



## Visual pattern mining in histology image collections using bag of features

Angel Cruz-Roa, Juan C. Caicedo, Fabio A. González\*

Bioingenium Research Group, Computer Systems and Industrial Engineering Department, National University of Colombia, Cra 30 No 45 03-Ciudad Universitaria, Faculty of Engineering, Building 453 Office 114, Bogotá DC, Colombia

### ARTICLE INFO

#### Article history:

Received 9 March 2011

Received in revised form 19 March 2011

Accepted 17 April 2011

#### Keywords:

Collection-based image analysis  
 Visual pattern mining  
 Visual knowledge discovery  
 Bag of features (BOF)  
 Visual-codebook feature selection  
 Kernel-based image annotation  
 Identification of visual patterns  
 Histology and histopathology images  
 Basal-cell carcinoma  
 Fundamental tissues

### ABSTRACT

**Objective:** The paper addresses the problem of finding visual patterns in histology image collections. In particular, it proposes a method for correlating basic visual patterns with high-level concepts combining an appropriate image collection representation with state-of-the-art machine learning techniques.

**Methodology:** The proposed method starts by representing the visual content of the collection using a bag-of-features strategy. Then, two main visual mining tasks are performed: finding associations between visual-patterns and high-level concepts, and performing automatic image annotation. Associations are found using minimum-redundancy-maximum-relevance feature selection and co-clustering analysis. Annotation is done by applying a support-vector-machine classifier. Additionally, the proposed method includes an interpretation mechanism that associates concept annotations with corresponding image regions.

The method was evaluated in two data sets: one comprising histology images from the different four fundamental tissues, and the other composed of histopathology images used for cancer diagnosis. Different visual-word representations and codebook sizes were tested. The performance in both concept association and image annotation tasks was qualitatively and quantitatively evaluated.

**Results:** The results show that the method is able to find highly discriminative visual features and to associate them to high-level concepts. In the annotation task the method showed a competitive performance: an increase of 21% in  $f$ -measure with respect to the baseline in the histopathology data set, and an increase of 47% in the histology data set.

**Conclusions:** The experimental evidence suggests that the bag-of-features representation is a good alternative to represent visual content in histology images. The proposed method exploits this representation to perform visual pattern mining from a wider perspective where the focus is the image collection as a whole, rather than individual images.

© 2011 Elsevier B.V. All rights reserved.

### 1. Introduction

Thanks to the development on acquisition methods and equipment and the accelerated progress in communications and computer technologies, there is an ever increasing availability of digital biomedical images [1]. Biomedical images are an important source of information, and a potential source of knowledge, for both routine clinical decision and biomedical research. Nevertheless, a thorough exploitation of this potential requires techniques able to automatically extract information and knowledge from this vast amount of data. This is an enterprise that has already started, but is far from being finished [2]. A great deal of work has been

done on the area of medical imaging, which is gradually moving from computer assisted image analysis systems, mainly based on image processing techniques [3], to fully automatic systems based on pattern recognition and machine learning methods [4]. Most of the work on automatic medical image analysis and interpretation has concentrated on individual images rather than on collections of images. Changing this perspective poses new, and potentially useful, questions: What are the relationships between the images? What are the common and distinctive characteristics among them? What are the implicit categories or groups that could be identified in the collection?

The questions discussed in the previous paragraph can be deemed as instances of a more general image understanding problem, in which the focus of the interpretation process is not an individual image, but the image collection as a whole. This introduces new challenges, but also provides new methods to extract hidden knowledge from data. This could have an important impact

\* Corresponding author. Tel.: +57 1 3165322; fax: +57 1 3165491.

E-mail addresses: [aacruzr@unal.edu.co](mailto:aacruzr@unal.edu.co) (A. Cruz-Roa), [jccaicedoru@unal.edu.co](mailto:jccaicedoru@unal.edu.co) (J.C. Caicedo), [fagonzalez@unal.edu.co](mailto:fagonzalez@unal.edu.co) (F.A. González).

on biomedical image analysis since it provides a new set of tools to automatically find interesting patterns in images, which are difficult to obtain when images are individually analyzed.

This paper addresses the problem of automatically extracting visual patterns from biomedical image collections. The successful solution of this problem requires to solve two main subproblems: first, to find an appropriate image representation that takes into account the structure of the image collection; second, to choose appropriate machine learning tools that, based on the image collection representation, could extract the most meaningful visual patterns. **The main contribution of the present work is a method that successfully solves these problems by combining state-of-the-art machine learning techniques for feature selection and pattern mining, along with a bag-of-feature approach to represent the image collection visual content.**

During the last few years, the bag-of-features (BOF) image representation has attracted great attention from the computer vision community. This approach is an evolution of texton-based representations and is also influenced by the bag-of-words representation for text classification and retrieval [5]. The BOF representation is an adaptive approach to model image structure in a robust way. In contrast to image segmentation, this approach does not attempt to identify complete objects inside images, which may be a harder task than the image classification itself. Instead, the BOF approach looks for small characteristic image regions allowing the representation of complex image contents without explicitly modeling objects and their relationships, a task that is tackled in another stage of the image analysis process. Briefly, the BOF approach works as follows: a set of small regions are extracted from all the images in the collection, these regions are represented by feature vectors, then a visual dictionary, a set of codewords, is built as a summary of these feature vectors, finally, each image in the collection is represented by the frequency of the dictionary codewords that it contains, i.e., each image is represented by a codeword histogram. In addition, an important advantage of the BOF approach is its adaptiveness to the particular image collection to be processed. Some of these properties are particularly useful for medical image analysis and, in fact, the BOF representation has been successfully applied to some problems in medical imaging. For instance, Tomassi et al. [6] adapted the BOF representation to effectively classify radiological images in an automatic image annotation task.

This work concentrates on histological images, and this is motivated by two main reasons: **first, histology images are particularly challenging from an image understanding point of view; in this type of images, visual patterns are generally a complex combination of fundamental visual features involving texture, color and shape [7]; second, the success of the BOF representation in other type of images, such as natural scenes or X-rays, is not a guarantee that it will perform well in histology images, so, the sole fact of testing this representation in this type of images is a contribution by itself.** The assessment and identification of biological parameters in histology materials is usually made by visual inspection of tissue samples, something that may be often an inadequate approach to extract the potential information [8]. Manual intervention holds the disadvantages of being subjective, laborious, and insufficient when more complex information is needed or is simply unknown. **Loukas [9] stresses the importance of computerized methods as an essential tool for interpreting data with major diagnostic value, and points out that there is little work on computational methods dedicated to the extraction of meaningful information from histology images.**

In our previous work [10,11] we presented preliminary results that suggested the potential of the BOF representation for histology images. The present work builds on these preliminary results performing a systematic experimentation and proposing a method for visual pattern mining in histology image collections. In particu-

lar, the proposed method is tested on two different histology image collections: one involving a wide range of images acquired from different organs and representative of normal fundamental tissues, and the other one including images used for diagnosis of a type of skin cancer called basal-cell carcinoma. In both cases, images have global annotations, corresponding to high-level concepts, but miss local or regional annotations. Taking into account that an image may involve different tissues and biological structures, the challenge is to identify the particular visual patterns that characterize the different associated high-level concepts. The experimental results show that the proposed method is successful on finding meaningful visual patterns. Particularly, the method was able to find: visual patterns that are highly correlated with high-level concepts in both data sets (e.g. cancer cells appearance for *cystic change* concept), the space location of these visual words in global annotated images (e.g. image regions related to *muscular tissue*), and groups of images with similar appearance and visual patterns (e.g. a group of images that share the same concept, *epithelial tissue*, and stain, *Masson's trichrome*). Additionally, in the automatic annotation task the BOF representation showed an improved performance over the baseline given by a global representation based on texture (Gabor, Zernike and Tamura). Specifically, the test *f*-measure, which combines precision and recall, is increased by 21% in the histopathology data set and 47% in the histology data set.

The paper is organized as follows: Section 2 reviews previous works in histology image representation and medical image annotation using BOF; Section 3 describes the details of the different stages of the image collection representation strategy based on BOF; Section 4 presents the proposed method for visual pattern mining; Section 5 describes the proposed automatic annotation strategy; Section 6 discusses the image acquisition process and data sets used; Section 7 presents the experimental results in both data sets; finally, conclusions and future work are discussed in Section 8.

## 2. Previous work

Microscopy image processing has been the subject of an important body of research since digital images started to be coupled to microscopes in the early 80s to acquire and analyze high quality images [8]. From signal processing operations to automated pathology grading, there is a wide range of applications and problems in microscopy images that brings together researchers of different disciplines.

One of the earliest applications of computer vision to histology images has been the characterization of cells within a slide. Quantifying cells or defining their boundaries in blood films or tissue slides is a time consuming and subjective task, yet, an important procedure in research and diagnosis activities. Automatic identification and measurement of single cell properties has been proposed as a mechanism to help experts make accurate and reproducible measurements in cell cultures [12] and infected blood films [13] among others. These methods showed to be 30 times faster than humans and up to 90% accurate in completing the task.

In a broad range of histology analysis tasks, the subject of study goes from individual cells to complete tissue regions. These cases are usually related to cancerous lesions, and the purpose of tissue analysis is to determine the stage of the lesion on different parts of the tissue, a procedure known as grading. Automatic grading is approached as an image segmentation problem to identify those regions that correspond to each grade [14]. Different pathologies may have different grading protocols and strategies to identify lesion stage. For instance, Doyle et al. [15] follows the Gleason system, which describes 5 increasingly malignant stages of cancer and uses discriminative learning to identify them according to several

tissue architectural features. In [16], neuroblastoma histological slides are processed for identifying pathological regions associated to three different grade subtypes, which are identified in a multiresolution framework. Other examples of tissue segmentation and tissue classification can be found in [17,18].

The main characteristic of the works described above, is that the image analysis concentrates on evaluating information in one image to segment cells or regions, which is still a very important and fundamental problem in histology image analysis. However, our approach is different from these works since we follow an image collection analysis strategy to extract meaningful information out of a set of images rather than process or segment tissues in individual slides.

Another branch of research in histology image analysis is the automatic image classification, annotation and retrieval. Since large numbers of digital histology slides are being stored more frequently, methods for automatic image organization and access strategies are becoming important. Histology image classification using multiple transformed features was evaluated by Orlov et al. [19], training classifiers that decide what overall category an image belongs to.

The design of image similarity measures for histology images has been approached by Tang et al. [20] and Naik et al. [21], to enable image retrieval systems to use semantic information. These approaches tend to exploit more systematically the information within a collection of images rather than processing just individual images. Learning a classification model and computing image similarities are tasks that require image collection analysis. However, these works are not intended to discover relationships between visual patterns or to reveal the image collection structure, which is the core of our study.

The present work is closely related to **a new emerging research area called bioimage informatics** [22,23], which comprises image processing, data mining and database visualization, extraction, searching, comparison and management of biomedical knowledge inside massive image collections. Peng [22] reviewed techniques and biomedical application of this area to high-throughput/high-content analysis of cellular phenotypes, atlas building for model organisms, understanding the dynamic processes in cells and living organisms, joint analysis using both bioimage informatics and other bioinformatics methods.

Mining of visual patterns in biomedical image collections has important applications in both research and clinical practice. For instance, Swedlow and Eliceiri [24] and Kvilekval et al. [25] present bioimage informatics tools for analysis and management of large biomedical data supporting collaborative research in molecular and cell biology. Another examples is the work of Madabhushi et al. [26,27] that proposes a method that combines multimodal information sources, including magnetic resonance imaging (MRI), digital pathology and protein expression, to support the prognosis and theragnosis of cancer patients.

In our study, we consider the BOF approach for image representation as a mechanism to analyze local image patterns from a whole collection perspective. This strategy has been previously used by other researchers to approach certain problems in medical image analysis, particularly, high level interpretation of radiology images. Bosch et al. [28] and Iakovidis et al. [60] used a BOF approach to deal with mammography images and X-ray images respectively. Tommasi et al. [6] and Avni et al. [29] have adapted BOF models to more general medical image collections, with different modalities, body parts and pathologies. To the best of our knowledge, our work is the first attempt to systematically evaluate the potential of a BOF model in histology images. An important difference with other works that apply BOF to medical images, is that, in almost all the cases, these works are mainly focused on the problem of automatic image annotation using discriminative models [6,29,30], however other works

propose the use of generative models taking advantage of this representation to learn the latent semantic of data using probabilistic latent semantic analysis (pLSA) and latent dirichlet allocation (LDA) [31,32]. Related works have been applied in biomedical and bioinformatics problems [33,34]. In histology, latent semantic analysis has not been extensively applied, for instance, pLSA has been used for dimensionality reduction in histopathology images but not for latent topic analysis [35].

In our work, the BOF approach is used to learn discriminative models for automatic image annotation, as well as for analyzing relationships between local visual patterns and image categories from a wider perspective, adding an interpretation layer that aims to explain image collection structures and that supports high-level decision making in histology.

### 3. Image collection visual content representation using BOF

The BOF framework is an adaptation of the bag-of-words scheme used for text categorization and text retrieval. The key idea is the construction of a codebook, i.e., a visual vocabulary in which the most representative patterns are codified as code-words or visual words. Then, the image representation as BOF is a histogram generated through a simple frequency analysis of each codeword inside the image. Csurka et al. [36] describe four steps to classify images using a BOF representation: (1) feature extraction and representation, (2) codebook construction, (3) the BOF representation of images, and, finally, (4) training of learning algorithms. Fig. 1 shows an overview of the three first steps. The BOF approach is a novel and simple method to represent image contents using collection dependent patterns. This is also a flexible and adaptable framework, since each step may be determined by different techniques according to the particular application domain needs. The following subsections describe these steps.

#### 3.1. Feature extraction and representation

In general, the BOF approach starts extracting small blocks (in the present work,  $8 \times 8$  pixels) from each image in the collection. There are two main alternatives for block extraction, partition of the image by a regular grid or extraction of blocks on salient points [37]. In this study the regular-grid-based extraction is used; this process take into account a large quantity of blocks, but reduces the probability of missing interesting patterns. Each extracted block must be represented by a set of features. There is a great variety of image descriptors proposed in the literature [38], we studied three different strategies that have produced good results when used in conjunction with the BOF representation [30,36]. The first strategy uses the raw block, i.e., the feature vector has 64 values corresponding to the luminance values of the corresponding pixels (thus, the color information is ignored). The advantage of this strategy is its simplicity and computational efficiency [37].

The second block-representation strategy is based on scale-invariant feature transform (SIFT) points [39]. This strategy uses a key-point detector based on the identification of interesting points in the location-scale space. This is implemented efficiently by processing a series of difference-of-Gaussian images. The final stage of this algorithm calculates a rotation invariant descriptor using predefined orientations over a set of blocks. SIFT points are used with the most common parameter configuration: 8 orientations and  $4 \times 4$  blocks of cells, resulting in a descriptor of 128 dimensions. The SIFT algorithm has demonstrated to be a robust key-point descriptor in different image retrieval and matching applications, since it is invariant to common image transformations, illumination changes and noise [28,39].

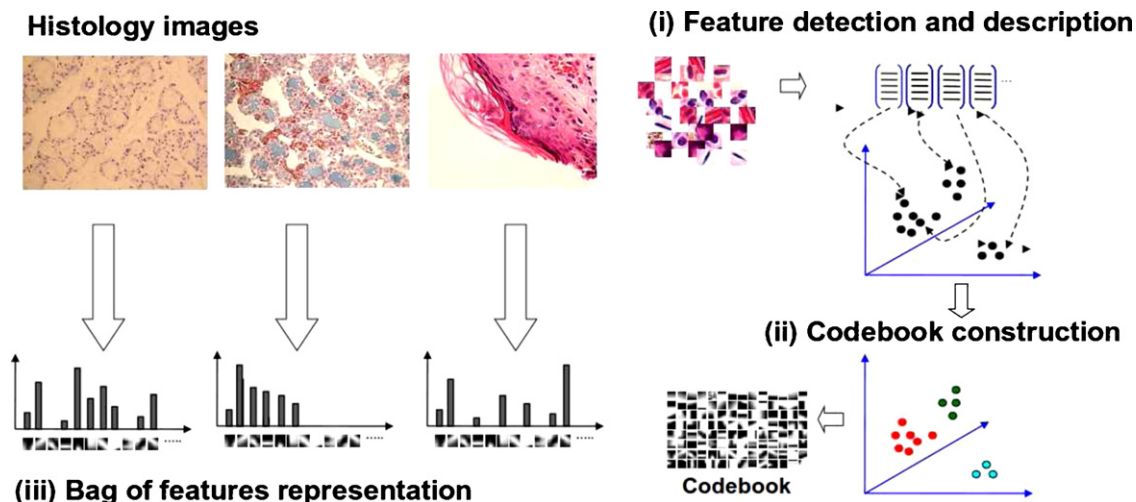


Fig. 1. Overview of the BOF approach to image representation.

Finally, the third strategy is the discrete cosine transform (DCT) [40,41] applied to each channel of the RGB color space by block. The descriptor is built merging the 64 coefficients from each one of the three channels. This strategy generates a visual word that takes into account color and texture information from local features described in an efficient way.

### 3.2. Codebook construction

The visual dictionary or codebook is built using a clustering or vector quantization algorithm applied to the set of block descriptors extracted from the image collection. All local features, over a training image set, are brought together independently of the source image and are clustered to learn a set of representative visual words from the whole collection. The  $k$ -means algorithm is used in this work to find a set of centroids that correspond to the codewords. Nowak et al. [37] have reported that the application of a clustering algorithm has not a big impact in the classification of natural images, compared with a random selection of codewords. However, this is not necessarily the case for histology images [10].

An important decision in the construction of the codebook is the size selection, that is, how many codewords are needed to represent image contents. According to different works on natural image classification, the larger the codebook size, the better [36,37]. However, Tomassi et al. [6] found that the size of the codebook is not a significant aspect in a medical image classification task. We evaluated different codebook sizes, to analyze the impact of this parameter in the pattern mining task.

## 4. Visual pattern mining

The main goal of data mining is to extract useful knowledge from large data bases. This knowledge is usually represented in terms of interesting patterns that uncover hidden, unexpected, and/or interesting relationships among data items. Data mining methods have been successfully applied to different types of data including transactional databases, web pages, and text documents [42]. Image collections are not an exception, in fact there have been some attempts to perform data mining in image databases [43–46], but the advancement has not been as fast as for the other types of data. The fact is that dealing with visual information is particularly challenging because of the semantic gap, i.e., the difficulty of finding a connection between low-level visual information and its conceptual interpretation. This gap has been a widely discussed topic and the focus of several works that propose methods to reduce it

[1,30,47,48]. In biomedical images, the gap may be even larger than in natural scene images or generic objects [49], due to the heterogeneity of these images, the complexity of the structures, and the specialized knowledge required to understand them.

This paper proposes a system to perform visual pattern mining that adapts particularly well to histology image collections. The input of the system is a set of images that have global annotations, which associate images with general conceptual classes. The main goal of the system is to find visual patterns that can be associated with the high-level annotations. The first type of visual patterns are individual visual codewords that are highly correlated with conceptual classes and that have a good discrimination performance. This codewords are selected by a feature selection and analysis process. In general, it is not possible to characterize conceptual classes by individual codewords, but by complex interactions between them. Thus, the next level of visual patterns combine various codewords and associate them with conceptual classes. This is accomplished by a biclustering analysis that brings out these complex interactions. The mined visual patterns can be used to understand how high-level concepts relate to low-level visual content, e.g. mapping discriminative words back to the image to identify characteristic regions associated to a particular concept. Additionally, this patterns can be used to automatically annotate new images. This is accomplished by an annotation module that uses the BOF codified images to train a supervised learning model (e.g., a support vector machine). The overall approach is depicted in Fig. 2, and the different steps involved are detailed in the next subsections. The annotation stage is discussed in Section 5.

### 4.1. Visual word discrimination analysis

The visual dictionary or codebook, as a whole, summarizes the set of visual patterns that are representative of the image collection. Some visual codewords are shared by all the conceptual classes and some others are associated with particular conceptual classes. We are interested in finding the latter kind of codewords, which are good representatives of particular classes, i.e., codewords with a high discriminative power. In the general scope of machine learning, this process is known as feature selection. There are different approaches to perform feature selection, one popular strategy is to choose those features (in this case, codewords) that have a high correlation or dependence with a particular class [50]. This approach is called *maximum relevance* feature selection in [51]. Mutual information (MI) is a popular approximation to measure feature relevance. In general, MI measures the depen-

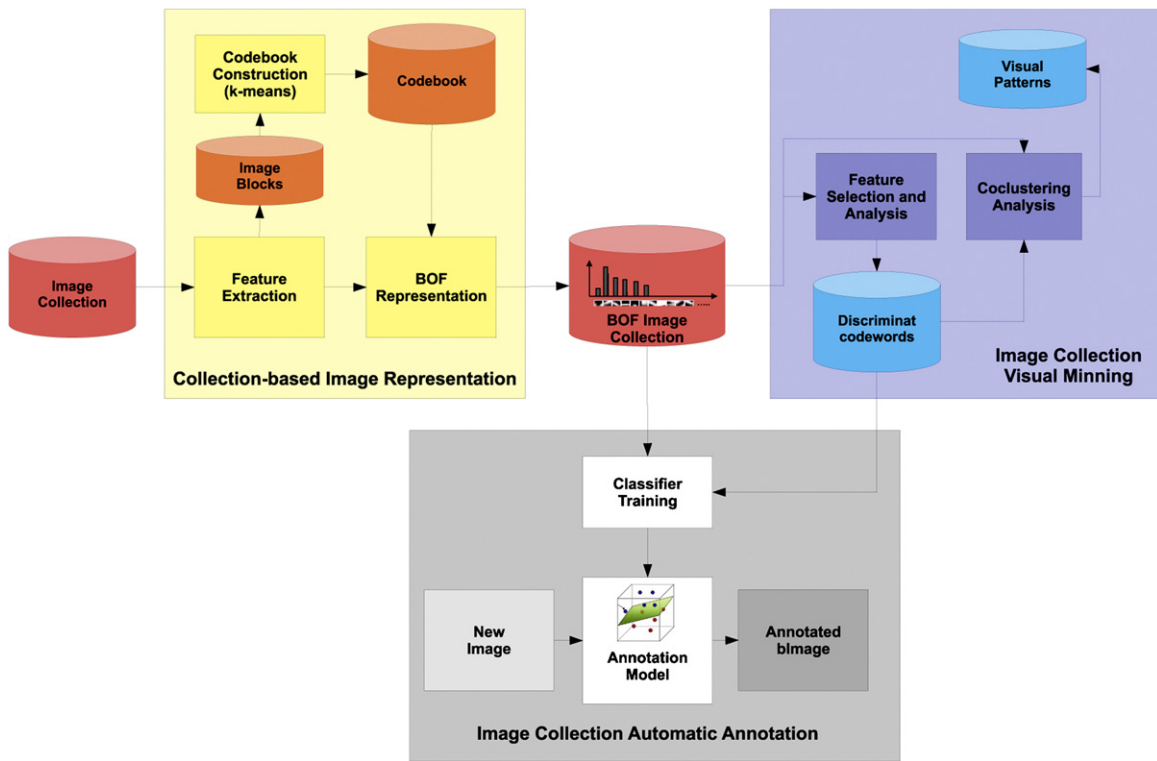


Fig. 2. Overview of the proposed method for visual pattern mining using BOF image representation.

dence between two random variables. In this context, each visual codeword ( $w_i$ ) and concept ( $c_j$ ) are assumed to be binary random variables that measure the presence/absence of a visual word or concept in a particular image block. MI is used to measure the relevance of a visual word with respect to a concept as follows:

$$I(c_j; w_i) = \sum_{w_i \in \{0,1\}} \sum_{c_j \in \{0,1\}} P(c_j, w_i) \log \left( \frac{P(c_j, w_i)}{P(c_j)P(w_i)} \right), \quad (1)$$

where  $P(c_j)$  is the probability that an individual block belongs to an image labeled with the concept  $c_j$ ,  $P(w_i)$  is the probability that one block corresponds to the visual codeword  $w_i$ , and  $P(c_j, w_i)$  is the joint probability.

One problem when using correlation or mutual information to measure the relevance of visual codewords with respect to concepts is that it ignores the interactions among visual codewords and concepts. Some visual codewords can be relevant for several concepts, so their discriminative power is low. A better criterion is provided by the minimum redundancy maximum relevance feature selection (mRMR) method proposed by Peng et al. in [51], which selects a subset of codewords (features) that maximizes the codeword-concept relevance, while minimizing the inter-codeword redundancy. The maximum relevance criterion is represented by the following expression:

$$\max_W D(W, c_j) = \max_W \frac{1}{|W|} \sum_{w_i \in W} I(w_i; c_j), \quad (2)$$

The minimum redundancy criterion is represented by the following equation:

$$\min_W R(W) = \min_W \frac{1}{|W|^2} \sum_{w_i, w_j \in W} I(w_i; w_j) \quad (3)$$

Then simultaneous optimization of both criteria is accomplished by defining a combined objective function  $\Phi(W, c_j)$  defined as follows:

$$\max_W \Phi(W, c_j) = \max_W D(W, c_j) - R(W) \quad (4)$$

The above objective function is solved by an incremental method that builds an optimal subset of features  $W$ . In this work, this method is applied to get a small subset of visual codewords from the large original visual codebook. This produces a set of codewords that collectively have a high discriminative power.

The Peng's method for feature selection [51] is useful to select a highly representative and discriminative subset of codewords. However, we are interested on finding which visual codewords from this subset are associated to which concept. To accomplish this, the degree of connection between visual features,  $w_i$ , and concepts,  $c_j$ , is estimated by the conditional probability  $P(c_j|w_i)$ , which indicates the probability of having the concept  $c_j$  in an image where the visual word  $w_i$  is present. This is calculated as follows:

$$P(c_j|w_i) = \frac{P(c_j = 1, w_i = 1)}{P(c_j = 1, w_i = 1) + P(c_j = 0, w_i = 1)} \quad (5)$$

The above expression is useful to determine the representative visual words per concept using the following strategy: first, the conditional probability of (5) is calculated for each concept and each visual word in the subset of the most relevant-discriminative visual words; second, each visual word is assigned to the concept with the highest conditional probability.

#### 4.2. Biclustering analysis

Biclustering (or coclustering) analysis is a data mining technique which allows simultaneous clustering by rows and columns of a data matrix. This method, with its respective graphic representation of data, is commonly applied in bioinformatics for gene expression analysis [52]. In this particular application field, biclus-

tering analysis is useful to correlate gene expression with different experimental samples (conditions). The different conditions may correspond to samples taken at different times in an experiment, samples from different organs, or samples from different individuals. This is accomplished by finding subsets of genes that are differentially expressed in particular subsets of conditions. The input data is represented as a matrix with rows corresponding to genes and columns to samples. A value at position  $D_{ij}$  represents the amount of expression of gene  $i$  in sample  $j$ . A cocluster (o bicluster) is a submatrix  $D_{IJ}$  defined by the intersection of a subset of genes,  $I \subseteq Rows$ , and a subset of columns,  $J \subseteq Columns$ . Depending on the application, biclusters may be required to exhibit particular properties: constant values, constant rows/columns, coherent values, etc. [52]. Different algorithms have been proposed to efficiently find the different types of biclusters, with a particular emphasis on biclusters with coherent values, which are of special interest in gene expression analysis [52–54].

In this work, we propose to apply biclustering to histology image analysis using the following approach: images are analogous to samples (or conditions) and visual words are analogous to genes. The data matrix is calculated using only the set of most discriminative visual codewords generated by the mRMR feature selection method described in the previous subsection. The main goal is to find biclusters that relate sets of images, which are conceptually connected, with sets of codewords. This goal translates into finding biclusters with high constant values. An agglomerative hierarchical clustering, using Euclidean distance and average linkage, is applied simultaneously to both images and visual codewords. The resulting dendrograms are drawn alongside the matrix representation, where the cells frequency values are represented with colors (blue for low frequency, red for high frequency). Good candidate biclusters can be spotted as rectangles with a uniform red color.

#### 4.3. Identification of characteristic regions

One important aspect of data mining models is their interpretability, i.e., the possibility of understanding how a model is capturing the semantics of a problem. For instance, an interpretable classification model must be able to explain why a particular sample was classified in a given class. In the context of visual pattern mining, this translates to models that are able to relate high level decisions, e.g. assigning an annotation, with particular visual patterns in the image. This is particularly useful in some application scenarios such as computer-aided medical image analysis.

Here we show how a BOF representation allows a higher degree of interpretability in contrasts with other image representations (global histograms, directional transforms, etc.). Specifically, a straightforward method to infer the regions related to global annotations in an automatic way is proposed. Thanks to the fact that BOF visual codewords are, by construction, local visual features, it is possible to map particular codewords back to their position in a particular image. If images are globally annotated (as it is the case for both data sets used in this study), there is not information about what regions of an image are ‘responsible’ for the particular annotations the image has. We can exploit the locality of the codewords to automatically infer this information as follows:

1. Given a new image, represent it using a particular BOF codebook that has been previously processed to find the most discriminative codewords per class.
2. If the image is not annotated, annotate it using, for instance, an annotation algorithm such as the one proposed in Section 5.
3. For a particular image annotation, we can identify codewords present in the image which are associated with the class indicated by the annotation. The resulting set of codewords could be

further filtered to keep only those codewords with the highest class conditional probability.

4. Highlight those image blocks that correspond to the set of selected codewords. The highlighted area corresponds to the region of interest.

### 5. Histology image annotation using BOF

Annotation is an important task in biomedical image analysis. Different works focus on the solution of this problem for different types of images. Section 2 discusses some of these works in the area of histology image classification. Despite the fact that the focus of this work is not the annotation problem, it is interesting to see how different representation alternatives, such as texture features and BOF representation, perform on this task. Works using BOF representation in biomedical images are not abundant and in the particular case of histology images they are even scarcer [10,11]. In order to evaluate the BOF representation we compare different strategies (varying the dictionary size and the type of visual word) against texture features that have been suggested for histology image analysis [8,9] including Gabor and Tamura features.<sup>1</sup>

In this section, we applied a supervised learning approach that uses a state-of-the-art learning algorithm to build a classifier for each one of the concepts in each data set. The classifiers use as input the BOF representation of the images and produce a real number that indicates whether the image exhibits the concept (1) or not (-1).

#### 5.1. Kernel methods

Classifiers used in this work are support vector machines (SVM), that receives as input a data representation implicitly defined by a kernel function [56]. Kernel functions describe a similarity relationship between the objects to be classified. The image representation that we are dealing with are histograms with code word frequencies generated by the BOF representation. In that sense, a natural choice of a kernel function would be a similarity measure between histogram structures. The histogram intersection kernel is the kernel function used in this work:

$$D_{\cap}(H, H') = \sum_{m=1}^M \min(H_m, H'_m),$$

where  $H$  and  $H'$  are the codeword frequency histograms of two images, calculated using a codebook with  $M$  codewords.

#### 5.2. Automatic annotation setup

Since one image can be classified in many classes simultaneously, the classification strategy is based on binary classifiers following the one-against-all rule. One classifier is trained per concept class in each data set. Training, including hyper-parameter tuning, is performed on the 80% of the data and the final assessment of classifiers is performed on the remaining 20%. The data set partition is done using stratified sampling in order to preserve the original distribution of examples in both data sets. This is particularly important due to the high class imbalance, especially in the histopathology data set. The SVM hyper-parameters are tuned-up using 10-fold cross validation on the training set choosing the parameters where the average performance was the best. Then, the whole training set is used to train the classifier with the selected hyper-parameters and finally the performance is reported over the

<sup>1</sup> Details of these descriptors are given in [55].

test set. Three different block representations were evaluated: SIFT, block-based and DCT, and three different codebook sizes: 150, 500 and 1000. The performance measures reported in this work are precision and recall to evaluate the detection rate of positive examples, since the class imbalance may produce trivial classifiers with high accuracy that do not recognize any positive example.

## 6. Histology image data sets used in this study

Histology is a fundamental area of biology that studies the anatomy of cells and tissues at the microscopic level in both plants and animals. The main tool for histology is the microscope (light or electron) that is used to examine thin tissue sections. Histology and histopathology<sup>2</sup> images are of great importance for medicine. They are a fundamental asset to determine the normality of a particular biological structure or to diagnose diseases like cancer. Histology courses are designed to train physicians in order to learn different tissue appearances, which vary according to the structure, function and cell organization in different organs of the body. These characteristics are usually highlighted with the help of different types of stains. Histology images are used both for fundamental biological research and for clinical decision making.

In this study, two different data sets that reflect both kind of applications are used. The basal-cell carcinoma data set, which will be denoted as the *histopathology data set*, is constituted by skin images (pathological or normal) stained with hematoxylin–eosin (HE). This data set was annotated by an expert identifying the presence of both normal and abnormal biological structures inside each image. One important characteristic of this data set is that, generally, these concepts correspond to small regions in the whole image, however, annotations are assigned to whole images and not to regions, making the task of automatic annotation even more challenging. On the other hand the fundamental tissues data set, which is denoted as the *histology data set*, is constituted by images of normal tissues colored with different stains (HE, periodic acid-Schiff (PAS), immunohistochemistry (IHC), Masson's trichrome, etc.) and different magnifications (10×, 20× and 40×). Images in this data set were also globally annotated by an expert with the type of fundamental tissue (connective, epithelial, muscular and nervous) that predominates in the image.

The challenges posed by each data set are different. In the histopathology data set the visual appearance of concepts is related with small biological structures that exhibit high variability, which is caused by the presence of pathological tissues associated with a skin cancer. The histology data set presents a high inter-tissue visual appearance variability caused by the different microscopy magnifications and stains used. In both data sets the visual appearance of tissues and biological structures changes according with type of cut of the biological sample (e.g. muscle fibers look like rounded cells in transverse cut whereas that same structure looks like elongated cells in oblique cut).

### 6.1. Basal-cell carcinoma data set

This data set has been previously used in an unrelated clinical study to diagnose a special skin cancer known as basal-cell carcinoma. Basal-cell carcinoma is the most common skin disease in white populations and its incidence is growing worldwide [57]. It has different risk factors and its development is mainly due to ultraviolet radiation exposure. Pathologists confirm whether or not this disease is present after a biopsied tissue is evaluated under microscope. In this evaluation, physicians aim to recognize

**Table 1**  
Image distribution by semantic concept in the histopathology data set.

Concept	#Images
<i>Pilosebaceous anexa</i>	145
<i>Cystic change</i>	67
<i>Elastosis</i>	125
<i>Eccrine glands</i>	148
<i>Lymphocyte infiltrate</i>	140
<i>Lesion with fibrosis</i>	90
<i>Necrosis</i>	52
<i>N-P-C, elastosis</i>	50
<i>N-P-C, infiltrate</i>	176
<i>N-P-C, pilosebaceous a.</i>	60
<i>Sanguineous vessel</i>	122

some characteristic patterns or complex mixes of patterns. This process is called differential diagnosis and it is mainly achieved by visual analysis. In [58], the structural patterns that characterize the basal-cell carcinoma are described and correspond to 11 different complex patterns. The database is composed of 1502 images globally annotated by experts. Each label corresponds to a histopathology concept which may be found in a basal-cell carcinoma image. An image may have one or several labels, i.e., different concepts may be recognized within the same image and the other way around. Fig. 3 shows a sample of images from four different concept classes in the data set and Table 1 shows the image distribution per class.

### 6.2. Fundamental tissues data set

This data set comprises images from different organs that are representative of the four fundamental tissues. The data set includes 2828 images annotated with a global description of the tissue type. The data set composition is as follows: 484 connective tissue images, 804 epithelial tissue images, 514 muscular tissue images, and 1026 nervous tissue images. The images show the four tissues in different stains (HE, Masson's trichrome, PAS, IHC, etc.) and at different magnifications and cuts. Fig. 4 shows two samples for each kind of fundamental tissues and Table 2 shows data distribution by concept.

## 7. Results

### 7.1. Histology and histopathology codebooks

Different codebooks were built for the two data sets using the process described in the previous subsections. Specifically, the three different feature extraction strategies combined with three different codebook sizes generated nine different codebooks per data set. For the histology data set a sample of 1000 images was randomly selected to generate a set of 1,536,000 blocks, which was used as input to the *k*-means clustering algorithm to build the codebook. The same process was applied to the histopathology data set starting from a sample of 1,280,000 blocks extracted from 1000 images. In both cases the *k*-means algorithm was run with *k* equal to 150, 500 and 1000, to generate the respective codebooks.

Fig. 5 shows the block-based and DCT-based codebooks of size 500, where the visual codewords extracted from the histopathol-

**Table 2**  
Image distribution by fundamental tissue in the histology data set.

Concept	#Images
<i>Connective</i>	484
<i>Epithelial</i>	804
<i>Muscular</i>	514
<i>Nervous</i>	1026

<sup>2</sup> Analysis of pathological tissues.

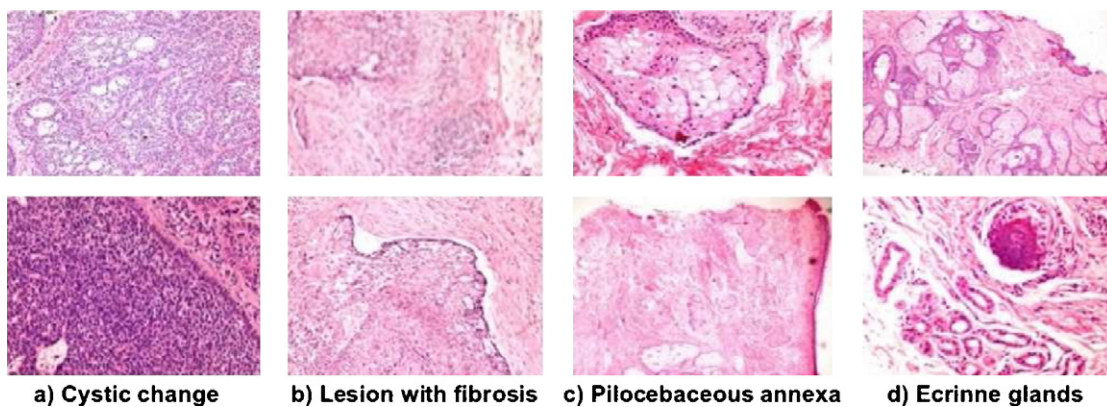


Fig. 3. Sample images from the histopathology data set for basal-cell carcinoma diagnosis exhibiting different normal and abnormal patterns.

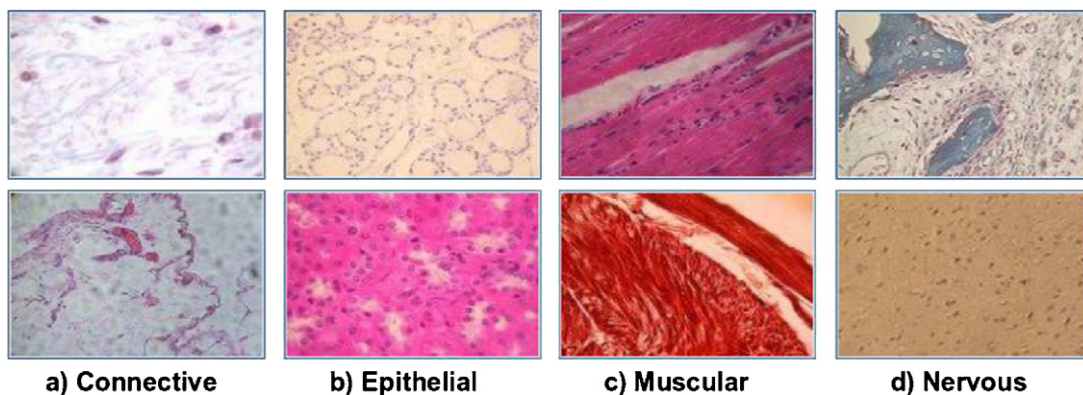


Fig. 4. Sample images for the four fundamental tissues in the histology data set.

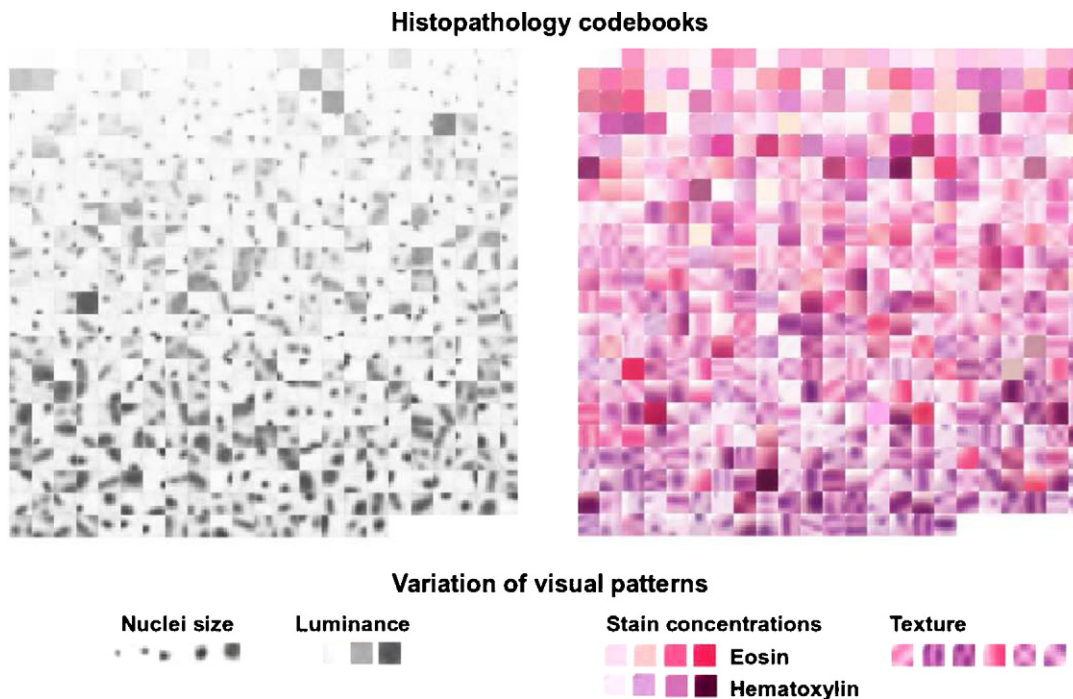


Fig. 5. Comparison of visual words in the codebooks of size 500 based on blocks (top-left) and DCT (top-right) from histopathology data set. In both codebooks the visual words are sorted, in descending order, by their frequency in the whole collection. Each codebook captures different visual patterns including variation in nuclei size, luminance, stain concentration and texture (bottom).



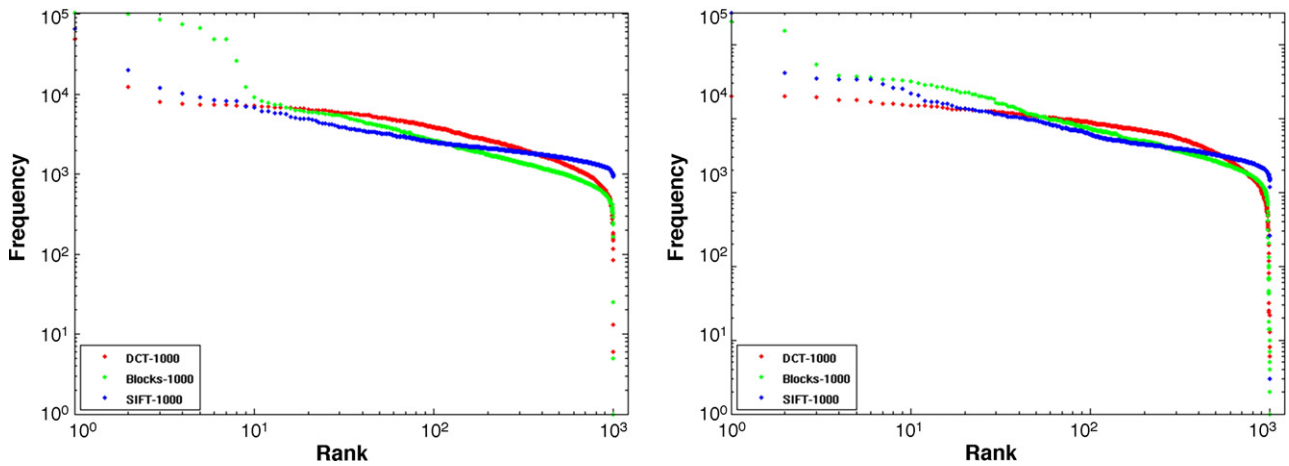


Fig. 6. The frequency of visual words against their rank for 1000-size codebook based on blocks, SIFT and DCT. Left, histopathology data set, and, right, histology data set.

ogy collection were sorted by frequency in descending order. For the block-based codebooks, the visual representation of codewords is generated by averaging the raw blocks in the cluster of the corresponding codeword. In the case of DCT-based codebooks, the visual representation is generated by calculating the inverse DCT of the centroid of the corresponding cluster.

In the block-based codebook we can appreciate detailed patterns associated with nuclei of different sizes, orientations and

luminance levels. In the DCT-based codebook the most frequent patterns correspond to color features related with the different concentrations of stain, followed by texture related patterns. Different color tones are associated with the cytoplasm (pink levels) and diffuse or more general patterns of nuclei (purple levels) according with *hematoxylin-eosin* stain. This behavior is also observed in the histology data set, the main difference is that this data set has a richer variety of color, which is a direct consequence of the higher number of stains present in it.

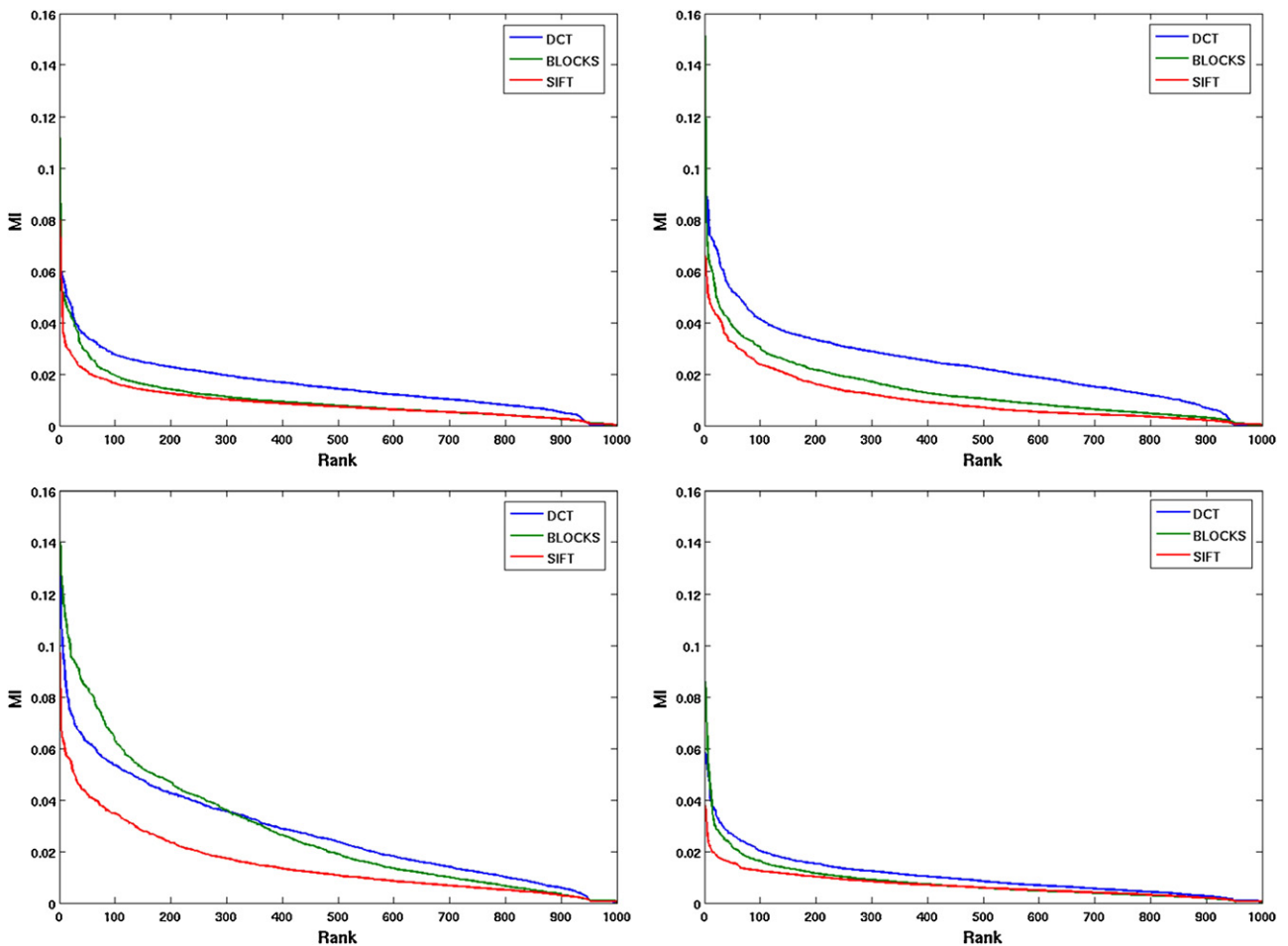
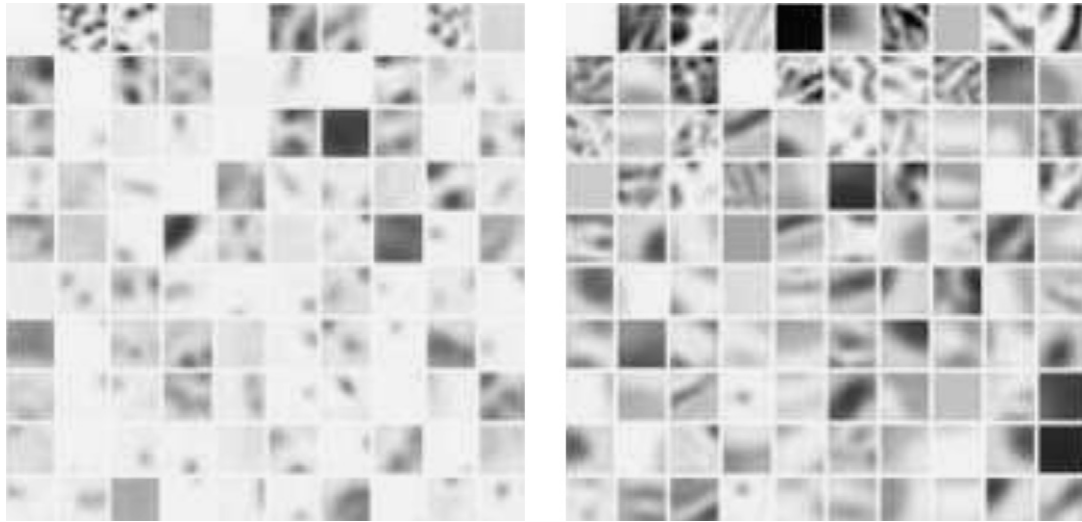
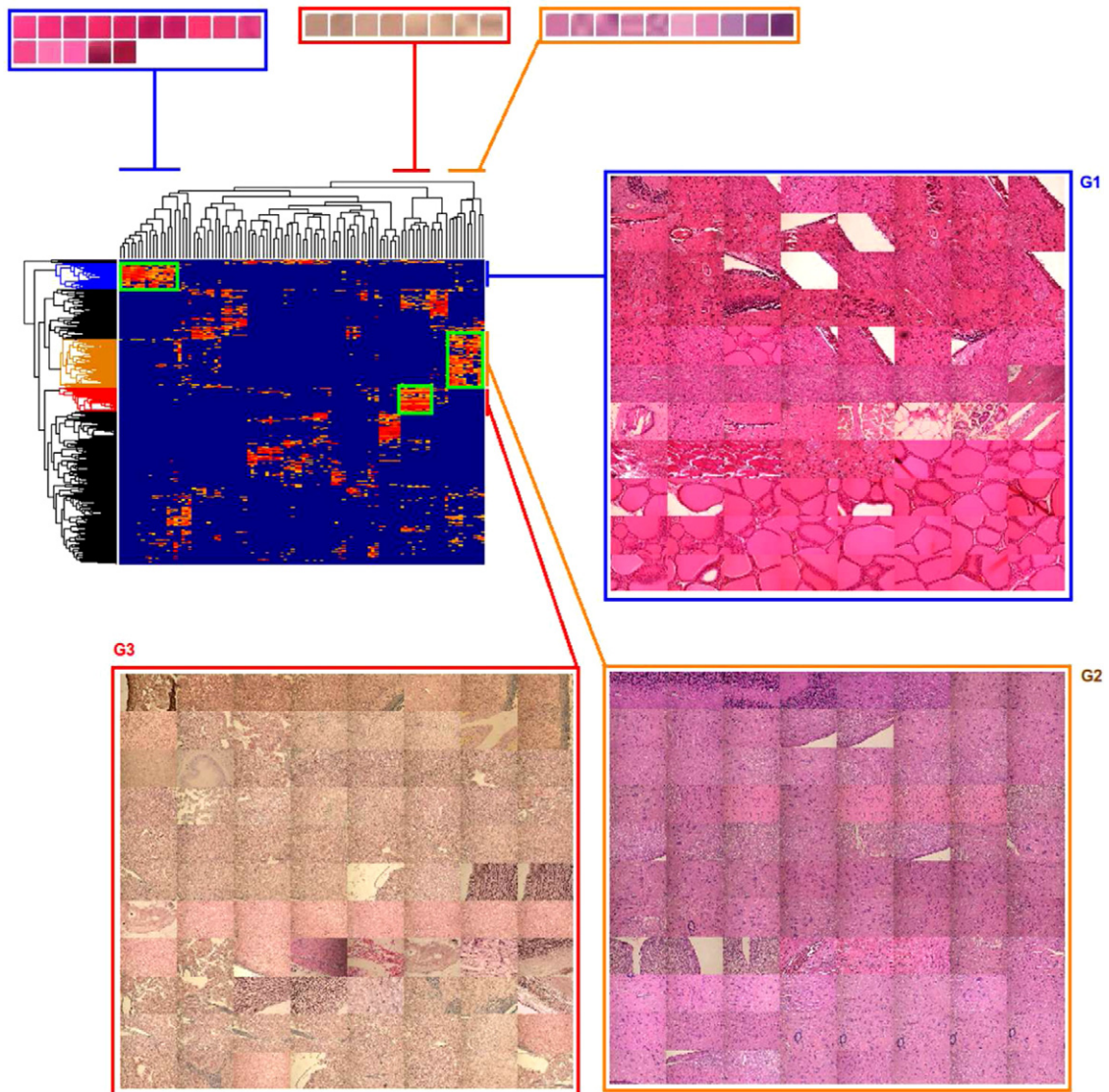


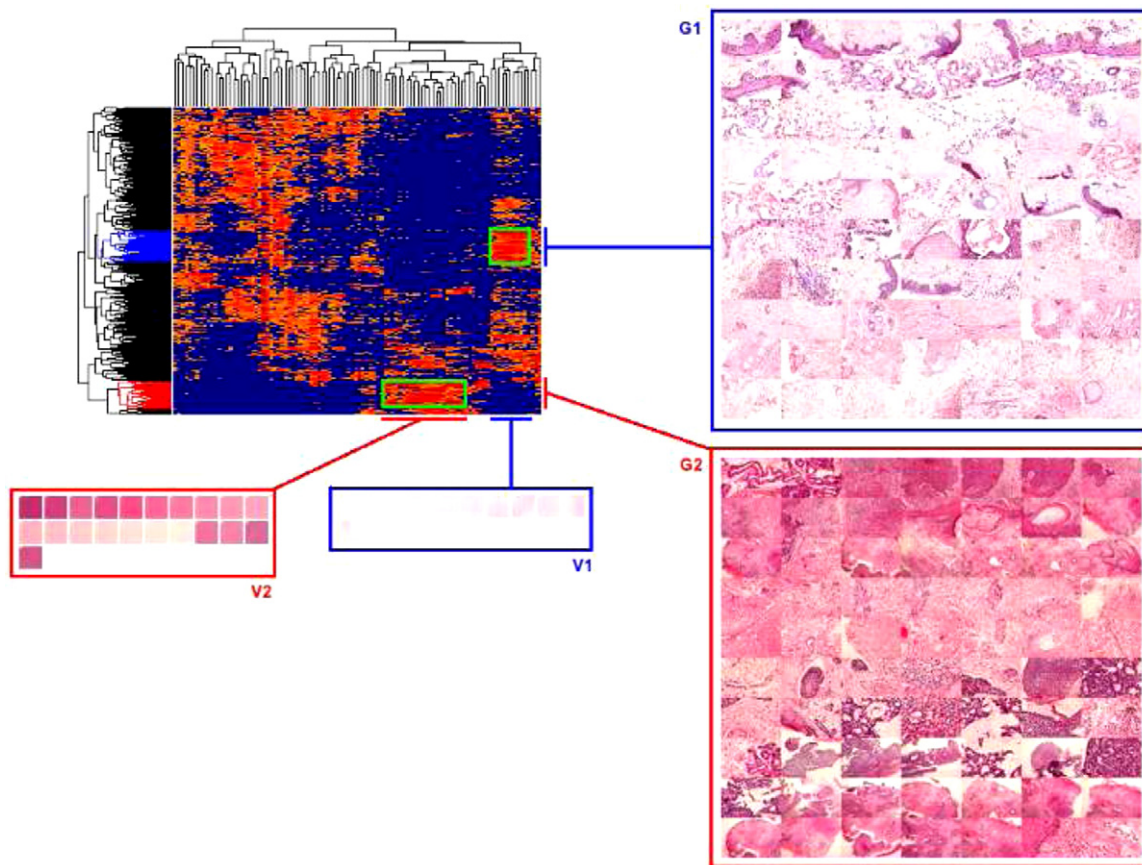
Fig. 7. Mutual information distribution for *muscular* (top-left) and *nervous* (top-right) concepts in the histology data set, and *cystic change* (bottom-left) and *necrosis* (bottom-right) in the histopathology data set. In all the cases, codewords are ranked according to their mutual information value against the respective concept.



**Fig. 8.** 100 visual words selected by mRMR method for both data sets, histopathology data set (left) and histology data set (right).



**Fig. 9.** Biclustering analysis of the histology data set using the 100 most discriminant DCT visual codewords. Rows correspond to images and columns to visual codewords. Biclusters appear as bright-red areas, which indicate that a set of related images share a set of related codewords. Three example biclusters are highlighted: G1 and G2 (mainly epithelial tissue images) and G3 (mainly nervous tissue images). The corresponding codewords are shown as well. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)



**Fig. 10.** Biclustering analysis of the histopathology data set using the 100 most discriminant DCT visual codewords. Rows correspond to images and columns to visual codewords. Biclusters appear as bright-red areas, which indicate that a set of related images share a set of related codewords. Two example biclusters are highlighted: G1 (mainly images with *eccrine glands*, *sanguineous vessel*, *lesion with fibrosis* and *pilosebaceous annexa* annotations, all of them normal) and G2 (images with pathological concepts *cystic change* and *N-P-C infiltrate*, and a normal concept, *sanguineous vessels*). The corresponding codewords are shown as well. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

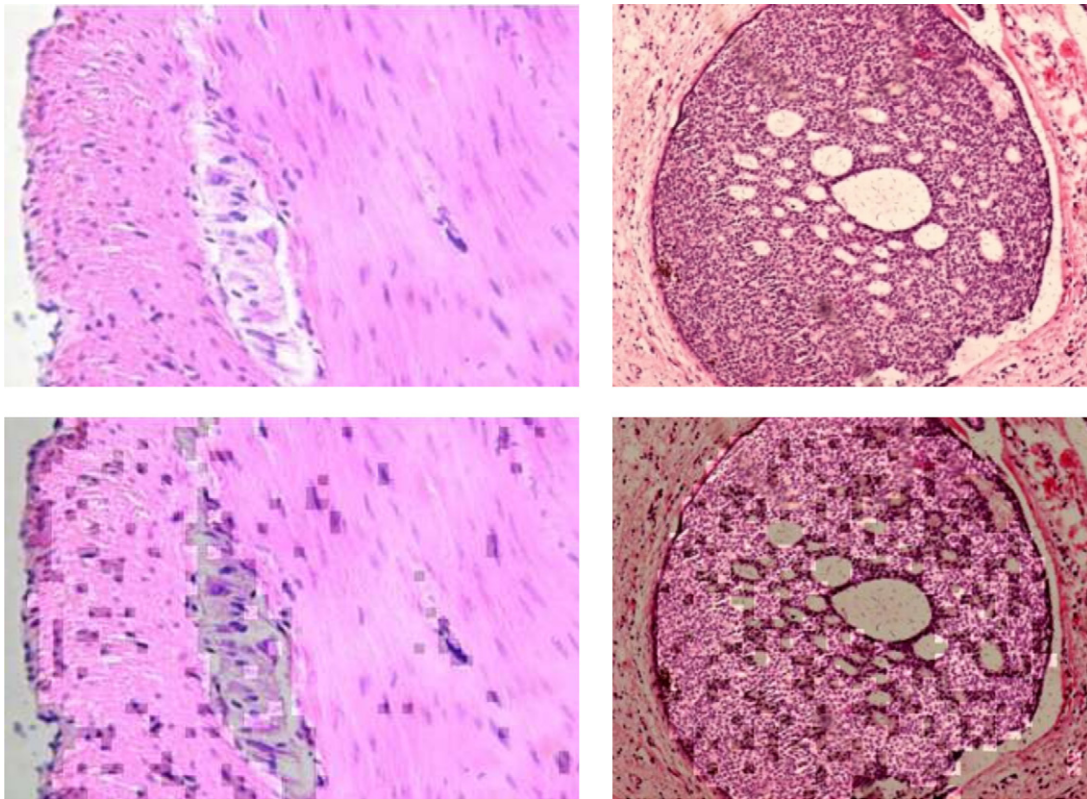
Regarding the codeword frequency distribution of the different codebooks, we want to determine whether the distribution follows the Zipf's law, which all the natural languages are known to satisfy. Zipf's law states that the frequency of any word is inversely proportional to its rank in a frequency table. Fig. 6 shows rank-vs-frequency log-log plots for both data sets and the three types of block representation in a 1000-size codebook. If the Zipf's law is satisfied the plot for each codebook should be a straight line. The central part of all the plots clearly follows this behavior, however the frequency of the least frequent codewords plunges. This behavior has been previously observed in non-medical image databases and it has been related to the fact that these codewords may be associated with noise and artifacts [59].

In the particular case of histology and histopathology images, high-frequency visual codewords can be associated with homogeneous regions, which are usually associated with the background or with tissues with low cell density, such as the stroma. These visual codewords are the analogous of *stop words* (such as prepositions, articles, connectives, etc.) in natural languages, which have a high frequency but contribute very less to the meaning. Likewise those visual words with lower frequencies falling abruptly in the plots of Fig. 6 are associated with patterns that are rare in the collection and may be related to noise/artifacts in the images or produced by the local minima of the  $k$ -means clustering algorithm.

## 7.2. Visual words related to semantic concepts

As discussed in Section 4.1, MI measures the association between concepts and visual words. MI was calculated between each visual codeword and each concept in both data sets. Different codebook sizes and different block representations were tested. In all the cases, codewords from the codebook with size 1000 exhibited higher MI values. This is (strongly) related to the fact that codewords have a higher specificity in larger codebooks. However, this does not mean that always a larger codebook is better [10]. The remaining results in this section were obtained with the 1000-size codebook. Fig. 7 shows the codeword MI distribution for four different classes, two per data set. As it can be seen, DCT-based codebooks exhibit higher MI values. This behavior is repeated through almost all the concepts in both data sets. One exception is the *cystic change* class, in the histopathology data set, where the top 300 MI values correspond to codewords using the block representation. The good performance of DCT representation is related to the fact that it can capture both texture and color information. Cystic change is a condition associated with the presence of basal-cell carcinoma, which manifests with a high density of large and dark nuclei. This seems to be better captured by the raw block representation.

The mRMR feature selection give us a subset of visual codewords with minimum redundancy and maximum relevance, which jointly are able to better discriminate the different conceptual classes in the data set. It was applied to the different codebooks correspond-



**Fig. 11.** Automatic detection of concept-related regions in images. Top images: original images from the histology data set (left) and histopathology data set (right). Bottom images: image blocks corresponding to visual codewords related with the conceptual class of the images are highlighted: *muscular tissue* (left) and *cystic change* (right). The identification is automatically performed using the procedure described in Section 4.3.

ing to the different representations. Fig. 8 illustrates the results with two subsets of 100 visual words each obtained with mRMR from both data sets using block-based representation. These subsets contain visual words that are characteristic of the different conceptual classes and that, collectively, can discriminate them.

To appreciate the difference between codewords and how they are able to discriminate the conceptual classes, the codewords were further classified according to their conditional probability  $P(C_j|w_i)$ , the probability that an image containing the codeword  $w_i$  belongs to the class  $C_j$ . A word is assigned to the class it better predicts, i.e.,  $\max_j P(C_j|w_i)$ . Table 3 shows the result for the block-based discriminative codewords for the histology data set. In general, the codewords have a high conditional probability that implies that they are highly related to the corresponding classes. The subsets of visual codewords associated to each class clearly represent the visual features that are distinctive of each conceptual class. For instance, visual codewords associated with muscular tissue present the different appearances of muscular fibers, whereas visual codewords associated to epithelial tissue present inter-cell contact points, which are characteristic of this type of tissue where cells are closely packed.

Table 4 shows the results for DCT-based visual codewords. In contrast with the results for block-based representation, these visual codewords do not codify texture information, but color information. The reason is that different stains are usually employed with different tissues, and each type of stain has a distinctive color. For example, skin images, which mainly correspond to epithelial tissue, are typically stained with hematoxylin–eosin that produces distinctive pink and purple tones. Connective tissue has a scarcer presence of nuclei and cells that causes low stain concentration. This is captured by visual codewords with brighter colors. Another important difference between the subsets of discriminative visual

codewords of both representations is that the DCT-based subset exhibits a more balanced distribution of discriminative codewords per class, as well as higher conditional probabilities. This means that this representation better captures the class differences and this is in fact corroborated in the next section, where it produces the best results in the annotation task. However, this does not mean that the DCT representation must be chosen over the block-based one in all the cases, on the contrary, they can be used in conjunction since they complementarily capture different aspects of the visual content.





Tables 5 and 6 show the two discriminative subsets obtained by mRMR using block and DCT-based representation respectively. The tables show the results for three representative conceptual classes: *cystic change*, *eccrine glands* and *NPC- Infiltrate*. Again, DCT representation exhibits a higher discriminative power than the block-based representation. Discrimination in this data set is definitely more challenging than in the histology data set. Notwithstanding, the codewords are able to capture some of the distinctive visual features of some conceptual classes. For instance, visual codewords linked to cystic change, a condition associated to the presence of basal-cell cancer, capture its distinctive agglomerative presence of nuclei and the particular dark-purple tint.

### 7.3. Biclustering analysis of semantic groups

The biclustering analysis provides an additional mechanism to detect interactions between image subsets and codewords subsets. It was applied to each data set using the subset of 100 visual words found by mRMR. The data matrix was preprocessed calculating the logarithm of the frequency values to emphasize the differences between low and high values. Fig. 9 shows the result for the histology data set using the DCT representation. Biclusters appear as

**Table 3**

Block-based visual codewords with highest conditional probabilities for each concept in histology data set. We can observe in each row the concept, followed by number of visual codewords selected with high conditional probability from the codebook, the range of class conditional probability values, and finally the first highest visual codewords.

Concept	#	$P(C_j W_i)_{min}$	$P(C_j W_i)_{max}$	Visual words
Connective	3	0.25	0.5	
Epithelial	21	0.3314	0.5698	
Muscular	18	0.3254	1	
Nervous	58	0.3501	1	



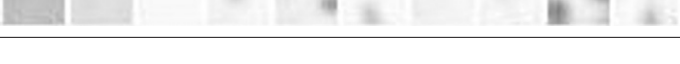
**Table 4**

DCT-based visual codewords with highest conditional probabilities for each concept in histology data set. We can observe in each row the concept, followed by number of visual codewords selected with high conditional probability from the codebook, the range of class conditional probability values, and finally the first highest visual codewords.

Concept	#	$P(C_j W_i)_{min}$	$P(C_j W_i)_{max}$	Visual words
Connective	19	0.3491	0.8631	
Epithelial	31	0.4646	0.9711	
Muscular	24	0.3706	0.9396	
Nervous	26	0.3447	0.9396	

**Table 5**

Blocks-based visual words with highest conditional probabilities for each concept in histopathology data set. We can observe in each row the concept, followed by number of visual codewords selected with high conditional probability from the codebook, the range of class conditional probability values, and finally the first highest visual codewords.




Concept	#	$P(C_j W_i)_{min}$	$P(C_j W_i)_{max}$	Visual words
Cystic change	26	0.1360	0.4	
Eccrine glands	3	0.1068	0.3497	
NPC - Infiltrate	52	0.1215	0.3207	

bright-red areas, which indicate that a set of related images share a set of related codewords. Brighter colors indicates a high frequency of these codewords in the images of the subset that can be detected. As an example, three different biclusters are highlighted showing the associated images (rows) and visual codewords (columns). Groups G1 and G2 correspond to epithelial images with two different types of stain, and group G2 mainly includes nervous tissue images.

Fig. 10 shows the biclustering of the histopathology data set using the 100 most discriminant DCT codewords. Two groups are highlighted G1 (blue) and G2 (red) with their corresponding visual words. The group G1 is mainly constituted by images with normal concepts such as *eccrine glands*, *sanguineous vessel*, *lesion with fibrosis* and *pilosebaceous annexa* which are characterized by the visual words with bright colors ranging from pink to white. On the other hand G2 is mainly composed of images exhibiting pathological con-

**Table 6**

DCT-based visual words with highest conditional probabilities for each concept in histopathology data set. We can observe in each row the concept, followed by number of visual codewords selected with high conditional probability from the codebook, the range of class conditional probability values, and finally the first highest visual codewords.

Concept	#	$P(C_j W_i)_{min}$	$P(C_j W_i)_{max}$	Visual words
Cystic change	14	0.1656	0.5493	
Eccrine glands	9	0.1032	0.5493	
NPC - Infiltrate	31	0.1131	0.3317	

**Table 7**  
Automatic annotation performance for both data sets: histopathology data set and histology data set. The average of precision, recall and  $f$ -measure on the test data sets are reported for texture features (Gabor-G, Zernike-Z and Tamura-T) and different combinations of block representations (blocks-B, SIFT-S and DCT-D) and codebook sizes for BOF representation. B+S+D corresponds to a BOF representation combining the three types of block representation.

	Texture			BOF-150			BOF-500			BOF-1000			
	G	Z	T	B	S	D	B	S	D	B	S	D	B+S+D
<i>Histopathology data set</i>													
Precision	0.14	0.00	0.28	0.32	0.33	0.44	0.44	0.36	0.56	0.40	0.40	0.59	0.70
Recall	0.09	0.00	0.03	0.14	0.18	0.28	0.19	0.13	0.23	0.16	0.11	0.23	0.25
$f$ -measure	0.10	0.00	0.06	0.19	0.22	0.34	0.25	0.18	0.32	0.22	0.17	0.32	0.30
<i>Histology data set</i>													
Precision	0.62	0.48	0.63	0.60	0.52	0.84	0.68	0.52	0.89	0.74	0.49	0.91	0.92
Recall	0.30	0.30	0.36	0.61	0.27	0.83	0.65	0.31	0.87	0.66	0.36	0.88	0.88
$f$ -measure	0.37	0.34	0.42	0.59	0.33	0.83	0.66	0.37	0.88	0.69	0.40	0.89	0.90

cepts such as *cystic change* and *N-P-C infiltrate*, as well as a normal concept, *sanguineous vessel*. Images in group G2 are related to a subset of visual codewords that mixes dark and light tones. The dark tones are mainly related to with high nucleus density, which is an indicator of tumor presence in images with *cystic change* and *N-P-C infiltrate*, whereas brighter colors are related to the absence of stain in holes like *glands* or *sanguineous vessels*.

#### 7.4. Identification of visual patterns on images

Fig. 11 shows the application of the region identification strategy shown in Section 4.3. Original images are shown at the top: an image from the histology data set exhibiting *muscular* tissue (left) and one image from the pathology data set exhibiting *cystic change* (right). The image blocks codified by visual codewords associated to the corresponding concepts were highlighted and the other image blocks were darkened. The result is shown on the bottom images. In the image with *muscular* tissue the highlighted region corresponds to a region that has an important presence of muscle fibers. In the image annotated with *cystic change*, the highlighted region exhibits a high profusion of cell nuclei, which is an important characteristic of this condition.

The results in this section illustrated the application of the proposed visual mining method to two different histology image collections. In both cases, the method was able to find relevant and meaningful visual patterns with a strong relationship with high-level concepts. The method combines different representation and analysis tools. However, the main contribution of the proposed approach does not relies on the individual methods used, but on the overall focus on the analysis of the image collection as a whole, rather than analysis of individual images.

#### 7.5. Automatic annotation performance

The annotation strategy was applied to both data sets. Different classifiers were trained for each class in each one of the data sets. The performance of each classifier was independently evaluated using the test images. The average precision and recall per class are reported in Table 7. The baseline is given by standard texture features (Gabor, Zernike and Tamura) calculated for the whole image. In all the cases, the BOF representation outperforms the baseline. Overall DCT representation exhibited the best performance among the individual features, followed by raw blocks, and leaving SIFT representation at the last place. SIFT features have been reported to produce good results in other type of images (e.g. natural scenes), but in the case of histology images the results are quite poor. The success of DCT representation is explained by the fact that it simultaneously captures both color and texture information, two important visual features in histology images. In general, the largest codebook produces the best results among the different representations in both data sets. However, the block representa-

tion reaches the top performance in the histopathology data set with a codebook of size 500. The annotation of the histopathology data set is clearly harder than the annotation of the histology data set. This is due to the fact that histopathology images involve more conceptual classes, some of them with very few samples, and some with complex visual structure that, probably, is not appropriately captured by the three representation strategies used. Additionally, a representation scheme combining the three types of BOF codebooks was tested. In this case, each image is represented by the concatenation of the three histograms. This strategy in fact produced better results in both data sets. In histopathology data set the improvement was 11% in precision and 2% in recall. In the histology data set the improvement was more discrete, 1% in precision and  $f$ -measure. The better performance exhibited by the combined BOF strategy in one of the data sets suggests the presence of complex visual patterns that involves different aspects of the visual appearance (color, texture, etc.). In conclusion, the BOF representation is a good alternative for histology image representation for automatic annotation tasks. This corroborates the main hypothesis in this work, that the local, distributed nature of BOF representation is able to capture the distinctive visual patterns in histology images.

## 8. Conclusions

The paper proposed a strategy to automatically extract visual patterns from a histology image collection. The foundation of the method is a BOF representation that builds a codebook which gathers the building blocks that explain the visual content of the image collection. A state-of-the-art feature selection process is applied to find a set of discriminative codewords. The codewords are related to high-level concepts individually, using conditional probabilities, and collectively, using biclustering.

The method was evaluated in two histology image data sets. Histology images are particularly difficult to analyze because of their high variability and complex visual structure. The method was able to successfully find visual patterns that could be related to high-level concepts. The experimental results also showed that the BOF representation is a valuable alternative for histology image representation.

The main contribution of the paper does not relies on the individual methods by themselves, but on the overall perspective that focuses on the analysis of the image collection as a whole. This novel perspective allows to use methods, such as biclustering, that traditionally have not been applied to the image analysis problem. This perspective does not replace traditional biomedical image analysis methods, but complement them. For instance, the method for automatically detecting concept-related regions in images, can extend a conventional annotation method by equipping it with an explanatory capability.

This is a first explorative work with encouraging results. It answers one important question related to the feasibility of performing visual pattern mining in histopathology image collections using a BOF representation. However, the results pose new questions which are the focus of our current and future work, including: exploration of new representation alternatives which take into account structural and multiscale information in order to capture biological and magnification variability, application to other type of biomedical images, use of other data analysis methods as latent semantic analysis, and data fusion from different sources.

## Acknowledgments

This work was partially funded by the project *Sistema distribuido de anotación automática y recuperación semántica de imágenes de histología* 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] Müller 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;73:1–23.
- [2] González FA, Romero E. *Biomedical Image Analysis and Machine Learning Technologies: Applications and Techniques*, 1st ed., Information Science Reference - Imprint of: IGI Publishing, Hershey, PA; 2009. Ch. 1. From Biomedical Image Analysis to Biomedical Image Understanding Using Machine Learning.
- [3] Bankman IN. *Handbook of medical imaging: processing and analysis*. 1st ed. San Diego, CA: Academic Press; 2000.
- [4] González FA, Romero E. *Biomedical Image Analysis and Machine Learning Technologies: Applications and Techniques*, 1st ed., Information Science Reference - Imprint of: IGI Publishing, Hershey, PA; 2009.
- [5] Lewis DD. Naive (bayes) at forty: The independence assumption in information retrieval. In: Nédellec C, Rouveirol C, editors. *Proceedings of ECML-98*, 10th European conference on machine learning, No. 1398. 1998. p. 4–15.
- [6] Tommasi T, Orabona F, Caputo B. Clef2007 image annotation task: an svm-based cue integration approach. In: Nardi A, Peters C, editors. *Working notes for the cross-language retrieval in Image Collections 2007 Workshop*. 2007.
- [7] Zheng L, Wetzel AW, Gilbertson J, Becich MJ. Design and analysis of a content-based pathology image retrieval system. *IEEE Transactions on Information Technology in Biomedicine* 2003;7(4):249–55.
- [8] Bonnet N. Some trends in microscope image processing. *Micron* 2004;35(8):635–53.
- [9] Loukas C. A survey on histological image analysis-based assessment of three major biological factors influencing radiotherapy: proliferation, hypoxia and vasculature. *Computer Methods and Programs in Biomedicine* 2004;74(3):183–99.
- [10] Caicedo J, Cruz-Roa A, González F. Histopathology image classification using bag of features and kernel functions. In: Combi C, Shahar Y, Abu-Hanna A, editors. *Proceedings of the 12th conference on artificial intelligence in medicine: artificial intelligence in medicine*, vol. 5651 of Lecture Notes in Computer Science. 2009. p. 126–35.
- [11] Cruz-Roa A, Caicedo JC, González FA. Visual pattern analysis in histopathology images using bag of features. In: Bayro-Corrochano E, Eklundh J-O, editors. *Proceedings of the 14th Iberoamerican conference on pattern recognition: progress in pattern recognition, image analysis, computer vision, and applications*, Vol. 5856 of Lecture Notes in Computer Science. 2009. p. 521–8.
- [12] Allalou A, van de Rijke FM, Tafrechi RJ, Raap AK, Wahily C. Image based measurements of single cell mtDNA mutation load. In: Ersboll B, Pedersen K, editors. *Image analysis*, vol. 4522 of Lecture Notes in Computer Science. 2007. p. 631–40.
- [13] Díaz G, González FA, Romero E. A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images. *Journal of Biomedical Informatics* 2009;42(2):296–307.
- [14] Diamond J, Anderson NH, Bartels PH, Montironi R, Hamilton PW. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human Pathology* 2004;35(9):1121–31.
- [15] Doyle S, Hwang M, Shah K, Madabhushi A, Feldman M, Tomaszewski J. Automated grading of prostate cancer using architectural and textural image features. In: 4th IEEE international symposium on biomedical imaging: from nano to macro, 2007. ISBI 2007. 2007. p. 1284–7.
- [16] 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 Recognition* 2009;42(6):1080–92.
- [17] 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. *Medical image analysis* 2009;13(1):156–66.
- [18] 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* 2009;55(1):169–83.
- [19] Orlov N, Shamir L, Macura T, Johnston J, Eckley DM, Goldberg IG. Wnd-charm: multi-purpose image classification using compound image transforms. *Pattern Recognition Letters* 2008;29(11):1684–93.
- [20] Tang H, Hanka R, Ip H. Histological image retrieval based on semantic content analysis. *IEEE Transactions on Information Technology in Biomedicine* 2003;7(1):26–36.
- [21] 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. *SPIE Medical Imaging: Computer-Aided Diagnosis* 2009;7260, 72603F1–12.
- [22] Peng H. Bioimage informatics: a new area of engineering biology. *Bioinformatics* 2008;24(17):1827–36.
- [23] Swedlow JR, Goldberg IG, Eliceiri KW. Bioimage informatics for experimental biology. *Annual review of biophysics* 2009;38(1):327–46.
- [24] Swedlow JR, Eliceiri KW. Open source bioimage informatics for cell biology. *Trends in Cell Biology* 2009;19(11):656–60.
- [25] Kvilekval K, Fedorov D, Obara B, Singh A, Manjunath BS. Bisque: a platform for bioimage analysis and management. *Bioinformatics* 2010;26(4):544–52.
- [26] Madabhushi A. Digital pathology image analysis: opportunities and challenges (editorial). *Imaging In Medicine* 2009;1(1):7–10.
- [27] Madabhushi A, Basavanally A, Doyle S, Agner S, Lee G. Computer-aided prognosis: predicting patient and disease outcome via multi-modal image analysis. In: *Proceedings of the 2010 IEEE international conference on biomedical imaging: from nano to macro, ISBI'10*. 2010. p. 1415–8.
- [28] Bosch A, Munoz X, Oliver A, Martí J. Modeling and classifying breast tissue density in mammograms. In: *Proceedings of the 2006 IEEE Computer Society Conference on computer vision and pattern recognition*, vol. 2 of CVPR '06. 2006. p. 1552–8.
- [29] Avni U, Greenspan H, Sharon M, Konen E, Goldberger J. X-ray image categorization and retrieval using patch-based visual words representation. In: *ISBI'09: proceedings of the sixth IEEE international conference on symposium on biomedical imaging*. 2009. p. 350–3.
- [30] Bosch A, Munoz X, Martí R. Review: which is the best way to organize/classify images by content? *Image and Vision Computing* 2007;25:778–91.
- [31] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 2001;42:177–96.
- [32] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *Journal of Machine Learning Research* 2003;3:993–1022.
- [33] Rogers S, Girolami M, Campbell C, Breitling R. The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2005;2:143–56.
- [34] Bicego M, Lovato P, Ferrarini A, DelleDonne M. Biclustering of expression microarray data with topic models. In: *Proceedings of the 2010 20th international conference on pattern recognition*, vol. 0 of ICPR '10. 2010. p. 2728–31.
- [35] Díaz G, Romero E. Histopathological image classification using stain component features on a pls model. In: Bloch I, Cesar R, editors. *Proceedings of the 15th Iberoamerican congress conference on progress in pattern recognition, image analysis, computer vision, and applications*, vol. 6419 of CIARP'10. 2010. p. 55–62.
- [36] Csurka G, Dance CR, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. In: *ECCV international workshop on statistical learning in computer vision*. 2004. p. 1–22.
- [37] Nowak E, Jurie F, Triggs B. Sampling strategies for bag-of-features image classification. In: Leonaridis A, Bischof H, Pinz A, editors. *Computer vision - ECCV 2006*, vol. 3954 of Lecture Notes in Computer Science. 2006. p. 490–503.
- [38] Li J, Allinson NM. A comprehensive review of current local features for computer vision. *Neurocomputing* 2008;71(10–12):1771–87.
- [39] Lowe DG. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 2004;60:91–110.
- [40] Kamiya Y, Takahashi T, Ide I, Murase H. A multimodal constellation model for object category recognition. In: Huet B, Smeaton A, Mayer-Patel K, Avrithis Y, editors. *Advances in multimedia modeling*, vol. 5371 of Lecture Notes in Computer Science. 2009. p. 310–21.
- [41] Deselaers T, Ferrari V. Global and efficient self-similarity for object classification and detection. In: *IEEE computer society conference on computer vision and pattern recognition*, CVPR 2010. 2010. p. 1633–40.
- [42] Han J, Kamber M. *Data mining: concepts and techniques*. Morgan Kaufmann; 2000.
- [43] Hsu W, Lee ML, Zhang J. Image mining: trends and developments. *Journal of Intelligent Information Systems* 2002;19:7–23.
- [44] Berlage T. Analyzing and mining image databases. *Drug Discovery Today* 2005;10(11):795–802.
- [45] Malik HH, Kender JR. Clustering web images using association rules, interest-ness measures and hypergraph partitions. In: *ICWE '06: proceedings of the 6th international conference on Web engineering*. 2006. p. 48–55.
- [46] Ribeiro M, Balan A, Felipe J, Traina A, Traina C. Mining statistical association rules to select the most relevant medical image features. In: Zighed D, Tsumoto S, Ras Z, Hacid H, editors. *Mining complex data*, vol. 165 of Studies in Computational Intelligence. 2009. p. 113–31.

- [47] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2000;22(12):1349–80.
- [48] Zhou XS, Zillner S, Moeller M, Sintek M, Zhan Y, Krishnan A, et al. Semantics and cbir: a medical imaging perspective. In: *Proceedings of the 2008 international conference on content-based image and video retrieval, CIVR '08*. 2008. p. 571–80.
- [49] Tommasi T, Orabona F, Caputo B. Discriminative cue integration for medical image annotation. *Pattern Recognition Letters* 2008;29(15):1996–2002.
- [50] Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research* 2003;3:1157–82.
- [51] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2005;27:1226–38.
- [52] Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE Transactions on Computational Biology and Bioinformatics* 2004;1(1):24–45.
- [53] Cheng Y, Church GM. Biclustering of expression data. In: Bourne PE, Gribskov M, Altman RB, Jensen N, Hope DA, Lengauer T, et al., editors. *Proceedings of the eighth international conference on intelligent systems for molecular biology*, vol. 8. AAAI Press; 2000. p. 93–103.
- [54] Prelic A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Grissem W, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 2006;22(9):1122–9.
- [55] Caicedo JC, Izquierdo E. Combining low-level features for improved classification and retrieval of histology images. *Transactions on Mass-Data Analysis of Images and Signals* 2010;2(1):68–82.
- [56] Shawe-Taylor J, Cristianini N. *Kernel methods for pattern analysis*. Cambridge, UK: Cambridge University Press; 2004.
- [57] Fletcher CDM. *Diagnostic histopathology of tumors*. Amsterdam: Elsevier Science; 2003.
- [58] Wong CSM, Strange RC, Lear JT. Basal cell carcinoma. *British Medical Journal* 2003;327:794–8.
- [59] Yang J, Jiang Y-G, Hauptmann AG, Ngo C-W. Evaluating bag-of-visual-words representations in scene classification. In: *MIR '07: proceedings of the international workshop on Workshop on multimedia information retrieval*. 2007. p. 197–206.
- [60] Iakovidis DK, Pelekis N, Kotsifakos EE, Kopanakis I, Karanikas H, Theodoridis Y. A pattern similarity scheme for medical image retrieval. *Information Technology in Biomedicine, IEEE Transactions* 2009;13(4):442–50.