Content Based Image Retrieval For Histology Image Collection Using Visual Pattern Mining

U.Ravindran, T.Shakila

Abstract - CBIR is trending to an enormous growth in the field of Artificial intelligence based on visual pattern mining in histology images extends the boundaries of CBIR in Genetic Research. This method starts by representing the visual content of the collection using a bag-of-feature 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. Additionally includes an interpretation mechanism that associates concept annotations with corresponding image. A neural network as an classifier, It can improve the classification accuracy more than 80%.

Index Terms — Content-based image retrieval, Visual pattern mining, Bag of features (BOF), Visual-codebook, feature selection, Image annotation, Histology and histopathology images, Fundamental tissues

1 INTRODUCTION

BIOMEDICAL images are an important Source of information, and a potential source of knowledge, for both routine clinical decision and biomedical research[2]. Nevertheless, a thorough exploitation of this potential requires techniques able to automatically extract information and knowledge from this vast amount of data done on the area of medical imaging, which is gradually moving from computer assisted image analysis systems, mainly based on image processing techniques, to fully automatic systems based on pattern recognition and machine learning methods. 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. The paper is organized as follows:

Section 1 presents the proposed method content based image retrieval [CBIR] for visual pattern mining; Section 2 describes the details of the different stages of the image collection representation strategy based on BOF; Section 3 an 4 describes the proposed automatic annotation, selection and classification strategy; finally, conclusions and future work are discussed in Section 6 and 7.

2 PREVIOUS WORK

The previous approach is composed of three main phases: a pre-processing step, which corrects luminance differences. A segmentation step that uses the normalized RGB color space for classifying pixels either as erythrocyte or background followed by an Inclusion-Tree representation that structures the pixel information into objects, from which erythrocytes are found. Finally, a two step classification process identifies infected erythrocytes and differentiates the infection stage, using a trained bank of classifiers. State-of-the-art nonlinear classifiers were evaluated for these phases: a multilayer perceptron neural network (MLP) [27] and a support vector machine (SVM).

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 a 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 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. 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. In proposed neural network can be used as an classifier for improving accuracy and speed than svm in both testing and training. Using Neural network can solve non-linear problem along this direction and many retrieval systems, both research [1] and commercial, have been built. Low-level features such as colors, textures and shapes of objects are widely used for CBIR. However, in specific applications, such as medical imaging, low-level features play a substantial role in defining the content of the data[7]. A typical content based image retrieval system is depicted in Figure 1. There is a growing interest in CBIR because of the limitations inherent in metadata-based systems, as well as the large range of possible uses for efficient image retrieval. Textual information about images can be easily searched using existing technology, but requires humans to personally describe every image in the database.

3 CONTENT BASED IMAGE RETRIEVAL

Content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) is the application of computer vision techniques to the image retrieval problem, that is, the problem of searching for digital images in large databases. Content-based means that the search will analyze the actual contents of the image rather than the metadata such as keywords, tags, and/or descriptions associated with the image. The term 'content' in this context might refer to colors, shapes, textures, or any other information that can be derived from the image. CBIR is desirable because most web based image search engines rely purely on metadata and this produces a lot of garbage in the results. Also having humans manually enter keywords for images in a large database can be inefficient, expensive and may not capture every keyword that describes the image. Thus a system that can filter images based on their content would provide better indexing and return more accurate results. In CBIR, images are automatically indexed by summarizing their visual contents through automatically extracted quantities, or features, such as color, texture or shape. Thus, low-level numerical features, extracted by a computer, are substituted for higher-level, text-based, manual annotations or keywords. In inception of CBIR, many techniques have been developed

4 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[10] is a histogram generated through a simple frequency analysis of each codeword inside the image.

The 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.

4.1 Feature Extraction and Representation

The BOF approach starts extracting small blocks (in the present work, 8×8 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.

The regular-grid-based extraction is used; this process takes 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.

Three different strategies that have produced good results when used in conjunction with the BOF representation. The first strategy uses the raw block, i.e., the feature vector has 64 values corresponding to the luminance values of the corresponding pixels.

The second block-representation strategy is based on scale invariant feature transform (SIFT) points. 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 4×4 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.

Finally, the third strategy is the discrete cosine transform (DCT) [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.

4.2 Code Book 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 code words. Clustering algorithm has not a big impact in the classification of natural images compared with a random selection of code words. However, this is not necessarily the case for histology images. An important decision in the construction of the codebook is the size selection, that is, how many code words are needed to represent image contents. According to different works on natural image classification, the larger the codebook size, the better. However found that the size of the codebook is not a
significant aspect in a medical image classification task.

4.2.1 VECTOR QUANTIZATION

Vector quantization is a classical quantization technique from signal processing which allows the modeling of probability density functions by the distribution of prototype vectors. It was originally used for data compression. It works by dividing a large set of points (vectors) into groups having approximately the same number of points closest to them. Each group is represented by its centroid point, as in k-means and some other clustering algorithms.

The density matching property of vector quantization is powerful, especially for identifying the density of large and high-dimensional data. Since data points are represented by the index of their closest centroid, commonly occurring data have low error, and rare data high error. This is why VQ is suitable for lossy data compression. It can also be used for lossy data correction and density estimation.

4.2.2 K-Mean Clustering

In data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results in a partitioning of the data space into Voronoi cells. The problem is computationally difficult (NP-hard), however there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum.

1. To minimize sum of squared Euclidean distances between points $x_i$ and their nearest cluster centers $mk$
   
   Algorithm:
   1. Randomly initialize K cluster centers
   2. Iterate until convergence
   3. Assign each data point to the nearest center
   4. Recompute each cluster center as the mean of all points

   $$D(X, M) = \sum_{\text{Cluster } k} \sum_{\text{point } i} (x_i - mk)^2$$
   
   These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data, however k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

4.3 FEATURE SELECTION AND ANALYSIS

Feature selection, also known as variable selection, feature reduction, attribute selection or variables/Subset selection, is the technique of selecting a subset of relevant features for building robust learning models. Feature selection is a particularly important step in analyzing the data from many experimental techniques in biology, such as DNA microarrays, because they often entail a large number of measured variables (features) but a very low number of samples. By removing most irrelevant and redundant features from the data.
Feature selection helps improve the performance of learning models by:
1. Alleviating the effect of the curse of dimensionality.
2. Enhancing generalization capability.
3. Speeding up learning process.

5 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. The main goal of the system is to find visual patterns that can be associated with the high-level annotations. The two analysis can be used in visual mining they are
1. Visual word discrimination analysis
2. Biclustering analysis

5.1 VISUAL WORD DISCRIMINATION ANALYSIS

Visual discrimination analysis used for good representatives of particular classes, i.e., codewords with a high discriminative power. In the general 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. Mutual information (MI) is a popular approximation to measure feature relevance.

6 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]. 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.

7 CLASSIFIER TRAINING

In image classification, an image is classified according to its visual content. For example, does it contain an airplane or not. An important application is image retrieval - searching through an image dataset to obtain (or retrieve) those images with particular visual content. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

A classification [19] task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time.
Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time.

The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating.

In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown.

A neural network used as a classifier, neural network consists of units (neurons), arranged in layers, which convert an input vector into some output. Each unit takes an input, applies a (often nonlinear) function to it and then passes the output on to the next layer. Generally the networks are defined to be feed-forward: a unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one unit to another, and it is these weightings which are tuned in the training phase to adapt a neural network to the particular problem at hand. This is the learning phase.

8 ANNOTATION

Automatic image annotation (also known as automatic image tagging or linguistic indexing) is the process by which a computer system automatically assigns metadata in the form of captioning or keywords to a digital image. This application of computer vision techniques is used in image retrieval systems to organize and locate images of interest from a database.

It can be regarded as a type of multi-class image classification with a very large number of classes - as large as the vocabulary size. Typically, image analysis in the form of extracted feature vectors and the training annotation words are used by machine learning techniques to attempt to automatically apply annotations to new images.

The first methods learned the correlations between image features and training annotations, then techniques were developed using machine translation to try to translate the textual vocabulary with the ‘visual vocabulary’, or clustered regions known as blobs.

9 HISTOLOGY IMAGE DATA SETS

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 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.
9.1 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 and at different magnifications and cuts.

<table>
<thead>
<tr>
<th>Concept</th>
<th>#Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connective</td>
<td>484</td>
</tr>
<tr>
<td>Epithelial</td>
<td>804</td>
</tr>
<tr>
<td>Muscular</td>
<td>514</td>
</tr>
<tr>
<td>Nervous</td>
<td>1026</td>
</tr>
</tbody>
</table>

10 CONCLUSION

A strategy automatically extracts 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 using 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 rely 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.

It 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.

11 FUTURE ENHANCEMENT

The 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.

REFERENCES


