Knowledge-Guided Semantic Indexing of Breast Cancer Histopathology Images

Adina Eunice Tutac[1,5], Daniel Racoceanu[1,6], Thomas Putti[2], Wei Xiong[1.4],
Wee-Kheng Leow[1,3] Vladimir Cretu[5]
[1]IPAL UMI CNRS 2955 Singapore, [2]NUH Singapore, [3]NUS Singapore, [4]I2R A-STAR Singapore,
[5]Politehnica University of Timisoara Romania, [6]University of Besançon France
newlifeadit@gmail.com, danielr@comp.nus.edu.sg

Abstract

Narrowing the semantic gap represents one of the most outstanding challenges in medical image analysis and indexing. This paper introduces a medical knowledge – guided paradigm for semantic indexing of histopathology images, applied to breast cancer grading (BCG). Our method improves pathologists' current manual procedures consistency by employing a semantic indexing technique, according to a rule-based decision system related to Nottingham BCG system. The challenge is to move from the medical concepts/ rules related to the BCG, to the computer vision (CV) concepts and symbolic rules, to design a future generic framework- following Web Ontology Language standards - for an semi- automatic generation of CV rules. The effectiveness of this approach was experimentally validated over six breast cancer cases consisting of 7000 frames with domain knowledge from experts of Singapore National University Hospital, Pathology Department. Our method provides pathologists a robust and consistent tool for BCG and opens interesting perspectives for the semantic retrieval and visual positioning.

1. INTRODUCTION

Within the last decade, histological grading [1] has become widely accepted as a powerful indicator of prognosis in breast cancer. Most grading systems currently employed for breast cancer combine criteria in nuclear pleomorphism, tubule formation and mitotic counts. In general, each grading criteria is evaluated by a score of 1 to 3 (3 being associated to the most serious case) and the score of all three components are added together to give the "grade". Breast Cancer Grading (BCG) [1] requires time and attention while classifying 100 cases/ day, each of them having around 2000 frames. Currently, BCG is achieved by visual examinations of pathologists. Such a manual work is time-consuming and inconsistent. According to these issues, developing an automatic grading system represents a strong medical requirement.

Such an automatic grading system should naturally be able to semantically index the images in line with the medical domain knowledge, and inspired from their real content. Content-based image indexing [3], [5] has been subject of significant researches in the context of medical imaging domain [4]. Solving the issue of the semantic gap [6] between the low level features and the high level semantic concepts [7] represents the cutting edge research [8], [9].

In this paper, we propose a solution to meet pathologist needs for automatic BCG. Beyond this, we further model the BCG-related medical knowledge into reasoning rules. These rules are embedded in semantic indexing approaches.

The proposed method provides pathologists a robust and consistent tool, as a second opinion, for breast cancer grading, using the Nottingham grading system [1] [2]. The effectiveness of the proposed approach has been validated in experiments over six breast cancer cases consisting of 7000 frames with domain knowledge from pathologist experts.

The paper is organized as follows. Section 2 introduces domain knowledge analysis by describing a synthesis of the breast cancer grading standard system and showing the importance of grading in breast cancer detection. Section 3 presents our grading approach model with the medical image indexing inspired rules. The semantic indexing of image features to give the local and the global grading is presented in section 4. Section 5 contains experiments and results leading to understanding semantic breast cancer image analysis, thus, to achieve the grading aim. Finally, the results and approaches are analyzed and conclusion/perspectives are indicated.

2. Domain Knowledge analysis

Breast cancer refers to a malignant tumor that has developed from cells within the breast. Breast cancer is a leading cause of death among women, and its incidence is rising. Although curable, especially when detected at early stages, breast cancer is expected to account for 28% of incident cancer and 20% of cancer deaths in women. A powerful marker in breast cancer detection is the breast cancer grading. Among the standard grading systems used all over the world, Nottingham Grading System (NGS) is preferred for the reason of providing more objective criteria for the three component elements of grading and specifically addresses mitosis counting in a more rigorous fashion. The three component NGS criteria are briefed below (see Figure 1):

- Tubule Formation (TF) - are referred as white blobs (lumina) surrounded by a continuous string of cell nuclei.
- Mitoses represent diving cells; the Mitosis Count (MC) score is assessed in the peripheral areas of the neoplasm (poor tubule formation areas) and it's based on the number of mitoses per 10 High Power Field's (HPF's).
- Nuclear Pleomorphism Score (NPS) - categorizes cells nuclei based on two features: size and shape.

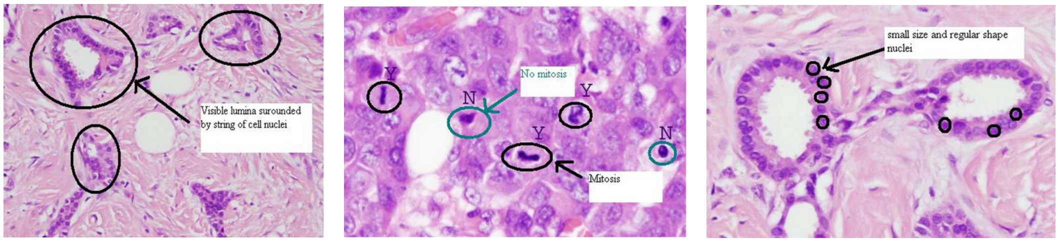

Visible lumina surounded by string of cell nuclei
Y
N
No mitosis
Y
Y
N
Mitosis
small size and regular shape nuclei

Figure 1. NGS synthesis a) TF with grade 1 b) Mitosis differentiation - the black arrow indicates mitosis; the green arrow indicates non-dividing cells c) Small size and regular shape nuclei- NPS grade 1

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

3. Breast Cancer Grading Rules-Based System Modelling

The purpose of this section is to introduce the approach used to step from the 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.

The modeling demarche has been leaded according to the Ontology Web Language (OWL) developed in the Semantic Web Protégé framework [10].

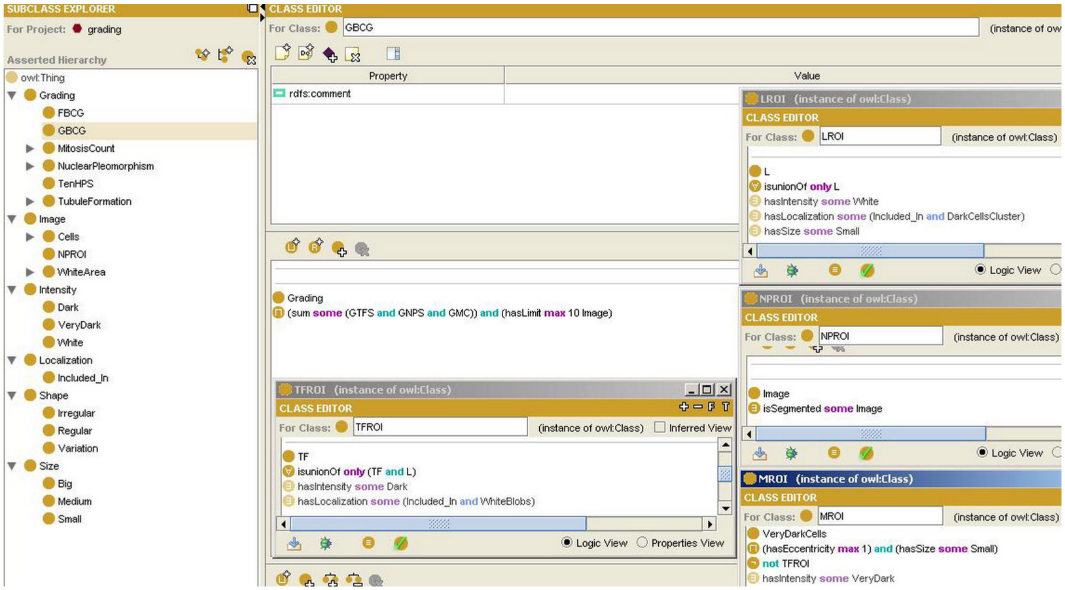
This section is structured in three parts, according to the main steps of the proposed approach: development of the correspondence between medical concepts and computer vision concepts with respect to the OWL standard; definition of intermediate CV rules and generation of the final Symbolic Rules by combination of the CV concepts and intermediate CV symbolic rules.

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 1). Therefore, to clearly define the rules, medical concepts are transformed into computer vision concepts. The CV concepts are then modeled as Protégé concepts, in standard types as classes, properties and symbolic rules (see Figure 2). Various instances for different classes are created as individuals, where specific values are assigned for classes and properties, respectively.

Table 1. CONCEPTS correspondence (Medical – CV Protégé)

Medical Concepts	CV concepts	Protégé concepts type
Slide	Image (digitized)	Super class
Grading	Grading	Super class
Cells	Cells	Class inherited from Image
CellsCluster	Union of Cells	Class inherited from Cells
DarkCellsCluster/ VeryDarkCellsCluster	Union of Cells	Class inherited from Cells with hasIntensity (property) Dark/VeryDark (instances of Intensity class)
Lumina	White compact segments of the Image included in union of dark Cells	Class inherited from WhiteBlobs with 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	Union of Cells/ Dividing Cells nuclei/ dimension & shape features of nuclei	Classes inherited from Grading
Local Grading/Global Grading (10 HPFs)	Grading computation for TF , MC, NP for a single image/ten images	Class inherited from TubuleFormation/ MC/ NPS



SUBCLASS EXPLORER

For Project: grading

Asserted Hierarchy

- owl:Thing
  - Grading
    - FBCG
    - GBCG
    - MitosisCount
    - NuclearPleomorphism
    - TenHPS
    - TubuleFormation
  - Image
    - Cells
    - NPROI
    - WhiteArea
  - Intensity
    - Dark
    - VeryDark
    - White
  - Localization
    - Included_In
  - Shape
    - Irregular
    - Regular
    - Variation
  - Size
    - Big
    - Medium
    - Small

CLASS EDITOR

For Class: GBCG (instance of ow

Property	Value
rdfs:comment	

Grading
(sum some (GTFS and GNPS and GMC)) and (hasLimit max 10 Image)

TFROI (instance of owl:Class)

CLASS EDITOR

For Class: TFROI (instance of owl:Class) □ Inferred View

TF
isunionOf only (TF and L)
hasIntensity some Dark
hasLocalization some (Included_In and WhiteBlobs)

● Logic View ○ Properties View

LROI (instance of owl:Class)

CLASS EDITOR

For Class: LROI (instance of owl:Class)

L
isunionOf only L
hasIntensity some White
hasLocalization some (Included_In and DarkCellsCluster)
hasSize some Small

● Logic View ○

NPROI (instance of owl:Class)

CLASS EDITOR

For Class: NPROI (instance of owl:Class)

Image
isSegmented some Image

● Logic View ○

MROI (instance of owl:Class)

CLASS EDITOR

For Class: MROI (instance of owl:Class)

VeryDarkCells
(hasEccentricity max 1) and (hasSize some Small)
not TFROI
hasIntensity some VeryDark

Figure 2. PROTEGE model of the Breast Cancer Grading computer vision symbolic rules and concepts

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.

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

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

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

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:

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

The TF_{ROI} symbolic rule creates the union of all *DarkCellsCluster* – with intensity and localization dependence and *L*. The *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

$$segment(\mathrm{Im}) = \{size(DarkCellsCluster_i), shape(DarkCellsCluster_i)\}$$

4. 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. 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 3).

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

4.1 Local grading computation

The local grading computation process uses functions and operators to define the required symbolic rules (see Table 2).

Table 2. FUNCTIONS used in criteria score symbolic rules

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

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

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.

4.2 Global grading computation

The global breast cancer grading is computed similar with each local score, but over 10 HPFs [1] (see Figure 3). The 10 HPFs specification appears as the upper limit at each computation of sum in the rules.

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

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

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

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

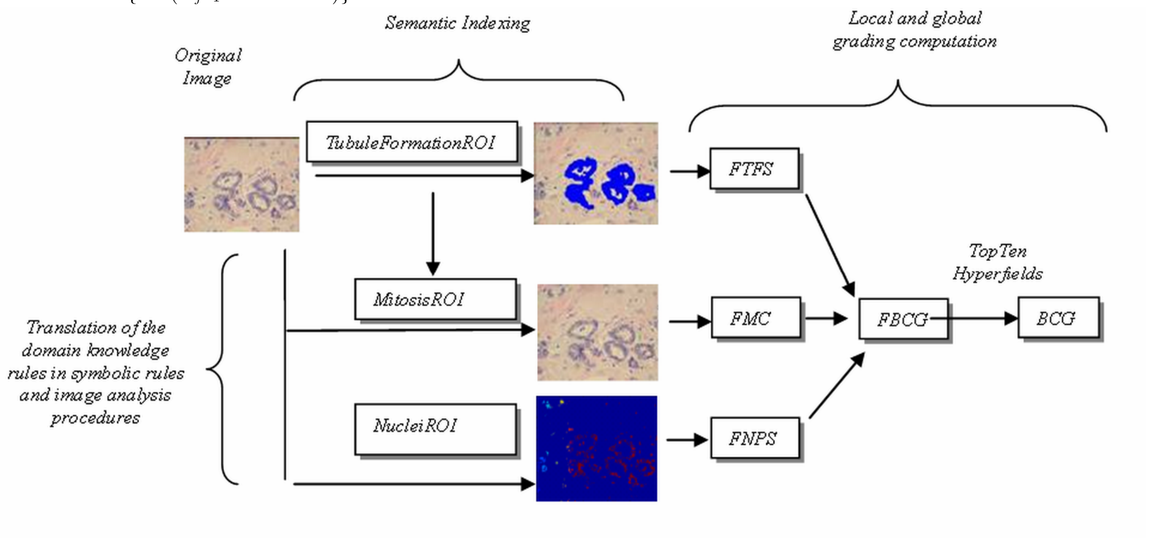

Semantic Indexing
Local and global grading computation
Original Image
TubuleFormationROI
FTFS
TopTen Hyperfields
Translation of the domain knowledge rules in symbolic rules and image analysis procedures
MitosisROI
FMC
FBCG
BCG
NucleiROI
FNPS

Figure 3. SEMANTIC Indexing in BCG Context

5. EXPERIMENTS & RESULTS

The experimental part consist in analyzing and indexing pathologic images of six breast cancer cases, consisting of 7000 frames scanned from the tumor tissue slides obtained through collaboration with the Pathology Department 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, have 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 3. 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 4. 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 5. 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 [11] given by the microscope 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.

6. Discussions, Conclusion and Perspectives

Even if strongly related to a particular application field and specific medical domain, the presented semantic labeling approach has a generic character. Indeed, in association with localization and quantitative information, this meaningful indexing allows to design semantic query content-based medical image retrieval systems, usable in evidence based medicine framework. 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 platforms, automatic semantic-query based visual positioning systems [12] present also a strong interest for the medical technicians and doctors in terms of time efficiency.

Finally, 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.

Acknowledgment

This project is partially supported by the ONCO-MEDIA[i] project. We would like to thank to Pr. Teh Ming, head of the Pathology Department of the National University Hospital from Singapore for his constant support in this joint study.

References

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

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

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

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

[11] 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

[12] 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

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