Visual Pattern Analysis in Histopathology Images Using Bag of Features

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

Bioingenium Research Group
Universidad Nacional de Colombia
{aacruzr,jccaicedoru,fagonzalezo}@unal.edu.co

Abstract. This paper presents a framework to analyse visual patterns in a collection of medical images in a two stage procedure. First, a set of representative visual patterns from the image collection is obtained by constructing a visual-word dictionary under a bag-of-features approach. Second, an analysis of the relationships between visual patterns and semantic concepts in the image collection is performed. The most important visual patterns for each semantic concept are identified using correlation analysis. A matrix visualization of the structure and organization of the image collection is generated using a cluster analysis. The experimental evaluation was conducted on a histopathology image collection and results showed clear relationships between visual patterns and semantic concepts, that in addition, are of easy interpretation and understanding.

1 Introduction

Medical research centers and medical schools today are facing the problem of analyzing huge volumes of images from ongoing studies and the normal clinical operation [7]. The amount of available visual information in medicine constantly grows and discovering visual patterns in a large collection of images is a challenging task. Currently, academic image collections for classroom study or advanced research in medicine are managed by an expert who carefully organize images according to domain knowledge criteria. However, these collections have no more than a few hundred images, since the capacity of human beings to deal with large data collections is limited. Computers are an important asset to support tasks such as the analysis of image structure [5] and the identification of common and distinctive visual patterns in large image collections[6].

A large collection of medical images may be organized according to several categories that describe anatomical or pathological properties, using metadata from a hospital information system or records from a medical research survey. So, given such a collection, the main goal is the characterization of those visual properties that are common to a set of semantically related images. In the context of this paper, this problem is denoted visual pattern analysis on an image collection. The identification of visual patterns on a collection of medical images may lead to a better understanding of biological structures and also to

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 521–528, 2009.

design computer aided diagnosis tools or educational applications to train new physicians [4]. Two main questions arise when dealing with the visual pattern analysis task: how does the system detect or identify patterns that compose image structures in the collection?, and how do those visual patterns relate with pathological concepts?.

In this paper we propose a framework to answer these two questions. First, to identify visual patterns inside an image collection, the use of a bag-of-features representation is proposed, in which a dictionary or codebook is defined by grouping features extracted from all individual images. This dictionary constitutes a representative set of the visual patterns in the image collection, that can be visually understood and interpreted by domain experts, a task that is not always possible using other variety of image representations. Second, the relationships between visual patterns and semantic concepts is analysed applying two complementary strategies: a correlation analysis and a cluster analysis. The correlation analysis allows to identify a set of visual patterns that are frequently associated with particular concepts, while the cluster analysis allows to visualize the distribution of patterns for similar images and the image collection structure. This framework has been applied to a collection of histopathology images showing how both, the feature dictionary and the subsequent analysis, are revealing the visual and semantic structure of the collection.

The bag-of-features representation has been successfully applied for classification of natural scenes [2] and medical images [6], but its applicability on histopathology images has been largely unexplored [1]. This paper also aims to evaluate the suitability of this approach for histopathology images under the proposed framework. The structure of this paper is as follows: Section 2 presents details of the bag-of-features approach. Section 3 discusses the identification of semantic relationships using correlation analysis and cluster analysis. Section 4 presents the experimental results on a histopathology image collection and finally Section 5 presents the conclusions and future work.

2 The Bag-of-Features Representation

The *bag-of-features* representation 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*, that is, a visual vocabulary, in which the most representative patterns are codified as *codewords* or visual words. Then, the image representation is generated through a simple frequency analysis of each *codeword* inside the image. This representation has been successfully applied in different image classification tasks. There are three main steps to build a *bag-of-features* representation [2]: (i) feature detection and description; (ii) codebook generation; and, finally; (iii) the *bag-of-features* construction. Figure 1 shows an overview of those steps. The *bag-of-features* approach is a novel and simple method to represent image contents using collection dependent patterns.

In this work the following strategy has been used to generate the bag of features representation for histopathology images: for feature detection, raw blocks

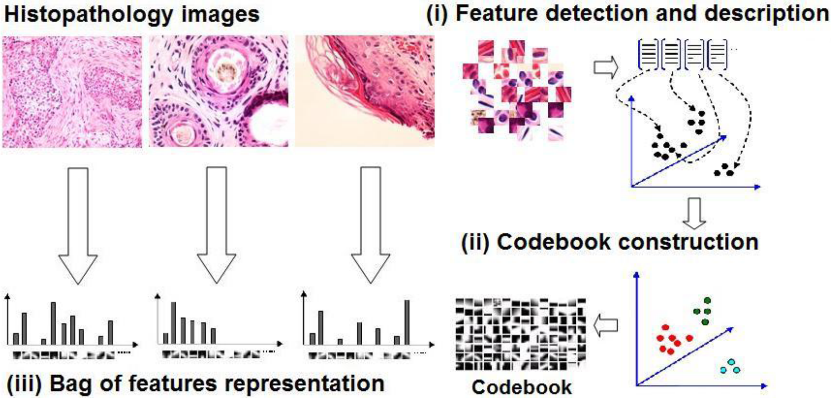
Histopathology images
(i) Feature detection and description
(ii) Codebook construction
(iii) Bag of features representation
Codebook

Fig. 1. Overview of the Bag of Features representation

are extracted from a regular grid on each image using 8 × 8 pixels per block. Each block is represented by the array of 64 gray level values which is used as feature vector. For codebook generation, the k-means algorithm is applied over the whole set of blocks. The size of the codebook, k, is an important parameter. It is expected that a moderately codebook size, in the order of hundreds, wold be enough to capture the most important patterns in the collection [1]. For the experimentation carried on in the present work, $k = 50$ was used based on previous findings in the same image collection [1]. Finally, the bag of features for each image is generated by counting the occurrence of visual words in the codebook.

3 Visual Pattern Analysis

The bag-of-features codebook constitutes a summary of the visual patterns present in the histopathology image collection. The hypothesis is that some of these visual patterns are related to histopathology concepts. In order to corroborate it, two strategies are applied, a correlation analysis and a cluster analysis.

3.1 Correlation Analysis

The goal of the correlation analysis is to measure the strength of the relationship between a particular visual pattern from the dictionary and a semantic concept. Images in the collection are known to be in one or several predefined categories or semantic classes. Then, we assume two random variables to analyse the correlation between them: semantic concepts and visual patterns. For semantic concepts the random variable is binary and indicates the presence or absence of the concept in the image. For visual words, the random variable is assumed continuous and corresponds to the relative frequency of the visual word in the image.

Following these assumptions, we can evaluate the correlation of visual patterns and semantic concepts. When a particular concept and a visual pattern are constantly exhibited in an image set, it is expected that the correlation between them has a positive value. On the other hand, if the visual pattern is not usually in those images that exhibit the concept, then a negative correlation is expected. Hence, the correlation analysis is useful to identify the set of most representative visual patterns associated to semantic concepts.

3.2 Clustering Analysis

A natural basis for organizing visual patterns is to group together those that share similar occurrence in images. The purpose of this cluster analysis is to generate a reordering of visual patterns to analyse the relationships with semantic concepts. Due to the large amount of images in a collection and also to a potential large dictionary of visual patterns, it is difficult to assimilate underlying relationships. Therefore, we follow a visual representation that is usually applied in bioinformatics to visualize and explore gene expression data in an intuitive manner for biologists [3]. We combine clustering methods with a graphical representation of the visual patterns in images by representing each occurrence value using a color in a matrix, as it is shown in Figure 4. A blue color indicates a low frequency of visual patterns in images, while a red color indicates a high frequency of the pattern. Other ranges of blue and yellow indicate intermediate frequencies. Each row in the matrix represents an image and each column represents a visual pattern.

We use agglomerative hierarchical clustering, with average linkage, to organize both, rows and columns in the matrix and the corresponding dendrogram is also drawn alongside the matrix representation. The distance measure applied in this work is Euclidean distance among bag-of-features representations (rows) and the occurrence of visual patterns in all images (columns). This analysis is expected to organize rows such that images in each group share a semantic concept. It highly depends on the bag-of-features representation, so that we can evaluate how good this representation is for semantic image contents. In addition, the column organization is expected to reveal the set of visual patterns that are related to particular semantic concepts.

4 Results

The image dataset used in this work is a set of histopathology images used to diagnose a special skin cancer known as basal cell carcinoma. This dataset has been used in previous studies for automatic image annotation and retrieval [1]. A subset of this collection has been selected to analyse the structure of 4 histopathology concepts (cystic change, lesion with fibrosis, morpheaform pattern and pilosebaceous annexa). This subset of images sums up to 348 images processed for this study (67, 90, 37 and 154 for each concept class respectively).

4.1 Correlation Analysis

The correlation analysis shows that some visual words are more relevant to identify some particular concepts than others. Figure 2 shows how the four concepts

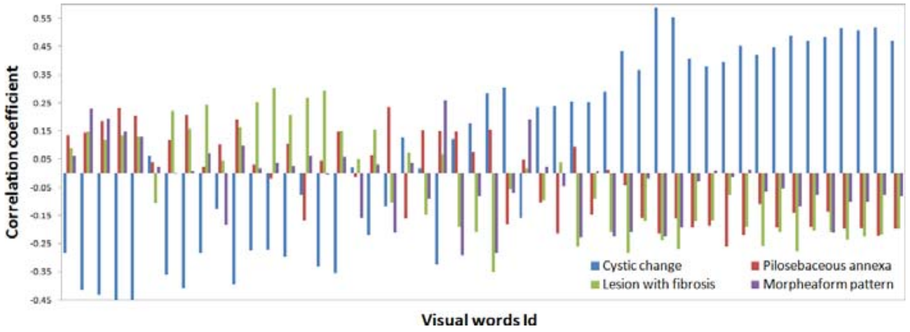

Fig. 2. Correlation coefficient measures between high-level concepts and visual words. Visual words in horizontal axis are sorted by frequency of occurrence from left to right in descending order.

Table 1. Ten visual words with highest correlation value for each concept

Concept	1	2	3	4	5	6	7	8	9	10
Cystic change										
Pilosebaceous annexa										
Lesion with fibrosis										
Morpheaform pattern										

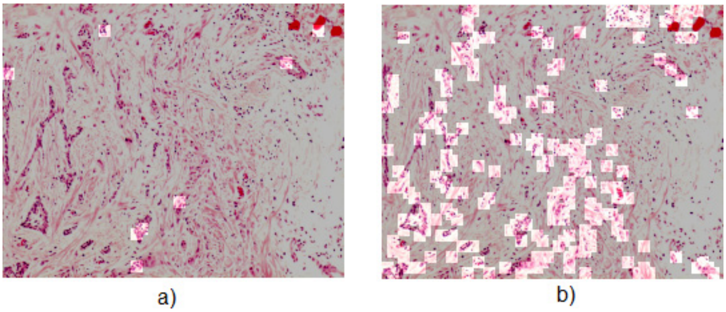

Fig. 3. Spatial location of visual patterns in an image in the category *lesion with fibrosis*. a) highlighted blocks are the ten most correlated visual patterns for the *cystic change* concept. b) highlighted blocks are the ten most correlated visual patterns for *lesion with fibrosis*.

are correlated with each visual pattern. Note that *cystic change* is highly correlated with a set of visual patterns that other concepts are not. It can also be observed for *lesion with fibrosis.* For all concepts it is possible to identify a set of highly correlated visual patterns, since the plot in general shows that patterns with high correlation with a concept present low correlation with others.

Table 1 shows top the ten visual words with highest correlation for each concept. The correlation analysis assigns to each concept a set of visual words. *Cystic change*, for example, is more correlated with dark elements and parts of big circular patterns, which is consistent with a notion of large and dense cells and nuclei. On the other hand, *Lesion with fibrosis* shows small gray points over a bright background.

Figure 3 shows the spatial location of visual patterns on an image of the category *lesion with fibrosis.* Relevant visual word are shown as blocks with a lighter color. Subfigure 3.a) highlights the top-ten visual patterns from the *cystic change* category, showing a low presence of those patterns. On the other hand Subfigure 3.b) highlights the top-ten visual patterns of the *lesion with fibrosis* category, which are clearly more frequent in the image.

4.2 Cluster Analysis

Cluster analysis allows to distinguish groups of related visual patterns and a general organization of images and concepts in the collection under the bag-of-features representation. It is achieved using a graphical representation of the data, indicating occurrence values in a colored matrix. Colors range from dark blue, indicating a very low frequency, to red, indicating high frequency values. To plot this matrix, the 6 most frequent visual words were ignored since they usually correspond to background and do not have discriminative power. Figure 4 shows the obtained matrix for all images in the analysed collection, with visual patterns from the codebook organized in columns, and images organized in rows. The clustering algorithm reordered rows and columns according to their similarity.

This matrix shows group of images related to groups of visual patterns. For instance, in Figure 4 a red box and a black box in the upper-left corner of the matrix shows two different groups of images with a high frequency of several visual patterns. In the vertical dendrogram these groups are colored with green and blue respectively and all of the images in them present the *cystic change* concept. The left side of the figure shows the images and the visual words associated with those regions of the cluster matrix. The orange box, in the same figure, shows how other images in the red portion of the vertical cluster present a high frequency for other visual words. In this group, there are images with other concepts, mainly *pilocebaceus annexa* and l*esion with fibrosis.*

The cluster analysis shows that it is possible to find visual patterns that can be associated with semantic concepts. The visual representation makes it easier the task of finding those visual patterns. In this particular example, the class of images tagged with the *cystic change* concept are clearly differentiated from the other classes by a characteristic set of low-level visual patterns associated with large cells and nuclei.

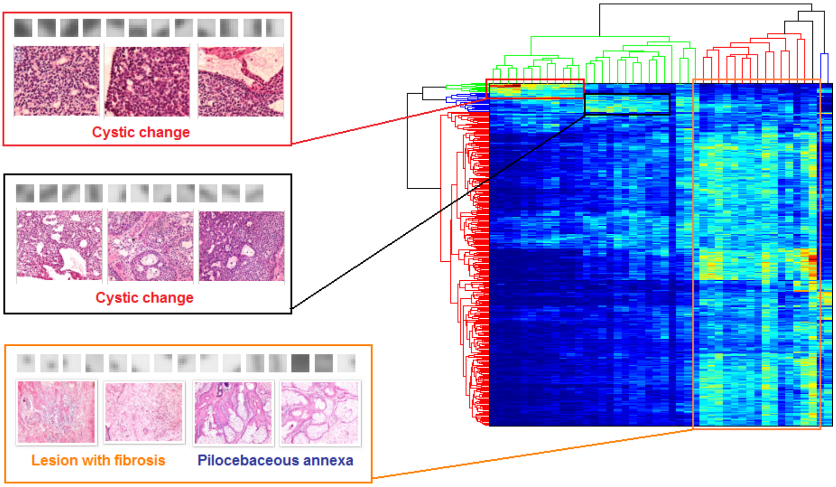

Fig. 4. Cluster analysis on the complete image dataset with 4 concepts. The rows of the matrix correspond to images and the columns correspond to visual words. The color of the matrix represent the frequency of the visual words for each image: blue represents low frequency, red represents high frequency. Both, images and visual words are clustered using hierarchical clustering. The result is represented by the vertical and horizontal dendograms. Three different regions of the matrix are marked by colored boxes. The corresponding visual words, concepts and sample images of these regions are detailed in the left side rectangles.

5 Conclusions and Future Work

This paper has presented a framework to identify and analyse visual patterns in a collection of medical images using a bag-of-features representation. The main hypothesis of this paper was that visual words, identified in the collection using the bag-of-features representation, can be related to semantic concepts in histopathology images. The hypothesis was corroborated by the exploratory experiments based on correlation and cluster analysis. These results suggest that this representation may be useful for analysis and understanding of histopathology images. The cluster analysis is analogous to the one used in bioinformatics to analyse gene array data, where the goal is, e.g., to find how a diseases relates to the presence or absence of a particular gene. In the image analysis context, visual words are analogous to genes with the important advantage that they could be directly related to specific regions of particular images. This kind of analysis is not possible with other image descriptors such as moments, histograms or transformation coefficients. In addition, these analysis may help to design and improve automatic tools to manage image collections, such as image retrieval systems. For instance, understanding the group of visual patterns that better describe a set of concepts, weighting schemes or pruning strategies may be applied in a more informed fashion.

References

1. Caicedo, J.C., Cruz, A., Gonzalez, F.: Histopathology image classification using bag of features and kernel functions. In: Combi, C., Shahai, Y., Abu-Hanna, A. (eds.) Artificial Intelligence in Medicine (AIME 2009). LNCS (LNAI), vol. 5651, pp. 126–135. Springer, Heidelberg (2009)
2. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision (2004)
3. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95(25), 14863–14868 (1998)
4. Iakovidis, D., Pelekis, N., Kotsifakos, E., Kopanakis, I., Karanikas, H., Theodoridis, Y.: A pattern similarity scheme for medical image retrieval. IEEE Transactions on Information Technology in Biomedicine (2008)
5. Ogiela, M.R., Tadeusiewicz, R.: Artificial intelligence structural imaging techniques in visual pattern analysis and medical data understanding. Pattern Recognition 36(10), 2441–2452 (2003)
6. Orabona, F., Caputo, B., Tommasi, T.: Clef 2007 image annotation task: An svm - based cue integration approach. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
7. Yang, G., Yu, X., Zhuang, X.: The current status and development of pattern recognition diagnostic methods based on medical imaging. In: IEEE International Conference on Networking, Sensing and Control, 2008. ICNSC 2008, pp. 567–572 (2008)