Journal of Biomedical Informatics 44 (2011) 519–528

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Content-based histopathology image retrieval using a kernel-based semantic annotation framework

Juan C. Caicedo [a], Fabio A. González [a,*], Eduardo Romero [b]

[a] Computer Systems and Industrial Engineering Department, National University of Colombia, Bogotá, Colombia
[b] Diagnostic Imaging Department, Medicine School, National University of Colombia, Bogotá, Colombia

ARTICLE INFO

Article history:
Received 27 January 2010
Available online 3 February 2011

Keywords:
Auto-annotation
Biomedical images
Histology
Histopathology
Image retrieval
Kernels
Kernel alignment

ABSTRACT

Large amounts of histology images are captured and archived in pathology departments due to the ever expanding use of digital microscopy. The ability to manage and access these collections of digital images is regarded as a key component of next generation medical imaging systems. This paper addresses the problem of retrieving histopathology images from a large collection using an example image as query. The proposed approach automatically annotates the images in the collection, as well as the query images, with high-level semantic concepts. This semantic representation delivers an improved retrieval performance providing more meaningful results. We model the problem of automatic image annotation using kernel methods, resulting in a unified framework that includes: (1) multiple features for image representation, (2) a feature integration and selection mechanism (3) and an automatic semantic image annotation strategy. An extensive experimental evaluation demonstrated the effectiveness of the proposed framework to build meaningful image representations for learning and useful semantic annotations for image retrieval.

1. Introduction

The use of digital imaging in histopathology has been rapidly rising during the last few years [1]. Pathology departments using digital microscopy equipments can share slides without sending the glass, and can include images in electronic reports and publications [2]. Then, large amounts of digital histopathology images are constantly acquired as part of the routine operation in these specialized centers. Image collections are stored using information technologies such as Picture Archiving and Communication Systems (PACS), but they remain archived in the long term, basically because after some time they result useless, since the main actual exploit is the one associated to the specific clinical case. Nevertheless, these large image collections are a potential source of information and knowledge, which may support educational activities, research studies and even the clinical decision making process itself, if the right tools to access these collections are developed [3].

Accessing a collection of histology images can be done using different query paradigms. For instance, using structured queries in conventional databases, using keywords in a text retrieval engine or using example images in a content-based image retrieval system [4]. In this paper, we consider the problem of retrieving histopathology images from the collection using example images as queries, that is, the user presents reference images to the system, and the system uses the visual content to match similar images in the collection. Using visual contents to search for images is considered a beneficial technology for next generation medical imaging systems [3], and is also considered one of the major challenges in image retrieval research. The problem underneath an image retrieval system is the mechanism for identifying relevant images, which is mostly a similarity measure between image contents.

Different similarity measures have been proposed and studied for medical image retrieval using low-level features, which focus mainly on characterizing visual properties that can be computed from pixels [5,6]. However, bridging the semantic gap [7] has become the main focus of image retrieval research, i.e., reducing the discrepancy between the information extracted by low-level features and the high level interpretations of human beings on the same images. This research has led to semantic representations of histology images to address the problem of identifying semantically related images rather than just visually similar images [8]. These strategies aim to provide better search results which are more likely to match contents in the same way as physicians would do.

This paper presents a framework to archive and retrieve histopathology images by content. To overcome the problem of delivering semantically valid images for a medical task, we propose an automatic image annotation framework that recognizes high-level

* Corresponding author. Address: Computer Systems and Industrial Engineering Department, National University of Colombia, Cra 30 45-03, Ciudad Universitaria, Edif. 453, Of. 114 Bogotá, Colombia.
E-mail addresses: jccaicedoru@unal.edu.co (J.C. Caicedo), fagonzalezo@unal.edu.co (F.A. González), edromero@unal.edu.co (E. Romero).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.01.011

concepts after analyzing image visual contents. The main contribution of this work is a strategy to generate multi-feature image representations for the automatic recognition of histopathology concepts. This strategy has been designed as a unified framework based on kernel methods theory, and includes three main aspects for semantic image content recognition: (1) multiple visual features to represent histology image contents (2) appropriate kernel functions to harness the structure of the input data, and (3) the optimal combination of multiple kernel functions according to the underlying image semantics. Kernel functions are fused using a weighted linear combination, whose weights are found by an optimization process that maximizes the correlation between low-level features, represented by kernel functions, and high-level semantic concepts.

The proposed strategy has been implemented and evaluated using a large database of real histopathology images, extracted from medical records of a pathology lab. An extensive validation was conducted using the ground truth provided by pathologists. The experimental evaluation showed that the semantic image annotation leads to an average improvement in the retrieval response of 57% when it is compared to visual search using only low-level features. Also, the results show that a multi-feature representation for visual contents can be progressively improved by operating kernel functions. We found that modeling feature structure and non-linear patterns with kernel functions is more likely to improve the discriminative power of multi-feature representation spaces. The contents of this paper are organized as follows: Section 2 presents a review of previous works related to histology image retrieval. Section 3 introduces the collection of histopathology images used in this study. Section 4 describes the proposed methods for automatic image annotation, based on kernel methods. The experimental setup and results are presented in Section 5. Finally, Section 6 presents the concluding remarks and future work.

2. Related work

Digital microscopy is a very broad field of active research, that ranges from image acquisition and compression [9] to automatic disease detection [10]. Content-based retrieval of microscopy and histology images is one of these areas of research that is receiving increasing attention from researchers. Image retrieval focuses on methods and tools for managing large collections of digital slides, providing effective access to all the available information, in contrast to other research areas that focus on processing individual images for making automatic decisions, such as automated grading [11,12] or tissue classification [13,14].

Image retrieval on pathology image collections was approached by Zheng et al. [15] using low-level features. Four different visual features were studied to measure the discriminative power of similarity measures to correctly identify relevant images given an example query. They reported a correlation between the computed similarity and pathological significance on the tested collection, without the use of domain knowledge. However, to scale up the system performance, low-level features may be insufficient. Tang et al. [8] investigated the role of semantic information to represent local image content in gastro-intestinal tissue images. Their method aim to assign a semantic annotation to each region on the image using machine learning algorithms. The main disadvantage of this approach is the need for manual annotations made on specific regions for a large enough sample of training images. Naik et al. [16] also approached the problem of histology image retrieval using semantic knowledge. They used multiple texture and architectural features of tissues, and employed a boosting algorithm to identify feature weights that maximizes retrieval and classification performance.

In this paper we address the problem of semantic image retrieval for histopathology images, using an automatic annotation strategy. Our previous work on histopathology image retrieval [17] showed the potential of using semantic features to represent image contents. In this paper we extend that work by generalizing the representation of visual contents in a set of multiple heterogeneous features rather than a unique feature vector. Also, we recast the image annotation problem in terms of kernel methods for image representation, feature selection and concept detection, as is presented in the following Sections.

3. Basal-cell carcinoma images

Images in this work have been used to diagnose a special kind of skin cancer known as basal-cell carcinoma. Basal-cell carcinoma is the most common skin disease in white populations and its incidence is growing world wide [18]. The histopathology collection is composed of 1502 images at 1280×1024 pixels, acquired under a Nikon microscope and stored in lossless JPG format.

The collection was studied and annotated by a pathologist to describe its contents, elaborating a data set with images and descriptions of their related concepts. Table 1 shows the list of 18 concepts and the number of available examples in the collection. One image may contain several concepts, that is, different biological structures are exhibited in one single image. Notice that Table 1 lists the number of images per concept, but not their co-occurrences. The total number of annotated images in the collection is about 900 corresponding to pathological cases, while the remaining 600 are images with normal skin tissue. This data set also shows a high imbalance between the number of examples exhibiting a concept and the rest of the collection.

The concept list also includes some structures that are not pathological such as *pilosebaceous units*, *eccrine glands* and *blood vessels*. One of the histopathology concepts reported in Table 1 is N–P–C, which is a convention for *Nodule, Palisading cells and Clefts* (N–P–C), which is a typical sign of basal cell carcinoma, not by the presence of any of them individually but by the manifestation of all three visual patterns together.

Fig. 1 shows examples of histopathology images with some specific regions in which the concepts can be observed. These examples show that one image can have more than one interesting pattern for pathologists. In addition, the Figure shows that normal biological structures have well-defined visual configurations, in contrast to pathological patterns that may appear with different visual variabilities. In particular, note that the dotted lines (green)

Table 1
Histopathology concepts and the corresponding number of examples in the data set.

Concept	Training	Test	Total
Blood vessel	96	26	122
Cystic change	46	21	67
Eccrine glands	100	48	148
Elastosis	92	33	125
Fibrosis	67	23	90
Lymphocyte inf.	101	39	140
Micronodules	35	6	41
Morpheaform pattern	29	8	37
N–P–C, elastosis	40	12	52
N–P–C, fibrosis	34	16	50
N–P–C, infiltration	133	45	178
N–P–C, pilosebaceous	27	11	38
N–P–C, trabeculae	10	4	14
Necrosis	27	5	32
Perineural invasion	5	1	6
Pilosebaceous unit	119	35	154
Thick trabeculae	45	15	60
Ulceration	10	5	15

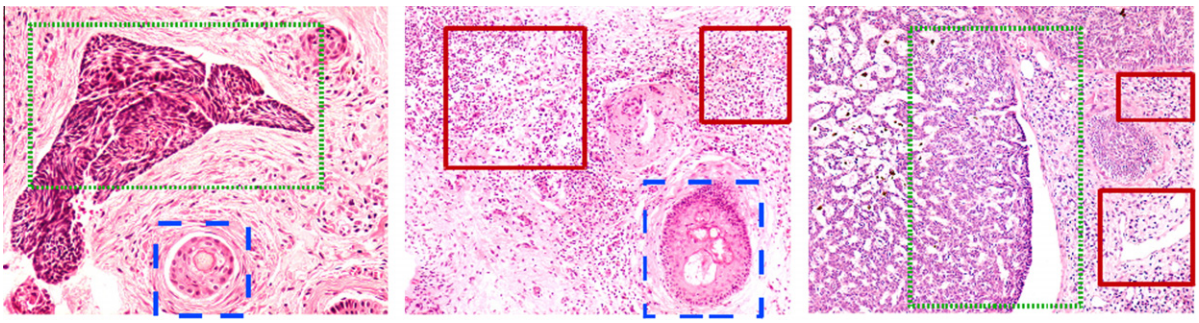


Fig. 1. Three histopathology image examples with some highlighted biological structures and pathological patterns. Dashed lines (blue) show pilocebaseus units, a normal biological structure in the skin. Dotted lines (green) show example regions of Nodule, Palisading cells and Clefts (NPC), a clear evidence of basal-cell carcinoma. Continuous lines (red) show regions with another clue to detect basal-cell carcinoma: lymphocyte infiltration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cover a portion of nodule that contrast with epithelial tissue with a cleft in the middle. However, the nodule structure in both cases looks different and also the contrasting tissue in the other side of the cleft. These are some examples of the variabilities that may be found in real histopathology images.

This data set was divided up into training (75%) and test (25%) sets, using stratified sampling as is shown in the table. The annotations provided by the pathologists are useful to automatically validate whether search results are relevant to the user information needs. This image collection has been previously used to test two different retrieval strategies: one that uses only visual similarity [19] and another that uses a semantic representation approach based on SVM classifiers with basic kernels [17].

4. Semantic image annotation and retrieval

The proposed strategy for histology image retrieval is oriented to produce a set of semantic image annotations through visual content analysis. Our framework aims to build a general and complete visual representation of images that can provide enough evidence of the presence or absence of certain histopathology concepts. Fig. 2 shows the three fundamental steps in our framework: first, the extraction of multiple visual features is performed on the input images. Second, the new content representation is build integrating all visual features using kernel functions. Third, this content representation is used to detect histopathology concepts. After generating automatic annotations, the result can be used to search images with similar annotations or just to index the input images in the retrieval system.

4.1. Image features

Feature extraction is an important task for image analysis and understanding and there are different approaches to address this problem [20]. Global features for characterizing whole scenes have been proposed using color histograms [21] and MPEG7 features [22]. Likewise, global descriptors such as textures and down-scale representations have been evaluated in medical imaging [23]. One important advantage of using a global image description strategy is that it is unnecessary to specify a model for objects or regions that images may contain. On the contrary, global features provide a holistic image representation that characterizes the composition of the whole image.

We modeled histopathology images as a set of global histogram features, taking into account that pathology patterns may have high visual variabilities that are characterized by different feature sets. For instance, as it is shown in Fig. 1, the NPC pattern, highlighted in dotted lines (green), contains a mixture of the textures inside the nodule, the edges provided by the cleft, and the density of the palisading cells. To describe the different visual characteristics of histopathology patterns, seven feature spaces have been selected: gray scale histogram, invariant feature histogram [24], local binary patterns [25], RGB color histogram [24], bag of SIFT features, Sobel histogram [26] and Tamura texture histogram [27].

These seven low-level features are complementary with respect to the kind of measure they do over image pixels, since they apply different computations to build the histograms. However, some of them measure similar visual properties on images, such as Tamura texture and local binary patterns, both modeling texture patterns, but using different approaches (statistical and deterministic, respectively). Also, the invariant feature histogram and SIFT features are intended to identify characteristics that are invariant to rotations and translations. The invariant feature histogram follows an integral approach that sums globally over rotation and create a histogram over translation [24]. On the other hand, the bag of SIFT features is based on a learned dictionary of rotation invariant visual patterns that are counted in each image to construct a histogram of frequencies [28].

The global features were chosen to create a general and broad repertory, in contrast to approaches that carefully select visual features for the specific problem at hand. The relevant features for

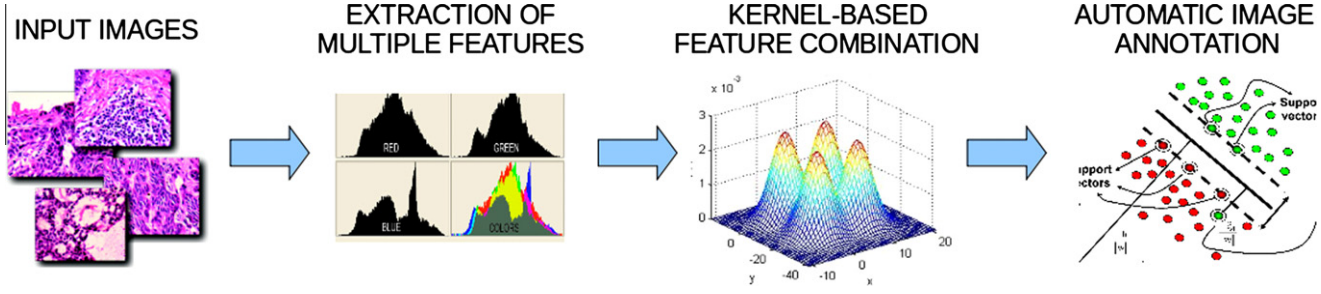


Fig. 2. Overview of the main steps of the proposed strategy for automatic image annotation.

each high-level concept are automatically chosen by a feature fusion process to be described below. All histogram features are global content descriptors that do not allow to identify spatial location of objects or patterns. This represents an advantage when dealing with histopathology images, in which pathology patterns are spread around the image such as lymphocyte infiltration, in which lymphocytes can be seen covering different tissue regions. Elastosis is another example of a stroma's property that can be seen along the complete tissue slide. In that sense, the set of histograms provides different measures to detect variations in the global image composition, that can be exploited to reveal its semantic meaning.

4.2. Kernel functions

Kernel methods are an alternative family of algorithms and strategies to perform machine learning [29]. One of the main distinctive characteristics of kernel methods is that they do not emphasize the representation of objects as feature vectors. Instead, objects are characterized implicitly by kernel functions that measure the similarity between two objects. A kernel function induces an implicit high-dimensional feature space where, in principle, it is easier to find patterns.

Informally, a kernel function measures the similarity of two objects. Formally, a kernel function, $k : X \times X \rightarrow \mathbb{R}$, maps pairs (x,z) from a set of objects X, the problem space, to the real space. A kernel function implicitly generates a map, $\Phi{:}X \rightarrow F$, where F corresponds to a Hilbert space, called the feature space. The dot product in F is calculated by k, specifically $k(x,z) = \langle \Phi(x), \Phi(z) \rangle_F$.

One can deal with histograms as simple data vectors, regardless their probability distribution properties. In that sense, we can calculate the dot product between histograms treating them as high dimensional feature vectors. This operation will be herein denoted as the identity kernel, since it induces a feature space that is equivalent to the input space. On the other hand, we can harness the structure of histogram data by evaluating the similarity measure between two histograms in a more meaningful way. The histogram intersection is a similarity function devised to calculate the common area between histograms as follows:

$$k_{\cap}(A,B) = \sum_{i=0}^{m} min(a_i, b_i) \tag{1}$$

where $A = (a_1 \cdots a_n)$ and $B = (b_1 \cdots b_n)$ are histograms. This similarity measure has been shown to satisfy the Mercer's properties [30]. This is important when using learning methods, such as SVM, since it guarantees the optimal solution of the associated convex optimization problem. Another advantage of this kernel is that it can be efficiently computed; in fact, Maji et al. [31] recently proposed a very efficient technique to train SVM that use the histogram intersection kernel.

Using the histogram intersection kernel with SVM, we are modeling a non-linear classification rule in a high-dimensional feature space [32]. This particular property of kernel method solutions, allows us to capture the high variability of visual patterns along the same semantic concept. This special property will be discussed in the next Subsection.

4.3. Combination of kernels

As discussed before, pathology concepts are characterized by different types of features including colors, textures and edges. Given two images, a similarity measure may be calculated by applying a kernel function to a pair of images represented by a particular type of feature histogram. For instance, when using the Gray Histogram, we can distinguish if an image has the same brightness level as another one, while using local binary patterns, we can evaluate if they have similar low-level tissue composition. This provides a repertory of kernels that compare images according to different visual properties. Now, we want to equip the classification system with the ability to adjust the importance of each feature when dealing with a particular semantic concept.

Formally, there is a set of kernels $\{k_i : X \times X \rightarrow \mathbb{R}\}_i$, where i indicates the type of visual features used to calculate the similarity. Notice that despite the fact that the different kernels use different features to calculate the similarity, all of them have the same domain, i.e., they are image kernels. The problem is how to use these different image kernels to calculate an overall similarity measure for images. The new similarity measure would correspond to a kernel function k_α that induces a new image representation space. k_α is defined as a linear combination of the n individual histogram kernels:

$$k_\alpha(x,z) = \sum_{i=1}^{n} \alpha_i k_i(x,z) \tag{2}$$

The weights α_i allow to parameterize the kernel giving higher or lower importance to each individual feature. Notice that the linear combination of kernel functions, associated to different visual features, is implicitly defining a new feature space whose structure may be adapted to better recognize a particular semantic concept. In particular, it has been shown that a linear combination of two kernels is a valid kernel provided that the associated weights are all positive. In addition, the linear combination of two kernels leads to a new feature space that is isomorphic to the Cartesian product of the individual feature spaces [29].

The problem now is to find a vector of weights α that maximizes the performance of the kernel k_α in an image classification task. In the case of histopathology images, different concepts require different classifiers that emphasize the appropriate visual features. Herein we use the kernel alignment strategy [33] to build an adapted kernel function for each concept. Each adapted kernel function is expected to emphasize those visual features that allow to better recognize the presence (or absence) of the corresponding concept in a given image.

Kernel-target alignment [33] measures how appropriate a kernel function is for solving a specific classification problem. In particular, the alignment of two kernels with respect to a sample S, is defined as:

$$A_S(k_1,k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle}\sqrt{\langle K_2, K_2 \rangle_F}} \tag{3}$$

where k_1, k_2 are kernel functions; K_1, K_2 are matrices corresponding to the evaluation of the kernel functions on a sample S; and $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product defined as $\langle A, B \rangle_F = \sum_i \sum_j A_{ij} B_{ij}$.

Given the binary labels for a training set, in which *1* indicates the presence of one selected concept and -1 indicates the absence of that concept in the image, we can build a target function to optimize the kernel alignment measure. Defining $y{:}X \rightarrow \{-1,1\}$ as the binary label for an image in X, the problem space, the target kernel k^* is then defined as $k^*(x,z) = y(x)y(z)$. The target kernel k^* is the optimal kernel for solving the given classification task, since it explicitly reveals whether the objects x and z are in the same class or not. The goodness of a given kernel k is measured in terms of how much it aligns with the target kernel in a training sample. Formally this is expressed as

$$A_S^*(k) = A_S(k, k^*) \tag{4}$$

The problem of finding appropriate weights for k_α then becomes the problem of finding the weights α that maximize the target alignment $A_S^*(k_\alpha)$. In [34], this problem is solved by transforming it to an equivalent quadratic programming problem and is the strategy followed in this work. This kernel combination strategy is

in fact a type of feature fusion task, but performed at the kernel level, making it an integral part of the learning process. The main advantage is that features are optimally combined during the learning process depending on the particular type of classification problem to be solved.

After combining the basic kernel functions, we also composed the resulting kernel with a Radial Basis Function (RBF) to emphasize non-linear patterns in the representation space. Given the optimally combined kernel k^*_α, we use it to compute the RBF kernel as follows:

$$k_G(x,z) = exp\left(-\left(k^*_\alpha(x,x) + k^*_\alpha(z,z) - 2k^*_\alpha(x,z)\right)/2\sigma^2\right) \quad (5)$$

4.4. SVM classifiers

Support Vector Machines (SVM) are linear classifiers whose decision function is a hyperplane in the feature space. For each histopathology concept we have modeled a new feature space using adapted combinations of kernel functions, resulting in a new kernel function to classify images of that particular concept. Then, we train a SVM for each concept using the corresponding adapted kernel function. All trained SVM classifiers are then arranged in the semantic image annotator. Since SVM are linear classifiers in the feature space, and our feature space models non-linear relationships between images, the resulting classification rule is non-linear in the input space [32]. This enables the automatic annotation module to capture high visual variabilities among the same semantic concept.

4.5. Semantic image annotator

The goal of an image annotation module is to analyze the visual image contents to produce a semantic interpretation. This interpretation corresponds to the assignment of several semantic labels. The image annotation module is an arrangement of SVM classifiers that detects the presence of pre-defined semantic concepts in images. The aim is to identify which labels are more appropriate to describe an image according to its visual content.

Semantic annotations are built using SVM outputs, but, instead of using binary labels that indicate whether or not an image contains a concept, a degree of presence or absence is modeled for each possible concept. Each image is assigned to a semantic feature vector in $\mathbb{R}^n$, where n is the number of concepts. Each component of the semantic feature vector is generated by applying a sigmoid function to the output $v_i, i \in \{1 \cdots n\}$ of the corresponding SVM:

$$f(v_i) = \frac{1}{1 + e^{-a(v_i+b)}} \quad (6)$$

The shape of the function (a and b parameters) has an important repercussion on the sensitivity of the semantic annotation process. Specifically, the sigmoid function parameters affect the trade-off between precision and recall. To optimize the retrieval performance a set of parameters (a,b) may be set for each individual concept. For our study, we used a unique set of parameters that maximizes the global mean average precision on the training data, making a general balance among all concepts. It simplifies the procedure to find good candidates and reduces the number of parameters for the indexing method.

Finally, the semantic similarity of two images is calculated by applying the Tanimoto coefficient to the semantic feature vectors describing the images. Given two semantic vectors v and υ, the Tanimoto coefficient is defined as:

$$T(v,\upsilon) = \frac{v \cdot \upsilon}{\|v\|^2 + \|\upsilon\|^2 - v \cdot \upsilon} \quad (7)$$

The Tanimoto coefficient evaluates the degree of coincidence between two vectors, which, in this context, is related to the common concepts of the two images being compared.

5. Experimental evaluation

The experimental evaluation process presented in this Section has two main goals: first, to evaluate the performance of the proposed kernel-based annotation framework on real histopathology images, second, to determine the impact on the retrieval performance, when using semantic annotations instead of using only low-level visual features.

5.1. Feature combination

The first step in the proposed framework is to build a new image representation based on kernel functions. In this work, for each of the 18 histopathology concepts, a new kernel is adapted. The feature combination strategy is applied to each class, using a 10-fold cross validation on the training data set to estimate the parameters of the kernel alignment algorithm that optimize the discerning capacity of the feature space. Histogram features were normalized using norm $\ell_1 = 1$, which produces discrete probability distributions instead of frequency histograms, and make the set of features comparable during the combination process. Fig. 3 shows the list of histopathology concepts with the obtained weights for each feature. We evaluated two basic kernel functions, named identity kernel k_I and Histogram Intersection Kernel $k_\cap$. Notice how each kernel function emphasizes differently the set of features, indicating that the discriminative power of each descriptor changes according to the way in which it is used. Also, the optimization algorithm assigns different weights to each concept, varying the way in which features are combined. Each concept obtains a different weight adjustment, since the corresponding set of positive examples have different visual configurations.

The final row in the Figure presents the sum of all weights across different concepts, revealing the general preference that the optimization algorithm had in terms of feature selection. For the identity kernel, the algorithm selected the LBP features as the more discriminative ones, whereas for the Histogram Intersection Kernel, the algorithm preferred SIFT features. In general, the more important visual features for this discrimination task were textures (LBP, SIFT and TAM), which shows consistency with previous findings for histology image representation. Nevertheless, in most of the cases, even though features can be ranked in a preference order, histopathology concepts require a combination of several visual features. Notice that just in a few cases, the weights for certain features is zero, suggesting that multiple visual features are complementary for recognizing histopathology concepts.

5.2. Automatic image annotation

The semantic image annotator is composed of 18 binary SVM classifiers that evaluate image contents under a kernel-based framework. The classification strategy is one-against-all, i.e., each classifier is learned independently of the others. It is specially useful since each image can be annotated using multiple labels. The classification module is first trained using 10-fold cross validation to estimate good parameters for each classifier. The parameter is chosen to maximize the f-measure per class, since we want to correctly annotate as many images as possible with high precision. Reported performance measures are precision, recall, *f*-measure and average-accuracy. The latter was computed by averaging the accuracies of the positive class and the negative class for each binary classification problem. In addition, reported measures are

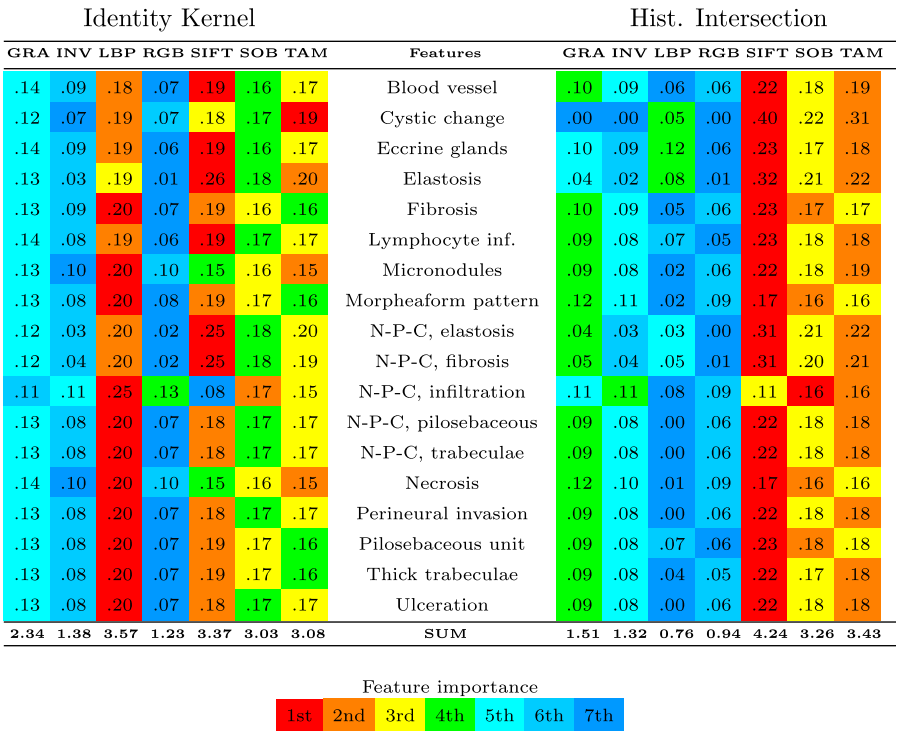


Fig. 3. Heat maps of weights assigned to features.

weighted-average scores among all classes according to the number of images in each class.

The experimentation includes the evaluation of two different strategies for building kernel functions: a direct combination of kernels adding functions with equal weights and an optimal combination of kernels using the kernel-target alignment framework. Each strategy evaluates four kernel functions as well: the identity kernel, the Histogram Intersection Kernel and the composition of these two kernels with the RBF. Experimental results are presented in Table 2 showing the performance measures of all evaluated strategies. The best overall performance is obtained by the optimal combination of features in terms of precision and average-accuracy while recall and F1 are better under the simple combination strategy. This same tendency can be observed when comparing the Histogram Intersection Kernel and the identity kernel. This basically means that the former discriminates more accurately while the latter annotates more correct images. Notice that the identity kernel deals with features as simple vectors whereas the Histogram Intersection Kernel exploits the structure of histogram data. On the other hand, RBF kernels show a considerably higher recall and better F1 and accuracy values, indicating the effectiveness of the RBF to highlight non-linear patterns in the feature space.

Table 2
Classification results on the test data set using different kernel functions.

Kernel function		Precision	Recall	F1	Accuracy
Direct feature combination					
k_I	Identity kernel	0.575	0.377	0.455	0.637
$k_\cap$	Hist. intersection	0.762	0.351	0.481	0.637
k_G	RBF kernel	0.444	0.615	0.516	0.720
$k_{G\circ\cap}$	RBF ∘ intersection	0.662	0.555	0.604	0.726
Optimal feature combination					
k_I^*	Identity kernel	0.567	0.382	0.457	0.643
$k_\cap^*$	Hist. intersection	0.771	0.365	0.496	0.645
k_G^*	RBF Kernel	0.442	0.609	0.512	0.714
$k_{G\circ\cap}^*$	RBF ∘ intersection	0.656	0.547	0.596	0.735

The average precision of the best model ($k_{G\circ\cap}^*$) is 66% and its recall is about 55%. There are different challenges to effectively recognize histopathology concepts in images such as the class imbalance, in which the reduced number of samples for a particular class, make it difficult to recognize a positive example among hundreds of negative ones. In addition, the high intra-class variability of images and subtle inter-class differences are also difficult to model, even when using multiple features.

Figures in Table 2 give an estimate of classification performance on new, unseen histopathology images. To evaluate the significance of differences between classification rates, we employed a McNemar's test [35] on the same test data. Table 3 presents the results for a number of tests comparing the performance of classifiers. We trained 18 classifiers, one per histopathology concept, with each kernel function. Then, we ran pairwise tests and account for the number of classifiers that are statistically superior and the number of classifiers in which other kernel function is better. We call it wins and losses in Table 3, and the difference is computed to determine which kernel provides more advantages to recognize histopathology concepts. It shows that the optimally combined Histogram Intersection Kernel, composed with RBF ($k_{G\circ\cap}^*$), has the largest number of significantly better classifiers with respect to the other kernel functions. Notice that kernels composed with RBF have a positive difference whereas simple kernels accumulate a negative difference in performance, suggesting that the RBF is an

Table 3
Evaluation of different kernels using McNemar's test. Eighteen classifiers are trained for each kernel, one per concept. A classifier based on a given kernel for a particular concept is compared against all other classifiers for the same concept. Cell numbers indicate the number of times that a classifier based on a particular kernel is significantly better or worse than classifiers based on other kernels.

Kernel	k_I	$k_\cap$	k_G	$k_{G\circ\cap}$	k_I^*	$k_\cap^*$	k_G^*	$k_{G\circ\cap}^*$
Wins	4	8	19	16	5	8	19	17
Losses	18	17	8	6	17	17	8	5
Difference	−14	−9	11	10	−12	−9	11	12

important factor to improve classification performance. In addition, it can be observed that the Histogram Intersection Kernel gets a higher score with respect to the identity kernel, as well as the aligned kernels with respect to non-aligned kernels.

Our experiments aimed to evaluate differences between classifiers that use different image representations, which are built by modeling high-dimensional feature spaces using kernel functions. Experimental results show that image representations using the Histogram Intersection Kernel provide a better performance than the identity kernel, mainly due to the way in which the former uses structured data. Also, the RBF kernel shows important performance improvements by highlighting non-linear patterns in the feature space. Finally, the optimal combination of kernels shows improvements in terms of absolute performance, even though these results do not provide enough evidence of significant differences.

5.3. Image retrieval

To evaluate the performance of the retrieval module, images in the test set are used as queries following a leave-one-out strategy, which amounts to approximately 520 different queries. Standard performance measures are used to evaluate the system response including mean average Precision (maPrec), precision at position k ($P(n = k)$), recall at position k ($R(n = k)$), and recall vs. precision plots [36]. The maPrec value is computed using the images that the algorithm retrieves until every relevant image has been found, i.e., until a 100% of recall is met. Reported values are the average results for the 520 test queries. The evaluation of the image retrieval system covers two main strategies to search for similar images: using low-level visual features and using semantic annotations.

5.3.1. Visual retrieval performance

A baseline model using similarity functions for low-level image features is included to compare experimental results. The model based on low-level features calculates the similarity between histograms to produce an image ranking using the Histogram Intersection Kernel as similarity measure [19]. Table 4 presents performance measures to compare the response of low-level features, in which SIFT features and Sobel histogram offer the better response. The Bag of SIFT features, that showed the better performance in terms of maPrec, has an important advantage with respect to the other set of visual features: it is based on a learned dictionary of visual patterns extracted from the whole collection, and accounts for an orderless representation of visual patterns in images. Then, these features provide invariance to both, rotation (given by the SIFT descriptor) and translation (given by the orderless spatial arrangement of the bag of features). The invariant feature histogram is also invariant to rotation and translation, however,it takes a geometric approach based on single image analysis. The strength of the bag of SIFT features resides on the collection-based dictionary construction as opposed to the single image analysis of the other set of features.

Nevertheless, the precision of all visual features decreases very fast as they return more images. This can be observed in the Table by comparing the precision at 1, $P(1)$, with respect to precision at 100, $P(100)$, i.e., the variation in precision along the first 100 results. None of the models can maintain a precision higher than 20%, which means that, in a first page showing 100 results, less than 20 images would be relevant. These results serve as baseline to evaluate the contribution of the proposed models.

5.3.2. Semantic retrieval performance

In the following experiments, both, the query images and the database images, have been automatically annotated by the system. Since these annotations rely on the kernel function used for classification, the retrieval system is evaluated according to the kernel strategy that generates the annotations. Again, two strategies are evaluated: the simple kernel combination and the optimal combination of kernel functions.

Consider the four performance measures reported in Table 5 to evaluate the retrieval response for all kernel functions. The notation for kernel functions is the same as that presented in Table 2. $P(1)$ is the precision of the first retrieved image averaged among all tested queries, which is used to evaluate early precision. All semantic models present a $P(1)$ greater than 0.50, meaning that, in more than half of the queries, these models retrieve a relevant image in the first position. This contrasts with visual features in which almost all models have a $P(1)$ less than 0.50. In addition, $P(100)$ shows how the precision changes among the first 100 results, in which all semantic models keep around 0.40, contrasting with visual feature models, which present a $P(100)$ around 0.15. It demonstrates the effectiveness of semantic models to bring more relevant images in the first pages of results. The measure $R(100)$ indicates the recall in the first 100 results, in which semantic models present values around 0.40 whereas only visual models are around 0.15, indicating that more relevant images are rapidly found by semantic models.

The last measure in Table 5 is mean average Precision (maPrec), which evaluates the long term precision of the model, that is, the average precision until every relevant image is found. This is the most standard performance measure in information retrieval to compare performance between systems and models. The values obtained by semantic models are around 0.18 whereas only visual features obtain values around 0.10, showing an average improvement of 57%. These results show an important improvement of the retrieval performance of semantic retrieval models over the visual-based retrieval models.

To evaluate the significance of the obtained results, we employed and Analysis of Variance (ANOVA) test on the maPrec values. We ranked all models by maPrec and evaluated groups of models to find classes with similar intra-class performance and significantly different inter-class performance. Table 6 presents the results of the test using a significance value $\alpha = 1\%$, showing that the difference between semantic models and visual features is statistically significant. It also shows 3 classes of semantic models with statistically different performance, whose partition is mainly due to the underlying kernel function. Each kernel function is a different image representation for the learning algorithms, and these results suggest that the most important factor to build an effective feature space is the use of an appropriate kernel function to exploit the structure of data and to highlight non-linear relationships. Notice that the Histogram Intersection Kernel in classes III and IV provides an absolute performance slightly better when it

Table 4
Retrieval performance measures for low-level visual features.

Measure	GRA	INV	LBP	RGB	SIFT	SOB	TAM
$P(1)$	0.32	0.19	0.36	0.50	0.46	0.54	0.42
$P(100)$	0.13	0.10	0.14	0.14	0.18	0.16	0.15
$R(100)$	0.13	0.10	0.14	0.14	0.17	0.16	0.14
maPrec	0.10	0.09	0.10	0.10	0.12	0.11	0.10

Table 5
Retrieval performance measures for all semantic models.

Measure	k_I	$k_\cap$	k_G	$k_{G\circ\cap}$	k_I^*	$k_\cap^*$	k_G^*	$k_{G\circ\cap}^*$
$P(1)$	0.56	0.60	0.62	0.68	0.53	0.60	0.58	0.68
$P(100)$	0.36	0.39	0.42	0.44	0.36	0.39	0.41	0.42
$R(100)$	0.35	0.38	0.41	0.43	0.35	0.38	0.41	0.42
maPrec	0.15	0.17	0.20	0.20	0.15	0.17	0.20	0.21

Table 6
Groups with significantly different performance according to the ANOVA test on mean average Precision (maPrec) values.

Class	Model		maPrec
I		Visual features	0.103
II	k_I^*	Aligned identity kernel	0.150
	k_I	Identity kernel	0.152
III	$k_\cap$	Hist. intersection	0.169
	$k_\cap^*$	Aligned hist. intersection	0.170
IV	k_G^*	Aligned RBF kernel	0.198
	k_G	RBF kernel	0.201
	$k_{G\circ\cap}$	RBF ∘ intersection	0.201
	$k_{G\circ\cap}^*$	Aligned RBF ∘ intersection	0.210

is obtained from an optimal combination of features. Even though the optimal combination of features does not show enough evidence to be considered as a factor that produces statistically significant differences, it has shown a positive impact in the automatic image annotation and retrieval tasks.

Another way to compare the performance of models is using the recall vs. precision plot, as is shown in Fig. 4. The parameter to used to generate the curves is the number n of nearest neighbors provided by the retrieval process. This Figure shows the performance of one model of classes I–III and two models of class IV, according to the statistically different classes presented in Table 6. It shows the differences in performance between models, as predicted by the ANOVA test. We selected two models of class IV to illustrate the differences between the optimal combination and direct combination of features, in which a slightly improved performance can be observed. We argue that, even though the proposed optimal feature combination strategy does not show a significant improvement, it has potential applications in the design and selection of feature sets for histology image representation. Also, this strategy may allow a further improvement of classification and retrieval results if the set of features is more targeted to describe specific histopathology properties, as opposed to the use of general purpose image features. Nevertheless, constructing image representations with kernel functions has allowed to integrate multiple visual features, to exploit feature structure, to integrate a feature selection strategy and to highlight non-linear patterns in the same framework, as an effective strategy for semantic histopathology image retrieval.

Fig. 5 shows an illustration of the differences between visual retrieval and semantic retrieval using the proposed methods. The query image is the first from left to right, and it is used to search for images exhibiting the *lymphocyte infiltrate* concept. The top five results are presented immediately after the query image, marked with blue squares if they are relevant or red squares if they are not. The results obtained using Sobel features as retrieval strategy share more appearance commonalities with the query than those provided by the semantic retrieval. However, the three last results of the visual retrieval are not relevant because they do not exhibit the target concept, while the results obtained with the semantic annotations are all relevant.

In summary, the response of the retrieval system is more appropriate when it is configured to search images using semantic annotations in contrast to the performance obtained using only low-level features, as the results have shown. It is important to notice that semantic annotations rely on the automatic analysis of visual image features, and the performance heavily depends on the image representation. That was in fact the main purpose of this study, to model and evaluate different factors to generate expressive feature spaces for histology images. These representations can be efficiently harnessed by learning algorithms, which extract high-level semantics from images and labels during training to be transferred to new, unseen images.

5.4. Discussions

The components of a system to retrieve histopathology images using an example image has been presented and evaluated. The system provides access to images according to the semantic content, which is generated by an automatic annotation module. The most remarkable characteristic of the proposed auto-annotation module is that it generates image representations in high dimensional feature spaces using kernel functions and multiple visual features, to better recognize histopathology concepts in images. The following are some specific benefits of the way in which we model the problem:

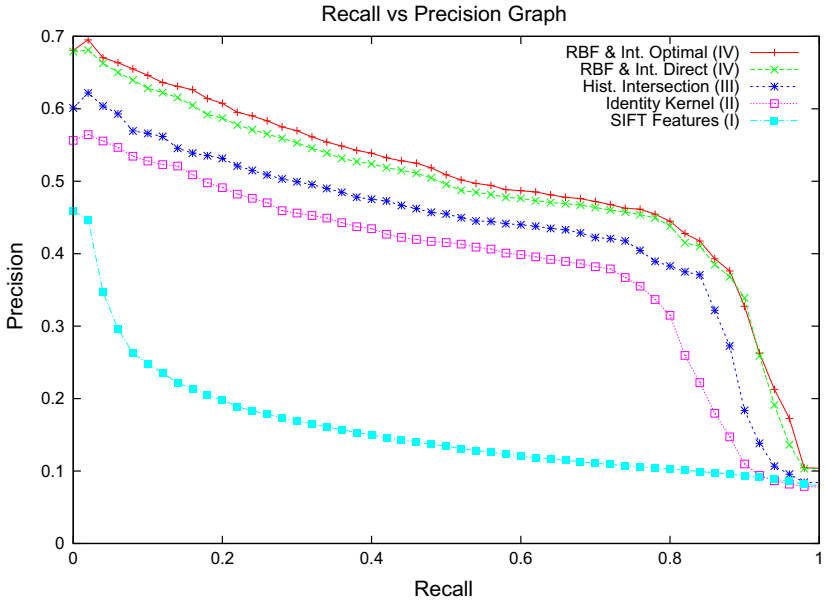
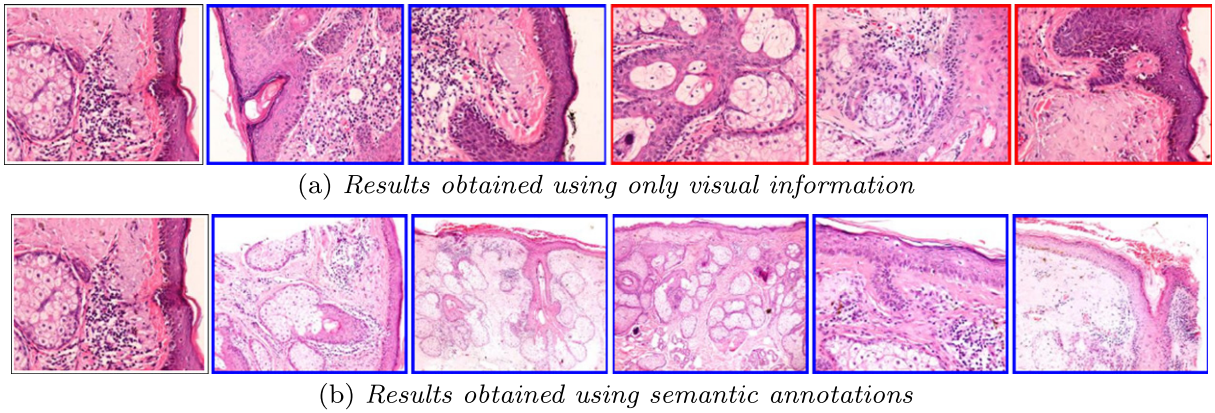


Fig. 4. Recall vs Precision graph comparing the retrieval performance of statistically different models. Two models of class IV are plotted to illustrate differences between the direct and optimal combination of kernels. The best performing visual feature (SIFT) is included as representative of class I.



(a) *Results obtained using only visual information*

(b) *Results obtained using semantic annotations*

Fig. 5. Illustration of a content-based query. The query is the first image from left to right. The top-5 results are shown in order of relevance from left to right. Results are marked with blue if they are relevant and with red if they are not. The query image is used to search for images with lymphocyte infiltrate.

1. Multiple features: Histology images are known to have objects that can be described using multiple features. Architectural features [13], textural features [37] and even colors [38] have been proposed to capture variabilities of image contents. We designed the annotation module to deal with multiple features of different nature, and implemented seven histograms in our studies to demonstrate the potential of this approach. These histogram features included textures, colors, edges and invariants, and each histogram has 256 or 512 bins, that are efficiently managed by our module.
2. Structured features: In our kernel-based framework visual features can have arbitrary structure as long as they are provided with a valid kernel function. We evaluated the identity kernel and the Histogram Intersection Kernel to process histogram features. In our study, the identity kernel can be regarded as an attempt to use the original descriptors as simple feature vectors and the linear combination of identity kernels can be understood as the concatenation of these vectors. Experimental results showed that the Histogram Intersection Kernel, which exploits the particular structure of histograms to evaluate a similarity measure, provides more accurate results in classification and retrieval tasks. Our model can be extended to include other structures such as trees and graphs in the visual feature set.
3. Combination of features: Since all visual descriptors are mapped to a high-dimensional feature space using a kernel function, we model the problem of feature combination as a problem of kernel functions combination, and in such a way, we generate combined feature spaces that integrate all the information. This strategy can be understood as a late fusion process as opposed to previous approaches for histology image classification and retrieval that concatenate features in a single feature vector [16,39], i.e., using an early fusion strategy. Our approach provides the advantage of considering the particular structure of each feature independently of the others, instead of mixing up everything in a unique vector. Furthermore, our combination approach can include the automatic weighting of features following a kernel alignment strategy. In our experiments, the latter procedure did not show a significant improvement in the final performance, however, we consider this extension as a tool of great potential to select more specialized visual descriptors and to design better image representations.
4. Highlighting non-linear patterns: The histopathology concepts included in our study showed to have high non-linearity in the feature space. This is observed by the large improvement on classification and retrieval performance that was obtained using the RBF kernel. This is an additional advantage of our framework, taking into account that the resulting image representation can be further improved just by operating kernel functions. Representing non-linear patterns is specially useful in image classification tasks, where learning algorithms need to separate complex regions in the feature space.
5. Semantic annotations: Our approach does not attempt to find a unique right class for every image. Instead, it generates multiple annotations according to the visual contents, allowing to extend the functionality to new required search terms. This characteristic makes it different to other approaches that consider just a few labels, as opposed to ours that considered 18 high-level concepts. In our study we only considered the query by example paradigm as the way to retrieve images, but using the automatically generated annotations, images can also be retrieved using a keyword-based strategy.
6. Semantic vs. visual retrieval: Visual features have been extensively used for image retrieval, and the community has found that the main problem using them is the semantic gap. The automatic analysis of visual image contents is at the core of the proposed strategy, and we found that the way in which visual features are used determines the final retrieval performance. Our study showed that the discriminative power of visual features highly depends on the kernel function used to train classifiers, since they allow learning algorithms to exploit feature structure and non-linear patterns. On the other hand, a standard visual retrieval approach only rank images using a similarity measure, i.e., finding nearest neighbors. The success of the proposed semantic retrieval approach is that it uses machine learning to translate non-linear patterns that can be found in visual feature spaces into a more explicit semantic format that is used to rank images efficiently.

6. Conclusions

This paper presented a novel strategy for automatic annotation of histopathology images. The proposed framework is entirely based on kernel methods, allowing to deal with multiple visual descriptors to build expressive feature spaces. The generated annotations are used to search images with similar annotations in an image retrieval system under the query-by-example paradigm. We implemented and evaluated the model following an extensive experimentation on real histopathology images. The proposed strategy to retrieve semantically valid results from a large collection of histopathology images showed an average improvement of 57% when compared to visual search, based on low-level features. In our future work, we consider the use of more specialized visual features for histology images to improve the final search

quality, and the automatic analysis of co-occurrence among annotations to differentiate between normal and abnormal images.

Acknowledgments

This work was partially funded by the project SISTEMA DISTRIBUIDO DE ANOTACION AUTOMATICA Y RECUPERACION SEMANTICA DE IMAGENES DE HISTOLOGIA number 1101-487-25779 of Ministerio de Educación Nacional de Colombia by Convocatoria COLCIENCIAS 487 de 2009. Also this work was partially funded by the project REPRESENTACIÓN Y CLASIFICACIÓN DE GRANDES COLECCIONES DE IMÁGENES MÉDICAS number 1101-489-25577 of Colciencias by Convocatoria COLCIENCIAS 489 de 2009.

References

[1] Kragel P, Kragel P. Digital microscopy: a survey to examine patterns of use and technology standards. In: Telehealth/AT '08: Proceedings of the IASTED International Conference on Telehealth/Assistive Technologies. Anaheim (CA, USA): ACTA Press; 2008. p. 195–7.
[2] Dennis T, Start RD, Cross SS. The use of digital imaging, video conferencing, and telepathology in histopathology: a national survey. J Clin Pathol 2005;58(3):254–8.
[3] Müller H, Michoux N, Bandon D, Geissbuhler A. A review of content-based image retrieval systems in medical applications–clinical benefits and future directions. Int J Med Inform 2004;73(1):1–23.
[4] Datta R, Joshi D, Li J, Wang JZ. Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 2008;40(2):1–60.
[5] Deselaers T, Keysers D, Ney H. FIRE – flexible image retrieval engine: ImageCLEF 2004 evaluation, multilingual information access for text. Speech Images (2005) 688–98.
[6] Müller H, Lovis C, Geissbuhler A. The medGIFT project on medical image retrieval. In: Proceedings of first international conference on medical imaging and telemedicine, Wuyi Mountain, China, 2005.
[7] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 2000;22(12):1349–80.
[8] Tang HL, Hanka R, Ip HHS. Histological image retrieval based on semantic content analysis. Inform Technol Biomed, IEEE Trans 2003;7(1):26–36.
[9] Iregui M, Gomez F, Romero E. Strategies for efficient virtual microscopy in pathological samples using JPEG2000. Micron 2007;38(7):700–13.
[10] Yu F, Ip H. Semantic content analysis and annotation of histological images. Comput Biol Med 2008;38(6):635–49.
[11] Doyle S, Hwang M, Shah K, Madabhushi A, Feldman M, Tomaszeweski J. Automated grading of prostate cancer using architectural and textural image features. In: ISBI, 2007. p. 1284–7.
[12] Kong J, Sertel O, Shimada H, Boyer KL, Saltz JH, Gurcan MN. Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation. Pattern Recogn 2009;42(6):1080–92.
[13] Tambasco M, Costello BM, Kouznetsov A, Yau A, Magliocco AM. Quantifying the architectural complexity of microscopic images of histology specimens. Micron 2009;40(4):486–94.
[14] Mosaliganti K, Janoos F, Irfanoglu O, Ridgway R, Machiraju R, Huang K, et al. Tensor classification of N-point correlation function features for histology tissue segmentation. Med Image Anal 2009;13(1):156–66.
[15] Zheng L, Wetzel AW, Gilbertson J, Becich MJ. Design and analysis of a content-based pathology image retrieval system. Inform Technol Biomed, IEEE Trans 2003;7(4):249–55.
[16] Naik J, Doyle S, Basavanally A, Ganesan S, Feldman MD, Tomaszewski JE, et al., A boosted distance metric: application to content based image retrieval and classification of digitized histopathology. In: SPIE medical imaging: computer-aided diagnosis, vol. 7260; 2009. p. 72603F1–12.
[17] Caicedo JC, Gonzalez FA, Romero E. A semantic content-based retrieval method for histopathology images. Inform Retrieval Technol LNCS 2008;4993:51–60.
[18] Wong CSM, Strange RC, Lear JT. Basal cell carcinoma. BMJ 2003;327:794–8.
[19] Caicedo JC, Gonzalez FA, Triana E, Romero E. Design of a medical image database with content-based retrieval capabilities. Adv Image Video Technol LNCS 2007;4872:919–31.
[20] Bosch A, Muñoz X, Martí R. Which is the best way to organize/classify images by content? Image Vision Comput 2007;25(6):778–91.
[21] Szummer M, Picard RW. Indoor–outdoor image classification, content-based access of image and video database, 1998. In: Proceedings., 1998 IEEE international workshop on; 1998. p. 42–51.
[22] Qi X, Han Y. Incorporating multiple SVMs for automatic image annotation. Pattern Recogn 2007;40(2):728–41.
[23] Gueld MO, Keysers D, Deselaers T, Leisten M, Schubert H, Ney H, et al. Comparison of global features for categorization of medical images. Med Imag 2004;5371:211–22.
[24] Siggelkow S. Feature histograms for content-based image retrieval. Ph.D. thesis, Albert-Ludwigs-UniversitÂšat Freiburg im Breisgau; 2002.
[25] Berman AP, Shapiro LG. A flexible image database system for content-based retrieval. Comput Vision Image Understanding 75.
[26] Mark ASA, Nikson S. Feature extraction and image processing. Elsevier; 2002.
[27] Deselaers T. Features for image retrieval. Ph.D. thesis, RWTH Aachen University. Aachen, Germany; 2003.
[28] Caicedo JC, Cruz A, Gonzalez F. Histopathology image classification using bag of features and kernel functions. In: Artificial intelligence in medicine conference, AIME 2009 LNAI 5651; 2009. p. 126–35.
[29] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. Cambridge University Press; 2004.
[30] Barla A, Franceschi E, Odone F, Verri A. Image kernels, pattern recognition with support vector machines. LNCS 2002;2388:617–28.
[31] Maji S, Berg AC, Malik J. Classification using intersection kernel support vector machines is efficient. In: IEEE conference on computer vision and pattern recognition, 2008. CVPR 2008. p. 1–8.
[32] Schölkopf B, Smola A. Learning with kernels. Support vector machines, regularization, optimization and beyond. The MIT Press; 2002.
[33] Cristianini N, Shawe-Taylor J, Elissee A, Kandola J. On kernel-target alignment. In: Advances in neural information processing systems, vol. 14; 2002. p. 367–73.
[34] Kandola J, Shawe-Taylor J, Cristianini N. Optimizing kernel alignment over combinations of kernel. Tech. rep., Department of Computer Science, Royal Holloway, University of London, UK; 2002.
[35] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation, vol. 10. MIT Press; 1998. p. 1895–923.
[36] Müller H, Müller W, Squire DM, Marchand-Maillet S, Pun T. Performance evaluation in content-based image retrieval: overview and proposals. Pattern Recogn Lett 2001;22(5):593–601.
[37] Diamond J. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. Human Pathol 2004;35(9):1121–31.
[38] 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. J Signal Process Syst 2009;55(1):169–83.
[39] Orlov N, Shamir L, Macura T, Johnston J, Eckley DM, Goldberg IG. WND-CHARM: multi-purpose image classification using compound image transforms. Pattern Recogn Lett 2008;29(11):1684–93.