magnification (10X), thus the image processing and analysis step is based on this scale. Another idea that was conveyed in this approach is the computation of local grading, at the first hand, in order to provide better results. The local grading for each frame is used further for the global grading, instead of computing directly the grading for the whole slide.

The algorithm segments images and processes the object recognition phase (feature extraction step) followed by the semantic classification criteria rules modeling. Thus, it is created a correspondence between the visual features and the semantic image labeling, in terms of *Mitosis*, *Nuclei* and *Tubule Formation* regions of interest - *ROIs*.

Image segmentation with gray scale conversion and adaptive tresholding obtains a collection of such ROI, meaningful for breast cancer grading and – more generally – for breast cancer evolution diagnosis/prognosis. The region selection is correlated with the model rules (see Figure 6).

Semantic indexing of concepts extracted from the image gives us the means to create the rules for the computation of local grading.

5.1 Local Grading Computation

The local grading computation process uses functions and operators to define the required symbolic rules (see Table V).

Table V. FUNCTIONS used in criteria score symbolic rules

Symbolic rules	Description
$f_{FTFS}(x) = \begin{cases} 1, x > 0.75 \\ 2, 0.10 < x < 0.75 \\ 3, x < 0.10 \end{cases}$	the TF score as reported in the pathologist rule
$f_{FMC}(x) = \begin{cases} 1, x < 9 \\ 2, 10 < x < 19 \\ 3, x > 19 \end{cases}$	the MC grade function with the NGS values
f(*Size*)+g(*Shape*)	The pleomorphism value off all nuclei

Frame Tubule Formation Score (FTFS) :

$$FTFS = \{f(Area(TF_{ROI})/Area(DarkCellsCluster))\}$$

Frame Mitosis Count (FMC):

$$FMC = \{f(count(M_{ROI}))\}$$

Frame Nuclear Pleomorphism Score (FNPS):

$$FNPS = \{\text{round}(\sum_{i=1}^{\text{count}(NPROI)} (\text{f}(Size) + \text{g}(Shape))/\text{count}(NP_{ROI}))\}$$

The local breast cancer grading (FBCG) rule

$$FBCGi = \{f(FTFSi + FMCi + FNPSi), i = frameID\}$$

represents the sum of the three values computed for each NGS criterion, over a single frame.

5.2 Global grading computation

The global breast cancer grading is computed similar with each local score, but over 10 HPFs [1] (see Figure 6). The 10 HPFs specification appears as the upper limit at each computation of sum in the rules:

$$GTFS = \left\{ f_{TF} \left(\frac{\sum_{j=1}^{10} Area(TF_{ROI_j})}{\sum_{j=1}^{10} Area(\text{Im}_j)} \right) \right\}$$

$$GMC = \left\{ f_{MC} \left(count \left\{ \sum_{j=1}^{10} M_{ROI_j} \right\} \right) \right\}$$

$$GNPS = \left\{ \frac{f_{NP} \left(\sum_{j=1}^{10} \left\{ \sum_{k=1}^{count(NP_{ROI_j})} \left(f(Size_{kj}) + gShape_{kj} \right) \right\} \right)}{\sum_{j=1}^{10} count(NP_{ROI_j})} \right\}$$

$$GBCG = \{f(GTFS_j + GMC_j + GNPS_j), j = \{1,...10\}\}$$

6 Experiments & Results

The experimental part consists in analyzing and indexing pathologic images of six breast core-biopsy cases stained with H&E marker, consisting of 7000 frames scanned from the tumor tissue slides and obtained from the Pathology Department of National University Hospital of Singapore (NUH). The database is composed by two sets: 1400 frames used for the training algorithm phase and 5600 frames used for the testing and validation phase. The slides were scanned on a sequence of frames at 10X40 (400X) magnification with a 1080 X 1024 resolution.

The set of histopathology slides, labeled by our medical partners, has been digitized into a number of hyperfields (frames). Each frame is then analyzed and a local grading is computed. According to this local grading, top ten images are automatically retrieved to provide the slide global grading.

Table VI. PATHOLOGIC visual grading and configuration of the training and testing database

Data type	Case ID	Tubule score	Nuclear score	Mitosis count	BCG (path)
Training database (1400 images)	1000	1	1	3	*1*
	2000	1	2	1	*1*
	4895	3	3	3	*3*
Testing database (5600 images)	5020	2	3	3	*3*
	5042	3	3	2	*3*
	5075	3	2	1	*2*

Table VII. Semi-AUTOMATIC grading results

Data type	Case ID	Tubule score	Nuclear score	Mitosis count	automatic BCG
Training database	1000	1	1	3	***1***
	2000	2	2	1	***1***
	4895	3	2	3	***3***
Testing database	5020	3	2	3	***3***
	5042	3	2	3	***3***
	5075	3	2	1	***2***

Table VIII. COMPONENT scores and global grading errors

Data base	Tubule score	Nuclear score	Mitosis count	Component scores error	Global BCG error
Training errors	11%	11%	0	7,33%	0%
Testing errors	11%	22%	0	11%	0%

We use Matlab programming environment to develop the method. The program is tuned to take into account the images' scale [22] given by the microscope [23] in the automatic acquisition phase. Local errors were registered in the training base for the tubule score in case 2000 and for the nuclear score in case 4895. In the testing database, local errors were obtained at the tubule score and nuclear score for the case 5020 and only for the nuclear score in 5042 case. Note that for the mitosis count there was no registration in either training or testing database which gives us a good confidence degree in the detection of mitosis. (100% automated detection). The most interesting fact is that, when computing the BCG for training and testing database respectively, local errors (7.33%, 11%) are not propagated to the global level (0), computed by a simple formula of matches from the total items. The good results obtained on the global grading are promising and allow us to envisage interesting generic perspectives of this approach.

7 Discussions, Conclusion and Perspectives

Despite being strongly related to a particular application field and a specific medical domain, the presented semantic indexing approach has a generic character. Indeed, in association with the following ontology validation procedure:

- Ontology segmentation - extract a subset of existing thesauri (i.e. NCI thesaurus)
- Ontology structuring - according to the OWL standards (related to Semantic Web standards); for this purpose we are using the same Protégé framework.
- Automatic Ontology verification - use of the Protégé reasoner in order to have a first evaluation of the ontology consistency
- Medical validation - consult the local medical collaborators (important, even if somehow subjective and partial)
- Ontology official validation - submit the ontology to the OBO web site (www.obofoundry.org/), agreeing that - once validated by them - our ontology will be published on their website.

this meaningful (medical domain relevant) semantic index allows to design semantic query content-based medical image retrieval systems, usable in translational approaches. These types of CBIR systems will certainly replace in the near future the actual query by example ones, based only on visual features.

In the context of virtual microscope platform, automatic semantic-query based visual positioning systems [24] present also a high interest for the medical technicians and doctors in terms of time efficiency. From the image processing and analysis standpoint, we envisage that a multi-scale approach will provide an improvement in terms of speed and time consuming task. This reasoning follows even closer the pathologist procedure and also helps in a better identification of the invasive area (neoplasm) thus the grading will be proceeded only on this specific part of the slide.

In addition, the purpose of generating computer vision (CV) concepts and symbolic rules from medical concepts/rules related to the breast cancer grading, with respect to OWL and the Semantic Web is seen as future generic perspectives for an assisted semi-automatic generation of CV rules and computer programs, starting from specific medical queries/rules. Obviously, a true maintenance mechanism has to be included in the future in the existing approach.

Apart coping with the semi-automatic breast cancer grading challenge, our approach emphasizes medical imaging vital importance for accurate "bona-fide" diagnosis assistance. Indeed, a virtual microscope platform based on our principle will allow pathologist to ensure a robust grading procedure, by providing the opportunity to benefit from a semi-automatic semantic annotation and further exploration of the lesions' neighborhood (region of interest-ROI) at different scales.

ACKNOWLEDGMENT

This project is supported by the ONCO-MEDIA[1] project.

REFERENCES

[1] F. Long, H. Zhang and D. Feng, "Fundamentals of Content-Based Image Retrieval", pp. 1-26, 2001

[2] R. Datta, D. Joshi, J. Li and J. Wang, "Image Retrieval: Ideas, Influences, and Trends of New Age", *ACM Transactions on Computing Surveys,* pp.1-66, 2008

[3] T. Deserno, S. Antani, and R. Long, "Gaps in content-based image retrieval", pp.1-11, 2007

[4] A. Smeulders, M. Worring, S. Santini, A. Gupta and R.Jain, "Content-Based Image Retrieval at the End of Early Years", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.22, no.12, pp. 1349-1480, 2000

[5] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler, "A Review of Content-Based Image Retrieval Systems in Medical Application- Clinical Benefits and Future Directions", *International Journal of Medical Informatics*, vol. 73, pp. 1-30, 2004

[6] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.29, no.3, 2007

[7] R. Zhao and W. Growski, "Bridging the Semantic Gap in Image Retrieval, Distributed Multimedia Databases: Techniques and Applications", 2002

[8] N. Vasconcelos, "From Pixels to Semantic Spaces: Advances in Content-Based Image Retrieval", pp.20-26, 2007

[9] Y. KAlfoglou, S.Dasmahapatra, D.Dupplow, B.Hu, P.Lewis, N. Shadbolt, "Living with the Semantic Gap: Experiences and Remedies in the Context of Medical Imaging", *1st International Conference on Semantics and Digital Media Technologies*, 2006

[10] S. Little and J. Hunter, "Rules-By-Example- A Novel Approach to Semantic Indexing and Querying of Images", *International Semantic Web Conference ISWC*, pp.534-548, 2004

[11] Y. Liu, N. Lazar, W. Rothfus, F. Dellaert, A. Moore, J.Schneider, and T.Kanade, "Semantic - based Biomedical Image Indexing and Retrieval", *Trends and Advances in Content- Based Image and Video Retrieval", Shapiro, Kriegel and Veltkamp ed.,* pp. 1-20, in press, 2004

[12] H. Tang, R. Hanka, and H.Ip, "Histological Image Retrieval Based on Semantic Content Analysis", *IEEE Transaction on Information Technology Medicine*, vol. 7, no. 1, 2003

[13] Tutac AE, Racoceanu D, Putti T, Xiong W, Leow W-K, Cretu V., "Knowledge-Guided Semantic Indexing of Breast Cancer Histopathology Images", BioMedical Engineering and Informatics: New Development and the Future, *Proceedings of the First International Conference on BioMedical Engineering and Informatics*, Editors: Yonghong Peng and Yufeng Zhang, Published by IEEE Computer Society, 27 - 30 May 2008, Sanya, Hainan, China, pp. 107-112

[14] The Protégé Ontology Editor and Knowledge Acquisition System - http://protege.stanford.edu/

[15] A. Tutac, "Histological Grading on Breast Cancer", *IPAL internal report 2007*, MIIRAD/IPAL – BCG, 2007

[16] C. Demir and B. Yener, "Automated cancer diagnosis based on histopathological images: a systematic survey", Tech Rep, 2005

[17] H. Jeong, T. Kim, H. Hwang and H-J. Choi, "Comparison of thresholding methods for breast tumor cells segmentation", in *Proc of 7th Int. Workshop on Enterprise networking and Computing in Healthcare Industry,* pp. 392-395, 2005

[18] M. Adawi, Z. Shehab, H. Keshk and M. Shourbagy, "A fast algorithm for segmentation of microscopic cell images", in *Proc. 4th Int. Conf. Inf. & Com. Tech,* 2006

[19] S. Petushi, F. Garcia, M. Haber, C. Katsinis, and A. Tozeren, "Large- Scale Computation on histology images reveal grade-differentiating parameter for breast cancer", pp. 1-11, 2006

[20] J.A. Beliën, J.P. Baak, P.J. van Diest and A.H. Ginkel, "Counting mitosis by image processing in feulgen stained breast cancer sections: the influence of resolution", *Cytometry,* vol.28 (2), pp.135-140, 1997

[21] D.L. McGuiness and F.van Harmelen, "OWL Web Ontology Language W3C Overview", pp. 1-26, 2004

[22] P. Van Osta, J.M. Geusebroek, K. Ver Donck, L. Bols, J. Geysen, and B. Romeny, "The Principles of Scale Space applied to structure and color in light microscopy", *Proceedings RMS*, vol. 37, no. 3, 2002

[23] I. Marandet, A. Tutac, "Smart Microscope User Guide", *IPAL internal report 2006*, MIIRAD/IPAL -µ-MediSearch, 2006

[24] G. Begelman, M. Lifshits, and E. Rivlin, "Visual Positioning of Previously Defined ROIs on Microscopic Slides*", IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, 2006

[1] ONCO-MEDIA (ONtology and COntext related MEdical image Distributed Intelligent Access) - ICT ASIA International Project – www.onco-media.com