HISTOPATHOLOGY IMAGE REPRESENTATION FOR AUTOMATIC ANALYSIS: A STATE-OF-THE-ART REVIEW

John Arevalo[3], Angel Cruz-Roa[2], Fabio A. González O[1].

[1]Ph.D., Associate professor, Universidad Nacional de Colombia, Computing systems and industrial engineering dept.

[2]M.Sc. Universidad Nacional de Colombia, Ph.D. Candidate

[3] M.Sc. Universidad Nacional de Colombia, Ph.D. Student

Corresponding: fagonzalezo@unal.edu.co

Recibido: Enero 7 de 2014 Aceptado: Abril 22 de 2014

Abstract

This paper presents a review of the state-of-the-art in histopathology image representation used in automatic image analysis tasks. Automatic analysis of histopathology images is important for building computer-assisted diagnosis tools, automatic image enhancing systems and virtual microscopy systems, among other applications. Histopathology images have a rich mix of visual patterns with particularities that make them difficult to analyze. The paper discusses these particularities, the acquisition process and the challenges found when doing automatic analysis. Second an overview of recent works and methods addressed to deal with visual content representation in different automatic image analysis tasks is presented. Third an overview of applications of image representation methods in several medical domains and tasks is presented. Finally, the paper concludes with current trends of automatic analysis of histopathology images like digital pathology.
Keywords: Histopathology; image analysis, computer-assisted; pattern recognition system; informatics computing, medical; state-of-the-art review

REPRESENTACIÓN DE IMÁGENES DE HISTOPATOLOGÍA UTILIZADA EN TAREAS DE ANÁLISIS AUTOMÁTICO: ESTADO DEL ARTE

Resumen

Este artículo presenta una revisión del estado del arte en la representación de imágenes de histopatología utilizada en tareas de análisis automático. El análisis de imágenes hispatológicas es importante en la construcción de herramientas para el diagnóstico asistido por computador, sistemas de mejoramiento automático de imágenes y sistemas de microscopía virtual, entre otras aplicaciones. Estas imágenes tienen una gran mezcla de patrones visuales con características particulares que hacen de su análisis una tarea difícil. El artículo discute estas particularidades, el proceso de adquisición y los retos particulares al realizar un análisis automático. En la segunda sección se presenta una revisión de trabajos y métodos recientes enfocados a la representación del contenido visual en diferentes tareas de análisis automático. En tercer lugar, se presenta una visión general de las aplicaciones para los métodos de representación en diferentes dominios médicos.

Finalmente el trabajo concluye con las actuales tendencias del análisis automático de imágenes de histopatología como la patología digital.
Palabras clave: Histopatología, análisis de imagen asistida por computador, sistema de reconocimiento de patrones, informática médica, estado del arte

REPRESENTAÇÃO DE IMAGENS HISTOPATOLÓGICAS PELO ANÁLISE AUTOMÁTICO: REVISÃO DO ESTADO DA ARTE

Resumo

Este artigo é uma revisão do estado da arte na representação de imagens histopatológicas utilizadas nas tarefas de análise automáticos. O análise de imagens histopatológicas é importante na construção de ferramentas para o diagnóstico assistido por computador, sistemas de melhoramento automático de imagens e sistemas de microscopia virtual. Essas imagens tem uma grande mistura de padrões visuais com caraterísticas particulares, que fazem do análise uma tarefa difícil. O artigo discute essas particularidades, o processo de aquisição, e os desafios particulares no momento de realizar uma análise automático. Na segunda seção se apresenta uma revisão dos trabalhos e métodos recentes, com foco à representação do conteúdo visual em diferentes tarefas de análise automático. Na terceira, se apresenta uma visão geral das aplicações para os métodos de representação em diferentes domínios médicos. Finalmente, o artigo conclui com as atuais tendências do análise automático de imagens histopatológicas como a patologia digital.
Palavras-chave: Histopatología, análise de imagens assistido pelo computador, sistema de padrão de reconhecimento, informática medica, revisão do estado da arte.

Introduction

Histology is an important area of biology which studies the cell anatomy and tissues of animals and plants in a microscopic level. Histology and histopathology images are very important for diagnosis purposes; these images are a fundamental resource to determine the state of a particular biological structure, to support diagnosis of diseases like cancer, or to analyze anatomy of cells and tissues. Medical doctors and biologists are trained in histology to interpret the appearance of tissues according with their structure, functionality and cellular organization at different organs, those features can be highlighted using several staining processes. This kind of images are used for both biological research and making clinical decisions, and commonly are used like ground truth for other studies such as x-ray and MRI [1], [2].

Automatic analysis of this kind of images is a relevant and prolific research area [3], [4], however, due to their visual richness and complex variety of structures, these images require specialized analysis depending on the organ and particular task. For example, Naghdy et al [5] considered a texture analysis using Gabor Filter to find relevant regions based on the fact that pathologists look at the magnified images of cervix biopsy, and grade cervical cancers based on the spread of abnormal cells into the epithelium layers. Colon cancer was analyzed in [6] by defining a homogeneity measure based on the texture of different parts of the tissue, which are characterized with the spatial organizations of its cellular and connective tissue components. It is noteworthy these organizations show differences, depending on the organ. To address prostate cancer diagnosis, Monaco et al [7] set gland's area as a feature, knowing it has discriminant capability for benign and malignant glands. A hierarchical algorithm was presented to detect cancerous regions at lower resolutions, then refines Gleason grades in higher resolutions. Doyle et al [8] presented a model for automatic grading of breast cancer, using combination of texture and topological features, results not only showed a discriminative capability, but also gained an interpretable model based on graph theory.

Thereby, each problem (medical domain and application) has a different and specialized method that cannot necessarily be extended to other contexts. In addition to this, histology image representation poses important challenges because of their high visual variability due to different acquisition processes, anatomical variability, staining, image magnification and the type of cut [9].

Nowadays, automatic methods for image representation and analysis have been successfully applied in medical imaging. For example, there are several computational methods which take advantage of the increasing assortment of public and large biomedical image databases [10]. In fact, new research areas like digital pathology [1] and bioimage informatics [11] are emerging. Digital pathology is an image-based environment focused on the research of histology images analysis based on pattern recognition and experimental workflow [1], whereas bioimage informatics comprises image processing, data mining and database visualization, extraction, searching, comparison and management of biomedical knowledge inside massive image collections [11].

Image representation is the first key stage in histopathology image analysis workflows. The main goal of this paper is to give an overview of current techniques used to describe the visual content of histopathology images and their applications. Firstly, the nature of histopathology images, their characterization and acquisition process is detailed in Section 2, this Section also discusses the main challenges posed by the automatic analysis of this kind of images from a computational point of view. A review of the state of the art on visual content representation of histopathology images is described in Section 3. Different medical applications are presented in Section 4. Finally, current trends and conclusions are discussed in Section 5.

Histopathology images

Digital images are pixels matrices with intensities of each color channel. This is the lowest-level representation which does not provide semantic information by itself. Direct inspection of this low-level information alone does not provide an evidence of, for instance, a particular tissue being a tumor or how many nuclei cells there are present in the image. The fact that the high-level semantic information is not immediately apparent from the low-level pixel data is known as the *semantic gap* [12]. In histopathology images, such gap is particularly important mainly because of the high visual variability resulting from the diversity of tissues and structures and the particularities of the acquisition process.

Characteristics of Histopathology Images

Histopathology images have normal and abnormal biological structures, morphological and architectural characteristics which could be identified by pathologists depending on their experience, but some structures are small with respect to tissue region, and relevant patterns generally have high visual appearance variability. Most of visual variability is inherent of biological structures and anatomy. In addition, acquisition process adds noise and visual variability on each stage. In order to visualize this process, it is useful to review the histopathology image acquisition process depicted in Fig. 1. First, the biological sample is taken from an organ. Then, a fixation process is done over the biopsy to assure chemical stability on the tissue and avoid post-mortem changes [13]. After this, it must be cut into sections that can be placed onto glass slides. The sections are stained to reveal cellular components by chemical reactions. The most common dyes used are Hematoxylin - Eosin (H&E) which stain cell nuclei in a dark blue or purple and cytoplasm and connective tissue in a bright pink. Finally, the section is cover slipped to be viewed and digitized with a microscope.

The particular process used for fixation and staining produces different visual effects in the acquired images. Additionally, acquired images are affected by luminance and other factors associated to the acquisition technology used. Taking into account that these images are a 2D projection of a 3D object (organs), the biological sample has a different appearance depending on its orientation (e.g. longitudinal, oblique or cross-sectional [14]). Depending on the region of interest, the laboratory personnel could take different images at different magnification. Usually, the image resolution is related to the size or scale of biological structures that can be seen in the histology slide. These human factors add a new source of variability. Main issues that contribute with visual heterogeneity of images are depicted in Fig. 2, such as image magnification, type of cut, luminance and stain concentration.

Magnification: Magnification refers to increasing the proportion of biological structures which are visible under the microscope according to the set of lenses. Conventional microscopes have a standard set of objectives 2X, 10X, 20X, 40X and 100X. First row in Fig. 2 shows the appearance of tongue muscle tissue stained with Masson's

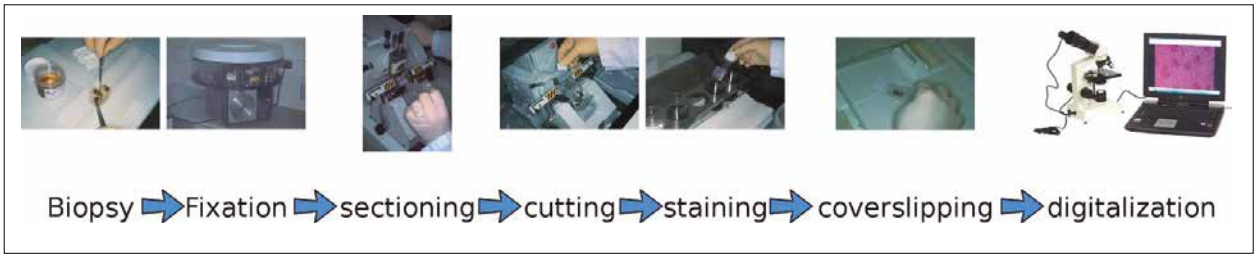

Fig. 1. *Acquisition workflow diagram*[1]

trichrome staining at 10X, 20X and 40X. It is clear that, even being the same organ, appearance of those images is highly variant identifying different structures at different magnifications.

Staining: It is an auxiliary technique in microscopy to improve the contrast in a biological sample seen under a microscope. Dyes are used in both, biology and medicine, to highlight biological structures from different tissues according to biomedical or diagnosis interest. There are several types of dyes according to chemistry properties of biological structures which must be spotlighted [15].

Depending on the region or biological structure of interest, some types of stains are more appropriated than others. So, it is possible to get different images and visual appearances from one biopsy depending on the type of stain. The second row in Fig. 2 shows region between dermis and epidermis at 40X magnification stained with H&E and Masson's trichrome respectively. It is noteworthy

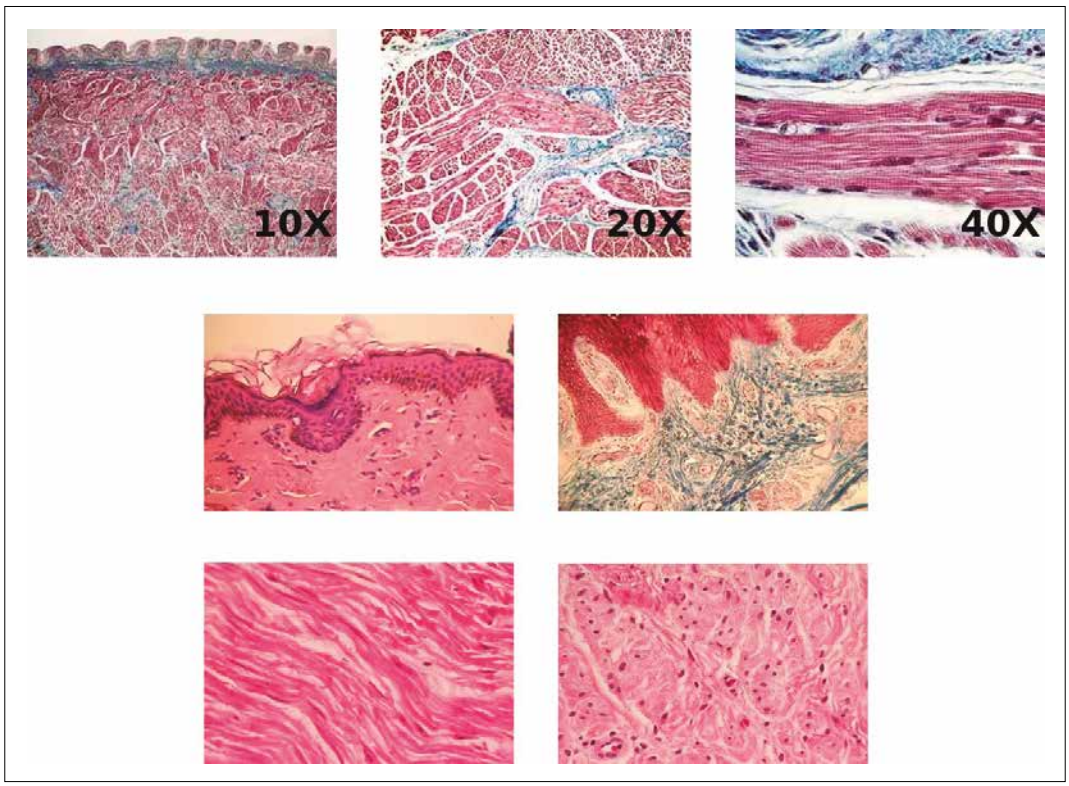

Fig. 2. *Visual variability in histology images*[2]*. First row shows variability due to magnification of same tongue muscle tissue. Second row shows variability due to staining and luminance. Third row shows variability of smooth muscle tissue due to section orientation (longitudinal and cross-section)*

[1] Adapted from http://library.med.utah.edu/WebPath/HISTHTML/HISTOTCH/HISTOTCH.html
[2] Partly taken from http://www.pathguy.com/histo/053.htm

how the performing of staining process affects in different way the luminance and the intensity of color channels. This behavior in histology images adds conceptual visual variability determined by the type of stain applied.

Slice orientation: The tissue appearance in histology images is related with the slice cut orientation. Last row in Fig. 2 shows how the tissue appearance changes when the cut is done in a longitudinal or cross-sectional orientation in smooth muscle tissue using the same staining (H&E). There are infinite possible cut orientations, which makes harder the characterization of a 3D organ by a thin 2D sample. Although there are standard protocols for slice cutting and histology image acquisition depending of organ, the final precision of the results depends on several and accumulative factors like the type of microtome used to take the slice, the composition of the sample, the experience of histotechnologist, the fixative used to harden the tissue, among others.

Histopathology image representation

Automatic Histopathology Image Analysis Process

A typical automatic histopathology image workflow is depicted in Fig. 3. The first step is image preprocessing in which raw image data is transformed in order to reduce visual variability and noise, as well as making it more suitable for the subsequent steps. Common tasks in this phase include pixels intensity normalization to deal with luminance artifacts, image scaling to reduce representation size and dimensionality reduction using techniques such as Principal Component Analysis (PCA).

The second step is feature extraction, whose purpose is to produce a more descriptive representation of the image making explicit important information which is not directly manifest from the raw pixels. There are two main types of features: region-based (local) descriptors and image-based (global) descriptors. Image-based descriptors try to describe the image as a single entity, while region-based methods assume image is composed of independent building blocks and, therefore the image can be modeled as a combination of such blocks. Sometimes, visual features in a region-based scheme are extracted from the most relevant parts of the image, commonly known as Regions of Interest (ROI), this can be achieved through segmentation techniques where background is removed to highlight relevant objects, or by applying techniques like Scale-invariant feature transform (SIFT) [16] to detect interesting points in the object. Either global or part-based, Feature extraction process calculates descriptors to describe in a concise way the content of the image.

In the third step interesting visual patterns are detected and identified, this is known as pattern recognition. This is usually accomplished using supervised machine learning methods. These methods learn a discriminative/classification model from annotated training data. The learned model is later used to classify new unseen images [17].

The performance of pattern recognition models is highly dependent on the features used; therefore picking the right features for a particular problem is the main motivation

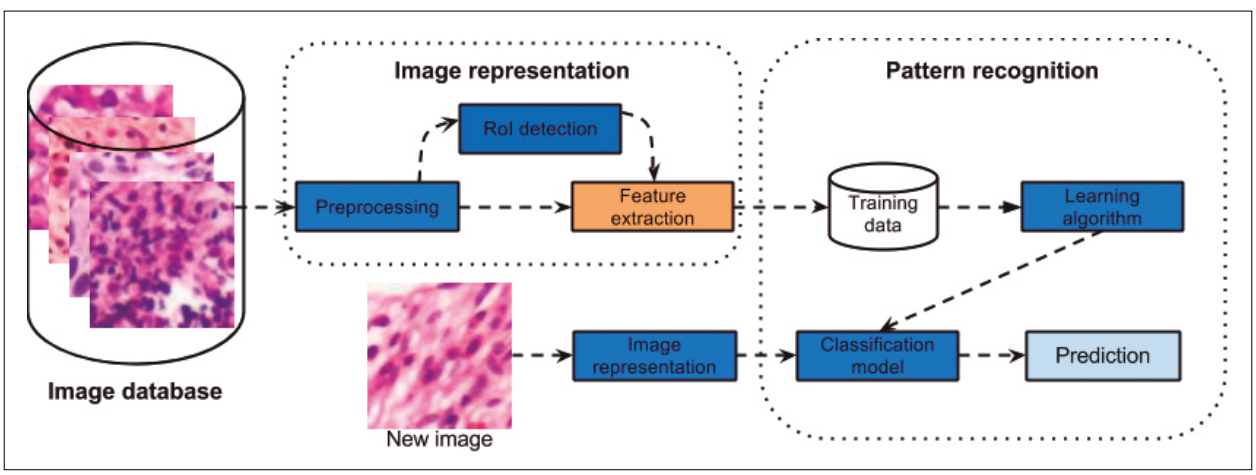

Fig. 3- *Typical histopathology automatic image analysis workflow.*

to provide a review of current representation techniques applied in histopathology domain.

Supervised machine learning algorithms applied to automatic image categorization require representative training sets of images. Because of the high histopathology image visual variability, a supervised learning method usually needs a large number of images from different patients to make successful generalizations [18]. In addition to visual variability, complex structures and patterns are presented in histology images. Then, different approaches have been applied to characterize and represent the visual information on histopathology images and solve particular problems (e.g. cancer detection and grading).

Visual features

A common characteristic present in many of the reviewed works is the fact that particular image analysis problems demand specific image representation schemes. The main reason is that different organs express different complex biological structures, increasing visual variability and making harder to close the semantic gap. Every approach needs to describe the image content within some data structure that captures its visual content and biological structures configuration. Different methods for visual features extraction have been proposed. An overview of the most common feature extraction schemes used for histopathology image representation in different works is listed in Table 1.

Visual features aim to describe the most relevant information to feed a specific Machine Learning (ML) algorithm, depending on the task and the type of the images, some features can be more efficiently than others. Visual content of images could be typically represented by using either all the pixels in the image for a unique representation (Image-based descriptors) or splitting the image and extract a set of features per each segment (region-based descriptors). Independently of ROI to be analyzed (i.e. whole image or segments), in histopathology image domain, feature extraction methods are usually classified as follows [3], [19].

Intensity features: This kind of features provides information of the gray level or color of pixels located in the ROI. This feature extraction approach uses different color spaces. Samsi et al [20], [21] used the Hue channel from the HSV (Hue-Saturation-Value) color space conversion of the original image, while Oszdemir et al [18] worked on white/pink/purple color dimension. In [22], 11 color models were evaluated to compare the incidence over the performance in a classification task, they found that there is no single model which works better than others in every case. This can be supported in the results reported on [23], [24], where the first one reported RGB color space performed better than others, while second one ensures adding CIELab color information may improve classification capability.

Morphological features: These features provide information about the size and shape of the described region, object or image. In [7] gland area was used as discriminative criteria to classify between benign or malignant, while Dundar et. al.[25] used perimeter as feature descriptor to characterize cell size in segmentation process. Taking advantage of pathologist's experience in diagnosis of Oral Submucous Fibrosis (OSF), Muthu et al [26] represented the image using morphological features like eccentricity,

Table 1. *Overview of feature extraction schemes applied on histology domain.*

Feature type	Detail	Applications
Intensities	Grayscale	[35]
	RGB	[18], [22], [35], [43], [47], [54], [67]
	HSV	[20–22], [35]
	CIEL*a*b	[22], [24], [67]
Morphological	Area, perimeter, shape, etc.	[7], [25–29], [68–72]
Topological	Voronoi Diagram, Delaunay Triangulation	[8], [30], [31], [69]
	Nearest Neighbor	[30], [31]
	Minimum Spanning Tree	[8], [30], [31], [69]
Texture	Haralick	[8], [20], [21], [26], [33–35]
	Gabor Filter	[5], [8], [45]
	Co-occurrence texture	[24], [52], [73]
	Haar wavelet coefficients	[41], [49]
	LBP	[44], [45]
	Bag of features	[36–43], [74]

perimeter, area equivalent diameter to describe cell nucleus. Not only classification tasks can be performed using this type of features, Dangott et al [27] built an automated system for differential white blood cell (WBC) counting based on 19 features such as area, perimeter, convex area, solidity, orientation and eccentricity. Also, these features have been used to construct Content-Based Image Retrieval (CBIR) to find prostate histopathology images based on morphological similarity [28]. As well as [29] and some of mentioned works, most of the morphological measures estimation are done based on a previous segmentation, and therefore its performance depends on the precision of such segmentation.

Topological features: Topological features provide information about some structure within the image. This description is particularly useful on the segmentation task, where graph theory is commonly the most used in this group of features. A graph is created with a set of points of interest (e.g. nuclei detection), a Voronoi Diagram (VD) uses this set to build a diagram by divide the region with approximately the same area with graph's vertices as center for each region. Delaunay Triangulation (DT) is built by fitting a set of triangles over this graph in such way that each circumcircle does not contain any vertex inside of them. Area, perimeter, angle and other geometrical features are calculated from this set of regions/triangles to describe visual content of such regions. Nearest Neighbor uses information like distance, density, closeness of N nearest neighbor as descriptor for each vertex in the graph. Lastly, Minimum Spanning Tree (MST) finds the shortest path in a graph, length of each edge can be used as descriptor. Basavanhally [30]concatenated up to 50 features from DT VD MST and nearest neighbor to train a classifier to distinguish low and high grades in breast cancer histopathology images. Madabhushi [31] proposed a method that extract information from multimodal information, including magnetic resonance imaging (MRI), digital pathology and protein expression combining DT, VD, MST and NN descriptors by graph embedding method to characterize spatial arrangement of nuclear structures and, in conjunction with a SVM classifier, distinguish samples with different levels of Lymphocytic infiltration in breast cancer. In [8] a model was proposed to grade cancer in breast images by feature combination of VD and DT using graph embedding. In [27] DT also was used to find possible edges that might correspond to the boundaries between segmented nucleus.

Texture features: These features provide information about the variation of intensities presented in the region, helping to tissue identification. Haralick features [32] suggests a set of 28 textural features defined by several equations, some of those related with statistical properties such like correlations, means, variances among others. Depending on the domain application, only some of those features can be used. Kuse et al [33] extracted 18 Haralick texture features to classify lymphocytes and non-lymphocytes, Chaddad [34] used 5 main Haralick's coefficients from Gray Level Co-Ocurrence (GLCM) Matrix for texture analysis and colon cancer cell detection. Cinar et al [35] extracted 7 Haralick features from normalized co-ocurrence matrix for grayscale, RGB and HSV color spaces to indexing content in a CBIR system. In [8] calculated 6 Haralick features were calculated to combine them with topological and color features to grade breast cancer.

Bag of features is an adaptive approach to model image structures using a dictionary learned from images patches. Each patch can be represented through different descriptors, image representation is built with a frequency histogram where each bin shows how related is the image with each visual word of the dictionary. This model has obtained success results for basal-cell carcinoma detection [36–40], medullobastoma [41] and renal cell carcinoma [42] classification. This kind of representation was used to build a CBIR system using Non Negative Matrix Factorization (NMF) by Vanegas et al [43].

Local Binary Pattern (LBP) is a texture descriptor that labels the pixels by thresholding the neighborhood against current pixel, set 1 if neighbor is above it, 0 otherwise, then these values are concatenated following a circle around evaluated pixel to obtain 8 binary digits, to finally convert them in a 10-base number. This operator has been applied in histopathology images to identify oral cancer [44], [45], as well as a descriptor for image indexing for retrieval systems [43].

Feature composition: To improve the performance of the applied methodology, several authors have built more complex features by combination, or selection depending on statistical or another analysis [4]. In [35] color and texture features were extracted and normalized with null mean and unit variance, and then Non-negative Matrix Factorization (NMF) model was used to combine such features by dimensionality reduction. In [26] five biological characteristics were suggested by experienced oncopathologists: nuclear changes, polymorphism, nuclear irregularity, hyperchromasia and nuclear texture, then these features were extracted based on that type of characteristics. An unsupervised feature extraction was applied to remove any bias towards certain features which might afterwards

affect the classification procedure. Also, in [44], three texture features were extracted: (LBP), Higher order spectra (HOS) and Laws Texture Energy (LTE), then applied Analysis of variance (ANOVA) to extract features prior to classification to verify discriminating capability.

The above features, summarized in Fig. 4, are much more complex than others reported on this document, and all of them require a prior design and parameters have to be fitted according with the objective and domain. Previous classification demonstrates efforts have been oriented to build specific design for each domain, even for each task. However this representation could not be scalable for other problems.

Histopathology Image Classification for Diagnosis and Grading Support

One of the most common applications for automatic histopathology image analysis is detection and grading of cancer. Main problems where automatic histopathology image analysis has been applied are described as follows.

Prostate cancer

In Colombia, prostate cancer is the most common cancer in men and it is also one of the leading causes of death by cancer (more than 8000 new cases and 2400 deaths per year [46]). This type of cancer is graded from 1 to 5 using Gleason grading method, and it is based on structural features of the tissue, where Grade 1 has well defined patterns, while 5 grade patterns are difficult to differentiate. Automatic grading can be performed in two stages, a segmentation process to extract glands from the background and to get morphological measures from them, and then train a classifier to recognize their patterns and perform annotation of each grade.

[47] applied PCA in RGB color decomposition to remove correlations between channels, then k-means clustering algorithm is applied with resultant components for k = 4 to identify lumen, nuclei, cytoplasm and stroma with expert guidance, finally, a region growing method was applied to complete the segmentation. After that, morphological features of segmented regions are used as input for an LDA classifier [7] convolved the image with a Gaussian Kernel, using peaks as seeds for a region growing procedure. Glands area are estimated and used jointly with a MRF (Markov Random Field) to train a Bayesian estimator and classify glands as either malignant or benign. [29], [48] performed a semi-automatic gland segmentation using low-level, high-level and specific domain information. Then morphological and topological features were extracted from regions and used to train a binary Support Vector Machine (SVM) model, for classifying between 3 and 4 grades. Another approach using extracted features directly from image into a classification model has been applied

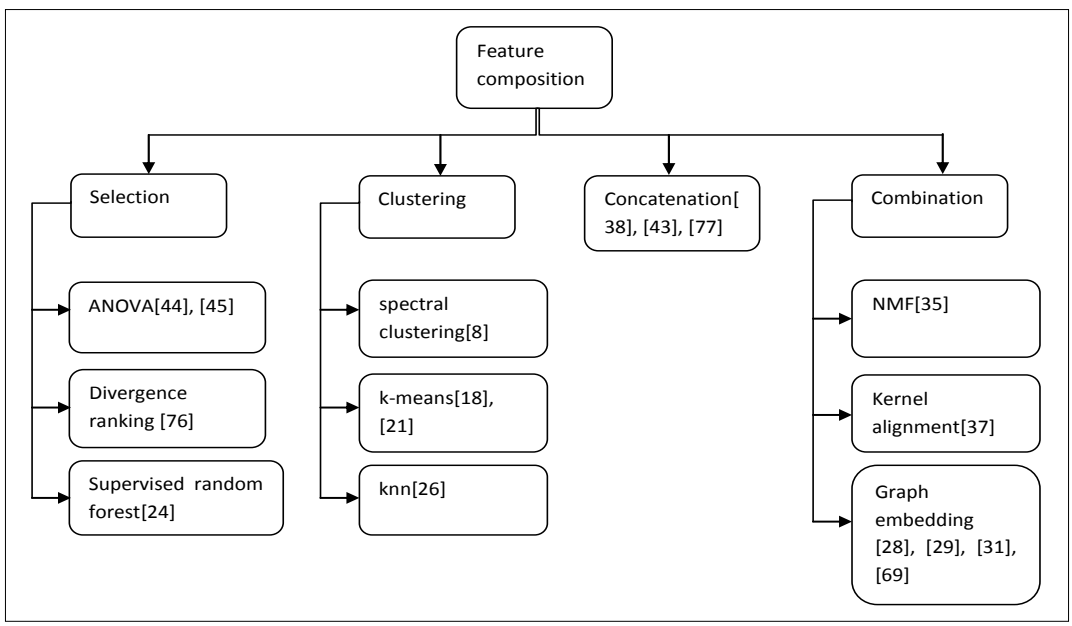

Fig. 4. *Overview of common approaches for feature composition in histopathology image representation.*

in this domain. This approach commonly uses texture features like co-occurrence matrix [24], Haar wavelets [49] and text on histogram [50].

Cervix cancer

Cervix cancer is the most prevalent cancer in Colombia, 17.5% of new cases in women each year, and 97% of these new cases were diagnosed with a primary tumor histology [51]. Cervix cancer refers to cancer forming tissues of the uterine cervix, and is graded into three categories: cervical intraepithelial neoplasia (CIN) 1, CIN 2 and CIN 3 corresponding to mild, moderate and severe dysplasia.

In [52] segmentation was performed with a supervised model using texture features and an SVM classifier at different resolutions, after that, morphological measures of nuclei are used to train a multiclass SVM model with RBF Kernel and to grade cervix cancer [5] proposed a gabor-filter-based segmentation with a supervised learning approach, resulting a pixel-by-pixel classification into four classes: basal, stroma, normal and abnormal cell. Then, clustering and median filters are applied to count normal and abnormal nuclei. Finally, ratio between normal and abnormal is used as discriminant feature to grade the image. On the other hand, a standalone segmentation was proposed in [53] to identify cervical nuclei based on HSV color and morphological features, while [54] applied GMM (Gaussian Mixture Models) and grayscale features.

Colon cancer

It is the third most common cancer in the world for both women and men, in Colombia it is the fourth for both. Also, it is the fifth in Colombia leading causes of death for cancer. Assessment of cancer grading is based on visual abnormalities found by pathologists; this grade is subjective because it depends on interpretation given by the expert. Automatic methods have been proposed to detect and grade colon cancer.

[18] presented a strategy to classify samples as normal, low-grade and high-grade cancerous, representing images with features extracted from windows centered in random sampling points, and a k-means clustering is performed to get template sequences. Training images are represented with these templates sequences, then a markovian model is trained with such representations. Experiments showed the proposed model performs better than raw features, even when there was less data to train. A segmentation approach was shown in [34] using five measures of Haralick texture features, to detect three types of cells: carcinoma, benign and intraepithelial neoplasia, using a neural network as classifier obtaining at the end a computational simplicity of segmentation which is performed in a very short time [6] presented an unsupervised object-oriented segmentation algorithm based on texture features. K-means clustering is performed on the color intensities to detect and define objects referred to connective tissues, luminal structures and epithelial cell components. Two homogeneity measures were defined, object size uniformity and object spatial distribution uniformity, as textures features for each object, finally a region-growing algorithm is carried out to completed segmentation process. The proposed method was compared against JSEG algorithm, a pixel-oriented approach [55], obtaining better performance in terms of specificity and sensitivity.

Basal cell carcinoma

Basal-cell carcinoma (BCC) is the most common skin disease in white populations, its incidence is growing worldwide [56] and it represents more than 59% of skin cancer cases in Colombia. It has different risk factors and its development is mainly due to ultraviolet radiation exposure. Pathologists confirm whether or not this disease is present after a biopsied tissue is evaluated under microscope. In this evaluation, physicians aim to recognize some characteristic patterns or complex mixes of patterns. This process is called differential diagnosis and it is mainly achieved by visual analysis. In [57], the structural patterns that characterize the basal-cell carcinoma are described and correspond to 11 different complex patterns.

[36] published one of the first models to diagnose BCC disease using BOF representation with two texture features: raw block and SIFT descriptor. This kind of representation can be extended to analyze semantic concepts by identifying visual patterns in BCC images [38], [40]. In order to support medical searching within BCC image collections, [58] proposed a method to ease its visualization and exploration using BOF with texture features. Looking for a more detailed diagnosis, [39] proposed an annotation model with probabilistic Latent Semantic Analysis (pLSA) and BOF representation as input features. [37] defined a kernel functions combination to build a semantic annotation framework, then such annotations were used to build a CBIR. Results showed an average improvement of 57% when compared to visual search based on low-level features and histogram intersection kernel as similarity measure. Other methods for automatic annotation for BCC collections has been proposed, [59], [60] applied NMF with BOF to build a latent

topic model with probabilistic support as main advantage for interpretability of results. A proper determination of Regions of interest (ROI) would allow to concentrate any processing effort on specific image areas to diagnose BCC disease, to find such RoIs [61] proposed a supervised model inspired by visual cortex areas of the brain and the way their perceive the world. Diaz and Romero [62] have proposed a microstructural tissue analysis in basal cell carcinoma images using a stain correction of H&E images and an automatic method to identify morphological and architectural features by square regions in images based on latent semantic analysis and support vector machines.

Open trends problems and challenges

The most important challenge is that unlike natural images where high-level semantic concepts are related with connected areas and objects, in histopathology images high-level semantic interpretations are related to pathological lesions, morphological and architectural features, structural configuration, cells and biological patterns organization (context), and not only the presence of patterns is important but also their absence is. Such factors depend on the amount of magnification and the type of organ, which encompass a complex mixture of visual patterns that allow deciding about the illness presence. Dealing automatically with these factors is still an open problem [59].

On the other hand, thanks to recent advances in microscopical acquisition technology (scanners and robotic microscopes) it has been possible to collect huge numbers of histopathology images and make them publicly available through publicly accessible image databases [63].Table 2 lists different open histopathology image databases accessible through the web. This trend is very positive for the progress of digital histopathology research, since it allows objective comparison of methods and strategies. However, standard evaluation protocols are required. An example effort in this direction is the publication of the MITOS dataset, by the International Conference on Pattern Recognition (ICPR), along with a set of evaluation metrics to encourage participation on their contest where the objective was to evaluate methods to automatically determine the mitotic count.

Nowadays one of the new challenges is to deal the growing size of these image collections (from hundred thousands to millions) and whole-slide-images sizes (20 GB or more) [64]. This phenomena is known as Big Data, and in pathology domain is opening new challenges, opportunities and approaches which are being addressed by a new area called *Digital Pathology* [65]. Finally, It should be noticed most of these Finally, It should be noticed most of these methods and frameworks reviewed in this work are addressed to support and to empower the diagnosis of pathologists rather than replace them [66].

Table 2. *List of public access histopathology image collections*

Dataset	Size	Web page	Source
HistologyDS[38]	2828 Images[1]	http://www.informed.unal.edu.co/histologyDS	Universidad Nacional de Colombia
The Cancer Genome Atlas	89464 files[2]	http://cancergenome.nih.gov/	National Cancer Institute (U.S.)
Stanford Tissue Microarray Database (TMA)	55874 images[3]	http://tma.im/cgi-bin/home.pl	Stanford University
MITOS	50 Images[4]	http://ipal.cnrs.fr/ICPR2012/	The Ohio State University
caIMAGE	910 Images	http://emice.nci.nih.gov/caimage	National Cancer Institute (U.S.)
Atlas of breast histopathology [75]	>150 Images.	http://www.webmicroscope.net/atlases/atlas_getting_started.asp	University of Tampere, Finland
Slide Library of Virtual pathology	5051 WSI Images[5]	http://www.virtualpathology.leeds.ac.uk/index.php	University of Leeds, England

1 http://www.informed.unal.edu.co/histologyDS is a subset of BiMed(http://www.informed.unal.edu.co/)
2 https://tcga-data.nci.nih.gov/datareports/statsDashboard.htm
3 http://tma.im/cgi-bin/viewStain.pl?op=antibody
4 http://ipal.cnrs.fr/ICPR2012/?q=node/5
5 http://www.virtualpathology.leeds.ac.uk/slidelibrary/index.php

Conflict of interest: The authors declare that they have no conflict of interest

Funding: This work was partially funded by projects "Multimodal Image Retrieval to Support Medical Case-Based Scientific Literature Search", ID R1212LAC006 by Microsoft Research LACCIR, "Diseño e implementación de un sistema de cómputo sobre recursos heterogéneos para la identificación de estructuras atmosféricas en predicción climatológica" Number 1225-569-34920 through Colciencias contract number 0213-2013 and "Convocatoria del programa nacional de proyectos para el fortalecimiento de la investigación, la creación y la innovación en posgrados de la Universidad Nacional de Colombia 2013 - 2015" with proposal number 18722. Cruz-Roa also thanks Colciencias for its support through a doctoral grant in call 528/2011. Arevalo also thanks Colciencias for its support through a doctoral grant in call 617/2013.

References

[1] Fuchs TJ, Buhmann JM. Computational pathology: challenges and promises for tissue analysis. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society. 2011;35(7–8): 515–30

[2] Hipp JD, Smith SC, Sica J, Lucas D, Hipp JA, Kunju LP, et al. Tryggo: Old norse for truth: The real truth about ground truth: New insights into the challenges of generating ground truth maps for WSI CAD algorithm evaluation. Journal of pathology informatics. 2012;3(1): 8

[3] He L, Rodney Long L, Antani S, Thoma GR. Histology image analysis for carcinoma detection and grading. Computer Methods and Programs in Biomedicine. 2012

[4] Gurcan M, Boucheron L, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: a review. IEEE reviews in biomedical engineering. 2009;2: 147–71

[5] Naghdy G, Ros M, Todd C, Norahmawati E. Cervical Cancer Classification Using Gabor Filters in 2011. IEEE First International Conference on Healthcare Informatics. Imaging and Systems Biology. 2011

[6] Tosun AB, Kandemir M, Sokmensuer C, Gunduz-Demir C. Object-oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection. Pattern Recognition. 2009;42(6): 1104–1112

[7] Monaco JP, Tomaszewski JE, Feldman MD, Hagemann I, Moradi M, Mousavi P, et al. High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models. Medical Image Analysis. 2010;14(4): 617–629

[8] Doyle S, Agner S, Madabhushi A, Feldman M, Tomaszewski J. Automated grading of breast cancer histopathology using spectral clusteringwith textural and architectural image features. In 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro. 2008. p. 496–499.

[9] Le Bozec C, Jaulent M, Zapletal E, Heudes D, Degoulet P. A visual coding system in histopathology and its consensual acquisition. Proceedings of the AMIA Symposium. 1999. p. 306.

[10] Cooper L. High performance image analysis for large histological datasets. 2009.

[11] Basu S. Some upcomming Challenges in Bioimage Informatics. 2012.

[12] Kalfoglou Y, Dasmahapatra S, Dupplaw D, Hu B, Lewis P, Shadbolt N. Living with the semantic gap: Experiences and remedies in the context of medical imaging. 2006.

[13] Hewitson T, Darby IA. Histology Protocols., vol. 611. Totowa, NJ: Humana Press; 2010

[14] Díaz G. Semantic Information Extraction from Microscopy Medical Images. National University of Colombia. 2010.

[15] Kiernan J. Histological and Histochemical Methods: Theory and Practice. 4th ed. Cold Spring Harbor Laboratory Press; 2008. p. 606.

[16] Lowe DG. Object recognition from local scale-invariant features. In Computer vision, 1999. The proceedings of the seventh IEEE international conference. 1999;2: 1150–1157.

[17] Alpaydin E. Introduction to Machine Learning. 2nd ed. The MIT Press; 2010.

[18] Ozdemir E, Sokmensuer C, Gunduz-Demir C. A resampling-based Markovian model for automated colon cancer diagnosis. IEEE transactions on bio-medical engineering. 2012; 59(1): 281–9

[19] Demir C, Yener B. Automated cancer diagnosis based on histopathological images: a systematic survey. Rensselaer Polytechnic Institute. 2005

[20] Samsi S, Lozanski G, Shanarah A, Krishanmurthy AK, Gurcan MN. Detection of Follicles From IHC-Stained Slides of Follicular Lymphoma Using Iterative Watershed. IEEE Transactions on Biomedical Engineering. 2010; 57(10): 2609–2612

[21] Samsi S, Krishnamurthy AK, Gurcan MN. An efficient computational framework for the analysis of whole slide images: Application to follicular lymphoma immunohistochemistry. Journal of Computational Science. 2012

[22] Mete M, Topaloglu U. Statistical comparison of color model-classifier pairs in hematoxylin and eosin stained histological images. En 2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology; 2009. p. 284–291.

[23] Mahmoud-Ghoneim D. Optimizing automated characterization of liver fibrosis histological images by investigating color spaces at different resolutions. Theoretical biology & medical modelling. 2011;8: 25

[24] DiFranco MD, O'Hurley G, Kay EW, Watson R, Cunningham P. Ensemble based system for whole-slide prostate cancer probability mapping using color texture features. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society. 2010;35(7-8): 629–45

[25] Dundar M, Badve S, Bilgin G, Raykar V, Jain R, Sertel O, et al. Computerized classification of intraductal breast lesions using histopathological images. IEEE Transactions on Biomedical Engineering. 2011;58(7): 1977–1984

[26] Muthu Rama Krishnan M, Chakraborty C, Paul R, Ray AK. Hybrid segmentation, characterization and classification of basal cell nuclei from histopathological images of normal oral mucosa and oral submucous fibrosis. Expert Systems with Applications. 2012; 39(1): 1062–1077

[27] Dangott B, Salama , Ramesh N, Tasdizen T. Isolation and two-step classification of normal white blood cells in peripheral blood smears. Journal of Pathology Informatics. 2012;3(1): 13, 2012.

[28] Sparks R, Madabhushi A, Sparks MA. Content-based image retrieval utilizing explicit shape descriptors: applications to breast MRI and prostate histopathology. In Progress in Biomedical Optics and Imaging - Proceedings of SPIE;2011. p. 79621I–79621I–13.

[29] Naik S, Doyle S, Madabhushi A, Tomaszewski J, Feldman M. Automated Gland Segmentation and Gleason Grading of Prostate Histology by Integrating Low-, High-level and Domain Specific Information. Workshop on Microscopic Image Analysis with Applications in Biology; 2007.

[30] Basavanhally A, Ganesan S, Shih N, Mies C, Feldman M, Tomaszewski J, et al. A boosted classifier for integrating multiple fields of view: Breast cancer grading in histopathology. Proceedings - International Symposium on Biomedical Imaging; 2011. p. 125–128.

[31] Madabhushi A, Agner S, Basavanhally A, Doyle S, Lee G. Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society. 2011;35(7): 506–14

[32] Haralick R, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics. 1973;3(6): 610–621

[33] Kuse M, Sharma T, Gupta S. A Classification Scheme for Lymphocyte Segmentation in H&E Stained Histology Images. In Recognizing Patterns in Signals, Speech, Images and Videos, vol. 6388, D. Ünay, Z. Çataltepe, and S. Aksoy, Eds. Springer Berlin / Heidelberg; 2010. p. 235–243.

[34] Chaddad A, Tanougast C, Dandache A, Al Houseini A, Bouridane A. Improving of colon cancer cells detection based on Haralick's features on segmented histopathological images. IEEE International Conference on Computer Applications and Industrial Electronics (ICCAIE); 2011. p. 87–90.

[35] Cinar Akakin H, Gurcan M. Content-based Microscopic Image Retrieval System for Multi-Image Queries.IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society; 2012.

[36] Caicedo J, Cruz-Roa A, Gonzalez F. Histopathology image classification using bag of features and kernel functions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2009;5651(LNAI): 126–135

[37] Caicedo J, Gonzalez F, Romero E. Content-based histopathology image retrieval using a kernel-based semantic annotation framework. Journal of Biomedical Informatics. 2011;44(4): 519–528

[38] Cruz-Roa A, Caicedo J, González F. Visual pattern mining in histology image collections using bag of features.Artificial intelligence in medicine. 2011;52(2): 91–106

[39] Díaz G, Romero E. Histopathological Image Classification Using Stain Component Features on a pLSA Model. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, vol. 6419, I. Bloch and R. Cesar, Eds. Springer Berlin / Heidelberg; 2010. p. 55–62.

[40] Cruz-Roa A, Caicedo J, Gonzalez F. Visual pattern analysis in histopathology images using bag of features. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2009;5856(LNCS): 521–528

[41] Galaro J, Judkins AR, Ellison D, Baccon J, Madabhushi A. An integrated texton and bag of words classifier for identifying anaplastic medulloblastomas. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society Conference; 2011. p. 3443–6

[42] Raza S, Parry RM, Sharma Y, Chaudry Q, Moffitt RA, Young A, et al. Automated classification of renal cell carcinoma subtypes using bag-of-features. In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC; 2010. p. 6749–6752

[43] Vanegas J, Caicedo J, Gonzalez F, Romero E. Histology Image Indexing Using a Non-negative Semantic Embedding. In Medical Content-Based Retrieval for Clinical Decision Support, vol. 7075, H. Müller, H. Greenspan, and T. Syeda-Mahmood, Eds. Springer Berlin / Heidelberg; 2012. p. 80–91

[44] Krishnan M, Venkatraghavan V, Acharya U, Pal M, Paul R, Min L, et al. Automated oral cancer identification using histopathological images: a hybrid feature extraction paradigm.Micron. 2012;43(2): 352–64

[45] Krishnan MR, Shah P, Choudhary A, Chakraborty C, Paul RR, Ray AK. Textural characterization of histopathological images for oral sub-mucous fibrosis detection.Tissue & cell. 2011;43(5): 318–30

[46] Ministerio de Salud y Protección Social (Colombia). Guías de atención integral en cancer; 2012

[47] Peng Y, Jiang Y, Eisengart L, Healy MA, Straus FH, Yang XJ. Computer-aided identification of prostatic adenocarcinoma: Segmentation of glandular structures. Journal of pathology informatics. 2011;2: 33

[48] Naik S, Doyle S, Agner S, Madabhushi A, Feldman M, Tomaszewski J. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2008. p. 284–287.

[49] Almuntashri A, Agaian S, Thompson I, Rabah D, Zin Al-Abdin O, Nicolas M. Gleason grade-based automatic classification of prostate cancer pathological images. IEEE International Conference on Systems, Man, and Cybernetics; 2011. p. 2696–2701.

[50] Khurd P, Bahlmann C, Maday P, Kamen A, Gibbs-Strauss S, Genega EM, et al. Computer-aided Gleason grading of prostate cancer histopathological images using texton forest. Proceedings / IEEE International Symposium on Biomedical Imaging: from nano to macro. IEEE International Symposium on Biomedical Imaging. 2010;14: 636–639

[51] Ministerio de Salud y Protección Social and E.S.E.Instituto Nacional de Cancerología (Colombia). Anuario estadístico 2010. Instituto nacional de cancerología; 2012.

[52] Wang Y, Crookes D, Eldin OS, Wang S, Hamilton P, Diamond J. Assisted Diagnosis of Cervical Intraepithelial Neoplasia (CIN). IEEE Journal of Selected Topics in Signal Processing. 2009;3(1): 112–121

[53] Zhang L, Chen S, Wang T, Chen Y, Liu S, Li M. A Practical Segmentation Method for Automated Screening of Cervical Cytology. International Conference on Intelligent Computation and Bio-Medical Instrumentation; 2011. p. 140–143.

[54] He L, Rodney L, Antani S, Thomas GR. Local and global Gaussian mixture models for hematoxylin and eosin stained histology image segmentation. International Conference on Hybrid Intelligent Systems; 2010. p. 223–228.

[55] Jing F, Li MJ, Zhang HJ, Zhang B. Unsupervised image segmentation using local homogeneity analysis,International Symposium on Circuits and Systems. 2003;2: II–456–II–459

[56] Miller SJ, Alam M, Andersen J, Berg D, Bichakjian CK, Bowen G, et al. Basal cell and squamous cell skin cancers. Journal of the National Comprehensive Cancer Network. 2010;8(8): 836–864

[57] Wong CS, Strange RC, Lear JT. Basal cell carcinoma.BMJ Clinical research. 2003;327: 794–8

[58] Camargo J, Caicedo J, Gonzalez F. Kernel-Based Visualization of Large Collections of Medical Images Involving Domain Knowledge. X Congreso Internacional de Interaccion Persona-Ordenador; 2009

[59] Cruz-Roa A, Romero E, Gonzalez F, Diaz G. Automatic annotation of histopathological images using a latent topic model based on non-negative matrix factorization. Journal of Pathology Informatics. 2011;2(2): 4

[60] Cruz-Roa A, Diaz G, Gonzalez F. A framework for semantic analysis of histopathological images using nonnegative matrix factorization. Computing Congress (CCC), 2011 6th Colombian; 2011. p. 1–7.

[61] Gutiérrez R, Gómez F, Roa-Peña L, Romero E. A supervised visual model for finding regions of interest in basal cell carcinoma images.Diagnostic pathology. 2011;6: 26

[62] Díaz G, Romero E. Micro-structural tissue analysis for automatic histopathological image annotation. Microscopy research and technique. 2011;75(3): 343–58

[63] Hipp JD, Sica J, McKenna B, Monaco J, Madabhushi A, Cheng J, et al. The need for the pathology community to sponsor a whole slide imaging repository with technical guidance from the pathology informatics community. Journal of pathology informatics. 2011;2: 31

[64] Ghaznavi F, Evans AJ, Madabhushi A, Feldman MD. Digital Imaging in Pathology: Whole-Slide Imaging and Beyond. Annual Review of Pathology: Mechanisms of Disease. 2012;8: 1

[65] Madabhushi A. Digital pathology image analysis: opportunities and challenges. Imaging in Medicine. 2009;1: 7–10

[66] Hipp J, Cheng J, Daignault S, Sica J, Dugan MC, Lucas M, et al. Automated area calculation of histopathologic features using SIVQ. Analytical cellular pathology. 2011;34(5): 265–75

[67] Nguyen K, Sabata B, Jain AK. Prostate cancer grading: Gland segmentation and structural features. Pattern Recognition Letters. 2012;33(7): 951–961

[68] Loménie N, Racoceanu D. Point set morphological filtering and semantic spatial configuration modeling: Application to microscopic image and bio-structure analysis. Pattern Recognition. 2012;45(8): 2894–2911

[69] Basavanhally AN, Ganesan S, Agner S, Monaco JP, Feldman MD, Tomaszewski JE, et al. Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. IEEE transactions on bio-medical engineering. 2010;57(3): 642–53

[70] Muthu Rama M, Pal M, Bomminayuni SK, Chakraborty C, Paul RR, Chatterjee J, et al. Automated classification of cells in sub-epithelial connective tissue of oral sub-mucous fibrosis-an SVM based approach. Computers in biology and medicine. 2009;39(12): 1096–104

[71] Lomenie N, Racoceanu D. Spatial relationships over sparse representations. In 2009 24th International Conference Image and Vision Computing New Zealand; , 2009.

[72] Xu J, Janowczyk A, Chandran S, Madabhushi A. A high-throughput active contour scheme for segmentation of histopathological imagery. Medical image analysis. 2011;15(6): 851–62

[73] Simsek A, Tosun A, Aykanat C, Sokmensuer C, Gunduz-Demir C. Multilevel Segmentation of Histopathological Images using Cooccurrence of Tissue Objects. IEEE transactions on bio-medical engineering. 2012;99: 1

[74] Ozdemir E, Sokmensuer C, Gunduz-Demir C. Histopathological image classification with the bag of words model. EEE 19th Signal Processing and Communications Applications Conference (SIU); 2011. p. 634–637

[75] Lundin M, Lundin L, Helin H, Isola J. A digital atlas of breast histopathology: an application of web based virtual microscopy. Journal of Clinical Pathology. 200;57(12): 1288–1291

[76] Muthu Rama M, Shah P, Chakraborty C, Ray A. Statistical Analysis of Textural Features for Improved Classification of Oral Histopathological Images. Journal of Medical Systems. 2012;36(2): 865–881

[77] Nguyen K, Sabata B, Jain A. Prostate cancer detection: Fusion of cytological and textural features. Journal of Pathology Informatics. 2011;2(2): 3