UNIVERSIDAD NACIONAL DE COLOMBIA

Medical Image Retrieval Using Multimodal Semantic Indexing

Jorge Andrés Vanegas Ramírez

Universidad Nacional de Colombia
Engineering School
Bogotá, D.C., Colombia
2013

Master's Thesis

Medical Image Retrieval Using Multimodal Semantic Indexing

by

Jorge Andrés Vanegas Ramírez

Submitted to the Engineering School of the Universidad Nacional de Colombia, in partial fulfillment of the requirements for the degree of

Master of Science in Systems and Computer Engineering

Under the guidance of

Fabio A. Gonzalez, Ph.D.
Associate Professor
Engineering School

MindLab Research Group
Universidad Nacional de Colombia
Engineering School
Bogotá, D.C., Colombia
2013

Acknowledgments

This thesis would not have been possible without the help, support, guidance and patience of my thesis advisor, Professor Fabio Gonzalez.

Also, I would like to express my very sincere gratitude to Dr. Juan C. Caicedo, for his support and good advice. And, I want to thank all the people who are part of the research groups MindLab and BioIngenium, who have become not only colleagues but also good friends.

Finally, but most important of all, I am grateful to my parents, Gustavo Vanegas and Belén Ramírez, who have always given me their unconditional support, encouragement and advice, so that I could concentrate on my thesis.

Abstract

Large collections of medical images have become a valuable source of knowledge, taking an important role in education, medical research and clinical decision making. An important unsolved issue that is actively investigated is the efficient and effective access to these repositories.

This work addresses the problem of information retrieval in large collections of biomedical images, allowing to use sample images as alternative queries to the classic keywords. The proposed approach takes advantage of both modalities: text and visual information. The main drawback of the multimodal strategies is that the associated algorithms are memory and computation intensive. So, an important challenge addressed in this work is the design of scalable strategies, that can be applied efficiently and effectively in large medical image collections.

The experimental evaluation shows that the proposed multimodal strategies are useful to improve the image retrieval performance, and are fully applicable to large image repositories.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Large collections of medical images have become a valuable source of knowledge, taking an important role in education, research and as support for making decisions and collaboration in pathologies identification [1, 2]. One of the main problems is that the size of the collections has been rising precipitously in recent years thanks to new acquisition facilities. Modern hospitals may produce more than 100,000 images a day; this is about 100 GB of data [3]. This generates huge repositories of valuable information, which in many cases is difficult to process and manage appropriately. This makes necessary the development of tools for an effective and an efficient access to this type of information. This has caused an increasing interest in this research field during the last years. These efforts have been embodied in several medical image retrieval systems [4], as are the ASSERT [5], CasImages, MedGIFT and IRMA [6] systems, among others.

Also, the awareness of the importance of such systems has increased in recent years. Imaging systems and image archives have often been described as an important economic and clinical factor in the hospital environment. Several radiological teaching files exist and radiology reports have also been proposed in a multimedia form. Currently, there exist several Web-interfaces to medical image databases [4]. Datasets of medical images have often been used for retrieval systems and the medical domain is often cited as one of the principal application domains for content-based access technologies in terms of potential impact.

This work addresses the problem of information retrieval in large collections of biomedical images. The main goal is to propose a strategy for image search that allows using keywords or sample images as queries. To accomplish this, the proposed strategy should allow processing and data management of the two modalities: text and visual features. One of the main problems to address is the construction of a multimodal search index, i.e., integrating visual and textual information in the same data structure. This scheme would provide the facility to search for images using keywords, as do classic search systems, or by example images, which in the medical case could be associated with a diagnostic image for which we wish to obtain similar images as a reference. This strategy would allow that even the images that have no related textual content can be recovered.

1.1 Thesis goals

The main goal of this research was to design and implement a strategy for biomedical image retrieval in collections with mixed information: images and text, using a multimodal search index. The following is the description of the general purposes of this research:

- To identify, collect and depurate the biomedical images collection with their textual descriptions.
- To adapt and implement automatic extraction methods for text documents in the collection.
- To adapt and implement image-processing methods to extract the visual features that represent the image content.
- To design and/or adapt a strategy to represent the contents of documents and images in a multimodal search index.
- To implement a prototype system for medical image retrieval using the multimodal index.
- To conduct a performance evaluation of the image retrieval algorithms proposed.

1.2 Main contributions

The following is the outline of the main contributions of this work:

1.2.1 Design and implementation of indexing and search strategies for medical images

A semantic model for histology images indexing is presented. This model finds the relationships between visual features and text terms directly. The method has demonstrated to be a good alternative for image indexing when the textual modality is clean and structured. This work was published in:

- Vanegas J.A., Caicedo J.C., González F.A. & Romero E., *Histology Image Indexing Using a Non-negative Semantic Embedding.* In Medical Content-Based Retrieval for Clinical Decision Support. MICCAI 2011. Volume 7075, 2012, pp 80-91, ISBN:978-3-642-28459-5, ISSN:0302-9743. Springer-Verlag GmbH Berlin Heidelberg.

A strategy for images representation by fusing visual and semantic features was proposed. The proposed fusion strategy is based on projecting semantic data to the visual feature space. Parts of this work were published in:

- Vanegas J.A., Caicedo J.C., González F.A., *Histology Image Indexing Combining Visual And Semantic Features.* 7th International Seminar on Medical Information Processing and Analysis. SIPAIM 2011.

1.2.2 Design of scalable indexing strategies

The proposed content based image retrieval system requires two computationally expensive tasks, that difficult its applicability to large collections of images. The first task is the representation of visual contents and the second task is learning the multimodal semantic representation.

For the representation of visual contents its is proposed a parallelized Bag-of-Features based on Map-Reduce architecture. This strategy allows to index large collections of images by dividing the computing workload in multiple processing units. This work was published in:

- Vanegas, J.A., Caicedo J.C, & González F., *Scalable Construction of a Bag of Features Representation Using the Map-Reduce Architecture.* The Latin American Conference on High Performance Computing. CLCAR 2012.

Finally, to achieve and scalable solution for learning a multimodal semantic representation, the Online Non-negative Semantic Embedding is proposed to reduce the computational requirements, both in terms of memory and processing time. This work was published in:

- Vanegas J.A and González F.A., *Large Scale Image Indexing using Online Non-negative Semantic Embedding.* In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2013, Vol. 8258, 2013, ISBN 978-3-642-41821-1, Springer-Verlag

1.2.3 Applications and other contributions

1.2.3.1 Histology image retrieval system

As part of this work, a prototype system for medical images retrieval was implemented, which is intended as a proof of concept for the proposed computational models. The system can be found online at `http://168.176.61.90:9100/mirs/`, and the source code is accessible at `http://code.google.com/p/bioingenium-large-scale-tools`

1.2.3.2 Large scale algorithms

As part of the results of this work, many algorithms for large scale image indexing, have been implemented. The source code of the implemented algorithms is accessible at `http://code.google.com/p/bioingenium-large-scale-tools`.

1.2.3.3 ImageCLEFmed challenge

The team at our research lab has participated in the ImageCLEFmed campaign, which is a medical image search contest for researchers. Some of the proposed representations have been tested in the context of this challenge. Experiments have been prepared for the 2012 version of this challenge, using improved text and visual representations. Both data modalities were employed separately in these experiments, and the results ranked our text strategy in the first place, among 54 experiments submitted by other research groups. Also, the visual strategy was ranked in the third place, among other 36 experiments.

The official results can be found in the website of ImageCLEFmed 2012 [1]

- Vanegas, J.A., Caicedo, J. C., Camargo, J., Ramos-Pollán, R., & Gonzalez, F. *Bioingenium at ImageCLEF 2012: Text and Visual Indexing for Medical Images.* CLEF (Online Working Notes/Labs/Workshop).

1.3 Thesis organization

Part 1: Preliminaries:

[1] http://www.imageclef.org/medical/2012

- *Chapter 1: Introduction.* This chapter presents an introduction to the research problem, its goals and the main contributions.

- *Chapter 2: Problem Statement.* This chapter exposes the general background and definition of the research problem. First, the scope of this work is defined and then, the main approaches in content-based image retrieval are presented and discussed.

- *Chapter 3: Basic Notions and Definitions.* This chapter presents detailed information about the kind of images to address in this work, and defines the required theoretical basis for the development of this work.

Part 2: Methods:

- *Chapter 4: Image Indexing using a Non-Negative Semantic Embedding.* This chapter presents a method for image indexing that uses the textual modality as a semantic representation by modeling the relationships between visual features and text terms directly. This method has demonstrated to be a good alternative for image indexing when the textual modality is clean and structured.

- *Chapter 5: Fusing Visual and Semantic Contents.* This chapter presents an strategy for image representation by combining visual and semantic features, in order to exploit the best properties from each data modality.

Part 3: Efficiency and Scalability.

- Chapter 6: *Parallelized Bag-of-Features.* In this chapter it is proposed a distributed implementation of the classical Bag-of-Features algorithm using a Map-Reduce architecture.

- Chapter 7: *Online Non-negative Semantic Embedding Model.* In order to make feasible the semantic embedding strategy to large images collections, in this chapter it is presented a reformulation of NSE as an online algorithm using stochastic gradient descent approach.

Part 4: Conclusions

- Chapter 8: *Conclusions and Future Work.* Presents a general discussion about the results of this research, followed by the main conclusions and the future work in this area.

Chapter 2

Problem statement

The current commercial systems for image searching on the web, allows to search images using keywords, and also provides links to find similar images in terms of visual content. The strategies used in these systems are mainly based on a single kind of information modality (visual only or text only). But the information is usually contained in one modality is complementary to the other, so that it would be very desirable to find an optimal way to combine this information, to obtain a more detailed representation of each image.

To provide an efficient access to large biomedical imaging repositories, it is proposed to design a strategy for image search that allows using keywords or sample images as queries. To accomplish this, the proposed strategy should allow processing and data management of the two modalities: text and visual features. One of the main problems to address is the construction of a multimodal search index, i.e., integrating visual and textual information in the same data structure. This scheme would provide the facility to search for images using keywords, as classical search systems do, or by example images, which in the medical case could be associated with a diagnostic image for which we wish to obtain similar images as a reference. This strategy would allow that even the images that have no related textual content can be retrieved.

2.1 Scope

This research studied the problem of Content-Based Image Retrieval and its application in the biomedical field. The main goal of this research is to propose an effective relevant image search method in a large collection of biomedical images exploiting multimodal information found in such repositories. Two main features have been investigated in the proposed strategy: effectiveness and efficiency. Effectiveness deals with the problem of retrieving useful results for the user, according to its information needs. Efficiency deals with interactive responses, defined as getting results with short delays. These features will be verified with several datasets of biomedical images.

2.2 Previous work

The search for a solution that enables effective and efficient access to large data repositories has had a great interest among researchers in the last decade. These efforts have been embodied in several medical image retrieval systems [4], as are the ASSERT [5], CasImages, MedGIFT and IRMA [6] systems, among others. Currently, image retrieval systems can be classified into 4 groups: text-based image retrieval, visual content image retrieval, semantic content image retrieval and composite systems [7].

2.2.1 Text-based image retrieval systems

This is the basic retrieval system and is the most commonly used in hospitals to organize medical images, known as PACS (Picture Archiving and Communication System). In this kind of system, the images are indexed by text known as metadata, which are handwritten annotations made by experts. However, this approach is quite poor because with only textual information, is not possible to represent all the visual content present in each image [4].

2.2.2 Visual content-based image retrieval systems

A CBIR (Content-Based Image Retrieval) system, unlike the traditional search systems, do not requires the use of meta-data attached to images such as keywords, tags or other associated descriptors that allows identification, but uses the actual content of the image that is visual information, such as colors, shapes, textures or any other information that may result from the image itself. The main important property of such features is to be invariant regarding spatial transformations. i.e.., should not depend on properties of scaling, translation and rotation.

Thus in order to find an image you can set some visual values as query parameters. A common technique is to present query parameters through some original image with the purpose of recovering more closely match. This is known as a query

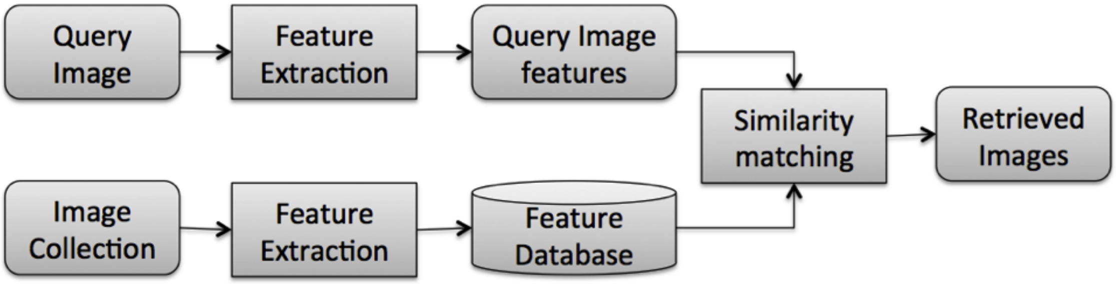


Figure 2.1: Content based image retrieval process.

for example. This in the medical case can be seen as an image diagnosis for which you want to find similar clinical cases.

The search by visual content requires computer vision methods for modeling the visual characteristics of the images and identify potentially relevant images. The two most important things to consider in the visual content-based systems are basically an appropriate descriptor for the specific type of images to be treated, and a strategy for ranking of the most relevant images based on some calculation of similarity. In Figure 2.1 it is shown the basic process in a content based image retrieval system.

The main drawback of finding images using visual features is that visual similarity does not necessarily produce valid results; it is because objects with similar colors or textures can have very different meanings. This problem is called "semantic gap" [8], and refers to the gap between visual features and high-level concepts (concepts that a human can interpret in the images). Since then, the problem of CBIR research has focused on modeling the semantic content of images.

Several studies have been reported in which are evaluated the applicability of content-based image retrieval in finding pathologies. [9] is a study that seeks to determine whether a patient has Alzheimer's disease through the analysis of visual content on magnetic resonance imaging. In histology and dermatology images, there have been studies focused on the detection of tumors [10]. In [11, 12, 13, 14] different strategies are used in the mammogram image analysis in order to identify possible calcifications and breast cancer. Some of the techniques used in these studies are presented below in more detail.

G. Shengwen et al. [15] presents a work focused in chest x-ray image databases, the images are classified into 8 categories: normal, pneumonia, phthisis, bronchitis, bronchiectasis, pneumothorax, lung cancer and pleural effusion. In order to propose a retrieval system for this kind of images, the paper address problems as salient points detection, visual feature extraction and detection of region of interest (ROI). In order to extract ROIs Shengwen et al. proposed an approach based on density distribution that takes into account the density differences in different parts of the chest such as clavicle and lung borders. For the image representations low-level features are used including: texture, Gabor feature vector using 4 scales and six orientations; and edge histograms, vectors composed by 5 categories, vertical, horizontal, 45° diagonal, 135° diagonal and isotropic. Before performing feature extraction, images are processed with a set of filters.

Rahman et al. [16] proposed an image representation based in local features. Authors present an image retrieval method based on SIFT (Scale-Invariant Feature-Transform), this technique seeks to represent the image taking into account only points of interest, and converts these into visual words, this technique is called Bag of key-points. The paper also presents other concepts such as SOM (Self-Organization Map) which is used for clustering of images using a fuzzy feature space.

S. Junding et al. [12] presents a study in mammography images where two main steps are proposed: first, an analysis based on regions of interest, and second, a more robust representation of texture features. Within the proposed methodology a preprocessing stage is provided basically aimed at removing useless information that only adds noise to the analysis:first, a low pass frequency filter that enhances the dominant texture structures is implemented, and secondly, restricts the analysis only for a so-called regions of interest, allowing lower computational cost and avoiding the negative influence of the complex environment. Then presents the descriptors used: Restrict Distortion, which describes the change of intensities between neighboring pixels and weighted moments. These two descriptors are used to build the index for the calculation of similarity.

K. Byrd et al. [11] shows a procedure used explicitly for mammography images, for which techniques are based on signal and image processing such as Fourier analysis, probabilistic analysis and digital filters. The proposed methodology called DMCBIR is based on the RGCwML which is an automatic boundary detection algorithm. Within the study four methods of similarity measure are used: the Hausdorff distance, Euclidean distance, L2 distance metric and WI (Williams Index) .

I. El-Naqa et al. [13] used low level features like local edges, grey level histograms. In this paper a similarity estimation based in two steps is used. Where in a first step, called classification, all images that don't have any kind of similarity are identified and immediately discard. This is performed for a basic classifier of low computational cost. In the second step, called regression, the unrejected images in the first step are processed with a more accurate classifier. For the first step two alternatives were used: Fisher Discriminant and Support Vector Machine (SVM), and in the second step General Regression Neural Network (GRNN) and SVMs were used.

S. Kinoshita et al. [14] proposed a preprocessing step in order to remove noise in the images. For the image representation a combination of several features are used: features of shape and granulometric are presented. For dimensionality reduction, the Principal Component Analysis technic (PCA) is performed.

P. Bugatti et al. [17] present an image retrieval technique that pretends to improve the performance using relevance feedback (RF), which takes into account the interactions with users. The system based in relevance feedback performs in three steps. First, the system retrieves the most similar images, second, the user judges the returned images as relevant and irrelevant, and third, the system adjusts the original query based in user feedback. The second and third steps are repeated until the user is satisfied with the system results. As distance measure two metrics are evaluated: Canberra and χ^2. The results obtained in the experimental evaluation show that the better distance measure is Canberra distance, that obtained the better performance and presents the lowest computational cost. The images are represented using two feature extractors: Haralick's texture an Zernike moments (the first 256 moments).

2.2.3 Semantic content-based image retrieval systems

In order to minimize the problem of semantic gap, the semantic-based image retrieval approach is proposed, i.e., using higher-level concepts rather than using crude representation of the visual data. There are several methods to achieve a semantic representation, which could be separated into two approaches: One approach in which the system adds a predetermined semantic structure and a second approach in which semantic concepts are learned in machine learning processes, in both approaches the result is a mapping between low-level features and the semantic space provided.

Under the approach that uses a predetermined semantic structure we can find jobs as [18, 19, 20], where the semantic representation is based on the construction of ontologies. And for the second approach, in [21, 22] the technique of Non-negative Matrix Factorization (NMF) was used in seeking to achieve a semantic representation by the latent factors emerging from the process of factorization, and in [23, 24, 25] the PLSA (Probabilistic Latent Semantics Analysis) model was used.

Q. Chen et al. [26] show a strategy for improving performance in the retrieval task using latent semantic indexing (LSI) in medical images, the dataset of this study corresponds to a set of 1345 images composed by gastroscope images with cancer and healthy tissues. Initially, the images are represented by low-level features. The features used are Color Histogram and Color Correlogram, the Color Correlogram represents the change in color correlation according to the change in distance. The main reason for using color features is that gastric cancer cells are clearly different in color with the rest of gastric tissues, while red is the predominating color of gastroscopic images, the gastric cancer cells are colored in yellow and red. Once the low-level representation is performed, this data is used to build a semantic latent indexing. The used method is SVD (Singular Value Decomposition). The main idea is to decompose an initial data matrix that is composed by a set of features vectors that represent each image. With the LSI technique it is possible to represent the original matrix with other matrix of lower rank. This technique is expected to achieve two goals: eliminate data redundancy and mitigate the noise.

2.2.4 Multimodal image retrieval systems

Usually, collections of image data can be accompanied by textual data, and in general by other kinds of data. These data could be combined in order to improve the retrieval process, since they may contain complementary information that allows better image representation.

To combine multiple information modalities, we have two alternatives called early fusion and late fusion [27]. In late fusion, the information for each modality is processed separately, carrying out indexing independently, and the information fusion takes place at the query time. In the case of early fusion, information combination is performed in the step of defining characteristics, modeling relationships between these, creating a new type of multimodal representation. This way, in the query stage it is enough to evaluate this new multimodal feature vector.

In [21, 22, 28, 24, 25] can be observed as well as using the algorithms of NMF and PLSA to find a semantic representation in an extended way to include textual and visual information together, showing a significant gain in data retrieval performance. A competition held annually by the ImageCLEF [29], which is focused on making an experimental evaluation of medical image retrieval, demonstrates the benefits of combining visual and textual modalities. Some of the most relevant works based in this approach are described below:

A. Alpkocak et al. [30] describe the system used by the DEMIR (Dokuz Eylul University Multimedia Information Retrieval) group for the task of retrieval of images for ImageCLEF 2011. The system is based on late fusion techniques for the

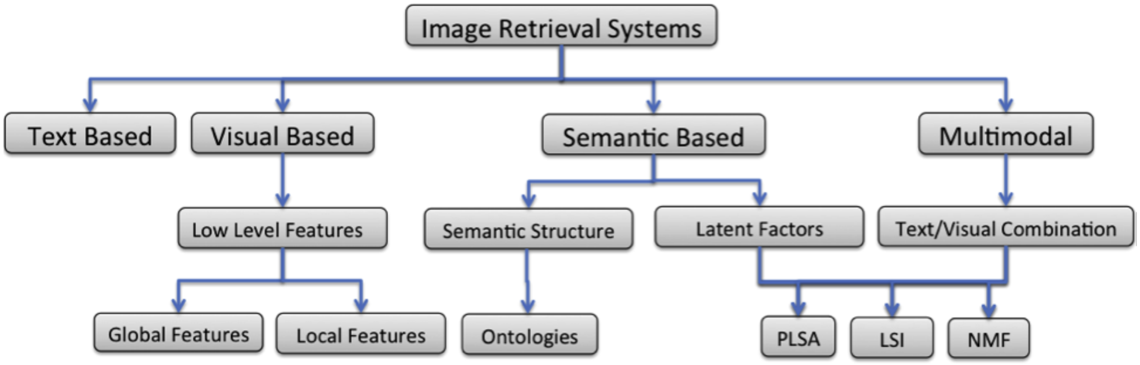


Figure 2.2: Image Retrieval Systems Classification.

combination of the multimodal information. The text processing is based on the Terrier system and the visual representation use low-level features. Visual characteristics evaluated are: EHD (Edge Histogram Descriptor), CEDD: EHD descriptor with color histogram, FCTH: diffuse version of CEDD, BTDH: similar to FCTH but it uses brightness values instead of color histogram. Among the different tests, it was found that the best visual descriptors were DSBB and CEDD. As a measure of distance, the Euclidean distance is used. For the fusion of the multimodal information, two late fusion techniques were introduced: Average and Weighted Average, which are basically linear combinations of score values obtained for each mode separately.

G. Csurka et al. [31] also used a late fusion approach. First, an initial filtering is performed using only textual information to retrieve the 1000 most relevant documents, then, these documents are ranked again by a linear combination of the values of visual and textual scores. For visual representation the Fisher Vector is used and for the text representation, 4 models were evaluated: 1) DIR: Smoothed Dirichlet Language Model; 2) LGD: Log-logistic Model Based Information; 3) SPL: Smoothed Power Law Model-Based Information; 4) AX: Lexical Entailment Model based IR. In the evaluation, the AX model was the one who showed better results. The experiments showed that the best results are obtained by giving more weight to textual information. A high value in a visual mode improves accuracy in the top results but reduces considerably the value of MAP.

2.3 Histology images

This work focuses in histological and histopathological images, and this is motivated by the fact that histology images are particularly challenging from an image understanding point of view. In this kind of images, visual patterns are generally a complex combination of fundamental visual features involving texture, color and shape [32, 33].

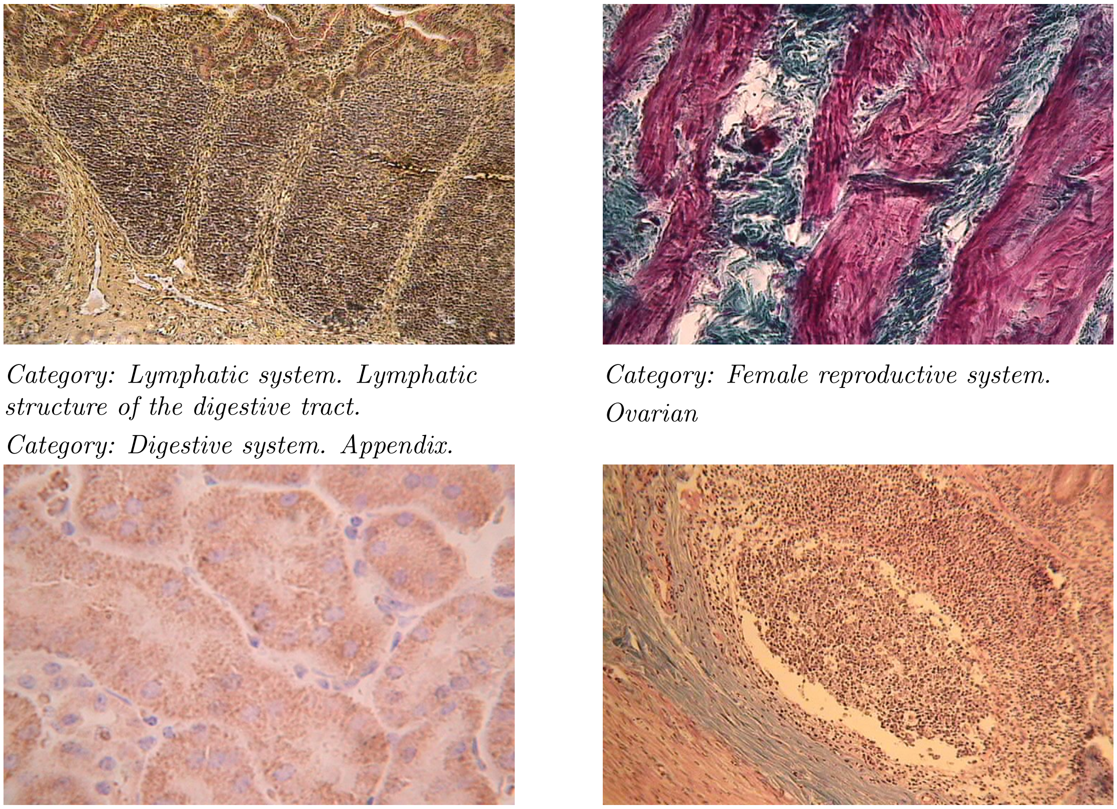
The major challenges encountered in this kind of images are as follows:

- The visual appearance of the biological tissues and structures changes according to the type of cutting of a biological sample. Although protocols exist for the acquisition process of the histology sheets in order to standardize cutting types, the accuracy of the cuts is defined by different technical elements.
- Visual variability for staining. One feature of the histology images is that they can using different types of stains in order to highlight specific biological structures. Thus, for the same given sample different types of stains can be used.
- Different magnifications. This presents a high variability of visual appearance for the same tissue because it contains images annotated with the corresponding concept in different magnifications.

Histology and histopathology images are important for medicine. These are a key asset to determine the normality of a particular biological structure, or to diagnose diseases such as cancer. In this work, two histology image collections were used for experimental evaluation. This collections are described in the subsequent subsections.

2.3.1 Histology atlas dataset

This dataset is composed of 2,641 images extracted from an atlas of histology for the study of the four fundamental tissues [34]. The collection includes photographs of histology slides acquired with a digital camera coupled to a microscope, using different magnification factors to focus important biological structures. Each of these images was annotated by an expert, indicating the biological system and organs that can be observed. The total number of different keywords that can be found in this data set is 46, which was obtained after a standardization of the vocabulary used to describe the semantic contents. The list of terms includes circulatory system, heart, lymphatic system and thymus, among others. Usually, images have just one term attached to it, but in several cases images can have various.



Category: Lymphatic system. Lymphatic structure of the digestive tract.

Category: Digestive system. Appendix.

Category: Female reproductive system. Ovarian

Category: Urinary system. Kidney.

Category: Lymphatic system. Lymphatic structure of the digestive tract.

Category: Digestive system. Ileus.

Figure 2.3: Sample images from histology atlas dataset. Images with associated terms.

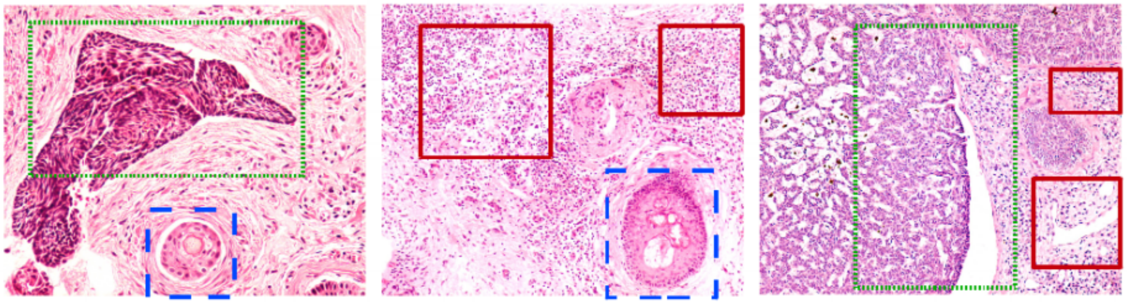


Figure 2.4: Some example images from the histopathology image collection of biological structures and pathological patterns highlighted. The dashed lines (blue) show pilosebaceous units, normal biological structure in the skin. The dotted lines (green) show regions with nodules, palisade cells and crevices (NEH). Solid lines (red) show regions with infiltration of lymphocytes, another indication of basal cell carcinoma.

2.3.2 Basal-cell carcinoma dataset

It is a histopathology image collection that has been used to diagnose a special kind of skin cancer known as basal-cell carcinoma [35]. This type of cancer is more common in fair-skinned populations and its incidence is increasing worldwide [36, 37]. It has different risk factors and its development is mainly due to ultraviolet radiation exposure.

The collection is composed of 1,502 images that were studied and annotated by a pathologist to describe its contents, elaborating a list with 18 terms. The list of keywords includes micronodules, elastosis, and fibrosis, among others. In this data set, one image may contain several keywords attached, that is, different biological structures are exhibited in one single image.

2.4 Conclusions

As we can see the amount of work related to this investigation field is extensive, and there is a wide range of proposed approaches that try to address this problem from different points of view. This is evidence of a field that still lacks a long way to go. In fact, looking more closely at each of the works, we can see that the results are still far from expected. This shows a very promising field for future research, especially in the field of medical imaging in which the benefits of finding a truly robust technique would be invaluable.

Also, according to the reviewed works, using a latent semantic representation and exploiting the multimodal information, it is possible to achieve a better understanding of the visual content of images, allowing to address the semantic gap problem, and thereby improving performance in information retrieval. It is important to highlight that most of the work in CBIR has been devoted to natural-scene images, with some exceptions discussed in this chapter. In the particular case of multimodal CBIR applied to biomedical images, the first works published in the area are ours and are integral part of this thesis.

Chapter 3

Basic Notions and Definitions

This chapter presents detailed information about the kind of images to be processed, analyzed and indexed in this work, also, the chapter introduces the required notation and fundamental concepts that form the basis of this work.

3.1 Image representation

An essential task in any content-based image retrieval (CBIR) is the representation of images, which has as objective to provide information of interest than just the values of the pixels located in a matrix. An image can be represented by visual content descriptor that can provide information like color, texture, shape, spatial relationship among others.

A visual content descriptor can be either global or local. A global descriptor uses the visual features of the whole image, whereas a local descriptor uses the visual features of regions or objects to describe the image content. In this work the bag-of-features representation is used, a local descriptor which has been found to be an effective representation for microscopy image analysis [38, 39].

3.1.1 Bag of visual words

The bag of visual words representation is an adaptation of the bag of words scheme used in text categorization and retrieval [40]. The main idea is to construct a codebook or visual vocabulary, in which the most representative patterns are encoded as visual words. In this way the representation of the image is generated through a simple frequency analysis of each codeword within the image. This representation has been successfully applied in various tasks of classification and retrieval [38, 41, 42]. There are three main steps in constructing a representation of bag of visual words

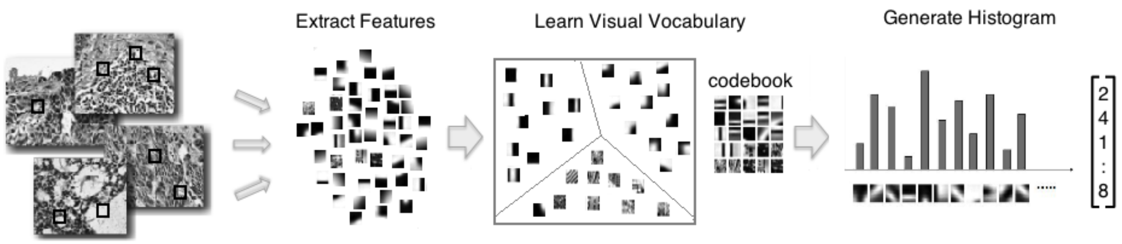


Figure 3.1: Overview of bag of visual words representation illustrating the three major steps: 1) local features are extracted from a set of training images, 2) the visual vocabulary is learned and 3) for each image occurrences of visual words in the dictionary were recorded in a histogram.

[43]: (*i*) visual words detection, (*ii*) visual dictionary generation and, finally, (*iii*) visual words quantization to construct the histogram. Figure 5.1 shows an overview of the steps.

3.1.1.1 First stage. Visual words detection

In general the bag of visual words algorithm begins with the extraction of small blocks. For the extraction of these blocks we can follow several strategies such as: extract blocks randomly placed in the image, extract blocks using a regular grid or extracting blocks belonging to certain regions of interest. In this work the extraction approach based on a regular grid is used, that generates a higher number of blocks and therefore implies a greater computational load, but also reduces the probability of losing interesting visual patterns. Finally, each of these blocks or visual words is represented by a descriptor. In this work we use the DCT descriptor for block representation, which has exhibited a good performance in annotation task for histology images [42].

DCT Descriptor. This type of representation takes into account information from color and texture in an efficient way. For each block of $n \times n$, the DCT (Discrete Cosine Transform) is applied to each of the three color components (Red, Green and Blue) . Finally, the descriptor is built by joining the n^2 coefficients of each of the three channels, obtaining a descriptor of $3n^2$ dimensions.

3.1.1.2 Second stage. Learning the visual vocabulary

The purpose of this second step is to define our visual vocabulary or dictionary, as a set of K visual words. This seeks to bring together the wide range of visual characteristics obtained in the initial stage, and reduce it to a representative set. For this purpose, the technique typically employs K-means clustering, which gives as a result a set of centroids that define the visual dictionary. This is an unsupervised learning technique that requires high computational power and high memory consumption, due to the high dimensionality of the data and the large number of iterations required until the algorithm converges to an optimal result.

3.1.1.3 Third stage. Visual words quantization to construct the histogram

Once we have the visual vocabulary, the last step is to translate all visual characteristics obtained in the first step to the set of visual words. This is done by calculating the distance of each feature vector to each centroid, and assign it to the nearest. Once the translation is completed, it takes a count of how many times a word is found in every visual image, generating a histogram of visual words, encoded as a vector of dimensions equal to the size of the vocabulary, which is the final representation for each image. Finally a collection of images can be represented as a matrix (imaging visual characteristics).

3.2 Text representation

The most widely used text representation for text retrieval and classification is the *bag of words* model. In this model, a text is represented as an unordered collection of words, disregarding grammar and even word order. Thus, the occurrence of each word is used as a feature. Finally, a document is represented as a vector that contains term frequencies. In both information retrieval and text classification, it is common to weigh terms by various schemes, the most of popular of which is tf–idf.

3.2.1 TF-IDF

The tf-idf weight(term frequency-inverse document frequency) is a statistical measure used to evaluate how important a word is to a document in a collection. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the collection. Tf–idf is the product of two statistics, term frequency and inverse document frequency. Various ways for determining the exact values of both statistics exist. In the case of the term frequency $tf(t,d)$, the simplest choice is to use the raw frequency of a term in a document. Other possibilities are boolean frequencies, i.e.,$tf(t,d) = 1$ if t occurs in d and 0 otherwise; and normalized frequency, to prevent a bias towards longer documents.

The inverse document frequency (idf) is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term.

$$idf(t, D) = log\left(\frac{|D|}{\{d \in D : t \in d\}}\right) \tag{3.1}$$

where, $|D|$ is the total number of documents in the collection and$\{d \in D : t \in d\}$ number of documents in the collection where the term t appears. Finally tf-idf is calculated as:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \tag{3.2}$$

3.2.2 Latent semantic techniques

A successful approach in information retrieval is the latent semantic indexing, which implies a transformation in the representation of the collection to a lower rank approximation, allowing the extraction of the underlying semantic structure. This approach alleviates noise in “term” usage and implicitly solves the problems of polysemy and synonymy. There are several methods to achieve a semantic representation, among the most relevant are the classical LSA method, which is based in the mathematical technique called singular value decomposition (SVD), the pLSA method and NMF method which follow a probabilistic model.

3.2.2.1 PLSA

Probabilistic latent semantic analysis (PLSA), also known as probabilistic latent semantic indexing (PLSI, especially in information retrieval circles) is a statistical technique for the analysis of two-mode and co-occurrence data. In effect, one can derive a low dimensional representation of the observed variables in terms of their affinity to certain hidden variables, just as in latent semantic analysis. PLSA evolved from latent semantic analysis, adding a sounder probabilistic model. Compared to standard latent semantic analysis which stems from linear algebra and downsizes the occurrence tables (usually via a singular value decomposition), probabilistic latent semantic analysis is based on a mixture decomposition derived from a latent class model. This results in a more principled approach which has a solid foundation in statistics.

3.2.2.2 Nonnegative matrix factorization

Matrix decompositions are useful to extract structural information from a collection of data samples. For an input matrix $X \in \mathbb{R}^{n \times l}$, containing l data samples with n features in its column vectors, Nonnegative Matrix Factorization (NMF) finds a low rank approximation of the data using non-negativity constraints:

$$X \approx WH$$
$$W, H \geq 0$$

where $W \in \mathbb{R}^{n \times r}$ is the basis of the vector space in which the data will be represented and $H \in \mathbb{R}^{r \times l}$ is the new data representation using r factors.

NMF finds the matrices W and H by solving the associated optimization problem that corresponds to minimizing the reconstruction error of the original data. In this work, the Lee and Seung's approach [44] is adopted to obtain the factorization, using the divergence criterion as objective function:

$$D(X|WH) = \sum_{ij} \left(X_{ij} log \frac{X_{ij}}{(WH)_{ij}} - X_{ij} + (WH)_{ij} \right) \tag{3.3}$$

which is zero when $X = WH$. This function may be regarded as the Kullback-Leibler Divergence between the two matrices as long as both are normalized in such a way that the sum of their values is equal to one. Then, the matrices may be considered to be probability distributions. Following this approach, an iterative algorithm which alternates the optimization of W and H uses the following multiplicative updating rules:

$$\begin{aligned} W_{ia} &= W_{ia} \frac{\sum_{\mu} H_{a\mu} X_{i\mu} / (WH)_{i\mu}}{\sum_{v} H_{av}} \\ H_{a\mu} &= H_{a\mu} \frac{\sum_{i} W_{ia} X_{i\mu} / (WH)_{i\mu}}{\sum_{k} W_{ka}} \end{aligned} \tag{3.4}$$

These rules are guaranteed to decrease the objective function and find at least a locally optimal solution to the factorization problem [45]. The NMF algorithm has been proposed for multimodal image indexing by taking the matrices of visual data and semantic annotations as input, following two strategies [46]: 1) NMF-mixed, which concatenates the two inputs in a unique matrix and 2) NMF-asymmetric, which decomposes the semantic data first and adapts the visual data afterwards. The main goal of either algorithm is to build a common latent factors representation for both data modalities and then employ it as an effective multimodal index.

The latent factors representation is achieved by setting the rank$r$ of the decomposition to some appropriate size, which simply amounts for the number of latent factors. Therefore, the matrix H determines the latent encoding for every image, and the matrix W provides the transformation from the space of original features to the latent factor representation.

One of the reasons NMF has had success in modeling data representations is because its ability to find parts of objects. When compared to standard latent semantic indexing or singular value decomposition (SVD), which had orthonormal restrictions and no constraints in sign, NMF gives more interpretable basis vectors [44] and finds better structural patterns in different collections of data [47]. This usually results in an improved performance in the underlying computational task.

3.3 Content based image retrieval evaluation

In order to compare the results of different works in the area, a group of tools that allows evaluation of different content image retrieval systems was proposed by Müller et al. [48]. This is a group of performance measures that have been adapted from the discipline of textual information retrieval.

3.3.1 Performance measures

Mainly there are two statistical measures: precision and recall. When it has completed a consultation process the entire set of images belonging to the system is divided into two groups, which are the group of recovered images and the image group omitted. In turn, the images of each group are divided into a group of images that are relevant as the search criteria above and that do not meet the criteria. Therefore:

For a given query topic, the precision is defined as the ratio between the number of retrieved images that are really relevant to the total retrieved images in a query.

$$Precision = \frac{|\{relevant\,documents\} \cap \{retrieved\,documents\}|}{|\{retrieved\,documents\}|} \tag{3.5}$$

And recall, is the ratio of relevant images that were correctly recovered, with respect to the total relevant images that should be recovered.

$$Recall = \frac{|\{relevant\,documents\} \cap \{retrieved\,documents\}|}{|\{relevant\,documents\}|} \tag{3.6}$$

Therefore, a good information retrieval system seeks to maximize both values. Precision and recall values can be plotted to give a precision-recall curve, as shown

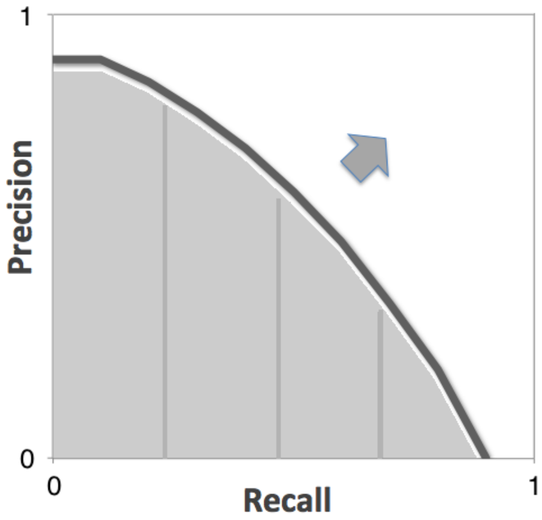


Figure 3.2: Precision-recall Curve

in the figure 3.2. In which an ideal result is that the precision value starts with a value of 1 and it remains even when the values of recall increases.

As we can see in figure 3.2 we can graph all values of precision in function of recall. Average precision (AveP) computes the average value of precision over the interval from $recall = 0$ to $recall = 1$ Mean average precision (MAP) for a set of queries is the mean of the average precision scores for each query.

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{3.7}$$

where Q is the number of queries.

The MAP value is one of the most important measures to consider, because it summarizes the overall behavior of a retrieval system, and this is the principal evaluation parameter in the most important benchmarking [29].

Chapter 4

Image Indexing using a Non-negative Semantic Embedding

This chapter presents a new method for indexing histology images using multimodal information, taking advantage of two kinds of data: visual data extracted directly from images and available text data from annotations performed by experts. The new strategy called Non-negative Semantic Embedding that extends the NMF algorithm, defines a mapping between visual and text data assuming that the space spanned by the text is a good enough representation of the images semantics.

The results presented in this chapter were published in *Workshop on Medical Content-Based Retrieval for Clinical Decision Support. MICCAI 2011*[49].

4.1 Non-negative semantic embedding

The methods based in NMF are oriented to model latent factors for multimodal data, that is, to find the hidden structure of the collection, which is assumed to be common between the two data modalities. We propose a simplified strategy that extends the NMF-asymmetric algorithm [46] to a setting in which the semantic encoding is already known.

NSE method is used when we assume that the semantic encoding is already known, i.e., in this model we assume that the space spanned by text terms is a good enough representation of image semantics, and we use it to index and represent all images in the collection. Then, we want to find a way to embed visual features in this semantic space, to index images with or without annotations. We formulate this problem as finding a linear transformation of the visual data imposing a non negativity constraint on the solution, as follows:

$$V = ST; S \geq 0 \tag{4.1}$$

Where, $S \in \mathbb{R}^{m \times n}$ is the transformation matrix representing the relationships between the visual and text modalities. n is the number of visual features and m is the number of text terms. The non-negativity constraint in this case enforces an additive reconstruction of visual features, since vectors in the matrix S can be thought of as parts of images that are combined according to the presence of associated labels. Notice that the encoding matrix, T, is the matrix of text annotations, and the vectors in S can also be directly interpreted as the visual features related to each text term.

Finding the matrix S when V and T are known is a convex problem under the divergence or euclidean criteria in the related minimization problem. However, we approximate the solution to this problem using the NMF updating rule for the matrix S. Instead of requiring a global optimum for this problem, which might result in overfitting to the training data, we accept a good approximate solution obtained from the Lee and Seung's approach [45]. The updating rule usually converges to a local minimum, but it may result in a better generalization with some robustness to intrinsic noise in the training data. Updating rule for S:

$$S_{ia} \leftarrow S_{ia} \frac{\sum_{\mu} T_{a\mu} V_{i\mu} / (ST)_{i\mu}}{\sum T_{av}} \tag{4.2}$$

We call this approach the Non-negative Semantic Embedding (NSE) taking into account that the semantic space is known in the problem and the resulting solution to embed image features on it is non-negative. It also differs from the Gonzalez et al. [22] approach in the sense that no latent factors are herein modeled, but instead, a semantic space is assumed from the given text terms space.

4.2 Image indexing and search

The indexing methods described above require a training phase to learn a mapping of images to the semantic or latent space. The result of that training phase is a matrix S, which contains the basis of the indexing space and serves as linear transformation to project new data. So, when new images are obtained without text annotations, either to be included in the collection or as queries, we can find the semantic representation for this image using visual features only.

Let v be the visual features of a new image, unseen during training. To embed this image in the semantic space, the following equation needs to be solved:

$$v = St \tag{4.3}$$

where S is the basis of the semantic space and t is the semantic representation of the new image. The image v will be embedded into the semantic space by finding the vector $t \geq 0$ that satisfies the equation. This is done by using the multiplicative updating rule for t while keeping the matrix S fixed.

This strategy applies for both, the NMF-based algorithms using latent factors and the NSE. So, now that we can represent images in the collection and query images in same space, the problem of finding relevant results reduces to the problem of matching images with similar representations. To do so, we employ the dot product as similarity measure, which gives a notion of the extent to which two images share similar components in the latent or semantic space. Finally, results are ranked in decreasing order of similarity and delivered to the user.

4.3 Experimental evaluation

The evaluation of NSE method was conducted using the Histology Atlas dataset. All the results were reported in [49].

We conducted retrieval experiments under the query-by-example paradigm to evaluate the proposed methods. A set of 100 images were randomly selected as queries from the database of 2,641 images used in this study. The remaining 2,541 images were used as the target collection to find relevant images.

We performed automatic experiments by sending a query to the system and evaluating the relevance of the results. A ranked image in the results list is considered relevant if it shares at least one keyword with the query. For this experiment, the evaluation was done using traditional measures of Image Retrieval, including Mean Average Precision (MAP) and the Recall-Precision plots.

4.3.1 Image Features

We build a bag-of-features representation for the set of histological images, as it has been found to be an effective representation for microscopy image analysis [38, 39]. We start by extracting patches of 8 × 8 pixels from a set of training images with an overlap of 4 pixels along the x and y axes. The DCT (Discrete Cosine Transform) transform is applied in each of the 3 RGB channels to extract the largest

21 coefficients and their associated functions. A k-means clustering is applied to build a dictionary of 500 visual terms. This bag-of-features configuration have shown good results with similar types of histology images [38, 39].

Once the vocabulary has been built, every image in the collection goes through the patch extraction process. Each patch from an image is linked to one visual term in the dictionary using a nearest neighbor criterion. Finally, the histogram of frequencies is constructed for each image. We experimentally found that 500 visual terms was enough to achieve a good performance, and that larger dictionaries do not provide significant improvements, but just more computational load.

4.3.2 Text annotations

In this data set of histology images, text annotations are clean and clearly defined terms from a technical vocabulary. Since the annotation process followed a systematic revision, there is no need to build a vector space model or to account for term frequencies. We build semantic vectors following a boolean approach, assigning 1 to the terms attached to an image and 0 otherwise. This leads to 46-dimensional binary vectors, which serve to build the text matrix.

4.3.3 Visual search

As a first experiment we retrieved histology images using only visual information as a baseline to assess improvements of other methods. The visual descriptors are based on the bag-of-features strategy, so images are represented by histograms of the occurrence of visual patterns in a dictionary. Direct visual matching is done by calculating the level of similarity between images using the histogram intersection similarity measure [50], as follows:

$$K_{HI}(x,y) = \sum_{i=1}^{n} min\left\{x_i, y_i\right\} \tag{4.4}$$

where x and y are images and x_i, y_i are the $i-th$ occurrence of the visual feature in these images, respectively. Using direct matching only, the system achieves a performance up to 0.210 in terms of MAP. An additional experiment using visual features was conducted to determine if latent factors learned only from visual data can help to improve over this baseline.

We applied an NMF decomposition on the visual matrix X_v, using different numbers of latent factors. In the best performing case, this strategy reaches a value of 0.172 of MAP. This suggest that the dimensionality reduction made by NMF leads to a loss of discriminative power of visual descriptors instead of helping to identify semantic patterns in the collection.

4.3.4 Multimodal search

Multimodal search aims to introduce text information during the search process, even though in our case queries are expressed using no keywords but example images. To this end, we employ three different algorithms based on NMF: NMF-mixed [46], NMF-asymmetric [46] and NSE, and compare their performance against visual search. For the first algorithm (NMF-mixed) the construction of a multimodal matrix was done by setting $\alpha = 0.5$ to give the same importance to visual and text data.

For the NMF-mixed and NMF-asymmetric algorithms we performed several experiments using different sizes of the latent semantic space, to experimentally determine an appropriate number of latent factors. In contrast, the NSE algorithm does not need to set this parameter. Figure 4.1 shows the result of exploring the number of latent factors for these algorithms, including reference lines for the visual search and NSE. It can be seen that the response of NSE outperforms by a large margin the response of all other strategies. In addition, the Figure 4.1 shows that besides NSE, only the NMF-asymmetric strategy is able to improve over the visual baseline.

The loss in performance of using NMF-mixed and NMF-visual can be explained by the difficulty of these strategies to find meaningful latent factors from the given input data. As was mentioned before, the NMF-visual fails to find semantic patterns in the collection leading to a decreasing of the discriminatory power of the full visual representation. The NMF-mixed shows a better achievement in this setting, which improves over NMF-visual due to the presence of text terms in the multimodal matrix. Still, the semantic patterns are not correctly modeled by multimodal factors because they have to deal with the reconstruction of visual features as well.

The two strategies that improve over the direct visual matching are NSE and NMF-asymmetric, which concentrate on exploiting text information as the semantic reference data. The NMF-asymmetric, in particular, decomposes the text matrix in an attempt to find meaningful relationships between text terms to build semantic latent factors. In our experiments, we decompose the matrix of 46 terms in 10, 20, 30 and 40 latent factors, from which the second choice presented the best performance. However, the improvement of NMF-asymmetric is modest with respect to NSE, indicating that latent factors are not a fundamental modeling aspect for this dataset.

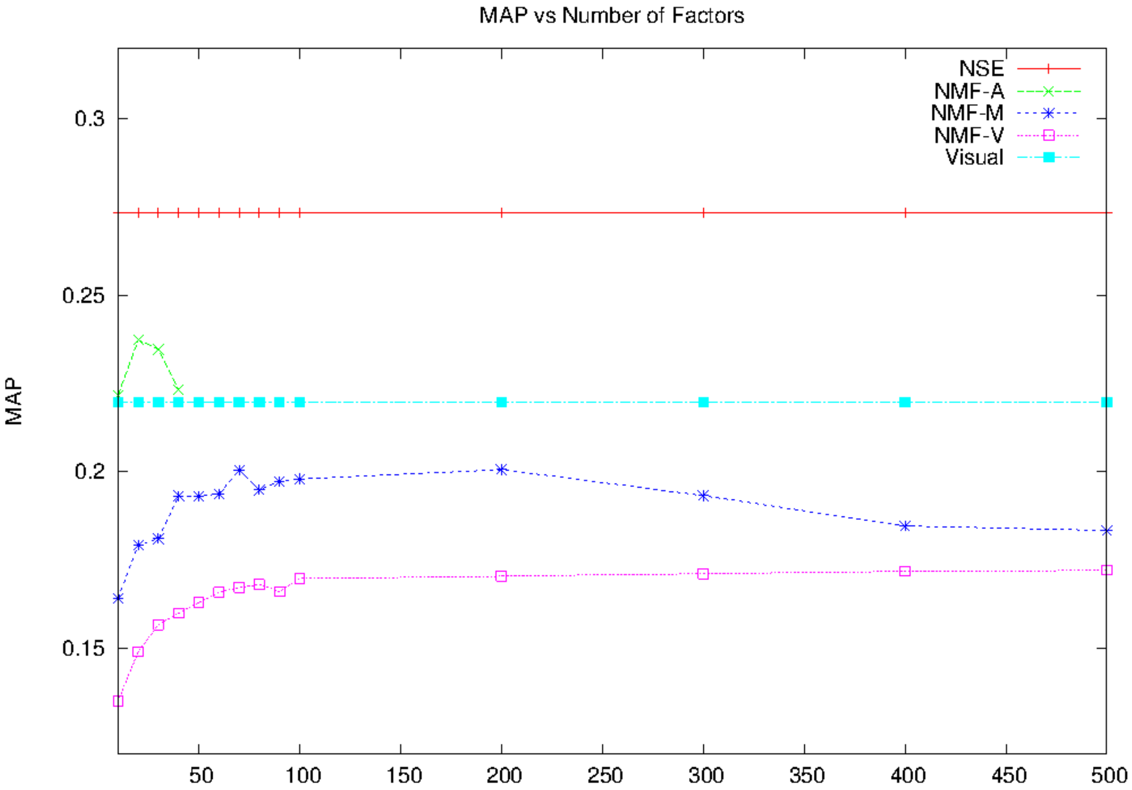


Figure 4.1: Comparison between NSE and other algorithms based on NMF. Latent factors vs. MAP

Another consideration of the modest improvement of NMF-asymmetric may be found in the two steps algorithm. When the first decomposition is done, an approximation error is generated since perfect reconstruction is not required. Then, the second decomposition builds on top of it to construct latent factors for visual features, introducing its own approximation error as well. The NSE algorithm simplifies the approach by learning a unique matrix that correlates visual and text data directly.

Another way to measure the performance of the evaluated algorithms is using the Recall vs. Precision graph. For this evaluation, we selected the number of latent factors that provided the best performance in the previous evaluation. We plot the interpolated Recall-Precision graph, in which one can observe the differences in precision along the retrieval process, i.e., while all relevant images are being retrieved to the user. Figure 4.2 shows the result of this evaluation, and reveals that the direct visual matching presents very good results in early precision, but also falls very fast as long as more relevant images are required.

The second best performance in the first positions of the results page is given by the NMF-visual approach, according to the Recall-Precision graph. What it actually

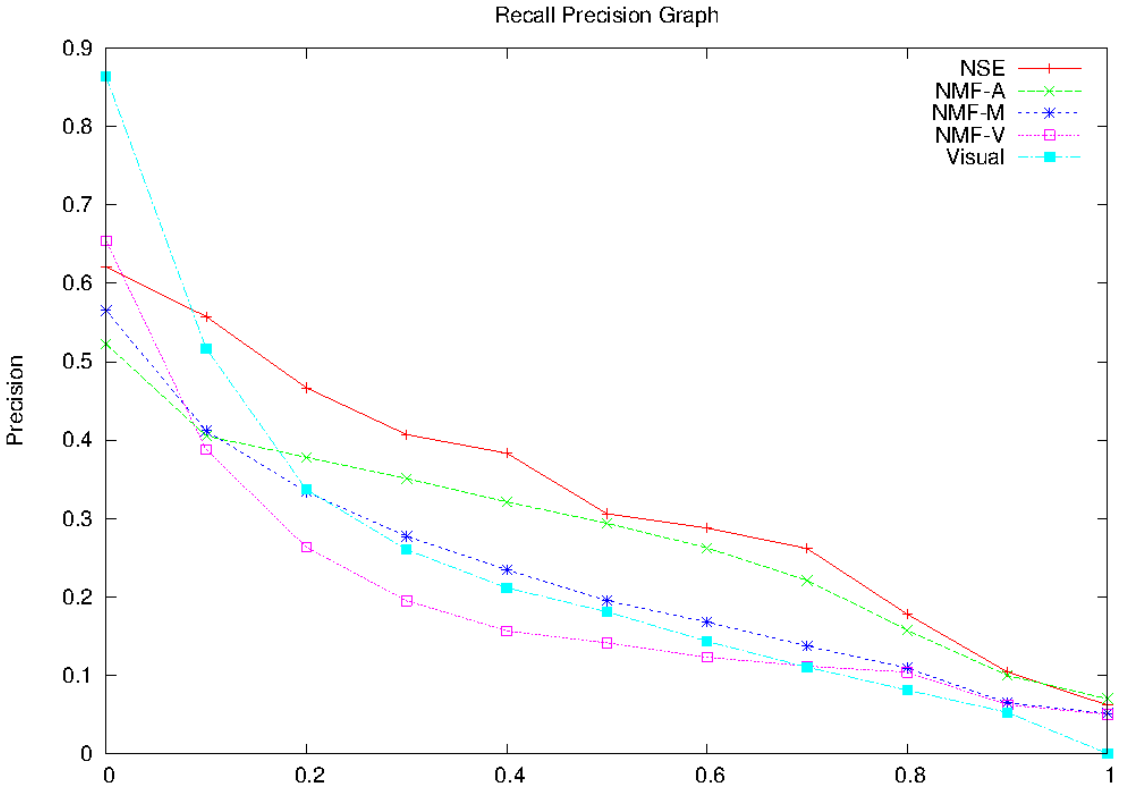


Figure 4.2: Comparison between NSE and other algorithms based on NMF. Recall vs Precision.

means, is that nearest neighbors under a visual similarity measure are very likely to be relevant in a histology image collection. This contrast with all other semantic indexing approaches (NMF-M, NMF-A and NSE) which are modeling structural patterns in the whole collection, rather than exploiting the local similarities of the dataset. Nevertheless, the performance of NSE and NMF-asymmetric showed a more consistent and sustained higher precision as long as the user explores more images in the results page.

The good performance obtained in early position with the direct visual matching method are due to the fact that in this collection several histology images come from sections that belong to the same histological plate or represent the same region with different zoom levels. Therefore, in this case, visual relationships directly denote semantic relations. But this behavior only occurs in the first retrieved images. In the Recall-Precision graph shows that values of recall greater than 10% (that in this collection represents an average of 20 images) the accuracy of the visual methods falls dramatically, while the NSE shows a more stable behavior.

Table 4.1 summarizes the findings of our experimental results, presenting the number of required latent factors, MAP measures, relative improvement in terms of

Table 4.1: Performance measures for NSE and other algorithms based on NMF

Method	Latent Factors	MAP	Improvement	P@10
Visual Matching	N/A	0.210	N/A	0.650
NMF-Visual	300	0.172	-18.8%	0.088
NMF-Mixed	70	0.201	-4.3%	0.252
NMF-Asymmetric	20	0.235	+11.9%	0.208
NSE	N/A	0.273	+30.0%	0.265

MAP w.r.t. the visual baseline and the precision at the first 10 results (P@10). The Table 4.1 shows that NMF-asymmetric and NSE provide a significant improvement in terms of MAP compared with only visual retrieval. It also confirms the dominant position of visual search in terms of early precision.

4.4 Conclusions

In this chapter we presented a method for image indexing that combines visual and text information using a variation of non-negative matrix factorization. In this method text annotations provide a semantic representation space where the visual content of images is embedded using NMF.

The experimental evaluation demonstrates an increasing in retrieval performance. The effectiveness of this method may be explained by the fact that it efficiently exploits the semantic information contained in clean text annotations. This means that this methodology is mainly applicable in cases where we have a controlled annotation process.

Chapter 5

Fusing Visual and Semantic Contents

Two strategies for image representation have been presented in the previous chapter. The first strategy is entirely based on visual features, to match for visually similar images. The second strategy is based on semantic data and the estimations of potential keywords for images without annotations. In this chapter a third strategy is introduced, based on multimodal fusion. The main goal of this scheme is to combine visual features and semantic data together in the same image representation to exploit the best properties from each data modality.

The results presented in this chapter were published in *SIPAIM 2011* [51].

5.1 Fusion by back-projection

The proposed fusion strategy is based on projecting semantic data to the visual feature space and then making a convex combination of both, visual and semantic representations. It can be understood as an early fusion strategy, since the representations are merged before its subsequent use.

The proposed approach follows the steps illustrated in Figure 5.1. So, assuming a histogram of visual features x_v and a vector of semantic data x_s, the fusion procedure generates a new image representation defined as:

$$x_f := \lambda S x_s + (1 - \lambda) x_v \tag{5.1}$$

where $x_f \in \mathbb{R}^n$ is the vector of fused features in the visual space and λ is the parameter of the convex combination that controls the relative importance of data modalities. This fusion approach takes the semantic representation of images and projects it back to the visual space using the reconstruction formula:

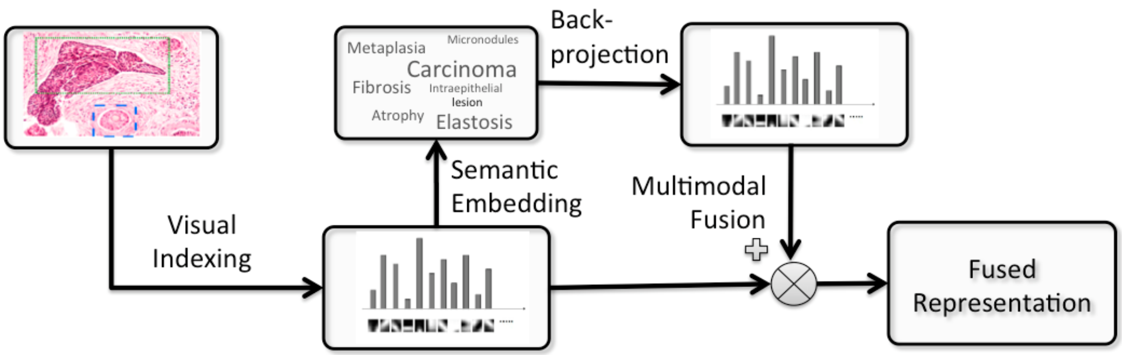


Figure 5.1: Overview of the proposed fused representation. From the input image to the final fused representation, three main processes are carried out: feature extraction, semantic embedding and multimodal fusion by back-projection.

$$\hat{x}_v := Sx_s \tag{5.2}$$

This back-projection is a linear combination of the column vectors in S using the semantic annotations as weights. In that way, the reconstructed vector $\hat{x}_v$ represents the set of visual features that an image should have according to the learned multimodal relationships in the image collection. Therefore, $\hat{x}_v$ and x_v highlight different visual structures of the same image, since $\hat{x}_v$ is a semantic approximation of the observed visual features.

5.2 Controlling modality importance

The parameter λ in the convex combination of the fusion strategy allows to control the importance of each data modality. The problem of assigning more weight to one or the other modality mainly depends on the performance that each modality offers to solve queries. More specifically, it depends on how faithfully one modality represents the true contents of an image. On the one hand, visual features may be inaccurate to represent high level semantic concepts, but good at representing low level visual arrangements. On the other hand, the semantic representation may be noisy or incomplete because of human errors or prediction discrepancies.

Now, the parameter λ is split in two different parameters to consider two kind of images: database images and query images. For both images, the semantic keywords are predicted by the learned NSE model. So, the prediction may be more accurate for images in the database since the model was learned using that data. For these images, the parameter λ will be called α throughout this paper.

On the other hand, query images will require a different parameter tuning since there is some uncertainty in the quality of their semantic predictions, and so, the original visual features may be more faithful to the true content. For these images, the parameter λ will be called β. This distinction is made to highlight that modality importance depends on how much we trust on the available modalities, and in the experiments run in this paper, it has been associated to images in the database and query images.

5.3 Searching in the fused space

The resulting fused representation lies in the visual feature space. So, in order to exploit the structure of the resulting representation, which inherits the structure of visual features, the histogram intersection defined in the previous chapter can be used as ranking function for search.

5.4 Experimental evaluation

The evaluation of the proposed fused strategy was conducted using the Renata and Carcinoma datasets. Following the same experimental setup defined in the previous chapter.

In both datasets a test set of images were randomly selected as queries, and the remaining images were used for training. In the histology atlas dataset 100 images were used as queries and in the basal-cell carcinoma dataset a 30% of all images were used as queries, i.e., 301 query images. In the experiments the semantic data associated to queries is not used during the search phase, but only for evaluation purposes. Our goal is to simulate visual queries using example images without associated metadata.

Images in both datasets are represented using the bag-of-features approach with a dictionary of 500 visual terms. For semantic data, a binary vector of m dimensions represents each document, where $m = 40$ in the histology atlas data set, and $m = 18$ in the basal-cell carcinoma data set. These are the total number of semantic keywords in each collection.

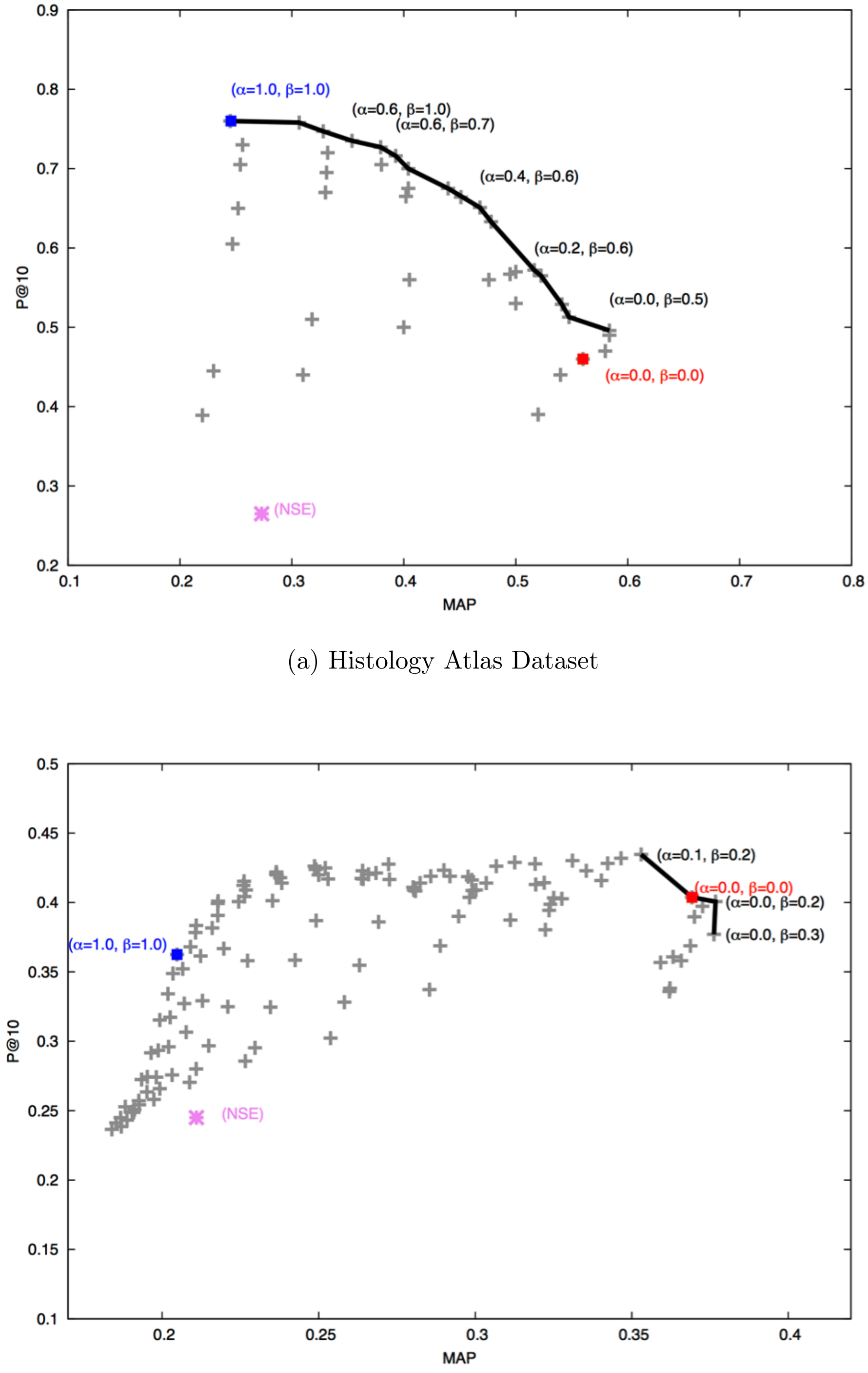
5.4.1 Setting Parameters for fused Search

As a first experiment, we want to review the impact of the two parameters α and β, which determine the weight of the visual data, for database images and query images, respectively. Since the fusion is achieved following a convex combination of both data modalities, the weight for semantic data is the complement of the value assigned to the visual data. By varying these two parameters, a different retrieval performance is obtained. In this setup, we are interested in optimizing the values of α and β to maximize simultaneously general retrieval precision (MAP) and early precision (P@10). This can be understood as a multi-objective optimization problem.

Since the number of parameters is just two, an exhaustive search was run varying α and β using a step of 0.1 in the interval [0.0, 1.0]. For each configuration, an evaluation of the resulting fused representation was made to measure MAP and P@10. The results for both data sets are presented in Figure 5.2. Points in the plot represent α and β pairs and their position in the cartesian plane reveals the obtained performance. The best solutions lie in the Pareto frontier which is shown as a black line in the plots. Note that the distribution of points for each data set looks different and so its Pareto frontier. The histology atlas data set has a wider optimal frontier with many parameter configurations providing a good performance trade-off, compared to the basal-cell carcinoma data set that has a narrower solution.

Two interesting configurations are highlighted in blue and red, which correspond to $\alpha = \beta = 1.0$ and 0.0, respectively. When the parameters are set to 1.0 the search process is configured to use only visual information, whereas with 0.0, it uses only semantic information. Notice that both configurations lie in opposite sides of the plot unveiling the tradeoff in performance between both data modalities. The blue point (a visual setup) provides a low MAP value with relatively high early precision, on the other hand, the red point (a semantic setup) provides high MAP performance but may decrease P@10 (early precision).

Some points in the Pareto frontier are labeled with the corresponding α and β values to illustrate good performing configurations. In both data sets, the solutions in the frontier tend to have a higher value for β with respect to α. This shows that query images require a slightly higher weight for the visual modality and database images require a slightly higher weight for the semantic modality. This finding supports the hypothesis that a good retrieval performance is achieved giving more importance to that modality that we can trust better, which is the visual data for query images because it is the observed data modality, while the semantic one is predicted by the model. On the other hand, for database images we can trust a little bit more to the semantic modality since this is the data that has been used for training the model.



(a) Histology Atlas Dataset

(b) Basal-cell Carcinoma Dataset

Figure 5.2: Performance on the retrieval task using different values of α and β. The x axis is MAP and the y axis is P@10. Points are pairs of values for α and β. Optimal configurations are on the Pareto frontier.

An interesting result is also illustrated in the plots: purple points correspond to the performance of the semantic search in the semantic space, as opposed to the semantic setup illustrated as a red point which is the performance of semantic search in the visual space, i.e., after a back-projection of the semantic data has taken place (see Equations 5.1 and 5.2). This result shows that just by back-projecting semantic data to the visual space, we are recovering important visual information that is exploited by the histogram intersection similarity during ranking, providing a significant boost in performance. This also provides evidence that focusing only on semantic data for histology image search leads to loosing important visual information and results in degraded performance.

5.5 Retrieval experiments

The following experiments aim to compare the performance of three search strategies: visual indexing, semantic indexing and multimodal indexing. Table 5.1 presents MAP and P@10 scores measured on the two evaluated data sets. These results show that the fused representation performs better than the visual and semantic retrieval strategies.

Consider the visual retrieval strategy as the baseline method in our setup, since no learning is employed to search images. Then, the semantic retrieval obtained by applying NSE to the database and query images improves in terms of MAP with respect to the baseline. Notice, however, that early precision as is measured by P@10, has decreased. This result indicates that the semantic search is able to find all relevant images in less result pages, but sacrificing the quality of the results in the very first page. The issue can be observed in both histology data sets. This finding supports the idea that summarizing images in a few keywords may lead to loss of discrimination power between images, as visual details are not available anymore. Actually, the good performance on early precision showed by the visual retrieval strategy suggests that very similar images with respect to only visual contents are likely to be relevant for users.

The multimodal indexing strategy produces the best performance in general search precision (MAP) and very competitive early precision (P@10) in the histology atlas data set. In the basal cell carcinoma data set, multimodal fusion outperforms the other methods regarding both criteria. This demonstrates the ability of the

Method	P@10	MAP	Improvement (MAP)
Visual Matching	0.7610	0.2451	N/A
NSE	0.2620	0.2704	+10.32%
NSE-BP ($\lambda = 0.6$, $\beta = 0.8$)	0.7590	0.3226	+31.61%

(a) Histology Atlas dataset

Method	P@10	MAP	Improvement (MAP)
Visual Matching	0.3615	0.2123	N/A
NSE	0.2757	0.2032	-4.28%
NSE-BP ($\lambda = 0.1$, $\beta = 0.2$)	0.4346	0.3530	+66.27%

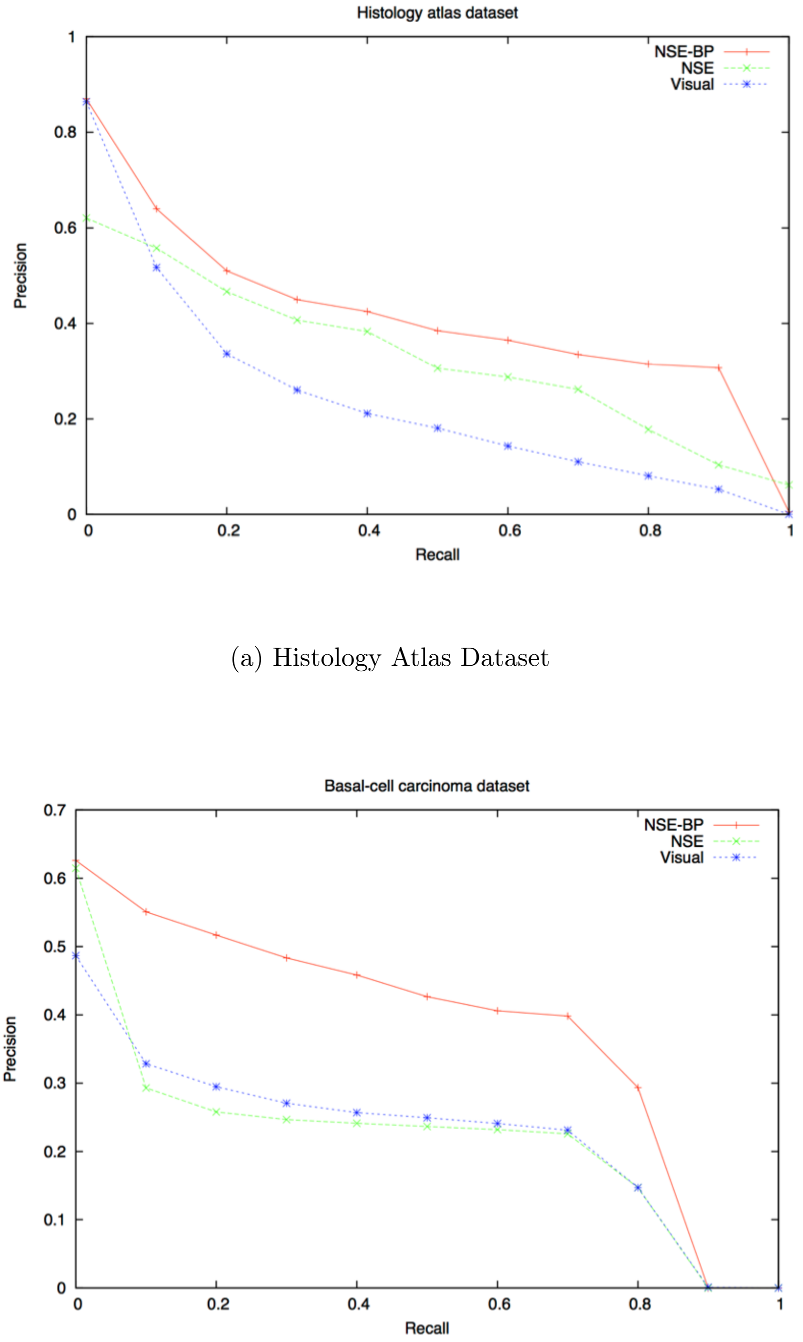
(b) Basal-cell Carcinoma Dataset

Table 5.1: Retrieval task performance for visual (Visual Matching), semantic (NSE) and fused (NSE-BP) representations.

proposed fusion strategy to harness visual and semantic information together to build an improved representation for images. The resulting representation is able to give information about visual details that can match with other images as well as general semantic concepts associated to them. Figure 5.3 presents the recall-precision graphs for these retrieval experiments, showing a significant improvement of the proposed multimodal fusion strategy. The Figure also allows to observe how the performance of the visual and semantic search overlap at some point during the retrieval process, due to the trade-off in performance between both modalities. Instead, the multimodal indexing produce consistently better results.

5.6 Conclusions

The main reason for studying the fusion of visual and semantic data is because they are complementary sources of information: while visual data tends to be ambiguous, semantic data tends to be very specific; and while visual data provides detailed appearance description, semantic data gives no clues on how an image looks like. So, depending on the fusion strategy, multimodal relationships become more useful for making decisions on data.



(a) Histology Atlas Dataset

(b) Basal-cell Carcinoma Dataset

Figure 5.3: Recall-Precision Graph for retrieval experiments on the Histology atlas data set and the carcinoma data set

Chapter 6

Parallelized Bag-of-Features

The bag-of-features is a descriptor that has become very popular and has been used actively in retrieval and annotation tasks, showing very good results [52]. The problem, is that computing this descriptor may result in a high computational cost, which makes it unfeasible to use on databases with hundreds of thousands of images. Due to this, it is proposed to extend this model to a distributed architecture, that allows a scalable solution that can be easily expandable by the addition of more resources to the computational infrastructure.

In this chapter a parallelization strategy is proposed based on the Map-Reduce Framework which has proven to be an extensible solution to solve the typical problems of working with large data sets.

The results presented in this chapter were published in *CLCAR 2012* [53].

6.1 Parallelized Bag-of-Features on a Map-Reduce architecture

Bag of Features representation has been successfully proved successful for different computer visions tasks, and it is very efficient in terms of memory usage in retrieval time, since the final result is a sparse representation. However, the representation construction process, requires a large computational cost. First, considering the initial step (visual words detection), a high amount of disk space is needed, due to the intermediate representation for a large number of patches or image subregions. For each image, about 1,000 patches can be extracted, resulting in an important storage factor if the image collection is large. The second stage (learning the visual vocabulary), demands large amounts of memory, since it is necessary to load and process a large sample of these feature vectors and compare them with each centroid, and furthermore, requires an excessive amount of computing time because the K-means algorithm needs numerous iterations until convergence is achieved. All this requirements make this model not a feasible choice for collections with hundreds of thousands of images.

To overcome this drawback, we consider a distributed strategy to extend the model for use in an architecture that achieves an efficient split of the computational load. For this purpose, we adopt the Map-Reduce architecture and reformulate the bag-of-features method to fit it naturally within the framework. The goal of the proposed approach is to compute the representation for a collection of images by harnessing computing resources in a cluster of machines dedicated to index images. White et al. [54] presented two strategies to compute the bag-of-features for web-scale image collections. However, no experimental evaluation was conducted by them neither to observe speedups nor to assess the quality of image retrieval. In this work, we present algorithmic details as well as comprehensive experiments to estimate the applicability of this approach to practical image search systems.

6.1.1 Map-Reduce framework

Map-Reduce is a framework proposed by Google to support distributed data processing. In simple terms, a program based on this architecture must have defined two functions: Map and Reduce. The Map function is applied in parallel to all input items, and Reduce function usually seeks to combine the results of processes performed in parallel by the MAP function. Both functions are defined with a structured data in key-value pairs. Map function takes a pair of data in one defined domain, and returns a list of pairs belonging to other domain:

$$Map(k1, v1) \to list(k2, v2) \tag{6.1}$$

The framework collects the pairs returned by all Maps and generate a group composed of pairs with the same key. The Reduce function is then applied to each group, which in turn produces a collection of values in the same domain:

$$Reduce(k2, list(v2)) \to list(v3) \tag{6.2}$$

The result of each Reduce call is typically one value or an empty return, though one call is allowed to return more than one value. Finally, the results of all Reduce calls are collected as the desired result list. Thus any algorithm can be implemented in this architecture if it can be defined in certain parts that can be processed independently in parallel and were only interested in combining the final result.

Chu et al. [55] show how is it possible to implement a parallelizable version of several machine learning algorithms using Map-Reduce and exploit a multi core architecture, only by rewriting the algorithms to achieve in a certain "Summation form". We follow some of these principles to design algorithms to compute the bag-of-features representation.

6.1.2 Bag-of-Features in a distributed architecture

Initially, we must note that the three main stages: feature extraction, visual vocabulary learning and histogram construction, should remain sequential to each other, since each stage depends on results obtained in the previous step, but within each stage we can define different parallelization strategies.

6.1.2.1 Distributed feature extraction

The objective in this step is to generate a set of feature vectors that describes each patch extracted from each image. The strategy of parallelization is very simple, since each image could be processed independently of the others. This stage can be divided basically into the following steps:

1. Extract a set of patches from each image
2. Generate a feature vector for each patch

The map and reduce stages can be defined as follows:

Map stage: process a subset of images, i.e., perform the feature extraction process for all patch set from each image. As result of the process a set of feature vectors is returned.

$$\begin{aligned} &Map_{fe}(image\,id, binary\,data) \\ &\rightarrow list(image\,id, feature\,vector) \end{aligned} \tag{6.3}$$

Reduce stage: In this case Reduce function takes no action in the data and the output list is identical to the input list.

$$\begin{aligned} &Reduce_{fe}(image\,id, list(feature\,vector)) \\ &\rightarrow list(feature\,vector) \end{aligned} \tag{6.4}$$

Figure 6.1 shows an overview of complete Map-Reduce cycle for this phase.

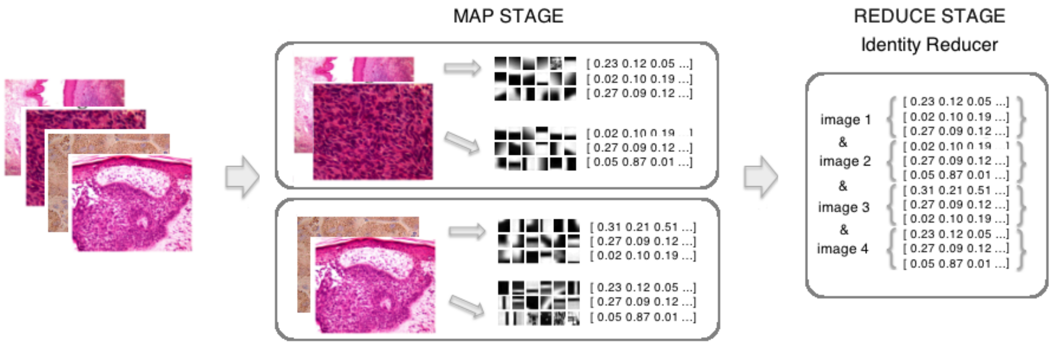


Figure 6.1: Map-Reduce cycle for feature extraction, 1. Map stage: perform the feature extraction process for all patch set, 2. Reduce stage: only groups together all result.

6.1.2.2 Learn visual vocabulary

For this step, a significant sample of all previously extracted feature vectors is used to learn the visual dictionary. The K-means algorithm can be computed in a distributed mode by delegating the distance computations to different workers. This stage can be divided basically into the following steps:

1. Initialize randomly a number k of centroids.
2. For each feature vector, calculate the nearest centroid using euclidian distance and assign the feature vector to this cluster (or visual word).
3. Re-compute centroids (mean of feature vectors in cluster)
4. Stop when there are no new re-assignments.

The map and reduce stages can be defined as follows:

Map stage: calculates the distances between samples and centroids; matches samples with the nearest centroid and assigns them to that specific cluster. Each map task is processed with a data block.

$$\begin{aligned} &Map_{lv}(visual\,word, feature\,vector)\\ &\rightarrow list(visual\,word, feature\,vector) \end{aligned} \tag{6.5}$$

Reduce stage: recalculates the centroid point using the average of the coordinates of all the points in that cluster. The associated points are averaged out to produce the new location of the centroid. The centroids configuration is given as feedback into the Mappers.

$$\begin{aligned} &Reduce_{lv}(visual\,word, list(feature\,vector))\\ &\rightarrow cluster = avg(feature\,vector) \end{aligned} \tag{6.6}$$

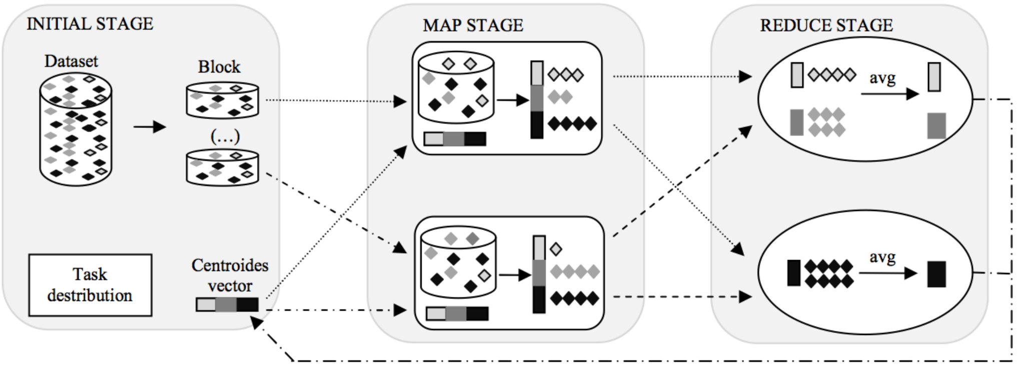


Figure 6.2: ”The three stages of a K-means job”, Rui Maximo Esteves, Rui Pais, and Chunming Rong. K-means clustering in the cloud – a mahout test. Advanced Information Networking and Applications Workshops, International Conference on, 0:514–519, 2011.

This Map-Reduce cycle is repeated until the centroids converge. The final centroids will be the representative vectors of each visual word. Figure 6.2 shows the three main stages of this step.

6.1.2.3 Quantize features using visual vocabulary

In this final step the final image representation is generated completing the following steps:

1. Translation of each feature vector to a visual word, using the previously learned codebook.

2. Calculate the histogram.

The map and reduce stages can be defined as follows:

Map stage: all features vector are compared with the learned visual codebook, and again the euclidian distance is used as proximity criterion to find the closest cluster, which will be assigned to the feature vector.

$$\begin{gathered} Map_{qf}(image\,id, feature\,vector) \\ \rightarrow list(visual\,word, 1) \end{gathered} \tag{6.7}$$

Reduce stage: counting is performed for all visual words found in Map stage.

$$\begin{gathered} Reduce_{qf}(visual\,word, list(visual\,word\,count)) \\ \rightarrow visual\,word\,count = \Sigma(visual\,word\,count) \end{gathered} \tag{6.8}$$

The result is a non-normalized histogram. Figure 6.3 shows an overview of complete Map-Reduce cycle for this phase.

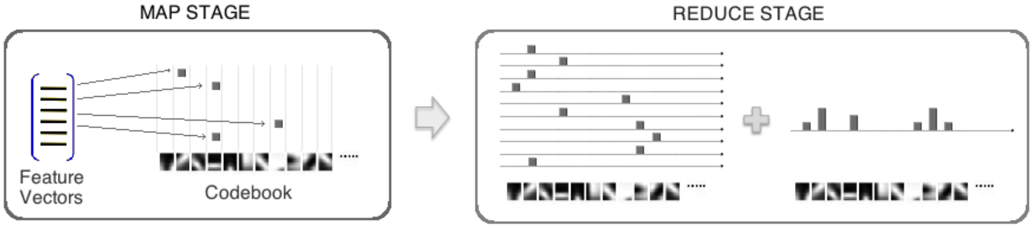


Figure 6.3: Map-Reduce cycle for Histogram Construction, 1. Map stage: perform the translation from feature vector to a visual word, 2. Reduce stage: count the number of occurrences of each visual word.

Finally, the distributed bag-of-features algorithm uses $2+n$ Map-Reduce cycles, where n is the number of cycles required until k-means algorithm converges.

6.1.3 Implementation

The distributed bag-of-features algorithm was written in Java programming language, and Implemented under Apache Hadoop framework, which is an open source implementation of Map-Reduce architecture. The k-means implementation of the Apache Mahout library [56] was used, which also provides various basic machine learning algorithms, that runs over a Hadoop system.

This framework has been configured in pseudo-distributed mode, allowing to run multiple instances in parallel on one machine, allowing to take advantage of multicore architecture.

The source code of the implemented algorithm, is accessible at `http://code.google.com/p/bioingenium-large-scale-tools`

6.1.4 Experimental evaluation

To have a basis for comparison, we defined a similar configuration to that used in last chapter [57] where a sequential bag-of-features was employed. The regular grid for feature extraction using patches of 8×8 pixels is used with an overlap of 4 pixels along the x and y axes. Each patch is processed using DCT (Discrete Cosine Transform) and is represented by the 21 largest coefficients.

In order to verify the obtained improvement using parallel processing, we run the proposed distributed bag-of-features algorithm, with a different number of processor cores. From a single core to 10 cores.

6.1.4.1 Experiment environment

The Map-Reduce framework was configured in pseudo-distributed mode, i.e., a single node, with a Intel(R) Xeon(R) CPU at 2.40GHz with 12 processor cores, and 32 GB of ram memory, with Apache Hadoop version 0.20.203 as Map-Reduce implementation installed over a Linux Ubuntu 10.04.

6.1.4.2 Visual search

We performed the retrieval experiment using direct visual matching, calculating the level of similarity between images using the histogram intersection similarity measure [50], as follows:

$$K_{HI}(x, y) = \sum_{i=1}^{n} min \{x_i, y_i\} \tag{6.9}$$

where x and y are images and x_i, y_i indicate the frequency of the $i-$th visual feature in each image respectively.

6.1.4.3 Time consumption

Table 6.1 shows the speedup obtained by increasing the number of cores. Processing time was calculated for each step. The results show a similar saving of time in each step. As the required time in the dictionary learning stage, depends on the required number of cycles until k-means algorithm converges, we calculated the average time spent in each cycle and we have calculated the total time spent if 52 iterations were required which is the amount used for a single core case. Table 6.1 shows that with the addition of an extra core, we can reduce by half the processing time, and may even further reduce processing time by adding more cores although the gain ratio is reduced.

Figure 6.4 indicates that under the Amdahl's law, approximately 85% of the algorithm is parallelizable. According to this percentage, we estimate that increasing the number of nodes, we could reach a speedup of about 6.6. Table 6.2 shows a comparison in retrieval performance between the proposed multicore version and the classic sequential BoF. In both cases we conducted 5 retrieval experiments and the results were averaged.

Table 6.1: Required computing time in a multicore infrastructure

Number of cores	**Feature extraction**		**Dictionary learning**		**Histogram representation**		**Totals**	
	Time (minutes)	**Speedup**	**Time** (minutes)	**Speedup**	**Time** (minutes)	**Speedup**	**Time** (minutes)	**Speedup**
1 core	66	–	344.5 (52) 65x5.3	–	90	–	**500.5**	**–**
2 cores	39	1.69	175.5 (61) 65x2.7	1.96	44	2.04	**258.5**	**1.93**
4 cores	31	2.12	110.5 (56) 65x1.7	3.12	31	2.90	**172.5**	**2.90**
10 cores	16	4.13	78 (43) 65x1.2	4.42	28	3.21	**122**	**4.10**

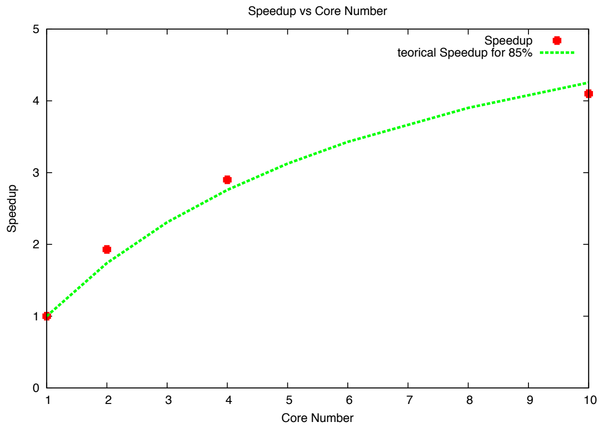


Figure 6.4: Speedup obtained by increasing the number of cores and Amdahl's law

6.1.4.4 Performance in retrieval tasks

As shown in Table 6.2 the proposed strategy does not present significant degradation in the quality of the representation, preserving the values of performance measures in a similar range. Variations may be due to the random initialization of the K-means algorithm, which still converges to a useful solution to construct the bag-of-features representation.

Table 6.2: Performance in retrieval tasks

Performance Metric	**Sequential BoF**	**Distributed BoF (10 cores)**
MAP	0.244±0.029	0.253±0.020
P@10	0.668±0.018	0.646±0.098
P@20	0.622±0.040	0.592±0.029

* Results reported for Visual Matching in[57]

6.2 Conclusions

We presented a strategy to extend the bag-of-features model in a distributed architecture allowing scalability to hundreds of thousands of images. The proposed parallelization strategy is based on the Map-Reduce Framework. This allows to achieve parallelism reformulating the algorithm as a combination of basic Map and Reduce functions, without altering the original algorithm. The experimental evaluation demonstrates important speedups by dividing the computing workload in multiple processing units. This allows to index large collections of images by harnessing the computing infrastructure in a dedicated cluster. Also, the resulting representation has no degradation in the quality of the representation, which translates in the same performance for the underlying image search task.

Although the step of learning the visual vocabulary is considerably reduced with this parallelization strategy, this is still expensive for a large number of images, due to the number of map-reduce cycles required until the algorithm converges. This can be improved by considering an online implementation of K-Means algorithm which shows a faster convergence than the conventional batch approach [58]. This is left for future research.

Chapter 7

Online Non-negative Semantic Embedding

The main drawback of current multimodal learning strategies is that associated algorithms are memory and computation intensive [59], which makes it difficult to use them in a large scale setup. For instance, the work of Romberg et al. [60] aims to build a multimodal index for a collection of 10 million Flickr images using a PLSA based algorithm. However, in their experimental setup, they only could apply the learning algorithm to a small sample of 10,000 images, losing the potential of such a vast amount of training data.

Recent works investigate the extent to which probabilistic models can be parallelized efficiently, overcoming underlying problems such as sharing data across workers and other memory restrictions [61]. Similar approaches have been proposed for parallelization of matrix factorization algorithms [62, 63] for web scale collections. However, it still requires large computational resources dedicated to decompose big matrices.

This Chapter presents an efficient version of the the NSE algorithm introduced in Chapter 4. That version of the algorithm is based on the NMF model, using the updating rule proposed by Lee et al. [45]. In order to make this algorithm applicable to large image collections, we reformulated it as a online learning algorithm allowing to reduce its computational requirements, both in terms of memory and processing time.

The results presented in this chapter has been submitted to *CIARP 3013.*

7.1 Online non-negative semantic embedding model

In order to make the NSE algorithm scalable, we want to reformulate the problem under an online formulation using stochastic gradient descent, which is a gradient descent optimization method for minimizing an objective function that is written as a sum of differentiable functions. In this context, we can formulate the problem of semantic embedding as the following optimization problem:

$$\min_{S \geq 0} \; d(V, ST) \tag{7.1}$$

where, $d(.,.)$ is a function that measures the difference between V and ST. The purpose is to find S that minimize this difference. The two most popular difference measures used in this kind of problems are the Frobenius norm (least squared loss function) and the Kullback-Leibler divergence. The optimization problem for the two kinds of measurement is solved below.

7.2 Frobenius norm optimization

Using the Frobenius norm, the optimization problem in Eq. 7.1 may be rewritten as:

$$\min_{S \geq 0} \; \left(\|V - ST\|_F^2 + \lambda R\left(S\right) \right) \tag{7.2}$$

where, $R(S)$ is a regularizer function and λ a regularization parameter. $\|\cdot\|_F$ is the Frobenius norm of a matrix:

$$\|V - ST\|_F^2 = \sum_{i=1}^{n} \sum_{k=1}^{l} \left(V_{ik} - \left(ST\right)_{ik}\right)^2 \tag{7.3}$$

where, n is the number of visual features and l is the number of training examples.

Defining $R(S)$ as the Frobenius norm of the solution matrix ($R(S) = \|S\|_F^2$) in the optimization problem, we get:

$$\min_{S \geq 0} \; \|V - ST\|_F^2 + \lambda \|S\|_F^2 \tag{7.4}$$

The gradient of the objective function $f(S)$ is:

$$\nabla f\left(S\right) = 2\left(ST - V\right)T^T + 2\lambda S \tag{7.5}$$

And the Updating rule for gradient descent optimization is:

$$S_{\tau+1} = S_\tau - \gamma \nabla f\left(S_\tau\right)$$

$$S_{\tau+1} = S_\tau + \gamma \left[\left(V - S_\tau T\right) T^{\mathsf{T}} - \lambda S_\tau\right] \tag{7.6}$$

where, γ is the step size.

7.3 Kullback-Leibler divergence optimization

Another popular optimization function for NMF is the generalized Kullback-Leibler divergence between V and ST [45], Although the KL-divergence equation is not symmetric, and therefore, it is not strictly a distance metric, this allows us to take advantage of the normalized visual and text representation that can be interpreted as probability distributions. Zhirong Yang et al. [64] show that projected gradient methods based in KL-divergence run faster and yield better approximation than others widely used NMF algorithms.

The corresponding optimization problem is as follows:

$$\min_{S \geq 0} \sum_{i=1}^{n} \sum_{k=1}^{l} \left(V_{ik} \log \left(\frac{V_{ik}}{(ST)_{ik}} \right) - V_{ik} + (ST)_{ik} \right) \tag{7.7}$$

The gradient of the optimization function $f(S)$ is:

$$\nabla f\left(S\right) = \left([1]_{n \times l} - \tfrac{V}{ST}\right) T^T$$

And finally the updating rule for gradient descent approach is:

$$S^{\tau+1} = S^\tau + \gamma \left[\left(\frac{V}{ST} - [1]_{n \times l} \right) T^T \right] \tag{7.8}$$

7.4 Non-negativity restriction

This algorithm requires a non-negativity restriction that can be easily incorporated by using a projected gradient strategy. The projection function maps a point back to the feasible region in each iteration [65].

With τ as the index of iterations, the projection function update the current solution S_τ to $S_{\tau+1}$ by the following rule:

$$S_{\tau+1} = P\left[S_\tau - \gamma \nabla f\left(S_\tau\right)\right]$$

$$P[s_{ij}] = \begin{cases} s_{ij} & if\ s_{ij} \geq 0, \\ 0 & if\ s_{ij} < 0, \end{cases} \tag{7.9}$$

7.5 Probabilistic interpretation

If each column of the V matrix is normalized, it can be interpreted as the conditional probability of finding a visual word i in a given document j. In the same way the T matrix can be interpreted as the conditional probability of finding a semantic term z in a given document j. And the S matrix can be interpreted as the probability of a visual word i is related with a given term z. This probabilistic model can be written as follows:

$$P(w_i \mid d_j) = \sum_{k=1}^{m} P(w_i \mid z_k) P(z_k \mid d_j) \tag{7.10}$$

7.6 Online formulation

The idea of online learning using stochastic approximations is to compute the new solution for each unknown in the problem using a single data sample at a time. Then, we can scan large data sets without memory restrictions, and this can be potentially scaled up to large image datasets. The updating rule has to be reformulated in such a way that it only depends on the τ-th sample (v_t , t_t visual and text features for the τ-th image). For the Frobenius norm, the updating rule is reformulated as follows:

$$S_{\tau+1} = S_\tau + \gamma \left[\left(v_\tau - S_\tau t_\tau \right) t_\tau^T - \lambda S_\tau \right] \tag{7.11}$$

With this rule, the transformation matrix is updated using only one image example with its corresponding text data at a time. Once the transformation matrix is calculated, the image can be discarded, keeping low memory usage requirements.

For the Kullback-Leibler divergence, the updating rule is reformulated as follows:

$$S_{\tau+1} = S_\tau + \gamma \left[\left(\frac{v_\tau}{S_\tau t_\tau} - [1]_{n\times 1} \right) t_\tau^T \right] \tag{7.12}$$

The resulting algorithm is shown in Algorithm 7.1. The algorithm 7.1 starts by randomly initializing the transformation matrix. Each iteration consists on updating the transformation matrix from an observed pair of visual and text features randomly obtained. Since the stochastic algorithm does not need to remember examples from

Algorithm 7.1 Online Non-negative Semantic Embedding

```
input $S^0$: initial transformation matrix, $\gamma_0$: initial step size, $N$ number of iterations
begin
  for $k = 1 : N$
    1. Step size calculation
       $\gamma_k = \gamma_0/(1 + \gamma_0 \lambda k)$
    2. Update transformation matrix
       $S_{\tau+1} = P\left[S_\tau - \gamma\left[\left(\frac{v_\tau}{S_\tau t_\tau} - [1]_{n\times m}\right) t_\tau^{\mathsf{T}}\right]\right]$
  end for
return $S_{\tau+1}$
end
```

previous iterations, it can process large data sets with very low memory usage. The step size used in this algorithm is a decreasing rate [66] that depends on the number of iterations and an initial learning rate γ.

A small variation of this algorithm is obtained by using several samples at each iteration instead of using only one. Experimental results show faster execution when using mini-batches instead of single examples, and also a better numerical stability for the solution.

7.7 Image indexing and search

A special indexing case is when images do not have attached text. This situation is very typical. For example, when users are interested in searching the database using example images as queries. A new image without text can be projected to the semantic space by finding the pseudo-inverse of the transformation matrix S^+.

$$S^+ = \left(S^T S + \beta I\right)^{-1} S^T$$

$$t = S^+ v \tag{7.13}$$

where, v is the visual representation of the new image and t is the semantic representation and β is a regularization parameter. In this way, we can search the database using an inferred text representation based in its visual features. Note the fact that this pseudo-inverse matrix has to be preprocessed only once and then it is stored in memory, making very efficient the projection process for a new image.

7.8 Experimental evaluation

This section is intended to evaluate the proposed algorithm in terms of performance in image retrieval task and its ability to scale up to large image collections.

7.8.1 Datasets

In order to evaluate the performance of the proposed algorithm we have used three different datasets with different size scales: Carcinoma dataset, Histology Atlas dataset and MIRFlickr 25000 dataset.

MIRFlickr 25000 dataset. The purpose of using this dataset is only in order to evaluate the proposed algorithm on a larger scale order. The MIRFlickr-25000 image data set is composed of 25,000 pictures downloaded from the popular online photo sharing service Flickr. These photos were collected directly from the web, to provide a realistic dataset for image retrieval research, with high-resolution images and associated metadata [67]. This image collection has been manually annotated using a set of 38 semantic terms.

7.8.2 Experimental setup

We conducted retrieval experiments under the query-by-example paradigm. In all datasets 20% of images were randomly selected as queries and the remaining images were used as the target collection to find relevant images.

We performed automatic experiments by sending a query to the system and evaluating the relevance of the results. A ranked image in the results list is considered relevant if it shares at least one keyword with the query. For this experiment, the evaluation was done using traditional measures of Image Retrieval, including Mean Average Precision (MAP) and the Recall-Precision plots.

Image Features: In all datasets we build a bag-of-features representation, with the following characteristics: We start by extracting patches of 8×8 pixels from a set of training images with an overlap of 4 pixels along the x and y axes. The DCT (Discrete Cosine Transform) transform is applied in each of the 3 RGB channels to extract the largest 21 coefficients and their associated functions. A k-means clustering is applied to build a dictionary, for Carcinoma and the Histology Atlas datasets we use 500 visual terms and for MIRFlickr we select a dictionary of 2000 features. Once the vocabulary has been built, every image in the collection goes through the patch extraction process. Each patch from an image is linked to one visual term in the dictionary using a nearest neighbor criterion. Finally, the histogram of frequencies is constructed for each image. We experimentally found that 500 visual terms for the Histology Atlas and Carcinoma and 2000 for MIRFlickr was enough to achieve a good performance, and that larger dictionaries do not provide significant improvements, but this would only add more computational load.

Text annotations: In these data sets the text annotations are clean and clearly defined terms from a technical vocabulary. Since the annotation process was followed by a systematic revision, there is no need to build a vector space model or to account for term frequencies. We build semantic vectors following a boolean approach, assigning 1 to the terms attached to an image and 0 otherwise. This leads to 46-dimensional binary vectors, for text representation in the Histology Atlas dataset, 18-dimensional binary vectors for Carcinoma dataset and 39-dimensional binary vectors for Flickr.

7.8.2.1 Convergence

The first evaluation conducted in this work is the analysis of convergence of the algorithm comparing the batch and online approaches.

Figure 7.1 shows the reconstruction error between the visual matrix and the multiplication of the textual matrix and the learned transformation matrix, as shown in the figure, the stochastic approach can achieve a very fast convergence rate in comparison with the batch algorithm. It needs only about 4 epochs to achieve convergence, on the other hand, the batch algorithm needs more than 50 epochs to achieve a similar reconstruction error. Even so, our goal is not to reconstruct exactly the visual matrix, but instead to obtain a transformation matrix the allows to find in the better way the relationship between the visual and text representation and serves as a bridge that could predict the possible text representation for images with only visual information, and use it to try to improve the retrieval performance. In order to do this, we perform retrieval experiments in several datasets shown as follows.

7.8.3 Retrieval performance

In order to evaluate the performance of the proposed algorithm, we compare the proposed online algorithm with the classical NSE (based in KL-divergence and using

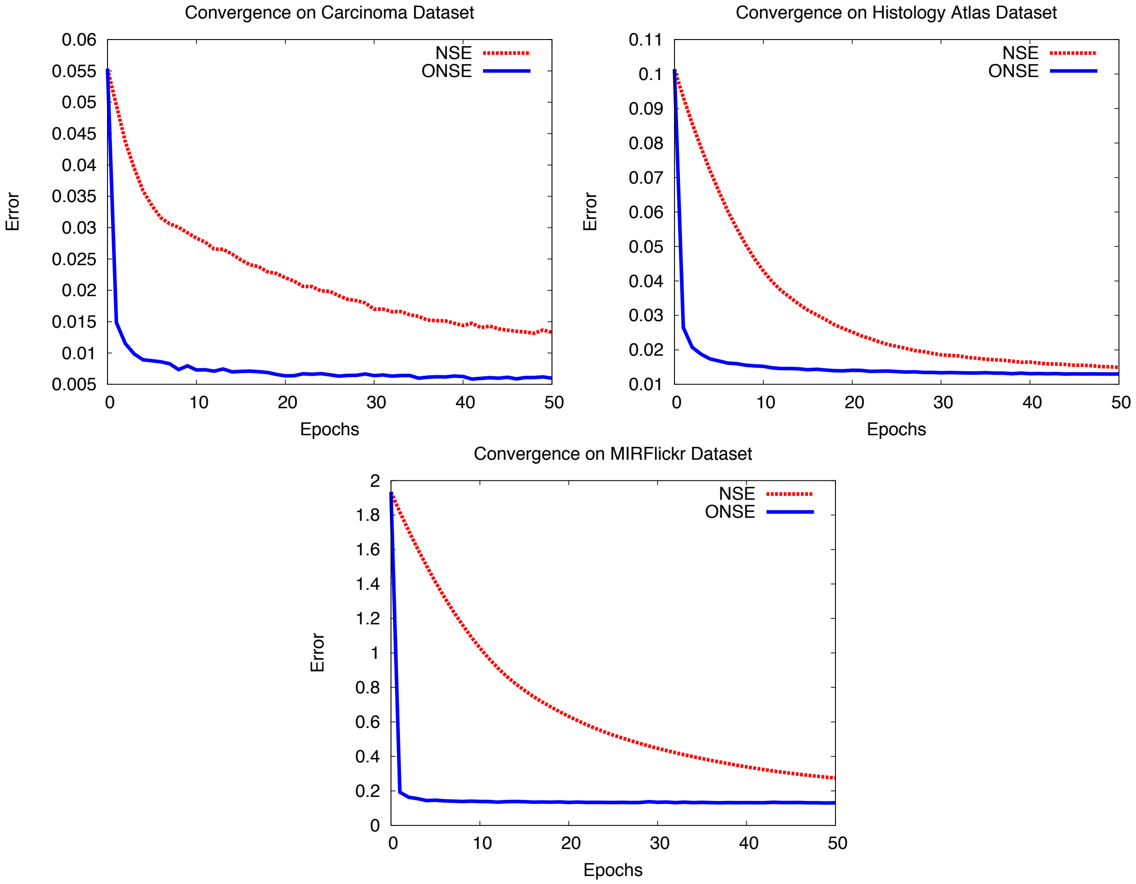


Figure 7.1: Convergence comparison for gradient descent approach between classical and stochastic formulations on Carcinoma, Histology Atlas and MIRFlickr data sets

	ONSE (FN)				ONSE (KL)			
Dataset	λ_0	γ	β	mini-batch size	λ_0	γ	β	mini-batch size
Carcinoma	2^{-3}	2^{-2}	2	16	2^{-5}	2^{-2}	2^{4}	16
Histology Atlas	2^{-4}	2^{-3}	1	16	2^{-6}	2^{-3}	2	16
MIRFlickr	2^{-7}	2^{-4}	1	32	2^{-8}	2^{-10}	2	32

Table 7.1: Results of parameter tuning for Online Non-negative Semantic Embedding (ONSE) using Frobenius Norm (FN) and Kullback-Leibler and divergence (KL) optimization approaches.

multiplicative updating rules) and the MMCR (Modified Multi-stage Convex Relaxation) proposed by Hsan et al. [68]. Although the MMCR algorithm was proposed mainly for annotation, it is possible to use its semantic score vector as a new representation for retrieval task. This evaluation consist in automatic experiments by sending a query to the system and evaluating the relevance of the results. A ranked image in the results list is considered relevant if it shares at least one keyword with the query.

7.8.4 Parameter tuning

Each algorithm has a set of parameters that can impact the quality of the resulting model. So, as preliminary evaluation, we focused in finding the better configuration for each algorithm in each particular dataset. We perform retrieval experiments using 10 fold cross-validation in the subset of 80% of the images that were not selected as queries. Finally, we select the configuration that performs best in average in all folds. The parameters that affect more drastically the quality of the results are the initial step size γ_0 and the regularization parameter λ. Improper settings of these parameters may cause the algorithm to converge very slowly or diverge. As a result, we present the best configuration for each dataset in the table 7.1.

Once, we had found the better configuration for this algorithm, we evaluate the proposed algorithms with the remaining 20% of images as test. So we use this 20% of images as queries and the 80% as finding objective. Table 7.2 summarizes the findings of our experimental results.

In all cases, a general improvement over visual baseline (direct visual matching using visual representation) is shown in MAP measure. We can also see that the ONSE algorithm based on Kullback-Leibler divergence, in all cases improves over the Least Squared based algorithm, showing that KL-divergence based algorithms are more suitable for this kind of tasks. Finally, with the exception of the Histology Atlas

	Carcinoma		**Histology Atlas**		**MIRFlickr**	
Algorithm	MAP	P@10	MAP	P@10	MAP	P@10
Visual	0.2236	0.3503	0.2107	0.6104	0.2505	0.4931
MMCR	0.3146	0.3322	0.5346	0.6030	0.3670	0.5063
NSE	0.3265	0.3249	0.4025	0.4148	0.3672	0.5079
ONSE (FN)	0.2621	0.2654	0.2621	0.2654	0.3405	0.4901
ONSE (KL)	0.3171	0.3651	0.3594	0.4439	0.3674	0.5065

Table 7.2: Image retrieval performance for all evaluated strategies. Reported measures are mean average precision (MAP) and precision at the first 10 results (P@10).

dataset, NSE, ONSE-KL and MMCR algorithms present a very similar performance.

7.8.5 Computational load

To show the ability to handle large scale dataset, we measure the time consumption against the other approaches. Table 7.3 shows the average time consumption for the training phase. Reported times are the result of running all algorithms 5 times in a CPU at 2.4 Ghz using only one core. The size of each dataset is also reported to observe how the algorithm complexity grows. NSE algorithm takes about 5 seconds to process the Carcinoma dataset, 9 to process the Histology Atlas collection and finally increases to 494 seconds for MIRFlicker. MMCR have the most time consuming, requiring about 2 seconds for Carcinoma, 14 for the Histology Atlas and 2834 for MIRFlickr. In contrast, the ONSE algorithm only requires 0.3 seconds for Carcinoma, 1.2 for the Histology Atlas and 27 for MIRFlickr. Thus for MIRFlickr dataset, ONSE algorithm is 18 times faster than NSE and 104 times faster than MMCR.

An important fact is that the main reason for the reduction of training time is that the number of required epochs until the ONSE algorithm converges is reduced drastically (convergence in all algorithms is verified by means of a minimum threshold required to improve the error in each epoch), in Carcinoma dataset we can see a reduction from 130 epochs for NSE to only 4 in the Online version. In general Bottou [69] shows that for a small collection, it is necessary to use very few epochs and for large collections, only one full scan is required.

7.8.6 Memory usage

Table 7.4 shows the maximum memory usage for each algorithm. The results show clearly that not only ONSE algorithm consumes less memory, but also the amount

Dataset	**Size**	**Algorithm**	**Epochs**	**Epoch Avg. Time** (sec)	**Total Avg. Time** (sec)
Carcinoma	1502	MMCR	8	0.2854	2.1878
		NSE	130	0.0411	5.3442
		ONSE (KL)	**4**	**0.0836**	**0.3345**
Histology Atlas	2641	MMCR	10	1.5351	14.2029
		NSE	90	0.1009	9.0869
		ONSE (KL)	**4**	**0.3027**	**1.2086**
MIRFlickr	25000	MMCR	10	283.4327	2834,3278
		NSE	200	2.4701	494.017
		ONSE (KL)	**2**	**13.755497**	**27.2188**

Table 7.3: Time consumption in training phase. Presents the amount of time required for each epoch (Epoch Avg. Time), and the total number of epochs and total average time required until the algorithm converges (Total Avg. Time).

		MMCR	**NSE (KL)**	**ONSE**
Dataset	Size	Memory (MB)		
Carcinoma	1502	232.6	188.8	**149.9**
Histology Atlas	2641	307.6	209.0	**158.8**
MIRFlickr	25000	2139.0	932.0	**293.0**

Table 7.4: Memory requirements in training phase, in megabytes (MB)

of memory required to increase the size of dataset does not increase as dramatically as in the other algorithms.

7.9 Conclusions

The main reason of the drastically reduced processing time of ONSE, is that the number of epochs in the algorithm is reduced to a constant, cutting down dramatically the time complexity. Our algorithm has been designed on top of stochastic learning theory that is guaranteed to converge in a few epochs for medium collections and just one full scan for large collection [69]. Furthermore, the proposed algorithm reduces the computational complexity decreasing the memory requirements. The only element that must be kept in memory is the transformation matrix, since visual and textual samples used in each update can be discarded, and it is only necessary to process a small number of vectors for each iteration. This makes the algorithm very suitable for large scale collections, since there is no practical limit of memory for scanning large image databases.

Chapter 8

Conclusions and Future Work

This work addressed the problem of retrieving histological images using as query an example image. Under this setup, the system relies mainly on processing the visual contents to find relevant images. But, matching visual similarities does not necessarily lead to meaningful results, a problem known as the semantic gap [8].

To overcome the semantic gap, a multimodal semantic indexing was proposed by exploiting additional information resources, such as image meta-data and accompanying text. An important issue is that in a very extensive collection, it is so difficult to ensure an adequate annotation for each image, and in general we are in a situation in which we have lots of images available and only a small portion of them is actually annotated.

In order to overcome this drawback we proposed strategies that takes advantage of the semantic information extracted from annotated images to improve the search process in the entire collection.

8.1 Semantic Representation with multimodal Information

This work presents a strategy to address the problem of indexing histological images, using the multimodal information drawn from two kinds of data: images and the available text from annotations. We introduce a strategy to find the relationships between these two data modalities, the Non-negative Semantic Embedding, which defines a mapping between visual and text data. Using this approach, the system is able to project new images to the space defined by the semantic annotations.

The experimental evaluation demonstrates that the NSE method increases the retrieval performance. The effectiveness of this method may be explained by the fact that it efficiently exploits the semantic information contained in text annotations. But we must take into account that this methodology is mainly applicable in cases where we have a controlled annotation process.

An important characteristic of the proposed method is that it naturally deals with different types of data: non-annotated images, images with multiple annotations, text queries, etc. This is accomplished by mapping both images and queries to a common semantic representation space.

The NSE method obtains a general increase in retrieval performance, but is surpassed in precision by the simple direct visual matching in the first results, where the images have the greatest visual similarity. In this way, NSE is useful in the situation where we are interested in finding a large amount of relevant results. But, if we are interested in retrieving a few results with a high precision it is better to use direct visual matching.

8.2 Fusing Visual and Semantic Contents

In order to improve the capacity of retrieving a greater amount of relevant results without losing precision in the first results, we proposed the NSE-BP (NSE Back projected) strategy that represents the images by combining visual and semantic features.

Fusing visual and semantic data is a good strategy because they are complementary sources of information: while visual data tends to be ambiguous, semantic data tends to be very specific; and while visual data provides detailed appearance description, semantic data gives no clues on how an image looks like. So, depending on the fusion strategy, multimodal relationships become more useful for making decisions on data.

8.3 Efficiency and scalability

The bag-of-features descriptor has shown very good results in retrieval and annotation tasks [52]. Also, the semantic learning strategies have proven to be an important tool to mitigate the semantic gap problem [8]. The main drawback of these strategies is that this kind of algorithms are memory and computation intensive [70].

To make the proposed strategies applicable to real world problems, we must ensure that these strategies are applicable to large collections of data. The strategies employed to achieve scalability are the parallelization of algorithms that distribute the computational load, and the reformulation as online algorithms which reduces the computational requirements, both in terms of memory and processing time.

For the bag-of-features we proposed a reformulation of the classical algorithm for a distributed architecture. This allows to achieve parallelism by reformulating the algorithm as a combination of basic Map and Reduce functions, without altering the original algorithm. The experimental evaluation demonstrates important speedups by dividing the computing workload in multiple processing units. This allows to index large collections of images by harnessing the computing infrastructure in a dedicated cluster. Also, the resulting representation has no degradation in the quality of the representation, which translates in the same performance for the underlying image retrieval task.

Although the step of learning the visual vocabulary is considerably reduced with this parallelization strategy, this is still expensive for a large number of images, due to the number of map-reduce cycles required until the algorithm converges. This can be improved by considering an online implementation of K-Means algorithm which shows a faster convergence than the conventional batch approach [58]. This is left for future research.

For NSE algorithm we proposed a reformulation as an online learning algorithm allowing to deal with large collections of data, achieving a significantly reduction in memory requirements and computational load, but keeping a competitive retrieval performance. The main reason that reduces the processing time is, that the number of necessary epochs in the ONSE decreases drastically. The algorithm approach that we follow in this work relies on recent theoretical advances for large scale learning, which provide guarantees about convergence and scalability [66, 71].

8.4 Future work

The fact that NSE and NSEBP are linear transformation models has several advantages, one of the most important is its simplicity, allowing easy implementation and scalability, but it also imposes significant restrictions that limit its flexibility. For example, semantics projection models could be extended by allowing a mapping in a nonlinear way which would allow more precise projection. Furthermore, the combination model by back-projection uses a simple linear combination between the two modalities, but it is possible that more complex combinations approaches can exploit better the relations between both modalities. Therefore, as a future work it would be interesting to explore non-linear strategies.

On the other hand, an important fact to keep in mind is that the two strategies used to achieve scalability of the algorithms can be used in combination. For instance, for the Bag-of-Features algorithm, we presented a parallelized strategy, but in addition, it can also be reformulated as an online algorithm. Likewise, we could further improve the ONSE algorithm on runtime and complexity, taking advantage of parallel infrastructures using parallelized formulation of stochastic gradient descent as is showed by Zinkevich et al.[72].

References

[1] Randall E. Millikan, Stuart O. Zimmerman, and Shida Jin. Tools for Study: National Databanking. In Cheryl T. Lee and David P. Wood, editors, *Bladder Cancer*, Current Clinical Urology, chapter 27, pages 295–302. Humana Press, Totowa, NJ, 2010.

[2] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *I. J. Medical Informatics*, 73(1):1–23, 2004.

[3] Henning Müller and Thomas M. Deserno. Content-based medical image retrieval. In *Biomedical Image Processing - Methods and Applications*. Springer, 2011.

[4] Payel Ghosh, Sameer Antani, L. Rodney Long, and George R. Thoma. Review of medical image retrieval systems and future directions. In *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6. IEEE, June 2011.

[5] Chi-Ren Shyu, Carla Brodley, Avi Kak, Akio Kosaka, Alex M. Aisen, and Lynn S. Broderick. ASSERT: A Physician-in-the-loop Content-Based Retrieval System for HRCT Image Databases. *Computer Vision and Image Understanding*, 75:111–132, 1999.

[6] Thomas M. Lehmann, Henning Schubert, Daniel Keysers, Michael Kohnen, and Berthold B. Wein. The IRMA code for unique classification of medical images. *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, 5033(1):440–451, 2003.

[7] S. Bhadoria and C.G. Dethe. Study of medical image retrieval. In *Data Storage and Data Engineering (DSDE), 2010 International Conference on*, pages 192 –196, feb. 2010.

[8] A W M Smeulders, M Worring, S Santini, A Gupta, and R Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[9] Mayank Agarwal and Javed Mostafa. Content-based image retrieval for Alzheimer's disease detection. In *2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 13–18. IEEE, June 2011.

[10] Henning Müller, Samuel Marquis, Gilles Cohen, and Antoine Geissbuhler. Lung CT analysis and retrieval as a diagnostic aid. In Rolf Engelbrecht, Antoine Geissbuhler, Christian Lovis, and George Mihalas, editors, *Proceedings of the Medical Informatics Europe Conference (MIE 2005)*, volume 116, pages 453–458, Geneva, Switzerland, 2005.

[11] Zeng J. Chouikha M. Byrd, K. An assessed digital mammography segmentation algorithm used for content-based image retrieval. volume 2, Guilin, 2007. cited By (since 1996) 0; Conference of 8th International Conference on Signal Processing, ICSP 2006; Conference Date: 16 November 2006 through 20 November 2006; Conference Code: 69656.

[12] Junding Sun and Zhaosheng Zhang. An Effective Method for Mammograph Image Retrieval. In *Computational Intelligence and Security, 2008. CIS '08. International Conference on*, volume 1, pages 190–193, 2008.

[13] I. El-Naqa, Yongyi Yang, N.P. Galatsanos, and M.N. Wernick. Content-based image retrieval for digital mammography. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 3, pages III–141 – III–144 vol.3, 2002.

[14] De Azevedo-Marques P.M.a Pereira Jr. R.R.a Rodrigues J.A.H.a Rangayyan R.M.c Kinoshita, S.K.a b. Content-based retrieval of mammograms using visual features related to breast density patterns. *Journal of Digital Imaging*, 20(2):172–190, 2007. cited By (since 1996) 16.

[15] Shengwen Guo and Jinshan Tang. Content based image retrieval from chest radiography databases. In *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*, pages 3–6, 2009.

[16] Md.M. Rahman, S K Antani, and G R Thoma. Biomedical image retrieval in a fuzzy feature space with affine region detection and vector quantization of a scale-invariant descriptor, 2010.

[17] Pedro H. Bugatti, Agma J. M. Traina, and Caetano Traina. Improving content-based retrieval of medical images through dynamic distance on relevance feedback. In *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6. IEEE, June 2011.

[18] M C Dï¿œaz-Galiano, M T Martï¿œn-Valdivia, and L A Ureï¿œa-Lï¿œpez. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in Biology and Medicine*, 39(4):396–403, April 2009.

[19] Gowri Allampalli-Nagaraj and Isabelle Bichindaritz. Automatic semantic indexing of medical images using a web ontology language for case-based image retrieval. *Engineering Applications of Artificial Intelligence*, 22(1):18–25, February 2009.

[20] D K Iakovidis, D Schober, M Boeker, and S Schulz. An ontology of image representations for medical image mining. In *Final Program and Abstract Book - 9th International Conference on Information Technology and Applications in Biomedicine, ITAB 2009*, 9th International Conference on Information Technology and Applications in Biomedicine, ITAB 2009, Deptartment of Informatics and Computer Technology, Technological Educational Institute of Lamia, GR-35100 Lamia, Greece, 2009.

[21] Jaafar BenAbdallah, Juan C. Caicedo, Fabio A. Gonzalez, and Olfa Nasraoui. *Multimodal Image Annotation Using Non-negative Matrix Factorization*. IEEE, August 2010.

[22] Fabio A González, Juan C Caicedo, Olfa Nasraoui, and Jaafar Ben-Abdallah. NMF-based multimodal image indexing for querying by visual example. In *ACM International Conference On Image And Video Retrieval CIVR2010*, pages 366–373. ACM Press, 2010.

[23] Yong Wang, Tao Mei, Shaogang Gong, and Xian-Sheng Hua. Combining global, regional and contextual features for automatic image annotation. *Pattern Recognition*, 42(2):259–266, February 2009.

[24] Zhixin Li, Zhiping Shi, Xi Liu, and Zhongzhi Shi. Modeling continuous visual features for semantic image annotation and retrieval. *Pattern Recognition Letters*, 32(3):516–523, February 2011.

[25] Zhixin Li, Zhiping Shi, Xi Liu, Zhiqing Li, and Zhongzhi Shi. Fusing semantic aspects for image annotation and retrieval. *Journal of Visual Communication and Image Representation*, 21(8):798–805, November 2010.

[26] Qin Chen, Xiaoying Tai, Baochuan Jiang, Gang Li, and Jieyu Zhao. Medical Image Retrieval Based on Latent Semantic Indexing. In *Computer Science and Software Engineering, 2008 International Conference on*, volume 4, pages 561–564, 2008.

[27] Adrien Depeursinge and Henning Müller. Fusion techniques for combining textual and visual information retrieval. In W Bruce Croft, Henning Müller, Paul Clough, Thomas Deselaers, and Barbara Caputo, editors, *ImageCLEF*, volume 32 of *The Springer International Series On Information Retrieval*, pages 95–114. Springer Berlin Heidelberg, 2010. 10.1007/978-3-642-15181-1_6.

[28] L Wang and B S Manjunath. A semantic representation for image retrieval. *International Conference on Image Processing Barcelona 2003*, 2:523–526, 2003.

[29] Henning Müller, Thomas Deselaers, Thomas Deserno, Paul Clough, Eugene Kim, and William Hersh. Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks. In *Evaluation of Multilingual and Multimodal Information Retrieval*, pages 595–608. Springer, 2007.

[30] Adil Alpkocak, Okan Ozturkmenoglu, and Tolga Berber. DEMIR at ImageCLEFMed 2011 : Evaluation of Fusion Techniques for Multimodal Content-based Medical Image Retrieval. *Working Notes of CLEF 2011*, 2011.

[31] Jacquet Guillaume Csurka Gabriela, Clinchant Stephane. XRCE's Participation at Medical Image Modality Classification and Ad-hoc Retrieval Tasks of ImageCLEF 2011. *Working Notes of CLEF 2011*, 2011.

[32] Lei Zheng, A. W. Wetzel, J. Gilbertson, and M. J. Becich. Design and analysis of a content-based pathology image retrieval system. *Trans. Info. Tech. Biomed.*, 7(4):249–255, December 2003.

[33] N. Bonnet. Some trends in microscope image processing. *Micron*, 35(8):635–653, December 2004.

[34] F. González, J. Caicedo, A. Cruz, J. Camargo, E. Romero, C. Spinel, D. Seligmann, and J. Forero. A web-based system for biomedical image storage, annotation, content-based retrieval and exploration. 2009.

[35] Juan C. Caicedo, Fabio A. Gonzalez, and Eduardo Romero. Content-based histopathology image retrieval using a kernel-based semantic annotation framework. *J. of Biomedical Informatics*, 44(4):519–528, August 2011.

[36] Christopher D. M. Fletcher. *Diagnostic Histopathology of tumors.* Elsevier Science, 2003.

[37] C.S.M. Wong, R.C. Strange, and J.T. Lear. Basal cell carcinoma. *BMJ*, 327(7418):794–8, 2003.

[38] Gloria Díaz and Eduardo Romero. Histopathological image classification using stain component features on a plsa model. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, LNCS 6419:55–62, 2010.

[39] Angel Cruz-Roa, Juan C. Caicedo, and Fabio A. Gonzalez. Visual pattern analysis in histopathology images using bag of features. *Iberoamerican Conference on Pattern Recognition*, LNCS 5856:521–528, 2009.

[40] Thorsten Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 137–142, London, UK, UK, 1998. Springer-Verlag.

[41] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

[42] Angel Cruz-Roa, Juan C. Caicedo, and Fabio A. González. Visual pattern mining in histology image collections using bag of features. *Artificial Intelligence in Medicine*, 52(2):91–106, June 2011.

[43] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and CÃ©dric Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[44] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[45] D D Lee and H S Seung. New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation. *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, 13(1):V–621–V–624, 2001.

[46] Juan C. Caicedo, Jaafar BenAbdallah, Fabio A. González, and Olfa Nasraoui. Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. *Neurocomput.*, 76(1):50–60, January 2012.

[47] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 267–273, New York, NY, USA, 2003. ACM.

[48] Henning Müller, Wolfgang Müller, David M. Squire, Stéphane Marchand-Maillet, and Thierry Pun. Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recognition Letters*, 22(5):593–601, April 2001.

[49] Jorge A. Vanegas, Juan C. Caicedo, Fabio A. GonzÃ¡lez, and Eduardo Romero. Histology image indexing using a non-negative semantic embedding. In Henning MÃŒller, Hayit Greenspan, and Tanveer Fathima Syeda-Mahmood, editors, *MCBR-CDS*, volume 7075 of *Lecture Notes in Computer Science*, pages 80–91. Springer, 2011.

[50] Michael J Swain and Dana H Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[51] Jorge A. Vanegas, Juan C. Caicedo, and Fabio A. Gonzalez. Histology image indexing combining visual and semantic features. 2011.

[52] Anna Bosch, Xavier Muñoz, and Robert Martí. Which is the best way to organize/classify images by content? *Image and Vision Computing*, 25(6):778–791, June 2007.

[53] Jorge A. Vanegas, Juan C. Caicedo, and Fabio A. Gonzalez. The latin american conference on high performance computing. clcar 2012. 2012.

[54] Brandyn White, Tom Yeh, Jimmy Lin, and Larry Davis. Web-scale computer vision using MapReduce for multimedia data mining. In *Proceedings of the*

Tenth International Workshop on Multimedia Data Mining, MDMKDD '10, New York, NY, USA, 2010. ACM.

[55] Cheng T. Chu, Sang K. Kim, Yi A. Lin, Yuanyuan Yu, Gary R. Bradski, Andrew Y. Ng, and Kunle Olukotun. Map-reduce for machine learning on multicore. In Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors, *NIPS*, pages 281–288. MIT Press, 2006.

[56] Rui Maximo Esteves, Rui Pais, and Chunming Rong. K-means clustering in the cloud – a mahout test. *Advanced Information Networking and Applications Workshops, International Conference on*, 0:514–519, 2011.

[57] Jorge A. Vanegas, Juan C. Caicedo, Fabio A. González, and Eduardo Romero. Histology image indexing using a non-negative semantic embedding. In *MCBR-CDS*, pages 80–91, 2011.

[58] LÃ©on Bottou and Yoshua Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7*, pages 585–592. MIT Press, 1995.

[59] Pulla Chandrika and C. V. Jawahar. Multi modal semantic indexing for image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 342–349, New York, NY, USA, 2010. ACM.

[60] Stefan Romberg, Rainer Lienhart, and Eva Hörster. Multimodal image retrieval. *International Journal of Multimedia Information Retrieval*, 1(1):31–44, April 2012.

[61] Raymond Wan, Vo N. Anh, and Hiroshi Mamitsuka. Efficient probabilistic latent semantic analysis through parallelization information retrieval technology. volume 5839 of *Lecture Notes in Computer Science*, chapter 38, pages 432–443. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2009.

[62] Rainer Gemulla, Erik Nijkamp, Peter J. Haas, and Yannis Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 69–77, New York, NY, USA, 2011. ACM.

[63] Chao Liu, Hung-chih Yang, Jinliang Fan, Li-Wei He, and Yi-Min Wang. Distributed nonnegative matrix factorization for web-scale dyadic data analysis on

mapreduce. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 681–690, New York, NY, USA, 2010. ACM.

[64] Zhirong Yang, He Zhang, Zhijian Yuan, and Erkki Oja. Kullback-leibler divergence for nonnegative matrix factorization. In *ICANN (1)*, pages 250–257, 2011.

[65] Chih jen Lin. Projected gradient methods for non-negative matrix factorization. Technical report, Neural Computation, 2007.

[66] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France, August 2010. Springer.

[67] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[68] Min-Hsuan Tsai, Jinjun Wang, Tong Zhang, Yihong Gong, and Thomas S. Huang. Learning semantic embedding at a large scale. In *ICIP*, pages 2497–2500, 2011.

[69] Léon Bottou and Yann LeCun. Large scale online learning. In *NIPS*, 2003.

[70] Pulla Chandrika and C. V. Jawahar. Multi modal semantic indexing for image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 342–349, New York, NY, USA, 2010. ACM.

[71] L Eon Bottou and Yann Le Cun. Large scale online learning. In *In NIPS*, page 2004. MIT Press, 2003.

[72] Martin Zinkevich, Markus Weimer, Alex Smola, and Lihong Li. Parallelized stochastic gradient descent. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2595–2603, 2010.

Appendix A

Histology Image Indexing Using a Non-negative Semantic Embedding

Vanegas J.A., Caicedo J.C., González F.A. & Romero E., Histology Image Indexing Using a Non-negative Semantic Embedding. In Medical Content-Based Retrieval for Clinical Decision Support. MICCAI 2011. Volume 7075, 2012, pp 80-91, ISBN:978-3-642-28459-5, ISSN:0302-9743. Springer-Verlag GmbH Berlin Heidelberg.

Histology Image Indexing using a Non-negative Semantic Embedding

Jorge A. Vanegas, Juan C. Caicedo, Fabio A. González, and Eduardo Romero

Bioingenium Research Group
National University of Colombia
{javanegasr,jccaicedoru,fagonzalezo,edromero}@unal.edu.co
http://www.bioingenium.unal.edu.co

Abstract. Large on-line collections of biomedical images are becoming more common and may be a potential source of knowledge. An important unsolved issue that is actively investigated is the efficient and effective access to these repositories. A good access strategy demands an appropriate indexing of the collection. This paper presents a new method for indexing histology images using multimodal information, taking advantage of two kinds of data: visual data extracted directly from images and available text data from annotations performed by experts. The new strategy called Non-negative Semantic Embedding that extends the NMF algorithm define a mapping between visual an text data assuming that the space spanned by the text is good enough representation of the images semantic. Evaluation of the proposed method is carried out by comparing it with other strategies, showing a remarkable image search improvement since the proposed approach effectively exploits the image semantic relationships.

1 Introduction

Digital microscopy is currently an important tool to support the decision making process in clinical and research environments. Instead of moving the glass around, specialists can share a digitized sample, allowing other physicians to navigate the slide by the use of a virtual microscope [10]. The digitization process results in very large files known as virtual slides, or alternatively, many different microphotographs taken from the same slide. A large database of these micropictures can support training, educational and research activities. Likewise, the great amount of information contained in these image collections and their accompanying meta-data is a potential resource to support the specialist's decision-making processes. However, accessing and retrieving images from a large collection is a high time consuming task. In this work, we consider the problem of retrieving histological images using as query an example image. Under this setup, the system relies mainly on processing the visual contents to find relevant images. Yet, matching visual similarities does not necessarily lead to meaningful results, a problem known as the semantic gap [11].

To overcome this problem, semantic indexing has been proposed by exploiting additional information resources, such as image meta-data and accompanying text. The main problem is that collecting that data is an arduous process which makes it very difficult to ensure an adequate annotation for each image in a very extensive collection. Therefore, we are in a situation in which we have lots of images available and only a small portion of them is actually annotated. So, the challenge is to design a method that takes advantage of the semantic information extracted from annotated images to improve the search process in the entire collection.

In this paper we address the problem of indexing histological images, using the multimodal information drawn from two kinds of data: images and the available text from annotations. We introduce a new strategy to find the relationships between these two data modalities, the Non-negative Semantic Embedding, which defines a mapping between visual and text data. Using this approach, the system is able to project new images to the space defined by the semantic annotations. This work presents two main contributions: first, a new method to index images using multimodal information, and second, an experimental evaluation of histological images which defines the role of semantic data in this particular field. The rest of this paper is organized as follows: Section 2 discusses the related work; Section 3 presents the structure of the histology image collection used in this work; Section 4 introduces the proposed method called Non-negative Semantic Embedding; Section 5 presents the experimental evaluation; and, finally, Section 6 presents some concluding remarks.

2 Related Work

Content-based retrieval has been evaluated for histological images using low-level visual features and similarity measures [14,2], allowing users to access the collection using a query-by-example (QBE) paradigm. The semantic analysis and automatic classification of histology images [13,1] can be used to improve the response of a retrieval system.

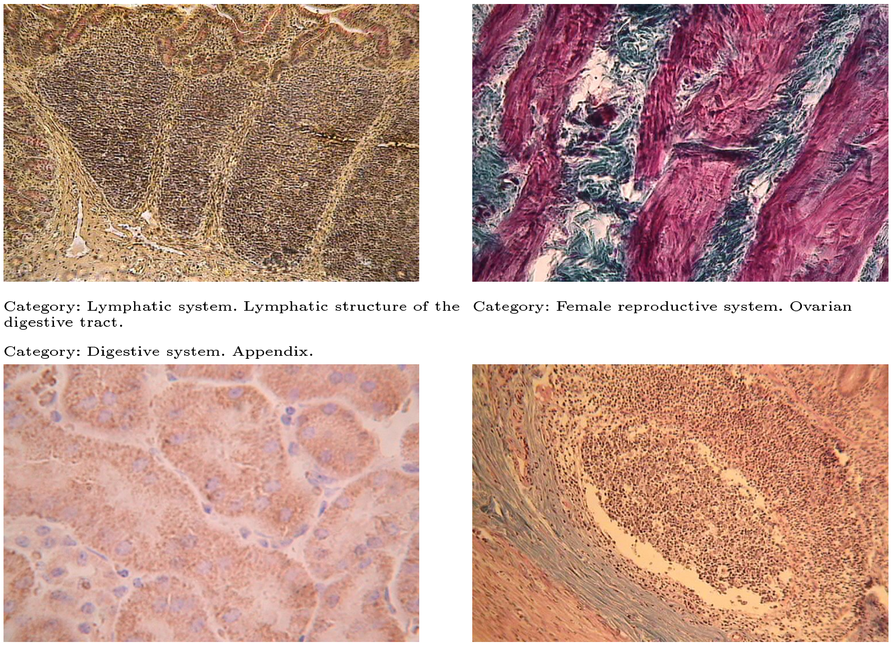
The ImageCLEFmed campaign organizes an experimental evaluation of medical image retrieval, providing access to a large dataset with multimodal data (images and text). The 2010 version of this event showed important benefits in performance of using multimodal data to retrieve relevant images [9]. Similarly, our approach intends to exploit multimodal data in histology image collections for building a semantic index that supports QBE.

Recently, a strategy for multimodal indexing of images was proposed to improve the performance of a retrieval system that works under the QBE paradigm [5]. This strategy models latent factors using Non-negative Matrix Factorization (NMF) to find meaningful relationships between visual and text data. We build on top of these ideas to propose a novel algorithm, the Non-negative Semantic Embedding, which is used to build a semantic index of histology images.

3 Histology Images

Images in this work have been collected to build an atlas of histology for the study of the four fundamental tissues. The collection comprises 20,000 images of these tissues, from different biological systems and organs, which were stained using Hematoxiline & Eosine and Immunohistochemical techniques (available at www.informed.unal.edu.co). The collection includes photographs of histology slides acquired with a digital camera coupled to a microscope, using different magnification factors to focus important biological structures. The main use of this collection is for educational and research purposes, and is accessible via Internet at www.informed.unal.edu.co.

From this large collection, a subset of 2,641 images was selected for this work. Each of these images was annotated by an expert, indicating the biological system and organs that can be observed. The total number of different categories is 46, after a standardization of the vocabulary used to describe the semantic contents. The list of terms includes circulatory system, hearth, lymphatic system and thymus, among others. Usually, images have just one category attached to it, but in several cases images can have more than one category . Figure 1 shows example images with the corresponding associated terms.



Category: Lymphatic system. Lymphatic structure of the digestive tract.

Category: Digestive system. Appendix.

Category: Female reproductive system. Ovarian

Category: Urinary system. Kidney.

Category: Lymphatic system. Lymphatic structure of the digestive tract.

Category: Digestive system. Ileus.

Fig. 1. Example images with associated terms

4 Matrix Factorization for Multimodal Indexing

Assume a discrete image representation for the visual contents using, for instance, a bag of features, i.e., a random collection of parts of the image contents whose distribution probability is approached by a simple frequentist approach, that is to say, the frequency of these patches is determined within the whole image collection. Similarly, assume a text representation for annotations, also approached by term frequencies in the attached annotations and also using a bag of words. Then, we can build two matrices to describe the occurrence of visual and textual features in the image collection. Let $X_v \in \mathbb{R}^{n \times l}$ be the matrix of visual features, where n is the number of visual features and l is the total number of images in the collection. Let $X_t \in \mathbb{R}^{m \times l}$ be the matrix of text term frequencies, where m is the number of terms or keywords.

Our goal is to uncover the underlying structure of these matrices to build an effective index for image search. The proposed strategy is based on the Non-negative Matrix Factorization (NMF) algorithm, to find a linear representation of the data, which must be non negative [7].

4.1 Non-negative Matrix Factorization

The main purpose of NMF is to find an approximation of the matrix X, in terms of two smaller matrices as follows:

$$X = WH \ . \tag{1}$$

where $X \in \mathbb{R}^{p \times l}$, $W \in \mathbb{R}^{p \times r}$, $H \in \mathbb{R}^{r \times l}$, p is the total number of available features, l is the number of images in the collection, and r is the rank of the decomposition. The matrix W is known as the basis matrix and H is known as the encoding matrix. This factorization is found by solving the associated optimization problem to minimize the squared error of the decomposition or the Kullback Leibler divergence between the original matrix and the reconstructed one.

As with González et al. [5], we use the divergence criterion in this work, following the Lee and Seung's approach [6] to find the decomposition with two multiplicative updating rules for W and H, respectively:

Updating rule for W

$$W_{ia} \leftarrow W_{ia} \frac{\sum_{\mu} H_{a\mu} X_{i\mu}/(WH)_{i\mu}}{\sum H_{av}} \ . \tag{2}$$

Updating rule for H

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_{i} W_{ia} X_{i\mu}/(WH)_{i\mu}}{\sum_{k} W_{ka}} \ . \tag{3}$$

4.2 Multimodal Indexing Via NMF

González et al. [5] proposed two ways of using NMF for multimodal indexing of images. The first strategy, the NMF-mixed, builds a multimodal matrix $X = \left[\alpha X_v^T \; (1-\alpha) X_t^T\right]^T$ with $\alpha \in [0, 1]$, that contains the visual and text data. Matrices are mixed after normalization of each vector to have L2-norm = 1. This matrix is then decomposed using NMF to model a set of latent factors (columns of the matrix W) that have features of both modalities.

The second strategy, the NMF-asymmetric, follows a two step process, starting with the decomposition of the text matrix. After that, the obtained encoding for images (columns of the matrix H) is fixed to find a second matrix of latent factors for visual contents. The goal is to find a semantic image representation based on the relationships between text terms only, and then find a function to project visual features to the same latent space. The two steps of the NMF-asymmetric strategy are as follows:

$$X_t = W_t H \ . \tag{4}$$

$$X_v = W_v H \ . \tag{5}$$

where W_t is the matrix of latent factors for text, W_v is the matrix of latent factors for visual features and H is the common representation for both modalities. The matrix W_v is obtained by running the multiplicative update for W only (Equation 2) since the matrix H is fixed.

4.3 Non-negative Semantic Embedding

The methods described in the previous Subsection are oriented to model latent factors for multimodal data, that is, to find the hidden structure of the collection, which is assumed to be common between the two data modalities. We propose a simplified strategy that extends the NMF-asymmetric algorithm to a setting in which the semantic encoding is already known.

We assume that the space spanned by text terms is a good enough representation of image semantics, and we use it to index and represent all images in the collection. Then, we want to find a way to embed visual features in this semantic space, to index images with or without annotations. We formulate this problem as finding a linear transformation of the visual data imposing a non negativity constraint on the solution, as follows:

$$X_v = W X_t ; W \geq 0 \ . \tag{6}$$

where $W \in \mathbb{R}^{n \times m}$, n the number of visual features and m the number of text terms. The non-negativity constraint in this case enforces an additive reconstruction of visual features, since vectors in the matrix W can be thought of as parts of images that are combined according to the presence of associated labels. Notice that the encoding matrix is the matrix of text annotations, and the vectors in W can also be directly interpreted as the visual features related to each text term.

Finding the matrix W when X_v and X_t are known is a convex problem under the divergence or euclidean criteria in the related minimization problem. However, we approximate the solution to this problem using the NMF updating rule for the matrix W. Instead of requiring a global optimum for this problem, which might result in overfitting to the training data, we accept a good approximate solution obtained from the Lee and Seung's approach. The updating rule usually converges to a local minimum, but it may result in a better generalization with some robustness to intrinsic noise in the training data.

We call this approach the Non-negative Semantic Embedding (NSE) taking into account that the semantic space is known in the problem and the resulting solution to embed image features on it is non-negative. It also differs from the Gonzalez et al. [5] approach in the sense that no latent factors are herein modeled, but instead, a semantic space is assumed from the given text terms space.

4.4 Image Indexing and Search

The indexing methods described above require a training phase to learn a mapping of images to the semantic or latent space. The result of that training phase is a matrix W, which contains the basis of the indexing space and serves as linear transformation to project new data. So, when new images are obtained without text annotations, either to be included in the collection or as queries, we can find the semantic representation for this image using visual features only.

Let y be the visual features of a new image, unseen during training. To embed this image in the semantic space, the following equation needs to be solved:

$$y = Wh \ . \tag{7}$$

where W is the basis of the semantic space and h is the semantic representation of the new image. The image y will be embedded into the semantic space by finding the vector $h \geq 0$ that satisfies the equation. This is done by using the multiplicative updating rule for h while keeping the matrix W fixed.

This strategy applies for both, the NMF-based algorithms using latent factors and the NSE. So, now that we can represent images in the collection and query images in same space, the problem of finding relevant results reduces to the problem of matching images with similar representations. To do so, we employ the dot product as similarity measure, which gives a notion of the extent to which two images share similar components in the latent or semantic space. Finally, results are ranked in decreasing order of similarity and delivered to the user.

5 Experiments and Results

5.1 Experimental Setup

We conducted retrieval experiments under the query-by-example paradigm to evaluate the proposed methods. A set of 100 images were randomly selected as queries from the database of 2,641 images used in this study. The remaining 2,541 images were used as the target collection to find relevant images.

We performed automatic experiments by sending a query to the system and evaluating the relevance of the results. A ranked image in the results list is considered relevant if it shares at least one keyword with the query. For this experiment, the evaluation was done using traditional measures of Image Retrieval, including Mean Average Precision (MAP) and the Recall-Precision plots.

Image Features We build a bag-of-features representation for the set of histological images, as it has been found to be an effective representation for microscopy image analysis [4,3]. We start by extracting patches of 8x8 pixels from a set of training images with an overlap of 4 pixels along the x and y axes. The DCT (Discrete Cosine Transform) transform is applied in each of the 3 RGB channels to extract the largest 21 coefficients and their associated functions. A k-means clustering is applied to build a dictionary of 500 visual terms. This bag-of-features configuration have shown good results with similar types of histology images [4,3].

Once the vocabulary has been built, every image in the collection goes through the patch extraction process. Each patch from an image is linked to one visual term in the dictionary using a nearest neighbor criterion. Finally, the histogram of frequencies is constructed for each image. We experimentally found that 500 visual terms was enough to achieve a good performance, and that larger dictionaries do not provide significant improvements, but just more computational load.

Text annotations In this data set of histology images, text annotations are clean and clearly defined terms from a technical vocabulary. Since the annotation process followed a systematic revision, there is no need to build a vector space model or to account for term frequencies. We build semantic vectors following a boolean approach, assigning 1 to the terms attached to an image and 0 otherwise. This leads to 46-dimensional binary vectors, which serve to build the text matrix.

5.2 Experiments

Visual Search As a first experiment we retrieved histology images using only visual information as a baseline to assess improvements of other methods. The visual descriptors are based on the bag-of-features strategy, so images are represented by histograms of the occurrence of visual patterns in a dictionary. Direct visual matching is done by calculating the level of similarity between images using the histogram intersection similarity measure [12], as follows:

$$K_{HI}(x,y) = \sum_{i=1}^{n} min\left\{x_i, y_i\right\} \quad . \tag{8}$$

where x and y are images and x_i, y_i are the i-th occurrence of the visual feature in these images, respectively. Using direct matching only, the system achieves a performance up to 0.210 in terms of MAP. An additional experiment using visual features was conducted to determine if latent factors learned only from visual data can help to improve over this baseline.

We applied an NMF decomposition on the visual matrix X_v, using different numbers of latent factors. In the best performing case, this strategy reaches a value of 0.172 of MAP. This suggest that the dimensionality reduction made by NMF leads to a loss of discriminative power of visual descriptors instead of helping to identify semantic patterns in the collection.

Multimodal Search Multimodal search aims to introduce text information during the search process, even though in our case queries are expressed using no keywords but example images. To this end, we employ three different algorithms based on NMF: NMF-mixed, NMF-asymmetric and NSE, and compare their performance against visual search. For the first algorithm (NMF-mixed) the construction of a multimodal matrix was done by setting α=0.5 to give the same importance to visual and text data.

For the NMF-mixed and NMF-asymmetric algorithms we performed several experiments using different sizes of the latent semantic space, to experimentally determine an appropriate number of latent factors. In contrast, the NSE algorithm does not need to set this parameter. Figure 2 shows the result of exploring the number of latent factors for these algorithms, including reference lines for the visual search and NSE. It can be seen that the response of NSE outperforms by a large margin the response of all other strategies. In addition, the Figure shows that besides NSE, only the NMF-asymmetric strategy is able to improve over the visual baseline.

The loss in performance of using NMF-mixed and NMF-visual can be explained by the difficulty of these strategies to find meaningful latent factors from the given input data. As was mentioned before, the NMF-visual fails to find semantic patterns in the collection leading to a decreasing of the discriminatory power of the full visual representation. The NMF-mixed shows a better achievement in this setting, which improves over NMF-visual due to the presence of text terms in the multimodal matrix. Still, the semantic patterns are not correctly modeled by multimodal factors because they have to deal with the reconstruction of visual features as well.

The two strategies that improve over the direct visual matching are NSE and NMF-asymmetric, which concentrate on exploiting text information as the semantic reference data. The NMF-asymmetric, in particular, decomposes the text matrix in an attempt to find meaningful relationships between text terms to build semantic latent factors. In our experiments, we decompose the matrix of 46 terms in 10, 20, 30 and 40 latent factors, from which the second choice presented the best performance. However, the improvement of NMF-asymmetric is modest with respect to NSE, indicating that latent factors are not a fundamental modelling aspect for this dataset.

Another consideration of the modest improvement of NMF-asymmetric may be found in the two steps algorithm. When the first decomposition is done, an approximation error is generated since perfect reconstruction is not required. Then, the second decomposition builds on top of it to construct latent factors for visual features, introducing its own approximation error as well. The NSE algorithm simplifies the approach by learning a unique matrix that correlates visual and text data directly.

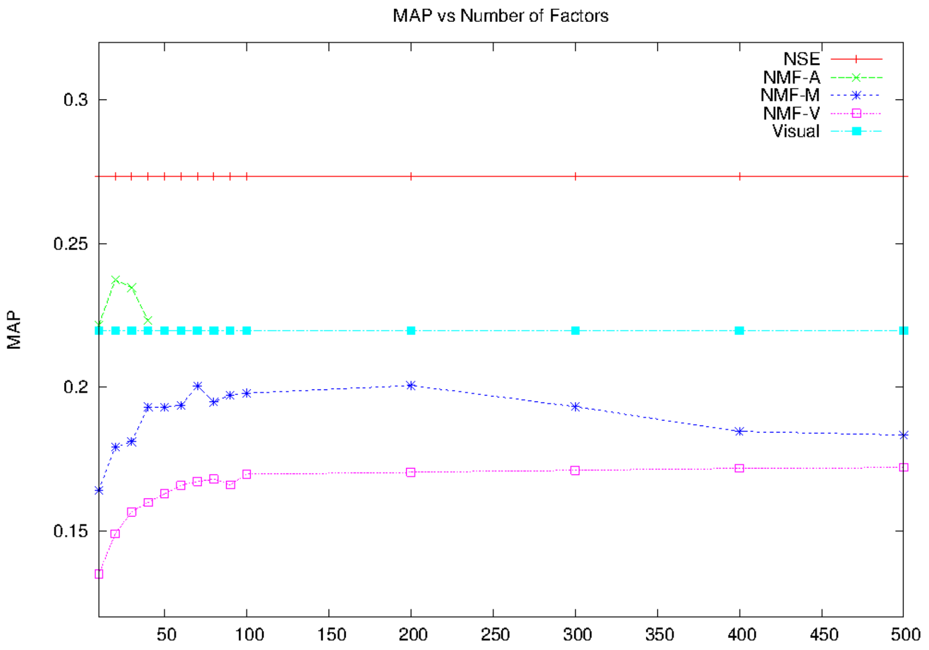


Fig. 2. Latent factors vs. MAP

Another way to measure the performance of the evaluated algorithms is using the Recall vs. Precision graph. For this evaluation, we selected the number of latent factors that provided the best performance in the previous evaluation. We plot the interpolated Recall-Precision graph, in which one can observe the differences in precision along the retrieval process, i.e., while all relevant images are being retrieved to the user. Figure 3 shows the result of this evaluation, and reveals that the direct visual matching presents very good results in early precision, but also falls very fast as long as more relevant images are required.

The second best performance in the first positions of the results page is given by the NMF-visual approach, according to the Recall-Precision graph. What it actually means, is that nearest neighbors under a visual similarity measure are very likely to be relevant in a histology image collection. This contrast with all other semantic indexing approaches (NMF-M, NMF-A and NSE) which are modeling structural patterns in the whole collection, rather than exploiting the local similarities of the dataset. Nevertheless, the performance of NSE and NMF-asymmetric showed a more consistent and sustained higher precision as long as the user explores more images in the results page.

The good performance obtained in early position with the direct visual matching method are due to the fact that in this collection several histology images come from sections that belong to the same histological plate or represent the same region with different zoom levels. Therefore, in this case, visual relationships directly denote semantic relations. But this behavior only occurs in the first retrieved images. In the Recall-Precision graph shows that values of recall greater than 10% (that in this collection represents an average of 20 images) the accuracy of the visual methods falls dramatically, while the NSE shows a more stable behavior.

Table 1 summarizes the findings of our experimental results, presenting the number of required latent factors, MAP measures, relative improvement in terms of MAP w.r.t. the visual baseline and the precision at the first 10 results (P@10). The Table shows that NMF-asymmetric and NSE provide a significant improvement in terms of MAP compared with only visual retrieval. It also confirms the dominant position of visual search in terms of early precision.

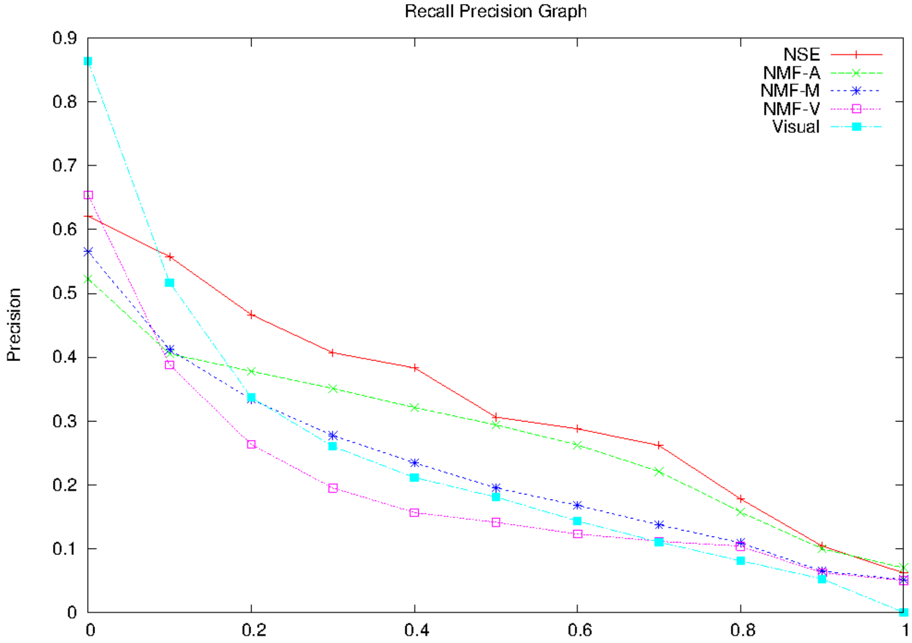


Fig. 3. Recall vs. Precision Graph

6 Conclusions

We presented a method for image indexing that combines visual and text information using a variation of non-negative matrix factorization. Text annotations provide a semantic representation space where the visual content of images is embedded using NMF.

The experimental evaluation demonstrates an increase in retrieval performance. The effectiveness of this method may be explained by the fact that it efficiently exploits the semantic information contained in text annotations. But we must take into account that this methodology is mainly applicable in cases where we have a controlled annotation process.

An important characteristic of the proposed method is that it naturally deals with different types of data: non-annotated images, images with multiple annotations, text queries, etc. This is accomplished by mapping both images and queries to a common semantic representation space. Another important advantage is that the method could be efficiently implemented by integrating the last developments on on-line matrix factorization (such as the method proposed by Mairal et al. [8]) which can deal with large amounts of data.

Table 1. Performance measures for all evaluated strategies

Method	Latent Factors	MAP	Improvement	P@10
Visual Matching	N/A	0.210	N/A	0.650
NMF-Visual	300	0.172	-18.8%	0.088
NMF-Mixed	70	0.201	-4.3%	0.252
NMF-Asymmetric	20	0.235	+11.9%	0.208
NSE	N/A	0.273	+30.0%	0.265

References

1. Juan C. Caicedo, Angel Cruz, and Fabio A. Gonzalez. Histopathology image classification using bag of features and kernel functions. In *Proceedings of the 12th Conference on Artificial Intelligence in Medicine: Artificial Intelligence in Medicine*, AIME '09, pages 126–135, Berlin, Heidelberg, 2009. Springer-Verlag.
2. Juan C. Caicedo, Fabio A. Gonzalez, Edwin Triana, and Eduardo Romero. Design of a Medical Image Database with Content-Based Retrieval Capabilities. *Advances in Image and Video Technology*, LNCS 4872:919–931, 2007.
3. Angel Cruz-Roa, Juan C. Caicedo, and Fabio A. Gonzalez. Visual pattern analysis in histopathology images using bag of features. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5856 LNCS:521–528, 2009. cited By (since 1996) 1; Conference of 14th Iberoamerican Conference on Pattern Recognition, CIARP 2009; Conference Date: 15 November 2009 through 18 November 2009; Conference Code: 83218.

4. Gloria Díaz and Eduardo Romero. Histopathological image classification using stain component features on a plsa model. In *Proceedings of the 15th Iberoamerican congress conference on Progress in pattern recognition, image analysis, computer vision, and applications*, CIARP'10, pages 55–62, Berlin, Heidelberg, 2010. Springer-Verlag.
5. Fabio A González, Juan C Caicedo, Olfa Nasraoui, and Jaafar Ben-Abdallah. NMF-based multimodal image indexing for querying by visual example. In *ACM International Conference On Image And Video Retrieval CIVR2010*, pages 366–373. ACM Press, 2010.
6. D D Lee and H S Seung. New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation. *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, 13(1):V–621–V–624, 2001.
7. W Liu, Nanning Zheng, and Xiaofeng Lu. Non-negative matrix factorization for visual coding. In *2003 IEEE International Conference on Acoustics Speech and Signal Processing 2003 Proceedings ICASSP 03*, volume 3, pages III–293–6. Ieee, 2003.
8. Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, March 2010.
9. Henning Müller, Paul Clough, Thomas Deselaers, and Barbara Caputo. *ImageCLEF: Experimental Evaluation in Visual Information Retrieval.* Springer, 2010.
10. Lucia Roa-Peña, Francisco Gómez, and Eduardo Romero. An experimental study of pathologist's navigation patterns in virtual microscopy. *Diagnostic pathology*, 5(1):71, January 2010.
11. A W M Smeulders, M Worring, S Santini, A Gupta, and R Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
12. Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991. 10.1007/BF00130487.
13. F. Yu and H. Ip. Semantic content analysis and annotation of histological images. *Computers in Biology and Medicine*, 38(6):635–649, June 2008.
14. Lei Zheng, A. W. Wetzel, J. Gilbertson, and M. J. Becich. Design and analysis of a content-based pathology image retrieval system. *Information Technology in Biomedicine, IEEE Transactions on*, 7(4):249–255, 2003.

Appendix B

Histology Image Indexing Combining Visual And Semantic Features

Vanegas J.A., Caicedo J.C., González F.A., Histology Image Indexing Combining Visual And Semantic Features. 7th International Seminar on Medical Information Processing and Analysis. SIPAIM 2011.

Histology Image Indexing Combining Visual And Semantic Features

Jorge A. Vanegas[a], Juan C. Caicedo[a], Fabio A. González[a]

[a]*BioIngenium Research Group, Universidad Nacional de Colombia, Bogotá, Colombia.*

Abstract

This paper presents a histology image indexing strategy for content-based image retrieval that takes advantage of multimodal information, i.e., visual data extracted from images and related text from annotations provided by experts. The proposed strategy combines the two modalities to build an improved, semantic-aware visual representation. The proposed strategy was tested on a set of histology images with annotations identifying biological systems and organs. The experimental evaluation shows and improvement on mean average precision of 80.5 % with respect to a strategy that exclusively uses the visual information, and an improvement of 38.8 % with respect to a previously proposed multimodal indexing strategy.

Keywords: Multimodal information, histology images, image indexing, information retrieval

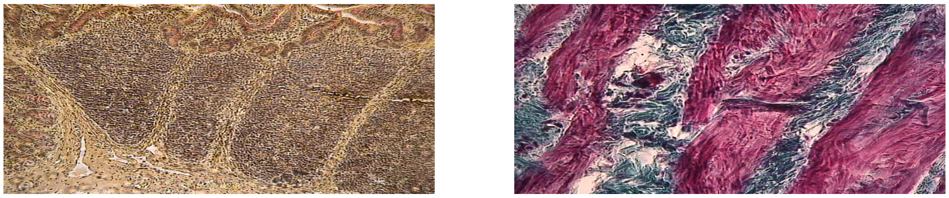
1. Introduction

Large collections of biomedical images are becoming a very valuable source of knowledge that may be helpful for research, education and for medical decision support. The problem is to find a good strategy that allows us to access these repositories in an efficient and effective way. Access to histological image repositories using a query-by-example (QBE) approach based on low-level visual features has been investigated in different works [1, 2]. Subsequent works [3, 4] showed that the semantic analysis of histology images can be used to improve the response of an image retrieval system. However, the semantic representation implies a dimensionality reduction, which may involve reduction in discrimination capacity during the information retrieval process. In this paper we propose a strategy to combine visual and semantic representation in order to overcome this problem.

This work presents two main contributions: first, we propose a new method for multimodal image visual content indexing enriched by the semantic information present in annotations, second, we propose a method to smoothly combine visual indexing and multimodal indexing to better exploit the advantages of each approach. The rest of the paper is organized as follows: Section 2 presents the structure of the histology image collection used in this work; Section 3 presents the basic concepts of semantic representations using matrix factorization; Section 4 presents the proposed method to combine the visual and sematic information; Section 5 presents the experimental evaluation; and, finally, Section 6 presents some concluding remarks.

2. Histology Images

The collection of histology images of this study come from an atlas of histology composed of 20,000 images of tissues from different biological systems and organs (accessible via Internet at www.informed.unal.edu.co). From this large collection, we selected for this work a subset of 2,641 images, that contains annotations for each image, indicating the biological system and observed organs, resulting in a total of 46 different categories.



Category: Lymphatic system. Lymphatic structure of the digestive tract.

Category: Digestive system. Appendix.

Category: Female reproductive system. Ovarian.

Figure 1: Example of histology images with associated terms

3. Matrix Factorization for Semantic Representation

Using a bag-of-words model for text and visual features [5], we can represent an image by two matrices: X_v and X_t. where $X_v \in \mathbb{R}^{n \times l}$ is the matrix of visual word frequencies, n is the number of visual words, l is the total number of images in the collection, $X_t \in \mathbb{R}^{m \times l}$ is the matrix of text term frequencies and m is the number of terms or keywords. The goal of semantic representation, is to discover the underlying structure of these matrices to build an effective index for image search. The semantic representation method called Non-negative Semantic Embedding that we use as one of the baselines in this paper, is based on the Non-negative Matrix Factorization (NMF) algorithm[6].

3.1. Non-negative Matrix Factorization

The main purpose of NMF is to find an approximation of a matrix X, in terms of two smaller matrices with the restriction of only allowing non-negative values, as follows:

$$X \approx WH \ . \tag{1}$$

where $X \in \mathbb{R}^{p \times l}$, $W \in \mathbb{R}^{p \times r}$, $H \in \mathbb{R}^{r \times l}$, p is the total number of available features, l is the number of images in the collection, and r is the rank of the decomposition. The matrix W is known as the basis matrix and H is known as the encoding matrix.

The factorization is found by applying two multiplicative updating rules for W and H following the Lee and Seung's approach [7] and by solving the associated optimization problem that corresponds to minimizing the squared error of the decomposition or the Kullback-Leibler divergence between the original matrix and the reconstructed one.

3.2. Non-negative Semantic Embedding

In [8] a method for image indexing is presented, called Non-negative Semantic Embedding (NSE), that assumes that the space spanned by text terms is a good enough representation of image semantics, and uses it to represent the images. In this approach the encoding matrix is directly the matrix of text annotations:

$$X_v = WX_t; W \geq 0 \ . \tag{2}$$

where $W \in \mathbb{R}^{n \times m}$, n the number of visual features and m the number of text terms. By solving W, we find a linear transformation that allows us to represent visual information in the semantic space given by the text annotations. The matrix W is obtained by running the multiplicative update only for W.

4. Combining Visual And Semantic Features

Using a semantic representation it is possible to find relationships between images that have a high visual variability, providing an improvement in the capacity of retrieval to get a greater amount of relevant results. The NSE method is used in [8] obtaining a general increase in retrieval performance, but is surpassed in precision by the simple direct visual matching in the first results, where the images have a greatest visual similarity. In this way, NSE is useful in the situation where we are interested in finding a large amount of relevant results. But, if we are interested in retrieving a few results with a high precision it is better to use direct visual matching.

In order to improve the capacity of retrieving a greater amount of relevant results without losing precision in the first results, we propose a strategy that represents the images by combining visual and semantic features. The new image representation can be described by the following equation:

$$X'_v = \lambda X_v + (1 - \lambda) W X_t \tag{3}$$

where X_v is the original visual feature matrix, X_t is the original text features matrix, W is the basis matrix of the semantic space calculated in the training phase, WX_t is the reconstructed visual matrix based on the semantic representation and $\lambda \in [0, 1]$ is a weighting parameter that controls the relative importance of the two representations.

4.1. Image Indexing and Search

When a new image y_v, without text annotations, arrives, either to be included in the collection or to be used as query, it is represented using the following equation:

$$y'_v = \beta y_v + (1 - \beta) W h \tag{4}$$

where $\beta \in [0, 1]$ is a weighting parameter, W is the basis matrix of the semantic space calculated in the training phase, and h corresponds to the representation of the image in the semantic space, which is found solving the equation: $y_v = Wh$.

5. Experiments and Results

5.1. Experimental Setup

In order to evaluate the proposed strategy, we performed retrieval experiments using the query-by-example paradigm. For this study we used a set of 2641 histology images. Of this group 100 images were randomly selected as queries. The remaining 2541 images were used as the target collection to find relevant images.

5.1.1. Visual Representation

The images are represented using the bag-of-features approach, as it has been found to be an effective representation for microscopy image analysis [9, 5]. We extracted blocks of 8x8 pixels from a set of training images with an overlap of 4 pixels along the x and y axes. The DCT (Discrete Cosine Transform) is applied in each of the 3 RGB channels and are used the largest 21 coefficients. A k-means clustering is applied to build a dictionary of 500 visual terms. This bag-of-features configuration have shown a good performance with similar types of histology images [9, 5].

5.1.2. Text annotations

Text annotations in this dataset are clean and clearly defined terms from a technical vocabulary. Each image is represented by a semantic vector following a boolean approach, assigning 1 to the terms attached to the image and 0 otherwise. This leads to 46-dimensional binary vectors, which serve to build the text matrix.

5.1.3. Similarity Measure

For the similarity measure, we used the histogram intersection similarity [10], which has proven to be a very suitable measure for comparing images in a bag-of-words representation:

$$K_{HI}(x,y) = \sum_{i=1}^{n} min\left\{x_i, y_i\right\} \tag{5}$$

where x and y are images and x_i , y_i are the i-th occurrence of the visual feature in these images.

5.2. Experiments

In order to find the better configuration for the combination of visual and semantic features, we performed automatic experiments by sending a query to the system and evaluating the relevance of the results with different values of λ and β. To evaluate the performance, we used two measures: Precision at 10 and MAP (Main Averange Precision), Figure 2 shows the results of these two measures for different configurations of λ and β. The points over the gray line belong to the Pareto frontier, i.e. the points over the

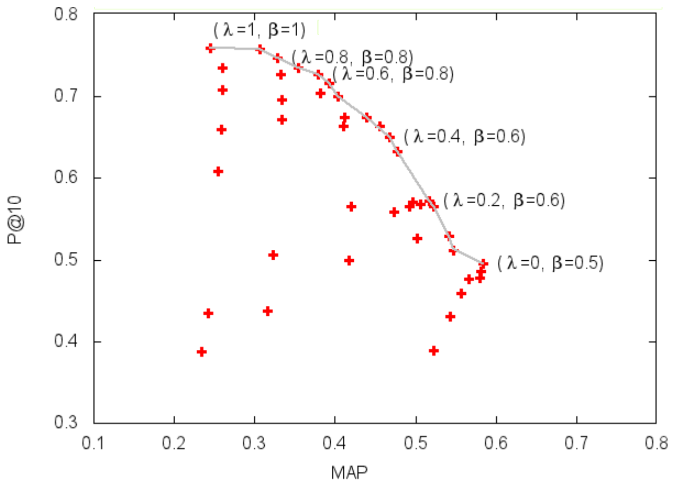


Figure 2: Pareto frontier

gray line are optimal solutions of this multiobjective optimization problem. Figure 2 shows the trade-off between the MAP performance and precision at 10, this result is expected according to [8], where is shown that semantic representation improves the capacity of recovering a greater amount of relevant results, achieving an increase in the value of MAP, but causes a reduction of precision in the first results. In the visual representation, the behavior is opposite, we obtained a good performance in early position, but the accuracy falls dramatically when we recover a greater amount of images.

An important result is that, with $\lambda = \beta = 0$, i.e. a purely semantic representation, we get a value of 0.47 of precision and a value of 0.579 of MAP, both are considerably higher than the reported values in [8] using NSE, this suggests that the transformation in the representation and the change in the method of similarity measure (in [8] the dot product is used to calculate the similarity in the semantic space) allows us to use the semantic information in a more appropriate way. This initial gain in the performance caused by the transformation in the representation, allows us to compromise some accuracy in order to get an improvement in the value of MAP.

According to the preliminary results, a good configuration is obtained with $\beta = 0.8$, and λ between 0.6 and 0.8, where we could get an improved in MAP compared to the direct visual matching without a considerable loss of precision. The performance of this configuration is evaluated in conjunction with the NSE method and the direct visual matching using the Recall vs. Precision graph (figure 3), where is showed that the performance of the

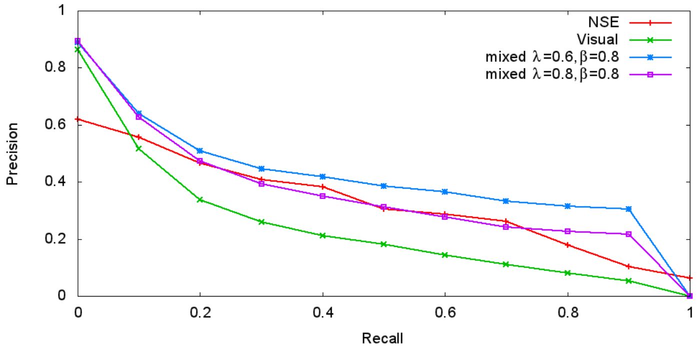


Figure 3: Recall vs. Precision Graph

Table 1: Performance measures for Visual Matching, NSE and Mixed Strategy

Method	P@10	MAP	Improvement
Visual Matching	0.650	0.210	N/A
NSE	0.273	0.265	+30.0%
Mixed $\lambda = 0.6$, $\beta = 0.8$	0.720	0.379	+80.48%

mixed strategy usig $\beta = 0.8$, and $\lambda = 0.6$ improve over the direct visual matching and NSE. Table 1 summarizes the findings of our experimental results.

6. Conclusions

We presented a strategy for image indexing that combines visual and semantic features. The proposed mixed strategy is able to take advantage of the benefits of each kind of representation. The experimental evaluation demonstrated that with the appropriate combination of visual and semantic information we obtained a significant increase in the general performance during the information retrieval process, reaching improvements over the direct visual matching and the NSE strategies.

Acknowledgments

This work was partially funded by the project Anotación Automática y Recuperación por Contenido de Imágenes Radiolgícas usando Semántica Latente number 110152128803 and project Sistema para la Recuperación de Imágenes Médicas utilizando Indexación Multimodal number 110152128767 by Convocatoria Colciencias 521 de 2010.

References

[1] L. Zheng, A. W. Wetzel, J. Gilbertson, M. J. Becich, Design and analysis of a content-based pathology image retrieval system, Information Technology in Biomedicine, IEEE Transactions on 7 (4) (2003) 249–255.

[2] J. C. Caicedo, F. A. Gonzalez, E. Triana, E. Romero, Design of a Medical Image Database with Content-Based Retrieval Capabilities, Advances in Image and Video Technology LNCS 4872 (2007) 919–931.

[3] F. Yu, H. Ip, Semantic content analysis and annotation of histological images, Computers in Biology and Medicine 38 (6) (2008) 635–649.

[4] J. C. Caicedo, A. Cruz, F. A. Gonzalez, Histopathology image classification using bag of features and kernel functions, in: Proceedings of the 12th Conference on Artificial Intelligence in Medicine: Artificial Intelligence in Medicine, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 126–135.

[5] A. Cruz-Roa, J. C. Caicedo, F. A. Gonzalez, Visual pattern analysis in histopathology images using bag of features, Iberoamerican Conference on Pattern Recognition LNCS 5856 (2009) 521–528.

[6] W. Liu, N. Zheng, X. Lu, Non-negative matrix factorization for visual coding, in: 2003 IEEE International Conference on Acoustics Speech and Signal Processing 2003 Proceedings, Vol. 3, Ieee, 2003, pp. III–293–6.

[7] D. D. Lee, H. S. Seung, New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation, 2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings 13 (1) (2001) V–621–V–624. arXiv:1010.1763.

[8] J. A. Vanegas, J. C. Caicedo, F. Gonzalez, E. Romero, Histology image indexing using a non-negative semantic embedding (2011).

[9] G. Díaz, E. Romero, Histopathological image classification using stain component features on a plsa model, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications LNCS 6419 (2010) 55–62.

[10] M. J. Swain, D. H. Ballard, Color indexing, International Journal of Computer Vision 7 (1991) 11–32.

Appendix C

Scalable Construction of a Bag of Features Representation Using the Map-Reduce Architecture

Vanegas, J.A., Caicedo J.C, & González F., Scalable Construction of a Bag of Features Representation Using the Map-Reduce Architecture. The Latin American Conference on High Performance Computing. CLCAR 2012.

Indexing Images with Bag-of-Features on a Map-Reduce Architecture

Jorge A. Vanegas
Universidad Nacional de Colombia
javanegasr@unal.edu.co

Juan C. Caicedo
Universidad Nacional de Colombia
jccaicedoru@unal.edu.co

Fabio A. Gonzalez
Universidad Nacional de Colombia
fagonzalezo@unal.edu.co

Abstract

Image indexing and search is becoming increasingly important in specialized fields where visual contents support the decision making process. Building a content-based index for image search poses computational challenges when the target collection is very large. This paper describes a methodology to compute the bag-of-features descriptor for indexing a medical image collection in a visual search engine. The proposed methodology is oriented to run in the Map-Reduce architecture, for achieving a proper computational load distribution that makes feasible the processing of large image collections. Experiments were conducted with a histology image dataset, to evaluate scalability and the quality of retrieval performance. Results show important speedups for computing image descriptors and no reduction in the quality of the representation, since the employed parallelization strategy does not have any approximation or simplification of the basic algorithm.

1 Introduction

The image acquisition facilities in recent years have increased the popularity of large repositories of this kind of data in different contexts such as personal photo collections, scientific data and medical imaging. In medical and clinical applications, these image collections have shown to be a valuable source of knowledge, taking an important role in education and research. In order to provide access to this collections of images, the design of search engines able to match visual contents is an active topic of research [5].

One of the problems of modern image collections is that their size is very large that it becomes difficult to process and work with them, using classic management tools. For instance, modern hospitals can produce about 120,000 images per day; this is about 100 GB of data [10] that should be indexed in a daily basis. Currently, we can generate huge repositories of valuable data but we are not able to process and manage them appropriately. So, the challenge is the development of tools for an effective and efficient access to this information.

A fundamental requirement for content-based image indexing systems is a good representation of visual contents. The bag-of-features is a descriptor that has become very popular and has been used actively in retrieval and annotation tasks, showing very good results [1]. The problem is that computing this descriptor may result in a high computational cost, which makes it unfeasible to use on databases with hundreds of thousands of images. Then, the challenge is to implement this model in a distributed architecture that allows for a scalable solution as more resources are added to the computational infrastructure.

This paper proposes a parallelization strategy based on the Map-Reduce Framework which has proven to be an extensible solution to solve the typical problems of working with large data sets. The goal of the proposed approach is to compute the representation for a collection of images by harnessing computing resources in a cluster of machines dedicated to index images. White et al. [15] presented two strategies to compute the bag-of-features for web-scale image collections. However, no experimental evaluation was conducted by them neither to observe speedups nor to assess the quality of image retrieval. In this work, we present algorithmic details as well as comprehensive experiments to estimate the applicability of this approach to practical image search systems.

The structure of this paper is as follows: Section 2 presents details of the bag-of-features model. Section 3 presents a brief introduction about Map-Reduce framework. Section 4 defines details about the parallelization strategy for bag-of-features algorithm and its implementation in Map-Reduce. Section 5 presents experimental results and Section 6 presents some concluding remarks.

2 Bag-of-Features Model

The bag-of-features model is an adaptation of the bag-of-words used for text categorization and text retrieval [7]. In this kind of representation the sequences of words are ignored and the number of occurrences of words in a document is the only information taken into account. The bag-of-features model is based on the idea that an image can be treated as a document, which consists of a set of words, which are visual characteristics that can be extracted from an image subregion.

The final representation of the image is a histogram that determines the number of times that each visual word was found in an image. So, the image is represented in a simplified way by a counting independent features, and the spatial relationships between these features are not taken into account. To build a bag-of-features representation three main steps are required: 1) Feature extraction, 2) Learn visual vocabulary and 3) Quantization of visual words for histogram construction [4]. Figure 1 shows an overview of the three steps.

Feature Extraction

The first step seeks to extract a set of local features within an image, i.e., to describe a set of small subregions or blocks from an image. This step presents an initial process consisting in the detection or selection of these blocks. The way to choose these blocks may vary. An alternative is to simply use a regular grid [8, 3], dividing the image into several patches of equal size. In this way all regions of the image are processed. Another alternative is to consider only points of interest [11]. Once the subregions are selected, some kind of descriptor to quantify the features is applied to them. The kind of descriptor to use in these blocks can be defined for convenience. For instance the raw block or more elaborate descriptors such as SIFT (Scale-Invariant Feature Transform) [9] or DCT (Discrete Cosine Transform) [14] could be used. In any case the result of this stage is a set of feature vectors.

Learn Visual Vocabulary

The purpose of this second step is to define our visual vocabulary or codebook as a set of K visual words. This seeks to group the wide range of visual features obtained in the initial stage and reduce them to a representative set. For this purpose the K-means clustering technique is usually employed, in which the final result is a set of centroids that define the visual dictionary.

Histogram Construction

Having defined our visual dictionary, the last stage is to translate all visual features obtained in the first step to visual words, this is done by calculating the distance of each feature vector to each centroid and assign it to the nearest one. Once the translation is done, a counting of how many times a visual word is found in each image is performed, thereby generating a histogram of K visual words.

The K-means clustering is simple but it has high time complexity when the data sets are large. In these circumstances the memory of a single machine can be a restriction.

Performance Considerations

This kind of representation has been proved succesful for different computer visions tasks. However, the representation construction process requires a large computational cost. First, considering the initial step, a high amount of disk space is needed due to the intermediate representation for a large number of patches or image subregions. In the second stage, learning the visual vocabulary demands large amounts of memory, since it is necessary to load and process a large sample of these feature vectors and compare them with each centroid, and furthermore requires an excessive amount of computing time because the K-means algorithm needs numerous iterations until the algorithm converges. All this requirements make this model not a feasible choice for collections with hundreds of thousands of images.

To overcome this drawback, we consider a distributed environment to extend the model for use in a architecture that achieves an efficient split of the computational load. For this purpose, we adopt the Map-Reduce architecture and reformulate the bag-of-features method to fit it naturally within the framework.

3 Map-Reduce Framework

Map-Reduce is a framework proposed by Google to support distributed data processing. In simple terms, a program based on this architecture must have defined two functions: Map and Reduce. The Map function is applied in parallel to all input items, and Reduce function usually seeks to combine the results of processes performed in parallel by the MAP function. Both functions are defined with a structured data in key-value pairs. Map function takes a pair of data in one defined domain, and returns a list of pairs belonging to other domain: $Map(k1, v1) \rightarrow list(k2, v2)$.

The framework collects the pairs returned by all Maps and generate a group composed of pairs with the same key. The Reduce function is then applied to each group, which in turn produces a collection of values in the same domain: $Reduce(k2, list(v2)) \rightarrow list(v3)$.

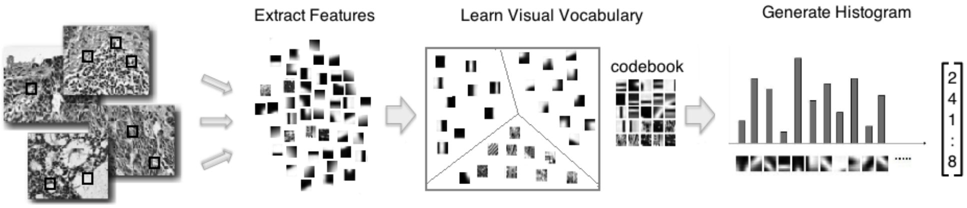


Figure 1. Overview of the bag-of-features representation illustrating the three steps: 1. Local features extracted from a training set, 2. Learning a visual vocabulary using k-means and 3. visual words occurrences are accounted in a histogram.

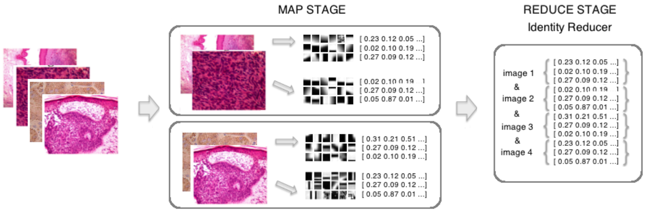


Figure 2. Feature extraction stage, 1. Map: perform the feature extraction for all patches, 2. Reduce: group together all results.

The result of each Reduce call is typically one value or an empty return, though one call is allowed to return more than one value. Finally, the results of all Reduce calls are collected as the desired result list. Thus any algorithm can be implemented in this architecture if it can be defined in certain parts that can be processed independently in parallel and were only interested in combining the final result.

4 Bag-of-Features in a Distributed Architecture

Initially we must note that the three main stages: feature extraction, visual vocabulary learning and histogram construction, should remain sequential to each other, since each stage depends on results obtained in the previous step, but within each stage we can define different parallelization strategies.

4.1 Distributed feature extraction

The objective in this step is to generate a set of feature vectors that describes each patch extracted from each image. The strategy of paralelization is very simple, since each image could be processed independently of the others. This stage can be divided basically into the following steps: 1) Extract a set of patches from each image. 2) Generate a feature vector for each patch.

The map and reduce stages can be defined as follows: ***Map stage***: process a subset of images, i.e., perform the feature extraction process for all patch set from each image. As result of the process a set of feature vectors is returned.

$$\begin{aligned} &Map_{fe}(image\,id, binary\,data) \\ &\rightarrow list(image\,id, feature\,vector) \end{aligned} \tag{1}$$

Reduce stage: In this case Reduce function takes no action in the data and the output list is identical to the input list.

$$\begin{aligned} &Reduce_{fe}(image\,id, list(feature\,vector)) \\ &\rightarrow list(feature\,vector) \end{aligned} \tag{2}$$

Figure 2 shows an overview of complete Map-Reduce cycle for this phase.

4.2 Learn visual vocabulary

For this step a significant sample of all previously extracted feature vectors is used to learn the visual dictionary. The K-means algorithm can be computed in a distributed mode by delegating the distance computations to different workers. This stage can be divided basically into the following steps: 1) Initialize randomly an number k of centroids. 2) For each feature vector, calculate the nearest centroid using euclidian distance and assign the feature vector to this cluster (or visual word). 3) Re-compute centroids (mean of feature vectors in cluster).4) Stop when there are no new re-assignments.

The map and reduce stages can be defined as follows:

Map stage: calculates the distances between samples and centroids; matches samples with the nearest centroid and assigns them to that specific cluster. Each map task is processed with a data block.

$$\begin{aligned} &Map_{lv}(visual\,word, featrure\,vector) \\ &\rightarrow list(visual\,word, featrure\,vector) \end{aligned} \tag{3}$$

Reduce stage: recalculates the centroid point using the average of the coordinates of all the points in that cluster. The associated points are averaged out to produce the new location of the centroid. The centroids configuration is given as feedback into the Mappers.

$$\begin{gathered} Reduce_{lv}(visual\,word, list(featrure\,vector)) \\ \rightarrow cluster = avg(featrure\,vector) \end{gathered} \tag{4}$$

This Map-Reduce cycle is repeated until the centroids converge. The final centroids will be the representative vectors of each visual word.

4.3 Quantize features using visual vocabulary

In this final step the final image representation is generated completing the following steps: 1) Translation of each feature vector to a visual word, using the previously learned codebook. 2) Calculate the histogram. The map and reduce stages can be defined as follows:

Map stage: all features vector are compared with the learned visual codebook, and again is used the euclidian distance as proximity criterion to find the closer cluster, which will be assigned.

$$\begin{gathered} Map_{qf}(image\,id, featrure\,vector) \\ \rightarrow list(visual\,word, 1) \end{gathered} \tag{5}$$

Reduce stage: counting is performed for all visual words found in Map stage.

$$\begin{gathered} Reduce_{qf}(visual\,word, list(visual\,word\,count)) \\ \rightarrow visual\,word\,count = \Sigma(visual\,word\,count) \end{gathered} \tag{6}$$

The result is a non-normalized histogram. Figure 3 shows an overview of complete Map-Reduce cycle for this phase.

Finally, the distributed bag-of-features algorithm uses $2 + n$ Map-Reduce cycles, where n is the number of cycles required until k-means algorithm converges.

4.4 Implementation

The distributed bag-of-features algorithm was written in Java programming language, and Implemented under Apache Hadoop framework, which is an open source implementation of Map-Reduce architecture. The k-means implementation of the Apache Mahout library [6] was used, which also provides various basic machine learning algorithms, that runs over a Hadoop system.

This framework has been configured in pseudo-distributed mode, allowing to run multiple instances in parallel on one machine, allowing to take advantage of multi-core architecture.

5 Experiments

5.1 Data set

In this work we used a set of 2,641 histology images. This selection of images comes from a collection of different biological systems and organs, which were stained using Hematoxiline & Eosine and Immunohistochemical techniques. The collection includes photographs of histology slides acquired with a digital camera coupled to a microscope, using different magnification factors to focus important biological structures. The main use of this collection is for educational and research purposes, and is accessible via Internet at http://www.informed.unal.edu.co. These images were classified in 46 categories defined by the biological system and organ shown. In [13] the bag-of-features model is evaluated in its sequential form, as a descriptor for this set of images.

5.2 Experimental Setup

In order to evaluate the proposed method, we conducted retrieval experiments under the query-by-example paradigm. A set of 100 images were randomly selected as queries from the database of 2,641 images used in this study. The remaining 2,541 images were used as the target collection to find relevant images.

We performed automatic retrieval experiments by sending a query to the system and evaluating the relevance of the results. A ranked image in the results list is considered relevant if it shares at least one keyword with the query. For this experiment, the evaluation was done using traditional image retrieval performance measures, including mean average precision (MAP) and early precision measures.

To have a basis for comparison, we defined a similar configuration to that used in [13] where a sequential bag-of-features was employed. The regular grid for feature extraction using patches of 8x8 pixels is used with an overlap of 4 pixels along the x and y axes. Each patch is processed using DCT (Discrete Cosine Transform) and is represented by the 21 largest coefficients.

In order to verify the obtained improvement using parallel processing, we run the proposed distributed bag-of-features algorithm, with a different number of processor cores. From a single core to 10 cores.

5.3 Experiment environment

The Map-Reduce framework was configured in pseudo-distributed mode, i.e., a single node, with a Intel(R) Xeon(R) CPU at 2.40GHz with 12 processor cores, and 32 GB of ram memory, with Apache Hadoop version 0.20.203 as Map-Reduce implementation installed over a Linux Ubuntu 10.04.

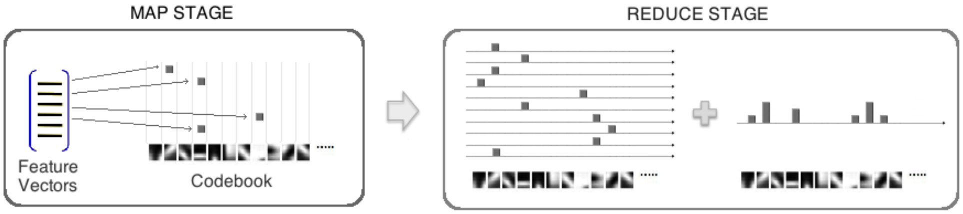


Figure 3. Histogram Construction stage, 1. Map: perform the translation from feature vector to a visual word, 2. Reduce: count the number of occurrences of each visual word.

Table 1. Required computing time in a multicore infrastructure

Number of cores	Feature extraction		Dictionary learning		Histogram representation		Totals	
	Time	Speedup	Time	Speedup	Time	Speedup	Time (minutes)	Speedup
1 core	66	–	344.5	–	90	–	**500.5**	-
2 cores	39	1.69	175.5	1.96	44	2.04	**258.5**	**1.93**
4 cores	31	2.12	110.5	3.12	31	2.90	**172.5**	**2.90**
10 cores	16	4.13	78	4.42	28	3.21	**122**	**4.10**

5.3.1 Visual Search

We performed the retrieval experiment using direct visual matching, calculating the level of similarity between images using the histogram intersection similarity measure [12], as follows:

$$K_{HI}(x, y) = \sum_{i=1}^{n} min\{x_i, y_i\} \quad . \tag{7}$$

where x and y are images and x_i, y_i indicate the frequency of the i-th visual feature in each image respectively.

5.4 Results

Table 1 shows the speed up obtained by increasing the number of cores. Processing time was calculated for each step. The results show a similar saving of time in each step. As the required time in the dictionary learning stage, depends on the required number of cycles until k-means algorithm converges, we calculated the average time spent in each cycle and we have calculated the total time spent if 52 iterations were required which is the amount used for a single core case. Table 1 shows that with the addition of an extra core, we can reduce by half the processing time, and may even further reduce processing time by adding more cores although the gain ratio is reduced.

The observed speedups in the last column of Table 1 were plotted and compared with the theoretical speedups of the Amdahl's law. This is observed in Figure 4, which indicates that in practice we can achieve 85% of parallelization. Theoretically, the algorithm to compute a bag-of-features is almost 100% parallelizable, so there is a 15% of loss in practice when running the algorithm on this architecture. This loss can be observed in all three stages of the algorithm, so it may be explained by communication delays and infrastructure management. We consider that this is not a critical loss when the image collection is very large, since on the other hand, we are overcoming memory limitations and other problems associated to managing large amounts of data.

The next evaluation that we conducted in this work is to observe the quality of the search process. Table 2 shows a comparison of retrieval performance between the proposed parallel version and the sequential results reported in [13]. In both cases we conducted 5 retrieval experiments with 100 queries each, resulting on 500 different query evaluations. The reported results are the average of these experiments. According to the results in Table 2 the variation in performance is very small for the repetitions of the experiments and the total number of queries. These variations

Table 2. Performance in retrieval tasks

Performance Metric	**Sequential BoF ***	**Distributed BoF (10 cores)**
MAP	0.244±0.029	0.253±0.020
P@10	0.668±0.018	0.646±0.098
P@20	0.622±0.040	0.592±0.029

* Results reported for Visual Matching in[13]

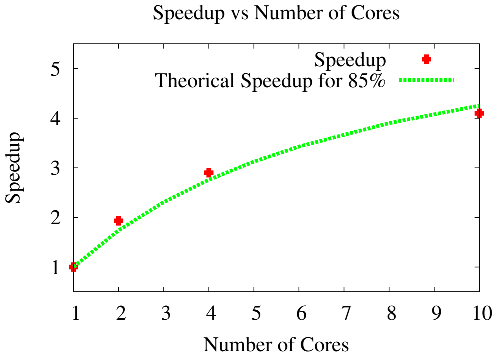


Figure 4. Speed up obtained by increasing the number of cores and Amdahl's law

may be due to the random initialization of the K-means algorithm, or the error rate at the final iteration. Yet, the algorithm converges to a useful solution to construct a bag-of-features representation and preserves the values of performance measures in a similar range.

6 Conclusions

We presented a strategy to implement the bag-of-features method in a distributed architecture following the Map-Reduce Framework. This allows to achieve parallelism by reformulating the algorithm without altering the original algorithm. The experimental evaluation demonstrates important speedups by dividing the computing workload in multiple processing units. This allows to index large collections of images by harnessing the computing infrastructure in a dedicated cluster. Also, the resulting representation has no degradation in the quality of the representation, which translages in the same performance for the underlying image search task.

The step of learning the visual vocabulary can be further improved by considering an online implementation of the K-Means algorithm, which shows a faster convergence than the conventional batch approach [2]. This is left for future research. The proposed algorithm has been implemented and evaluated using the Apache Hadoop framework and we provide access to the source code via Internet at http://code.google.com/p/bioingenium-large-scale-tools.

Acknowledgement

This work was partially funded by the project Anotación Automática y Recuperación por Contenido de Imágenes Radiológicas Usando Semántica Latente, number 110152128803 and project Sistema para la Recuperación de Imágenes Médicas Utilizando Indexación Multimodal, number 110152128767 by Convocatoria Colciencias 521 de 2010.

References

[1] A. Bosch, X. Muñoz, and R. Martí. Which is the best way to organize/classify images by content? *Image and Vision Computing*, 25(6):778–791, June 2007.

[2] L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7*, pages 585–592. MIT Press, 1995.

[3] A. Cruz-Roa, J. C. Caicedo, and F. A. Gonzalez. Visual pattern analysis in histopathology images using bag of features. *Lecture Notes in Computer Science*, pages 521–528, 2009.

[4] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *ECCV*, pages 1–22, 2004.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, April 2008.

[6] R. M. Esteves, R. Pais, and C. Rong. K-means clustering in the cloud – a mahout test. *Advanced Information Networking and Applications Workshops*, pages 514–519, 2011.

[7] S. G. and M. M. J. Orderless document representation: frequencies of words from a dictionary. 1983.

[8] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR 2005*, pages 524–531, Washington, DC, USA, 2005. IEEE Computer Society.

[9] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV '99*, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.

[10] H. Müller and T. M. Deserno. Content-based medical image retrieval. In *Biomedical Image Processing - Methods and Applications*. Springer, 2011.

[11] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *ICCV*, 2005.

[12] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991.

[13] J. A. Vanegas, J. C. Caicedo, F. A. González, and E. Romero. Histology image indexing using a non-negative semantic embedding. In *MCBR-CDS*, pages 80–91, 2011.

[14] C. wah Ngo, T. chuen Pong, and T. Chin. Exploiting image indexing techniques in dct domain. In *IAPR*, pages 1841–1851, 1998.

[15] B. White, T. Yeh, J. Lin, and L. Davis. Web-scale computer vision using MapReduce for multimedia data mining. In *MDMKDD '10*, MDMKDD '10, New York, NY, USA, 2010. ACM.

Appendix D

Large Scale Image Indexing using Online Non-negative Semantic Embedding

Vanegas J.A and González F.A., Large Scale Image Indexing using Online Non-negative Semantic Embedding. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2013, Vol. 8258, 2013, ISBN 978-3-642-41821-1, Springer-Verlag

Large Scale Image Indexing using Online Non-negative Semantic Embedding

Jorge A. Vanegas and Fabio A. González

MindLab Research Group, Universidad Nacional de Colombia, Bogotá, Colombia
{javanegasr, fagonzalezo}@unal.edu.co

Abstract. This paper presents a novel method to address the problem of indexing a large set of images taking advantage of associated multimodal content such as text or tags. The method finds relationships between the visual and text modalities enriching the image content representation to improve the performance of content-based image search.
This method finds a mapping that connects visual and text information that allows to project new (annotated and unannotated) images to the space defined by semantic annotations, this new representation can be used to search into the collection using a query-by-example strategy and to annotate new unannotated images. The principal advantage of the proposed method is its formulation as an online learning algorithm, which can scale to deal with large image collections. The experimental evaluation shows that the proposed method, in comparison with several baseline methods, is faster and consumes less memory, keeping a competitive performance in content-based image search.

1 Introduction

Large online collections of images are becoming common, thanks to the fast advance in acquisition, storage and communication technology. These collections are potential source of knowledge, but an effective and efficient access to them is fundamental to harness this potential. The classic way to search for images is by typing keywords on a search engine, but in many cases it is desirable to search by providing an example image. This approach, called content-based image retrieval, has been studied during the last two decades resulting in important progress . However, it is well known that matching visual features alone may lead to results with lack of semantic validity [16]. In this paper we address the problem of indexing the visual content of an image collection, enriching it with the semantic information provided by text annotations. The method presented in this papers learns relationships between visual features and text keywords co-occurring in images. A successful strategy to find these relationships is to build a common semantic representation space where both image and text content are embedded. This has been previously approached using different methods: Latent Semantic Analysis (LSA) [8], Latent Dirichlet Allocation (LDA) [1], Non-negative Matrix Factorization (NMF) [4] and Non-negative Semantic Embedding (NSE) [19], among others. The main drawback of most semantic learning strategies is that the algorithms are memory and computation intensive [7]. In order to address this drawback, it is proposed a reformulation of the NSE algorithm as an online learning process, which scales up to data collections with a vast amount of samples.

This work presents two main contributions: first, a reformulation of the NSE algorithm to make it scalable to large image collections, and second, an experimental evaluation of the algorithm performance in a content-based image retrieval task. The rest of this paper is organized as follows: Section 2 discusses the related work; Section 3 introduces the proposed method called Online Non-negative Semantic Embedding (ONSE); Section 4 presents the experimental evaluation; and, finally, Section 5 presents some concluding remarks.

2 Related Work

The strategy of finding relationships between visual and text representations has been extensively studied in the last years, specially focused in the task of image annotation. However many of the proposed algorithms have been designed without considering a large scale setup [15,10,11]. In some cases, these algorithms can be scaled up by relying on parallelized implementations and assuming the availability of abundant computational resources. However, this can be expensive, tricky and hard to accomplish.

There are some works that try to make semantic embedding approaches suitable for large scale collections. For example, Hsan et al. [18] propose to utilize multi-modality cues by incorporating visual and textual information as embedded objects, by using a simple linear projection to approximate the embedding functions, solving a non-smooth convex optimization problem. Their goal is to make the method (called Modified Multi-stage Convex Relaxation, MMCR) suitable for large scale image collections by reformulating the basic algorithm in some way that is possible to reduce the time complexity and the amount of storage, achieving a significant reduction in time complexity. Also, Jason Weston et al. [20] present a scalable architecture, proposing methods that learn to represent images and annotations jointly in a low dimension embedding space. To make training time efficient, they propose a loss function based in stochastic gradient descent (SGD) approach. Likewise, Juan Caicedo et al. [6] propose multimodal matrix factorization algorithms based on SGD to decompose a training data set, and find correspondences between visual patterns and text terms in large image collection.

The proposed algorithm in this work is based on a stochastic gradient descent approach, which, according to the work of Bottou [2], requires very little time to reach a predefined expected risk. This makes the strategy suitable for large scale learning problems, providing guarantees about convergence and scalability [2,3].

3 Online Non-negative Semantic Embedding Model

When the image associated text has a rich and clean semantic interpretation (e.g. tags provided by experts), the text representation may be used directly as the semantic space. So the problem of finding a common semantic representation for both visual and text content is reduced to map the visual content to the semantic space defined by the tags. A method that follows this strategy is the Non-negative Semantic Embedding (NSE) [19].

3.1 Non-negative semantic embedding

If the visual and semantic representations are vectors, a database of images can be represented with two matrices by joining the corresponding vectors of visual and semantic features as columns of the matrices. Let $V \in \mathbb{R}^{n \times l}$ be the matrix of visual features, where n is the number of visual patterns in the bag of features representations and l the number of images in the collection, and let $T \in \mathbb{R}^{m \times l}$ be the matrix of text terms, with m the number of keywords in the terms dictionary. NSE is used when we assume that the semantic encoding is already known, and we use it to index and represent all images in the collection. We formulate this problem as finding a linear transformation of the visual data imposing a non negativity constraint on the solution: $V \approx ST; S \geq 0$. Where, $S \in \mathbb{R}^{n \times m}$ is the transformation matrix representing the relationships between the visual and text modalities. The non-negativity constraint in this case enforces an additive reconstruction of visual features, since vectors in the matrix S can be thought of as parts of images that are combined according to the presence of associated labels. Notice that the vectors in S can be interpreted as the visual features related to each text term. Our purpose is to solve the problem under an online formulation using stochastic gradient descent, which is a gradient descent optimization method for minimizing an objective function that is written as a sum of differentiable functions. In this context, we can formulate the problem of semantic embedding as the optimization problem of $\min_{S \geq 0} d(V, ST)$. Where, $d(.,.)$ is a function that measures the difference between V and ST. The purpose is to find S that minimize this difference.

3.2 Kullback-Leibler divergence optimization

A popular measure function for NMF is the generalized Kullback-Leibler divergence between V and ST [14], Although the KL-divergence equation is not symmetric, and therefore, it is not strictly a distance metric. This allows to take advantage of the normalized visual and text representation that can be interpreted as probability distributions. Zhirong Yang et. al [21] show that projected gradient methods based in for KL-divergence runs faster and yields better approximation than others widely used NMF algorithms. The updating rule for gradient descent approach with τ as the index of iterations and γ as the step size is:

$$S_{\tau+1} = S_{\tau} + \gamma \left[\left(\frac{V}{ST} - [1]_{n \times l} \right) T^{\mathsf{T}} \right] \quad . \tag{1}$$

This algorithm requires a non-negativity restriction that can be incorporated by using a projected gradient strategy. The projection function maps a point back to the feasible region in each iteration [13], updating the current solution S_{τ} to $S_{\tau+1}$ by the following rule:

$$S_{\tau+1} = P[S_{\tau} - \gamma \nabla f(S_{\tau})]; \quad P[s_{ij}] = \begin{cases} s_{ij} & if\, s_{ij} \geq 0, \\ 0 & if\, s_{ij} < 0, \end{cases} \quad . \tag{2}$$

3.3 Online formulation

The idea of online learning using stochastic approximations is to compute the new solution for each unknown in the problem using a single data sample at a time. Then,

Algoritmo 1 Online Non-negative Semantic Embedding

```
input $S^0$: Initial transformation matrix, $\gamma_0$: initial step size, $N$: number of iterations
for $k = 1$ to $N$ do
    1. Step size calculation: $\gamma_k = \gamma_0/(1+\gamma_0\lambda k)$
    2. Update transformation matrix: $S_{\tau+1} = P\left[S_\tau - \gamma_k\left[\left(\frac{v_\tau}{S_\tau t_\tau} - [1]_{n\times m}\right)t_\tau^{\mathsf{T}}\right]\right]$
end for
return $S_{\tau+1}$
```

we can scan large data sets without memory restrictions. The updating rule has to be reformulated in such a way that it only depends on the τ-th sample (v_t , t_t, visual and text features for the τ-th image). The updating rule is reformulated as follows:

$$S_{\tau+1} = S_\tau + \gamma\left[\left(\frac{v_\tau}{S_\tau t_\tau} - [1]_{n\times 1}\right)t_\tau^{\mathsf{T}}\right] \quad . \tag{3}$$

The resulting algorithm (Algorithm 1) starts by randomly initialization of the transformation matrix. Each iteration consists on updating the transformation matrix from an observed pair of visual and text features randomly obtained. The step size used in this algorithm is a decreasing rate [2] that depends on the number of iterations and an initial learning rate γ_0. A small variation of this algorithm is obtained by using several samples at each iteration instead of using only one. Experimental results show faster execution when using mini-batches instead of single examples, and also a better numerical stability for the solution.

3.4 Image Indexing and Search

A special indexing case is when images do not have attached text. An example of this situation is when users are interested in searching the database using example images as queries. A new image without text can be projected to the semantic space by finding the pseudo-inverse of the transformation matrix (S^+) .

$$t = S^+ v; \quad S^+ = \left(S^{\mathsf{T}} S + \beta I\right)^{-1} S^{\mathsf{T}} \quad . \tag{4}$$

where, v is the visual representation of the new image, t is the semantic representation and β is a regularization parameter. In this way we can searching the database using an inferred text representation based in its visual features. This pseudo-inverse matrix has to be preprocessed only once and storing in memory, making very efficient the process of projection for a new image. Finally, the ranking function for semantic search is based on the histogram intersection similarity[17].

4 Experiments and Results

4.1 Datasets

The performance of the proposed algorithm was evaluated using three different datasets with different sizes:

Carcinoma dataset. The Carcinoma dataset is a histopathology image collection that has been used to diagnose a special kind of skin cancer known as basal-cell carcinoma [5]. It is composed of 1,502 images that were studied and annotated by pathologists to highlight various tissue structures and relevant diagnostic information, elaborating a list with 18 terms. These images were acquired at various magnification levels, including 8X, 10X and 20X, and stored at 1280×1024 pixels. The list of keywords includes terms like micro-nodules, elastosis, and fibrosis, among others.

Histology dataset. The Histology dataset is composed of 2,641 images extracted from an atlas of histology for the study of the four fundamental tissues [19]. The collection includes photographs of histology in different magnification factors (10X, 20X and 40X). The resolution of these images is about 800×500 pixels. Each of these images was annotated by an expert, indicating the biological system and organs that can be observed. The total number of different keywords in this data set is 46.

MIRFlickr 25000 dataset. The MIRFlickr-25000 image dataset is composed of 25,000 pictures downloaded from the popular online photo sharing service Flickr. These photos were collected directly from the web, to provide a realistic dataset for image retrieval research, with high-resolution images and associated metadata [12]. This image collection has been manually annotated using a set of 38 semantic terms.

4.2 Experimental Setup

We conducted retrieval experiments under the query-by-example paradigm. In all datasets 20% of images were randomly selected as queries and the remaining images were used as the target collection to find relevant images. We performed automatic experiments by sending a query to the system and evaluating the relevance of the results. A ranked image in the results list is considered relevant if it shares at least one keyword with the query. The evaluation was done using traditional measures of image retrieval, including precision at 10 and mean average precision (MAP).

Image Features. In all datasets we build a bag-of-features representation, with the following characteristics: Patches of 8×8 pixels are extracted from a set of training images with an overlap of 4 pixels along the x and y axes. The DCT (Discrete Cosine Transform) transform is applied in each of the 3 RGB channels to extract the largest 21 coefficients. (DCT-based visual codewords has been found to be an effective representation for microscopy image analysis [9]). A k-means clustering is applied to build a dictionary. For Carcinoma and Histology datasets we use 500 visual terms and for MIRFlickr we select a dictionary of 2000 features (larger dictionaries do not provide significant improvements, but just more computational load). Once the vocabulary has been built, every image in the collection goes through the patch extraction process. Each patch from an image is linked to one visual term in the dictionary using a nearest neighbor criterion. Finally, the histogram of frequencies is constructed for each image.

Text annotations. In these data sets the text annotations are clean and clearly defined terms from a technical vocabulary and these represent directly the semantic space. We build semantic vectors following a boolean approach, assigning 1 to the terms attached to an image and 0 otherwise. This leads to 46-dimensional binary vectors, for text representation in the Histology dataset, 18-dimensional binary vectors for Carcinoma dataset and 39-dimensional binary vectors for Flickr.

4.3 Retrieval Performance

In order to evaluate the performance of the proposed algorithm, we compare the proposed online algorithm with the classical NSE and the MMCR (Modified Multi-stage Convex Relaxation) proposed by Hsan et. al [18]. Although the MMCR algorithm was proposed mainly for annotation, it is possible to use its semantic score vector as a new representation for retrieval task.

Parameter Tuning. The proposed algorithm has a set of parameters that can impact the quality of the resulting model. Improper settings of these parameters may cause the algorithm converge slowly or diverge. So, as preliminary evaluation, we perform an exploration of these parameters by retrieval experiments using cross-validation 10 fold in the subset of 80% of the images that were not selected as queries. And, we select the configuration that perform better in average in all folds (Table 1).

Table 1. Results of parameter tuning for Online Non-negative Semantic Embedding (ONSE).

Carcinoma				**Histology**				**MIRFlickr**			
λ_0	γ	β	**Mini-batch size**	λ_0	γ	β	**Mini-batch size**	λ_0	γ	β	**Mini-batch size**
2^{-5}	2^{-2}	2^{4}	16	2^{-6}	2^{-3}	2	16	2^{-8}	2^{-10}	2	32

Once, we had found the better configuration, we evaluate the proposed algorithm with the remaining 20% of images as test. So we use this 20% of images as queries and the 80% as finding objective. Table 2 summarizes the findings of our experimental results. In all cases, a general improvement over visual baseline (direct visual matching using visual representation) is shown in MAP measure. And, with the exception of the Histology dataset NSE, ONSE-KL and MMCR algorithms, present a very similar performance.

Table 2. Image retrieval performance. Reported measures are Mean Average Precision (MAP) and Precision at the first 10 results (P@10).

Algorithm	**Carcinoma**		**Histology**		**MIRFlickr**	
	MAP	**P@10**	**MAP**	**P@10**	**MAP**	**P@10**
Visual	0.2236	0.3503	0.2107	0.6104	0.2505	0.4931
MMCR [18]	0.3146	0.3322	0.5346	0.6030	0.3670	0.5063
NSE [19]	0.3265	0.3249	0.4025	0.4148	0.3672	0.5079
ONSE	0.3171	0.3651	0.3594	0.4439	0.3674	0.5065

4.4 Computational Load

Table 3 shows the average time consumption for the training phase. Reported times are the result of running all algorithms 5 times in a computer with 4 GB of ram memory and a CPU at 2.4Ghz using only one core. The size of each dataset is also reported to observe how the algorithm complexity grows. NSE algorithm take about 5 seconds to process the Carcinoma dataset, 9 to process the Histology collection and finally increases to 494 seconds for MIRFlickr. MMCR have the most time consuming, requiring about 2 seconds for Carcinoma 14 for Histology and 2834 for MIRFlickr. In contrast, the ONSE algorithm only requires 0.3 seconds for Carcinoma, 1.2 for Histology and 27 for MIRFlickr. Thus for MIRFlickr dataset, ONSE algorithm is 18 times faster than NSE and 104 times faster than MMCR.

Table 3. Time consumption in training phase: Time required for each epoch (Epoch Avg. Time) and the total average time required until convergence (Total Avg. Time). The algorithm presented in this paper (ONSE) is compared against MMCR [18] and NSE [19].

Dataset	Size	Algorithm	Epochs	Epoch Avg. Time (sec)	Total Avg. Time (sec)
Carcinoma	1502	MMCR	8	0.2854	2.1878
		NSE	130	0.0411	5.3442
		ONSE	**4**	**0.0836**	**0.3345**
Histology	2641	MMCR	10	1.5351	14.2029
		NSE	90	0.1009	9.0869
		ONSE	**4**	**0.3027**	**1.2086**
MIRFlickr	25000	MMCR	10	283.4327	2834,3278
		NSE	200	2.4701	494.017
		ONSE	**2**	**13.755497**	**27.2188**

The main reason for the reduction of training time, is, that the number of required epochs until the ONSE algorithm converges is reduced drastically (convergence in all algorithms is verified by means of a minimum threshold required to improve the error in each epoch). For instance, in the carcinoma dataset the NSE algorithm required 130 full scans to the training set and the online version only needed 4. In general, Bottou [3] shows that for a small collection, it is necessary to use very few epochs and for large collections, one full scan is enough. Furthermore, the proposed algorithm reduces the memory requirements, since the only element necessary to keep in memory is the transformation matrix, since visual and textual samples used in each update can be discarded,.

5 Conclusions

We presented an approach for large image indexing that takes advantage of text annotations to provide a semantic representation space where the visual content of images is embedded. This approach is a reformulation of NSE as an online learning algorithm allowing to deal with large collections of data, achieving a significantly reduction in memory requirements and computational load, but keeping a competitive retrieval performance.

Acknowledgements. This work was partially funded by projects Anotación Automática y Recuperación por Contenido de Imágenes Radiológicas Usando Semántica Latente, No. 110152128803 by Colciencias 0521-2010, Diseño e implementación de un sistema de cómputo sobre recursos heterogéneos para la identificación de estructuras atmosféricas en predicción climatológica, No. 1225-569-34920 by Colciencias 0213-2013 and Multimodal Image Retrieval to Support Medical Case-Based Scientific Literature Search, ID R1212LAC006 by Microsoft Research LACCIR.

References

1. K. Barnard, P. Duygulu, D. Forsyth, N. D. Freitas, D. M. Blei, J. K, T. Hofmann, T. Poggio, J. Shawe-taylor. Matching words and pictures. *JMLR*, 3:1107–1135, 2003.
2. L. Bottou. Large-scale machine learning with stochastic gradient descent. *COMPSTAT'2010*, strony 177–187, Paris, France, August 2010. Springer.
3. L. Bottou, Y. LeCun. Large scale online learning. *NIPS*, 2003.
4. J. C. Caicedo, J. BenAbdallah, F. A. González, O. Nasraoui. Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. *Neurocomput.*, 76(1):50–60, Sty. 2012.
5. J. C. Caicedo, A. Cruz, F. A. Gonzalez. Histopathology image classification using bag of features and kernel functions. *AIME 2009*, strony 126–135, Berlin, Heidelberg, 2009. Springer-Verlag.
6. J. C. Caicedo, F. A. González. Online matrix factorization for multimodal image retrieval. *Iberoamerican Congress on Pattern Recognition CIARP*, 2012.
7. P. Chandrika, C. V. Jawahar. Multi modal semantic indexing for image retrieval. CIVR '10, strony 342–349, New York, NY, USA, 2010. ACM.
8. Q. Chen, X. Tai, B. Jiang, G. Li, J. Zhao. Medical image retrieval based on latent semantic indexing. CSSE '08, strony 561–564, Washington, DC, USA, 2008. IEEE Computer Society.
9. A. Cruz-Roa, J. C. Caicedo, F. A. González. Visual pattern mining in histology image collections using bag of features. *AIME*, 52(2):91–106, Czerw. 2011.
10. C. Fang, L. Torresani. Measuring image distances via embedding in a semantic manifold. *ECCV*, strony 402–415, Paz. 2012.
11. M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. *In ICCV*, 2009.
12. M. J. Huiskes, M. S. Lew. The mir flickr retrieval evaluation. *MIR '08*, New York, NY, USA, 2008. ACM.
13. C. jen Lin. Projected gradient methods for non-negative matrix factorization. Raport instytutowy, Neural Computation, 2007.
14. D. D. Lee, H. S. Seung. Algorithms for non-negative matrix factorization. *In NIPS*, strony 556–562. MIT Press, 2000.
15. A. Makadia, V. Pavlovic, S. Kumar. A new baseline for image annotation. D. A. Forsyth, P. H. S. Torr, A. Zisserman, redaktorzy, *Computer Vision - ECCV 2008*. Springer, 2008.
16. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain. Content-based image retrieval at the end of the early years. *TPAMI*, 22(12):1349–1380, 2000.
17. M. J. Swain, D. H. Ballard. Color indexing. *IJCV*, 7:11–32, 1991.
18. M.-H. Tsai, J. Wang, T. Zhang, Y. Gong, T. S. Huang. Learning semantic embedding at a large scale. *ICIP*, strony 2497–2500, 2011.
19. J. A. Vanegas, J. C. Caicedo, F. Gonzalez, E. Romero. Histology image indexing using a non-negative semantic embedding. *MCBR-CDS 2011*. Springer, 2011.
20. J. Weston, S. Bengio, N. Usunier. Large scale image annotation: Learning to rank with joint word-image embeddings. *ECML*, 2010.
21. Z. Yang, H. Zhang, Z. Yuan, E. Oja. Kullback-leibler divergence for nonnegative matrix factorization. *ICANN*, strony 250–257, 2011.

Appendix E

Bioingenium at ImageCLEF 2012: Text and Visual Indexing for Medical Images

Vanegas, J.A., Caicedo, J. C., Camargo, J., Ramos-Pollán, R., & Gonzalez, F. Bioingenium at ImageCLEF 2012: Text and Visual Indexing for Medical Images. CLEF (Online Working Notes/Labs/Workshop).

Bioingenium at ImageCLEF 2012: Textual and Visual Indexing for Medical Images

Jorge A. Vanegas, Juan C. Caicedo, Jorge E. Camargo, Raul Ramos-Pollán, and Fabio A. González

Bioingenium Research Group
Universidad Nacional de Colombia
{javanegasr, jccaicedoru, jecamargom, rramosp, fagonzalezo}@unal.edu.co
http://www.bioingenium.unal.edu.co

Abstract. This paper describes the participation of the Bioingenium research group of Universidad Nacional de Colombia in the ImageCLEF2012 Medical Retrieval challenge, specifically in the ad-hoc image-based retrieval task. The methods used for solving textual and visual queries with which we submitted uni-modal runs are described. They were ranked 1st and 3rd respectively. These results have been obtained by using our own implementation of Okapi-BM25 weighting scheme for text retrieval, and by adding spatial layouts to the CEDD descriptors for visual retrieval. We also used these uni-modal features to learn multimodal representations using matrix factorization for solving visual queries. Despite the potential of multimodal indexes to improve the quality of visual queries, these experiments were not as successful as uni-modal indexes. We discuss the main findings of all these experiments.

Keywords: Image Retrieval, Medical Images, Multimodal Indexing.

1 Introduction

This paper describes the participation of the Bioingenium research group of Universidad Nacional de Colombia in the 2012 version of the Medical Image Retrieval challenge at ImageCLEF [7]. Our first motivation was to investigate the extent to which textual and visual indexes may be improved for searching in the collection of medical images, using keywords and visual examples separately. We aimed at designing suitable textual and visual representations by extending models that were successful in previous years, and preparing these representations for subsequent multimodal analysis.

For text indexing, we developed our own implementation of Okapi-BM25, which allows to determine limits on the number of terms used in the vector representation, by pruning irrelevant terms and keeping the most informative ones. For visual indexing, we introduced a spatial pyramid of CEDD features, making recursive partitions of the image and computing descriptors in each subregion. This representation extends the popular CEDD descriptor with spatial information, which results in an improved performance. We also implemented spatial extensions for bag-of-features histograms.

Our second motivation was to build an enhanced image index using both modalities, but for searching with visual examples only. The goal was to learn a multimodal representation that incorporates textual and visual information in the database, and then predict the multimodal representation for queries using visual features. This represents a very challenging problem since the medical image collection, with more than 300K images, constitutes a very large training set that poses computational difficulties for most learning algorithms. Other problems arise from this large multimodal image collection, such as the high dimensionality of textual and visual representations, and the presence of noise.

The results obtained with uni-modal strategies were successful in the general pooling, which ranked first in the case of textual queries among 54 other submissions, and third in the case of visual queries among 36 experiments. We consider that the multimodal indexing submission was not successful since it did not improve upon our own visual indexing strategy, which was the original goal. However, further experiments conducted off competition demonstrate interesting improvements. We believe that further research in this front may help to design more accurate image search systems working with the query-by-visual-example paradigm.

The structure of this paper is as follows: Section 2 briefly describes the medical image collection, Section 3 describes our text indexing approach, Section 4 presents the visual indexing strategies, Section 5 discusses the multimodal indexing approach for visual queries, and finally, conclusions and future works are outlined in Section 6.

2 The Data Set

The medical retrieval task of ImageCLEF 2012 is based on a subset of PubMed Central papers, containing 305,000 images extracted from biomedical articles. Participants have access to the selected images as well as all content of the corresponding articles. This year, a set of 22 topics was released for evaluation of the retrieval systems, where each one is composed of a variable number of images and associated text in 4 languages.

3 Text Indexing

Images in the collection belong to a medical article, so they can be indexed using the surrounding text content. Our goal was to build a term-document matrix using a vector space model with the Okapi-BM25 weighting scheme [6]. We developed an indexing tool using the Natural Language Toolkit for Python [1], which provides a clean API and extensive functionalities for common text processing tasks.

The text representation adopted in this work included information from the title of the paper and the image caption, which can be found in the XML file corresponding to each image in the data set. With that, a text corpus for the image collection was built, and standard text processing operations were applied, including tokenization, stemming, and stop-word removal. These operations determined the initial list of indexing terms.

We designed a prunning criterion to discard irrelevant terms from the initial list, thus, preserving only the most informative ones. A limit for the number of terms was established depending on their document frequency. If it is outside of a predefined interval, the term is removed from the indexing list. The thresholds were computed according to a minimum and maximum number of documents in which a term is allowed to occur. The criterion is as follows:

$$keep(t) = \begin{cases} true & \text{if } min < \mathrm{df}_t < max \\ false & \text{otherwise} \end{cases} \tag{1}$$

where df_t is the number of documents that contain the term t, and min and max are parameters that define the minimum and maximum number of documents in which the term should appear. The definitive list of indexing terms is obtained by applying this rule, which is very useful to limit the dimensionality of the resulting vector space for indexing.

The term-document matrix is built using term frequencies in each document, and Okapi-BM25 [6] is used to highlight the importance of the most relevant terms. Usually, BM25 is used as a ranking function that involves different factors including: term frequencies, inverse document frequencies, and the length of both, the document and the query. However, in our approach, we wanted to use the ideas of BM25 as a term weighting scheme so that we can apply further processes to the term-document matrix (such as multimodal fusion). The following equation describes the BM25-based term weighting used:

$$\mathrm{weight}(t,d) = \left[log \frac{N}{\mathrm{df}_t} \right] \cdot \left[\frac{(k_1+1)\,\mathrm{tf}_{t,d}}{k_1\left((1-b)+b\left(\frac{L_d}{L_{avg}}\right)\right)+\mathrm{tf}_{t,d}} \right] \tag{2}$$

where $\mathrm{tf}_{t,d}$ is the frequency of term t in document d, L_d and L_{avg} are the length of document d and the average document length in the collection, respectively, and, k_1 and b are positive tuning parameters to calibrate the term frequency scaling. We fixed $k_1 = 1.5$ and $b = 0.75$ according to the suggestions presented by Manning et al. [6]. For queries, we only used term frequencies without weighting, and the dot product similarity score was employed for document ranking.

3.1 Results

We submitted 2 textual runs using the indexing strategy described above, with the goal of evaluating the difference in performance after pruning the list of indexing terms. In both cases, we set a minimum frequency of 20 documents in which a term should be present to keep it in the list. The first experiment used 20,000 documents as maximum frequency, and the second experiment used 5,000 documents. These parameters resulted in vector spaces with approximately 28,000 and 18,000 dimensions, respectively.

Table 1. Retrieval performance of the submitted runs in the Medical Ad-hoc Image-Based Retrieval Task, using textual queries.

Run	Position	MAP	P@10
unal.text.bm25.20000	1	0.2182	0.3409
unal.text.bm25.5000	14	0.2045	0.2955
Extra (50,000)	N.A.	0.1991	0.3318

An additional experiment was evaluated after the competition finished to assess the contribution of the pruning strategy with respect to a longer list of terms. This experiment used 50,000 documents as maximum frequency, and produced a list of 29,000 terms. Notice how the number of indexing terms is controlled by the use of these two parameters, which remove rare terms as well as too common terms. We used this property to control the dimensionality of the resulting term-document matrix for further analysis, as is described later in Section 5.

Table 1 reports performance measures for the three experiments, the two first submitted to the official pooling and a third experiment run after the challenge. These results show the impact of the pruning strategy in the precision of the retrieval task, showing how the performance decreases by keeping or removing the wrong terms. The best response was obtained by the index limited by a frequency of 20,000 documents. This result ranked first in the category of textual experiments, and is the second best performance overall in the poolings for adhoc image-based medical retrieval.

Our second submission used 18,633 indexing terms, resulting in a significant dimensionality reduction, but also an important reduction in performance. This difference dropped the MAP performance in about 6%, leaving this experiment in the position number 14. However, notice that keeping more terms than those actually needed, can hurt the general retrieval precision even more. The additional experiment shows that a slight increase in the number of terms resulted in a decrease in performance of about 9%.

4 Visual Indexing

Our research group is currently leading an initiative to develop a framework for large scale image analysis for academic and scientific applications. The framework, named BIGS[8], is implemented in the Java programming language and integrates a wide variety of image processing tools, including feature extraction and learning algorithms. One of the most remarkable characteristics of BIGS is that it can easily run in a distributed environment with heterogeneous computing resources, from laptop and desktop computers to high-performance servers.

Obtaining a good quality representation for image contents in large databases is a challenging task, and BIGS was used to tune up image indexes by conducting experiments on the ImageCLEFmed 2011 data set. The experiments were run on different servers scattered throughout our lab, using BIGS to process all images stored on an HBase NoSQL database[1]. In spite of the large size of the image collection, having an lightweight experimental lifecycle as provided by BIGS was key to be able to gain understanding on how to better tune up image indexes.

As a result, we designed two indexes for content-based image retrieval for this year's data set, focusing on including spatial information in the representation, since it can help to better discriminate medical image arrangements. The Color and Edge Directivity Descriptor (CEDD) [3] was used as basic low-level characteristic in both indexes, since it has demonstrated good performance in image retrieval tasks, while keeping a small and compact representation.

4.1 Spatial Pyramid CEDD

The CEDD descriptor is a compact representation of the image content, consisting in a histogram of 144 bins to codify information of colors and edges. The small size of this descriptor makes it an excellent choice for indexing large scale image collections. This descriptor has been previously evaluated in the context of medical image retrieval at ImageCLEF, exhibiting a competitive performance due to the variety of image modalities and visual configurations in this data set.

We extended this representation by computing the CEDD descriptor in a recursive partition of the image in quadrants, forming a pyramid of spatially organized regions [5]. We employed a configuration using the full image plus 2 pyramid levels, which results in 21 spatially distributed regions, ending up in a visual representation with 3,024 features. These descriptors were computed from high-resolution images, i.e., as they are distributed in the ImageCLEFmed data set.

This descriptor was computed by the BIGS framework using 40 workers deployed in several computers at our lab. The total time required to index the full image collection of 305,000 images using this strategy was 37 minutes. Finally, the similarity between two images is calculated on this descriptor using the Tanimoto coefficient. Assuming that x and y are vector representations of the spatial pyramids for two images, this is computed as:

$$T_D = t(x,y) = \frac{x^T y}{x^T x + y^T y - x^T y} \tag{3}$$

4.2 Spatial Bag-of-Features

An image index using the bag-of-features representation [4] was introduced in our experiments as well. The bag-of-features methodology is comprised of 3 main procedures: extraction of local features from images, construction of a dictionary of visual words, and the computation of the histogram for each image. Spatial layouts can be added to enhance the representation with the relative position of words in the image plane. In that sense, this representation incorporates local,

[1] http://hbase.apache.org/

low-level information of images as well as global, spatially distributed arrangements.

For local features, we extracted blocks of 32×32 pixels on a regular grid and the CEDD descriptor is computed in these patches. The k-means algorithm is used to cluster a large sample of patches extracted from the collection, for building a dictionary of 5,000 visual terms. The histogram is constructed by counting the occurrence of dictionary words in each image. Besides the global counting of visual patterns, each image is also split in 3 horizontal, non-overlapping strips, and an additional histogram is computed there to estimate the spatial distribution of visual words. This results in four bag-of-features histograms that are bounded together in a single image descriptor with 20,000 features.

This representation was also computed using the BIGS framework with 40 workers deployed in several computers at our lab. The total time required to extract this representation for all images in the collection was 116 minutes, which is less than one hour and a half. The similarity measure computed for this representation is the histogram intersection, for two images with histograms x and y:

$$K_{HI}(x,y) = \sum_{i=1}^{n} min\left\{x_i, y_i\right\} \tag{4}$$

4.3 Visual Queries with Multiple Images

The topics proposed for this year's challenge included 22 different queries with multiple images, some of them with 6 or even 7 example images. Since a single ranking is required for queries with multiple image examples, a similarity integration rule was employed. The similarity score for a database image d with respect to a multi-image query $q = \{q_1, q_2, ..., q_n\}$, is obtained as follows:

$$score(d,q) = \sum_{k=1}^{n} similarity(d, q_k) \tag{5}$$

4.4 Results

We submitted two runs, one using the spatial pyramid of CEDD features and another with the spatial bag-of-features. The results are reported in Table 2, and shows that the spatial pyramid obtains a significantly better performance than the bag-of-features, both in general precision (MAP) and early precision (P@10 and P@30). The difference can also be observed in the positions obtained by these experiments in the general poolings, the spatial pyramid was ranked 3rd, whereas the bag-of-features was ranked 14th.

The spatial pyramid extension for the CEDD descriptor demonstrated to be an effective representation to discriminate more relevant images in this task. In addition, computing the spatial pyramid did not result in an excessive load of both, computational effort and representation length. This representation is still very light to compute with respect to the bag-of-features and keeps a compact descriptor with about 3,000 features.

In our preliminary experiments, we observed that adding a spatial layout on the image representation improves the performance of the medical image retrieval task. The two visual representations proposed in this work include spatial information using recursive computations of the same descriptor in partitions of the image. One of the reasons the spatial pyramid CEDD presented better performance than the bag-of-features is because of the level of granularity in the recursive partition, that allows to introduce more spatial details. This can be achieved because of the short length of the original CEDD descriptor, as opposed to the large dictionary of visual features that we employed in these experiments.

Table 2. Performance measures of the submitted runs in the Medical Ad-hoc Image-Based Retrieval Task for visual queries.

Run	**Position**	**MAP**	**P@10**	**P@30**
unal.visual.pyramidal.cedd.tanimoto	3	0,0073	0,0636	0,05
unal.visual.spatial.bof.3x1	14	0,0033	0,0455	0,0364

5 Multimodal Indexing for Visual Queries

One of our motivations to design textual and visual indexes for medical image collections is to develop a multimodal framework to integrate both modalities in a common representation. We focus our attention to the specific case of enhancing visual search functionalities by introducing available text information into the visual index. Thus, the goal is to improve the retrieval response using multimodal information even when users search with example images only.

In this work, we employed a multimodal latent factors model proposed in [2] for learning the relationships between visual features and text terms. The method is based on a matrix factorization algorithm, that proceeds with a multimodal decomposition of the visual and text matrices on a training data set. The matrix factorization problem is defined as follows:

$$\min_{P,Q,H} \frac{1}{2}\left(\|V - PH\|_F^2 + \|T - QH\|_F^2 + \lambda\left(\|P\|_F^2 + \|Q\|_F^2 + \|H\|_F^2\right)\right) \tag{6}$$

where $V \in \mathbb{R}^{n\times\ell}$ is the matrix of n visual features for ℓ training examples, $T \in \mathbb{R}^{m\times\ell}$ is the matrix of textual information with m terms, $P \in \mathbb{R}^{n\times r}$ is the transformation from the visual space to a multimodal space with r factors, $Q \in \mathbb{R}^{m\times r}$ is the transformation from the textual space to the multimodal space, and $H \in \mathbb{R}^{r\times\ell}$ is the multimodal latent representation for the training images. λ is a regularization parameter for this learning problem.

The solution to this problem presented in [2] is an online matrix factorization algorithm that can be scaled up to large data sets. This is specially useful for the

ImageCLEFmed 2012 collection, which has a large number of images that can be used for learning multimodal relationships between visual and textual information. When the linear transformation functions P and Q have been learned, new images can be projected to the multimodal space using the following equation:

$$h = \left(P^T P + \xi Q^T Q\right)^{-1} P^T v \tag{7}$$

where h is the multimodal representation for an image with visual features v, and ξ is a regularization parameter. The purpose of using these algorithms is to obtain a multimodal latent factor representations for all images, even if they do not have available text annotations, as may be the case of the queries. Using the multimodal representation, the ranking of images is computed using the dot product similarity measure, which indicates the extent to which two images share the same latent factors.

This strategy has demonstrated to be an effective method to learn multimodal relationships from image collections with attached texts, resulting in a data-driven representation for images that incorporates both modalities. Previous studies have shown important performance gains for these approaches, since visual features are complemented by the semantics of text descriptions, providing an enhanced mechanism of representing images.

5.1 Results

To construct a multimodal index for image search, we employed the matrices of text terms and visual features described in Sections 3 and 4, respectively. More specifically, we used the term-document matrix with 18,000 terms weighted with Okapi-BM25, and the visual matrix with 3,024 spatial pyramid CEDD features. One of the reasons we were interested in designing textual and visual indexes with bounded dimensionality is to reduce the computational cost of learning multimodal relationships.

The online multimodal matrix factorization (OMMF) algorithm was trained with the full collection of images in this challenge, i.e., using the 305,000 images with their corresponding text annotations. An implementation of the algorithm in the Java programming language was employed, which decomposed the matrices of 18,000 rows for text data, and 3,024 rows for visual data, with 305,000 columns in both cases, in 131 minutes in average. This algorithm has been designed to learn from as many examples as possible in a short time.

To tune up the learning algorithm and determine appropriate parameters for the factorization, experiments were conducted with the ImageCLEFmed 2011 collection. We found good parameters to solve queries in the previous year's challenge, that included 600 multimodal latent factors and other regularization parameters as needed. The criteria to select parameters for this algorithm is to observe improvements with respect to the direct visual matching, i.e., with respect to the visual indexing methods presented in Section 4, since the queries used in this experiment are also based on example images only.

Table 3. Performance results for multimodal indexing to solve visual queries. The first row reports the baseline method based on visual features only. The second row presents the results of the run submitted to the official poolings. The third row reports the result of an additional experiment run off competition.

Run	Position	MAP	Improvement	P@10	Rel-Ret
unal.visual.pyramidal.cedd.tanimoto	3	0,0073	N.A.	0,0636	117
unal.cedd.factorization.600	19	0,0024	-67.1%	0,0091	45
Additional experiment	N.A.	0.0087	+19.2%	0.0182	137

With the parameters that showed improvements in the 2011 collection, we prepared and submitted a run to the official poolings. Table 3 reports the results of this submission, as well as two other experiments for comparison. The first experiment in the Table is our baseline method, based on direct matching of visual features. The second result is the performance of the prepared run that has shown a decrease in performance with respect to the baseline. This loss is mainly explained by the use of parameters tuned to improve the performance in the 2011 challenge.

There are several differences between the challenge of 2011 and 2012. First, the nature of the proposed topics varied significantly, as this year's queries included more example images per topic, in average. Second, the size of the collection was increased, which resulted in bigger matrices in both dimensions. Third, this year's visual queries seem to be more difficult to answer, judging by the relative decrease in MAP observed in the results from 2011 to 2012. All these aspects may require a different configuration for the learning algorithm, in order to make it effective to retrieve more relevant results.

The results reported in the third row of Table 3 present the performance measures for an additional experiment run off competition to estimate the potential of the OMMF algorithm to improve upon the baseline. This result was obtained by tuning the algorithm parameters more appropriately for this year's task, and shows an important relative improvement.

The main goal of a multimodal algorithm in this context is to extract meaningful relationships between visual features and text terms. An additional challenge that makes the multimodal indexing strategy difficult to setup correctly, is attributed to the properties of the textual modality, which is very noisy and unstructured. Extracting semantic information useful for image analysis in this condition is still a very interesting research problem that requires further analysis.

6 Conclusions And Future Work

This paper presented the participation of the Bioingenium research group of Universidad Nacional de Colombia in the ad-hoc image-based medical retrieval task at ImageCLEF 2012. We submitted 5 runs: 2 textual and 3 visual, from which one was ranked first in the text modality and another was ranked third in the visual modality. These results were obtained by incorporating simple and effective extensions to well-known strategies for this task. We also explored multimodal indexing to answer visual queries, which is a very challenging and interesting research problem, that still requires further analysis. We believe that this is a promising research direction for improving image search systems, and the study of these models are the focus of our future research.

One of the main difficulties of this year's challenge was the size of the database, which required efficient computational tools to process and index the collection. In this work, we supported all of our visual indexing experiments on a distributed computing framework for large scale image analysis, named BIGS [8]. This framework allowed us to accelerate the exploration of visual indexing strategies, and investigate new image representation designs, such as the spatial pyramid CEDD that ranked third among 36 other experiments. We also used online learning algorithms for extracting multimodal relationships efficiently by training with the full collection of medical images in short execution times.

Acknowledgments

This work was partially funded by the project Anotación Automática y Recuperación por Contenido de Imágenes Radiológicas Usando Semántica Latente, number 110152128803 and project Sistema para la Recuperación de Imágenes Médicas Utilizando Indexación Multimodal, number 110152128767 by Convocatoria Colciencias 521 de 2010.

References

1. Steven Bird. Nltk: The natural language toolkit, 2002.
2. Juan C. Caicedo and Fabio A. Gonzalez. Online matrix factorization for multimodal image retrieval. In *17th Iberoamerican Congress on Pattern Recognition, CIARP*, 2012.
3. Savvas A. Chatzichristofis and Yiannis S. Boutalis. Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *Proceedings of the 6th international conference on Computer vision systems*, ICVS'08, pages 312–322, Berlin, Heidelberg, 2008. Springer-Verlag.
4. G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.
5. Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.
6. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA, 2008.
7. Jayashree Kalpathy-Cramer Dina Demner Fushman Sameer Antani Ivan Eggel Müller, Alba Garcia Seco de Herrera. Overview of the imageclef 2012 medical image retrieval and classification tasks. 2012.

8. Raul Ramos-Pollán, Fabio A. González, Juan C. Caicedo, Angel Cruz-Roa, Jorge E. Camargo, Jorge A. Vanegas, John E Arévalo, Paola K Rozo, Santiago A Pérez, Jose D Bermeo, and Juan S Otálora. BIGS: A framework for large-scale image processing and analysis over distributed and heterogeneous computing resources. In *IEEE International Conference on eScience. To appear*, 2012.