Published in final edited form as:
IEEE Trans Inf Technol Biomed. 2012 July ; 16(4): 758–769. doi:10.1109/TITB.2012.2185829.

Content-based Microscopic Image Retrieval System for Multi-Image Queries

Hatice Cinar Akakin and
Department of Biomedical Informatics, The Ohio State University, Columbus, OH, 43210 USA and Department of Electrical and Electronics Engineering, Anadolu University, Eskisehir, Turkey

Metin N. Gurcan [Senior Member, IEEE]
Department of Biomedical Informatics, The Ohio State University, Columbus, OH, 43210 USA

Hatice Cinar Akakin: haticecinarakakin@anadolu.edu.tr; Metin N. Gurcan: metin.gurcan@osumc.edu

Abstract

In this paper, we describe the design and development of a multi-tiered CBIR system for microscopic images utilizing a reference database that contains images of more than one disease. Proposed CBIR system uses a multi-tiered approach to classify and retrieve microscopic images involving their specific subtypes which are mostly difficult to discriminate and classify. This system enables both multi-image query and slide-level image retrieval in order to protect the semantic consistency among the retrieved images. New weighting terms, inspired from information retrieval (IR) theory, are defined for multiple-image query and retrieval. Performance of the system was tested on a dataset including 1666 imaged high power fields (HPF) extracted from 57 Follicular Lymphoma (FL) tissue slides with three subtypes and 44 Neuroblastoma (NB) tissue slides with four subtypes, where each slide is semantically annotated according to their subtypes by expert pathologists. By using leave-one-slide out testing scheme, the multi-image query algorithm with the proposed weighting strategy achieves about 93% and 86% of average classification accuracy at the first rank retrieval, outperforming the image-level retrieval accuracy by about 38 and 26 percentage points, for FL and NB diseases, respectively.

Index Terms

Content-based image retrieval; microscopy multi-image queries; weighting scores; information retrieval

I. Introduction

THANKS to the technical advances in diverse modalities such as X-Ray, CT and MRI, and their common use in clinical practice, the number of medical images is increasing every day. These medical images provide essential anatomical and functional information about different body parts for detection, diagnosis, treatment planning and monitoring as well as medical research, and education. Exploration and consolidation of the immense image collections require tools to access structurally different data for research, diagnostics and teaching. Picture archival and communication systems (PACS) provide the hardware and software for the storage, retrieval and management of radiological images [1]. However, such systems use the patient information, and/or modality to index and search the images; the content of the image is not utilized. Content-based Image Retrieval (CBIR) systems [2], [3], [4], [5], [6], [7] for medical images are important to deliver a stable platform to catalogue, search and retrieve images based on their content.

Although several CBIR projects exist for radiology [8], [9], [10] and several other projects are underway, there is an acute need for a comprehensive and flexible CBIR system for microscopic images with direct implications for the field of pathology and cancer research. Microscopic images present novel challenges because they *i*) are large in size *ii*) demonstrate high degree of visual variation and the often low visual distinctiveness between classes due to large variation in preparation (e.g. staining, thickness), *iii*) show huge biological variation. Therefore, a well-designed CBIR system for microscopic images can be extremely useful resource for cancer research, diagnosis, prognosis, treatment and teaching. In other words, such a system can *i*) assist pathologists in their diagnosis and prognosis, *ii*) potentially help to reduce inter- and intra-reader variability in clinical practice for the diseases, especially those with complicated classification, *iii*) help cancer researchers in better understanding of cancer development, treatment monitoring and clinical trials, *iv*) train future generation of researchers by providing consistent, relevant and always available support and assistance. In this paper, we describe the design and the development of a multi-tiered CBIR system for microscopic images from a reference database that contains more than one disease.

To provide a motivating example and to test the ideas developed in this work, images in our reference database include sample regions cropped from digitizedhematoxylin and eosin (H&E) stained whole-slides. Neuroblastoma (NB) and Follicular Lymphoma (FL) tissue images have been collected as part of our ongoing projects for both diseases. The input images to our system are digitized using a Scope XT digitizer (Aperio, San Diego, CA, USA) at 40× magnification. FL tissue slides were collected from the Department of Pathology, The Ohio State University in accordance with an IRB (Institutional Review Board) approved protocol. NB whole-slide tissue samples were collected from the Children’s Oncology Group slides. According to the recent medical statistics FL accounts for 20–25% of non-Hodgkin lymphomas in the US [11] and affects predominantly adults, particularly the middle-aged and elderly. FL cases are stratified to three histological grades from low risk to high risk category as follows: Grade-I, Grade-II and Grade-III. NB is the most common extracranial solid cancer in childhood and in infancy. According to the International Neuroblastoma Classification System, NB tissues are mainly divided into two subtypes such as Stroma-rich or Stroma-poor based on the degree of Schwannian stroma development [12]. Additionally, stroma-poor tissue has three subtypes such as Undifferentiated, Poorly Differentiated and Differentiating. These subcategories as well as the mitosis karryorrhexis index are used for prognostication.

Annotation of microscopic images, e.g. H&E stained pathology slides, with subtypes of the main disease needs an expert pathologist to select pathology-bearing regions, or regions of interests (ROIs) from the whole slide. Then each selected region is annotated semantically by giving a score according to its visual qualitative characteristics. For example, the number of centroblasts or mitotic-karryorrhectic cells can establish a score to interpret the underlying subtype of that disease. The final decision on the grade or subtype of the disease for the whole slide is given after considering the annotations of all sample regions, i.e., the average subtype-related score over all sample regions is assigned as the final score of that whole slide. Considering the extremely large sizes of microscopic images, it is obvious that manual annotation of these images is a time-consuming process and those annotated images may not be easily available for clinical use. Therefore, one of the aims of this study is to organize the annotated microscopic images in a database and utilize these images for the training of a CBIR system for microscopic images with different disease types and with their subtypes.

The novel aspects of our multi-tiered approach are: 1) it retrieves the most similar disease types in the slide-level rather than in the image-level by enabling multi-image queries in order to ensure the consistency among the retrieved images, 2) Slide-level scores are weighted in a sophisticated way by modifying the *term frequency – inverse document frequency* weighting concepts of information retrieval (IR) theory [13] to decrease the sensitivity of the proposed CBIR system to erroneously annotated sample images in the database. These aspects were designed to mimic the evaluation methodology of pathologists when they review a whole slide microscopic image. Since in real medical applications, especially for microscopic images at high magnifications, the query object is more likely to be a set of sample images extracted from a whole slide image rather than being a single image, the multi-image query model suits perfectly for our case. It has been also proved that, query by multi-images leads to more scalable and satisfactory query performances by overcoming the limitation on the specification of image content of single-image queries [14], [15].

In CBIR systems, images are typically represented with feature vectors extracted using low-level image processing techniques [8], [9], [16]. However, similarities in feature vector level does not always guarantee the semantic similarity, (i.e., interpretations of images according to their predefined categories), between query image and retrieved images. This is known as *the semantic gap problem* [17], [18]. In this paper, we will explore the effect of slide-level retrieval system with multiple query images in order to increase the semantic relevance of query image set and retrieved images.

A general flowchart of the proposed CBIR system is illustrated in Fig. 1. It shows the main steps of the CBIR algorithm, e.g., feature extraction, major disease type classification (1^{st} Tier), image retrieval according to the subtypes of the diseases (2^{nd} Tier).

The rest of the paper is organized as follows; Section II presents related works on CBIR methods for medical images. Section III explains the features extracted from the database images. Two-tier retrieval approach is explained in Section IV. Database description and results of the experiments are discussed in Section V. Conclusions are drawn in Section VI.

II. Related work

Most of the commercial search engines (e.g. Google, Yahoo!, Bing Image Search) are built around a semantic search, i.e. the user needs to type in a series of keywords and the images in those databases are also annotated using keywords, the match is accomplished primarily through these keywords. CBIR systems have been developed in the recent years to organize and utilize the valuable image sources effectively and efficiently for diverse collections of images. Most of the recent CBIR systems in biomedicine [5], [8], [9], [19], [20] are designed to classify and retrieve images according to the anatomical categories of their content, i.e., head or chest X-ray images or abdominal CT images. For example, the ASSERT system [5] was designed for high resolution computed tomography (CT) images of the lung where each set of feature was extracted from the pathology-bearing regions. Similarly, CBIR for CT images of three types of liver lesions was investigated by incorporating semantic features observed by radiologist as well as features computationally extracted from the images [8]. Previously, a prefiltering approach [9] was proposed to reduce the search space of query images by categorizing the images using multi-class support vector machines and fuzzy c-mean clustering. Twenty different modality specific semantic categories based on body region, and orientation differences and the database for retrieval included microscopic images of leukemia, Alzheimer’s disease, bacterial meningitis and skin lesions were used for retrieval. The retrieval after prefiltering was done according to main disease categories only, which is similar to the first-tier of our two-tiered approach. In another study [19], expectation-maximization algorithm was used to generate clusters of block-based low-level features extracted from radiographic images. Then, the similarity between two clusters was estimated as a function of the similarity of both their structures and the measure components. Pourghassem and et. al [20] proposed two level hierarchical medical image classification method. The first level was used to classify the images into the merged and non-merged classes. They tested their algorithm on medical X-ray images of 40 classes. Although this is a two-level hierarchical classification, it is different from our approach because only the merged classes were evaluated in the second level to be classified with MLP classifiers into one of 40 classes.

Traditional indexing and search strategies used in radiological systems are not directly applicable in the context of digital microscopy, since it is not obvious how to define a primary key or major anatomical structure for such images. To complicate things further, most known structures (e.g. cells, its components, tissue, etc.) are much more complex and require more detailed analysis than that would be needed at the higher resolutions and scale of radiological images. The feature extraction from microscopic images is also challenging because these images are composed of varying textures, overlapping structures, and different cell constituents even for the same disease types.

For the last decade, a few CBIR systems for the microscopic images have been developed for clinical use [6], [21], [22], [7], [16]. Mehta et al. designed a region specific retrieval system based on sub-image query search on whole slide images by extracting scale invariant features on the detected points of interests and 80% of match was achieved with the manual search for prostate H&E images [22] in the top five searches. In another study, image-level retrieval of four special types of skin cancer [21] was performed by constructing a visual word dictionary under a bag-of-features approach in order to represent a relationship between visual patterns and semantic concepts. Zheng et al. [6] proposed a CBIR system based on the weighted similarities of four feature types such as color histogram, image texture, Fourier coefficients and wavelet coefficients. The retrieval performance of their system was tested using agglomerative cluster analysis for different pathology image categories and the best retrieval performance was observed for prostate query images.

Recently, Yang et al. [7] developed a web-based system called *PathMiner* which includes automatic segmentation, CBIR and classification modules to assist diagnostics in pathology. They evaluated the classification performance of their system on five different blood cells such as chronic lymphocytic leukemia (CLL), mantle cell lymphoma (MCL), follicular center cell lymphoma (FCC), and acute lymphocytic leukemia (ALL) and acute myelogenous leukemia (AML) by using support vector machine (SVM) classifiers with texton histogram features and 87.27% of classification accuracy was achieved on an open-set with large variations in staining characters.

Most of the CBIR approaches designed for microscopic images have their own specific application area, specific feature extraction technique or a specific similarity measure for the evaluation. For example, disease-specific CBIR systems [22], [21], [16] have been developed for clinical decision support of specific diseases while some of the CBIR systems were designed for the classification of different types of pathology images, i.e., liver tissue, prostate tissue, breast tissue, lymph node and etc [6].

Although many promising CBIR approaches were developed for medical applications, there are still gaps in terms of image contents, retrieval methodology, performance evaluations and their application areas [17], [18], which make this research area an open problem for further studies. Particularly, the majority of the retrieval methodology of the published CBIR techniques focused on image-level retrieval either by choosing or defining an appropriate distance metric to compare the feature vectors from the query and database images [8], [16], [23]. However, multi-image query based retrieval is more suitable for challenging medical CBIR applications. Especially, microscopic images at high magnifications require multi-image queries in order to specify the query images more efficiently. Therefore, our CBIR method will focus on defining a retrieval methodology for multi-image queries, which can be also applicable for any type of multi-image query and retrieval application.

In summary, our approach focuses on one modality, which is the digital brightfield microscopic images of tissue slides and it does not aim to provide a way to search and index generic medical image collections. It differs from the existing microscopic CBIR methods mainly in two aspects. First, two different diseases (FL and NB) are processed within a CBIR system with their high level semantic annotations. The framework can also be extended to several other diseases. Second, our approach enables multi-image queries instead of one image query and provide a slide-level retrieval by keeping the slide-level consistency among retrieved images by using weighting scores depending on the image-level rank order and distributions of the subtypes over the reference dataset.

III. Feature Extraction

In this section, we will explain the feature extraction techniques that we employed to the images in our database.

A. Low level feature extraction

There are many factors affecting the performance and accuracy of CBIR systems, such as choosing more discriminative features, similarity measurement criteria, query formulation and so on. In order to design an effective CBIR system, the initial step in our work is to extract discriminative features from the images in the reference database. These features will also be calculated for query images.

One of the most discriminating characteristics of microscopic images is color, especially when compared to most common radiological images, which are mostly gray-level. Due to the high resolution of microscopic images, subtle changes in characteristics of cells, combinations of cells, structures and tissues can also be differentiated from each other by texture characteristics. Therefore, for our CBIR design, we heavily make use of color and texture characteristics and extract these features using low level image feature extraction techniques.

1. *Color features:* H&E images have considerably limited color spectrum, i.e., there are few dominant colors (hues of blue and pink) as shown on the sample images in Fig. 4 and Fig. 5. Therefore, in order to better represent the limited color information in more detail we used two more color spaces in addition to red-green-blue (RGB) color space. These additional color spaces are CIELab (Lab) and Hue-Saturation-Value (HSV) color spaces. In the Lab color space L corresponds to illumination and a and b channels corresponds to color opponents. Thus, features extracted from the Lab space characterize the intensity and color information of images separately [24]. On the other hand, the HSV color space is known with similarity to the human conceptual understanding of colors. Besides this, HSV space can separate the chromatic and achromatic components, i.e., hue (H) channel distinguish colors, saturation channel (S) represents the percentage of white light added to a pure color space and value (V) refers to intensity of perceived light [24]. For each channel of a given color space; mean value and standard deviation is computed as first and second order statistics features. In total, 18 (2 features × 3 channels × 3 color spaces) color features are extracted from each image. Additionally, mean value, standard deviation, skewness, kurtosis, maximum and

minimum values, energy and entropy values are computed for gray level intensity image. In summary, 26 color and gray-scale features are extracted using three different color spaces for a given image.

2. *Texture features:* Microscopic images with different disease types and subtypes can be distinguished via their homogeneity or texture characteristics. To capture the discriminative texture information, we investigated several texture feature extraction methods in the literature [25], [26], [27]. Co-occurrence histograms are the most frequently used method for texture feature extraction [4], [27], [28], [29]. They can be defined as a sample of a joint probability density of intensity levels of two pixels separated by a given displacement. The distribution in the histograms depend on the rotation angle and distance relationship between pixels. Once the co-occurrence histogram is computed, various features can be extracted related to texture characteristics, lower and higher order statistics, information theory related features and correlation measure. As a consequence, we extracted the following features: mean, standard deviation, contrast, correlation, energy, entropy and homogeneity from the normalized co-occurrence histograms for each *RGB* and *Lab* color channels and gray-level images. In addition, mean value, homogeneity and entropy values are extracted from the difference histograms [30] of the normalized co-occurrence matrix. For a given image, a total of 80 texture-based features are extracted using RGB, HSV color spaces and gray level intensities. It should be noted that, average of the co-occurrence histograms for eight different directions, i.e., 0°, ±45°, ±90°, ±135°, 180°, are calculated in order to obtain rotation invariant features. It should be noted that the images are at the same magnification level therefore, no scaling of the features is needed.

Once all color and texture features are extracted, they are concatenated to form a 106 dimensional feature vector. After feature extraction, a Z-score normalization is applied to each extracted feature in the feature vector by subtracting the mean of that feature followed by dividing to the standard deviation of that feature computed over the reference dataset. This normalization step converts all extracted features to a common scale with an average of zero and standard deviation of one. Then normalized feature vectors (NF) are stored for further CBIR processes. When the query image set is given to the system, the system will employ the same feature extraction and normalization procedure to the query images.

Instead of analyzing the contribution of extracted features based on selected color spaces or texture features by using feature selection algorithms, we preferred to use subspace projection method in order to represent the feature vectors more sparsely by decreasing the correlation among the features. In the literature, subspace projection methods have been widely used for dimensionality reduction and feature extraction. They are popular to analyze structures where large amount of correlated numerical data is available. Nonnegative Matrix Factorization (NMF) [31] is one of the data driven subspace projection method, which aims to factorize a data matrix into basis vectors and their combiner coefficients. They perform better for features extracted from partially represented data [32]. In our case, features from different color spaces and texture features can be assumed to be features of a partially represented data. Using a training data set, F_{DS} with size lxT, the m basis vectors, columns of W, are obtained as:

$$F_{DS} \approx W.H, \quad (1)$$

Here, l is the length of the feature vector, T is the number of samples in the dataset and m ($m < l$) is the size of NMF features. In the factorization in Eq. 1, the columns of the lxm matrix W stand for the basis vectors and the columns of the mxT matrix H determine how the basis vectors are activated to reconstruct the feature matrix F_{DS}. The columns of H represent the NMF-based feature vectors of the corresponding data. The classification of a test feature vector F_Q is based on its NMF features given by $h = W^+ F_Q$. The number of columns m in the (basis) matrix W was determined for each disease type empirically during training stage. In this study, the implementation of NMF code was based on the projected gradient method [33].

IV. Two-tier retrieval approach for multi-image queries

Our CBIR system operates at two tiers. In the first tier, the designed classifier categorizes the query image/images into one of the major disease types such as FL and NB. Once the disease category of the image is determined, the search for the query image can be carried out among the category relevant subtypes in the subsequent tier. For example, when the query image belongs to NB disease, database images in the first tier will be filtered according to the NB disease category. Then the subsequent search will be only performed on the NB category subset to retrieve the images from the correct category of the query images.

In the second tier, we will use our proposed multi-image query and retrieval methodology to retrieve the images from the reference database in the order of their image-level visual similarities by preserving the slide level semantic similarity.

A. First tier: Classification of disease type with SVM

A support Vector Machine (SVM) type classifier was employed to categorize the query image into one of the major disease type such as NB or FL using the extracted features which are explained in Section III-A. SVM classifiers are well founded in statistical learning theory and have been successfully used for various classification tasks in computer vision. Their purpose is to find a decision hyperplane for a binary classification problem by maximizing the margin, which is the distance between the hyperplane and the closest data points of each class in the training set that are called support vectors. The hyperplane is chosen among all the possible hyperplanes through a complex combinatorial problem optimization, so that it maximizes the distance (called the margin) between each class and the hyperplane itself. As SVMs are restricted to binary classification, several strategies are developed to adapt them for multiclass classification problems [34] such as one-against-all classification and pairwise classification.

In our SVM classifier, we selected the radial basis function which is one of the most frequently used kernels and it gives better results than other kernels for the categorization of our data. Libsvm Matlab code [35] was used in the experiments of this study.

B. Second tier: Slide level image retrieval

In this part of the CBIR algorithm, we proposed a two-level retrieval system, in the first level the search is performed similar to traditional CBIR systems such that the images are retrieved based on their image-level similarities. In the second level, the images will be retrieved according to their similarities in the slide-level. Once the category of the query image is detected in the first tier, further search is performed on the prefiltered database which includes only the sample images of the detected disease category. As we described in Section V-A, each disease has higher level semantic annotations based on their histological grades such as Grade-I, Grade-II and Grade-III in FL disease or differentiating levels such as Stroma-rich, Undifferentiated, Poorly Differentiated and Differentiating in NB disease. Therefore, it is necessary to retrieve images related to their higher level semantic characteristics in order to provide more accurate results to the user of the CBIR system.

Algorithm 1 summarizes the image-level search and Fig. 2 illustrates a sample nearest neighbor search scheme for a given query image set in image level. Here we used the term of *image-set* in order to represent multiple images in one query. Note that, image set may include only one image or several images cropped from one tissue slide. The distance between each image of query Q and the individual images in the dataset are computed using the correlation distance measure as shown in Equation 2.

$$\begin{aligned} Dist(F_{Q^n}, F_{DS^t}) &= 1 - \{Correlation\{F_{Q^n}, F_{DS^t}\}\} \\ &= 1 - \left\{ \frac{\langle (F_{Q^n}), (F_{DS^t}) \rangle}{\left\| F_{Q^n} \right\| \left\| F_{DS^t} \right\|} \right\}, \\ t=1, \ldots, T, \quad T = |F_{DS}| &\text{ and } n=1, \ldots, N, \quad N = |Q|, \end{aligned} \tag{2}$$

where N is the number of individual query entities in the given query image set Q, T is the number of images in the reference dataset DS, F_{Q^n} represents the feature vector of the n^{th} query image, F_{DS^t} represents the feature vector of the t^{th} image of the given dataset, $\langle \cdot, \cdot \rangle$ is the inner product, $\| \cdot \|$ is the L_2 norm and $| \cdot |$ is the cardinality.

Algorithm 1 provides us the frequency of similar images per image in the dataset to a given query image set or a slide in terms of scores. Scores are computed by summing the number of occurrences of each image in the dataset for a KNN search of that query image set. The output of this algorithm is the traditional image-level based retrieving of most similar images from the given dataset and their image level scores.

In our alternative approach to image-level retrieval, we propose to retrieve similar images from the database by keeping the slide level semantic grade among the retrieved images. For this purpose, we introduced a slide-level retrieval methodology, which is summarized in Algorithm 2. The conventional way of ranking the similarity of slides to a given query image-set is by sorting the similarity scores of the reference slides independent from their subtypes and retrieving the highest scored slides from the database, which means that subtypes of the slides are considered equally important. In our proposed approach, the first step is to scale the score of each slide by assigning different weight parameters based on subtype frequencies over the reference database. For example, in our dataset the number of slides per subtype is not equal, i.e., FL Grade-I has 15 slides while FL Grade-III has 22 slides. Therefore our algorithm assigns higher weights to the slides of FL Grade-I since its frequency is lower than FL Grade III. Similarly, the number of images per slide is varying among the slides. In order to make a sophisticated and intelligent relevance ranking system, it is necessary to take into account those statistical variations among slides and subtypes. The computational model illustrating all intermediate levels of the proposed slide-level CBIR system is given in Fig. 3 for a sample query image set.

Assigning weights to each slide and to each subtype based on the distribution (or frequency) of images per slide and distribution of slides per subtype is motivated by similar approaches in information retrieval theory [36]. In information theory, “term frequency (*tf*)” refers to the frequency of an index term in a reference document and “inverse document frequency (*idf*)” is inversely proportional to the number of documents containing that index term [37], [38], [39], [40] and they are used to assign weights for each term of the documents before computing similarity. However, in our case we do not have definite terms (i.e. words in documents) but we have scores representing the unweighted similarities between the query image set and the reference slides. Therefore, we adapted these concepts to assign weights to normalize the similarity scores of each slide and each subtype depending on the slide-level and image-level statistics of the dataset (e.g. the number of images per slide or the number of slides per subtype).

In our slide-level retrieval system, we redefined scores in terms of image term frequency (*itf*) which corresponds to normalized number of image count of a particular slide for a given query set.

Additionally, inverse slide frequency (*isf*) is inversely proportional to the number of slides per subtype, and it gives lower weights to the slides occurring in a larger set of subtype. Equation 3 represents the calculation of *isf* per subtype.

$$Subtype\_isf^{c}=\log\frac{\sum_{c=1}^{C}S^{c}}{S^{c}} \qquad (3)$$

where $c = 1, \ldots, C$, and S^c is the number of total slides for the c^{th} disease subtype.

Algorithm 2 summarizes the proposed weighting score approach. In order to take into account the rank of the slides in terms of their *itf* scores (*Score_itf*), we assigned a weighting term, called *Rank_weight*, to each subtype. First, *Score_itf* values were sorted in descending order and top K_2 of the sorted scores are summed according to their subtypes. These summed scores represent the *Rank_weight* term for each subtype. Basically, *Rank_weight* term corresponds to the proportion of summed *itf* scores within the top K_2 *itf* scores per subtype. The purpose of *Rank_weight* is to increase the likelihood of retrieving the subtype of the highest scored slides by assigning higher weights to the slide scores of that subtype. Notice that, unlike *Score_itf* and *Rank_weight*, *Subtype_isf* term depends only on the statistics of the dataset and hence it can be computed offline independent from the query image set.

V. Dataset and experimental results

A. Annotated microscopic image dataset

Table I lists the details of the database that we used in this study and Fig. 4 shows randomly selected sample images belonging to different histological grades of FL cases. The number of cropped images per slide is between 11 and 30 for FL cases and 7 and 35 for NB cases. For FL slides, a team of experienced hematopathologists selected about ten random microscopic high power fields (HPF) to interpret the disease grade in terms of the average number of centroblasts per HPF. Note that, for both FL and NB, we use internationally accepted and clinically practiced standards. For FL, our collaborating pathologists use the World Health Organization (WHO) grading system. For NB, the International Neuroblastoma Classification System, invented by our collaborator Dr. Hiroyuki Shimada, is used. In our database each HPF corresponds to one image and each slide belongs to one patient. Note that, in order to simplify the terminology of this paper, we used “image set” and “slide/patient” pairs interchangeably. The consensus of pathologists is used to stratify cases into their histological grades. The sizes of the cropped images are 1353×2168 pixels for FL cases and 1024×1024 and 1712×952 pixels and for NB cases.

For NB slides, pathologists pick the representative regions (images) from the whole slide and examine those regions at higher magnifications. The final decision for the differentiation grade of the entire slide is based on the grades of the sample images selected from that slide. Due to this differentiation grades, NB disease is differentiated to two subcategories such as Stroma Rich (SR) and Stroma Poor (SP). Stroma Poor subtype has three more subtypes such as Differentiating (D), Poorly Differentiated (PD) and Undifferentiated (UD). In total, NB disease has four subtypes. Figure 5 illustrates sample images cropped from different slides with different differentiation grades of NB to give an idea about their visual appearances.

Because of the heterogeneous characteristic of these tumors, all image-level annotations may not match with the annotation of the entire slide, which causes intra-slide variations. Additionally, there may be variations among inter-and intra-readings of pathologists, because of which, FL Grade-I and FL Grade-II subtypes [41] and NB-PD and NB-UD subtypes [25] are the most confused subtypes of the FL and NB diseases, respectively.

B. Results of the first tier

The organization of the test set and training set is performed in patient (slide) independent manner. In other words, none of the images of a test slide is included in the training set in order to obtain more realistic results both in first tier and second tier experiments.

In the experiments of the first tier evaluation, we randomly selected five FL and five NB slides for each test set and the remaining slides were used for the training of the SVM classifier. In total, ten slides were randomly chosen for test set and 91 slides were used for training. In order to comprehensively test and train all NB and FL sample slides with different test sets, we repeated the testing scheme until all the slides were used as a test slide, an approach similar to the leave-one-out testing.

The classification results of the first tier, summarized as a confusion matrix in Table II, were computed as the overall average over 50 test repetitions. These results were obtained with SVM classifiers trained with normalized features. The classification accuracies were evaluated in two different ways. One way is to evaluate the results at the image level, such that each image is classified independent from the other images of that test slide. The other way is to interpret the results at the slide-level by combining the decisions on all images of a test slide using decision fusion rules. Here, the majority rule is employed to the assigned classes of the test images to determine the slide level classification of that image set. In other words, the majority of the assigned classes for each test image is chosen as the representative class for that given slide. It is observed that, 0.6% of FL test images (6 images) were classified as NB at image level and after majority voting, all FL slides were classified correctly. For NB case, 5.3% of NB test images (38 images) were misclassified. After majority voting, all NB test slides except one with differentiating grade were correctly classified. It is noticed that, all images of that NB slide were also misclassified at the image level. This misclassified NB slide was used with both NB and FL slides in order to evaluate the retrieval accuracy in the second tier of the algorithm in case of a misclassified slide.

C. Results of the second tier

After determining the classes of query slides in the first tier, the next step is to retrieve the most relevant images from the database according to the main disease type of the query image set. Leave-one-slide-out cross validation testing scheme was employed for each disease type separately such that at each round one tissue slide with all corresponding images were used as a query image-set and the images of the remaining slides were used as the reference dataset for that query.

The organization of the performed experiments for the second tier is shown in Fig 6. For the slide level retrieval, we used the proposed weighted scores to rank the slides according to their relevancy to a given query slide. In order to assess the performance of *Rank_weight* on the retrieval system, we evaluated the experiments both with (Slide-level II) and without (Slide-level I) using this weighting term.

We used both precision and *A*rea *U*nder *Presicion* versus *R*ecall *C*urves (*AUPRC*) to measure the retrieval accuracy in our experiments. For a query Q, let K be the number of retrieved images and B be the number of relevant images among K retrieved images for that query. Then precision (P) and recall (R) values are calculated as in Equation 4.

$$P=\frac{B}{K} \text{ and } R=\frac{B}{M}. \quad (4)$$

Here M represents the number of all relevant images in the dataset to the query Q. PR curves can jointly represent the false alarms and dismissals for different K values in one plot. In this paper, the retrieved image or slide is considered to be correct if its semantic annotation (subtype of the disease) is same as the semantic annotation of the query image set.

When we analyzed the reults given in Tables III–VI and Fig. 8 for the retrieval of NB and FL slides, we had the following observations:

Retrieval for FL disease:

- In FL case the proposed weighting scheme results a higher retrieval accuracy for all subtypes of the FL disease when compared with image-level retrieval. The retrieval performance improvement between image-level retrieval and slide-level retrieval schemes is shown in Table III. The average retrieval performance for both NMF and NF features increases gradually from Image-level to Slide-level II, e.g., 30, 45 and 54 $AUPRC$ values were achieved for Image-level, Slide-level I and Slide-level II schemes with NMF features, respectively.
- Table IV presents the confusion matrix of NMF features in terms of precision values with top rank retrieval indices. Each row of the confusion matrix represents the precision values for the corresponding retrieval indices for the actual disease type (given in the first column of each row). For example, when the search image belongs “Grade I” and the retrieval rank is 3 as shown in Table IV, the correct retrieval accuracy is 77.9% and 22.1% were retrieved from other grades, e.g., 13.3% of the samples were retrieved from Grade II and 8.8% of them were retrieved from Grade III diseases. Grade I and Grade II are the most confused subtypes of FL disease, which is also the case clinically. Grade III achieved 95.5% of precision for the first rank. An average classification accuracy of 93% was achieved for FL diseases for the first rank retrieval. It should be noted that, even though one NB slide was misclassified as an FL slide, it was consistently retrieved at the last rank so that this misclassified slide did not degrade the performance of the retrieval.

Retrieval for NB disease:

- The retrieval performance for NB slides was improved with the proposed weighting scheme as suggested (Tables V–VI). Especially, for SR and PD subtypes of the NB disease, higher precision values were achieved when compared with D and UD cases. A possible explanation for this observation is that, PD and UD subtypes are the most commonly confused subtypes because of their high visual similarities [25].
- Even though visual similarities between PD and UD cases are high, the proposed score weighting approach with NMF features improved the retrieval accuracy for about 32 $AUPRC$ points for PD subtype by using weighted score with $Rank\_weight$ (Slide-level II) when compared with image-level retrieval accuracy. On the other hand, although UD case was the most difficult subtype to classify among all subtypes, 12 and 20 $AUPRC$ points improvement was achieved via Slide-level I and Slide-level II aprroaches, respectively, when compared with image-level retrieval.

- Average $AUPRC$ retrieval accuracies of NB slides for image-level, Slide-level I and Slide-level II methods are 39, 63 and 72, respectively.

NF versus NMF features:

- We compare the retrieval performance of these two feature types in Tables III and V. As suggested, NMF features perform slightly better than NF features, such that, 4 and 8 average $AUPRC$ points for improvement for NB slides and 1 and 3 average $AUPRC$ points improvement for FL slides were achieved by using Slide-level I and Slide-level II weighting approaches, respectively.

Comparison of precision values for Slide-level II and image level retrieval with NMF features is shown in Fig. 7. The proposed weighting strategy achieved about 93% and 86% of average classification accuracy at the first rank retrieval, outperforming the traditional image-level retrieval accuracy by about 38 and 26 percentage points, for FL and NB diseases, respectively.

In order to analyze the effect of the number of query images on the retrieval accuracy, we conducted an extra experiment on FL-NMF features with Slide-level II weighting approach. The number of query images was increased from one to maximum number of available images for a given query slide. The number of images per FL slides varies between 11 to 30. Figure 8 shows the precision values as a gray level image, where bright pixels represent higher precision values (i.e. pure white indicates a precision of 1 and pure black indicates 0). It is observed from this figure that, as the number of query images was increased the retrival accuracy was also increased.

An important point for the efficiency of this proposed approach is the parameter selection, i.e., K and K_2 parameters used to compute weightings of the scores, where K represents the number of searched images in image-level retrieval which is used further for computing both unweighted and weighted scores of the slides and K_2 is used to compute the $Rank\_weight$ parameter. In order to find the best parameters, we conducted an exhaustive search. We ran the proposed CBIR algorithm for $K = 1, \ldots, T$, and $K_2 = 1, \ldots, 7$ where T is the total number of images in the reference dataset. For NB disease, $K = 21$ and $K_2 = 5$ gives the best retrieval results while for FL disease, $K = 40$ and $K_2 = 5$ gives the best retrieval results. Different number of NMF features, i.e., 10 to 100, were tested for both FL and NB cases and best performances were obtained with 40 NMF features for NB and FL cases. Therefore, it is necessary to select these parameters separately for each main disease type during training.

VI. Conclusion

In this paper, we have presented a novel content-based microscopic image/slide retrieval algorithm. We have demonstrated that by using the proposed weighting scheme inspired by information retrieval theory, the slide-level retrieval performance of the CBIR system is considerably better than the traditional image-level retrieval accuracy for all seven subtypes of two challenging diseases, which have inter-and intra-reading semantic variations, intra-slide semantic variations and inter-subtype visual similarities. In the first tier, only one slide among 44 NB slides is misclassified and in the second tier, about 26 percentage points of improvement was achieved on the classification accuracy at the first rank retrieval over all diseases by using the proposed score weighting strategy. This CBIR system can enable the user, e.g, a pathologist, to select multiple HPF regions from a suspected tissue and submit those images as a query to the CBIR system and retrieve the most relevant slides with their semantic annotations with higher accuracies. The results, achieved under those challenging conditions, are also promising for automatic and unsupervised selected query images based on their HPF regions. Application of the proposed weighting strategy, inspired by the information retrieval theory is not limited to microscopic images only, and can be also useful for any type of multi-query search and content-based retrieval systems.

In our future work, we will *i*) investigate more effective texture and color feature extraction methods [42], [43], [44] *ii*) improve the robustness of the system by increasing the number of patients/slides in the database, *iii*) enhance the diversity of the database by including microscopic images from different disease types, *iv*) evaluate the performance of the system on automatically selected HPF regions for the query and finally *v*) develop a multi-purpose web-based tool for training future generations of researchers by providing consistent, relevant and always available support and assistance for the challenging diseases and finally help cancer researchers in better understanding of cancer development, treatment monitoring and clinical trials.

Acknowledgments

The authors would like to thank Dr. Gerard Lozanski for insightful discussions, Robert Schmidt and Amie Draper for developing the graphical user interface. The project described was supported in part by Award Number R01CA134451 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, or the National Institutes of Health.

References

1. Choplin RH, Boehme JM, Maynard CD. Picture archiving and communication systems: an overview. Radiographics. 1992; 12(1):127–9. [PubMed: 1734458]
2. Muller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medical applications clinical benefits and future directions. International Journal of Medical Informatics. 2004; 73(1):1–23. [PubMed: 15036075]
3. Hsu W, Long LR, Antani S. Spirs: a framework for content-based image retrieval from large biomedical databases. Studies in health technology and informatics. 2007; 129(Pt 1):188–92. [PubMed: 17911704]
4. Tang HL, Hanka R, Ip HHS. Histological image retrieval based on semantic content analysis. IEEE Transactions on Information Technology in Biomedicine. 2003; 7(1):26–36. [PubMed: 12670016]
5. Shyu CR, Brodley CE, Kak AC, Kosaka A, Aisen AM, Broderick LS. Assert: A physician-in-the-loop content-based retrieval system for hrct image databases. Computer Vision and Image Understanding. 1999; 75(1–2):111–132.
6. Zheng L, Wetzel A, Gilbertson J, Becich M. Design and analysis of a content-based pathology image retrieval system. IEEE Transactions on Information Technology in Biomedicine. 2003; 7(4): 249–255. [PubMed: 15000351]
7. Yang L, Tuzel O, Chen W, Meer P, Salaru G, Goodell L, Foran D. Pathminer: A web-based tool for computer-assisted diagnostics in pathology. Information Technology in Biomedicine, IEEE Transactions on. May; 2009 13(3):291–299.
8. Napel SA, Beaulieu CF, Rodriguez C, Cui J, Xu J, Gupta A, Korenblum D, Greenspan H, Ma Y, Rubin DL. Automated retrieval of ct images of liver lesions on the basis of image similarity: method and preliminary results. Radiology. 2010; 256(1):243–52. [PubMed: 20505065]
9. Rahman MM, Bhattacharya P, Desai BC. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. IEEE Transactions on Information Technology in Biomedicine. 2007; 11(1):58–69. [PubMed: 17249404]
10. Akgul C, Rubin D, Napel S, Beaulieu C, Greenspan H, Acar B. Content-based image retrieval in radiology: Current status and future directions. Journal of Digital Imaging. 2010:1–15. [PubMed: 20024595]
11. Indolent Follicular Lymphoma. Lymphoma Research Foundation; 2008.
12. Shimada H, Ambros IM, Dehner LP, Hata J-i, Joshi VV, Roald B, Stram DO, Gerbing RB, Lukens JN, Matthay KK, Castleberry RP. The international neuroblastoma pathology classification (the shimada system). Cancer. 1999; 86(2):364– 372. [PubMed: 10421273]

13. Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. 1. Addison Wesley; May. 1999
14. Iqbal, Q.; Aggarwal, JK. Feature integration, multi-image queries and relevance feedback in image retrieval. 6th International Conference on Visual Information Systems on DMS; 2003. p. 467-474.
15. Bjoerge, T.; Chang, E. Why one example is not enough for an image query. Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on; 2004. p. 253-256.
16. Naik, J.; Doyle, S.; Basavanhally, A.; Ganesan, S.; Feldman, MD.; Tomaszewski, JE.; Madabhushi, A. A boosted distance metric: application to content based image retrieval and classification of digitized histopathology. Karssemeijer, N.; Giger, ML., editors. Vol. 7260. SPIE; 2009.
17. Deserno T, Antani S, Long R. Ontology of gaps in content-based image retrieval. Journal of Digital Imaging. 2009; 22:202–215. [PubMed: 18239964]
18. Zhou, XS.; Zillner, S.; Moeller, M.; Sintek, M.; Zhan, Y.; Krishnan, A.; Gupta, A. Semantics and CBIR: a medical imaging perspective. Proceedings of the 2008 international conference on Content-based image and video retrieval; New York, NY, USA: ACM; 2008. p. 571-580.
19. Iakovidis D, Pelekis N, Kotsifakos E, Kopanakis I, Karanikas H, Theodoridis Y. A pattern similarity scheme for medical image retrieval. Information Technology in Biomedicine, IEEE Transactions on. 2009; 13(4):442–450.
20. Pourghassem H, Ghassemian H. Content-based medical image classification using a new hierarchical merging scheme. Computerized Medical Imaging and Graphics. 2008; 32(8):651–661. [PubMed: 18789648]
21. Bayro-Corrochano, E.; Eklundh, J-O. Visual Pattern Analysis in Histopathology Images Using Bag of Features, ser Lecture Notes in Computer Science. Vol. 5856. Springer; Berlin / Heidelberg: 2009.
22. Mehta, N.; Alomari, RS.; Chaudhary, V. Content based sub-image retrieval system for high resolution pathology images using salient interest points. Int. Conf. Proc IEEE EMBS; 2009. p. 3719-22.
23. Yang L, Jin R, Mummert L, Sukthankar R, Goode A, Zheng B, Hoi SCH, Satyanarayanan M. A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval. IEEE Trans Pattern Anal Mach Intell. 2010; 32(1):30–44. [PubMed: 19926897]
24. Gonzalez, RC.; Woods, RE. Digital Image Processing. 2. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc; 2001.
25. Sertel O, Kong J, Shimada H, Çatalyürek ÜV, Saltz JH, Gurcan MN. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. Pattern Recognition. 2009; 42(6):1093–1103. [PubMed: 20161324]
26. Chun YD, Kim NC, Jang IH. "Content-based image retrieval using multiresolution color and texture features," Multimedia. IEEE Transactions. 2008; 10(6):1073–1084.
27. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on. 1973; 3(6):610–621.
28. Ballerini L, Li X, Fisher RB, Aldridge B, Rees J. Content-based image retrieval of skin lesions by evolutionary feature synthesis. EvoApplications. 2010; (1):312–319.
29. Rahman, MM.; Antani, SK.; Thoma, GR. A classification-driven similarity matching framework for retrieval of biomedical images. Proceedings of the international conference on Multimedia information retrieval; New York, NY, USA: ACM; 2010. p. 147-154.
30. Unser M. Sum and difference histograms for texture classification. IEEE Trans Pattern Anal Mach Intell. Jan.1986 8:118–125. [PubMed: 21869331]
31. Lee DD, Seung HS. Learning the parts of objects by nonnegative matrix factorization. Nature. 1999; 401(6755):788–91. [PubMed: 10548103]
32. Soukup D, Bajla I. Robust object recognition under partial occlusions using nmf. Computational Intelligence and Neuroscience. 2008; 2008
33. Lin C-J. Projected gradient methods for non-negative matrix factorization. Neural Computation. 2007; 19:2756–2779. [PubMed: 17716011]
34. Hsu C-W, Lin C-J. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks. 2002; 13(2):415–425. [PubMed: 18244442]

35. Chang, C-C.; Lin, C-J. LIBSVM: a library for support vector machines. 2001. software available at http://www.csie.ntu.edu.tw/cjlin/libsvm

36. Aizawa A. “An information-theoretic perspective of tf idf measures,” Inf. Process Manage. Jan. 2003 39:45–65.

37. Kowalski, G. Information Retrieval Architecture and Algorithms. Springer; US: Indexing; p. 95-139.

38. Jones KS. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation. 1972; 28:11–21.

39. Elkan, C. Deriving tf-idf as a fisher kernel. 12th International Conference on String Processing and Information Retrieval (SPIRE; 2005. p. 296-301.

40. Metzler, D. Generalized inverse document frequency. Proceeding of the 17th ACM conference on Information and knowledge management, ser. CIKM ’08; 2008. p. 399-408.

41. Sertel O, Kong J, Catalyurek U, Lozanski G, Saltz J, Gurcan M. Histopathological Image Analysis Using Model-Based Intermediate Representations and Color Texture: Follicular Lymphoma Grading. Journal of Signal Processing Systems. Apr; 2009 55(1):169–183.

42. Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell. Jul.2002 24:971–987.

43. Luo J, Crandall D. Color Object Detection Using Spatial-Color Joint Probability Functions. Image Processing, IEEE Transactions on. 2006; 15(6):1443–1453.

44. Liu GH, Zhang L, Hou YK, Li ZY, Yang JY. Image retrieval based on multi-texton histogram. Pattern Recogn. Jul.2010 43:2380–2389.

Biographies

Hatice Cinar Akakin received her PhD degree in Electrical and Electronics Engineering Department of Bogazici University, Istanbul, Turkey in 2010. She was a teaching and research assistant at Bogazici University, Signal and Image Processing Laboratory (BUSIM) from 2004 to 2010. During her PhD, she got experience in developing algorithms for face image analysis from low level image processing to high level interpretations. She is currently working as a post-doctoral researcher at The Ohio State University, Columbus, OH. She is also an academic staff of the Electrical and Electronics Engineering Department at Anadolu University, Eskisehir, Turkey, as a research assistant. Dr. Cinar Akakin’s research interests are in the areas of computer vision, image and video analysis and machine learning.

Metin N. Gurcan is an Associate Professor of Biomedical Informatics at the Ohio State University and the director of the Clinical Image Analysis Lab. Dr. Gurcan received his BSc. and Ph.D. degrees in Electrical and Electronics Engineering from Bilkent University, Turkey and his MSc. Degree in Digital Systems Engineering from the University of Manchester Institute of Science and Technology, England. From 1999 to 2001, he was a postdoctoral research fellow and later a research investigator in the Department of Radiology at the University of Michigan, Ann Arbor. Prior to joining the Ohio State University in October 2005, he worked as a senior researcher and product director at a hightech company, specializing in computer-aided detection and diagnosis of cancer from radiological images. He is the author of over 75 peer-reviewed publications and has two patents in computer-aided diagnosis in volumetric imagery. Dr. Gurcan is the recipient of the British Foreign and Commonwealth Organization Award, NCI caBIG Embodying the Vision Award, Childrens Neuroblastoma Cancer Foundation Young Investigator Award, The OSU Cancer Center REAP Award, and is a senior of member of IEEE, SPIE and RSNA.

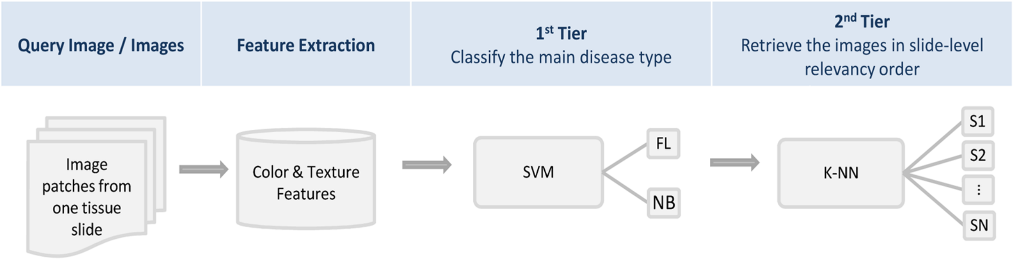
Query Image / Images
Feature Extraction
1st Tier
Classify the main disease type
2nd Tier
Retrieve the images in slide-level relevancy order
Image patches from one tissue slide
Color & Texture Features
SVM
FL
NB
K-NN
S1
S2
⋮
SN

Fig. 1.

The general flowchart for the CBIR system for a given query image or images

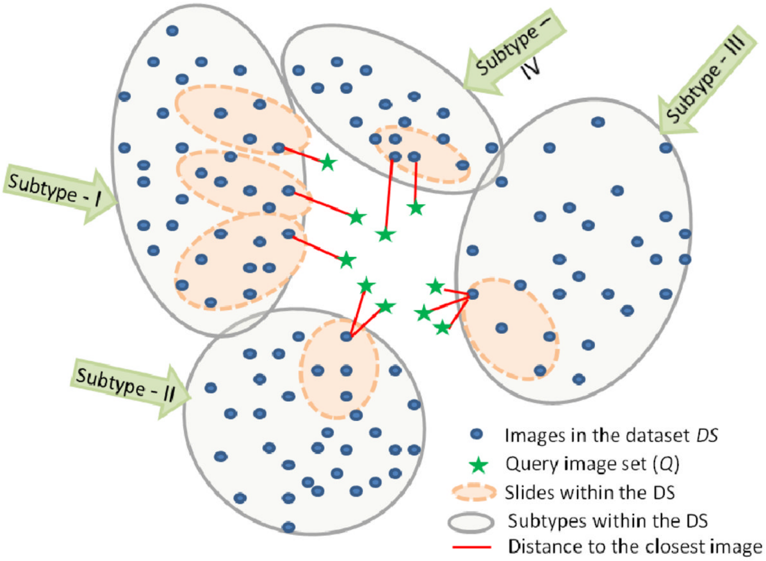

Fig. 2.

Sample image-level nearest neighbor search scheme for a given query image-set

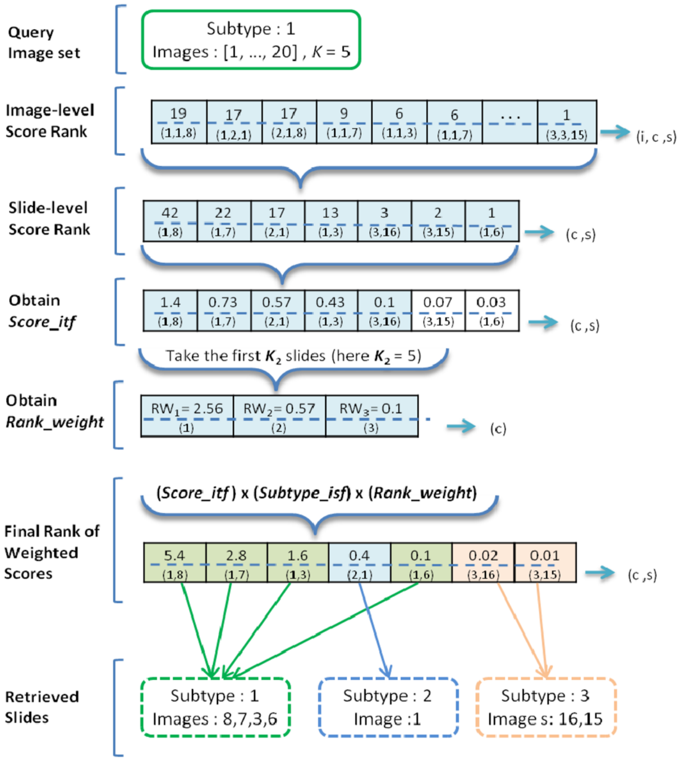

Fig. 3.

Computational model representing the transition from image level-scores to slide-level retrieval, where i =image number, c = subtype number and s = slide number. Here, the query Q is an image set with 20 images belonging to subtype 1. Image-level scores, slide-level scores, *Score_itf*, *Rank_weight*, relevancy rank of slides with weighted scores are computed respectively for the given sample query.

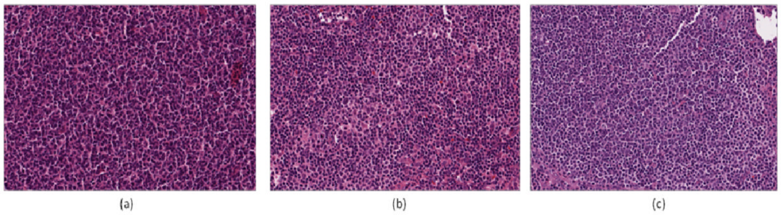

Fig. 4.

Sample H&E stained FL images associated with the three grades. (a) Grade - I (b) Grade - II and (c) Grade - III.

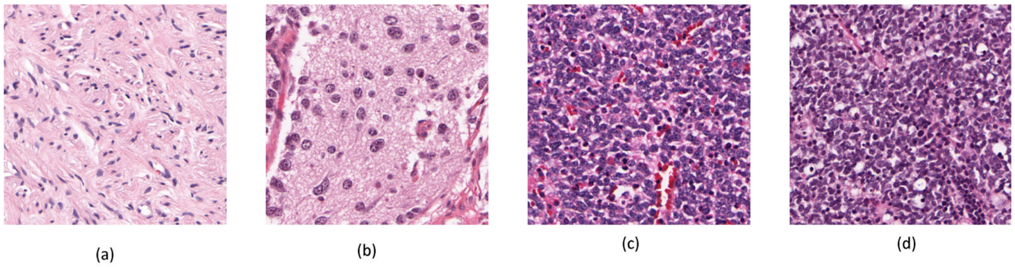

Fig. 5.
Sample NB images associated with the four grades. (a) SR, (b) D (c) PD and (d) UD.

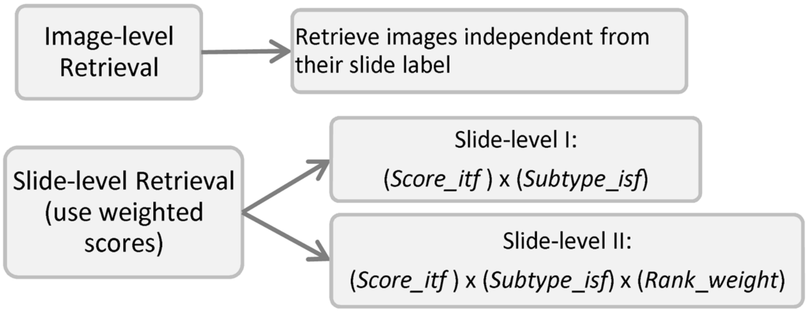
Image-level Retrieval
Retrieve images independent from their slide label
Slide-level Retrieval (use weighted scores)
Slide-level I:
(*Score_itf*) x (*Subtype_isf*)
Slide-level II:
(*Score_itf*) x (*Subtype_isf*) x (*Rank_weight*)

Fig. 6.
The second tier experimental scheme

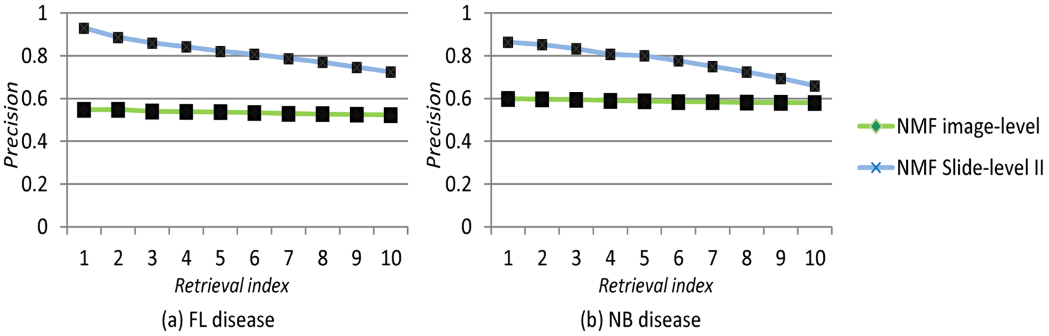

Fig. 7.

Comparison of average precision values for Slide-level II and image-level retrieval algorithms for FL and NB diseases.

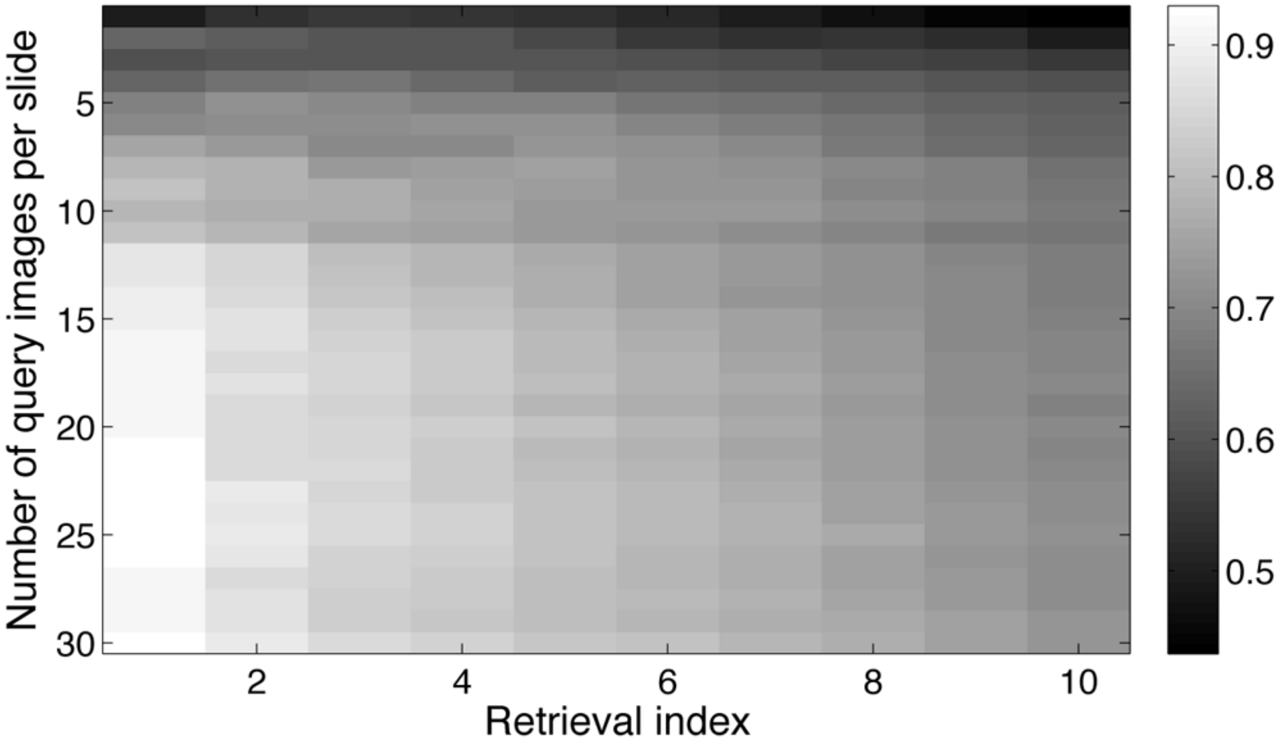

Fig. 8.

Analysis of number of query images per slide with top rank retrieval indices in terms of precision values. The precision values are represented as a gray level image.

TABLE I

Distribution of main diseases and their subtypes in the database

total # of images	**Follicular Lymphoma (943)**			
Subtypes	Grade I (269)	Grade II (372)	Grade III (302)	
Number of slides (Patients)	15	20	22	
total # of images	Neuroblastoma (723)			
Subtypes	Stroma Rich (174)	Stroma Poor (549)		
		Differentiating (163)	Poorly Differentiated (193)	Undifferentiated(193)
Number of slides (Patients)	13	8	12	11

TABLE II

Average classification results (%) for the 1^{st} tier over 50 test repetitions

	FL - image	NB - image	FL - slide	NB - slide
FL	99.4	0.6	100	0
NB	5.3	94.7	2.3	97.7

TABLE III

AUPRC statistics (mean ± standard deviation) for FL cases

	NMF Features			NF Features		
FL	**Image-level**	**Slide-level I**	**Slide-level II**	**Image-level**	**Slide-level I**	**Slide-level II**
Grade I	0.24 ± 0.15	0.40 ± 0.18	0.45 ± 0.23	0.23± 0.13	0.39 ± 0.18	0.42 ± 0.25
Grade II	0.38 ± 0.24	0.53 ± 0.15	0.64 ± 0.18	0.37 ± 0.23	0.50 ± 0.14	0.62 ± 0.19
Grade III	0.26 ± 0.15	0.42 ± 0.19	0.49 ± 0.25	0.25 ± 0.14	0.41 ± 0.19	0.48 ± 0.27
Average	0.30 ± 0.19	0.45 ± 0.18	0.54 ± 0.24	0.29 ± 0.18	0.44 ± 0.18	0.51 ± 0.25

TABLE IV

Confusion matrix for FL cases with top rank retrieval indices and precision values

FL-NMF	Grade I				Grade II				Grade III			
Retrieval index	**1**	**3**	**5**	**10**	**1**	**3**	**5**	**10**	**1**	**3**	**5**	**10**
I	86.8	77.9	71	53.4	6.6	13.3	18.8	29.3	6.6	8.8	10.2	17.3
II	5	8.3	11	13.5	95	91.7	89	77.5	0	0	0	9
III	0	1.5	2.7	3.6	4.5	12.1	13.6	15.5	95.5	86.4	83.7	80.9

TABLE V

AUPRC statistics (mean ± standard deviation) for NB cases

	NMF Features			NF Features		
NB	**Image-level**	**Slide-level I**	**Slide-level II**	**Image-level**	**Slide-level I**	**Slide-level II**
D	0.45 ± 0.19	0.61 ± 0.26	0.67 ± 0.32	0.43 ± 0.17	0.62 ± 0.26	0.66 ± 0.30
PD	0.37 ± 0.16	0.57 ± 0.13	0.69 ± 0.15	0.36 ± 0.15	0.50 ± 0.10	0.56 ± 0.13
SR	0.38 ± 0.27	0.81 ± 0.15	0.90 ± 0.13	0.40 ± 0.26	0.75 ± 0.16	0.81 ± 0.14
UD	0.39 ± 0.22	0.51 ± 0.24	0.59 ± 0.28	0.42 ± 0.20	0.48 ± 0.22	0.53 ± 0.24
Average	0.39 ± 0.21	0.63 ± 0.22	0.72 ± 0.25	0.40 ± 0.20	0.59 ± 0.21	0.64 ± 0.23

TABLE VI

Confusion matrix for NB cases with top rank retrieval indices and precision values

NB-NMF	D				PD				SR				UD			
Retrieval index	**1**	**3**	**5**	**10**	**1**	**3**	**5**	**10**	**1**	**3**	**5**	**10**	**1**	**3**	**5**	**10**
D	75	70.8	67.5	43.8	0	4.2	12.5	33.7	12.5	12.5	12.5	17.5	12.5	12.5	7.5	5
PD	0	0	0	0.8	100	88.9	85	64.2	0	0	0	0	0	11.1	15	35
SR	0	2.5	6.1	10	0	0	0	0	100	97.5	93.9	90	0	0	0	0
UD	9	3	3.6	16.4	27.3	27.3	29.1	28.1	0	0	0	0	63.7	69.7	67.3	55.5

Algorithm 1

Image-level Retrieval

```
for for a given query image-set $Q$, with $K$ retrieval do
  $Image\_Score\ (1 \ldots T) = [0 \ldots 0]$
  // Initially score of each image in the corresponding dataset equals to zero.
  for $n = 1$ to $N$ do
    for $t = 1$ to $T$ do
      $distance(n, t) = Dist(F_{Q^n}, F_{DS^t})$
    end for
    $ind = sort(distance)$ in descending order,
    Retrieve and display the $K$-closest images to the user
    and
    $Image\_Score(ind[1 : K]) = Image\_Score(ind[1 : K]) + 1$
    // Add 1 to the score of $K$-nearest images which has the smallest distance from the corresponding dataset to the query image
  end for
  // Scores are accumulated if $N > 1$
end for
```

Algorithm 2

Slide-level Retrieval

```
  for for a given query image-set $Q$, with $K$ image-level retrieval do
    Perform Algorithm 1
    // Do not display the retrieved images to the user
  end for
  for $c$ = 1 to $C$ do
    for $s$ = 1 to $S^c$ do
          $Score\_itf^{(c,s)} = \sum_{i=1}^{I^{(c,s)}} Image\_Score_i^{c,s} / I^{(c,s)}$
    end for
  end for
  // Sort $Score\_itf$ in descending order and select top $K2$ to compute $Rank\_weight$
for $c$ = 1 to $C$ do
          $Rank\_weight^{(c)} = \sum_{k=1}^{K2} Sorted\_Score\_itf^{(c,k)}$
  end for
  for $c$ = 1 to $C$ do
    for $s$ = 1 to $S^c$ do
          $Weight\_Score^{(c,s)} = Score\_itf^{(c,s)} *$
          $Subtype\_isf^{c} * Rank\_weight^{(c)}$
    end for
  end for
  // Sort the weighted scores in descending order
  // Display the user $n$-highest scored slides
```