
# Multimodal Information Spaces for Content-based Image Retrieval

**Juan C. Caicedo**

Universidad Nacional de Colombia
Facultad de Ingeniería
Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2012

**Doctoral Thesis**

# Multimodal Information Spaces for Content-based Image Retrieval

**by**

Juan Carlos Caicedo Rueda

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

Doctor of Engineering
Systems and Computer Engineering

Advisor:
Fabio A. Gonzalez

Bioingenium Research Group

Universidad Nacional de Colombia
Facultad de Ingeniería
Departamento de Ingeniería de Sistemas e Industrial
Bogotá D.C., 28 Sep 2012

To my parents Germán and Beatriz.

# Aknowledgements

I would like to thank my advisor Professor Fabio González for his guidance, patience and advice. He introduced me to the field of machine learning and information retrieval research, and inspired me to give the most out of myself. Also, profesor Eduardo Romero who was always open to discussions and willing to support my research career. This thesis has been a rewarding experience thanks to them. I would like to also thank my parents and siblings for their endless love and support.

I also thank all the people in the Bioingenium Research group, from whom I learned and shared over these years. Thanks to Gloria Díaz and Francisco Gómez for their endless conversations about meaning and impact of research. To Fabio Martínez and Andrea Rueda for their unconditional friendship and support. To Jorge Camargo and Angel Cruz for being available for both, hard work and good fun. To many other students and coworkers I had the opportunity to work and share with (in alphabetical order): Angélica Sandoval, Alejandro Riveros, Carlos Vargas, Cesar Sánchez, Edwin Niño, Jorge Vanegas, José Moreno, Juan Galeano, Laura Arévalo and Raul Ramos. I am very grateful to all of them, they made Universidad Nacional de Colombia a very interesting place to be.

I thank the various sources of funding I have had over the years: Universidad Nacional de Colombia, Centro de Telemedicina, and various Colciencias projects. Final thanks to Ebroul Izquierdo who received me as an associate researcher in Queen Mary, University of London during the winter of 2009. To Sing Bing Kang and Ashish Kapoor for their guidance during my research internship at Microsoft in the summer of 2010. And to Sergey Ioffe who supervised my internship work at Google during the summer of 2011. Working with them has been a great experience to grow and shape my research interests on various image analysis topics.

# Abstract

Image collections today are increasingly larger in size, and they continue to grow constantly. Without the help of image search systems these abundant visual records collected in many different fields and domains may remain unused and inaccessible. Many available image databases often contain complementary modalities, such as attached text resources, which can be used to build an index for querying with keywords. However, sometimes users do not have or do not know the right words to express what they need, and, in addition, keywords do not express all the visual variations that an image may contain. Using example images as queries can be viewed as an alternative in different scenarios such as searching images using a mobile phone with a coupled camera, or supporting medical diagnosis by searching a large medical image collection. Still, matching only visual features between the query and image databases may lead to undesirable results from the user's perspective. These conditions make the process of finding relevant images for a specific information need very challenging, time consuming or even frustrating.

Instead of considering only a single data modality to build image search indexes, the simultaneous use of both, visual and text data modalities, has been suggested. Non-visual information modalities may provide complementary information to enrich the image representation. The goal of this research work is to study the relationships between visual contents and text terms to build useful indexes for image search. A family of algorithms based on matrix factorization are proposed for extracting the multimodal aspects from an image collection. Using this knowledge about how visual features and text terms correlate, a search index is constructed, which can be searched using keywords, example images or combinations of both. Systematic experiments were conducted on different data sets to evaluate the proposed indexing algorithms. The experimental results showed that multimodal indexing is an effective strategy for designing image search systems.

**Keywords:** *Image Databases, Indexing Methods, Image Search, Multimodal Data Analysis, Machine Learning, Pattern Recognition, Matrix Factorization.*

# Resumen

Las colecciones de imágenes hoy en día son muy grandes y crecen constantemente. Sin la ayuda de sistemas para la búsqueda de imágenes esos abundantes registros visuales que han sido recolectados en diferentes areas del conocimiento pueden permanecer aislados sin uso. Muchas bases de datos de imágenes contienen modalidades de datos complementarias, como los recursos textuales que pueden ser utilizados para crear índices de búsqueda. Sin embargo, algunas veces los usuarios no tienen o no saben qué palabras utilizar para encontrar lo que necesitan, y adicionalmente, las palabras clave no expresan todas las variaciones visuales que una imagen puede tener. Utilizar imágenes de ejemplo para expresar la consulta puede ser visto como una alternativa, por ejemplo buscar imágenes con teléfonos móviles, o dar soporte al diagnóstico médico con las imágenes de los pacientes. Aún así, emparejar correctamente las características visuales de la consulta y las imágenes en la base de datos puede llevar a resultados semánticamente incorrectos. Estas condiciones hacen que el proceso de buscar imágenes relevantes para una necesidad de información particular sea una tarea difícil, que consume mucho tiempo o que incluso puede ser frustrante.

En lugar de considerar solo una modalidad de datos para construir índices de búsqueda para imágenes, el uso simultáneo de las modalidades visual y textual ha sido sugerido. Las modalidades no visuales pueden proporcionar información complementaria para enriquecer la representación de las imágenes. El objetivo de este trabajo de investigación es estudiar las relaciones entre los contenidos visuales y los términos textuales, para construir índices de búsqueda útiles. Este trabajo propone una familia de algoritmos basados en factorización de matrices para extraer los aspectos multimodales de una colección de imágenes. Utilizando este conocimiento acerca de cómo las características visuales se correlacionan con los términos textuales, se construye un índice que puede ser consultado con palabras clave, imágenes de ejemplo o por combinaciones de estas dos. Se realizaron experimentos sistemáticos en diferentes conjuntos de datos para evaluar los algoritmos de indexamiento propuestos. Los resultados muestran que el indexamiento multimodal es una estrategia efectiva para diseñar sistemas de búsqueda de imágenes.

**Palabras clave:** *Bases de datos de imágenes, Métodos de indexación, Búsqueda de Imágenes, Análisis de datos multimodal, Aprendizaje de Máquina, Reconocimiento de Patrones, Factorización de Matrices.*

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This thesis addresses the problem of indexing large image collections to provide effective and efficient content-based access. An effective image search system requires the ability to associate high-level concepts to images, in order to discriminate between relevant and non-relevant results. This ability is not natural for computer systems, and this condition poses the research challenge of modeling strategies to bridge the gap between visual signals and abstract interpretations made by humans.

This work proposes a family of algorithms for learning the relationships between visual features and text data, and to make these relationships useful for indexing image collections. The main goal is to construct a mixed content representation of images, determined simultaneously by text descriptions and visual characteristics. In that way, the representation can incorporate high-level, semantic information that may be found attached to images as well as low-level, appearance information extracted directly from images. This work proposes strategies to merge the two data sources for building multimodal representations, with the intention of improving the quality of the results in an image search system.

## 1.1 Motivation

Image search systems are becoming a pervasive technology in every field in which visual information is used to support processes and decisions, including arts, forensics, medicine, military and scientific applications, among others. Even for personal photo collections and social networks, simple image search engines are available. And the main reason for this is that image databases are becoming massive and keep growing constantly. The social web hosts billions of images from all around the world, while specialized fields all together collect millions of images every day.

That volume of information cannot be analyzed thoroughly and regularly by people as it is updated, so lots of potentially useful records may remain inaccessible without a proper search engine. Depending on the kind of information required, users may need to query for specific pictures using keywords or example images. The latter may be

more appropriate for some domains in which available records are at hand, or just for exploring the web using camera phones.

The main challenge of an image search system is to identify and retrieve from the database the most relevant images for a given query. There are several difficulties that make this goal a challenging problem. Searching for images using keywords can be hard for users if they do not have the right terms in mind. Besides, real world databases do not have clean text descriptions for every image, and some of the available annotations may be incomplete, imprecise or noisy. On the other hand, when users search for images using example pictures, correctly matching visual contents may be hard because of the *semantic gap*.

The semantic gap is known as the inability of computers to understand pictures as naturally as people do. Human beings can extract and recognize objects in very complex scenes, and can even associate these objects to generic concepts and abstract ideas. Understanding pictures is a task that people can do effortlessly to discriminate useful material when they are searching for something. The ability to associate visual structures with abstract ideas would be very helpful for a search system to find the right images for complex queries, yet, this is something that cannot be explicitly implemented using conventional programming notions.

The fundamental assumption of this dissertation is that text resources can provide information closer to high-level conceptual interpretations made by humans about visual information. The visual data modality is usually associated to a low-level appearance model for images, while the text data modality usually describes symbolic representations or high-level concepts present in the associated images. In that sense, the problem of understanding images from a computational perspective, can be approached by modeling the relationships between images and texts.

The algorithms and strategies proposed in this work are oriented to building an image representation that simultaneously contain information extracted from the visual and textual modalities. This is called a multimodal representation. Then, the main hypothesis is that by combining the two data modalities, a better response can be obtained from an image search system. The core research problem approached in this work is learning the correspondences between both data modalities and making them useful to index an image collection as well as to represent and interpret incoming queries.

## 1.2 Problem Statement

The problem studied in this dissertation is the design of effective and efficient strategies for image indexing and search. In particular, this work focuses on learning multimodal relationships between visual and text data. The main research question of this work is: how to automatically identify associations between images and text to improve the response of an image retrieval system, using the richness of visual data and the semantics of text annotations?.

### 1.2.1 Learning Multimodal Aspects

Images and text descriptions are not paired randomly. We can assume that images have been placed in a document together with their text description by a writer who wanted to express an idea. In that case, the image is an illustration of the idea and the text provides its narrative from an abstract but precise perspective. A similar situation happens with medical images, when physicians provide a diagnosis and attach a visual signal as complementary evidence. Even in photo-sharing web sites or social networks, some tags are assigned to pictures to indicate context, emotions, places and specific objects.

Without loss of generality, we can assume a joint probability distribution that generates pairs of images and texts following the same patterns of a human writer. From this distribution, an image of a *tall building* is more likely to be observed with words such as *city*, *office*, and *hotel*, and less likely to be observed with words such as *animal*, *forest*, and *waterfall*. This work assumes that the structure of this distribution can be learned from data, more specifically, using machine learning algorithms on a collection of image-text pairs. A potential solution has three main components: a model of the relationships between both modalities, an algorithm to learn the unknown parameters of the model, and an inference or prediction function to compute the multimodal representation.

The model of the relationships between visual and text contents may be designed using supervised learning, which allows to define classification functions that predict whether observed visual features are correlated with text terms. This strategy is able to model the relationships between the two data modalities, and well established algorithms for learning their parameters may be used, such as Support Vector Machines. However, this offers part of the solution only, since the representation obtained with classification functions is primarily semantic, i.e., the goal is set to predict text terms instead of building a multimodal representation.

The key point for modeling multimodal aspects is to note that visual contents and text descriptions are both partial views of common underlying concepts. Images are signals with abundant visual details including colors, textures, shapes and variations or combinations of them. Texts are composed of symbols including letters, words and sentences, usually associated to high-level ideas or interpretations. Both are complementary information of an observed phenomenon, which is not explicitly codified in any of the two data modalities, but instead, is hidden or latent in their relationships. For instance, the term *car*, is a symbolic representation of an abstract object, and only when visual features are added, such as colors and shapes, the object starts to become a concrete instance that can be observed in the real world.

Therefore, we can assume the existence of several hidden aspects that are described by both data modalities together, which are called multimodal aspects or multimodal factors in this work. Multimodal aspects are by definition meaningful relationships between some visual features and some text terms, and the problem of automatically discovering them is at the core of this research work. Unsupervised learning strategies,

such as probabilistic graphical models or matrix factorization algorithms are powerful tools to model and learn latent variables. However, these methodologies are usually oriented to operate on a single data modality, a condition that may limit their use on a multimodal setup. The main research challenge of this thesis is to extend the notions of latent factor analysis to the problem of learning multimodal aspects.

Finally, the problem of building multimodal representations consists on determining the extent to which an image exhibit each of the multimodal aspects. This can be understood as measuring the level of expression of each multimodal aspect in an image, and since these measurements unveil the underlying multimodal structure of an image, this can be used to represent simultaneously its visual characteristics and high-level semantic contents. This representation may be used in an image search system to compare multimodal contents and locate relevant images for a given query.

### 1.2.2 Research Challenges

The main challenges to realize a model for multimodal learning, as was described in the previous subsection, are associated to the complexity of image collections. In this work, three variables that define the complexity of an image database have been identified: application domain, structure of the text modality, and database size. These variables and their relationships are depicted in Figure 1.1, which also illustrates datasets used through this research work[1].

The Figure illustrates four regions that represent different degrees of complexity associated with indexing and searching multimodal image collections. The larger the database and the less structured the text modality, the more complex are the tasks of indexing and searching the collection.. Small image databases with predefined categories and labels may be easier to model, since more prior information can be introduced to exploit known regularities. The same can be said for an image retrieval system in a specific domain; more assumptions about image contents and semantic distributions can be hold when the domain is narrow.

The first important challenge of this work is to propose methods for learning multimodal aspects that can be applied under realistic and natural conditions, as those that can be found in real world databases. The evolution of algorithms and methods proposed in this work follow the path of the blue arrow depicted in Figure 1.1, which moves from small databases with controlled text data, toward large image collections with natural language descriptions. The latter is a situation harder to deal with, but more faithfully represents practical conditions.

Notice that the arrow does not cross the region of large databases with categorical or multi-label text descriptions. This condition is not considered realistic, as collecting clean labels for large amounts of images is a very costly effort. Nevertheless, research initiatives have been started recently to investigate this collections, such as the ImageNet[2] initiative, which is depicted in the Figure as reference only.

[1] Except for the ImageNet data set, which is illustrated for reference only

[2] http://www.imagenet.org

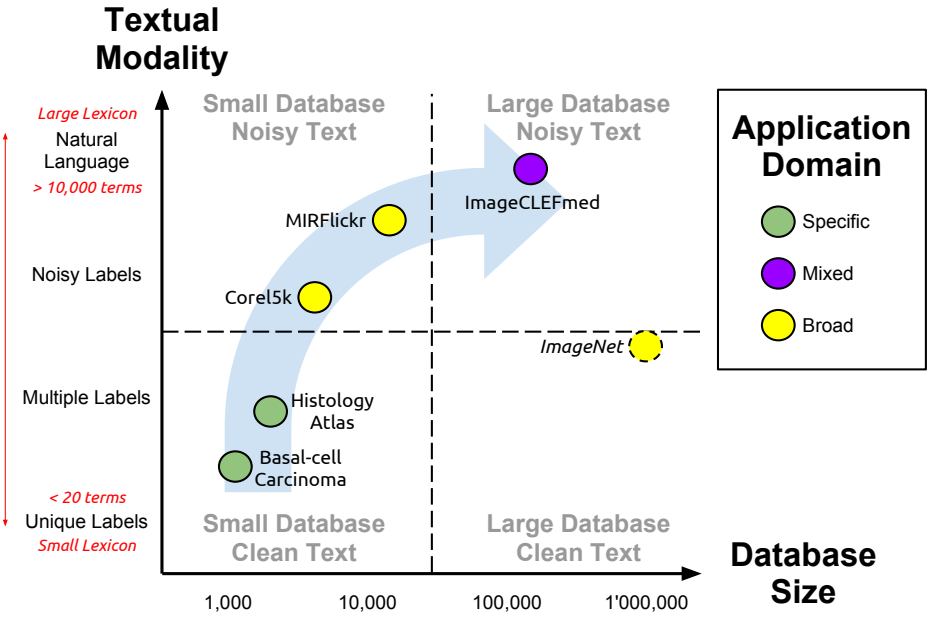



Figure 1.1: Illustration of the complexity of image databases.

#### 1.2.2.1 Textual Modality

A common step in image analysis is the definition of categories or labels for images, which is a process also known as manual annotation. These labels determine in a summarized fashion the understanding that humans may have of visual contents in images, and are very useful for recognition and categorization tasks. For broad application domains that involve natural scenes or web photos, labels may be collected for large image collections using crowdsourcing services, such as the *Amazon Mechanical Turk*[3], as has been done for the ImageNet collection [1]. However, for specific application domains that require specialized knowledge to evaluate and analyze images, such as the medical field, this may be much more expensive.

Collecting clean and controlled labels for large databases is, in general, not practical for building image search systems. Instead, text resources naturally attached to images may be easily found in different image databases. Figures in books and scholarly papers come with captions. Newspapers and web pages include images with context descriptions. Medical images are associated to health records, among other examples. Even though these texts are not as clean as manually assigned labels, their contents are likely associated to semantic or high-level descriptions for images. The main difficulty of using these text resources is extracting their underlying semantics.

Raw text contents also require normalization and appropriate representation. The Information Retrieval (IR) and Natural Language Processing (NLP) communities have established very well studied models for representing text documents, the most promi-

[3]http://aws.amazon.com/mturk/

nent being the Vector Space Model [2]. Determining a suitable representation for text contents in an image collection depends on the structure of these texts, as they can be very noisy, such as tags assigned by web users, semi-structured, such as captions and tittles extracted from scholarly articles, or very clean and well defined, such as categories and semantic keywords. That structure may require different weighting schemes to highlight important terms and diminish redundant or noisy words.

Real world image databases may have abundant unstructured text resources that may be exploited for understanding visual contents, and this represents a very challenging problem.

#### 1.2.2.2 Database Size

Another challenge for designing multimodal search systems is the ability of the algorithms to process large volumes of data, more specifically, the ability to extract multimodal relationships from large image collections in a reasonable amount of time. This is a challenging computational problem that can make the difference between a feasible solution and an impractical one.

The number of available images in a real world database may be very large. Actually, if the database is small, perhaps a search system is not needed since people may scan images quickly and filter out those that are not useful. So, when an image search system is required, it means that the number of archived records is beyond the ability of a person to keep track of their contents. Without a system that indexes all their records, and without the mechanisms to query and access specific images, these data will remain just archived and unused. And if an image is required, the cost of finding it might be very high.

A suitable method for indexing a large collection of images requires to extract multimodal relationships from massive amounts of examples to end up on a comprehensive model. Modern image retrieval benchmarks used for research currently include hundreds of thousands or even millions of images to reflect the importance of this problem. And even though researchers can rely on the assumption that computers get faster or algorithms can be parallelized, there are also smart ways to employ computational resources to achieve the goal of processing large amounts of data that can be extended to a multimodal setup.

#### 1.2.2.3 Application Domain

The application domain poses the interesting question of how to extract and represent pure visual contents from images, which has been widely studied in a large variety of fields. The difficulty of modeling visual contents is inversely proportional to the specificity of the application domain. Specific and narrow application domains allow to incorporate precise and discriminative features, which can involve prior knowledge about the patterns that have to be highlighted in an image. Broad domains may require generic features that account for a wider range of visual transformations.

When the structural variability of visual contents in a collection of images is very narrow, simple transformations and normalizations can be applied to align and match pixels directly, as may be the case of face recognition [3]. For domains with larger variability of visual contents, feature extraction usually helps to balance for deformations, rotations, translations and scaling of target objects in images. Under these circumstances, representations based on parts of objects have been investigated recently [4, 5].

Obtaining a good representation for image contents is an active area of research and it directly impacts the performance of any task involving visual signals. Adopting an appropriate visual representation is one of the main challenges faced in this work.

## 1.3 Contributions

This work presents several contributions to solve each of the problems described above for learning multimodal image representations. They are described briefly in the following subsections, and references to published works are provided.

### 1.3.1 Algorithms for Multimodal Learning

The main contributions of this work are the models associated to learning multimodal relationships between images and texts. The following are the models and approaches studied in this research.

#### *Supervised Learning for Image Auto-Annotation*

The first proposed method for modeling the relationships between images and texts was based on supervised learning. This model assumes that each of the terms in the text vocabulary may be related to visual features using a classification function. This work was published in:

- **Caicedo J.C.**, González F.A., Romero E. *Content-based histopathology image retrieval using a kernel-based semantic annotation framework.* Journal of Biomedical Informatics 44(4): 519-528. 2011
- **Caicedo J.C.**, González F.A., Romero E. *A Semantic Content-Based Retrieval Method for Histopathology Images.* Information Retrieval Technology. AIRS 2008, LNCS 4993, pp. 51–60, 2008
- **Caicedo J.C.**, González F.A., Romero E., Triana E. *Design Of A Medical Image Database With Content-Based Retrieval Capabilities.* Advances in Image and Video Technology, PSIVT 2007, LNCS 4872, pp. 919-931, 2007

### *Latent Factors for Learning Multimodal Relationships*

Two novel unsupervised learning algorithms were proposed to model the relationships between text terms and visual features using latent factors. These algorithms are based on Nonnegative Matrix Factorization, and are known as *NMF-Asymmetric* and *NMF-Mixed*. This work was published in:

- **Caicedo J.C.**, Ben-Abdallah J., González F.A., Nasraoui O. *Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization*. Neurocomputing 76(1): 50-60. 2012
- González F.G., **Caicedo J.C.**, Nasraoui O. and Ben-Abdallah J. *NMF-based Multimodal Image Indexing for Querying by Visual Example*. ACM CIVR 2010
- Ben-Abdallah J., **Caicedo J.C.**, González F.A., Nasraoui O. *Multimodal Image Annotation Using Non-negative Matrix Factorization*. IEEE/WIC/ACM International Conference on Web Intelligence, 2010, 128-135

### *Latent Semantic Kernels*

Following the ideas of latent factors to find multimodal relationships, an algorithm that represents images and texts using kernel functions was proposed. This algorithm has the advantage of exploiting the specific structure of data in the feature space, from where the latent factors are extracted. This work was published in:

- **Caicedo J.C.**, Moreno J.G., Niño E.A. and González F. *Combining Visual Features and Text Data for Medical Image Retrieval Using Latent Semantic Kernels*. ACM MIR 2010

### *Semantic Embeddings*

An extension of the proposed NMF-based algorithms was introduced, named the *Non-negative Semantic Embedding*. This work models the relationships between visual features and text terms directly, instead of using latent factors. This demonstrated to be a useful approach when the textual modality is clean and structured. Parts of this work have been published in:

- Vanegas J., **Caicedo J.C.**, González F.A., Romero E. *Histology Image Indexing Using a Non-negative Semantic Embedding.* Workshop on Medical Content-Based Retrieval for Clinical Decision Support. MICCAI 2011.

Additional extensions and other parts of this work are included in a submission to a Journal:

- **Caicedo J.C.**, Vanegas J., González F.A. *Histology Image Search Using Multimodal Fusion*. IEEE Transactions in Medical Imaging. *Submitted*. 2012

#### *Backprojection of Semantic Information*

An additional concept introduced in this work is the fusion by backprojecting semantic information to the visual space. Since the proposed algorithms for latent factors and semantic embedding are based on matrix factorization and subspace modeling, functions to project data to an alternative space can be formulated. Parts of this work were published in:

- **Caicedo J.C.**, González F.A. *Image Retrieval Using Multimodal Fusion based on Matrix Factorization.* International Conference on Multimedia Retrieval. ICMR 2012

Also, this extension of multimodal fusion has been included in the journal submission mentioned above: *Histology Image Search Using Multimodal Fusion.*

### 1.3.2 Large Scale Learning

Learning multimodal relationships may require the analysis of very large image databases. An extension of the matrix factorization algorithms has been proposed to efficiently deal with big data sets. The proposed approach is based on online learning principles and provides an efficient implementation with little memory requirements and drastically reduced execution times using a single CPU. Parts of this work have been published in:

- **Caicedo J.C.**, González F.A. *Online Matrix Factorization for Multimodal Image Retrieval.* 17th Iberoamerican Congress on Pattern Recognition. CIARP 2012

An extended version of this work is also being prepared for a Journal submission:

- **Caicedo J.C.**, González F.A. *Large Scale Multimodal Image Indexing via Online Matrix Factorization.* Journal of Multimedia Information Retrieval. *To be submitted.* 2012

### 1.3.3 Modeling Visual Contents

Each application domain requires its own content representation, specially if it corresponds to a particular field such as histology images. In this work several image representations were proposed mainly for medical image analysis.

#### *Bag-of-features for Histology Images*

Parts of the multimodal analysis described in this work for histology images has been conducted on top of a bag-of-features representation. Histology images are particularly textured due to the structure of tissues and arrangements of cells, which define visual patterns that may be organized in a dictionary. Early research was developed in the frame of this thesis to evaluate the performance of bag-of-features for histology images. These works where published in:

- **Caicedo J.C.**, Cruz-Roa A., and González F. *Histopathology Image Classification Using Bag of Features and Kernel Functions.* AIME 2009, vol. LNAI 5651, pp. 126135, 2009.

- **Caicedo J.C.**, Camargo J.E. and González F.A., *Content-based Access to Medical Image Collections.* Biomedical Image Analysis and Machine Learning Technologies: Applications and Techniques, IGI Global, 2009.

#### *Kernel-based Image Representations*

Kernel functions provide a natural framework to combine features in several ways, and this is specially useful for image analysis when multiple features can be computed, such as colors, textures, and edges. Alternative strategies using combinations of kernel functions were also explored, including latent semantic analysis and normalized combinations. These works have been published in:

- Moreno J.G., **Caicedo J.C.**, González F.A. *A Kernel-based Multi-feature Image Representation for Histopathology Image Classification.* Acta Biológica Colombiana 15 (3), 251-260, 2010.

- **Caicedo J.C.** and Izquierdo E. *Combining Low-level Features for Improved Classification and Retrieval of Histology Images.* International Conference on Mass Data Analysis of Images and Signals in Medicine, 2010

### 1.3.4 Applications and Other Contributions

Several applications of the work presented in this dissertation have been explored in the frame of other research projects and peer collaboration. The following are some highlights:

#### *A Live Histology Image Retrieval System*

Parts of this work were implemented and tested in a project for histology image collection management. The system can be found online at http://www.informed.unal.edu.co and all its features, which include, but are not limited to content-based image search, are discussed by González et al.:

- González F.A, **Caicedo J.C.**, Cruz-Roa A., Camargo J.E., Romero E., Spinel C., Seligmann D., Forero J. *A Web-Based System for Biomedical Image Storage, Annotation, Content-Based Retrieval and Exploration.* Microsoft eScience, 2009.

#### *Exploratory Image Search*

The proposed methods to generate a multimodal representation for images have been also used in other projects to investigate interaction mechanisms between users and an image collection. Camargo et al. discuss their potential application in this context:

- Camargo J., **Caicedo J.C.**, and González F.A. *Multimodal Visualization Based On Nonnegative Matrix Factorization.* ECDL 2010. LNCS, Volume 6273/2010, 429-432.
- Camargo J., **Caicedo J.C.**, Chavarro A. and González F.A. *Kernel-Based Strategy for Exploratory Image Collection Search.* IEEE CBMI 2010

### *Visual Pattern Mining*

Parts of the proposed approaches were also introduced to explore applications in visual pattern mining for histology images. The main purpose in this work was to thoroughly analyze relationships between visual patterns and categorical labels by introducing a layer of interpretation. Cruz-Roa et al. discuss the main findings:

- Cruz-Roa A, **Caicedo J.C.**, González F.A. *Visual Pattern Mining in Histology Image Collections Using Bag of Features.* Journal Artificial Intelligence in Medicine. Vol 52, pp. 91-106. 2011
- Cruz-Roa A., **Caicedo, J.C.**, and González F. *Visual Pattern Analysis in Histopathology Images Using Bag of Features.* CIARP 2009, vol. LNCS 5856, pp. 521-528, 2009.

### *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, specially focusing on the search task using visual examples. In the 2007 version, the strategy was mainly based on low-level visual features and supervised learning oriented to predict image modality. Moreno et al. discuss the main results obtained in the 2010 version, introducing a multimodal latent factors space. These works have been published in:

- Moreno J.G., **Caicedo J.C.**, González F.G. *Bioingenium at ImageCLEFmed 2010: A Latent Semantic Approach.* Cross Language Image Retrieval, ImageCLEF 2010
- **Caicedo J.C**, González F.A. and Romero E. *Content-based Medical Image Retrieval Using Low-level Visual Features and Modality Identification.* In proc. CLEF2007, LNCS 5152, pp. 615–622, 2008

Additional 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[4], and a conference paper is currently

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

being prepared to report these findings.

## 1.4 Thesis Organization

The remaining chapters of the thesis are organized as follows:

- *Chapter 2: A review of Image Search Methodologies.* This chapter discusses previous works related to image search systems, providing a review of three main approaches followed by the research work done in content-based image retrieval. This overview allows to understand why multimodal indexing strategies are an interesting research direction for building image search systems.

- *Chapter 3: Content-Based Histopathology Image Retrieval Using a Kernel-based Semantic Annotation Framework.* This chapter addresses the problem of building a semantic image representation for image search, using supervised learning behind an automatic annotation system. This model explores the relationships between visual features and semantic annotations using classification functions, and focuses on constructing an enhanced visual representation by combining various features to improve the performance of classifiers. The research presented in this chapter approaches the problems of visual representations for specialized imaging domains, and proposes a supervised model to learn multimodal relationships for clean and structured text labels.

- *Chapter 4: Multimodal Representation, Indexing, Automated Annotation and Retrieval of Image Collections via Non-negative Matrix Factorization.* A framework for modeling multimodal latent factors is presented in this Chapter, considering the problem of image collections with unstructured text contents. The proposed framework builds a latent factors space that represent simultaneously visual contents and text annotations. Various image search tasks are evaluated under this framework, including image retrieval with various query paradigms and automatic image annotation.

- *Chapter 5: Histology Image Search Using Multimodal Fusion.* An extension of the latent factors model is presented in this Chapter, to consider the problem of finding multimodal relationships in collections with clean and structured textual resources. This extension is also based on matrix factorization, and learns the relationships between visual and text terms directly, instead of using a latent factors space. To accomplish multimodal fusion, query images are first mapped to the text space and projected back to the visual space, then both the original visual representation and the backprojected representation are combined in a fused multimodal representation.

- *Chapter 6: Large Scale Multimodal Image Indexing via Online Matrix Factorization.* This Chapter presents an extension of the latent factor model to learn

from large amounts of examples, following online learning principles. The chapter studies computational issues of a matrix factorization algorithm for multimodal learning, and proposes an efficient model to extract meaningful relationships from very large image collections. The proposed model exhibits fast convergence rates and improved retrieval performance.

- *Chapter 7: Conclusions.* The final chapter presents the main conclusions and discussions of the dissertation, summarizes the main contributions, and highlights the most important findings. Also, some future research directions are presented and discussed.

# Chapter 2

# A Review of Image Search Methodologies

## 2.1 Content-Based Image Retrieval

In the very early stage of computerized image databases, the process of organizing pictures for office and home usage was completely made by hand. Some commercial systems provided image collections with thousands of records, carefully indexed by categories and using several keywords. Examples of such collections were the ClipArt Gallery of Microsoft and the Corel database. Due to the error prone process and immense effort required to organize an image database by hand, researchers began to work on alternative systems that could automatically analyze visual contents to search for images.

The QBIC [6] and the Photobook [7] are some examples of the first steps made toward automated manipulation of image contents to support search tasks. The basic idea underneath these models is to compute visual characteristics directly from images, and then evaluate a similarity measure between them. A large body of research was developed to improve the capabilities of image retrieval systems under that model, exploring various strategies to represent image contents and modeling specialized similarity measures. Riu et al. [8] and Veltkamp et al. [9] provide comprehensive surveys on the methodologies and techniques used during that period, which may be understood as the infancy of image retrieval research [10].

All the focus was given to extract and characterize visual properties for images, which was then understood as a limitation for providing practical image search systems. Smeulders et al. [11] had a clear influence on the subsequent progress made on the field by defining the problem of the semantic gap:

> "*The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation.*"

This implies that human interpretations are not evident from the signal representa-

tion of an image, and that other sources of knowledge should be considered to design computational models that can react as users expect these systems to do.

## 2.2 Semantic Image Retrieval

Instead of indexing images using visual contents, a lot of work was turned to index images using keywords and text annotations. The main assumption is that building an image index with the right keywords may help users to retrieve images that truly represent what they are searching for. The main research problem now is how to select and assign keywords for images automatically. Text annotations are not cleanly or systematically assigned to images, and many images can have no attached text.

The computer vision community influenced the design of mechanisms to build semantic image indexes, by bringing methods to predict words associated to images. If a system can automatically generate keywords according to what can be seen on images, all images could be indexed and searched. A coocurrence model for image parts and words was proposed by Mori et al. [12] to achieve that goal. Later, Duygulu et al. [13] proposed an auto-annotation model for images based on machine translation principles, assuming that image segments are terms in one language and words are terms in another language.

Many other works have investigated the problem of automatic image annotation which is considered to be one of the most challenging and important problems in computer vision [14, 15, 16, 17, 18, 19, 20]. Despite all this effort, the problem is still considered to be unsolved and the applicability of these methods in real world systems is debatable [21].

But besides the problem of accuracy of image auto-annotation algorithms, which most surely will continue to increase, there is a pronounced limitation of keywords to entirely communicate what images do along all their visual details. Yeh et al. [22] discuss that an image is worth a thousand words to interact with visual search services, a concept studied simultaneously by Fan et al. [23]. Later, Rasiwasia et al. [24] proposed the query-by-semantic example (QSE) that combines an image auto-annotation model with the visual example paradigm.

Thus, building a useful image search system is not only a problem of generating accurate keywords for images. Pictures can have several semantic interpretations, and summarizing their visual contents on a few words restricts the possibilities of finding useful information. An effective image search system should consider both, visual impressions of images as well as their potential semantic interpretations.

## 2.3 Multimodal Fusion

Atrey et al. [25] presents a comprehensive survey on the field of multimodal fusion for a variety of multimedia applications. They refer to the problem of multimodal fusion as the integration of multiple signals to perform semantic decisions. In the specific

domain of image retrieval, the available data modalities are visual features and text annotations, whereas the semantic decision may be understood as determining whether an image in the database is relevant to the user's query.

Smeulders et al. [11], Lew et al. [26] and Datta et al. [10] have suggested in different review papers published during the last decade, that multimodal fusion is a promising research direction to approach the problem of image search. In particular, they emphasize that combining visual information with text resources attached to images, may result in useful search indexes. Increasing interest has been shown on this topic, and several contributions model different aspects of the problem.

One of the earliest attempts to combine both data modalities for building an image index was conducted by La Cascia et al. [27] using a database of web images and text extracted from the corresponding HTML page. They reported improvements on the precision of the search task as well as new ways to allow users to interact with an image collection. Other models for multimodal analysis on image collections have been more oriented to approach the automatic image annotation problem [15, 28].

Recently, several multimodal image retrieval models have been proposed in the literature, from which it is important to distinguish between two general approaches: late multimodal fusion and early multimodal fusion. The former refers to the combination of both data modalities after each has been processed separately. Examples of late fusion approaches include linear combinations of similarity scores [29], and combinations of decisions made by different classifiers [30]. Early fusion, on the other hand, is oriented to combine multimodal data before its subsequent use. Examples of early fusion include latent semantic indexing [31] and concatenations of feature vectors [32].

One of the practical differences in the way researchers approach the problems of image auto-annotation and multimodal fusion is on the process of collecting the target text resources. Image auto-annotation models are usually oriented to gather clean and very structured keywords associated to categories [1]. Multimodal fusion usually deals with very unstructured text sources, which describe ideas in natural language sentences or noisy, uncontrolled tags [33]. The reason multimodal fusion has been considered a promising research direction is because it is relatively cheap to collect large databases of images with unstructured text descriptions, whereas building structured databases is very costly. However, the former is definitely a harder condition to take advantage of.

There is no clear consensus in the literature about how multimodal fusion should be modeled for image indexing, and which design has the potential to fully exploit the multimodal interactions in image collections. The goal of this dissertation is to contribute to the better understanding of the problem and to propose solutions to overcome its limitations.

# Chapter 3

# Content-Based Histopathology Image Retrieval Using a Kernel-based Semantic Annotation Framework

*This work has been published in the Journal of Biomedical Informatics [34].*

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

## 3.1 Introduction

The use of digital imaging in histopathology has been rapidly rising during the last few years [35]. Pathology departments using digital microscopy equipments can share slides without sending the glass, and can include images in electronic reports and publications [36]. Then, large amounts of digital histopathology images are constantly acquired

as part of the routine operation in these specialized centers. Image collections are stored using information technologies such as Picture Archiving and Communication Systems (PACS), but they remain archived in the long term, basically because after some time they result useless, since the main actual exploit is the one associated to the specific clinical case. Nevertheless, these large image collections are a potential source of information and knowledge, which may support educational activities, research studies and even the clinical decision making process itself, if the right tools to access these collections are developed [37].

Accessing a collection of histology images can be done using different query paradigms. For instance, using structured queries in conventional databases, using keywords in a text retrieval engine or using example images in a content-based image retrieval system [10]. In this Chapter, we consider the problem of retrieving histopathology images from the collection using example images as queries, that is, the user presents reference images to the system, and the system uses the visual content to match similar images in the collection. Using visual contents to search for images is considered a beneficial technology for next generation medical imaging systems [37], and is also considered one of the major challenges in image retrieval research. The problem underneath an image retrieval system is the mechanism for identifying relevant images, which is mostly a similarity measure between image contents.

Different similarity measures have been proposed and studied for medical image retrieval using low-level features, which focus mainly on characterizing visual properties that can be computed from pixels [38, 39]. However, bridging the semantic gap [11] has become the main focus of image retrieval research, i.e. reducing the discrepancy between the information extracted by low-level features and the high level interpretations of human beings on the same images. This research has led to semantic representations of histology images to address the problem of identifying semantically related images rather than just visually similar images [40]. These strategies aim to provide better search results which are more likely to match contents in the same way as physicians would do.

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

The proposed strategy has been implemented and evaluated using a large database

of real histopathology images, extracted from medical records of a pathology lab. An extensive validation was conducted using the ground truth provided by pathologists. The experimental evaluation showed that the semantic image annotation leads to an average improvement in the retrieval response of 57% when it is compared to visual search using only low-level features. Also, the results show that a multi-feature representation for visual contents can be progressively improved by operating kernel functions. We found that modeling feature structure and non-linear patterns with kernel functions is more likely to improve the discriminative power of multi-feature representation spaces. The contents of this Chapter are organized as follows: Section 3.2 presents a review of previous works related to histology image retrieval. Section 3.3 introduces the collection of histopathology images used in this study. Section 3.4 describes the proposed methods for automatic image annotation, based on kernel methods. The experimental setup and results are presented in Section 3.5. Finally, Section 3.6 presents the discussions and Section 3.7 presents the concluding remarks and future work.

## 3.2 Related Work

Digital microscopy is a very broad field of active research, that ranges from image acquisition and compression [41] to automatic disease detection [42]. Content-based retrieval of microscopy and histology images is one of these areas of research that is receiving increasing attention from researchers. Image retrieval focuses on methods and tools for managing large collections of digital slides, providing effective access to all the available information, in contrast to other research areas that focus on processing individual images for making automatic decisions, such as automated grading [43, 44] or tissue classification [45, 46].

Image retrieval on pathology image collections was approached by Zheng et al. [47] using low-level features. Four different visual features were studied to measure the discriminative power of similarity measures to correctly identify relevant images given an example query. They reported a correlation between the computed similarity and pathological significance on the tested collection, without the use of domain knowledge. However, to scale up the system performance, low-level features may be insufficient. Tang et al. [40] investigated the role of semantic information to represent local image content in gastro-intestinal tissue images. Their method aim to assign a semantic annotation to each region on the image using machine learning algorithms. The main disadvantage of this approach is the need for manual annotations made on specific regions for a large enough sample of training images. Naik et al. [48] also approached the problem of histology image retrieval using semantic knowledge. They used multiple texture and architectural features of tissues, and employed a boosting algorithm to identify feature weights that maximizes retrieval and classification performance.

In this Chapter we address the problem of semantic image retrieval for histopathology images, using an automatic annotation strategy. Our previous work on histopathology image retrieval [49] showed the potential of using semantic features to represent

Table 3.1: Histopathology concepts and the corresponding number of examples in the data set

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

image contents. In this Chapter we extend that work by generalizing the representation of visual contents in a set of multiple heterogeneous features rather than a unique feature vector. Also, we recast the image annotation problem in terms of kernel methods for image representation, feature selection and concept detection, as is presented in the following Sections.

## 3.3 Basal-cell Carcinoma Images

Images in this work have been used to diagnose a special kind of skin cancer known as basal-cell carcinoma. Basal-cell carcinoma is the most common skin disease in white populations and its incidence is growing world wide [50]. The histopathology collection is composed of 1,502 images at 1,280×1,024 pixels, acquired under a Nikon microscope and stored in lossless JPG format.

The collection was studied and annotated by a pathologist to describe its contents, elaborating a data set with images and descriptions of their related concepts. Table 3.1 shows the list of 18 concepts and the number of available examples in the collection. One image may contain several concepts, that is, different biological structures are exhibited in one single image. Notice that Table 3.1 lists the number of images per

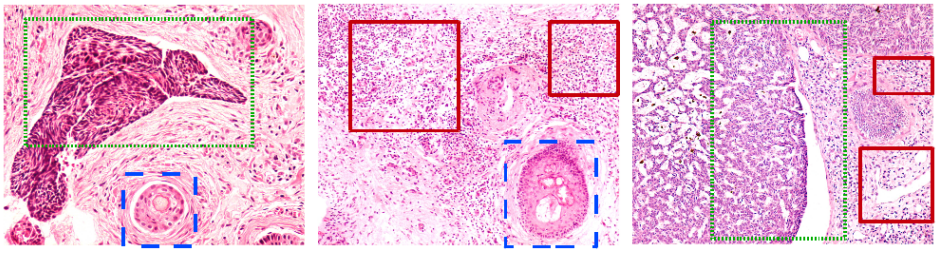


Figure 3.1: Three histopathology image examples with some highlighted biological structures and pathological patterns. Dashed lines (blue) show pilocebaseus units, a normal biological structure in the skin. Dotted lines (green) show example regions of Nodule, Palisading cells and Clefts (NPC), a clear evidence of basal-cell carcinoma. Continuous lines (red) show regions with another clue to detect basal-cell carcinoma: lymphocyte infiltration.

concept, but not their co-occurrences. The total number of annotated images in the collection is about 900 corresponding to pathological cases, while the remaining 600 are images with normal skin tissue. This data set also shows a high imbalance between the number of examples exhibiting a concept and the rest of the collection.

The concept list also includes some structures that are not pathological such as *pilosebaceous units*, *eccrine glands* and *blood vessel*s. One of the histopathology concepts reported in Table 3.1 is N-P-C, which is a convention for *Nodule, Palisading cells and Clefts* (N-P-C), which is a typical sign of basal cell carcinoma, not by the presence of any of them individually but by the manifestation of all three visual patterns together.

Figure 3.1 shows examples of histopathology images with some specific regions in which the concepts can be observed. These examples show that one image can have more than one interesting pattern for pathologists. In addition, the Figure shows that normal biological structures have well-defined visual configurations, in contrast to pathological patterns that may appear with different visual variabilities. In particular, note that the dotted lines (green) cover a portion of nodule that contrast with epithelial tissue with a cleft in the middle. However, the nodule structure in both cases looks different and also the contrasting tissue in the other side of the cleft. These are some examples of the variabilities that may be found in real histopathology images.

This data set was divided up into training (75%) and test (25%) sets, using stratified sampling as is shown in the table. The annotations provided by the pathologists are useful to automatically validate whether search results are relevant to the user information needs. This image collection has been previously used to test two different retrieval strategies: one that uses only visual similarity [51] and another that uses a semantic representation approach based on SVM classifiers with basic kernels [49].

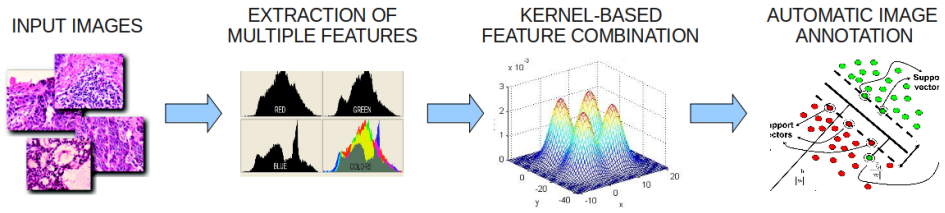



Figure 3.2: Overview of the main steps of the proposed strategy for automatic image annotation

## 3.4 Semantic Image Annotation and Retrieval

The proposed strategy for histology image retrieval is oriented to produce a set of semantic image annotations through visual content analysis. Our framework aims to build a general and complete visual representation of images that can provide enough evidence of the presence or absence of certain histopathology concepts. Figure 3.2 shows the three fundamental steps in our framework: first, the extraction of multiple visual features is performed on the input images. Second, the new content representation is build integrating all visual features using kernel functions. Third, this content representation is used to detect histopathology concepts. After generating automatic annotations, the result can be used to search images with similar annotations or just to index the input images in the retrieval system.

### 3.4.1 Image features

Feature extraction is an important task for image analysis and understanding and there are different approaches to address this problem [52]. Global features for characterizing whole scenes have been proposed using color histograms [53] and MPEG7 features [54]. Likewise, global descriptors such as textures and down-scale representations have been evaluated in medical imaging [55]. One important advantage of using a global image description strategy is that it is unnecessary to specify a model for objects or regions that images may contain. On the contrary, global features provide a holistic image representation that characterizes the composition of the whole image.

We modeled histopathology images as a set of global histogram features, taking into account that pathology patterns may have high visual variabilities that are characterized by different feature sets. For instance, as it is shown in Figure 3.1, the NPC pattern, highlighted in dotted lines (green), contains a mixture of the textures inside the nodule, the edges provided by the cleft, and the density of the palisading cells. To describe the different visual characteristics of histopathology patterns, seven feature spaces have been selected: gray scale histogram, invariant feature histogram [56], local binary patterns [57], RGB color histogram [56], bag of SIFT features, Sobel histogram [58] and Tamura texture histogram [59].

These seven low-level features are complementary with respect to the kind of mea-

sure they do over image pixels, since they apply different computations to build the histograms. However, some of them measure similar visual properties on images, such as Tamura texture and Local Binary Patterns, both modeling texture patterns, but using different approaches (statistical and deterministic, respectively). Also, the invariant feature histogram and SIFT features are intended to identify characteristics that are invariant to rotations and translations. The invariant feature histogram follows an integral approach that sums globally over rotation and create a histogram over translation [56]. On the other hand, the bag of SIFT features is based on a learned dictionary of rotation invariant visual patterns that are counted in each image to construct a histogram of frequencies [60].

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

### 3.4.2 Kernel functions

Kernel methods are an alternative family of algorithms and strategies to perform machine learning [61]. One of the main distinctive characteristics of kernel methods is that they do not emphasize the representation of objects as feature vectors. Instead, objects are characterized implicitly by kernel functions that measure the similarity between two objects. A kernel function induces an implicit high-dimensional feature space where, in principle, it is easier to find patterns.

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

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

$$k_{\cap}(A, B) = \sum_{i=0}^{m} min\,(a_i, b_i) \tag{3.1}$$

where $A = (a_1...a_n)$ and $B = (b_1...b_n)$ are histograms. This similarity measure has been shown to satisfy the Mercer's properties [62]. This is important when using learning methods, such as SVM, since it guarantees the optimal solution of the associated convex optimization problem. Another advantage of this kernel is that it can be efficiently computed; in fact, Maji et al. [63] recently proposed a very efficient technique to train SVM that use the histogram intersection kernel.

Using the histogram intersection kernel with SVM, we are modeling a non-linear classification rule in a high-dimensional feature space [64]. This particular property of kernel method solutions, allows us to capture the high variability of visual patterns along the same semantic concept. This special property will be discussed in the next Subsection.

### 3.4.3 Combination of kernels

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

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

$$k_\alpha(x, z) = \sum_{i=1}^{n} \alpha_i k_i(x, z) \tag{3.2}$$

The weights $\alpha_i$ allow to parameterize the kernel giving higher or lower importance to each individual feature. Notice that the linear combination of kernel functions, associated to different visual features, is implicitly defining a new feature space whose structure may be adapted to better recognize a particular semantic concept. In particular, it has been shown that a linear combination of two kernels is a valid kernel

provided that the associated weights are all positive. In addition, the linear combination of two kernels leads to a new feature space that is isomorphic to the Cartesian product of the individual feature spaces [61].

The problem now is to find a vector of weights $\alpha$ that maximizes the performance of the kernel $k_\alpha$ in an image classification task. In the case of histopathology images, different concepts require different classifiers that emphasize the appropriate visual features. Herein we use the kernel alignment strategy [65] to build an adapted kernel function for each concept. Each adapted kernel function is expected to emphasize those visual features that allow to better recognize the presence (or absence) of the corresponding concept in a given image.

Kernel-target alignment [65] measures how appropriate a kernel function is for solving a specific classification problem. In particular, the alignment of two kernels with respect to a sample $S$, is defined as:

$$A_S(k_1, k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle} \sqrt{\langle K_2, K_2 \rangle_F}}, \tag{3.3}$$

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

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

$$A^*_S(k) = A_S(k, k^*) \tag{3.4}$$

The problem of finding appropriate weights for $k_\alpha$ then becomes the problem of finding the weights $\alpha$ that maximize the target alignment $A^*_S(k_\alpha)$. In [66], this problem is solved by transforming it to an equivalent quadratic programming problem and is the strategy followed in this work. This kernel combination strategy is in fact a type of feature fusion task, but performed at the kernel level, making it an integral part of the learning process. The main advantage is that features are optimally combined during the learning process depending on the particular type of classification problem to be solved.

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

$$k_G(x, z) = exp\left(-\left(k^*_\alpha(x, x) + k^*_\alpha(z, z) - 2k^*_\alpha(x, z)\right)/2\sigma^2\right) \tag{3.5}$$

### 3.4.4 SVM Classifiers

Support Vector Machines (SVM) are linear classifiers whose decision function is a hyperplane in the feature space. For each histopathology concept we have modeled a new feature space using adapted combinations of kernel functions, resulting in a new kernel function to classify images of that particular concept. Then, we train a SVM for each concept using the corresponding adapted kernel function. All trained SVM classifiers are then arranged in the semantic image annotator. Since SVM are linear classifiers in the feature space, and our feature space models non-linear relationships between images, the resulting classification rule is non-linear in the input space [64]. This enables the automatic annotation module to capture high visual variabilities among the same semantic concept.

### 3.4.5 Semantic Image Annotator

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

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

$$f(\nu_i) = \frac{1}{1 + e^{-a(\nu_i + b)}} \tag{3.6}$$

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

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

$$T(\nu, \upsilon) = \frac{\nu \cdot \upsilon}{\|\nu\|^2 + \|\upsilon\|^2 - \nu \cdot \upsilon} \tag{3.7}$$

| Identity Kernel | | | | | | | | Hist. Intersection | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GRA | INV | LBP | RGB | SIFT | SOB | TAM | Features | GRA | INV | LBP | RGB | SIFT | SOB | TAM |
| .14 | .09 | .18 | .07 | .19 | .16 | .17 | Blood vessel | .10 | .09 | .06 | .06 | .22 | .18 | .19 |
| .12 | .07 | .19 | .07 | .18 | .17 | .19 | Cystic change | .00 | .00 | .05 | .00 | .40 | .22 | .31 |
| .14 | .09 | .19 | .06 | .19 | .16 | .17 | Eccrine glands | .10 | .09 | .12 | .06 | .23 | .17 | .18 |
| .13 | .03 | .19 | .01 | .26 | .18 | .20 | Elastosis | .04 | .02 | .08 | .01 | .32 | .21 | .22 |
| .13 | .09 | .20 | .07 | .19 | .16 | .16 | Fibrosis | .10 | .09 | .05 | .06 | .23 | .17 | .17 |
| .14 | .08 | .19 | .06 | .19 | .17 | .17 | Lymphocyte inf. | .09 | .08 | .07 | .05 | .23 | .18 | .18 |
| .13 | .10 | .20 | .10 | .15 | .16 | .15 | Micronodules | .09 | .08 | .02 | .06 | .22 | .18 | .19 |
| .13 | .08 | .20 | .08 | .19 | .17 | .16 | Morpheaform pattern | .12 | .11 | .02 | .09 | .17 | .16 | .16 |
| .12 | .03 | .20 | .02 | .25 | .18 | .20 | N-P-C, elastosis | .04 | .03 | .03 | .00 | .31 | .21 | .22 |
| .12 | .04 | .20 | .02 | .25 | .18 | .19 | N-P-C, fibrosis | .05 | .04 | .05 | .01 | .31 | .20 | .21 |
| .11 | .11 | .25 | .13 | .08 | .17 | .15 | N-P-C, infiltration | .11 | .11 | .08 | .09 | .11 | .16 | .16 |
| .13 | .08 | .20 | .07 | .18 | .17 | .17 | N-P-C, pilosebaceous | .09 | .08 | .00 | .06 | .22 | .18 | .18 |
| .13 | .08 | .20 | .07 | .18 | .17 | .17 | N-P-C, trabeculae | .09 | .08 | .00 | .06 | .22 | .18 | .18 |
| .14 | .10 | .20 | .10 | .15 | .16 | .15 | Necrosis | .12 | .10 | .01 | .09 | .17 | .16 | .16 |
| .13 | .08 | .20 | .07 | .18 | .17 | .17 | Perineural invasion | .09 | .08 | .00 | .06 | .22 | .18 | .18 |
| .13 | .08 | .20 | .07 | .19 | .17 | .16 | Pilosebaceous unit | .09 | .08 | .07 | .06 | .23 | .18 | .18 |
| .13 | .08 | .20 | .07 | .19 | .17 | .16 | Thick trabeculae | .09 | .08 | .04 | .05 | .22 | .17 | .18 |
| .13 | .08 | .20 | .07 | .18 | .17 | .17 | Ulceration | .09 | .08 | .00 | .06 | .22 | .18 | .18 |
| 2.34 | 1.38 | 3.57 | 1.23 | 3.37 | 3.03 | 3.08 | SUM | 1.51 | 1.32 | 0.76 | 0.94 | 4.24 | 3.26 | 3.43 |

Feature importance

1st 2nd 3rd 4th 5th 6th 7th

Figure 3.3: Heat maps of weights assigned to features

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

## 3.5 Experimental Evaluation

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

### 3.5.1 Feature Combination

The first step in the proposed framework is to build a new image representation based on kernel functions. In this work, for each of the 18 histopathology concepts, a new kernel is adapted. The feature combination strategy is applied to each class, using a

10-fold cross validation on the training data set to estimate the parameters of the kernel alignment algorithm that optimize the discerning capacity of the feature space. Histogram features were normalized using norm $\ell_1 = 1$, which produces discrete probability distributions instead of frequency histograms, and make the set of features comparable during the combination process. Figure 3.3 shows the list of histopathology concepts with the obtained weights for each feature. We evaluated two basic kernel functions, named Identity Kernel $k_I$ and Histogram Intersection Kernel $k_\cap$. Notice how each kernel function emphasizes differently the set of features, indicating that the discriminative power of each descriptor changes according to the way in which it is used. Also, the optimization algorithm assigns different weights to each concept, varying the way in which features are combined. Each concept obtains a different weight adjustment, since the corresponding set of positive examples have different visual configurations.

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

### 3.5.2 Automatic Image Annotation

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

The experimentation includes the evaluation of two different strategies for building kernel functions: a direct combination of kernels adding functions with equal weights and an optimal combination of kernels using the kernel-target alignment framework. Each strategy evaluates four kernel functions as well: the Identity Kernel, the Histogram Intersection Kernel and the composition of these two kernels with the RBF. Experimental results are presented in Table 3.2 showing the performance measures of

Table 3.2: Classification results on the test data set using different kernel functions

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

all evaluated strategies. The best overall performance is obtained by the optimal combination of features in terms of precision and average-accuracy while recall and F1 are better under the simple combination strategy. This same tendency can be observed when comparing the Histogram Intersection Kernel and the Identity Kernel. This basically means that the former discriminates more accurately while the latter annotates more correct images. Notice that the Identity Kernel deals with features as simple vectors whereas the Histogram Intersection Kernel exploits the structure of histogram data. On the other hand, RBF kernels show a considerably higher recall and better F1 and accuracy values, indicating the effectiveness of the RBF to highlight non-linear patterns in the feature space.

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

Figures in Table 3.2 give an estimate of classification performance on new, unseen histopatholgy images. To evaluate the significance of differences between classification rates, we employed a McNemar's test [67] on the same test data. Table 3.3 presents the results for a number of tests comparing the performance of classifiers. We trained 18 classifiers, one per histopathology concept, with each kernel function. Then, we ran pairwise tests and account for the number of classifiers that are statistically superior and the number of classifiers in which other kernel function is better. We call it wins and losses in Table 3.3, and the difference is computed to determine which kernel provides more advantages to recognize histopathology concepts. It shows that the optimally combined Histogram Intersection Kernel, composed with RBF ($k_{G\circ\cap}^*$), has the largest number of significantly better classifiers with respect to the other kernel

Table 3.3: Evaluation of different kernels using McNemar's test. 18 classifiers are trained for each kernel, one per concept. A classifier based on a given kernel for a particular concept is compared against all other classifiers for the same concept. Cell numbers indicate the number of times that a classifier based on a particular kernel is significantly better or worse than classifiers based on other kernels.

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

functions. Notice that kernels composed with RBF have a positive difference whereas simple kernels accumulate a negative difference in performance, suggesting that the RBF is an important factor to improve classification performance. In addition, it can be observed that the Histogram Intersection Kernel gets a higher score with respect to the Identity Kernel, as well as the aligned kernels with respect to non-aligned kernels.

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

### 3.5.3 Image Retrieval

To evaluate the performance of the retrieval module, images in the test set are used as queries following a leave-one-out strategy, which amounts to approximately 520 different queries. Standard performance measures are used to evaluate the system response including mean average Precision (maPrec), precision at position $k$ ($P(n = k)$), recall at position $k$ ($R(n = k)$), and recall vs. precision plots [68]. The maPrec value is computed using the images that the algorithm retrieves until every relevant image has been found, i.e. until a 100% of recall is met. Reported values are the average results for the 520 test queries. The evaluation of the image retrieval system covers two main strategies to search for similar images: using low-level visual features and using semantic annotations.

Table 3.4: Retrieval performance measures for low-level visual features

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

#### 3.5.3.1 Visual Retrieval Performance

A baseline model using similarity functions for low-level image features is included to compare experimental results. The model based on low-level features calculates the similarity between histograms to produce an image ranking using the Histogram Intersection Kernel as similarity measure [51]. Table 3.4 presents performance measures to compare the response of low-level features, in which SIFT features and Sobel histogram offer the better response. The Bag of SIFT features, that showed the better performance in terms of maPrec, has an important advantage with respect to the other set of visual features: it is based on a learned dictionary of visual patterns extracted from the whole collection, and accounts for an orderless representation of visual patterns in images. Then, these features provide invariance to both, rotation (given by the SIFT descriptor) and translation (given by the orderless spatial arrangement of the bag of features). The invariant feature histogram is also invariant to rotation and translation, however,it takes a geometric approach based on single image analysis. The strenght of the bag of SIFT features resides on the collection-based dictionary construction as opposed to the single image analysis of the other set of features.

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

#### 3.5.3.2 Semantic Retrieval Performance

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

Consider the four performance measures reported in Table 3.5 to evaluate the retrieval response for all kernel functions. The notation for kernel functions is the same as that presented in Table 3.2. $P(1)$ is the precision of the first retrieved image averaged among all tested queries, which is used to evaluate early precision. All semantic mod-

Table 3.5: Retrieval performance measures for all semantic models

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

Table 3.6: Groups with significantly different performance according to the ANOVA test on mean average Precision (maPrec) values.

| Class | | Model | maPrec |
|---|---|---|---|
| I | | Visual Features | 0.103 |
| II | $k_I^*$ | Aligned Identity Kernel | 0.150 |
| | $k_I$ | Identity Kernel | 0.152 |
| III | $k_\cap$ | Hist. Intersection | 0.169 |
| | $k_\cap^*$ | Aligned Hist. Intersection | 0.170 |
| IV | $k_G^*$ | Aligned RBF Kernel | 0.198 |
| | $k_G$ | RBF Kernel | 0.201 |
| | $k_{G\circ\cap}$ | RBF $\circ$ Intersection | 0.201 |
| | $k_{G\circ\cap}^*$ | Aligned RBF $\circ$ Intersection | 0.210 |

els present a $P(1)$ greater than 0.50, meaning that, in more than half of the queries, these models retrieve a relevant image in the first position. This contrasts with visual features in which almost all models have a $P(1)$ less than 0.50. In addition, $P(100)$ shows how the precision changes among the first 100 results, in which all semantic models keep around 0.40, contrasting with visual feature models, which present a $P(100)$ around 0.15. It demonstrates the effectiveness of semantic models to bring more relevant images in the first pages of results. The measure $R(100)$ indicates the recall in the first 100 results, in which semantic models present values around 0.40 whereas only visual models are around 0.15, indicating that more relevant images are rapidly found by semantic models.

The last measure in Table 3.5 is mean average Precision (maPrec), which evaluates the long term precision of the model, that is, the average precision until every relevant image is found. This is the most standard performance measure in information retrieval to compare performance between systems and models. The values obtained by semantic models are around 0.18 whereas only visual features obtain values around 0.10, showing an average improvement of 57%. These results show an important improvement of the retrieval performance of semantic retrieval models over the visual-based retrieval models.

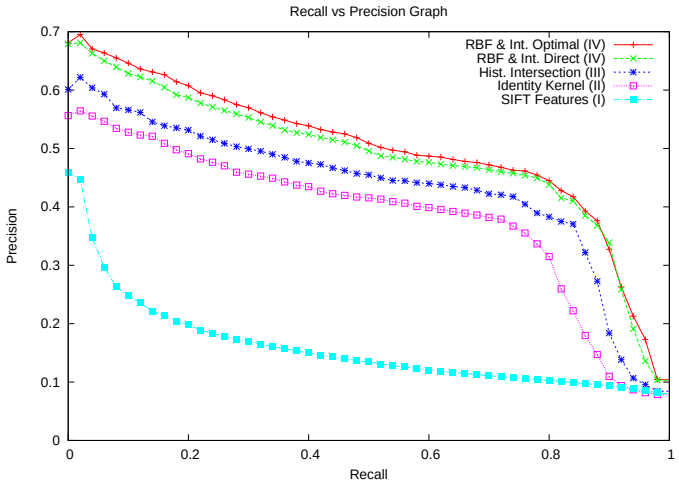



Figure 3.4: Recall vs Precision graph comparing the retrieval performance of statistically different models. Two models of class IV are plotted to illustrate differences between the direct and optimal combination of kernels. The best performing visual feature (SIFT) is included as representative of class I.

To evaluate the significance of the obtained results, we employed and Analysis of Variance (ANOVA) test on the maPrec values. We ranked all models by maPrec and evaluated groups of models to find classes with similar intra-class performance and significantly different inter-class performance. Table 3.6 presents the results of the test using a significance value $\alpha = 1\%$, showing that the difference between semantic models and visual features is statistically significant. It also shows 3 classes of semantic models with statistically different performance, whose partition is mainly due to the underlying kernel function. Each kernel function is a different image representation for the learning algorithms, and these results suggest that the most important factor to build an effective feature space is the use of an appropriate kernel function to exploit the structure of data and to highlight non-linear relationships. Notice that the Histogram Intersection Kernel in classes III and IV provides an absolute performance slightly better when it is obtained from an optimal combination of features. Even though the optimal combination of features does not show enough evidence to be considered as a factor that produces statistically significant differences, it has shown a positive impact in the automatic image annotation and retrieval tasks.

Another way to compare the performance of models is using the recall vs. precision plot, as is shown in Figure 3.4. The parameter to used to generate the curves is the number $n$ of nearest neighbors provided by the retrieval process. This Figure shows the performance of one model of classes I, II and III and two models of class IV, according to the statistically different classes presented in Table 3.6. It shows the differences in

performance between models, as predicted by the ANOVA test. We selected two models of class IV to illustrate the differences between the optimal combination and direct combination of features, in which a slightly improved performance can be observed. We argue that, even though the proposed optimal feature combination strategy does not show a significant improvement, it has potential applications in the design and selection of feature sets for histology image representation. Also, this strategy may allow a further improvement of classification and retrieval results if the set of features is more targeted to describe specific histopathology properties, as opposed to the use of general purpose image features. Nevertheless, constructing image representations with kernel functions has allowed to integrate multiple visual features, to exploit feature structure, to integrate a feature selection strategy and to highlight non-linear patterns in the same framework, as an effective strategy for semantic histopathology image retrieval.

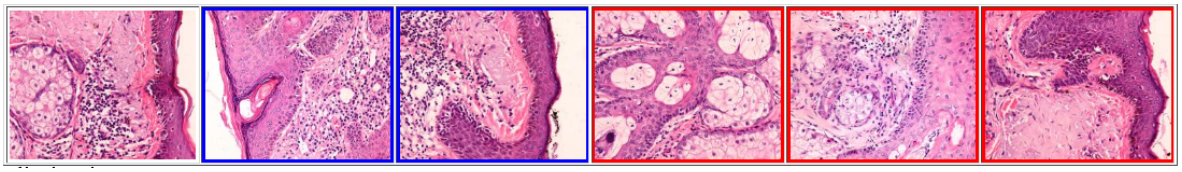
Figure 3.5 shows an illustration of the differences between visual retrieval and semantic retrieval using the proposed methods. The query image is the first from left to right, and it is used to search for images exhibiting the *lymphocyte infiltrate* concept. The top five results are presented immediately after the query image, marked with blue squares if they are relevant or red squares if they are not. The results obtained using Sobel features as retrieval strategy share more appearance commonalities with the query than those provided by the semantic retrieval. However, the three last results of the visual retrieval are not relevant because they do not exhibit the target concept, while the results obtained with the semantic annotations are all relevant.

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

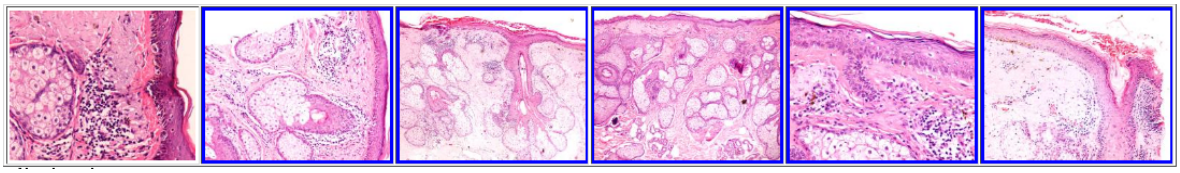
## 3.6 Discussions

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

1. Multiple features: Histology images are known to have objects that can be described using multiple features. Architectural features [45], textural features [69]



(a) *Results obtained using only visual information*



(b) *Results obtained using semantic annotations*

Figure 3.5: Illustration of a content-based query. The query is the first image from left to right. The top-5 results are shown in order of relevance from left to right. Results are marked with blue if they are relevant and with red if they are not. The query image is used to search for images with lymphocyte infiltrate.

and even colors [70] have been proposed to capture variabilities of image contents. We designed the annotation module to deal with multiple features of different nature, and implemented seven histograms in our studies to demonstrate the potential of this approach. These histogram features included textures, colors, edges and invariants, and each histogram has 256 or 512 bins, that are efficiently managed by our module.

2. Structured features: In our kernel-based framework visual features can have arbitrary structure as long as they are provided with a valid kernel function. We evaluated the Identity Kernel and the Histogram Intersection Kernel to process histogram features. In our study, the Identity Kernel can be regarded as an attempt to use the original descriptors as simple feature vectors and the linear combination of Identity Kernels can be understood as the concatenation of these vectors. Experimental results showed that the Histogram Intersection Kernel, which exploits the particular structure of histograms to evaluate a similarity measure, provides more accurate results in classification and retrieval tasks. Our model can be extended to include other structures such as trees and graphs in the visual feature set.

3. Combination of features: Since all visual descriptors are mapped to a high-dimensional feature space using a kernel function, we model the problem of feature combination as a problem of kernel functions combination, and in such a way, we generate combined feature spaces that integrate all the information. This strategy can be understood as a late fusion process as opposed to previous approaches for histology image classification and retrieval that concatenate features in a single feature vector [48, 71], i.e. using an early fusion strategy. Our approach provides

the advantage of considering the particular structure of each feature independently of the others, instead of mixing up everything in a unique vector. Furthermore, our combination approach can include the automatic weighting of features following a kernel alignment strategy. In our experiments, the latter procedure did not show a significant improvement in the final performance, however, we consider this extension as a tool of great potential to select more specialized visual descriptors and to design better image representations.

4. Highlighting non-linear patterns: The histopathology concepts included in our study showed to have high non-linearity in the feature space. This is observed by the large improvement on classification and retrieval performance that was obtained using the RBF kernel. This is an additional advantage of our framework, taking into account that the resulting image representation can be further improved just by operating kernel functions. Representing non-linear patterns is specially useful in image classification tasks, where learning algorithms need to separate complex regions in the feature space.

5. Semantic annotations: Our approach does not attempt to find a unique right class for every image. Instead, it generates multiple annotations according to the visual contents, allowing to extend the functionality to new required search terms. This characteristic makes it different to other approaches that consider just a few labels, as opposed to ours that considered 18 high-level concepts. In our study we only considered the query by example paradigm as the way to retrieve images, but using the automatically generated annotations, images can also be retrieved using a keyword-based strategy.

6. Semantic vs. visual retrieval: Visual features have been extensively used for image retrieval, and the community has found that the main problem using them is the semantic gap. The automatic analysis of visual image contents is at the core of the proposed strategy, and we found that the way in which visual features are used determines the final retrieval performance. Our study showed that the discriminative power of visual features highly depends on the kernel function used to train classifiers, since they allow learning algorithms to exploit feature structure and non-linear patterns. On the other hand, a standard visual retrieval approach only rank images using a similarity measure, i.e. finding nearest neighbors. The success of the proposed semantic retrieval approach is that it uses machine learning to translate non-linear patterns that can be found in visual feature spaces into a more explicit semantic format that is used to rank images efficiently.

## 3.7 Conclusions

This Chapter presented a novel strategy for automatic annotation of histopathology images. The proposed framework is entirely based on kernel methods, allowing to

deal with multiple visual descriptors to build expressive feature spaces. The generated annotations are used to search images with similar annotations in an image retrieval system under the query-by-example paradigm. We implemented and evaluated the model following an extensive experimentation on real histopathology images. The proposed strategy to retrieve semantically valid results from a large collection of histopathology images showed an average improvement of 57% when compared to visual search, based on low-level features. In our future work, we consider the use of more specialized visual features for histology images to improve the final search quality, and the automatic analysis of co-occurrence among annotations to differentiate between normal and abnormal images.

# Chapter 4

# Multimodal Representation, Indexing, Automated Annotation and Retrieval of Image Collections via Non-negative Matrix Factorization

*This work has been published in Neurocomputing [72].*

Massive image collections are increasingly available on the Web. These collections often incorporate complementary non-visual data such as text descriptions, comments, user ratings and tags. These additional data modalities may provide a semantic complement to the image visual content, which could improve the performance of different image content analysis tasks. This Chapter presents a novel method based on non-negative matrix factorization to generate multimodal image representations that integrate visual features and text information. The proposed approach discovers a set of latent factors that correlate multimodal data in the same representation space. We evaluated the potential of this multimodal image representation in various tasks associated to image indexing and search. Experimental results show that the proposed method highly outperforms the response of the system in both tasks, when compared to multimodal latent semantic spaces generated by a singular value decomposition.

## 4.1 Introduction

Most of the effort on web-content mining has concentrated on textual (and hypertextual) data. However, visual information is an important component of the web content nowadays. In particular, the advent of the Web 2.0, has been accompanied by an explosion of multimedia content. Specialized sites, such as Flickr and Picassa,

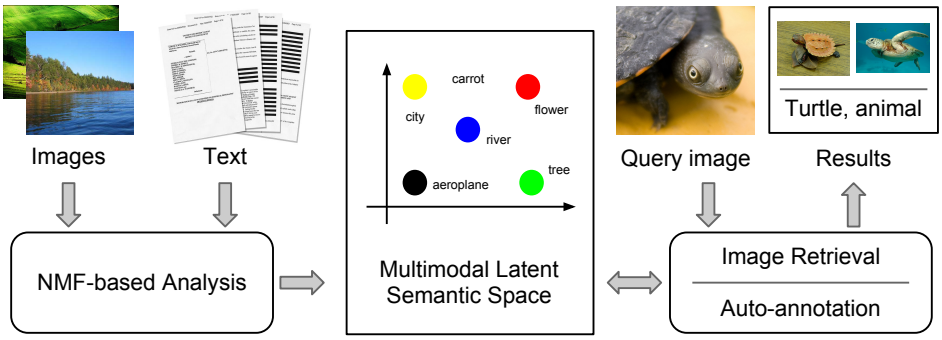



Figure 4.1: Overview of the proposed approach. NMF-based algorithms process image and text features to generate multimodal latent semantic space, in which both data modalities are represented together. Image retrieval and auto-annotation methods exploit the latent representation to find semantically related images and valid automatic annotations, respectively.

host billions of pictures uploaded by users. Other types of sites that allow users to upload visual content include: social networking sites, such as Facebook and MySpace, community-generated content, such as Wikipedia, and individual-generated content, such as in the blogosphere and Twitter.

The most salient characteristic of web image collections is that they come with a wide variety of associated data, such as text descriptions, tags, ratings and user comments. The availability of different sources of information brings the possibility to involve semantic evidence during the analysis of visual content in image collections, which is specially useful when considering the semantic gap [11], i.e. the discrepancy between visual features and semantic interpretations. Therefore, the combination of these data sources together with visual characteristics of images, has received increasing attention from the research community in multimedia processing. The main problem is to take advantage of different data modalities to enable computer systems with the ability to make appropriate decisions according to the high-level task, which is known in the literature as *multimodal fusion* [25].

In this Chapter, we consider the problem of building a multimodal image representation that combines two data modalities: visual patterns extracted from images and text terms extracted from attached text descriptions. The proposed strategy mines the relationships between these two modalities to construct a unified representation based on Latent Semantic Analysis (LSA) principles. We propose a solution based on Non-negative Matrix Factorization (NMF) to construct a latent-factor-based representation that can be spanned using text terms or visual features. We formulate a set of NMF-based algorithms for multimodal image analysis, which generates a joint visual-textual representation that is useful to approach different image analysis tasks.

The main contribution of our work is an NMF-based model to index multimodal data. In this work, multimodal collections are composed of images and some associated text descriptions. These image collections can be built from many different web sources,

including Flickr and Picassa, in which several text descriptions for images can be identified using information extraction techniques. More details about the representation and pre-processing of each separate data modality are given in Section 4.3. Then, given a database of images with the corresponding text annotations, the multimodal analysis is performed using NMF algorithms as depicted in Figure 4.1.

The NMF-based strategy generates the latent semantic space using both data modalities. The goal is to find a set of latent factors that explain the underlying structure of the collection and the relationships between multimodal features. The latent representation is computed using a training data set composed of objects exhibiting both modalities. New objects can later be projected to the latent semantic space even if they do not have both data modalities. In consequence, the multimodal latent semantic model can deal with images without text or text without images, which is specially useful to address several image analysis tasks and retrieval. The proposed models and their properties are presented in Section 4.4.

We evaluate the proposed strategy on two different tasks to demonstrate the potential of the multimodal representation: image indexing and automatic image annotation. Our method projects the input visual features to the multimodal latent semantic space to allow the subsequent analysis. Thus, in addition, we develop a set of algorithms for image search and image auto-annotation that are presented in Section 4.5.

An experimental evaluation was conducted using two image collections: Corel 5k [13], a collection of photographs with several tags and categories, and MIRFlickr 25000 [73], a data set of images downloaded from Flickr.com with the corresponding user generated tags and some additional labels provided as ground truth. The experimental setup and results are presented in Section 4.6, which shows that our proposed model outperforms baseline strategies. The final discussions are presented in Section 4.7 and the concluding remarks are presented in Section 4.8. Portions of this work have been previously reported in [74, 75].

## 4.2 Relation to Previous Work

The use of multiple data modalities for multimedia analysis has become an important research topic during the last years. A comprehensive survey of the many research aspects of multimodal fusion for automatic multimedia analysis can be found in [25], which includes applications in audio, image and video processing using multiple data sources to achieve semantic decisions. Also, in the particular field of image retrieval, Datta et al. [10] discussed the importance of multimodal fusion for image indexing. The construction of systems that make semantic decisions using heterogeneous data sources is the ultimate goal of multimodal fusion.

Two main strategies can be considered for combining multimodal information: late fusion and early fusion. Late fusion, also known as rank aggregation or fusion at a decision level, consists in processing each data source separately during the indexing phase, and the multimodal integration takes place during the query phase. The work of

Ah-Pine et al. [76] is an example of similarity combination to achieve multimodal access in image collections, using pseudo relevance feedback to re-rank images in different applications. On the other hand, early fusion, or fusion at a feature level, consists in modeling feature relationships to create a new multimodal representation, so that during the decision phase, the only task to do is usually analyzing multimodal features [77, 78]. Our work is categorized as an early fusion strategy for multimodal image analysis.

Latent topic analysis has been used to model the relationships between multimodal data, specifically images and text annotations. A set of generative models that use latent variables have been proposed to predict missing captions given unlabed images [79, 15]. These works are based on extensions of the Latent Dirichlet Allocation (LDA) model, in which a set of hidden factors are assumed to explain the associations between the two data types. Later, Monay and Gatica-Perez [80] proposed a simplified aspect model based on Probabilistic Latent Semantic Analysis (PLSA) to index and annotate images by jointly processing visual features and text data.

More recent works follow a latent topic analysis using matrix factorization approaches. Hare et al. [81] proposed a linear algebraic technique based on Singular Value Decomposition (SVD) to learn a semantic space for image features and textual descriptions. This method is a multimodal extension of Latent Semantic Indexing (LSI) for image retrieval that results in a semantic space suitable for image search. Latent topic analysis using matrix factorization has recently drawn of wide interest in information retrieval and image analysis. In particular, NMF algorithms have been used to analyze visual data to discover object classes [82] and to find correlations between image tags [83]. Other applications of NMF decompositions for visual data include [84, 85, 86].

All past works are different from ours since they are focused on processing either visual features or text annotations rather than exploiting multimodal interactions between both data types. Our work is the first one, according to our knowledge, that addresses the multimodal indexing problem using an NMF-based algorithm. In [74] and [75], we addressed the problems of multimodal image indexing and automatic image annotation, respectively. The present Chapter builds upon these works by proposing a unified method for solving both problems, and performing a systematic and extended experimental evaluation.

## 4.3 Multimodal Image Collections

Assume an image collection with attached unstructured text annotations. An excerpt of text may be identified from the source document for each image using information extraction techniques to locate captions, to parse image names and tool-tips, among others [87]. After the information extraction step, each image has an associated unstructured set of text terms. In our model, it is not required that every image has a text description, since it is reasonable to find many of them without surrounding text,

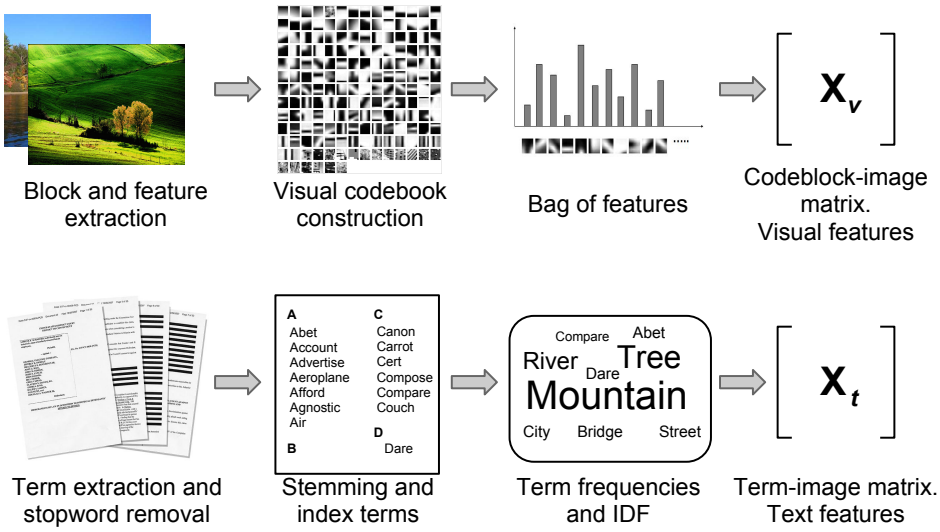



Figure 4.2: Data pre-processing. Each data modality is processed separately to construct two matrices, which are the base of the multimodal analysis.

or it might be difficult to identify a reliable piece of text to be attached. However, it is assumed that there are enough annotated images to perform a multimodal analysis. These collections containing portions of images without text descriptions will be referred to as partially annotated image collections, and that portion of the database will be indexed later after the multimodal analysis has taken place.

We assume the availability of a large enough sample of images taken from the collection together with their associated text descriptions extracted from the source documents. This sample will be called the training data set. Hence, the first step of multimodal analysis is to prepare each individual data modality separately in the training set. This preprocessing step aims to build two matrices, one with visual data and the other one with text data. Each image is represented as a vector in the visual matrix, while its corresponding text description is represented as a vector in the text matrix. In principle, any vector representation for images and text descriptions is allowed as long as all their features are non-negative. Then, the matrices $X_v$ for visual features and $X_t$ for text features will be non-negative matrices as well.

Different representations for visual image contents are naturally non-negative [52]. In this work, we adopt a bag-of-features approach to represent image content. We extract blocks of $8 \times 8$ pixels from a set of training images with an overlap of 4 pixels along the $x$ and $y$ axes to build a set of training blocks. Each block is processed in the three RGB color channels using the Discrete Cosine Transform (DCT) and the 21 largest coefficients per channel are used as features, leading to a block descriptor of 63 features with color and texture information [80]. The $k$-means algorithm is applied to the block set to construct a vocabulary of $n$ visual terms, which serve as reference vectors to quantize feature vectors extracted from blocks in any image. In that way, a histogram with the frequencies of each visual term is constructed for all images in

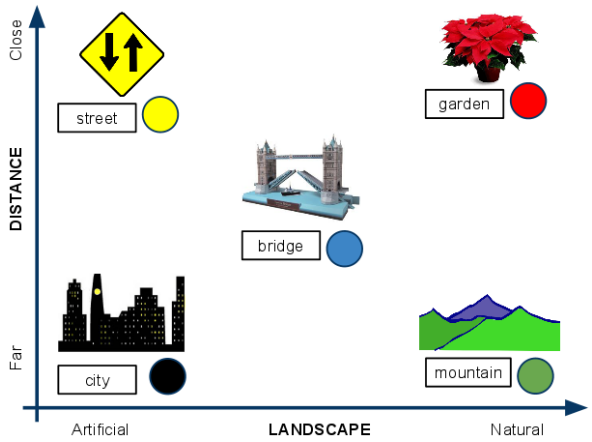



Figure 4.3: Illustration of the latent factor model. Hypothetical images, text terms (rectangles) and visual features (circles) are represented according to their relationships with latent factors in axes $x$ and $y$. From left to right, the factor Landscape changes from artificial to natural scenes. From bottom to top, the factor Distance indicates whether images were acquired from far away or close to an object.

the collection. Using this representation, the matrix of visual data is built, which is denoted by $X_v \in \mathbb{R}^{n \times l}$, where $n$ is the number of visual features and $l$ is the number of images in the collection.

Similarly, text descriptions are processed to construct a matrix of text data denoted by $X_t \in \mathbb{R}^{m \times l}$, with $m$ being the number of text features. Since text descriptions are considered in this work as unstructured annotations, with no restriction in vocabulary, syntax or even language, some natural language processing is required. In the general case, standard text processing operations can be applied to clean up meaningless terms and define the set of index terms. We start by transforming all words to lower-case, removing punctuation as well as stop-words and applying stemming operations [2]. After that, a vector with the frequencies of each index term for all text descriptions is built, to construct the matrix of text data. This matrix of term frequencies can be further improved using Inverse Document Frequency (IDF) weighting to highlight the importance of words along the corpus. In any case, the resulting matrix still meets the non-negativity requirement for our indexing approach.

## 4.4 Multimodal Latent Factor Analysis

Each of the two matrices described above, $X_v$ and $X_t$, provide information about the occurrence of different features for each image in the collection[1]. The concept of occurrence is borrowed from the bag-of-words and bag-of-features models, in which values

[1]Feature vectors for each image are normalized to have L2 norm $\ell_2 = 1$, in both, visual and text matrices.

account for the number of times a particular element appears within the image. In general, these values may be understood as occurrence ratios between features and images, so each image is characterized by the occurrence of certain features and each feature is characterized by the images in the collection in which they appear.

The purpose of a latent factor model is to try to explain these occurrences by characterizing both images and features using a set of $r$ factors inferred from occurrence patterns. For images, discovered factors might measure dimensions such as natural scenes versus buildings, or amount of forest cover as opposed to man-made objects; less well-defined dimensions such as illumination conditions or distance to focused objects; or completely uninterpretable dimensions. For features, each factor measures how much a feature appears in images related to the corresponding factor. We assume these factors to form new meaningful dimensions to organize images in the collection. Figure 4.3 illustrates this idea for a simplified example with two dimensions.

Since in our model, images, visual features and text terms can be represented together in a joint latent factor space, their relationships can be explained using the inner product between their corresponding representations. Let $h_i \in \mathbb{R}^r$ be the representation of an image $i$, and $w_f \in \mathbb{R}^r$ be the representation of a feature $f$, both in the latent factor space. The elements of $h_i$ measure the extent to which the image $i$ expresses latent factors. The elements of $w_f$ measure the extent to which the feature $f$ appears in images associated to the corresponding factors. Thus, the resulting dot product between these representations models the occurrence of the feature $f$ within the image $i$, as follows:

$$x_{f,i} = w_f^T h_i = \sum_{k=1}^{r} (w_f)_k (h_i)_k \tag{4.1}$$

From an image collection point of view, the occurrence patterns can be expressed using matrix notation in the following way:

$$X = WH \tag{4.2}$$

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, $r$ is the number of latent factors and $l$ is the number of images in the collection. Notice that both features and images have a vector representation using $r$ latent factors in the matrices $W$ and $H$ respectively. Also, the matrix $W$ is considered as the basis of the latent space, since each image in $X$ is represented through a linear combination of $W$'s columns using the coefficients in $H$. Thus, the main problem is to compute these representations out of the original feature matrix $X$, i.e., to find a factorization of $X$ in terms of $W$ and $H$.

### 4.4.1 Latent Factors via Singular Value Decomposition

A common approach to compute the latent factors in information retrieval is using Singular Value Decomposition (SVD). This strategy consists of estimating a rank-reduced

factorization of the feature matrix in terms of its eigenvectors and eigenvalues,

$$X = U\Sigma V^T \tag{4.3}$$

where $U$ and $V$ are orthonormal matrices. To generate a semantic space using this decomposition, the eigenvalues in $\Sigma$ are sorted in decreasing order to preserve the first $r$ largest eigenvalues while the rest are set to zero. This is equivalent to dividing the matrices of the decomposition as $U = [U_r U_e]$; $V^T = [V_r V_e]^T$ and splitting $\Sigma$ in two matrices, $\Sigma_r$ a squared matrix with the first $r$ selected eigenvalues and $\Sigma_e$ with the remaining eigenvalues. Then, the matrix factorization can be rewritten as:

$$X = U\Sigma V^T = U_r\Sigma_r V_r^T + U_e\Sigma_e V_e^T \tag{4.4}$$

Assuming that $X$ has $r$ independent factors, it can be shown that the best rank-$r$ approximation to $X$, in the least squares sense, is given by $X_r = U_r\Sigma_r V_r^T$. Using this low rank approximation to $X$, the basis of the latent factor space is $W = U_r$ and the latent image representation is $H = \Sigma_r V_r^T$.

Under this scheme, the basis of the latent factor space is composed of a set of orthogonal vectors from the matrix $U$. The criterion used to select the set of vectors is based on the size of the corresponding eigenvalues in $\Sigma$, since the larger the eigenvalue, the larger the feature variance of the collection in that direction, and the better the low-rank approximation. In that sense, the latent factors obtained using SVD are orthogonal factors that maximize the variance of the data representation.

### 4.4.2 Latent Factors via Non-negative Matrix Factorization

The observed occurrence values in the feature matrix $X$ may be modeled directly by learning the factor matrices $W$ and $H$ for features and images respectively, using alternative matrix factorization techniques to SVD. Two main requirements are herein considered to approximate the matrix factorization in Equation 4.2. First, the resulting basis for the latent factor space should allow non-orthogonal vectors as well as orthogonal ones, as long as they correspond to structural patterns in the feature matrix. Second, the matrix approximation must allow non-negative values only, both for the basis vectors and the codifying vectors too.

Notice that an approximation of the matrix factorization is allowed rather than requiring an exact matrix factorization. Then, the matrix factorization in this problem, may be expressed as $X \approx WH$, or more precisely, $X = WH + E$, where $E$ is a matrix of approximation errors. Thus, this can be approached as an optimization problem to find $W$ and $H$ that minimizes the Frobenius norm of the error matrix $\|E\|^2 = \|X - WH\|^2$, that is, the difference between the original matrix and its approximation.

Another objective function to evaluate the factorization approximation is the Kullback-Leibler (KL) divergence:

$$D(X|WH) = \sum_{i,j}\left(X_{i,j} log\frac{X_{i,j}}{(WH)_{i,j}} - X_{i,j} + (WH)_{i,j}\right). \tag{4.5}$$

This objective corresponds to the KL divergence between the empirical distribution of features in the matrix $X$ and the model distribution $WH$. This amounts to projecting the observed occurrences on the subspace spanned by the factors based on the KL-divergence [88]. This is different from a squared error based on Frobenius norm which would result in an orthogonal projection, resulting in a more appropriate model to deal with the data representations considered in this work: bag-of-features histograms for both visual and textual content, which can be interpreted as probability distributions.

According to Liu et al [89], minimizing the KL divergence is obtained by maximizing the likelihood of observing the data matrix under the assumption that it follows a Poisson Model. Minimizing the Frobenius norm is obtained by maximizing the likelihood of observing a Gaussian distributed data matrix. The Poisson distribution offers a more faithful model, compared to the Gaussian distribution, for representing "counts", which is exactly the case for our bag of words and bag of features representation for the text and visual content of our multimodal data, respectively.

Therefore, the formulation of the optimization problem to approximate a Non-negative Matrix Factorization (NMF), using the divergence objective function, is formally denoted as:

$$\begin{aligned} &min_{W,H} D(X|WH) \\ &s.t.\ W, H > 0 \end{aligned} \tag{4.6}$$

This optimization problem is convex in $W$ only or $H$ only, but it is not convex in both variables together. However, there are different techniques from numerical optimization that can be applied to obtain a good approximate solution. In particular, an algorithm to find $W$ and $H$ simultaneously based on multiplicative updates, has been shown to yield non-increasing objective values and to converge to stable solutions [90]. These multiplicative update rules are as follows:

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i W_{ia} X_{i\mu} / (WH)_{i\mu}}{\sum_k W_{ka}} \tag{4.7}$$

$$W_{ia} \leftarrow W_{ia} \frac{\sum_\mu H_{a\mu} X_{i\mu} / (WH)_{i\mu}}{\sum_v H_{av}} \tag{4.8}$$

Computing $W$ and $H$ in this way provides a good tradeoff between ease of implementation and speed of convergence. It is straightforward to see that the multiplying factors used to update matrices $W$ and $H$ become equal to1 when $X = WH$, which means that perfect factorization is necessarily a fixed point of the update rules. The proof of convergence for this strategy was provided by Lee and Seung [90] using an auxiliary function that is similar to the one used to prove the convergence for the Expectation Maximization algorithm.

The factorization is set to decompose the input matrix into a fixed number of factors defined by the parameter $r$. In our setup, the number of factors (that defines the complexity or order of the factorization model) is found using cross validation on the training data to compare the performance of models with different complexity. Tan and

Fevotte [91] presented a framework to perform model order selection for NMF, using an automatic relevance determination framework. This requires setting prior distributions on the elements of $W$ and $H$, so the method allows to recover an effective dimensionality of the latent space. Their work was developed to address the factorization using the KL-divergence, but can be extended to include other types of metrics.

### 4.4.3 Multimodal Latent Factors

The image database is composed of two data modalities, herein denoted by $X_v \in \mathbb{R}^{n \times l}$ and $X_t \in \mathbb{R}^{m \times l}$. The former is a matrix whose rows are indexed by $n$ visual features and whose columns correspond to the $l$ images in the database. The latter has $m$ rows to represent text terms and $l$ columns for images as well. The construction of a latent semantic space may be done by decomposing the matrix of images that can have only visual features or only text annotations. However, to generate a semantic space for image indexing, we are interested in exploiting multimodal relationships. Thus, two strategies to construct multimodal image representations are proposed as follows.

#### 4.4.3.1 Mixed multimodal representation

This strategy consists of the construction of a multimodal matrix $X = [\alpha X_v^T \ (1 - \alpha) X_t^T]^T$, with $\alpha \in [0, 1]$, a weighting parameter that controls the relative importance of the two data modalities. We set $\alpha = 0.5$ in our experiments (to give the same importance to visual and text data), unless otherwise stated. Then, the matrix is decomposed using NMF as follows:

$$X_{(n+m)\times l} = W_{(n+m)\times r} H_{r \times l} \tag{4.9}$$

where the subindices indicate the dimensions of the matrices, $W$ is the basis of the latent space, i.e., the latent factors, in which each multimodal object is represented by a linear combination of the $r$ columns of $W$. These factor proportions are codified in the columns of $H$. The mixed multimodal representation aims to find correlations between features of both modalities, i.e. to find relationships between visual features and text terms, since both of them are aligned in the same feature matrix $X$. A similar approach using SVD was proposed by [92], in which visual features and text terms are aligned to generate a multimodal latent semantic representation.

#### 4.4.3.2 Asymmetric multimodal representation

The previous strategy decomposes the multimodal information by building a multimodal matrix with visual features and text terms. It has been reported in the literature that text descriptions tend to provide a more reliable information source to extract semantic information for image retrieval than visual features [80]. Evidence of this fact can be observed in different image retrieval challenges that provided data sets with images and text descriptions, and final pollings show a dominant position of text-based approaches

[93, 94]. Thus, we present an asymmetric algorithm for the construction of the latent semantic space that first derives a semantic image representation from text data, and then follows an adaptation of the visual representation to fit the semantic one. In other words, the text information plays the role of a leader compared to the visual content. However, both modalities are exploited to discover the semantic representation.

The algorithm has two main steps to construct the semantic space:

1. *Building a semantic image representation*: This step decomposes the text matrix using the NMF algorithm:
$$X_t = W_t H_t \tag{4.10}$$
In this case, the $m \times r$ matrix $W_t$ contains the vectors of a basis in which text terms are correlated with $r$ latent semantic factors. After this step, the semantic representation for all images in the data set is codified in the matrix $H_t$.

2. *Adapting the latent space basis*: To complete the basis of the multimodal latent space, the construction of a basis for visual features is adapted to match the previously obtained semantic representation. That is, we find a matrix $W^v_{n\times r}$ which spans the semantic space using visual features instead of text terms, as follows:
$$X_v = W_v H_t \tag{4.11}$$

Notice that in the second step, the matrices $X_v$ and $H_t$, the visual features of the training collection and their latent representation, are already known. Then, the problem of finding $W_v$ may be accomplished by computing the multiplicative updates for $W$ while fixing $H$. We refer to this computation step as the *adaptation algorithm* in the rest of the Chapter.

The convergence of this modified algorithm to obtain the factorization $X_v = W_v H_t$ can be understood by analyzing the optimization problems in each step. In the first step of this algorithm, the latent representation for training images is obtained in $H_t$, running update rules in Equations 4.7 and 4.8, which have been shown to converge to a local minimum [95, 90]. In the second step of our asymmetric approach, the matrix $H_t$ is fixed, and we solve it by running the update rule in Equation 4.8 only. It is known that by fixing one of the matrices in the factorization makes the problem of fining the unknown convex [90], thus a global minimum can be found. However, this will be a global minium with respect to the previously found matrix, which is a local minimum of the first problem.

## 4.5 Image Indexing and Auto-Annotation

### 4.5.1 Image Indexing

The main goal of image indexing is to generate an image representation that can be used to match similar contents given a particular information need. After the multimodal

| *Input data* | *Projection* |
|---|---|
| $y$: multimodal | $y = Wh$ |
| $y_v$: visual | $y_v = W_v h$ |
| $y_t$: text | $y_t = W_t h$ |

Table 4.1: Projection of different input data to the latent space in the proposed framework.

decomposition has been done using the algorithms described above on a training set, all other images in the collection have to be indexed, i.e., all of them have to be projected to the latent semantic space. Consider a partially annotated image collection, in which images with and without associated text can be found. Also, since the system is based on a multimodal index, we can consider three different ways to query the system: using example images, using keywords, and using both. Be it for indexing images or to process queries, we can project all data to the latent space using the strategies presented in Table 4.1.

Any input data has to be projected to the latent space by finding $h > 0$ in terms of the corresponding basis of the multimodal space. When multimodal data is available, the full factor basis matrix $W$ is used. When only visual or text data is available, the corresponding sub-matrix of the basis ($W_v$ for visual data, or $W_t$ for text data) is used. The projection of $y$ on the latent space, $h$, is found by using the multiplicative updating rules for $h$ while keeping the matrix $W$ fixed. We refer to this computation step as the *codification algorithm*, since new data and the basis of the latent space are given to find the multimodal latent representation.

Once images and queries have been projected onto the multimodal latent space, a similarity measure is needed to identify relevant results. We use the dot product as similarity measure, and results are ranked in decreasing order of similarity. Notice that the dot product in the multimodal latent space gives a notion of the extent to which two vectors share similar values for latent factors.

### 4.5.2 Image Auto-Annotation

The problem of image auto-annotation is to assign a set of keywords to an image that do not have any attached text. As in the case of image search using visual queries, we need to project images to the multimodal latent space using the codification algorithm described in the previous subsection. Then, to find the annotation words that best describe an unannotated image, we first need to compute its similarity to training images as follows:

$$z = h^T H \tag{4.12}$$

where $h$ is the latent representation of the unannotated image, $H$ is the representation of the training images in the semantic space, and $z \in \mathbb{R}^l$ is a vector of similarity scores. To output annotation words to the query image, we compute their counts in the set of most similar training images. Then the annotation words are ranked according to their counts. This step is similar to what has been proposed in other works ([96, 97]) and is useful here to assess the benefits of using the NMF latent space construction compared to other multimodal based methods.

## 4.6 Experimental Evaluation

### 4.6.1 Datasets

#### 4.6.1.1 Corel 5k dataset

The Corel 5K image database is composed of 5,000 images in 50 categories and has been manually annotated using a text vocabulary of 371 terms [13]. This data set has been used as a benchmark in automatic image annotation and retrieval research, allowing the comparison of several strategies during the last years *[81, 14, 80, 24]*. The bag of features approach is used to represent visual contents using a dictionary of 2,000 visual patterns. The conventional experimental protocol is followed in this work, regarding testing and training partitions. The data set was split into three subsets with 4,000 images for training, 500 images for validation and 500 for final tests. The training set is used to feed the learning algorithms while the validation set is used for parameter tuning.

#### 4.6.1.2 MIRFlickr 25000 dataset

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 [73]. The data set comes with the Flickr tags given by users, which can be considered as low level, noisy text. By processing this content, a 2,105-word dictionary is defined based on the most frequent terms. The bag-of-features approach is used to represent visual content using a dictionary of 1,000 visual patterns. This image collection has also been manually annotated using a set of 38 semantic terms provided as ground truth. The annotation vector has binary elements indicating whether the photo can be described by the term or not. These are considered as high level textual descriptions.

We follow the conventional training-validation scheme, using 15,000 images for training and the remaining 10,000 images for testing [98]. For the retrieval experiments, 1,000 random images are taken as queries from the test set. Then, the database for these experiments contains 24,000 images, from which 15,000 have text annotations and the remaining 9,000 have only visual features. This is a quite realistic setup, in which

the image collection is partially annotated, and the multimodal analysis needs to be generalized to the remaining images.

### 4.6.2 Visual indexing

We evaluate the response of an image retrieval system that uses only visual features to retrieve images from the database when a visual query is provided. Our content-based descriptor is based on the bag-of-features approach, so images are represented by histograms of the occurrence of visual patterns in a codebook. To match image content using this representation, the histogram intersection is used as a similarity measure.

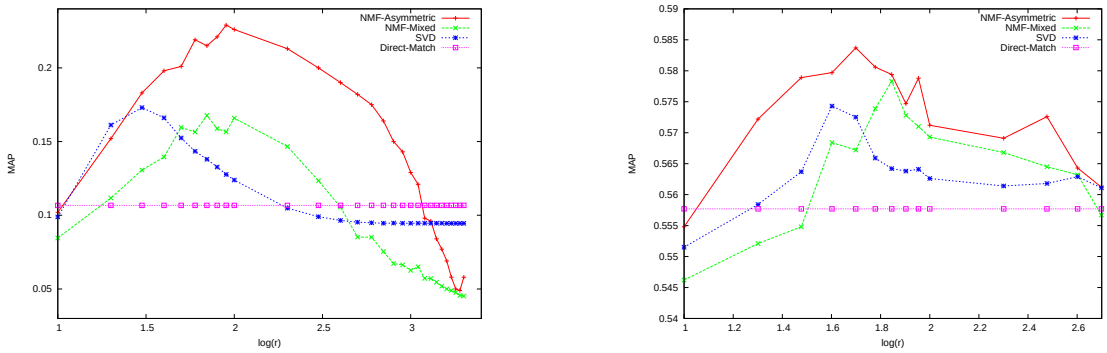
In addition to this visual matching strategy, we evaluate the performance of the system using a latent semantic space built upon visual information only. The NMF and SVD algorithms are fed with the matrix of visual descriptors and the search is performed in the semantic space using the dot product as similarity measure (essentially the same as cosine since the data is normalized to unit 1). To evaluate the performance of the retrieval response, the Mean Average Precision (MAP) is computed in each experiment. In the Corel 5k data set, an image in the results is considered relevant if it shares the same category with the query. For the MIRFlickr data set, an image in the results is considered relevant if it shares at least one semantic label with the query.

Experimental results on the Corel data set showed that direct matching performs better than using a visual latent space either with SVD or NMF[74]. The maximum performance using only visual data reaches a MAP value of 0.101. On the MIRFlickr data set, visual latent spaces perform slightly better than direct matching, reaching a MAP value of 0.557. This strategy does not take advantage of the text annotations in the collection, and works as a baseline to measure the improvement of the proposed multimodal approach.

### 4.6.3 Multimodal indexing

The multimodal analysis is performed using the training data by applying three algorithms: SVD mixed, NMF mixed and NMF asymmetric. Afterwards, all images and queries are indexed in the latent factors' space and the evaluation is carried out by observing the performance in terms of MAP. The experiments follow the Query by Example Paradigm (QBE), to evaluate the response of an image retrieval system that indexes images using multimodal data, even though the expected queries have only visual information. This evaluation challenges the algorithm's ability to retrieve semantically valid results in the absence of text annotations in the query.

Figure 4.4 presents the performance on the validation sets for all indexing strategies for both data sets, using different sizes of the latent space. All multimodal strategies show the construction of improved indexes for image search based on the QBE paradigm. Overall, NMF-mixed and SVD multimodal presented very similar performance in both data sets. The results show that the proposed NMF-asymmetric indexing algorithm achieves a better performance with respect to all other models. This shows that the




(a) Corel 5k data set

(b) MIRFlickr data set

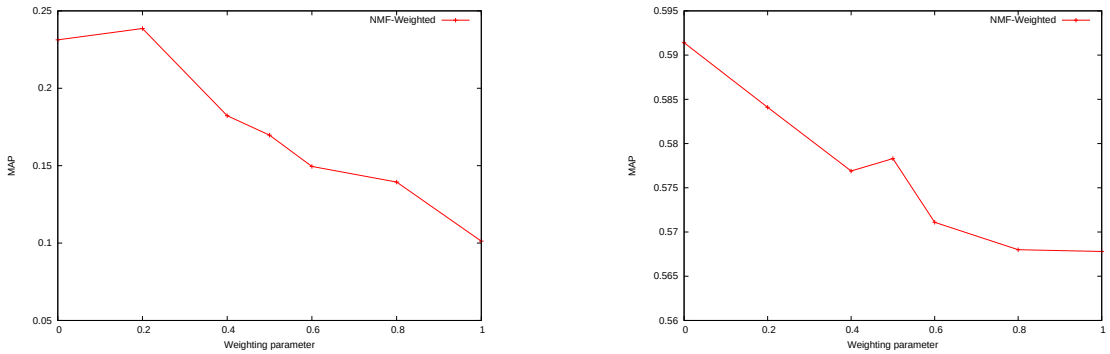
Figure 4.4: Mean Average Precision performance of all indexing strategies evaluating different sizes of the latent semantic space in the training dataset.

Table 4.2: Retrieval performance on the test sets for Corel 5k and MIRFlickr

| | Corel 5k | | MIRFlickr | |
|---|---|---|---|---|
| **Model** | **MAP** | **Gain** | **MAP** | **Gain** |
| Direct matching | 0.1071 | N/A | 0.5577 | N/A |
| SVD Mixed | 0.1780 | 66.2% | 0.5743 | 2.98% |
| NMF Mixed | 0.1727 | 61.2% | 0.5783 | 3.67% |
| NMF asymmetric | 0.2369 | 121.2% | 0.5837 | 4.67% |

proposed strategies effectively involve text semantics in the organization of visual patterns using multimodal factors. Notice that in most of the cases, the number of latent factors required to observe an improved performance is relatively low (less than 100 for both collections).

Table 4.2 shows MAP values for all the evaluated models, computed on the test sets in both collections. Experiments on the Corel data set show large improvements when using a multimodal index to search with images without text data. The relative improvement in the MIRFlircker data set is modest compared to the one observed in the Corel data set. This is because the MIRFlickr is an image collection extracted from a real online service, hence it is more noisy and challenging. This provides more realistic conditions and is also a bigger dataset with a ground truth given by multiple semantic labels rather than clearly defined categories. Despite all these differences, the proposed strategies show better performance with respect to baseline models, resulting in an improved retrieval response.




(a) Corel 5k data set

(b) MIRFlickr data set

Figure 4.5: Effect of the weighting parameter, $\alpha$, on the performance of the NMF-mixed algorithm. The weighting parameter controls the contribution of the visual and text data modalities to the initial construction of the latent space. $\alpha = 0$ corresponds to the assymetric multimodal representation, which is built exclusively using text data, and greater values indicate a greater contribution of visual modality.

### 4.6.4 Weighted Multimodal Indexing

In Section 4.4.3.1 the NMF-mixed algorithm was introduced with a weighting parameter that controls the relative importance of visual and text data modalities. In this Section we investigate the impact of alternative weightings to find the multimodal latent factors. The weighting parameter $\alpha$ allows a flexible configuration of the multimodal decomposition. When $\alpha = 1.0$, only visual information is considered in the matrix factorization algorithm. When $\alpha = 0.0$, the text data is the only one used to build the latent factor space, which is basically the same as the NMF-asymmetric algorithm. Intermediate values for $\alpha$ may result in different performances according to the contribution of each modality.

Figure 4.5 presents the performance response of the NMF-mixed algorithm when the weighting parameter $\alpha$ is changed. Both datasets, Corel 5k and MIRFlickr, show basically the same tendency: as long as more weight is given to the text modality, a better response in terms of MAP is observed. This is mainly due to the evaluation protocol followed for both datasets, which is based on a ground truth that relies on semantic categories or semantic labels. Text annotations are usually more correlated to these semantic representations, and might be considered closer to the human interpretations than visual features alone. However, we believe that under other evaluation scenarios, which may include more perceptual or subjective criteria, giving more weight to the visual modality may be of benefit to the underlying task. Examples of these alternative scenarios include exploratory image search [99] and visual pattern mining [100, 101].

| | Query type | | |
|---|---|---|---|
| **Database** | Visual | Multimodal | Keywords |
| Fully annotated | 0.2289 | 0.3345 | 0.3746 |
| Non-annotated | 0.1709 | 0.2211 | 0.2205 |

Table 4.3: Mean Average Precision for different types of queries on the Corel 5k data set.

### 4.6.5 Answering Multimodal Queries

The previous Sections have evaluated the response of the multimodal indexing system under the Query by Example paradigm, i.e., assuming that users express their information need using example images. That evaluation allows to assess the influence of multimodal data in the visual retrieval task, and enables the system to give more meaningful results when there is no other clue other than a visual example. This search mode may support the operation of online image search services for users with camera phones and other mobile devices that are used to capture an image and then look for similar ones.

However, the proposed multimodal retrieval system can handle other types of query paradigms, as was mentioned in Section 4.5.1, including query by keywords and even multimodal queries, i.e., queries with visual examples and text descriptions. Hence, the following experiments aim to evaluate other types of query paradigms to search in the multimodal index. We consider two main scenarios: first, an image collection that is fully annotated, in other words, text descriptions are available for every image in the database; second, a non-annotated image database, for which the multimodal analysis has been extended from a training sample. We used the NMF-asymmetric algorithm on these experiments on two subsets of the Corel 5k data set (fully annotated and non-annotated).

Table 4.3 reports MAP figures for this evaluation. Visual-only queries refer to the QBE paradigm evaluated in previous sections. Notice that when these visual queries are enhanced with some keywords to provide a multimodal query, the system response is improved. This is consistent with the notion that, as long as users provide more clues about their information need, the system should be able to retrieve more relevant results. In the case of text-only queries the situation is different regarding the type of collection that has been queried. On a fully annotated collection, the multimodal index performs better using keywords-only than using a multimodal query. However, when the collection do not have any text annotation, multimodal queries work slightly better than keywords-only.

This shows that the multimodal index is able to support multiple query paradigms even if the image collection is partially annotated. Then, other methods that rely on visual content only, would provide poor information because of the semantic gap, and methods that rely on text content only, would leave large portions of the images inaccessible. The proposed multimodal index can handle all these situations in a unified

| Collection | Query | Visual | Multimodal | Ground Truth |
|---|---|---|---|---|
| Corel 5k |  | water **sky** **plane** tree | **plane** **jet** clouds **sky** | sky plane jet |
| Corel 5k |  | **water** tree people grass | **buildings** **water** people sun | water sky buildings |
| MIRFlickr |  | male **people** **structures** people_r1 male_r1 **female** sky transport animals car | **plant_life** **flower** **flower_r1** indoor sky **structures** **people** animals **female** female_r1 | female flower flower_r1 people plant_life structures |
| MIRFlickr |  | **plant_life** structures animals **tree** flower transport **water** male people dog | **plant_life** **tree** **sky** tree_r1 structures **water** **river** **clouds** lake sea | clouds plant_life river river_r1 sky tree water |

Figure 4.6: Illustration of the generated annotations using visual content versus using multimodal content for training. Relevant tags are highlighted in blue.

fashion.

### 4.6.6 Image Auto-Annotation

Automatic annotation is performed in the proposed framework by searching similar images in the latent factors' space and selecting frequent terms associated to the top results. Auto-annotation experiments were carried out using the training-validation scheme to tune up parameters (number of factors and number of nearest neighbors). Annotations on the Corel 5k data set are chosen from the 371 terms of the attached text [13, 81, 14, 80]. Annotations on the MIRFlickr data set are chosen from all 38 semantic labels (relevant and potential) [98, 73]. Performance is measured using standard precision scores that indicate how many relevant tags have been assigned to query images.

Table 4.4 reports annotation performance in terms of MAP for both datasets, using the three multimodal strategies. These results are from the best performing configurations and show that NMF-asymmetric has a better performance than the mixed strategies. The annotation performance has a direct relationship with the retrieval performance since the underlying approach to assign terms is based on searching similar images. However, the appropriate factorization changes significantly according to the

Table 4.4: Annotation performance on the test sets for Corel 5k and MIRFlickr. The best MAP value is reported along with the number of factors to achieve that performance.

| | Corel 5k | | MIRFlickr | |
|---|---|---|---|---|
| **Model** | **MAP** | **Factors** | **MAP** | **Factors** |
| SVD Mixed | 0.2663 | 50 | 0.5617 | 100 |
| NMF Mixed | 0.2948 | 100 | 0.5775 | 40 |
| NMF asymmetric | 0.3180 | 300 | 0.6215 | 50 |

set of annotation terms: notice that the number of factors required to achieve the best performance is related to the number of annotation terms in the case of the asymmetric strategy. This suggests that these factors are learned to correlate the features required to identify annotation terms.

Figure 4.6 illustrates some example images with the corresponding annotations generated by a model based on visual data and another model based on multimodal data. The top four terms are shown for the Corel 5k data set, while the top ten terms are shown for the MIRFlickr data set. Ground truth annotations are presented as well to compare the performance of the models. It can be observed that by introducing text data in the factorization process, the quality of the annotation improves both in terms of rank and recall.

### 4.6.7 Computational Issues

The proposed strategy based on NMF algorithms require processing large matrices in real scenarios. In any of the NMF decompositions, there are basically 3 matrices involved in the process: $X_{p\times l} = W_{p\times r}H_{r\times l}$. The approximation to this factorization is achieved by computing iterative updates on $W$ and $H$, using multiplicative rules as was presented in Section 4.4.2. So, considering an implementation using matrix multiplications, and $k$ iterations until convergence, the time complexity of the algorithm is $O(p \times l \times r \times k)$. Notice that the complexity increases with the number of features $p$, the number of training examples $l$ and the number of latent factors $r$.

In practice, matrix multiplication can gain substancial performance using paralellization due to the nature of these operations. Our implementation was developed in Matlab and run on a server machine with 32 GB of RAM memory and 8 CPU cores. The largest matrix factorization in our experiments was done for the MIRFlickr data set, using 15,000 training examples, 5,000 features (visual + text) and 500 latent factors. In average, this decomposition takes 15.2 min. using all cores and 3GB of memory. In the case of the Corel 5k data set, the largest decomposition involving 4,500 training examples, 2,300 features (visual + text) and 500 latent factors, takes about 3.7 min. in average.

In search time, every image has already been indexed in the latent factor space, which in our experiments, turned out to be small, between 40 and 100 factors. A

similarity search using the dot product is linear in the number of images on the database, which takes several milliseconds using our implementation in Python. This search can be also scaled up easily to compute similarities in parallel and using advanced indexing techniques.

## 4.7 Discussion

### 4.7.1 Multimodal representations

Joint image-text analysis has been of wide interest since the seminal work of Barnard et al [79]. Most of the research efforts have been oriented to enable text-based image retrieval or to predict relevant annotations [13, 14, 102, 92, 80, 103, 81, 78, 98]. Our work is oriented to build a multimodal representation for images that integrates visual features and text terms for solving a variety of tasks. In most of our experiments, we used the text modality as an auxiliary source of information rather than using it as the target element to be predicted or explained by the models. The introduction of text information in the image representation was shown to provide important improvements in all the experimental setups presented in this work, demonstrating the potential of the proposed approach.

Our model provides a unified framework to deal with multimodal representations useful to approach a wide variety of tasks for image indexing and search. Few works have explored other tasks beyond keyword-based search, such as querying using semantic examples and pictures from mobile phones [23, 22, 24]. We do not restrict our evaluation to keyword-based search or semantic examples, instead we demonstrated the potential of multimodal representations for image indexing on fully and partially annotated collections, for searching with different query paradigms and for performing auto-annotation. We believe that a truly multimodal image management system should support all these diverse capabilities in an integrated framework and that is precisely one of the contributions of our work.

The evaluation carried out in this work was mainly oriented to measure the performance of the multimodal representation to make semantic decisions, i.e., retrieve relevant results or provide high-level annotations. However, the proposed multimodal representation may be of great benefit for other tasks such as perceptual image analysis or subjective evaluations, including visual pattern mining, image collection visualization and exploratory image search.

### 4.7.2 Latent Factors via NMF

We investigated the potential of matrix factorization to build multimodal latent factors to represent images. Latent factors can be obtained using a variety of methods including Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). We discuss some differences and similarities between

NMF and these other three approaches, that have also been used to model joint image-text data sets.

Hare et al. [92] used an LSI-like approach to build semantic spaces using visual and text data. We consider that modeling orthogonal factors is a restrictive strategy, since some latent factors might be related to each other, indicating meaningful patterns in the collection [104]. In addition, the basis and the latent representation are not restricted in terms of the sign that their dimensions can have, that is, latent factors may have positive or negative components. Different image and text representations are naturally non-negative as was discussed in Section 4.3. Thus, from a data analysis view-point, it is reasonable to model the structure of the collection using non-negative representations as well. Lee and Seung [95] demonstrated how to learn parts of objects by restricting the element representations to be non-negative. These properties lead to a more meaningful image representation, since latent factors can be seen as meaningful parts than can be combined in an additive way to understand image content.

Monay and Gatica-Perez [80] proposed modelling semantic aspects using PLSA to integrate both visual and text data. Although NMF and PLSA have been shown to optimize the same objective function, Ding et al. [105] emphasize the fact that they are different algorithms and converge to different solutions. PLSA deals with data from a statistical viewpoint, using Maximum Likelihood estimation to find an approximate latent representation. NMF models data from a sub-space viewpoint, using optimization strategies to approximate the matrix decomposition. Therefore, multimodal factors in PLSA consist of probabilities associated to features and terms, whereas NMF provide multimodal factors as vectors with visual features and terms which again, can be seen as meaningful parts of objects in the image collection.

Blei and Jordan [15] modelled the joint distribution of text terms and image segments using LDA, based on a generative model in which the visual data is the primary modality and is generated first. Thus, conditioned on the topics used for an image, text terms are then generated. This design explains the observed data following the process of an annotator, in which images are observed and then annotated. In most of our experiments, the NMF-asymmetric strategy performed better than NMF-mixed, suggesting that the text data should be used as the primary modality to approach semantic decisions (see previous Section for a discussion about evaluation). A generative model for our NMF-asymmetric algorithm would generate text data first and later the visual one. This design would explain the data following a painter process, in which the text is given and the pictures are then painted. This difference between Blei and Jordan's model and ours, motivates further research to better understand joint image-text modelling.

## 4.8 Conclusions

We presented an approach for building multimodal image representations using Non-negative Matrix Factorization as a method to create latent semantic factors where data

of different modalities can be associated. The two data modalities involved in our work are visual features and text terms. The main goal of the proposed multimodal representation is to reduce the semantic ambiguity of the visual content. Since we have experimentally found that textual data is more reliable for building topics of images, we introduced the asymmetric NMF method that exploits text data first and then adapts the basis of the visual descriptors. This can be seen as an enforcement of visual patterns to be organized according to the text semantics.

The experimental evaluation showed the potential of the proposed multimodal representation to approach a wide variety of tasks associated to image indexing in a unified framework. We demonstrated how to build multimodal factors, to extend them to partially annotated collections, to search using multiple query paradigms and to annotate images automatically, all supported by the same core methodology. All the evaluations were carried out using two standard real data sets, and the proposed approach consistently performed better than baseline methods. This work shows that matrix factorization strategies can be effectively used to model multimodal latent factors and also that multimodal representations are very useful to perform multiple image collection analysis tasks.

# Chapter 5

# Histology Image Search Using Multimodal Fusion

*The work presented in this chapter has been submitted to IEEE Transactions in Medical Imaging.*

This Chapter proposes a novel method for histology image search, which indexes images by combining visual contents and available semantic annotations. The system has been specially designed to search using example images as queries. The key component of the system models relationships between visual features and semantic annotations using a semantic embedding, which is learned from annotated images by nonnegative matrix factorization. The embedding is used in two directions: from visual to semantic annotations, to project unannotated image queries to the semantic space, and from semantic to visual data, to map back semantic annotations to the visual space. Images are finally represented fusing the semantic enriched visual representation with the original visual representation. Experiments have been carried out using two different histology data sets, and the results show consistent and significant improvements in search performance under the query-by-example paradigm.

## 5.1 Introduction

Digital pathology consists of a series of technologies to acquire, store, visualize, analyze and share histology samples in digital format, contributing to the preservation of all information related to pathology cases. These technologies make easy to exchange histology images and enable pathologists to rapidly study multiple samples from different cases without having to unpack the glass [35]. The increasing adoption of digital repositories for microscopy images can easily end up in large databases with thousands of records, which, besides of just archiving images, can also be exploited to support decision making processes in clinical and research activities. However, to take advantage of these collections, specialized tools for efficient and effective image search are required.

This Chapter presents a method for indexing and searching histology images that allows to present an example image to retrieve the most similar ones from a database. An image retrieval system for histology slides may allow clinicians and researchers to explore large collections of records previously evaluated and diagnosed by other physicians. When a new slide is being observed, a camera coupled to the microscope can capture the current view, send the picture to the retrieval system and show results in a connected computer. These results can help to clarify structures in the observed image, explore previous cases and, in general, support the decision making process.

To provide content-based retrieval support in image collections, different techniques were proposed during the past decade [10], some of them based on simply matching low-level visual features and some of them more elaborated trying to represent the image semantics. The main problem of approaches based on low-level visual features is that they usually fail to capture the high-level semantic of images, producing many non-relevant results. This problem is known as the semantic gap [11]. A number of different approaches attempt to bridge this gap, usually using machine learning methodologies. The general idea is to learn a model that connects low-level features with high-level semantic content. The most prominent approach is based on automatic image annotation [79], which analyzes visual contents to generate appropriate keywords that describe the image. Afterward, the system can retrieve images with similar labels. This approach has been investigated for histology images as well [40, 34].

Even though a system is able to match semantically related images using automatic annotations, two main problems appear when using this approach: first, the system loses the notion of visual similarity, since the search ends up relying entirely on keywords. Second, most of the systems for automatic image annotation need to train one classifier for each associated keyword, making an implementation hard to setup for real world vocabularies. In medical imaging, specialized terms with precise meanings actually help experts to communicate and determine accurate diagnosis. The MeSH controlled vocabulary for indexing journal articles in life sciences has about 30,000 heading subjects, each with the corresponding descriptions and synonyms. Using a detailed description to annotate medical images requires the ability to scale up methods to manage large vocabularies.

In this Chapter, we propose a novel method for indexing histology images using a multimodal fusion approach, that is, combining two data modalities: visual features and semantic annotations. Instead of using keywords as categories to learn classification functions, the proposed method uses them as an additional data source that represents images. The proposed method has two important features: first, it deals with annotated and unannotated images in a consistent way, exploiting those images with annotations to build a semantic embedding and allowing the semantic representation of unannotated images; second, it can deal with large annotation vocabularies since it does not need to build a classifier per concept.

The method is based in a matrix factorization algorithm that finds relationships between visual and semantic representations, making the two data sources exchangeable from one space to another. We further exploit this property by fusing both data

modalities in the same vector space, obtaining as a result the combined multimodal representation for images, which, according to the experimental evaluation, has an important impact on the retrieval performance of the system.

An experimental evaluation was conducted on two data sets of microscopy images, one of histology tissues whose descriptions use 40 different terms [106], and another of basal cell carcinoma with 18 terms [34]. Experimental results show a significant improvement of the proposed method in terms of retrieval precision, demonstrating the ability of the multimodal fusion to simultaneously bring discrimination of visual contents and semantic meanings. The contents of this Chapter are organized as follows: Section 5.2 discusses relevant related works in histology image retrieval. Details about the histology image data sets are presented in Section 5.3. Section 5.4 introduces the proposed algorithms and methods. The experimental evaluation and results are presented in Section 5.5. Finally, discussions are presented in Section 5.6 and concluding remarks are presented in Section 5.7.

## 5.2 Previous Work

The automatic analysis of histology images comprises different purposes and techniques. From image classification [107] to automatic pathology grading [43], the large amount of microscopy images in medicine is pushing for computer methods that allow to assess and manage visual collections to support the decision making process in the clinical practice. This work is focused on image search and retrieval technologies, which serve as a mechanism to identify related and useful histology images. In an image retrieval system, a semantic organization of images is required to successfully identify relevant histology images and filter out non-related ones. The following subsections discuss main approaches toward semantic image retrieval.

### 5.2.1 Supervised learning

Supervised learning algorithms are very common to recognize image semantics by assigning labels to them after analyzing visual contents. The approach consists in computing visual features from images and then train classifiers that separate a few image categories. The design of features are usually oriented to model textures [108, 109], an important characteristic to be analyzed on microscopy images, as well as combinations of multiple features [48, 34]. Then, classifiers are trained to recognize a limited number of categories.

Meng et al. [108] used parts of the IICBU Biological Image Repository to classify images into 3 to 4 categories using principal component classifiers. Naik et al. [48] used Support Vector Machine (SVM) classifiers to distinguish between 3 different breast tissue types. Orlov et al. [109] classified fluorescence microscopy images of HeLa cells with 10 different labels using Nearest Neighbor classifiers. The largest number of categories studied in microscopy images has been 18, for annotating basal cell carcinoma

images using SVM classifiers [60], and 20 for detecting fine semantic features in the Gastro-Intestinal track [40].

The main goal of these methodologies is to categorize input images in a set of labels that may help to organize an image database. However, if an image retrieval system relies only on predicted labels to find relevant images, its functionality would be very limited due to the small number of supported terms. Also, a significant tuning up effort would be required to maximize the accuracy of recognition rates for new categories. Thus, this approach may have problems to scale up to real world applications.

An additional drawback of these methods is that the transformation from visual contents to strict semantic keywords leads to a loss of useful optical information for the search process, because images are summarized in a few keywords and visual details are not considered anymore. Then, users can find images related to the same category, but the system is not aware of their specific visual arrangements.

### 5.2.2 Multimodal Fusion

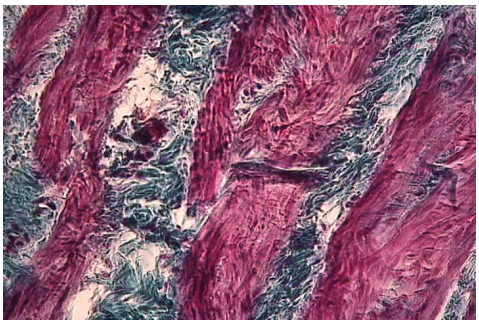
The combination of visual features and semantic information has been approached in some medical image retrieval tasks, such as finding useful images in academic journal repositories [110]. These strategies combine text descriptions or image captions along with visual characteristics to find relevant results, a procedure known as multimodal image retrieval. Cramer and Hersh [30] demonstrate significant improvements of multimodal fusion in a collection of medical images from the ImageCLEFmed challenge. However, their strategy assumes that the user's query is composed of example images as well as a text description. If users only provide example images, because they do not know precise terms or just because of any practical reason, the system does not have any other choice rather than matching pure visual contents.

This work is focused on combining semantic terms and visual features in a fused image representation. An important component of the proposed strategy is the ability to use the same representation for images that do not have text annotations. In that way, the system can handle example image queries as well as database images without semantic metadata. In this work, we build on top of Nonnegative Matrix Factorization algorithms recently proposed to find relationships in multimodal image collections [111]. We extend these ideas to propose a novel algorithm for fusing multimodal information in histology image databases with an arbitrary number of terms. Up to our knowledge, our work is the first multimodal approach specifically oriented to histology image retrieval.

Parts of this work were reported in [112], and this Chapter extends our previous work in two ways: first, the notion of multimodal fusion by back-projection is introduced, which allows to effectively combine visual and semantic representations for histology image indexing. Second, a more comprehensive experimentation was carried out, using more data and additional evaluations and discussions.

*Lymphatic structure of the digestive tract.*

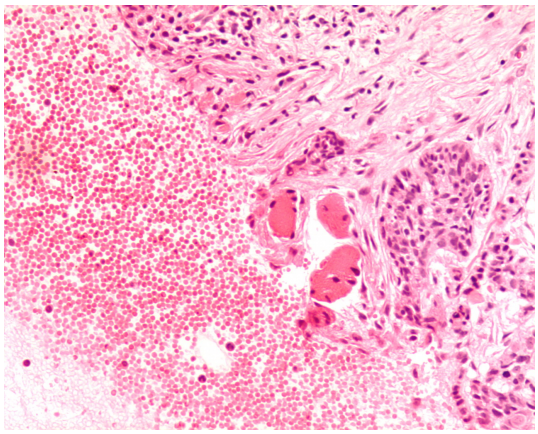
*Digestive system. Appendix.*



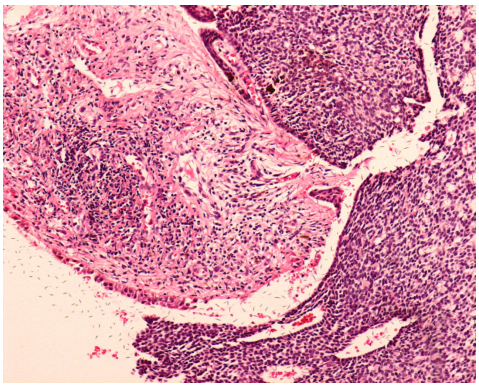
*Female reproductive system.*

*Ovary.*

(a) Histology Atlas Data Set.



*Lymphocyte infiltrate. Micro-nodules.*



*Nodules. Cystic change.*

(b) Basal-cell Carcinoma Data Set.

Figure 5.1: Sample images from two histology datasets.

## 5.3 Histology Images

In this work, two different histology image collections were used for experimental evaluation. The first data set is composed of 2,641 images extracted from an atlas of histology for the study of the four fundamental tissues [106]. 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 40, which were 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 keywords.

The second data set is a histopathology image collection that has been used to diag-

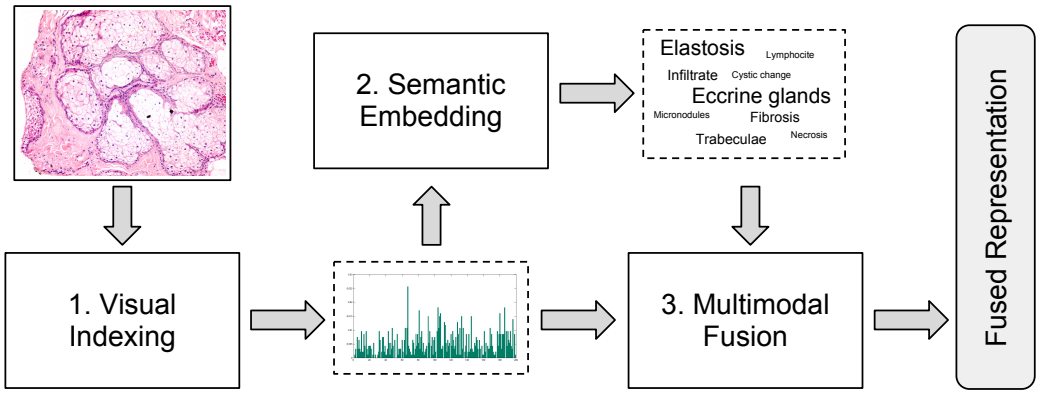



Figure 5.2: Overview of the proposed fused representation. From the input image to the final fused representation, three main processes are carried out: visual indexing, semantic embedding and multimodal fusion.

nose a special kind of skin cancer known as basal-cell carcinoma [60]. 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 terms like *micronodules*, *elastosis*, and *fibrosis*, among others. In this data set, one image usually contains several keywords attached to it, that is, different biological structures are exhibited in one single image.

An important difference between both collections is that the former contains images from healthy, normal tissue slides, used to study biological structures in histology, whereas the second data set contains pathological cases of carcinoma patients. However, one common characteristic is that both data sets have a relatively long list of predefined semantic terms, which includes specialized keywords relevant to image contents in each collection. Figure 5.1 illustrates some example images from both data sets.

## 5.4 Multimodal Histology Image Retrieval

The search method proposed in this work is based on a multimodal representation of images that combines visual features with semantic information. Figure 5.2 presents an overview of the proposed approach, which is comprised of three sequential stages: 1) visual indexing, 2) semantic embedding and 3) multimodal fusion. Three image representations are obtained throughout the process: 1) visual features, 2) semantic annotations and 3) the proposed fused representation. The retrieval engine can be set up to search using any of the three representations. The following subsections present technical details about the stages, the representations and how the retrieval engine works using each of those components.

### 5.4.1 Visual Indexing

The first step toward a multimodal representation is building features for visual contents. A large variety of methods have been investigated to extract and represent visual characteristics in histology images. Be it for automated grading [43], classification [60] or image retrieval [40], two important features are usually modeled: color and texture. Color features exploit useful information associated to staining levels, which are natural biomarkers for pathologists. Texture features exploit regularities in biological structures, since tissues are highly textured. In this work, a bag-of-features representation is used, following the setup proposed in[106], which models both, color features and textures.

The bag of features representation used in this work involves the following three main processes:

1. **Local feature extraction:** each image in the collection is processed to extract a set of local features in small regions. In this work, images are down-scaled to 256 pixels in the largest dimension (height or width). Then, patches of $8 \times 8$ pixels are extracted in a regular grid with an overlap of 4 pixels in the vertical and horizontal directions. Each patch contains local information of the image, which may contain parts of biological structures, such as cells or nuclei. In each patch, the Discrete Cosine Transform (DCT) is computed at each RGB color channel, and the largest 21st coefficients per channel are kept as patch descriptor, preserving the transform basis identifier. Then, the final descriptor for each patch is a sparse vector with 63 non-zero values.

2. **Dictionary construction:** A dictionary of patches is built to model the distribution of features that can be potentially found in images of a histology collection. The dictionary construction is achieved using a k-means algorithm to cluster a large sample of DCT features extracted from the collection. We set the algorithm to find 500 clusters and the centroids of each cluster are used as codeblocks.

3. **Histogram computation:** The final stage of the bag-of-features is the computation of a summarized representation of features in images. This representation is a histogram that accounts for the frequency of dictionary codeblocks inside images. Local DCT features in one image are mapped to the nearest element in the dictionary and the histogram is built by accumulating the occurrence of codeblocks. This representation provides a distribution of visual patterns learned from the whole image collection, and this property, together with the color-texture description of local features, allows to discriminate visual contents of histology images.

Indexing images by visual content in the retrieval system means that all searchable images in the collection, as well as all query images, are represented using the bag-of-features histogram. Then, the retrieval system requires a similarity function to rank images in the collection by comparing them with the features of the query. Since

this representation is a probability distribution of visual patterns, the most natural way to compare these features is using a similarity measure appropriate for probability distributions.

The histogram intersection is a measure for estimating the commonalities between two non-parametric probability distributions represented by histograms. It computes the common area between both histograms, obtaining a maximum value when both histograms are the same distribution and zero when there is nothing in common. The histogram intersection is defined as follows:

$$k_{\cap}(x, y) = \sum_{i=1}^{n} min\{x_i, y_i\} \tag{5.1}$$

where $x$ and $y$ are histograms and the sub-index $i$ represents the $i$-th bin in each histogram of a total of $n$. This similarity measure has been shown to be a valid kernel function for machine learning applications [113], and has been successfully used in many different computer vision tasks [114].

### 5.4.2 Semantic Embedding

In this work, we assume the availability of semantic annotations assigned to images in the database, in other words, a set of images has been collected with attached keywords describing various semantic aspects of histology slides. The purpose of using semantic annotations in the proposed framework is to learn the relationships between visual features and keywords, so the system can predict keywords on example images used as queries. The goal is to make the retrieval results more accurate by representing images using semantic terms as opposed to only visual descriptors.

The proposed strategy is based on a matrix factorization algorithm, which allows to model factors for representing the observed data. This Chapter extends the notions of multimodal image indexing using Nonnegative Matrix Factorization (NMF) originally proposed in [72], by introducing the Nonnegative Semantic Embedding and demonstrating its potential to construct a more effective representation for histology images. The following subsections present the definitions of the proposed method.

#### 5.4.2.1 Data representation

The previous section introduced the bag-of-features representation for visual image contents. Likewise, semantic data is herein represented as a bag-of-words following a vector space strategy, commonly used in natural language processing [2]. First, the dictionary of indexing terms is identified and selected from the total list of available keywords. Then, assuming a dictionary with $m$ terms, each image is represented as a vector in $\mathbb{R}^m$, in which each dimension accounts for the frequency of the corresponding semantic term if it is found attached to the image. Using this representation, each image can have as many semantic terms assigned as needed. Also, the size of the semantic dictionary is not limited and can be easily extended.

Since both, visual and semantic representations are vectors, a database of images can be represented with two matrices by stacking the corresponding vectors of visual and semantic features as columns of the matrices. The notation used in the following sections sets the matrix of visual data for a collection of $l$ images as $X_v \in \mathbb{R}^{n \times l}$, where $n$ is the number of visual patterns in the bag of features representations. The matrix of semantic terms for the same collection is $X_s \in \mathbb{R}^{m \times l}$, with $m$ the number of keywords in the semantic dictionary.

#### 5.4.2.2 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 $W \in \mathbb{R}^{n \times l}$ is the basis[1] 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 [95] is adopted to obtain the factorization, using the divergence criterion as objective function:

$$D(X|WH) = \sum_{ij} \left( X_{ij} log \frac{X_{ij}}{(WH)_{ij}} - X_{ij} + (WH)_{ij} \right) \tag{5.2}$$

which is zero when $X = WH$. This function may be regarded as the Kullback-Leibler Divergence (KLDiv) 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:

$$W_{ia} = W_{ia} \frac{\sum_{\mu} H_{a\mu} X_{i\mu} / (WH)_{i\mu}}{\sum_{v} H_{av}} \tag{5.3}$$

$$H_{a\mu} = H_{a\mu} \frac{\sum_{i} W_{ia} X_{i\mu} / (WH)_{i\mu}}{\sum_{k} W_{ka}} \tag{5.4}$$

These rules are guaranteed to decrease the objective function and find at least a locally optimal solution to the factorization problem [90]. The NMF algorithm has

[1]The terms "basis" is slightly abusive here, since the vectors in the matrix $W$ are not necessarily linearly independent and the set of vectors may be redundant.

been proposed for multimodal image indexing by taking the matrices of visual data and semantic annotations as input, following two strategies [72]: 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 afterward. 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 [95] and finds better structural patterns in different collections of data [104]. This usually results in an improved performance in the underlying computational task.

#### 5.4.2.3 Learning the Nonnegative Semantic Embedding

The core idea of the proposed semantic embedding is to find a direct relationship between visual features and semantic terms instead of modeling them through a latent factor space. The problem is formulated as finding an embedding of semantic terms that allows to reconstruct the visual features of observed images, using a linear transformation with non-negativity constraints, as follows:

$$X_v \approx W X_s \tag{5.5}$$

$$W \geq 0$$

where $W \in \mathbb{R}^{n \times m}$ is a matrix that approximately embeds visual features in the space of semantic terms. Instead of extracting a latent factors structure from the data, the proposed strategy fixes the latent encoding (the matrix $H$ in NMF) as the known semantic representation for images in the collection, $X_s$. This can be understood as requiring the latent factors to match exactly the semantic representation of images, resulting in a scheme for learning the structure of visual features that correlate with keywords.

For this problem, the divergence criterion described in Equation 5.2 is adopted as well. In this case, the optimization problem is convex and can be solved efficiently following gradient descent or interior point strategies. In this work, the matrix $W$ is learned with the first multiplicative updating rule of NMF in Equation 5.3, since this is a rescaled gradient descent approach that uses a data-dependent step size. Notice that the alternating optimization is no longer needed, since the second matrix has been fixed in this algorithm. Then, the solution for NSE is found by iteratively running only

the updating rule for $W$ for a number of iterations or until certain error reduction is reached.

The matrix $W$ is learned using a training data set with available annotations. Be it for those images in the collection that do not have any attached keyword, for new images added to the collection, or for example images used as queries, the learned matrix enables the system to represent any of those using semantic terms as well, since the relationships between visual features and keywords has been extracted from training data.

#### 5.4.2.4 Applying the Nonnegative Semantic Embedding to Unannotated Images

To recover the semantic information of an image without keywords, the following equation has to be solved for $x_s$:

$$x_v \approx W x_s \tag{5.6}$$
$$x_s \geq 0$$

where $x_v$ is the observed vector of visual features and $W$ is the semantic embedding. The non-negativity restriction follows the nature of the semantic data, which is the probability distribution of keywords associated to one image. Again, this problem is formulated as minimizing the divergence between the visual data and its reconstruction, and regarding the non-negativity restriction, the solution can be efficiently approximated using the second updating rule of Equation 5.4 in an iterative fashion.

By solving for $x_s$, the structure of a semantic representation for new images can be predicted. So, the system can complete missing annotations for images without any keyword in the database, and also, can represent visual queries with semantic terms for matching them with images in the database. During search time, the ranking function for a semantic search is based on the cosine similarity, defined as follows:

$$k_{cos}(x, y) = \frac{xy^T}{\|x\| \, \|y\|}$$

where $x$ and $y$ are vectors in the semantic space. This similarity measure is commonly used in text information retrieval and natural language processing [2].

### 5.4.3 Fusing Visual and Semantic Contents

Two strategies for image representation have been presented in the previous sections. The first strategy is entirely based on visual features, to match for visually similar images. The second strategy is based on semantic data and estimations of potential keywords for images without annotations. In this section we introduce the third strategy, 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.

#### 5.4.3.1 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 5.2. 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:

$$x_f := \lambda W x_s + (1 - \lambda) x_v \tag{5.7}$$

where $x_f \in \mathbb{R}^n$ is the vector of fused features in the visual space and $\lambda$ 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:

$$\hat{x}_v := W x_s \tag{5.8}$$

This back-projection is a linear combination of the column vectors in $W$ 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, according to Equations 5.6 and 5.8.

#### 5.4.3.2 Controlling modality importance

The parameter $\lambda$ in the convex combination of the fusion strategy (Eq. 5.7) 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 $\lambda$ 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 $\lambda$ will be called $\alpha$ throughout this Chapter.

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 $\lambda$ will be called $\beta$. This distinction is made to highlight that modality importance

depends on how much we trust the available modalities, and in the experiments run in this Chapter, it has been associated to images in the database and query images.

#### 5.4.3.3 Searching in the fused space

Notice that 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 Equation 5.1 is used as ranking function for search.

## 5.5 Experiments and Results

In order to evaluate the proposed method we conducted retrieval experiments in both data sets under the query-by-example paradigm. The main goal is to evaluate the hypothesis that a semantically enriched representation of images, using the proposed semantic embedding and fusion approach, improves the retrieval performance.

### 5.5.1 Experimental setup

We performed automated 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 information retrieval measures, including Mean Average Precision (MAP), recall-precision plots and precision at the first 10 results (P@10 or early precision).

#### 5.5.1.1 Queries

For both datasets, a set of randomly sampled query images was used for test, and the remaining images were used for training. In the histology atlas data set, 100 images were used as queries, and in the basal-cell carcinoma data set, a 30% of all images were used as queries, i.e., 301 query images [34]. In all our 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.

#### 5.5.1.2 Data representation

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.

Table 5.1: Retrieval performance using three different semantic indexing methods: Non-negative Semantic Embedding (NSE), NMF-Mixed [72] and NMF-Asymmetric [72]. NMF-based methods indicate the number of Latent Factors (LF) used to achieve that performance.

| **Histology Atlas data set** | | |
|---|---|---|
| **Method** | **P@10** | **MAP** |
| Visual matching (baseline) | 0.761 | 0.245 |
| NSE | 0.262 | 0.270 |
| NMF-Mixed (300 LF) | 0.252 | 0.201 |
| NMF-Asymmetric (20 LF) | 0.208 | 0.235 |
| **Basal-cell Carcinoma data set** | | |
| **Method** | **P@10** | **MAP** |
| Visual matching (baseline) | 0.362 | 0.212 |
| NSE | 0.276 | 0.203 |
| NMF-Mixed (400 LF) | 0.198 | 0.182 |
| NMF-Asymmetric (20 LF) | 0.226 | 0.181 |

### 5.5.2 Semantic Indexing

This paper proposes a novel method for semantic indexing, NSE, as well as a method for multimodal fusion, NSE-BP. The goal of this section is to evaluate the former. The NSE extends the ideas of multimodal indexing proposed in [72], so this Section compares performance with NMF-Mixed and NMF-Asymmetric algorithms for semantic indexing.

The main difference of NSE with respect to NMF-Mixed and NMF-Asymmetric is that the latter strategies model relationships between visual and semantic data through latent factors, whereas the former does it directly. Then, we run the NMF-based algorithms to index the two histology image data sets in the latent factors space, using several sizes of the latent space. Table 5.2 presents the results of this experiments reporting the best performance obtained with each method. These results show that NSE produces a better performance compared to NMF-Mixed and NMF-Asymmetric, both, for P@10 and MAP. Notice also that none of the semantic indexing methods improve early precision with respect to the visual matching baseline.

The Table reports the best performance for NMF-Mixed using 300 latent factors on the histology atlas data set and 400 latent factors on the basal-cell carcinoma data set. However, the NMF-Mixed algorithm, which performs a factorization on the joint matrix of visual and semantic data, does not improve upon the baseline in any of the histology data sets. The NMF-Asymmetric algorithm provides the best performance using just 20 latent factors, and is not able to improve performance either.

The NMF-based multimodal indexing algorithms fail to produce a good semantic representation for images in the histology data sets mainly because latent factors do not help to model the structure of the semantic space for these vocabularies. Latent factors are very useful to model patterns in unstructured text data, such as the paragraphs

surrounding images in web pages, which has been the proposed evaluation setup for NMF-Mixed and NMF-Asymmetric [72]. However, in the histology image data sets used in this work, the vocabularies of semantic data are specialized keywords in a well defined categorical structure that is useful to identify specific visual structures in histology images. Thus, the NSE algorithm does a better job at finding the relationships between visual features and semantic keywords directly instead of using intermediate latent factors.

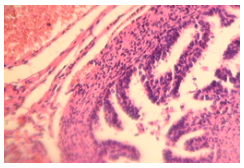
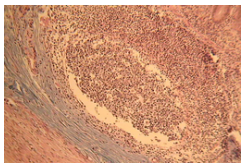
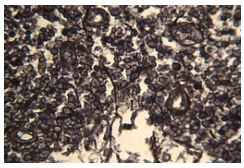
Figure 5.3 illustrates some labels recovered from the semantic representation generated by NSE for several histology images. An experiment was conducted on both data sets to embed images in the semantic space and then, applying a threshold on the predicted semantic vector for identifying coordinates with high response. Since these coordinates correspond to keywords in the semantic vocabulary for each data set, these values may be interpreted as a probability [2] of the label being attached to the image. Even though the core of our approach is not a discriminative model as many previous approaches are, predictions work very well in a number of cases. These predictions are the ones that the proposed fusion strategy incorporates in the visual image representation. This result also demonstrate the ability of the proposed algorithm to effectively find relationships between visual patterns and semantic keywords.

### 5.5.3 Setting Parameters for NSE-BP

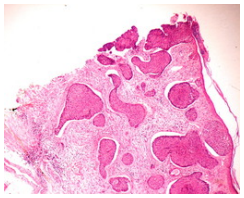
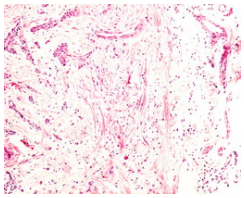
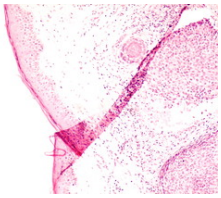
NSE-BP fuses visual and semantic contents by back-projecting images represented in the semantic space to the visual space. Recall from Section 5.4.3 that this fusion process requires to set a couple of parameters to determine the relative importance of each data modality. In this Section we discuss the impact of the two parameters $\alpha$ and $\beta$, 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 $\alpha$ and $\beta$ 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 $\alpha$ and $\beta$ 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 5.4. Points in the plot represent $(\alpha, \beta)$ 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

[2] Using an appropriate normalization.

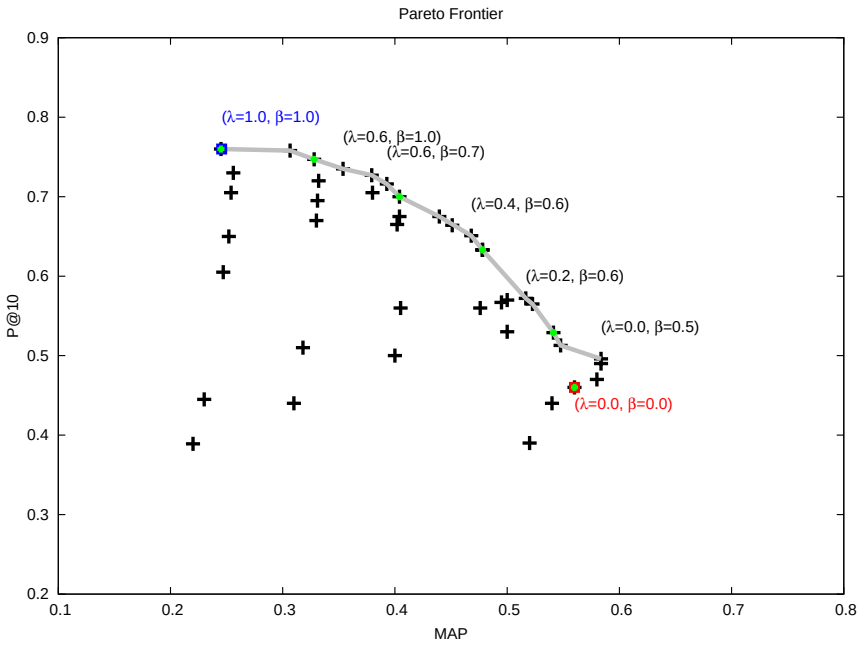
|  |  |  |
|---|---|---|
| 1. Female Reproductive System: Fallopian Tube | 1. Lymphatic System: Lymphatic: Gastrointestinal Tract.<br>2. Digestive System: Ileus | 1. Lymphatic System: Lymph Node |
| 1. Female Reproductive System: Fallopian Tube *[0.9030]*<br>2. Male Reproductive System: Scrotum *[0.2349]* | 1. Digestive System: Ileus *[1.5909]* | 1. Lymphatic System: Lymph Node *[0.7962]* |

(a) Histology atlas data set.

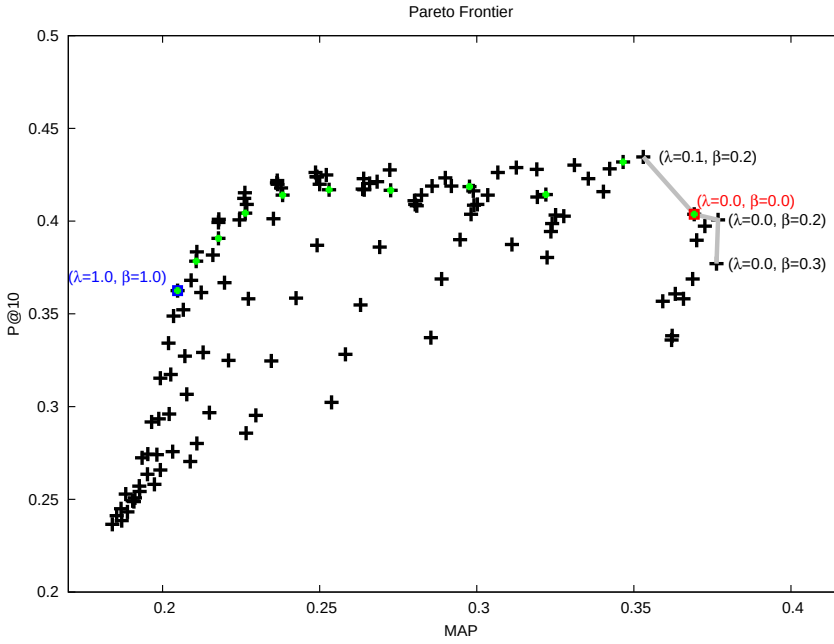
|  |  |  |
|---|---|---|
| 1. NPC Fibrosis<br>2. NPC Infiltration<br>3. NPC Elastosis | 1. Morpheaform pattern<br>2. Fibrosis | 1. NPC Pilosebaceous |
| 1. Pilosebaceous unit *[0.5974]*<br>2. NPC Elastosis *[0.4816]*<br>3. Morpheaform pattern *[0.4215]*<br>4. Micronodules *[0.1600]* | 1. Blood vessels *[0.8096]*<br>2. Morpheaform pattern *[0.2986]* | 1 NPC Pilosebaceous *[0.3755]*<br>2. NPC elastosis *[0.3701]*<br>3. Morpheaform pattern *[0.2754]*<br>4. Micronodules *[0.2464]* |

(b) Basal-Cell carcinoma data set

Figure 5.3: Example images with original keywords and predictions produced by NSE. Top row presents images, middle row presents original keywords, and bottom row presents predicted keywords. Predictions are reported with the score obtained from the projection, and they are in green text if correctly assigned and in red text otherwise.




(a) Histology atlas data set




(b) Basal-cell carcinoma data set

Figure 5.4: Performance on the retrieval task using different values of $\alpha$ and $\beta$. The $x$ axis is MAP and the $y$ axis is P@10. Points in the plot are $(\alpha, \beta)$ pairs. Optimal configurations are on the Pareto frontier.

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 $\alpha = \beta = 0.0$, 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 trade-off 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 $\alpha$ and $\beta$ values to illustrate good performing configurations. In both data sets, the solutions in the frontier tend to have a higher value for $\beta$ with respect to $\alpha$. 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. These findings support 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 rely more on the semantic representation of these images since they have been used for training the model.

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 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 5.7 and 5.8). 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.

### 5.5.4 Retrieval Experiments

The following experiments aim to compare the performance of three search strategies: visual indexing (Visual), semantic indexing (NSE) and multimodal indexing using back-projection (NSE-BP), as was described in Section 5.4. Table 5.2 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

Table 5.2: Retrieval performance for visual (Visual Matching), semantic (NSE) and fused (NSE-BP) representations.

| **Histology atlas data set** | | | |
|---|---|---|---|
| **Method** | **P@10** | **MAP** | **Improvement (MAP)** |
| Visual Matching | 0.7610 | 0.2451 | N/A |
| Semantic: NSE | 0.2620 | 0.2704 | +10.32% |
| Fusion: NSE-BP ($\lambda = 0.6$ $\beta = 0.8$ ) | 0.7590 | 0.3226 | +31.61% |
| **Basal-cell carcinoma data set** | | | |
| **Method** | **P@10** | **MAP** | **Improvement (MAP)** |
| Visual Matching | 0.3615 | 0.2123 | N/A |
| Semantic: NSE | 0.2757 | 0.2032 | -4.28% |
| Fusion: NSE-BP ($\lambda = 0.1$ $\beta = 0.2$) | 0.4346 | 0.3530 | +66.27% |

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 highly likely to be relevant for users.

The multimodal indexing strategy (NSE-BP) 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. Table 5.2 also presents a relative improvement in MAP of the proposed methods with respect the visual baseline, indicating important gains in the precision of the retrieved results. To determine whether these improvements are statistically significant, a hypothesis testing using analysis of variance (ANOVA) was conducted for both performance measures, P@10 and MAP. Table 5.3 summarizes these findings.

In general, the ANOVA tests, conducted with a significance value $\alpha = 5\%$, indicated that there is significant difference between the performance scores obtained by the three models. A closer look between the significance of paired models reveals other interesting results. In both data sets, the statistical difference between visual matching (Visual) and semantic search (NSE) is not significant in terms of MAP. This means, that no improvement or loss in general precision is obtained when using either method. However, the tests indicate that there is a significant difference between Visual and NSE in terms of P@10, i.e., in early precision. And since the NSE produces a loss in early precision according to Table 5.2, we conclude that the loss is significant.

Comparing the performance of Visual Matching (Visual) and the multimodal fused search (NSE-BP), the tests indicate no significant difference between them in terms of P@10 and a significant difference in terms of MAP. This shows that the proposed

Table 5.3: Results of the ANOVA test for performance measures obtained by the three search methods, on both data sets. Statistical differences in MAP and P@10 are reported, and methods are sorted by their corresponding performance from best to worse. Tests were performed with a significance value $\alpha = 5\%$.

**Histology Atlas data set**

| **General Precision** | | | | **Early Precision** | | | |
|---|---|---|---|---|---|---|---|
| Method | MAP | Diff.? | P-value | Method | P@10 | Diff.? | P-value |
| 1. Fusion | 0.3226 | Yes | 0.0422 | 1. Visual | 0.7610 | No | 0.9626 |
| 2. Semantic | 0.2704 | | | 2. Fusion | 0.7590 | | |
| 3. Visual | 0.2451 | No | 0.2550 | 3. Semantic | 0.2620 | Yes | 4.7e-20 |

**Basal-cell Carcinoma data set**

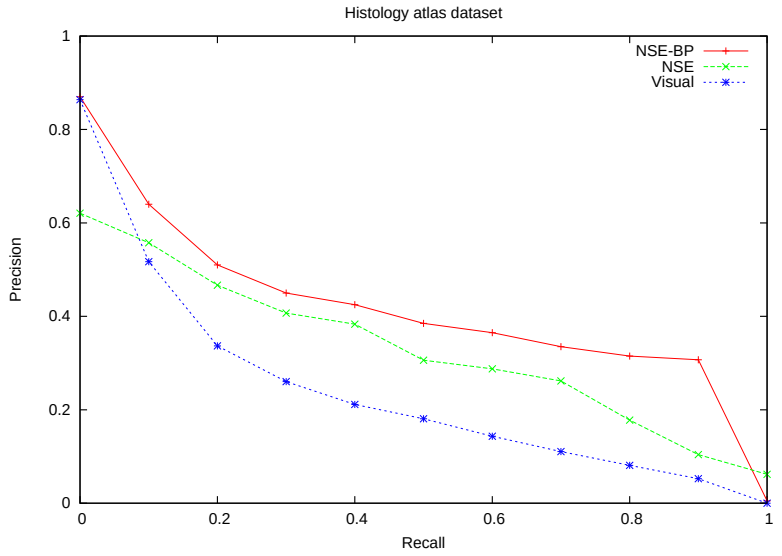
| **General Precision** | | | | **Early Precision** | | | |
|---|---|---|---|---|---|---|---|
| Method | MAP | Diff.? | P-value | Method | P@10 | Diff.? | P-value |
| 1. Fusion | 0.3530 | Yes | 3.0e-12 | 1. Fusion | 0.4346 | Yes | 0.0133 |
| 2. Visual | 0.2123 | | | 2. Visual | 0.3615 | | |
| 3. Semantic | 0.2032 | No | 0.4571 | 3. Semantic | 0.2757 | Yes | 2.0e-05 |

NSE-BP strategy outperforms the general precision of the visual matching strategy without losses in early precision. A final comparison between the NSE and NSE-BP also indicates significant differences between their performances, suggesting that multimodal fusion also improves with respect to a search method based on keywords only.

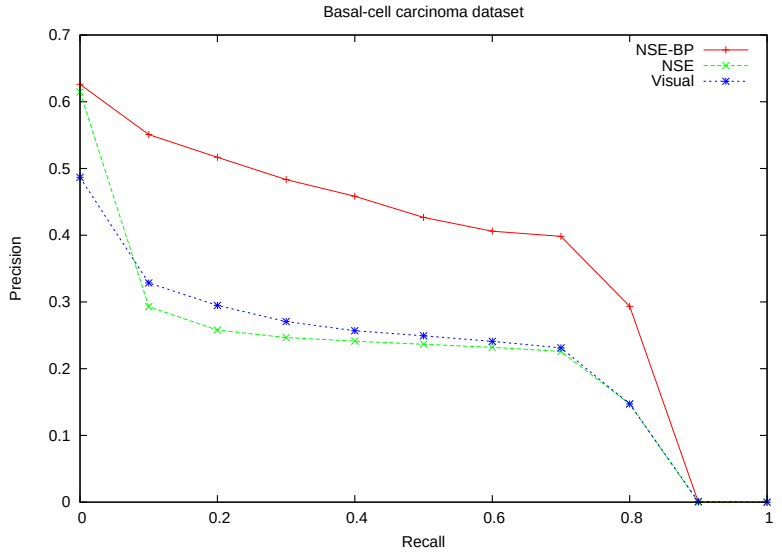
This demonstrates the ability of the proposed fusion strategy to harness visual and semantic information together to build an improved representation for images. The resulting representation is able to give information about visual details that can match other images as well as general semantic concepts associated to them. Figure 5.5 presents the recall-precision graphs for these retrieval experiments, which allow to observe the relative improvements from another perspective. The Figure shows 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.

Figure 5.6 present example retrieval results using the three evaluated methods on the histology atlas data set and the carcinoma data set, respectively. A query image is presented along with the top 4 results returned by each evaluated method. Result images with red frame are non-relevant images whereas images with blue frame are relevant results for the query. Relevant images retrieved by the visual indexing scheme are in general very alike to the query, and share many structural commonalities. However, strong visual matching does not always lead to correct answers for users.

The semantic indexing strategy finds relevant images that do not necessarily match the example visually, but share common keywords. Notice that this strategy does not




(a) Histology atlas data set




(b) Basal-cell carcinoma data set

Figure 5.5: Recall-Precision Graphs for retrieval experiments to compare the performance of the three search methods: Visual Matching, Semantic Indexing and Fused Indexing.

improve upon the baseline in the top 4 results, providing more evidence that translating images to keywords is a hard task that may lead to information loss. The multimodal indexing strategy produces a more balanced response with relevant images that may partially match visual structures and also share similar semantics.

The strength of the multimodal fusion strategy lies in a very important property of the proposed model: the reconstruction principle. The proposed semantic embedding is a matrix factorization algorithm that aims to reconstruct the matrix of visual data from a set of keywords, in other words, images are embedded in a semantic space indexed by keywords. The transformation function between the visual and semantic spaces is a matrix whose columns are prototypical images associated to keywords. Thus, to reconstruct the feature vector of an image, a linear combination of these prototypical images is achieved using as weights the values of associated keywords. This reconstruction is of course a rough approximation to the real observed visual features, however, it highlights those features that images with the same keywords should have.

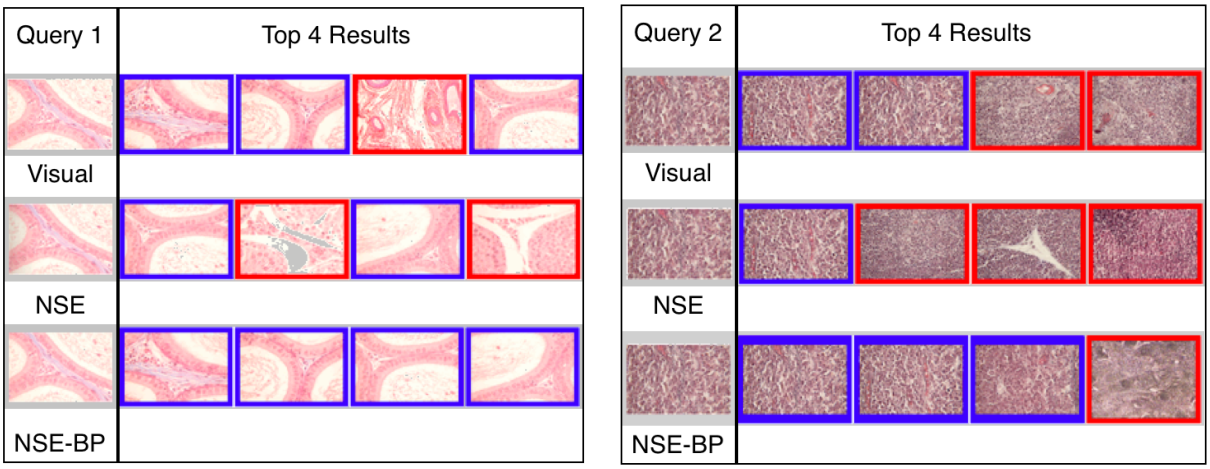
The back-projection step during the multimodal fusion takes advantage of this principle and transforms predicted keywords back to the visual space. This reconstruction works as a semantic smoothing function for the visual data when the fusion takes place. Then, matching smoothed visual features can be successfully achieved by using again the histogram intersection similarity to harness the specific structure of non-parametric probability distribution of visual features.

## 5.6 Discussions

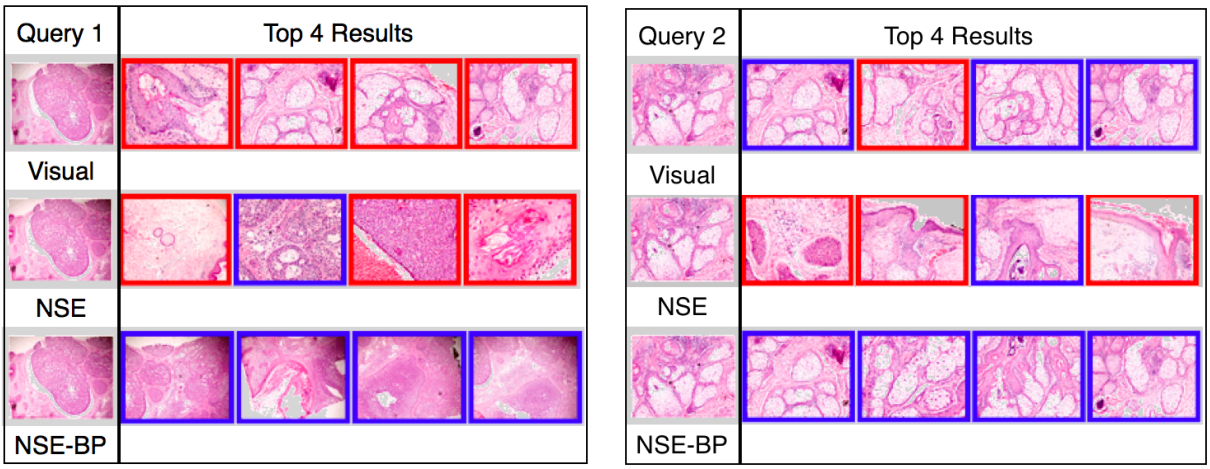
### 5.6.1 Other Multimodal Fusion Approaches

The proposed framework provides a principled tool for fusion of visual contents and semantic resources in histology images. The method models cross-modal relationships through a nonnegative data embedding which has the interesting property of making the two modalities exchangeable from one space to another. This property may be understood as a translation scheme between two languages that express the same idea in different ways. One of the languages is visual, which communicates optical details found in images, and the other language is semantic, which represents high-level interpretations of images. These two views of the same data are complementary and are fused to build a better image representation.

This paper presents the first work that approaches the problem of histology image retrieval following a multimodal setup. Previous works for semantic retrieval of histology images are mainly oriented to train classifiers for recognizing biological structures in images [40, 108, 48]. That strategy can be understood as a translation from the visual space to the semantic space without the possibility of a translation in the opposite way, and thus, without a clear fusion procedure. Experimental results in this work have shown that focusing on semantic data only may lead to losses of information and performance for image search. Multimodal fusion strategies have been used for




(a) Example queries on the histology atlas data set




(b) Example queries on the basal-cell carcinoma data set

Figure 5.6: Illustration of search results using the three search methods: Visual Matching (Visual), Semantic Indexing (NSE) and Fused Indexing (NSE-BP). The first row of each example presents the query image. The next images are the top 4 results, which are framed in blue if the image is relevant to the query and in red if the result is not relevant.

other medical image modalities, including late fusion of similarity measures [115, 110] and fusion of classifier decisions [116, 30]. However, these approaches are not oriented to model the relationships between data modalities as our approach does, and actually assume the availability of visual and semantic data for all images, either in the database or queries, which is not always a realistic scenario.

Multimodal fusion of visual and semantic data has also been recently proposed by several authors [72, 117, 118] for generic web photo collections, whose methodologies model the problem using latent factors to fuse both contents. Our work differs from theirs in the sense that we opted to model cross-modal relationships directly as opposed to through latent factors, which makes a better job for clean and structured vocabularies such as those available in medical image collections. Also, in our model the multimodal fusion is explicitly achieved in the visual space, which contrast with an implicit fusion done in a latent space.

### 5.6.2 Why Multimodal Fusion Works

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. In the proposed strategy, there are two important properties that make it useful for histology image indexing: global vs. local structure and a query expansion effect.

#### 5.6.2.1 Global vs. Local Collection Structure

The proposed framework starts by learning the relationships between visual features and semantic data, which later becomes a model for predicting the structure of the complementary modality when only one of the two is observed. On the one hand, for observed images, this model can approximately tell which labels an image should have. On the other hand, for an observed semantic keyword, the model can tell which visual features are more likely associated with that keyword. Therefore, the model contains global knowledge extracted from the training image collection, which is mainly associated to semantic data.

The fusion process also includes information observed only for each specific image, which provides particular hints that apply for one image and possibly some others locally similar to it. This local information is obtained from the visual similarity measure, which discriminates similar appearances in the image collection. Thus, during the fusion process, global knowledge about the multimodal structure of the image collection is recovered for each image and is incorporated in its representation together with the original visual features. The beneficial outcome of incorporating global and local information for making semantic decisions has also been observed in recommender systems

[119].

#### 5.6.2.2 Query Expansion Effect

Our setup for image retrieval considers example images as queries. Since the visual content representation used in this work is based on a bag of features, an analogy with text vocabularies may help to explain the effects of multimodal fusion.

Visual features in the dictionary of codeblocks may be understood as visual words representing specific visual arrangements or configurations. One specific pattern is a low-level word that may have different meanings from a high-level or semantic perspective. This problem is known in natural language processing as synonymy and can reduce the ability of an information retrieval system to retrieve all relevant documents [111]. Also, different visual words may be related to the same high-level meaning, which is known as polysemy, and usually decreases retrieval precision, i.e., the ability of the system to retrieve only relevant documents [111].

Experimental results in Section 5.5.4 are consistent with these definitions, since a visual synonymy effect is observed when the retrieval system is based on visual features only (the lowest MAP score). On the other hand, a visual polysemy effect is observed when using back-projected semantic data, with a higher MAP score but lower early precision (P@10). Thus, the back-projection of semantic data is able to disambiguate visual words by introducing other visual words semantically correlated to the query, and so correcting the synonymy effect. Here is when the visual query expansion happens. Besides, when both modalities are combined, the polysemy effect can also be corrected, if appropriate weights are assigned.

### 5.6.3 On Scaling Up to Large Semantic Vocabularies

Previous works on histology image retrieval are mainly based on classifiers trained to recognize several biological structures [40, 109, 60, 48, 108, 34]. To transfer these methodologies to real world system implementations, a significant tuning up effort would be required, since each classifier may have its own optimal configuration. The proposed method is a more principled and unified approach that integrates all semantic labels together for learning multimodal relationships. This makes an implementation simpler and ready to scale up for new keywords, as long as the corresponding example images are available.

We reproduced the experimental setup of [34] on the basal-cell carcinoma data set, and obtained a MAP score of 0.162 using our approach. This contrast to a MAP score of 0.170 reported by them, using SVM classifiers with histogram intersection kernels and optimally combined features. Our approach produces a very competitive result with a simpler strategy, and the difference in performance might be related to different aspects, such as the use of multiple descriptors as well as optimized parameters in that work. Instead of optimally combining multiple features for each of the 18 keywords, and then training and tuning 18 different classifiers, the proposed method makes a

single training step for learning multimodal relationships with no parameters. Fusion parameters are required later during the search phase, which can be tuned up online even while the system is already running.

## 5.7 Conclusions

This paper presents a novel strategy for multimodal fusion on histology image collections, which combines visual features and semantic data following a principled approach. This strategy comprises three main components, including a bag-of-features representation for visual contents, a semantic embedding algorithm to model the relationships between visual features and semantic metadata, and a scheme for predicting semantic data and combining it with visual features. The proposed algorithms for semantic embedding and fusion are based on Nonnegative Matrix Factorization which is a useful tool for modeling data patterns.

Experimental results show a consistent good performance in different evaluations explored in this paper, including retrieval precision with varying parameters, comparisons of retrieval strategies and illustrations of practical examples. This work has shown that representing images with semantic data provides only part of the information required to retrieve relevant images. In general, the proposed fused representation for histology images has demonstrated to be effective for correctly matching images under the query by example paradigm. Besides an improved performance, our approach results in a simple strategy to simultaneously learn from many semantic terms, which may be available for histology image collections.

Future research directions include the evaluations of these algorithms for large scale learning and indexing, considering image collections with more images and more semantic terms, both, for histology as well as other medical imaging modalities.

# Chapter 6

# Large Scale Multimodal Image Indexing via Online Matrix Factorization

*This work is prepared to be submitted to the Journal of Multimedia Information Retrieval.*

This Chapter addresses the problem of indexing a large set of images that could have structured or unstructured annotations. The main goal is to use the annotations to enrich the image content representation to improve the performance of content-based image search systems. The proposed method builds a multimodal semantic representation that fuses both the visual and textual content of images. This is accomplished by an online matrix factorization algorithm that finds a set of latent factors to encode the visual and textual content of the image collection. This semantic representation can be used to search the collection using a query-by-example strategy and to assign annotations to new images without keywords. The most remarkable characteristic of the proposed method is its formulation as an online learning algorithm, which is highly efficient in the use of memory resources and converges very quickly running in a single CPU. This allows to scale up the applicability of multimodal analysis to very large image collections. Experimental results on three different data sets show that this strategy produces improved retrieval performance with low computational effort.

## 6.1 Introduction

Image retrieval systems have became more pervasive during the last years, demonstrating applications in many different contexts, such as web image search, medical image analysis, mobile applications, and scientific imaging, among others [10]. Web users find images by typing some keywords on a search engine, while physicians and mobile users

provide image examples of what they are searching for. So, be it for informal activities or serious decisions, image retrieval systems have the potential to support many processes involving visual signals.

Being able to retrieve images automatically allows users to quickly browse image collections and make informed decisions relevant to their specific task. Without such a tool, finding images would be infeasible since modern image collections are very large and unstructured. The main challenge when designing and building image search systems is to ensure that the visual contents of the results are semantically valid for the user's query. Then, if a user is searching for images of dogs, the system should filter out images without dogs, and also, should not introduce images with cats even if they look very similar.

For a search engine accepting example images as queries, the content-based approach has been studied during the last two decades [9, 26] resulting in important progress. Very efficient systems can be built to find images using visual features in quite large collections [120, 121]. The accuracy of these systems has raised for some specific tasks, being the most successful the detection of near-duplicate images. However, it is well known that matching visual features alone may lead to results with lack of semantic validity [11].

To overcome this problem, the use of semantic resources has been introduced during the indexing stage of images, using two fundamentally different strategies: automatic image annotation [13] and multimodal indexing [25]. The first strategy aims to learn classification functions that evaluate image contents and automatically generate labels for images. The second strategy intents to build a representation that incorporates both, visual and semantic information about images. The main difference between these strategies is that the former requires manually labeled data whereas the latter only requires unstructured attached texts.

The purpose of multimodal image indexing is to generate a data-driven image representation, which simultaneously incorporates visual information as well as potential semantic descriptions. The potential source of semantic information for images can be easily found in different collections, including images from books and scholarly articles as well as web images. So, instead of requiring human intervention to define and assign labels for images, an algorithm is designed to automatically discover image categories from visual features and text data, and incorporate that information in the image signature.

This paper proposes a multimodal matrix factorization algorithm, which takes as input visual and text features from a collection of images and produces a latent factor representation. The resulting representation is used as indexing structure for image search under the query-by-example paradigm, i.e., users send example pictures to retrieve relevant ones from the database. The proposed approach focuses on learning a semantic representation for images, following an unsupervised, data-driven strategy instead of learning from labels or predefined categories. This framework can handle images without text annotations using a projection function to the latent factor space, which may be understood as a function that implicitly completes text annotations for

these images.

The design of the algorithm has been inspired by recent theoretical works on large scale learning, which suggest that the main bottleneck for mining modern collections of data is computing time rather than the number of samples [122]. Then, the proposed factorization algorithm has been formulated as an online learning process, which allows to scale up its applicability to data collections with a vast amount of examples. Experimental evidence shows fast convergence rates, allowing to obtain useful models with a single pass over the training data and little memory requirements.

In addition, results demonstrate the effectiveness of the proposed algorithm on three different benchmarks for image retrieval, achieving improved performances with minimum computational effort. This is a very remarkable achievement, which makes the proposed strategy applicable for real world systems based on large image collections. Up to our knowledge, this work presents the first Online algorithm for Multimodal Matrix Factorization, which we call OMMF throughout this paper.

The structure of the paper is as follows: the related work is discussed in Section 2. Section 3 presents the formulation of the proposed algorithm and its extensions. Section 4 reports experimental evaluations and results. Discussions are presented in Section 5 and Section 6 present concluding remarks.

## 6.2 Previous Works

In this paper, multimodal learning is considered as the problem of finding or discovering in an automatic or unsupervised way, the relationships between visual features on images, and text keywords attached to them. Then, the learned relationships are used for organizing an image database by indexing images with or without text descriptions. Under this definition, multimodal strategies that do not learn the relationships between these two data modalities are not directly related to this work. For a comprehensive survey of multimodal strategies in multimedia applications, the reader can refer to [25]. This Section reviews research works of multimodal learning methods oriented to model image-text relationships, as well as potential computational strategies to scale these models to large collections of data.

### 6.2.1 Learning Multimodal Relationships

One of the earlier attempts to modelling the relationships between images and words is the seminal work of Barnard et al. [79], which introduced the multimodal Latent Dirichlet Allocation (mmLDA) algorithm to learn the joint distribution of image regions and words. Several subsequent works have proposed related probabilistic algorithms to approach this problem, including multilayer Probabilistic Latent Semantic Analysis [123], and topic regression multimodal LDA [124].

The core concept underneath all these methods is to model latent variables for discovering the hidden structure of the data, using different prior distributions and

different dependencies between variables. And that idea has shown to be a powerful tool for learning multimodal relationships. However, the main bottleneck of all these models is their computational cost, which make them difficult to implement and scale for large image collections.

Matrix factorization algorithms have also been proposed for modelling latent structures, and have been used for learning multimodal interactions. For instance, Chandrika and Jawahar [31] evaluated a multimodal Latent Semantic Indexing (mmLSI) algorithm, based on Singular Value Decomposition (SVD) along with mmPLSA, obtaining improved results with both algorithms.

Several algorithms for learning multimodal relationships using Nonnegative Matrix Factorization (NMF) have been recently proposed by Caicedo et al. [72, 125] and Akata et al. [117]. These algorithms, the asymmetric NMF (aNMF) and the multimodal NMF (mNMF), were proposed simultaneously in independent works, and share a similar structure. One of their main differences is the underlying cost function that each algorithm optimizes, the former uses NMF under the Kullback Leibler Divergence (NMF-KLDiv) and the latter uses NMF under the Frobenius norm.

NMF-KLDiv has been demonstrated to be equivalent to the PLSA algorithm since they optimize the same objective function [105]. This sets interesting theoretical insights to model matrix decompositions for learning latent factors. However, even though matrix factorization can provide an alternative framework for multimodal learning, these algorithms still require significant computational resources for large scale learning.

### 6.2.2 Large Scale Multimodal Learning

The main drawback of current multimodal learning strategies is that the associated algorithms are memory and computation intensive [31], which makes it difficult to use in a large scale setup. For instance, the work of Romberg et al. [118] 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 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 [126]. Similar approaches have been proposed for parallelization of matrix factorization algorithms [127, 89] for web scale collections. However, it still requires large computational resources dedicated to decompose big matrices. More efficient strategies can be formulated to exploit the structure of matrix decompositions, such as the one proposed in this article. Besides, generic distributed matrix factorization algorithms are not applicable for a multimodal learning setup, which has two input matrices instead of one.

A different approach is presented by Mairal et al. [128] who proposed an online algorithm for sparse dictionary learning, which allows to process datasets of millions of instances with low memory consumption. This demonstrates the potential of online

learning for matrix factorization with the use of little computational resources. That algorithm has been proposed for different matrix factorization setups, mainly oriented to sparse coding and data compression. Our work follows a similar approach, in the online learning sense, but focusing the formulation toward multimodal analysis instead of sparse coding.

## 6.3 Multimodal Matrix Factorization

Consider the problem of building an image representation using visual features and text data. The collection of images is assumed to have text descriptions attached to most of the images, and that information has no particular structure and is possibly noisy. Many modern collections of images fall under that definition, such as images on the web with surrounding text, pictures from scholarly articles or books with their corresponding captions and definitions, and medical images with attached diagnosis and health records, among others. A collection of images tied to texts can be built by crawling these multiple sources of information.

From any of those sources, assume an unknown joint probability distribution for images and texts, from which aligned pairs can be drawn. Some of these pairs may have a higher probability than others, for instance, a picture of a car is likely to appear with keywords such as *vehicle*, *motor*, *wheels*, and less likely to have keywords such as *leg*, *food*, or *wallet*. We want to learn the structure of the probability distribution by observing these pairs and model the visual-text relationships using latent factors. The main assumption of the proposed model is that one latent factor describes a meaningful relationship between a small set of text keywords and a specific set of visual features.

Figure 6.1 illustrates an overview of learning multimodal latent factors from an image collection. The two available data modalities are in the sides of the picture, labeled with $V$ for visual and $T$ for text , which are aligned or paired. Latent factors are depicted in the middle of the Figure as correspondences between some visual arrangements and keywords that are potentially related to them. The proposed model aims to learn the parameters of two transformation functions, $P$ and $Q$, that correlate modalities with latent factors. These functions take as argument a distribution of features (visual and textual, respectively) and produce the distribution of latent factors that are more likely associated with these features.

The problem of learning the two functions $P$ and $Q$ is formulated as a matrix factorization problem in this work. Notice that these functions define a common representation space for both data modalities, which is the space of multimodal latent factors. Modelling a common representation space for two data modalities is also at the core of other multimodal learning algorithms [15, 118, 124].

This work assumes a bag-of-features for visual contents, which represents images using histograms with the frequency of visual patterns [52]. On the other hand, to represent text contents associated to images, a valid Vector Space Model is assumed [2], which may account for the frequency of keywords, for instance. The two sources of

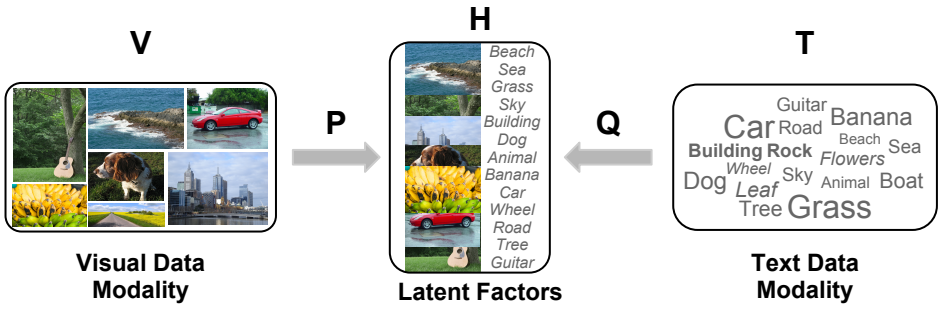



Figure 6.1: Overview of the proposed approach. The two data modalities are correlated through a latent factors space.

information are organized in two matrices to allow for collection-based analysis. The goal is to analyze the relationships between these two matrices and to extract meaningful patterns for building the desired multimodal representation.

The following is the notation used throughout this paper. Let $v_i$ be the vector of $n$ visual features for the $i$-th image. Let $t_i$ be the vector of $m$ text features for the same image. The number of training samples in a collection with available visual and text features is $\ell$. Then, the matrix of visual features is noted by $V = [v_1 \dots v_\ell] \in \mathbb{R}^{n \times \ell}$ and the matrix of text features is $T = [t_1 \dots t_\ell] \in \mathbb{R}^{m \times \ell}$. Abussing the notation a little bit, we will indicate the fact that a vector $v_i$ corresponds to a column of $V$ by $v_i \in V$, $t_i \in T$ is interpreted in the same fashion.

### 6.3.1 Problem statement

The problem of multimodal analysis in this work is set to build a data-driven representation for images in an unsupervised fashion, using as input the visual features and text annotations of images. This is formulated as finding the common factors between the two data modalities to span a new representation space combining both aspects. The multimodal factor decomposition consists on solving the following optimization problem:

$$\min_{P,Q,H} \frac{1}{2} \|V - PH\|_F^2 + \|T - QH\|_F^2 + \lambda \left( \|P\|_F^2 + \|Q\|_F^2 + \|H\|_F^2 \right) \tag{6.1}$$

where $P \in \mathbb{R}^{n \times r}$ and $Q \in \mathbb{R}^{m \times r}$ are linear functions encoding the relationships between $n$ visual features and $r$ factors, and between $m$ text features and $r$ factors, respectively; $H \in \mathbb{R}^{r \times \ell}$ is the matrix whose column vectors are the latent representation for a sample of size $\ell$; and $\lambda$ is a regularization parameter that penalizes arbitrary large magnitudes of the three matrices, in the Frobenius norm sense.

This optimization problem is equivalent to optimizing the following empirical cost

function when $j = \ell$:

$$\min_{P,Q} f_j(P,Q) := \frac{1}{j}\sum_{i=1}^{j} L\left(v_i, t_i, P, Q\right) + \lambda\left(\|P\|_F^2 + \|Q\|_F^2\right) \tag{6.2}$$

where $L$ is a loss function that should be small if $P$ and $Q$ correctly model the relationships of $v_i \in V$ and $t_i \in T$, and is defined as follows:

$$L\left(v_i, t_i, P, Q\right) = \min_{h\in\mathbb{R}^r} \frac{1}{2}\left(\|v_i - Ph\|_2^2 + \|t_i - Qh\|_2^2 + \lambda\|h\|_2^2\right) \tag{6.3}$$

where $h \in \mathbb{R}^r$ is the latent factor representation for the pair $(v_i, t_i)$. This function finds the latent factor representation that allows to reconstruct simultaneously the two original feature vectors with minimal error. This formulation suggests a solution strategy that alternates the optimization of $P$, $Q$ and $H$. Two possible solutions to this problem have been proposed recently; the first approached based on a Nonnegative Matrix Factorization (NMF) framework, using multiplicative updating rules [117], and the second, based on first order gradient descent [129].

### 6.3.2 Online Multimodal Matrix Factorization

In this work we are not interested in optimizing the empirical cost function for learning $P$ and $Q$, but instead, the expected cost:

$$f(P,Q) := \mathbb{E}_{(v,t)}\left[L(v,t,P,Q)\right] = \lim_{\ell\to\infty} f_\ell(P,Q) \tag{6.4}$$

where the expectation is taken relative to the joint probability distribution $\mathbf{p}(v,t)$, since these feature vectors are not observed alone, but in pairs or aligned. It has been shown that Stochastic Gradient Descent (SGD) algorithms can directly optimize the expected cost function of a learning problem, achieving faster convergence than second-order batch methods [122]. Furthermore, Bottou [130] shows that SGD is a bad optimization algorithm in the sense that it would require longer times to reach a predefined accuracy on a training set, but requires very little time to reach a predefined expected risk. This property makes it very suitable for large scale learning problems, i.e., when the limiting factor is computing time rather than number of training examples.

Figure 6.2 illustrates the operation of the online matrix factorization algorithm for updating $P$, one of the model parameter matrices. Assuming that the number of columns in the matrix $V$ is very large or potentially infinite, we can not solve the problem for the full matrix. Instead, the columns of the matrix are randomly permuted, and then, they are used one at a time for learning the structure of the matrix $P$. To complete the equation, the column vectors of the matrix $H$ are also computed on the fly with respect to the observed vector $v$ and the current solution for $P$. After updating the model parameters in the matrix $P$, both vectors, $v$ and $h$ are discarded, and new vectors are scanned.

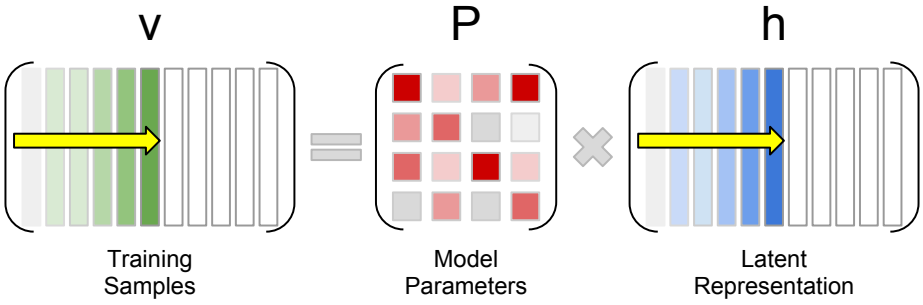



Figure 6.2: Illustration of an online matrix factorization algorithm that randomly scans a large matrix one vector at a time to update the parameters of the model.

#### 6.3.2.1 Algorithm Outline

The proposed algorithm uses stochastic learning approximations to make the multimodal decomposition scalable to very large training data sets. It follows the structure of an alternating optimization algorithm, taking turns between two main stages: 1) computing the latent factor representation for one observed vector, and 2) updating the modality transformation matrices. The procedure is presented in Algorithm 6.1.

The algorithm starts by randomly drawing a pair of feature vectors from the multimodal probability distribution. Since the stochastic algorithm does not need to remember examples from previous iterations, it can process large data sets with very low memory usage. It can be shown that Equation 6.5 is the solution for the problem in Equation 6.3. Also, a proof of convergence of the algorithm can be obtained by considering the independent solutions for $P$ and $Q$.

The first stage of the alternating algorithm consists on computing the latent factor representation for the observed pair of visual and text features. To compute this representation, the analytic solution of Equation 6.3 is obtained by fixing $P$ and $Q$ as the versions of the matrices known in the iteration $k$. The resulting solution is expressed in terms of these two matrices, the observed feature vectors and the regularization parameter, as shown in Equation 6.5. Notice that this latent representation is computed for just one example taken from the distribution, so it is not affected by the step size of the stochastic algorithm. In addition, in the same way as the feature vectors are discarded after the iteration $k$ is completed, this latent representation is discarded as well.

The second stage consists on updating the modality transformation matrices. In this part of the algorithm, the stochastic approximation is carried out by using the information of the observed visual and text features together with the corresponding latent representation. The updating rules for $P$ and $Q$ are the first order gradient descent rules, obtained from the partial derivatives of the objective function in Equation 6.1, and expressed in terms of one observed example, as shown in Equation 6.7 and 6.8. The step size used in this algorithm is a decreasing rate [130] that depends on the number of iterations, the regularization parameter and an initial learning rate (see Equation 6.6).

**Algorithm 6.1** Online Multimodal Matrix Factorization (OMMF)

INPUTS:
  $p(v,t)$: multimodal data distribution
  $\lambda \in \mathbb{R}$: regularization parameter
  $P_1, Q_1$: initial transformation matrices
  $\gamma_0$: initial step size
  $T$: number of iterations.
BEGIN
1. FOR $k$=1:$T$
2.  Draw $(v_k, t_k)$ from $p(v,t)$
3.  Compute latent factor representation:

$$h_k = (\lambda I + P_k^T P_k + Q_k^T Q_k)^{-1}(P_k^T v_k + Q_k^T t_k) \tag{6.5}$$

4.  Compute current step size:

$$\gamma_k = \gamma_0/(1 + \gamma_0 \lambda k) \tag{6.6}$$

5.  Update matrix $P$:

$$P_{k+1} = P_k + \gamma_k \left( (v_k - P_k h_k)\, h_k^T - \lambda P_k \right) \tag{6.7}$$

6.  Update matrix $Q$:

$$Q_{k+1} = Q_k + \gamma_k \left( (t_k - Q_k h_k)\, h_k^T - \lambda Q_k \right) \tag{6.8}$$

7. ENDFOR
8. return $P_{k+1}, Q_{k+1}$
END
OUTPUTS: $P$ and $Q$

#### 6.3.2.2 Minibatch Extension

A slight variation of this algorithm is obtained by using several samples at each iteration instead of using only one. This is known as the minibatch extension [128]. The proposed algorithm benefits particularly from this extension, since the operation in the first stage (Equation 6.5) involves the computation of an inverse matrix. Experimental results show faster execution when using minibatches instead of single examples, and also a better numerical stability for the solution of Equation 6.5.

#### 6.3.2.3 Weighted Modalities

The formulation of the objective function admits weights for data modalities for controlling their relative importance when learning the multimodal latent factors. In this work, we explore a weighted variation of this algorithm by using a convex combination between the terms for the reconstruction error of the modalities in Equation 6.1. This introduces an additional parameter $\alpha$ as follows:

$$\min_{P,Q,H} (1-\alpha) \|V - PH\|_F^2 + \alpha \|T - QH\|_F^2 + \lambda \left( \|P\|_F^2 + \|Q\|_F^2 + \|H\|_F^2 \right) \tag{6.9}$$

Thus, the new updating rules are:

$$\begin{array}{rcl} h_k & = & (\lambda I + (1-\alpha) P_k^T P_k + \alpha Q_k^T Q_k)^{-1} ((1-\alpha) P_k^T v_k + \alpha Q_k^T t_k) \\ P_{k+1} & = & P_k + \gamma_k \left( (1-\alpha) \left( v_k - P_k h_k \right) h_k^T - \lambda P_k \right) \\ Q_{k+1} & = & Q_{k+1} = Q_k + \gamma_k \left( \alpha \left( t_k - Q_k h_k \right) h_k^T - \lambda Q_k \right) \end{array} \tag{6.10}$$

This parameter can be useful to emphasize the algorithm on learning from the cleaner or more discriminative modality. The parameter $\alpha$ has been defined here to represent the weight of the text data. The weight of the visual modality corresponds to the complementary value of the convex combination.

#### 6.3.2.4 Other Extensions

Additional extensions of this algorithm also include Nonnegative Matrix Factorization, which can be easily incorporated by introducing a projection function that maps a point back to the positive feasible region in each iteration [131]. Also, the algorithm can be used to factorize an input with only one data modality, in which case the unused transformation matrix can be just dropped from the algorithm in every place. This is equivalent to giving a weight of one (1) to the required data modality and zero (0) to the other.

### 6.3.3 Image Indexing and Search

The proposed Online Multimodal Matrix Factorization (OMMF) algorithm learns the modality transformation matrices to project visual and text data to the latent space. After processing a large data set, these two matrices contain the knowledge extracted

from the joint probability distribution of the multimodal data, and codify the underlying structure and connexions between visual features and text annotations. Now, to effectively index the image collection, an additional processing step is required to generate the final multimodal representation.

For examples taken from the training database, that have both, visual features and text data, the multimodal latent representation is computed on the fly using the available versions of the transformation matrices $P_k$ and $Q_k$, in each iteration $k$. However, that latent representation is discarded as these matrices evolve during the training phase. Then, an additional pass through the data is required to compute the multimodal latent representation again for every image, using Equation 6.5. The same can be done for new images that were not part of the training stage, but that have both data modalities and are required to be searchable in the system.

An 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, acquired using a camera phone or a similar device. A new image without text can be projected to the multimodal latent space by solving the following optimization problem:

$$\min_h \|v - Ph\|_2^2 + \xi \|Qh\|_2^2 \tag{6.11}$$

where $h$ is the latent representation and $\xi$ is a regularization parameter to penalize the magnitude of the recovered text annotations. Notice that this objective aims to minimize the reconstruction error of the observed visual features, using the learned latent representation. Also, the latent representation is useful to recover the predicted text data directly. The following analytic solution can derived for the problem in Equation 6.11:

$$h = \left(P^T P + \xi Q^T Q\right)^{-1} P^T v \tag{6.12}$$

The projection matrix of visual features to the latent space can be precomputed for fast indexing of new images. In our experiments, we evaluate the retrieval performance under the query-by-example paradigm, i.e., queries are image examples without text descriptions. This approach is useful to assess the generalization ability of the proposed learning algorithm to construct image representations for pictures with partial data, which is a very common scenario in real world applications.

Finally, since all images are represented in the multimodal latent space, the dot product is used as matching criterion, as it estimates the extent to which two vector representations share the same latent factors.

## 6.4 Experiments and Results

An experimental evaluation was carried out to assess the properties of the proposed model in practice. Three different image collections were selected, with varying sizes and structures. The evaluation includes convergence properties to determine how fast

the algorithm achieves a stable solution. The quality of the factorization was measured using precision scores from information retrieval, and comparison with other methods were made. Also, the impact of different parameters is discussed and experimentally observed, and finally, computational time over small and large image colletions is also reported.

### 6.4.1 Data sets

The performance of the proposed framework has been evaluated using three different image retrieval benchmarks: Corel 5k [13], MIRFlickr [98] and ImageCLEFmed 2011 [132]. The size of these benchmarks are $5,000$, $25,000$ and $230,000$ images respectively, introducing 3 different orders of magnitude on the amount of available data for our experiments. All these collections have attached texts for images in the training set and include test sets or predefined queries to evaluate retrieval performance.

#### 6.4.1.1 Corel 5k

With 5,000 images from the Corel collection, this is a small scale data set used as benchmark in different image retrieval experiments. The data set is split in 3 standard subsets to allow other researchers reproduce results: 4,000 images for training, 500 images for validation and 500 images for test. This collection is organized in 50 different categories with 100 images in each category. In addition, the collection was manually annotated by researchers exploring image auto-annotation techniques. These annotations are composed of a dictionary with 374 terms, and each image has between 2 to 5 attached terms.

For this collection, a bag-of-features is used to represent visual contents, using the same setup as in [72]. Small patches are densely sampled from each image and represented using the Discrete Cosine Transform (DCT) coefficients for the three RGB color channels. A dictionary of 2,000 codeblocks was built and each image is represented as a histogram with their occurrences. For text features, the dictionary of 374 terms has been employed as free text annotations, building a binary vector for each image that indicates whether the corresponding term is associated or not. Validation and test images are assumed to have no text annotations and are used as queries as well. A database image is considered relevant to a query if it shares the same global category with the query.

#### 6.4.1.2 MIRFlickr

This collection is composed of 25,000 images crawled from the Flickr website [98]. This data set has different subsets primarily composed of 15,000 images for training and 10,000 for testing. These images have been labeled by researchers using 39 different annotations, some of them considered as potential and others as relevant. In our experiments, we evaluate the system using all 39 annotations. In addition, the data set has user tags taken from the Flickr website, which have been assigned by web users. These

tags are very noisy and unstructured, and have been used in this work as the text data modality.

In the same way as the Corel 5k data set, in this collection the visual contents are represented using a bag of DCT features with a dictionary of 2,000 codeblocks. However, the text data is processed in a different way, using standard natural language processing techniques. We employed stop-word removal and stemming to build the list of indexing terms, and finally, the Term-Frequency Inverse-Document-Frequency (TF-IDF) weighting scheme was applied. This resulted in a vector space model of approximately 1,300 dimensions. In our experiments, we divided the training set to use 10,000 images to populate the database, and the remaining 5,000 images as queries. The rest of the experimental evaluation follows the setup described in [98].

#### 6.4.1.3 ImageCLEFmed 2011

This is one of collections used in the CLEF campaign for multilingual and multimodal information retrieval. This collection is composed of 230,000 medical images extracted from scholarly articles in medicine [132]. Each image has associated text extracted from the corresponding paper, including the title and caption. The type of images is very wide, including x-ray, MRI, PET, microscopy, photos, graphs and diagrams, among others. Each year, the organizers propose a set of 30 queries which have to be used to rank the corresponding collection. When the campaign is done, a set of relevance judgements are used to rank the experiments of research groups.

For this collection, we computed a Spatial Pyramid of CEDD features, which consists of a recursive partition of images in quadrants [133], and extracting CEDD features [134] in each subregion. We employed a pyramid with 2 levels, which results in 21 spatially distributed regions, ending up in a visual representation with 3,024 features. These features were selected due to the large variety of image modalities in the collection, and the good performance that other researchers, using this descriptor, have reported for this task. Text features are built using a similar procedure for the MIRFlickr data set, but applying the BM25 weighting scheme [2] instead of TF-IDF. This resulted in a vector space of approximately 13,000 different terms. During search, an image is considered relevant according to the relevance judgements employed in 2011.

### 6.4.2 Convergence

The first set of experiments are oriented to observe the speed of convergence of the proposed algorithm by comparing it to two alternative multimodal decomposition algorithms that optimize the empirical cost function presented in Equation 6.3: a first order gradient descent algorithm (batch) [129], and a multimodal Nonnegative Matrix Factorization (mNMF) [72]. All three algorithms evaluated in this experiment are formulated to minimize the Frobenius norm of the difference between the original and reconstructed matrices. For these experiments, the data to be processed needs to fit in main memory since the updating rules of the batch and mNMF algorithms assume the

availability of the full matrices to compute parameter updates.

Thus, this evaluation was run on the Corel 5k and MIRFlickr data sets, which have the following dimensionalities:

- **Corel 5k:** using 4,000 training examples. Visual matrix $V \in \mathbb{R}^{2000\times4000}$ with 2,000 features, text matrix $T \in \mathbb{R}^{371\times4000}$ with 371 terms, visual transformation $P \in \mathbb{R}^{2000\times50}$ and text transformation $Q \in \mathbb{R}^{371\times50}$ with 50 latent factors.

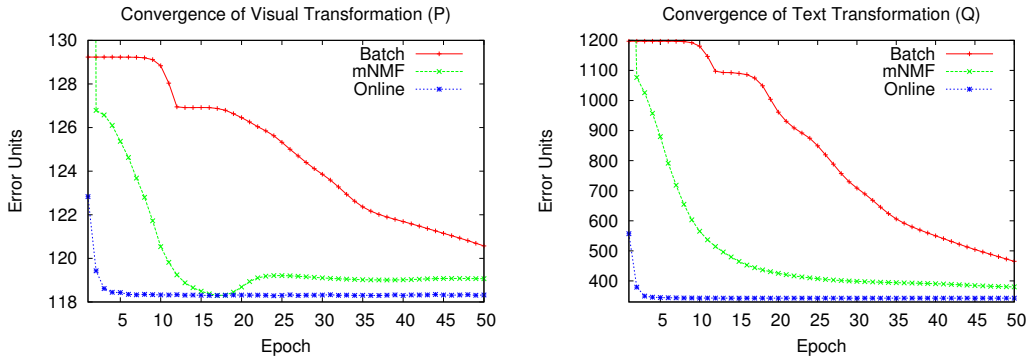
- **MIRFlickr:** using 10,000 training examples. Visual matrix $V \in \mathbb{R}^{2000\times10000}$ with 2,000 features, text matrix $T \in \mathbb{R}^{1391\times10000}$ with 1,391 terms, visual transformation $P \in \mathbb{R}^{2000\times500}$ and text transformation $Q \in \mathbb{R}^{1391\times500}$ with 500 latent factors.

Before running the experiments, the input matrices $V$ and $T$ are normalized in order to have columns with L1-norm = 1, which, for the kind of features used in this work, produces vectors of probability distributions. Matrices $P$ and $Q$ are randomly initialized within the interval $[0.0, 0.1]$ for all algorithms, and in all cases, the initialization was made using the same random seed. In this evaluation, one epoch is considered to be a complete usage of the training set for updating the model parameters ($P$ and $Q$). Experiments were run for 50 epochs with all algorithms and the reconstruction error of the visual and text matrices were measured on the training data at each epoch.

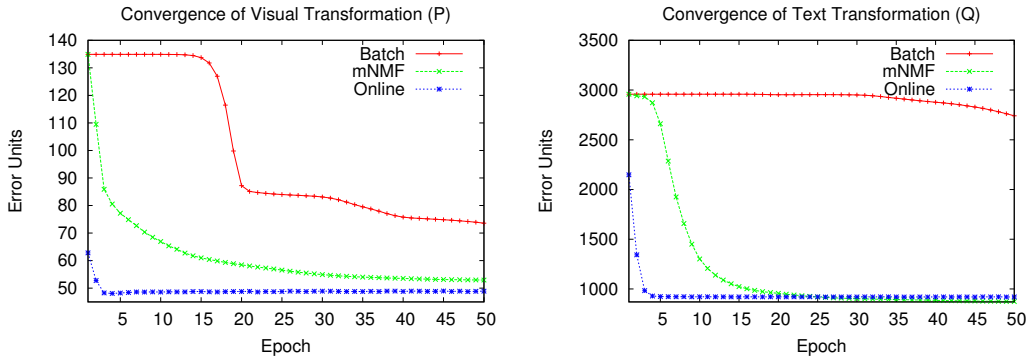
Figure 6.3 presents the error evolution of the two data modalities along the number of epochs. The proposed OMMF converges very soon before 5 epochs for both data modalities. The batch algorithm shows the worst convergence rate, something expected since it is based on a first order gradient descent strategy. The mNMF algorithm is in general faster to converge than the batch approach. Notice that mNMF tends to converge to the same error rate achieved by the OMMF algorithm in both modalities, which is explained by the similarity of their objective functions.

These algorithms deal with the simultaneous decomposition of two inputs: visual and text matrices. We can say that an algorithm converges as soon as the error rate for both matrices has achieved a steady solution. For the mNMF algorithm, the plots show that convergence may be apparently reached for one of the data modalities, but the other still needs some improvement. In the frame of 50 epochs run for these experiments, the batch and mNMF algorithms have not converged yet, likely requiring a number of additional epochs to reach the solution.

These results show that the proposed algorithm can achieve a very fast convergence rate, with the additional advantage of little memory consumption, since the full matrix is not required for training. It is important to note that minimum reconstruction error is not the ultimate goal of the proposed algorithm. The actual goal for the proposed strategy is to learn useful relationships between visual features and text annotations to improve image search performance. Therefore, image search experiments were conducted to observe the evolution of retrieval performance.

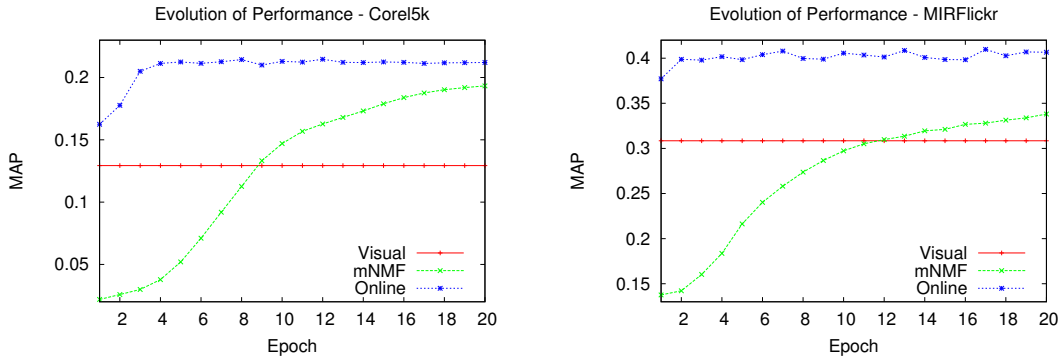



(a) Corel5k data set




(b) MIRFlickr data set

Figure 6.3: Convergence speed of the OMMF algorithm compared to the batch and mNMF algorithms. Top row: Corel 5k. Bottom row: MIRFlickr.




(a) Corel5k data set (b) MIRFlickr data set

Figure 6.4: Evolution of the Retrieval Performance.

### 6.4.3 Evolution of Retrieval Performance

Results in the previous subsection showed that very few epochs are needed to achieve a stable model. The following experiments aim to evaluate the quality of performance on the search task by measuring retrieval precision at each epoch of the algorithm. So, the solution obtained by the algorithm at a particular epoch is used to index query images as well as database images and compute the corresponding rankings. The resulting rankings are objectively assessed using Mean Average Precision (MAP), a common performance measure for information retrieval [2].

The same procedure is conducted with the mNMF algorithm for comparison in the evolution of performance, running the experiment for 20 epochs in each data set. Also, a baseline with no learning is employed to compare the gain in performance of the two evaluated algorithms, which consists of direct visual matching between query images and database images using visual features.

Figure 6.4 presents plots of the evolution of performance for the Corel5k data set as well as the MIRFlickr data set. These results provide another point of view for the convergence evaluation presented in the previous subsection and also confirm the main finding: very few epochs of the proposed algorithm are needed to achieve the best performance. Experiments on both data sets show that starting from the third epoch the proposed online algorithm reaches a stable retrieval performance, which does not improve or decrease significantly if it keeps running for more epochs.

Notice that OMMF improves upon the visual-matching baseline since the very first epoch, i.e., with just one pass over the training data an improved performance is achieved. In contrast, the evolution observed for the mNMF algorithm starts providing a performance below the baseline, and requires about 10 epochs to reach it. It is very likely that many more epochs are needed to get a similar performance to the one obtained by OMMF, since the progress observed in 20 epochs is slowing down asymptotically.

Hence, the proposed OMMF algorithm demonstrates useful properties to learn

visual-text relationships from large image collections. These results suggest that an iterative process is not necessary, as it is usually expected for matrix factorization algorithms. The result is also consistent with the properties of stochastic gradient descent algorithms provided by Bottou and Bousquet [122]. They demonstrate that using online learning, a single pass on the training set is asymptotically efficient when the size of the training sample is very large.

The experiments presented in this subsection suggest that two to five passes are needed on the Corel5k and MIRFlickr data sets, mainly due to their relatively small size (thousands of samples). However, the algorithm still enjoys fast convergence even in those small collections. Also, notice that the performance at the first iteration starts closer to the best value on the MIRFlickr collection in contrast to Corel5k, showing how more data can help to get better results with few epochs. Separate experiments conducted with the ImageCLEFmed data set, which is composed of hundred of thousands of examples, also showed that only one pass over the training data is enough to achieve the best performance.

### 6.4.4 Parameter Tuning

The proposed OMMF algorithm has a set of parameters that can impact the quality of the resulting model. Some of these parameters model the underlying structure of the data and others control convergence properties. These parameters are set in an empirical basis depending on the specific data collection. The following results aim to provide a better understanding of the role of these parameters, and potential strategies to explore their impact in the model.

#### 6.4.4.1 Controlling Convergence

During training, there are three parameters that can be combined to control the convergence of the algorithm. These parameters, in order of importance, are: initial step size $\gamma_0$, regularization parameter $\lambda$, and minibatch size. $\gamma_0$ is the most important parameter since it can determine whether the algorithm converges or not, regardless the other two. According to Equation 6.6, $\gamma_0$ is the seed for estimating the decreasing function of the step size. If $\gamma_0$ is too big the algorithm may diverge or converge too early, if $\gamma_0$ is too small convergence may take an unusual number of epochs to converge.

Remember that only one epoch should be enough to achieve both, convergence and good performance. So, when exploring the value of $\gamma_0$, convergence to a reasonable error rate should be observed with one epoch only. In our experiments, we explored values of this parameter of the form $2^i$ with $i \in [-7, 3]$. For each data set a different good parameter was found, whose values are reported in Table 6.1.

The second important parameter is $\lambda$, which determines the amount of regularization on the learned matrices determining the quality of the resulting factorization. Since this penalizes the norm of the model matrices, it can give more importance to reconstructing the original matrices (a small value) or obtaining a more general, smooth model (a

Table 6.1: Set of parameters to control convergence used for each image collection. In general, these values provided good convergence rates as well as good retrieval performance.

| **Data set** | $\gamma_0$ | $\lambda$ | **Minibatch** | $\xi$ |
|---|---|---|---|---|
| *Corel5K* | 0.125 | 0.1 | 128 | 0.01 |
| *MIRFlickr* | 2.0 | 0.05 | 512 | 0.001 |
| *ImageCLEFmed* | 0.3 | 0.001 | 512 | 0.001 |

large value). By varying this parameter, the algorithm converges to different solutions resulting in different error rates and retrieval performances. We explored this parameter using powers of 10, and measured the performance of the resulting factorization. Results showed that no large regularization is needed in any of the three datasets, indicating that the algorithm can be set to reduce the reconstruction error with a small value of $\lambda$.

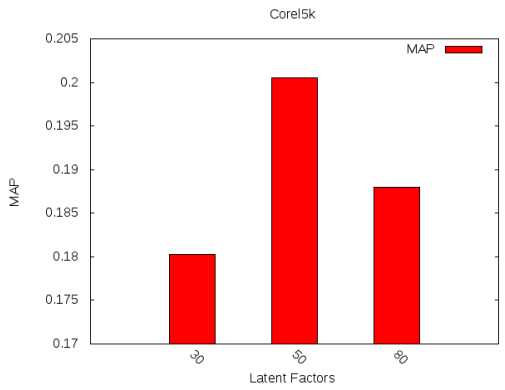
The third parameter to control convergence is the minibatch size, which can help to accelerate the speed of the algorithm. Usually, moving this parameter does not require to adjust the other two, unless a very large change is made. We explored this parameter using powers of two, from 8 to 512. Since this is the number of simultaneous vectors used during one update in the model, it speeds up the processing when its value is large, resulting in less computational load when scanning the whole data set. In general, we found that the size of the minibatch does not affect the quality of the factorization or the final error rate and its only advantage is on runtime speedup. An exception has to be made when the size of the minibatch is close to the number of samples in the training set, since this makes the algorithm to start behaving like a conventional batch learning algorithm.

A final regularization parameter is used in the function to project new images to the latent space. We explored the impact of this parameter, using again powers of 10, and found that a small value is best in all cases. The set of parameters that give good convergence properties on each data set are reported in Table 6.1.

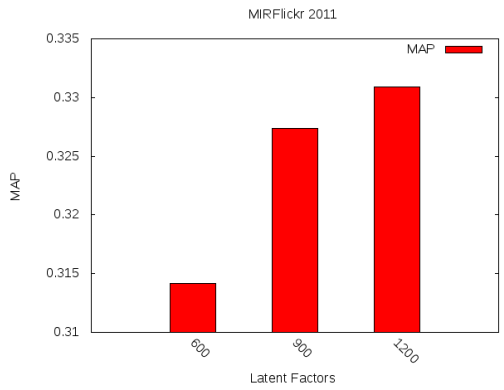
### 6.4.5 Unveiling Latent Collection Structure

In the proposed factorization model, the algorithm provides two parameters that are associated to the underlying structure of the image collection: the number of latent factors and the relative importance of each data modality. These two parameters have the potential of directly improving the performance for the image search task.

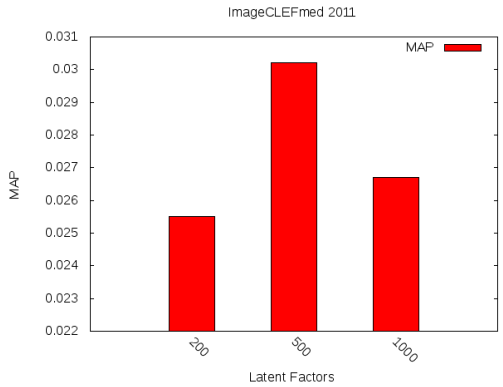
The proposed model is a tool to discover a common latent space for visual and text data, in which their relationships are codified. This is done through a decomposition of the original matrices by projecting the input data to a new space of lower dimensionality. A priori, the number of latent factors can be set as a portion of the features available in the input spaces to force correlations between them. However, setting this parameter carefully can greatly improve the performance of the underlying task. Figure




(a) Corel5k data set




(b) MIRFlickr data set



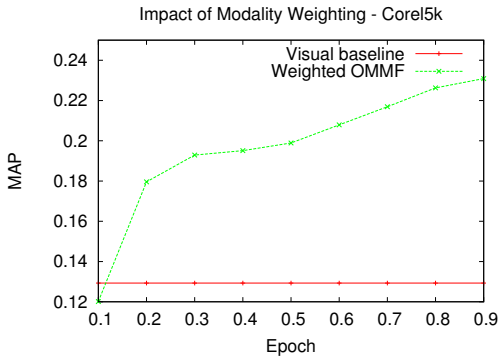

(c) ImageCLEFmed 2011

Figure 6.5: Impact of the number of latent factors in the retrieval performance for three data sets. Notice that each data set requires a different number of latent factors and the scale of useful dimensionalities is sensible to the data structure.

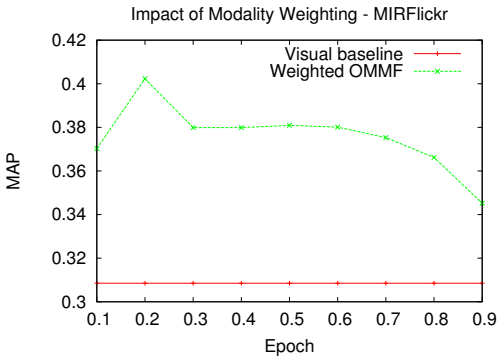
6.5 shows the impact of the number of latent factors for the three data sets used in this experimental evaluation.

Bars in the Figure represent Mean Average Precision (MAP) values obtained for experiments in each data set. The Corel5k data set requires very few latent factors, in the order of tens, showing a best performance with 50 latent factors. More or less factors do not perform as good, suggesting that 50 factors correctly model a semantic latent structure of this data. Actually, the Corel5k data set is organized in 50 categories that are used as ground truth. The MIRFlickr data set, on the other hand, requires significantly more latent factors to improve performance, in the order of thousands. Since this is a more complex data set, more latent factors help to organize patterns between visual features and text data. A similar pattern is shown in the ImageCLEFmed data set, which shows good retrieval response with 500 latent factors.

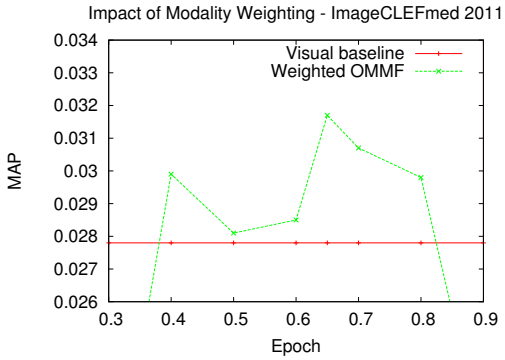
The other parameter that models underlying structure of the image collection is the




(a) Corel5k data set




(b) MIRFlickr data set




(c) ImageCLEFmed 2011 data set

Figure 6.6: Impact of moving the relative importance of each data modality on retrieval performance. The higher the text weight, the lower the importance of the visual modality. The performance is higher when the more informative modality is favored.

relative importance of the two data modalities, or modality weight $\alpha$ (see Equations 6.9 and 6.10). This is specified in the objective function with a convex combination of the terms associated to the reconstruction of the visual and text matrices. If the parameter is not used or set to 0.5, both data modalities are considered equally important to define the latent space. However, if the parameter is moved to favor one modality, the algorithm is biased to include more information extracted from that modality into the structure of the latent space.

This is specially useful when one data modality is cleaner or more discriminative than the other, allowing to boost the search performance and obtaining more accurate results. Experiments were run to evaluate the impact of this parameter in the three data sets, in all cases varying the parameter from 0.1 to 0.9 to gradually move the relative importance from the visual modality to the text modality. A total of 5 experiments were run and the obtained MAP score was averaged and plotted. Figure 6.6 presents the results of these experiments.

The results show that the Corel5k data set has a very strong preference for the text

data modality, which is organized in a dictionary with 374 keywords. These keywords are clean descriptions of what can be seen in these images and correlate very well with the 50 ground truth categories used for evaluation. That explains the better performance when the text weight is higher. The MIRFlickr data set, on the other hand, obtains a drop in performance when the text modality is prefered. This is explained by the noisy tags that are assigned to Flickr images by users, which are not controlled by any means, and so, extracting information from that data is harder. It is interesting to note that the best performance for the MIRFlickr collection is achieved giving a weight of 0.2 for text data, and the obtained MAP is significantly better than using 0.1, which is mostly visual. So, even though it is hard to extract useful knowledge from unstructured text, the algorithm can effectively make it.

The final plot shows results for the ImageCLEFmed data set, which presents a non-smooth balance between both modalities. However, the best performances are slightly biased towards the text modality, achieving the best performance with 0.65. Notice that for this data set, the latent space is not very useful when the parameter is set to the extremes of the axis, i.e., when one modality is given too much preference. This confirms that the good performance is achieved by exploiting multimodal interactions in the collection.

Note that giving more preference to one data modality does not mean that the other modality is useless. In the extreme case when the text modality is prefered over the visual one (such as in the Corel5k data set), it is still very important to include the visual modality information in the latent space, since new images are projected based on what is observed in their visual features. For more complex collections, such as MIRFlickr and ImageCLEFmed, good balances between both modalities are achieved by adjusting the parameter and exploiting the particular structure of the collection.

### 6.4.6 Image Search Benchmark

The goal of OMMF is to learn a new representation for images that combines visual data and text annotations. The multimodal representation is herein used for indexing images in a retrieval system that allows to search using example pictures, that is to say, the system is asked to retrieve images semantically related to an unknown, new image that has no keywords or text descriptions.

Since query images are used without text to search the database, the first step towards obtaining a list of results is to project these query images to the multimodal latent space, following the procedure presented in Section 6.3.3. Afterwards, images are matched in the multimodal latent space using the dot product as similarity function. The list of results is evaluated by measuring the precision of the output, i.e., a score that indicates the extent to which the results are correct according to the predefined relevance judgements. This score is the Mean Average Precision (MAP).

Previous subsections have presented experimental evaluations using as performance measure the precision of the retrieval task. In this section, experiments are oriented to evaluate the potential of the proposed algorithm to improve retrieval precision with

Table 6.2: Comparison of retrieval performance using different methods.

| Strategy | MAP Corel5k | Improv. | MAP MFlickr | Improv. |
|---|---|---|---|---|
| Visual search (baseline) | 0.1293 | N.A | 0.3085 | N.A |
| mNMF (Akata et al. [117]) | 0.2099 | 62.3% | 0.3672 | 19.0% |
| aNMF (Caicedo et al. [72]) | 0.2277 | 76.1% | 0.4258 | 38.0% |
| OMMF | 0.2352 | 81.9% | 0.4044 | 31.1% |

respect to other algorithms and benchmarks, under the same experimental conditions.

#### 6.4.6.1 Small Scale Search

This evaluation aims to compare the performance of OMMF with respect to two recently proposed multimodal indexing approaches: the multimodal NMF (mNMF) [117] and asymmetric NMF (aNMF) [72, 125]. The mNMF algorithm is based on Nonnegative Matrix Factorization and optimizes the same objective function modelled in this work, with the difference that non-negativity constraints were added to the optimization problem and multiplicative updating rules were derived for obtaining the solution.

The aNMF algorithm is also based on Nonnegative Matrix Factorization but has two important differences with the proposed approach: first, this is a two stages algorithm that decomposes the text matrix first, and then adapts the visual data modality to the structure of the latent space. Second, these decompositions solve for the minimum Kullback-Leibler divergence between the original and reconstructed matrices. It has been shown that an NMF decomposition based on the KL divergence is equivalent to the Probabilistic Latent Semantic Analysis (PLSA) algorithm [105], so, this approach is representative of other multimodal image indexing strategies based on PLSA [118, 123].

Since these two algorithms are based on updating rules that use the full input matrices, they are not suited for large scale learning. Thus, the Corel5k and MIRFlickr data sets are used in this experiment. An additional strategy for searching the image collection using the query-by-example paradigm is directly matching visual contents. Therefore, visual matching using the Histogram Intersection similarity [113] is used as the baseline method that does not require any kind of learning for indexing images. This allows to assess the contribution of using multimodal representations for image search.

All algorithms are fed with the same underlying visual representation for images as well as the same text representation. For details about visual and text features for each data set, please refer to Section 6.4.1. The number of latent factors and other parameters that apply were optimized to get the best performance for each algorithm. The mNMF and aNMF algorithms were run for a maximum of 100 epochs each. Table 6.2 reports MAP scores for each of the evaluated algorithms for the Corel5k and MIRFlickr data sets.

These results show that all three multimodal indexing algorithms are able to improve

the search response with respect to the baseline in a significant way. This important improvement demonstrates the potential of building multimodal representations for image search. The best performance on the Corel5k data set was obtained by OMMF with about 80% of improvement with respect to the baseline and about 3% with respect to the second best multimodal algorithm.

On the MIRFlickr data set, OMMF also shows a significant improvement that outperforms the baseline as well as the mNMF algorithm. However, the best performance on this data set was obtained by aNMF. One of the reasons the proposed OMMF is able to outperform mNMF, even though both optimize a very similar objective function, is because of its ability for fast convergence. Experimental results on Sections 6.4.2 and 6.4.3 showed that mNMF forwards rather slowly and approaches the solution asymptotically.

The results from OMMF and aNMF on the MIRFlickr data set are in the same order of MAP score as well as percentual improvement, which is more than 30% with respect to the baseline. The relative difference between both results is about 5% in favor of aNMF. The good performance presented by this algorithm can be mainly explained by the difference in the objective function of the underlying optimization problem. Note that aNMF, however, cannot scale up to process large data collections as easily as the proposed algorithm does. So, even though its potential to model multimodal relationships in small data collections is good, bringing its applicability to real world scenarios can be very hard.

#### 6.4.6.2 Large Scale Search

The following experiments aim to demonstrate the potential of OMMF for large scale image indexing. We conducted experiments with the ImageCLEFmed 2011 database, which is composed of 230,000 images, and compared the obtained results with those reported after the challenge ended. For these experiments, the full image collection was used for training, i.e., no sampling was applied for computational efficiency, so the algorithm could take full advantage of all the available data.

The ImageCLEFmed challenge provides an image collection of medical images extracted from scholarly journals and propose a set of 30 queries to retrieve the most relevant ones from that collection. Images in the database are provided with text surrounding them on papers, such as caption and titles. Queries are designed to also have both, example images as well as text descriptions. There are three strategies to answer queries in the ImageCLEFmed challenge: textual, visual and mixed. In our setup, we evaluate the performance of visual queries.

Table 6.3 presents performance measures obtained for this experiment. We started with a visual baseline whose MAP score is 0.0278, using Spatial Pyramid CEDD features (see Section 6.4.1) and the Tanimotto coefficient as similarity measure. Then, a multimodal space was learned from the collection by applying the OMMF algorithm, and then a ranking is computed in the latent factors space. The MAP score for the 30 queries increased to 0.0337, 21% more regarding the baseline, which is a significant

Table 6.3: Retrieval Performance for the ImageCLEFmed 2011 challenge.

| Strategy | MAP | # Relevant Images |
|---|---|---|
| Visual Baseline | 0.0278 | 682 |
| Online Multimodal Factorization | 0.0337 | 659 |
| Top 1 (Visual) @ ImageCLEFmed 2011 | 0.0338 | 717 |
| Top 2 (Visual) @ ImageCLEFmed 2011 | 0.0322 | 689 |

improvement for this task. The Table also reports the top 2 results from the competition of the year 2011 [132], showing that our result is very close to the top 1. These top results were completely based on visual features, so by integrating them in our framework, additional improvements may be observed.

Notice that to learn multimodal interactions from the given collection, the input matrices for this task are in the order of 230,000 columns and 20,000 rows, which do not fit in main memory. Therefore, other multimodal indexing algorithms, such as mNMF, aNMF or others based on PLSA, cannot be directly applied to this collection.

### 6.4.7 Computational Complexity

Matrix factorization algorithms can be expensive to compute because of their intrinsic computational complexity. For a multimodal matrix factorization algorithm the time complexity is $O(rp\ell k)$, where $r$ is the number of latent factors, $p = n+m$ is the number of visual and text features, $\ell$ is the number of training examples, and $k$ is the number of epochs. This is true for the proposed OMMF as well as NMF-based algorithms, due to matrix multiplications that dominate the algorithm complexity.

The proposed algorithm reduces the computational complexity of a multimodal factorization in two ways: first, memory requirements to compute parameter updates are very small. Two things are needed to be in memory, on the one hand the model parameters (matrices $P$ and $Q$), and on the other hand, a small number of vectors that are currently being processed. This makes the applicability of the algorithm very suitable for large scale collections, since there is no practical limit of memory for scanning large image databases. A potential implementation of this algorithm for a very large image collection can even extract visual and text features from images on the fly, use them for learning and then discard everything to go on with new samples.

The second advantage, but not least important, is that the number of epochs in the algorithm is reduced to a constant, cutting down dramatically time complexity. Our algorithm has been designed on top of stochastic learning theory that is guaranteed to converge in a single pass on the training set [130], and this property was experimentally verified in previous subsections of this paper. While other algorithms require hundreds of iterations over the training data, the proposed algorithm requires only one full scan for large data collections or very few for small collections.

We implemented the OMMF algorithm as well as the two other NMF-based algorithms in Matlab to conduct small scale experiments. An additional implementation

Table 6.4: Wall clock times to complete the multimodal decomposition for the three data sets.

| | **Corel5k** | **MIRFlickr** | **ImageCLEFmed** |
|---|---|---|---|
| Vectors | 4,000 | 10,000 | 230,000 |
| Factors | 50 | 1,200 | 600 |
| Visual Features | 2,000 | 2,000 | 3,024 |
| Text Features | 371 | 1,391 | 13,000 |
| mNMF | 130.21 seg | 113.15 min | — |
| aNMF | 128.93 seg | 58.52 min | — |
| OMMF | 2.11 seg | 7.01 min | 1.49 hrs |

was written in Java to process large scale experiments, since this programming language offers built-in libraries to naturally control dynamic I/O. A native interface for BLAS, named jblas[1], was used inside the Java implementation, since the BLAS package offers state-of-the-art performance for linear algebra computations.

Table 6.4 reports wall clock times required by all algorithms to learn the model that presented the best performance in the last subsection. The mNMF and aNMF were set to run for 100 epochs, since this was the number of epochs used to get the reported precision results. Reported times are the result of running all algorithms in a single CPU at 2.8Ghz. The size of each data set is also reported to observe how the algorithm complexity grows with the theoretical variables that affect it. NMF-based algorithms take about one hundred seconds to process the Corel5k data set and about one hour to process the MIRFlickr collection. In contrast, the OMMF algorithm makes 2 epochs on the Corel5k data set in 2 seconds, and 5 epochs on the MIRFlickr data set in 7 minutes.

In addition, OMMF was the only one able to process the ImageCLEFmed data set due to the memory constraints that apply for other algorithms. The OMMF was set to run for one single epoch for processing this large collection, finishing in less than one and a half hours using the Java implementation on a single CPU. That was enough to obtain the retrieval precision reported in the last section. Assuming that the full matrix of the ImageCLEFmed data set fits in memory, the computing time required by aNMF or mNMF to process this collection would be in the order of days. These results demonstrates the potential of the proposed algorithm to scale up to very large data collections.

## 6.5 Discussions

The experimental evaluation presented in this Section has shown that OMMF can build improved multimodal image representations for image search and achieve very competitive retrieval performance. Importantly, that is accomplished with very low computational resources, in contrast to other multimodal learning algorithms. The

[1]http://www.jblas.org

following Subsections present discussions about the main characteristics of the proposed OMMF algorithm.

### 6.5.1 Modelling Latent Factors

The proposed algorithm finds multimodal latent factors from an aligned image-text collection, by learning the relationships between both data modalities. Different strategies have been proposed to model the structure of image-text relationships in the same direction, mainly based on latent factors using probabilistic models [15, 31, 124, 118] or matrix factorization [98, 117, 72]. Most of these strategies work under the assumption that visual features and text annotations share a common latent factor space. From a probabilistic point of view, some of them have differences and commonalities in the underlying generative model.

The model proposed by Blei and Jordan [15] assumes that the visual modality is generated first, and the text modality comes later. On the other hand, Caicedo et al. [72] proposed an algorithm following the opposite direction, generating the text first, and the visual modality later. The approach presented in this paper considers both data modalities simultaneously, which follows the design of other similar works [81, 31, 117]. There is no clear consensus on the research community about the correct design for generating multimodal latent factors, and we believe this will require further fundamental research to better understand the impact of one design or another.

An important contribution of our work in that direction is a formulation that introduces a weighting term to account for the relative importance of each modality, allowing to better exploit asymmetries between both data sources and potential meaningful patterns. As the experimental results have shown, properly controlling this parameter allows to obtain major improvements for image search. This suggests that the discussion is not about which data modality is exploited first, but that their contributions to learn multimodal relations may vary. So, it is likely that each data modality may require its own modelling before finding correspondences between them, as was suggested by Putthividhy et al. [124].

An additional interesting result observed from the image search benchmark in Section 6.4.6 is that the aNMF algorithm [72] presented better performance than the proposed OMMF on the MIRFlickr data set. These algorithms present many differences, such as an asymmetric design and non-negativity constraints. However, we believe that the most important difference between both algorithms is the underlying cost function, which is the KL-Divergence for aNMF and the Frobenius norm for OMMF. Since a NMF algorithm that minimizes the KL-Divergence has been theoretically related to PLSA, this provides a more probabilistic grounding for aNMF. This suggest a possible improvement to OMMF, to optimize the KL-divergence instead of the Frobenius norm. This is part of our future work.

### 6.5.2 Large Scale Learning

This work makes an important contribution to the problem of large scale multimodal learning. Most of the previous proposals for learning multimodal relationships have been designed without considering a large scale setup [15, 31, 124, 118, 81, 117]. Sometimes, 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, since in a multimodal setup two sources of information have to be handled, and sharing information among computing nodes is usually an issue for several algorithms.

The approach that we follow in this work relies on recent theoretical advances for large scale learning, which provide guarantees about convergence and scalability [130, 122]. The online formulation of the matrix factorization algorithm that we have presented requires little computational resources to deal with large collections of data. Experimental results show that processing a collection with hundred of thousands of samples can be attained in less than two hours using a single CPU and no special memory requirements. It is important to highlight that our proposal focuses on using computational resources wisely instead of needing large infrastructures for running experiments. This actually makes room for further improvements on runtime and complexity, using for instance, a parallelized formulation of stochastic gradient descent [135].

Thus, our proposal breaks a computational barrier for large scale multimodal analysis, providing a feasible solution for huge image collections. Recent studies in machine learning and statistical data analysis have also shown that algorithms able to process abundant data quickly, can learn more precise models than algorithms optimized for specific small collections [136]. The proposed OMMF algorithm is an approach oriented to take advantage of that fact, by providing an unsupervised learning strategy for producing an improved multimodal image representation.

## 6.6 Conclusions

The Online Multimodal Matrix Factorization algorithm has been presented. The algorithm takes as input two aligned matrices of visual features and text occurrences representing the contents of an image collection, and learns the relationships between these two data modalities. These relationships are encoded in multimodal latent factors automatically extracted by the algorithm. New images without text annotations can be projected to the latent factors space, which implicitly predicts some keywords. The algorithm has been used in this work to index image databases, which are searched using a query-by-visual-example approach.

The main contribution of this work is the formulation of the multimodal matrix factorization as an online learning algorithm, which has little memory requirements and low computational load. This property makes the algorithm very convenient for processing and indexing large collections of images. An experimental evaluation was

conducted on three data sets with different amounts of images and associated texts. Results showed the ability of the algorithm to learn models with fast convergence rates and to produce very competitive retrieval performance. The largest experiment run in this evaluation involved 230,000 images with associated text descriptions, which were all included during the learning stage. The algorithm learned a model in 1.5 hours using a single CPU and resulted in a significant improvement in search precision.

This work has also provided some insights for modelling multimodal latent factors using matrix factorization. Experimental evidence shows that weighted modalities reveal structural patterns in a collection and can also improve the quality of the retrieval results. This suggest that there is a tradeoff for considering the influence of each data modality on the resulting latent factors. Also, the use of Kullback-Leibler Divergence as objective function in the minimization problems seems to offer promising improvements for learning, and that could be incorporated in the proposed algorithm structure. These are some future research directions, along with the study of parallelized stochastic gradient descent algorithms for matrix factorization.

# Chapter 7

# Conclusions

This thesis has presented models and strategies to address the problem of multimodal image indexing for building improved image search systems. This thesis focused on building a multimodal representation that combines the richness of visual data and the precise semantics of texts terms. Building this representations has been approached as a problem of learning the relationships that connect both data modalities in particular image databases. Different strategies were proposed and systematically evaluated on different multimodal image collections. The results show that introducing information extracted from the text modality in the image representation improves the quality of the retrieval results, even in scenarios where queries are exclusively visual.

Three research challenges were identified to learn multimodal relationships: the structure and chacracteristics of the text modality, the size of the database, and the type of application domain. Solutions to the main problem of multimodal learning have been proposed, considering various conditions for these three challenges. The main contribution of this research work is a family of algorithms based on matrix factorization, that have extended the notions of latent factor analysis to multimodal setups. The following subsections discuss different aspects of the addressed problems and of the strategies used to tackle them.

## 7.1 Image Search

This dissertation has studied the problems of image indexing to build useful visual search systems, making contributions to bridge the semantic gap through multimodal representations, and approaching practical associated problems such as the structure of the textual modality, the size of the database, and the application domain. The proposed methods have demonstrated to be effective for retrieving relevant images, and allowed to build content-based retrieval systems that deal with images without attached text in a principled way. This is very important to fully index a partially annotated collection of images, and specially to consider the query-by-example paradigm for searching.

A system to search using example images as queries have potential applications in

several domains, such as browsing the web or personal photo collections using example pictures acquired with the camera phone. In the medical practice, diagnostic images for new patients can be send as queries to retrieve similar cases from the hospital archive. Forensics and scientific research may also benefit from these tools. Other query paradigms can be seamlessly supported as well in the proposed frameworks, such as keyword-based queries, as was presented in Chapter 4. Combinations of keywords and example images are also allowed, and may be useful for instance, in specialized software readers that allow to search using images in a document, by sending the example figure as well its caption to the search engine.

All these functionalities allow to build improved image search systems, with the ability to understand visual examples and find related images with complete sense according to their high-level interpretations. The main contribution presented in this work was oriented to enable retrieval systems for this situation, by mining multimodal relationships in the target collections. We envision a multimodal search engine for images that can also process queries in natural language to state conditions on image contents, not in the surrounding text. Therefore, building an image search engine that can manipulate concepts and visual patterns to provide full content-based access is still an important open research problem.

## 7.2 Learning Multimodal Relationships

In this work, multimodal image indexing is approached by designing extensions of matrix factorization algorithms for multimodal latent factor analysis. This approach can be understood as learning a common subspace for two data modalities, whose basis reveal underlying relationships between them. Discovering the structure of this subspace allows to generalize multimodal relationships to new images without text. A family of algorithms that consider different structures of the text data modality have been presented.

When the text is composed of clean and structured annotations, multiple classification functions can be used to predict labels on images, as was presented in Chapter 3. This has been a popular approach for semantic image indexing, but still faces some limitations and problems as was discussed in Chapter 2. To overcome some of these limitations, a direct data embedding from visual features to semantic terms was presented in Chapter 5: *The Nonnegative Semantic Embedding*. This model can be also understood as a supervised learning approach, in which predictions contain the full distribution of terms for an image, having potential connections with structured output prediction [137].

For image collections with noisy and unstructured text descriptions, latent factor models based on Nonnegative Matrix Factorization were presented in Chapter 4: The *NMF-Asymmetric* and *NMF-Mixed*. These models can be understood as unsupervised learning strategies to discover multimodal aspects from two modality inputs. These algorithms can also be cast as probabilistic models [105], opening interesting possibilities

to study new alternatives to unveiling the relationships between both data modalities.

The main goal in any of these cases is to find links between visual features and their potential semantic interpretations, which can be found explicitly or implicitly in text data. This work has shown that the goal can be achieved under different textual conditions, from clean to noisy, and from structured to unstructured. This result encourages to further study the more challenging case of dealing with natural language text, since this can be cheaply obtained from many sources, and does not require supervision to build training data sets. Approaching this situation may require more elaboration of methods to filter and represent the text data modality, and more robust models to avoid noise, ambiguities and imprecisions. This represents a very interesting and challenging direction of future research.

## 7.3 Multimodal Representations

Modeling image contents is a very important problem in computer vision research, and it is mainly oriented to structure visual characteristics in a particular representation. Building content representations is traditionally accomplished using a single data modality based on the computation of visual features from pixels. Sometimes multiple features are combined together to produce an enhanced representation that integrates different characteristics such as colors, textures and edges, as was presented in Chapter 3.

Even using multiple visual features, the resulting representation is still a low-level description of image contents and does not include explicit semantics. The methods and algorithms proposed in this thesis are focused on computing image representations using visual features that are organized according to the semantics of text terms. Chapter 4 presents the construction of multimodal factors as the basic features for representing images in a latent space, and Chapter 5 presents fusion of visual and semantic features by projecting representations from one space to another.

The resulting multimodal representations have shown to be improved descriptions of image contents, in the sense that they allow to discriminate images in a more semantic fashion, similarly to the criterion followed by humans to discard non-relevant information. These representations have been primarily used in this work to compute ranking scores in an image search engine. However, they could be also used for other visual analysis tasks, such as categorization, clustering or event detection.

The most appealing property of a multimodal representation, according to the methods presented in this work, is that they are in fact data-driven representations. So, by feeding the algorithms with visual features and text terms the representations are automatically adjusted with respect to the underlying multimodal relationships. This can be understood as extending the frameworks for image analysis using a complementary semantic axis, which may allow to improve the accuracy of high-level decisions in practical systems.

## 7.4 Large Scale Multimodal Learning

Processing and indexing large image collections may be very difficult if the number of images is very large. The main scalability problem is on the use of learning algorithms since their computational complexity is not usually linear and they may demand high memory and CPU resources. A typical procedure to obtain a feasible solution is sampling a small portion of the image collection to train models. For instance, Romberg et al. [118] extracted a sample of 10,000 images for learning multimodal relationships from the collection that they planed to index, which consisted of 10 million images.

Besides the evident waste of available data to learn from, it is not clear how to select a representative, non-biased sample to train generic models. A better solution may be to harness the whole training set. Distributed computing may be used to share the load among various machines to speedup the process. However, this may pose significant efforts to configure computing infrastructures and develop practical programs.

This work has presented a learning strategy following online learning principles in Chapter 6, which demonstrated to be several orders of magnitude faster to learn multimodal relationships than batch models, running on a single CPU with low memory usage. The very fast convergence rate is achieved by setting the optimization problem to solve for the expected cost function in a stochastic process, instead of solving for minimum empirical cost. This has been demonstrated to offer asymptotically faster convergence when the limitation of the learning algorithm is time rather than the available amount of data [122].

This view of constraining the algorithm in terms of time instead of data, is a very useful property for analyzing and indexing large image collections, which are constantly growing. So, the results presented in this work demonstrate that large scale multimodal learning can be efficiently achieved by appropriately formulating the problem, instead of assuming automatic scalability by means of parallelization or distributed computing. Actually, with the current convergence rate of the proposed algorithm, parallelized online learning is a potential research direction, since it would further speed up the process, theoretically in a proportional rate to the number of available machines [135].

# Bibliography


[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[2] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2008.79

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained Part-Based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2009.167

[5] X. Xia, W. Yang, H. Li, and S. Zhang, "Part-Based object detection using cascades of boosted classifiers," in *Computer Vision - ACCV 2009*, ser. Lecture Notes in Computer Science, H. Zha, R.-i. Taniguchi, and S. Maybank, Eds. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2010, vol. 5995, ch. 52, pp. 556–565. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12304-7_52

[6] C. W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, "QBIC project: querying images by content, using color, texture, and shape," C. W. Niblack, Ed., vol. 1908, no. 1. SPIE, 1993, pp. 173–187. [Online]. Available: http://dx.doi.org/10.1117/12.143648

[7] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233–254, 1996. [Online]. Available: http://dx.doi.org/10.1007/BF00123143

[8] Y. Rui, T. S. Huang, and S.-f. Chang, "Image retrieval: Past, present, and future," in *Journal of Visual Communication and Image Representation*, vol. 10, 1997, pp. 1–23.

[9] R. C. Veltkamp and M. Tanase, "Content-based image retrieval systems: A survey," 2000. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.7344

[10] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, pp. 1–60, April 2008.

[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," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000. [Online]. Available: http://dx.doi.org/10.1109/34.895972

[12] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proceedings of the International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[13] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Computer Vision - ECCV 2002*, ser. Lecture Notes in Computer Science, 2002, ch. 7, pp. 349–354.

[14] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. New York, NY, USA: ACM, 2003, pp. 119–126.

[15] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, ser. SIGIR '03. New York, NY, USA: ACM, 2003, pp. 127–134. [Online]. Available: http://dx.doi.org/10.1145/860435.860460

[16] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2. IEEE, 2004, pp. 1002–1009. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2004.1315274

[17] Y. Jin, L. Khan, L. Wang, and M. Awad, "Image annotations by combining multiple evidence & wordNet," in *Proceedings of the 13th annual ACM international conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: ACM, 2005, pp. 706–715. [Online]. Available: http://dx.doi.org/10.1145/1101149.1101305

[18] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on*

*Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, Mar. 2007. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2007.61

[19] J. Li and J. Z. Wang, "Real-Time computerized annotation of pictures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 985–1002, Jun. 2008. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2007.70847

[20] C. Wang, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2009, pp. 1903–1910. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2009.5206800

[21] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ser. CHI '04. New York, NY, USA: ACM, 2004, pp. 319–326. [Online]. Available: http://dx.doi.org/10.1145/985692.985733

[22] T. Yeh, K. Grauman, K. Tollmar, and T. Darrell, "A picture is worth a thousand keywords: image-based object search on a mobile platform," in *CHI '05 extended abstracts on Human factors in computing systems*, ser. CHI EA '05. New York, NY, USA: ACM, 2005, pp. 2025–2028. [Online]. Available: http://dx.doi.org/10.1145/1056808.1057083

[23] X. Fan, X. Xie, Z. Li, M. Li, and W. Ma, "Photo-to-search: using multimodal queries to search the web from mobile devices," in *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*. Hilton, Singapore: ACM, 2005, pp. 143–150.

[24] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *Multimedia, IEEE Transactions on*, vol. 9, no. 5, pp. 923–938, 2007.

[25] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, April 2010.

[26] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 2, no. 1, pp. 1–19, Feb. 2006. [Online]. Available: http://dx.doi.org/10.1145/1126004.1126005

[27] M. La Cascia, S. Sethi, and S. Sclaroff, "Combining textual and visual cues for content-based image retrieval on the world wide web," 1998, pp. 24–28. [Online]. Available: http://dx.doi.org/10.1109/IVL.1998.694480

[28] P. Duygulu, M. Baştan, and D. Forsyth, "Translating images to words for recognizing objects in large image and video collections," in *Toward Category-Level Object Recognition*, ser. Lecture Notes in Computer Science, J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, Eds. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2006, vol. 4170, ch. 14, pp. 258–276. [Online]. Available: http://dx.doi.org/10.1007/11957959_14

[29] R. Nuray and F. Can, "Automatic ranking of information retrieval systems using data fusion," *Inf. Process. Manage.*, vol. 42, no. 3, pp. 595–614, May 2006. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2005.03.023

[30] J. K. Cramer and W. Hersh, "Multimodal medical image retrieval: image categorization to improve search precision," in *Proceedings of the international conference on Multimedia information retrieval*, ser. MIR '10. New York, NY, USA: ACM, 2010, pp. 165–174. [Online]. Available: http://dx.doi.org/10.1145/1743384.1743415

[31] P. Chandrika and C. V. Jawahar, "Multi modal semantic indexing for image retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, ser. CIVR '10. New York, NY, USA: ACM, 2010, pp. 342–349. [Online]. Available: http://dx.doi.org/10.1145/1816041.1816091

[32] S. Ayache, G. Quénot, and J. Gensel, "Classifier fusion for SVM-based multimedia semantic indexing," in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science, G. Amati, C. Carpineto, and G. Romano, Eds. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2007, vol. 4425, ch. 44, pp. 494–504. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-71496-5_44

[33] B. Sigurbjörnsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proceedings of the 17th international conference on World Wide Web*, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 327–336. [Online]. Available: http://dx.doi.org/10.1145/1367497.1367542

[34] J. C. Caicedo, E. Romero, and F. A. González, "Content-Based Histopathology Image Retrieval Using a Kernel-based Semantic Annotation Framework," *Journal of Biomedical Informatics*, vol. 44, pp. 519–528, Feb. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2011.01.011

[35] P. Kragel and P. Kragel, "Digital microscopy: a survey to examine patterns of use and technology standards," in *Telehealth/AT '08: Proceedings of the IASTED International Conference on Telehealth/Assistive Technologies*. Anaheim, CA, USA: ACTA Press, 2008, pp. 195–197. [Online]. Available: http://portal.acm.org/citation.cfm?id=1722803

[36] T. Dennis, R. D. Start, and S. S. Cross, "The use of digital imaging, video conferencing, and telepathology in histopathology: a national survey." *Journal of clinical pathology*, vol. 58, no. 3, pp. 254–258, March 2005.

[37] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications–clinical benefits and future directions," *International Journal of Medical Informatics*, vol. 73, no. 1, pp. 1–23, February 2004.

[38] T. Deselaers, D. Keysers, and H. Ney, "FIRE - Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation," *Multilingual Information Access for Text, Speech and Images*, pp. 688–698, 2005.

[39] H. Müller, C. Lovis, and A. Geissbuhler, "The medGIFT project on medical image retrieval," in *Proceedings of First International Conference on Medical Imaging and Telemedicine, Wuyi Mountain, China*, 2005.

[40] H. L. Tang, R. Hanka, and H. H. S. Ip, "Histological image retrieval based on semantic content analysis," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 7, no. 1, pp. 26–36, 2003.

[41] M. Iregui, F. Gomez, and E. Romero, "Strategies for efficient virtual microscopy in pathological samples using JPEG2000," *Micron*, vol. 38, no. 7, pp. 700–713, October 2007.

[42] F. Yu and H. Ip, "Semantic content analysis and annotation of histological images," *Computers in Biology and Medicine*, vol. 38, no. 6, pp. 635–649, June 2008.

[43] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski, "Automated grading of prostate cancer using architectural and textural image features," *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*, pp. 1284–1287, Apr. 2007. [Online]. Available: http://dx.doi.org/10.1109/ISBI.2007.357094

[44] J. Kong, O. Sertel, H. Shimada, K. L. Boyer, J. H. Saltz, and M. N. Gurcan, "Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation," *Pattern Recognition*, vol. 42, no. 6, pp. 1080–1092, June 2009.

[45] M. Tambasco, B. M. Costello, A. Kouznetsov, A. Yau, and A. M. Magliocco, "Quantifying the architectural complexity of microscopic images of histology specimens," *Micron*, vol. 40, no. 4, pp. 486–494, June 2009.

[46] K. Mosaliganti, F. Janoos, O. Irfanoglu, R. Ridgway, R. Machiraju, K. Huang, J. Saltz, G. Leone, and M. Ostrowski, "Tensor classification of N-point correlation function features for histology tissue segmentation." *Medical image analysis*, vol. 13, no. 1, pp. 156–166, February 2009.

[47] L. 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*, vol. 7, no. 4, pp. 249–255, 2003.

[48] J. Naik, S. Doyle, A. Basavanally, S. Ganesan, M. D. Feldman, J. E. Tomaszewski, and A. Madabhushi, "A Boosted Distance Metric: Application to Content Based Image Retrieval and Classification of Digitized Histopathology," *SPIE Medical Imaging: Computer-Aided Diagnosis*, vol. 7260, pp. 72 603F1–12, 2009.

[49] J. C. Caicedo, F. A. Gonzalez, and E. Romero, "A Semantic Content-Based Retrieval Method for Histopathology Images," *Information Retrieval Technology*, vol. LNCS 4993, pp. 51–60, 2008.

[50] C. S. M. Wong, R. C. Strange, and J. T. Lear, "Basal Cell Carcinoma," *BMJ*, vol. 327, pp. 794–798, 2003.

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

[52] A. Bosch, X. Muñoz, and R. Martí, "Which is the best way to organize/classify images by content?" *Image and Vision Computing*, vol. 25, no. 6, pp. 778–791, June 2007.

[53] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, pp. 42–51, 1998.

[54] X. Qi and Y. Han, "Incorporating multiple SVMs for automatic image annotation," *Pattern Recognition*, vol. 40, no. 2, pp. 728–741, February 2007.

[55] M. O. Gueld, D. Keysers, T. Deselaers, M. Leisten, H. Schubert, H. Ney, and T. M. Lehmann, "Comparison of global features for categorization of medical images," *Medical Imaging*, vol. 5371, pp. 211–222, 2004.

[56] S. Siggelkow, "Feature Histograms for Content-Based Image Retrieval," Ph.D. dissertation, Albert-Ludwigs-UniversitÂšat Freiburg im Breisgau, 2002.

[57] A. P. Berman and L. G. Shapiro, "A Flexible Image Database System for Content-Based Retrieval," *Computer Vision and Image Understanding*, vol. 75, 1999.

[58] A. S. A. Mark S. Nikson, *Feature Extraction and Image Processing*, Newnes, Ed. Elsevier, 2002.

[59] T. Deselaers, "Features for Image Retrieval," Ph.D. dissertation, RWTH Aachen University. Aachen, Germany, 2003.

[60] J. C. Caicedo, A. Cruz, and F. Gonzalez, "Histopathology Image Classification Using Bag of Features and Kernel Functions," *Artificial Intelligence in Medicine Conference, AIME 2009*, vol. LNAI 5651, pp. 126–135, 2009.

[61] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[62] A. Barla, E. Franceschi, F. Odone, and A. Verri, "Image Kernels," *Pattern Recognition with Support Vector Machines*, vol. LNCS 2388, pp. 617–628, 2002.

[63] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, 2008, pp. 1– 8.

[64] B. Schölkopf and A. Smola, *Learning with kernels. Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press, 2002.

[65] N. Cristianini, J. Shawe-Taylor, A. Elissee, and J. Kandola, "On kernel-target alignment," in *Advances in Neural Information Processing Systems 14*, vol. 14, 2002, pp. 367–373.

[66] J. Kandola, J. Shawe-Taylor, and N. Cristianini, "Optimizing Kernel Alignment over Combinations of Kernel," Department of Computer Science,Royal Holloway, University of London, UK, Tech. Rep., 2002.

[67] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural computation, MIT Press*, vol. 10, pp. 1895–1923, 1998.

[68] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: overview and proposals," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 593–601, April 2001.

[69] J. Diamond, "The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia," *Human Pathology*, vol. 35, no. 9, pp. 1121–1131, September 2004.

[70] O. Sertel, J. Kong, U. Catalyurek, G. Lozanski, J. Saltz, and M. Gurcan, "Histopathological Image Analysis Using Model-Based Intermediate Representations and Color Texture: Follicular Lymphoma Grading," *Journal of Signal Processing Systems*, vol. 55, no. 1, pp. 169–183, April 2009.

[71] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley, and I. G. Goldberg, "WND-CHARM: Multi-purpose image classification using compound image transforms," *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1684–1693, August 2008.

[72] J. C. Caicedo, J. BenAbdallah, F. A. González, and O. Nasraoui, "Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization," *Neurocomputing*, vol. 76, pp. 50–60, Jan. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2011.04.037

[73] M. J. Huiskes and M. 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: ACM, 2008.

[74] F. A. González, J. C. Caicedo, O. Nasraoui, and J. Ben-Abdallah, "Nmf-based multimodal image indexing for querying by visual example," in *Proceedings of the ACM international Conference on Image and Video Retrieval (Xi'an, China, July 05 - 07, 2010) ACM CIVR '10.*, 2010, pp. 366–373.

[75] J. Ben-Abdallah, J. C. Caicedo, F. A. GonzÃ¡lez, and O. Nasraoui, "Multimodal image annotation using non-negative matrix factorization," in *Proceedings of the IEEE Web Intelligence Conference (Toronto, Canada, Aug. 31, 2010) IEEE-WIC 10*, 2010.

[76] J. Ah-Pine, M. Bressan, S. Clinchant, G. Csurka, Y. Hoppenot, and J.-M. Renders, "Crossing textual and visual content in different application scenarios," *Multimedia Tools and Applications*, vol. 42, no. 1, pp. 31–56, March 2009.

[77] R. Agrawal, W. Grosky, and F. Fotouhi, "Image retrieval using multimodal keywords," *Multimedia, International Symposium on*, vol. 0, pp. 817–822, 2006.

[78] H. Song, X. Li, and P. Wang, "Multimodal image retrieval based on annotation keywords and visual content," *Computer-Aided Software Engineering, International Workshop on*, vol. 0, pp. 295–298, 2009.

[79] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, Mar. 2003. [Online]. Available: http://portal.acm.org/citation.cfm?id=944965

[80] F. Monay and D. Gatica-Perez, "Modeling semantic aspects for cross-media image indexing," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 10, pp. 1802–1817, August 2007.

[81] J. S. Hare, S. Samangooei, P. H. Lewis, and M. S. Nixon, "Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces," in *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*. New York, NY, USA: ACM, 2008, pp. 359–368.

[82] J. Tang and P. H. Lewis, *Non-negative matrix factorisation for object class discovery and image auto-annotation*. New York, New York, USA: ACM Press, Jul. 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1386352.1386370

[83] S. K. Gupta, D. Phung, B. Adams, T. Tran, and S. Venkatesh, *Nonnegative shared subspace learning and its application to social media retrieval*. New York, New York, USA: ACM Press, Jul. 2010. [Online]. Available: http://portal.acm.org/citation.cfm?id=1835804.1835951

[84] M. Cooper and J. Foote, "Summarizing video using non-negative similarity matrix factorization," in *IEEE Workshop on Multimedia Signal Processing*, 2002.

[85] W. Liu, N. Zheng, and X. Lu, "Non-negative matrix factorization for visual coding," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003.

[86] D. Liang, J. Yang, and Y. Chang, "Supervised non-negative matrix factorization based latent semantic image indexing," *CHINESE OPTICS LETTERS*, vol. 4, pp. 272–274, 2006.

[87] A. F. Smeaton and I. Quigley, "Experiments on using semantic distances between words in image caption retrieval," in *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 1996, pp. 174–180.

[88] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 50–57. [Online]. Available: http://dx.doi.org/10.1145/312624.312649

[89] C. Liu, H. C. Yang, J. Fan, L. W. He, and Y. M. Wang, "Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 681–690. [Online]. Available: http://dx.doi.org/10.1145/1772690.1772760

[90] D. D. Lee and H. S. Seung, "Algorithms for Nonnegative Matrix Factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[91] V. Y. F. Tan and C. Fevotte, "Automatic relevance determination in nonnegative matrix factorization," in *SPARS'09 - Signal Processing with Adaptive Sparse Structured Representations*, 2009.

[92] J. Hare, P. Lewis, P. Enser, and C. Sandom, "A linear-algebraic technique with an application in semantic image retrieval," *Lecture Notes in Computer Science*, vol. 4071, p. 31, 2006.

[93] T. Tsikrika and J. Kludas, "Overview of the wikipediamm task at imageclef 2009," *Cross Language Evaluation Forum (CLEF) Working Notes*, 2009.

[94] H. M
"uller, I. Eggel, S. Bedrick, S. Radhouani, B. Bakke, C. Kahn Jr, and W. Hersh, "Overview of the clef 2009 medical image retrieval track," *Cross Language Evaluation Forum (CLEF) Working Notes*, 2009.

[95] Lee, Daniel D. and Seung, H. Sebastian, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, October 1999.

[96] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 316–329.

[97] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid, "Image annotation with tagprop on the mirflickr set," in *MIR '10: Proceedings of the international conference on Multimedia information retrieval*. New York, NY, USA: ACM, 2010, pp. 537–546.

[98] J. Hare and P. Lewis, "Automatically annotating the mir flickr dataset: Experimental protocols, openly available data and semantic spaces," in *MIR '10: Proceedings of the international conference on Multimedia information retrieval*. ACM, March 2010, pp. 547–556.

[99] J. Camargo, J. Caicedo, and F. González, "Multimodal image collection visualization using non-negative matrix factorization," *Research and Advanced Technology for Digital Libraries*, pp. 429–432, 2010.

[100] L. Xie and S. F. Chang, "Pattern mining in visual concept streams," *Multimedia and Expo, IEEE International Conference on*, vol. 0, pp. 297–300, 2006. [Online]. Available: http://dx.doi.org/10.1109/ICME.2006.262457

[101] J. Han, "Data mining for image/video processing: a promising research frontier," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*, ser. CIVR '08. New York, NY, USA: ACM, 2008, pp. 1–2. [Online]. Available: http://dx.doi.org/10.1145/1386352.1386353

[102] M. Inoue, "On the need for annotation-based image retrieval," in *Workshop on Information Retrieval in Context*, 2004.

[103] P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. M
"uller, "Overview of the imageclef 2006 photographic retrieval and object annotation tasks," *Evaluation of Multilingual and Multi-modal Information Retrieval*, pp. 579–594, 2007.

[104] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. New York, NY, USA: ACM, 2003, pp. 267–273.

[105] C. Ding, T. Li, and W. Peng, "On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing," *Computational Statistics & Data Analysis*, vol. 52, no. 8, pp. 3913–3927, Apr. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.csda.2008.01.011

[106] A. Cruz-Roa, J. C. Caicedo, and F. A. González, "Visual pattern mining in histology image collections using bag of features," *Artificial Intelligence in Medicine*, vol. 52, no. 2, pp. 91–106, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.artmed.2011.04.010

[107] N. Bonnet, "Some trends in microscope image processing," *Micron*, vol. 35, no. 8, pp. 635–653, Dec. 2004. [Online]. Available: http://dx.doi.org/10.1016/j.micron.2004.04.006

[108] T. Meng, L. Lin, M.-L. Shyu, and S.-C. Chen, "Histology image classification using supervised classification and multimodal fusion," in *2010 IEEE International Symposium on Multimedia*. IEEE, Dec. 2010, pp. 145–152. [Online]. Available: http://dx.doi.org/10.1109/ISM.2010.29

[109] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley, and I. G. Goldberg, "WND-CHARM: Multi-purpose image classification using compound image transforms," *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1684–1693, Aug. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2008.04.013

[110] H. Müller and J. Kalpathy-Cramer, "The ImageCLEF medical retrieval task at ICPR 2010," in *Proceedings of the 20th International Conference on Pattern Recognition*, 2010, pp. 3284–3287.

[111] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Computing Surveys*, vol. 44, no. 1, Jan. 2012. [Online]. Available: http://dx.doi.org/10.1145/2071389.2071390

[112] J. A. Vanegas, J. C. Caicedo, F. A. González, and E. Romero, "Histology image indexing using a non-negative semantic embedding," in *MCBR-CDS'11 Proceedings of the Second MICCAI international conference on Medical Content-Based Retrieval for Clinical Decision Support*, ser. LNCS, 2012, vol. 7075, ch. 8, pp. 80–91. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-28460-1_8

[113] A. Barla, F. Odone, and A. Verri, "Histogram Intersection Kernel for Image Cassification," *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 3, pp. 513–16, 2003.

[114] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, 2005. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2005.239

[115] H. Müller, T. Deselaers, T. Deserno, P. Clough, E. Kim, and W. Hersh, "Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks," in *Evaluation of Multilingual and Multi-modal Information Retrieval*. Springer, 2007, pp. 595–608.

[116] T. Lehmann, M. Guld, T. Deselaers, D. Keysers, H. Schubert, K. Spitzer, H. Ney, and B. Wein, "Automatic categorization of medical images for content-based retrieval and data mining," *Computerized Medical Imaging and Graphics*, vol. 29, no. 2-3, pp. 143–155, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.compmedimag.2004.09.010

[117] Z. Akata, C. Thurau, and C. Bauckhage, "Non-negative matrix factorization in multimodality data for segmentation and label prediction," *16th Computer Vision Winter Workshop*, 2011.

[118] S. Romberg, R. Lienhart, and E. Hörster, "Multimodal image retrieval," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 1, pp. 31–44, Apr. 2012. [Online]. Available: http://dx.doi.org/10.1007/s13735-012-0006-4

[119] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009. [Online]. Available: http://dx.doi.org/10.1109/MC.2009.263

[120] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0. Los Alamitos, CA, USA: IEEE, Jun. 2007, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2007.383172

[121] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of GIST descriptors for web-scale image search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, ser. CIVR '09. New York, NY, USA: ACM, 2009, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1145/1646396.1646421

[122] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Optimization for Machine Learning*, S. Sra, S. Nowozin, and S. J. Wright, Eds. MIT Press, 2011, pp. 351–368. [Online]. Available: http://leon.bottou.org/papers/bottou-bousquet-2011

[123] R. Lienhart, S. Romberg, and E. Hörster, "Multilayer pLSA for multimodal image retrieval," in *CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval*. New York, NY, USA: ACM, 2009, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1145/1646396.1646408

[124] D. Putthividhy, H. T. Attias, and S. S. Nagarajan, "Topic regression multi-modal latent dirichlet allocation for image annotation," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 3408–3415, 2010. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2010.5540000

[125] F. González, J. Caicedo, O. Nasraoui, and J. Ben-Abdallah, "Nmf-based multimodal image indexing for querying by visual example," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2010, pp. 366–373.

[126] R. Wan, V. N. Anh, and H. Mamitsuka, "Efficient probabilistic latent semantic analysis through parallelization information retrieval technology," ser. Lecture Notes in Computer Science, G. G. Lee, D. Song, C.-Y. Lin, A. Aizawa, K. Kuriyama, M. Yoshioka, and T. Sakai, Eds. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2009, vol. 5839, ch. 38, pp. 432–443. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04769-5_38

[127] R. Gemulla, E. Nijkamp, P. J. Haas, and Y. 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*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 69–77. [Online]. Available: http://dx.doi.org/10.1145/2020408.2020426

[128] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010. [Online]. Available: http://portal.acm.org/citation.cfm?id=1756008

[129] J. Caicedo and F. González, "Multimodal fusion for image retrieval using matrix factorization," in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*. ACM, 2012, p. 56.

[130] L. Bottou, "Large-Scale machine learning with stochastic gradient descent," in *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, Y. Lechevallier and G. Saporta, Eds. Paris, France: Springer, Aug. 2010. [Online]. Available: http://leon.bottou.org/papers/bottou-2010

[131] C.-j. Lin, "Projected gradient methods for non-negative matrix factorization," in *Neural Computation*, vol. 19, 2007, pp. 2756–2779. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.65.5655

[132] J. Kalpathy-Cramer, H. Müller, S. Bedricks, I. Eggel, A. G. Seco de Herrera, and T. Tsikrika, "Overview of the CLEF 2011 medical image classification and retrieval tasks," 2011.

[133] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, vol. 2. Los Alamitos, CA, USA: IEEE, Oct. 2006, pp. 2169–2178. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.68

[134] S. A. Chatzichristofis and Y. 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*, ser. ICVS'08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 312–322. [Online]. Available: http://portal.acm.org/citation.cfm?id=1788559

[135] M. Zinkevich, M. Weimer, A. Smola, and L. Li, "Parallelized stochastic gradient descent," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds., 2010.

[136] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1109/MIS.2009.36

[137] G. Bakir, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan, Eds., *Predicting Structured Data*. The MIT Press, 2007.