UNIVERSIDAD NACIONAL DE COLOMBIA

Multimodal Non-linear Latent Semantic Method for Information Retrieval

Víctor Hugo Contreras Ordoñez

Universidad Nacional de Colombia
Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2019

Multimodal Non-linear Latent Semantic Method for Information Retrieval

Víctor Hugo Contreras Ordoñez

Submitted to the Engineering School of the Universidad Nacional de Colombia, in partial fulfillment of the requirements for the degree of:
Master in Systems and Computer Engineering

Advisor:
Fabio A. González Ph.D.
Co-Advisor:
Jorge A. Vanegas Ph.D.

Research field:
Computer Science
Research Group:
MindLab Research Group

Universidad Nacional de Colombia
Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2018

Dedication

To my parents Luis and Bidalba; and my brother Alfonso.

Acknowledgement

This thesis would not have been possible without the contributions of many people and institutions that support this work.

First I would like to thank my thesis advisors, Dr. Fabio A. González, and Dr. Jorge A. Vanegas, for their invaluable advice, support, help, understanding, confidence and encourage to continue learning and working. Also, I would recognize people in MindLab research group and I would specially like to thank Juan Sebastian Lara for his great contributions to this work, Juan Sebastian Otalora, John Arevalo, Oscar Julián Perdomo, Sebastian Sierra, and Andrés Rosso.

I want to thank COLCIENCIAS for funding this research with student grant Jóvenes investigadores 761/2016 and Universidad Nacional de Colombia engineering faculty for his funding, professors advice, access to academic and research events, access to infrastructure, libraries and bibliography information, and for contribution to publishing this research.

Finally, but not less important, I want to thank my family, my father Luis Ernesto, my mother Bidalba and my brother Alfonso for their unconditional love, confidence, support, encourage and advice.

Abstract

Multimodal information retrieval is an information retrieval sub-task where queries and database target elements are composed of several modalities or views. A modality is a representation of a complex phenomena, captured and measured by different sensors or information sources, each one encodes some information about it. Each modality representation contains complementary and shared information about the phenomenon of interest, this additional information can be used to improve information retrieval process. Several methods have been developed to take advantage of additional information distributed across different modalities. Some of them exploit statistical properties in multimodal data to find correlations and implicit relationships, others learn heterogenous distance functions, and others learn linear and non-linear projections that transform data from original input space to a common latent semantic space where different modalities are comparable. In spite the attention dedicated to this issue, multimodal information retrieval is still an open problem. This thesis presents a multimodal information retrieval system designed to learn several mapping functions to transform multimodal data to a latent semantic space, where different modalities are combined and can be compared to build a multimodal ranking and perform multimodal information retrieval task. Additionally, a multimodal kernelized latent semantic embedding method is proposed to construct a supervised multimodal index, integrating multimodal data and label supervision. This method can perform mappings to three different spaces where several information retrieval task setups can be performed.

The proposed system and method were evaluated in a multimodal medical case-based retrieval task where data is composed of whole-slide images of prostate tissue samples, pathologist's text report and Gleason score as a supervised label. Multimodal data and labels were combined to produce a multimodal index. This index was used to retrieve multimodal information and achieves outstanding results compared with previous works on this topic.

Non-linear mappings provide more flexibility and representation capacity to the proposed model. However, constructing the non-linear mapping in a large dataset using kernel methods can be computationally costly. To reduce the cost and allow large scale applications, the budget technique was introduced, showing good performance between speed and effectiveness.

Keywords: Multimodal information retrieval, Multimodal latent semantic embedding, matrix factorization, kernel methods, Multimodal fusion, Multimodal information retrieval system.

Resumen

La búsqueda y recuperación de datos multimodales es una importante tarea dentro del campo de búsqueda y recuperación de información, donde las consultas y los elementos de la base de datos objetivo están representados por un conjunto de modalidades, donde cada una de ellas captura un aspecto de un fenómeno de interés. Cada modalidad contiene información complementaria y común a otras modalidades. Con el fin de tomar ventaja de la información adicional distribuida a través de las distintas modalidades han sido desarrollados muchos algoritmos y métodos que utilizan las propiedades estadísticas en los datos multimodales para encontrar correlaciones implícitas, otros aprenden a calcular distancias heterogéneas, otros métodos aprenden a proyectar los datos desde el espacio de entrada hasta un espacio semántico común, donde las diferentes modalidades son comparables y se puede construir un ranking a partir de ellas.

En esta tesis se presenta el diseño de un sistema para la búsqueda y recuperación de información multimodal que aprende varias proyecciones no lineales a espacios semánticos latentes donde las distintas modalidades son representadas en conjunto y es posible realizar comparaciones y medidas de similitud para construir rankings multimodales. Adicionalmente se propone un método kernelizado para la proyección de datos a un espacio semántico latente usando la información de las etiquetas como método de supervisión para construir in índice multimodal que integra los datos multimodales y la información de las etiquetas; este método puede proyectar los datos a tres diferentes espacios semánticos donde varias configuraciones de búsqueda y recuperación de información pueden ser aplicadas.

El sistema y el método propuestos fueron evaluados en un conjunto de datos compuesto por casos médicos, donde cada caso consta de una imagen de tejido prostático, un reporte de texto del patólogo y un valor de Gleason score como etiqueta de supervisión. Combinando la información multimodal y la información en las etiquetas se generó un índice multimodal que se utilizó para realizar la tarea de búsqueda y recuperación de información por contenido obteniendo resultados sobresalientes.

Las proyecciones no-lineales permiten al modelo una mayor flexibilidad y capacidad de representación. Sin embargo calcular estas proyecciones no-lineales en un conjunto de datos enorme es computacionalmente costoso, para reducir este costo y habilitar el modelo para procesar datos a gran escala, la técnica del budget fue utilizada, mostrando un buen compromiso entre efectividad y velocidad.

Keywords: Multimodal information retrieval, Multimodal latent semantic embedding, matrix factorization, kernel methods, Multimodal fusion, Multimodal information retrieval system.

Content

List of Figures

List of Tables

1 Introduction

Information retrieval has attracted the attention of researchers for many years. Since the early 1950s to the present, a considerable amount of effort has been put to solve this issue effectively and efficiently. The primary goal of an information retrieval system is to satisfy a user's information need, giving an ordered collection of relevant elements.

Nowadays, information retrieval systems have gained particular relevance due to the fast growth of data collections in size and variety. The rapid developments in the internet and communications technologies have turned users from mere content consumers to producers, the content also has changed from text/hypertext in the beginnings to multimedia content nowadays. This pattern is also observed in a wide variety of domains from enterprise applications to medical information systems.

The pervasive presence of multimedia content and its quick growth makes necessary the development of information retrieval systems specialized in indexing and retrieving multimedia content. This kind of content poses a challenge known as "semantic gap", this is, there is not a direct relationship between its representation and its meaning. Several approaches have been proposed to overcome this challenge; some are based on feature engineering, others learn better content representations using latent semantic modeling, others use multi-view information that is common in multimedia content to learn a common semantic representation.

Usually, multimedia content contains multimodal information where the same phenomena is represented across multiple modalities, for example, images come with a text description like tags, captions or paragraphs; audios come with the author, gender, year and lyrics information; videos come with subtitles, transcriptions, image thumbnail, among others. Complementary information in each modality can be used to construct a multimodal semantic index and thus improve the retrieval system's performance including additional information about the phenomena within the retrieval and modeling process.

The main goal of this thesis is to research and to develop a multimodal information retrieval system based on a non-linear latent semantic strategy to construct a multimodal index taking advantage of additional information provided by multimodal data to improve retrieval performance.

1.1 Motivation

Currently, multimedia data is pervasive; this kind of data is everywhere from web pages that combine text, images, video, and audio to specialized multimedia platforms like video streaming services, podcast applications. These technologies combined with the ”web 2.0” paradigm where the user’s role changed from mere content consumer to a more active role of consumer/producer (prosumer), have contributed to the fast increase of multimedia data on the internet. As the amount of data is incremented more difficult is to find relevant information, to solve this issue it is necessary to develop high-quality information retrieval systems specialized in processing multimedia content.

Information retrieval systems have played a central role on the internet since the introduction of world wide web standard in 1989; in its beginnings, the internet was mostly composed of text/hypertext documents for that reason first web searchers were designed as text information retrieval systems with great success. However today the web is composed of multimedia data with heterogeneous sources. The first step to extract knowledge from this vast and valuable data resource includes identify and access relevant documents within large data collections; this task must be performed for future web searches that must be based on multimedia information retrieval systems.

Multimedia and multimodal information retrieval systems have application in different areas from research to industry. The following list briefly describes some application fields where multimedia information retrieval systems play an important role:

- *Medical and health care:* Medical domains generates a large volume of multimedia and multimodal information every day, for example, Müller et al. [61] described as in the radiology department of the General University Hospital of Geneva in the year 2002, 12000 images were generated per day, during the same year in the department of cardiology was generated 1 terabyte of cardiac images. Additional to medical images huge amount of textual data is generated, this text data includes clinical reports, medical histories, patient records.
  Large collections of multimedia and multimodal data in the medical domain contain valuable information that can contribute to improving health care services, accelerate the diagnosis process, train medical staff and support medical research. To enable these applications it is necessary to construct a robust information retrieval system that combines multimodal information to determine relevant medical cases to analyze in each situation.

- *Crime prevention:* Technology has a central role in crime prevention and surveillance, for example, security cameras are everywhere and produce many video hours where only some short events must be audited to provide security, prevent crimes and identify suspects. Extract this information from many video hours is a difficult task and must be executed by an automated system, this system is a multimedia information retrieval system specialized in video processing.

- *Enterprise applications:* Enterprises must compete locally and globally. Due to this pressure enterprises must improve their decision-making process. To take better decisions it is necessary to have the best information, and getting this information from different sources is the purpose of a multimodal information retrieval system.
- *E-commerce:* Online stores are becoming more popular, customers demand from these sites better multimodal retrieval systems that help them to find the right product using text descriptions, names or photos of similar products.
- *Copyright protection and plagiarism prevention:* Information retrieval systems provide a ranked list of similar elements, using these results, automated systems can determine automatically the amount of similarity between multimedia artistic work, including paintings, draws, videos, music, and documents. Similarity values can be used to establish an originality judgment, and protect the author's work from plagiarism.

Applications listed above are just some of the many where multimodal and multimedia retrieval systems can be applied.

1.2 Problem statement

Kernel methods have been demonstrated high-quality results and ability to learn complex relationships hidden in data, with a strong mathematical foundation they draw the attention of the research community and practitioners, and they were widely adopted and combined with several machine learning approaches and applied to many challenge issues with much success. Despite its advantages, kernel methods have several disadvantages related to its ability to scale efficiently for huge datasets.

Information retrieval systems had had a huge impact on the modern world as the amount of information grows at an exponential scale, find relevant information becomes more challenge and essential task. Successful information retrieval systems required precision and scalability characteristics to be able to fulfil user requirements. Many information retrieval and recommender systems use a low dimensional representation to construct indexes and perform retrieval or recommendation, to find these representations many algorithms are applied, Non-Negative Matrix Factorization is one of these methods with good performance, but scalability problems due to its need to keep in memory the full term-document matrix.

Nowadays data is not limited to a unique form and format, now data is represented in several formats, for instance, an article contains information in form of text, images, diagrams, and tables, we need to process these representations as a whole to get the valuable information contained within data. Many information retrieval systems focused on only one data representation or modality independently lost crucial information of complementary representations, this limitation constitutes an improvement opportunity for these systems. For these reasons, in this thesis, a solution to this problematic situation has proposed.

The main goal of this thesis is to research and develop effective and efficient methods and strategies to construct multimodal information retrieval systems. These systems perform the information retrieval task over multimodal data collections that are common in many application fields like web search engines, biomedical applications, e-commerce, multimedia indexing, among others.

Currently, multimedia data is pervasive and contains valuable information that can improve decision and business process. However, extracting information from multimedia data is a difficult task due to the "semantic gap", this is, the disparity between its computational representation and its semantic content. To narrow the semantic gap many approaches have been proposed, some of them use machine learning algorithms to learn the complex relationships between input features and semantic content. Despite significant advances in contributions to reducing the semantic gap, this is still an open problem where more research is necessary.

The present research tries to develop a multimodal information retrieval system and indexing algorithm that takes advantage of multi-view representation common in multimedia data to learn non-linear projection that model complex relationships between input features and latent semantic content. This important issue poses the following research questions:

- **How to model complex relationships between features and semantic content in multimodal data?**
- **How to combine multimodal information to construct a joint representation?**
- **How to perform a non-linear latent semantic embedding on multimodal data?**
- **How to use tag and class labels to improve semantic representations and narrow the semantic gap?**
- **How to design a large-scale multimodal information retrieval system?**
- **How to use kernel methods to induce non-linear mappings?**
- **How the kernel selection affects the algorithm performance?**

- **How the budget size impacts on retrieval effectiveness?**

1.3 Objectives

1.3.1 General goal

To develop and to evaluate a strategy to perform multimodal retrieval using non-linear latent semantic embeddings methods.

1.3.2 Specific objectives

- To design a multimodal latent semantic embedding strategy to construct a multimodal index based on kernel methods to allow non-linear modeling.
- To develop a prototype information retrieval system to perform multimodal information retrieval.
- To evaluate the performance of the proposed retrieval system in a multimodal retrieval task.

1.4 Results and Contributions

The main contributions of this thesis can be summarized as follows:

- **Contreras, V. H.**, Lara, J. S., Perdomo, O. J., Gonzalez, F.A., *Supervised Online Matrix Factorization for Histopathological Multimodal Retrieval.* Published manuscript.
  This work proposes a multimodal information retrieval system that uses MKSE-CM algorithm as an indexing strategy. The proposed method was tested over a multimodal medical dataset.

- **Contreras, V. H.**, Lara, J. S., Gonzalez, F. A., *Multimodal Kernel Semantic Regression for Medical Case Retrieval.* Manuscript in preparation.
  This paper presents an extension of MKSE-CM method adapted to perform a multimodal regression over Gleason scores on a medical dataset. The proposed indexing algorithm takes advantage of domain knowledge encoded within Gleason scores and allows to perform retrieval using three different projection spaces.

Vanegas, J. A., **Contreras, V. H.**, Escalante, H. J., Gonzalez, F. A. *Supervised On-line Kernel Semantic Embedding for Cross-Modal Retrieval.* Manuscript submitted.

This paper describes the Multimodal Kernel Semantic Embedding for Cross-Modal retrieval (MKSE-CM) method, that takes as input feature representations of two modalities and learns a non-linear latent semantic embedding to a common hidden space where data samples from different modalities can be compared. This model is a supervised method and takes advantage of label information to improve and align the latent semantic representation. Additionally, this algorithm can predict labels on test data and use these predictions as an indexing space. This work was jointly developed with Jorge Vanegas PhD and Fabio González PhD. My main contribution was in proposing and implementing semantic alignment on the method, performing experiments and helping with the paper writing.

1.5 Outline

The remain chapters of this thesis are organized as follows:

- *Chapter 2* Background and definitions: Introduces general definitions and theoretical concepts related to kernel methods, Latent semantic embedding, and large scale strategies.

- *Chapter 3* Related Work: presents a review of the related work on the topic of non-linear latent semantic embedding applied to multimodal information retrieval.

- *Chapter 4* Multimodal information retrieval System for medical case retrieval: In this chapter we proposed a multimodal information retrieval system and applied then to medical case retrieval problem.

- *Chapter 5* Multimodal Kernelized Latent Semantic Regression: This chapter introduces a new indexing algorithm based on non-linear latent semantic embedding.

- *Chapter 6* Conclusions and Future work: This chapter closes the discussion and presents the conclusions obtained from the accomplished work. Future work subsection suggests additional research opportunities on this field.

2 Background and Definitions

This chapter describes notation, definitions, and concepts used through this thesis document. The first part of this chapter defines the information retrieval task and its main variants, followed by kernel methods definition and latent semantic embeddings process description, and finally, the similarity and the performance measures employed in information retrieval are explained.

2.1 Information retrieval

Information retrieval task (IR) can be defined as the procedure of finding and access relevant material within a vast data collection in response to a user's information need expressed through a query [56]. Usually, these data collections are unstructured.

According to the query and the collection modality setups, the information retrieval task can be classified into two big groups: content-based retrieval (CBR), where only one modality is involved; and multimodal retrieval, where two or more modalities interact. This distinction is important due to depending on the modality representation and the interaction between the queries and the target collection the information retrieval process requires a different approach. This classification and some of its related task are shown in Figure **2-1**.

2.1.1 Content-based retrieval

Content-based retrieval is an information retrieval sub-task where queries and target data collection are expressed in the same modality. This setup is also known as query by example (QBE) because the query is an example of expected results [32].
In this retrieval setup as the queries and the target collection share a modality representation, a direct similarity measurement can be performed after a feature extraction process.
Content-based image retrieval, document retrieval, music retrieval, 3D model retrieval are some examples of this task.

Content-Based image retrieval

Content-based image retrieval is a prototypical example of QBE systems; in this setup, an image is provided as a query example, and the system's response is an ordered list of images

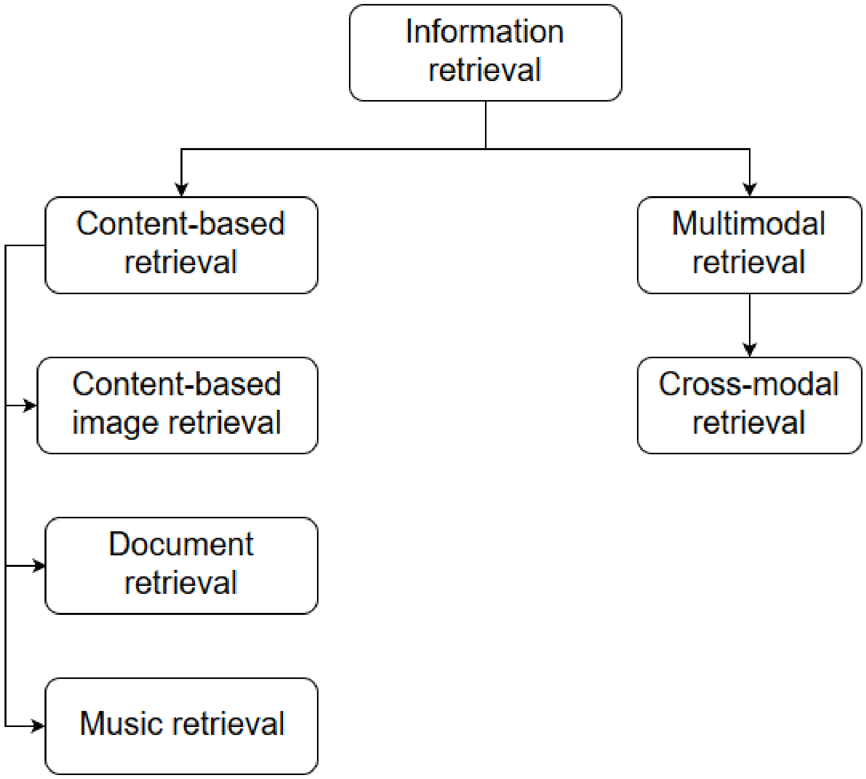

Figure 2-1: Information retrieval sub-task classification according to modal interaction. Left side shown content-based retrieval sub-task where only one modality is involved; on right side multimodal information retrieval sub-task where two or more modalities interact.

with similar content.

Document retrieval

In document retrieval, the query and the target database are text documents. Queries can be a keyword list, short phrases or a paragraph.

Music retrieval systems

In music information retrieval systems, queries are digital audio files or symbolic music notation, and target collection is composed of audio files representation.

2.2 Multimodal information retrieval

Multimodal data arise from the measure and capture information about the same entity of interest using different sensors, each sensor produces a representation, modality or view of the entity; and interaction of two or more modalities constitute a multimodal representation [44].

Currently, multimodal data is a common data representation, for example, internet articles like Wikipedia are composed of illustrative images and textual descriptions. These two modalities are different expressions of the same abstract entity, the article's topic.

2.2.1 Multimodal information retrieval

In multimodal information retrieval (MIR) task both query and target collection are constituted by multimodal data. MIR systems use data fusion techniques to construct joint models and multimodal indexes and thus perform the retrieval task.

2.2.2 Cross-modal information retrieval

Cross-modal information retrieval is a MIR sub-task with a special relevance due to its applications as a proxy task for more complex ones, like scene understanding, multimodal question answering, automatic image description, automatic video captioning, among others. In the cross-modal information retrieval setup, queries are expressed in one modality and target elements are expressed in another modality, for example, queries can be images and target elements can be suitable text descriptions that match query content.

2.3 Kernel methods

Kernel methods allow to work in a feature space by implicitly defining a non-linear mapping Φ from input space $\mathbb{R}^n$ to an inner product space F know as *feature space*. The feature space is usually a high-dimensional space; even feature space can be an infinite-dimensional space. Once data have been embedded into feature space, linear algorithms can be applied to them to discover patterns and relationships [72].

$$\Phi : \mathbb{R}^n \to F \tag{2-1}$$

The best-known kernel method is the support vector machine (SVM). The SVM is a linear method designed to find the optimal hyperplane that separates two classes with the maximum margin. However, efficiency and effectiveness of SVM it is limited to be applied in linearly separable data collections, to overcome this limitation SVM was combined with kernels methods increasing its generalization capacity; non-linear data can be separated mapping them to an appropriate feature space where they are linearly separable [69].

kernel trick

The main advantage of kernels methods lies in the *"kernel trick"*. The kernel trick performs an implicit projection to feature space using a *kernel function*. Compute the kernel function is equivalent to calculate the inner product in feature space without executing the projection process explicitly [69].

$$k(x, y) =< \Phi(x), \Phi(y) > \tag{2-2}$$

2.4 Latent semantic embedding methods

The principal purpose of latent semantic embedding algorithms is to project data from a high-dimensional space to a low-rank space where similarities between data samples are preserved. These algorithms usually use statistical and mathematical methods to model the underlying topics which constitute the bases of semantic space thus when an item is projected from input space to a latent semantic space its content is represented as a linear combination of latent topics.

Latent semantic embeddings are widely used in information retrieval task due to its advantages: from latent representation construct an index to speed up retrieval process is easier, perform similarity measures in latent space is faster than in input space and requires less computational resources [18].

Many algorithms produce latent semantic embeddings, some of the best-known methods are latent semantic analysis (LSA) [18, 45, 41] algorithm that decomposes the term-document matrix using singular value decomposition (SVD) to find the latent representation [17], probabilistic latent semantic analysis (PLSA) [36, 37] models the hidden variables as a probabilistic mixture and non-negative matrix factorization (NMF) recognized for its ease interpretability [47, 48, 38, 87].

2.4.1 Non-Negative Matrix Factorization

Non-Negative Matrix factorization (NMF) algorithm is an unsupervised method that perform matrix decomposition to learn a latent semantic embedding for data samples. For a given input matrix $X \in \mathbb{R}^{n \times d}$ with n features and d samples, the matrix factorization problem can be posed as find two matrices $W \in \mathbb{R}^{n \times l}$ and $H \in \mathbb{R}^{l \times d}$ whose product approximate input matrix $X \approx WH$; values for W and H can be found solving the following optimization problem [35, 95]:

$$\begin{aligned} \underset{W,H}{\text{minimize}} \quad & \|X - WH\|_F \\ \text{subject to} \quad & W \geq 0 \; H \geq 0 \end{aligned} \tag{2-3}$$

where $W \in \mathbb{R}^{n \times l}$ is known as the basics matrix, $H \in \mathbb{R}^{l \times d}$ is known as the encoding matrix, non-negative constraints ensure interpretability and cluster properties, l is a hyper-parameter that determines the number of latent topics in latent semantic space. Columns in H matrix contains the latent representation of documents in data collection, using this representation an index can be constructed using latent topics as indexers.

2.5 Similarity and distances measurements

Information retrieval systems and methods presented in this thesis learn a non-linear map from input space to latent semantic spaces, to construct a rank result list, the system must calculate a similarity value between query and target elements and using this value ordered the result list. In this section, we describe the principal distance and similarity measures employed to calculate the similarity score in information retrieval systems.

In this thesis, we employ kernel methods to learn a latent semantic embedding, these embeddings are vector spaces with inner product operation, vector spaces with this features are known as Hilbert spaces, in these spaces similarity can be calculated using similarity functions, such as *cosine similarity*, or also can be calculated from distances functions since those are inversely related to similarity scores. While similarity score normally varies between 0 and 1, being 0 not at all similarity and 1 total similarity; the distances functions vary in ranges from 0 to infinite, in this case, 0 means not distances or perfect similarity, whereas bigger values represent less similarity.

The Minkowski distances is one of the most common distance functions use to calculate the distance between two vectors $X = [x_1, \ldots, x_n] \in \mathbb{R}^n$ and $Y = [y_1, \ldots, y_n] \in \mathbb{R}^n$ within a vector space. The generalized vector distance function or Minkowski distance in $\mathbb{R}^n$ vector space can be defined as follows [42, 33, 29]:

$$D(X, Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} \tag{2-4}$$

where $p \in \mathbb{R}$ *and* $p \geq 1$.
The most common vector distances employed in applications are Euclidean distance where $p = 2$ and Manhattan distance where $p = 1$.

Cosine similarity is another metric frequently used to measure the similarity between vectors on Hilbert space. This similarity function measures the cosine of the angle between two vectors in inner product space. Its range varies from -1 to 1, as near to 1 greater the similarity. Cosine similarity can be described as follows [73]:

$$similarity = cos(\theta) = \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\|\|\mathbf{Y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2-5}$$

From a statistical point of view, we found centered correlation (CC) or Pearson's Correlation Coefficient (PCC) similarity function. This function computes the linear correlation between two random variables **X** and **Y**. Its range varies between -1 and 1, being values near to -1 an inverse linear relationship, values near to 1 represents a direct linear relationship and values near to 0 means, not linear relationship found. PCC can be defined as follows [5]:

$$CC = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{2-6}$$

where $\overline{x}$ is the mean value of X, and $\overline{y}$ is the mean value of Y, $cov(X, Y)$ is the covariance of variables X, Y, σ_X and σ_Y are the standard deviations of variables X and Y respectively. As is shown in equation 2-6 CC also can be calculated as the cosine similarity of centered variables.

2.6 Performance measures

The fundamental purpose of an information retrieval system is to satisfy the user's information need, but How to measure the fulfillment of this goal? To answer this question, we need to measure the system effectiveness using a set of performance metrics widely accepted and employed in this field that allows an objective comparison between IR systems. In the following, we describe the principal performance measures applied to evaluate information retrieval systems.

2.6.1 Precision

Precision (P) computes the fraction of relevant items retrieved over the total number of retrieved elements. This metric does not take into account the order in which results are presented [57, 80].

$$Precision = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ items\ retrieved} = P(relevant|retrieved) \tag{2-7}$$

2.6.2 Recall

Recall (R) computes the fraction of relevant items retrieved over the total number of relevant items. This metric does not take into account the order in which results are presented [57, 80].

$$Recall = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ relevant\ items} = P(retrieved|relevant) \tag{2-8}$$

2.6.3 Precision@k

Human attention is a scarce resource for that reason it is not common for a user to review every item in the result list. Good information retrieval systems present relevant items first in the result list. Precision at k (P@k) computes the precision value considering only the k items at the top of the result list, where k is a fixed value cut off point. Some common values for k are 5, 10, 30 and thus we have P@5, P@10, P@30 values [57].

2.6.4 bpref

The binary preference-based measure (bpref) computes a preferred relation whenever relevant items are retrieved before non-relevant within a ranked results list [80].

$$bpref = \frac{1}{R}\sum_{r}(1 - \frac{|n\ ranked\ higher\ than\ r|}{min(R, N)}) \tag{2-9}$$

Where R is the total number of relevant items for a given query, N is the total number of non-relevant items, r is the number of relevant retrieved items and n is the number of non-relevant items retrieved.

2.6.5 Average precision

Average precision (AP) computes the average of precision values for each recall level in ranked result list. This measure combines precision and recall for a given query and takes into account order in result list [90].

$$AP = \frac{\sum_r P@r}{R} \tag{2-10}$$

Where R is the total number of relevant items, r is the position of last relevant item retrieved, $P@r$ computes precision at $r - th$ item.

2.6.6 Mean Average Precision

Mean Average Precision (MAP) is the most extended and widely used performance metric for evaluating information retrieval systems. This measure computes the arithmetic mean of AP values for a set of queries [57, 80]. MAP takes into account the order of ranked results list and summaries system's response to a set of queries into only one value between 0 and 1 where the best results are near to 1. MAP is frequently used to compare the performance of different retrieval systems given the same queries.

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} AP_j \tag{2-11}$$

Where $|Q|$ is the cardinal of queries set.

2.6.7 Geometric Mean Average Precision

Geometric Mean Average precision (GMAP) is the geometric mean of AP values. This measure is commonly used to highlight improvements in low-performance topics [80, 4].

$$GMAP = \sqrt[n]{\prod_{n} AP_n} \tag{2-12}$$

3 Related Work

This chapter presents a summary of related works on multimodal information retrieval systems, multimodal data fusion approaches, kernel methods, and large-scale solutions for multimodal latent semantic embeddings.

3.1 Multimodal data fusion

Multimodal information retrieval is an important information retrieval task where queries and target database elements are represented using multimodal data. Multimodal data is composed of several media representations that capture an aspect of a complex phenomenon.

Each modality has its own data representation, scale, and feature extraction process, for this reason, combining the information distributed in the different modalities to improve information retrieval system's performance is a challenge. To solve this challenge three main approaches have been proposed: early fusion, late fusion, and intermediate fusion. Lahat et al. performed an extensive review on multimodal information retrieval, this review also contains main definitions describe above and use cases applications on health care, medical diagnosis, audio-visual multimodal retrieval, meteorological monitoring, among others [44].

3.1.1 Early fusion

In the early fusion approach, feature representations are combined to produce a joint representation, for this reason, this approach is also known as *feature level fusion* [59]. This approach has been used in many applications like video analysis and retrieval, audio and text retrieval, multimodal question answering, image-text retrieval, etc [40].

In the multimodal video retrieval field, we found Snoek et al. work where they tested early and late fusion approaches in the video retrieval task, early fusion approach combines visual, audio and text modalities concatenating them into a multimodal representation, then the multimodal representation is introduced into a supervised learning algorithm to perform retrieval. They concluded that late fusion had slightly better results in several semantic categories, however, for some categories, early fusion has a significant advantage over late fusion [74].

Depeursinge and Müller in ImageCLEF contest summarize methods employed to solve the proposed multimodal retrieval challenge they mention *curse of the dimensionality* as one of the main weakness of early fusion approaches, that weakness can be mitigated using feature weights combined [20].

3.1.2 Late fusion

In the late fusion approach, the outputs, decisions or similarity rankings of different retrieval systems are combined to produce a final response[58]. In this approach, each modality is processed by independent subsystems that produce their own decisions; then these decisions are combined using various strategies, for example, a convex combination of rankings, voting systems, weighted sums, among others for this reason this approach is also known as *decision level fusion* [59].

Late fusion approach has been applied to multimodal video retrieval in this experiment Snoek et al. use late fusion to combine decisions for image, audio and text, they found that late fusion approach slightly improves results for several semantic categories but performance it is limited for some categories where early fusion outperform this approach [74].

Guenes and Picardi studied the effect of late and early fusion on the performance of affect recognition system, combining face and body features. They concluded that late fusion by averaging decisions worked better than early or unimodal features, also they reported an especial advantage of late fusion allowing dynamically integrate face and body information showing promising and robustness results [31].

Ye et al. introduce a robust late fusion with rank minimization method where multiple model predictions are combined into a matrix and a late fusion is executed throughout matrix decomposition of joint matrix decision, this approach has shown a robust behaviour and is a viable alternative to averaging, voting and other traditional late fusion approaches [89].

Zheng et al. introduced a query-adaptative late fusion method for multimodal image retrieval that perform a late fusion at score-level showing feasibility of score-level fusion and outperformed other fusion schemes [94]. Bruno and Marchand-Maillet introduce a late fusion approach to combine clusters modelling their latent relationships and produce a multiview probabilistic cluster that outperforms early fusion approach [11]. Liu et al. proposed a multi-view clustering approach based on late fusion incomplete clustering matrices are imputed as the multiview clustering is constructed [54]. Sample specific late fusion method was introduced by Liu et al. this method learnt a late fusions weights for each sample obtaining higher weights for similar multimodal samples and a low one for negative or different semantic class, similar to the proposed method in this thesis [51].

Late fusion approaches had shown more flexibility than early fusion methods and in several tasks outperforms early approaches.

3.1.3 Intermediate fusion

Additional to late and early fusion approaches there are additional approaches to fusion multimodal heterogeneous data, this alternative approaches can be denominated as *intermediate fusion approach.* Intermediate fusion approaches are different from late and early fusion approaches in that the fusion of modalities is performed not at the feature or the decision level, this fusion is performed in an intermediate-level, for example, the method proposed in this thesis execute fusion at latent semantic embedding, combining latent representations of each modality.

In several cases, intermediate fusion approaches combine multimodal information inside a model, at the training phase. For example, we found some techniques like model co-training[91] where two models are trained simultaneously and they share weights and combine results from model layers to complement learning with learned information in other models. In the field of Multimodal representation learning, we found the Gated Multimodal Unit (GMU) developed by Arevalo et al. and applied successfully to movie gender classification using poster images and users reviews, this unit can be easily integrated into deep learning models [2].

3.2 Multimodal information retrieval methods

This section summaries several approaches employed to execute multimodal information retrieval task.

3.2.1 CCA and KCCA

First approaches to performing multimodal information retrieval task were based on statistical methods like Canonical Correlation Analysis (CCA) that find the directions where the correlation between two datasets is maximized, some works have used the matrices learned by CCA model as projection matrices from original input space to a joint latent semantic space where similarity measure can be computed and then used to produce a multimodal ranking[67].

CCA model was extended to perform non-linear mappings using kernel methods, in this form we find KCCA (Kernel CCA) that has been employed to perform a non-linear mapping from input space to a feature space where these two modalities are compatible and a similarity function can be applied on them [65].

3.2.2 Multimodal NMF

Non-negative matrix factorization NMF is a matrix decomposition method usually employed in blind source separation, spectral clustering, information retrieval, and recommender systems, among others. This model was extended to perform multimodal fusion. First approaches to multimodal-NMF were found in González et al. works [28, 12, 82] where a matrix factorization is performed first in one modality and then preserving previously learned matrix H the new factorization process is realized over other modalities, this approach was used to index and retrieval multimedia data and for medical information retrieval.

3.2.3 Deep learning approaches

Deep learning has attracted the attention of many researchers and currently is one of the hot topics in machine learning and applications, for this reason, many deep learning approaches have been proposed for multimodal information retrieval task.

Rastegar et al. present a multimodal deep learning framework with cross weights (MDL-CW), in this work the authors concluded that this method where weights among modalities are shared and a multimodal representation is reached throughout concatenate layer outperforms traditional multimodal retrieval methods such as KCCA [68]. Shao et al. proposed a deep CCA model combined with hypergraph semantic embedding (HSE)producing a DCCA-PHS model with competitive performance, in addition to this model the author introduced a search-based similarity measure inspired by PageRank algorithm independent from modal origin [71].

Srivastava et al. present a Multimodal Deep Belief Neural Network (DBN) that can learn a joint probability distribution from multimodal data, with this learned distributions the model can generate missing modalities and samples from each modality, their results show that this model can outperform traditional models such as Support Vector Machines and Latent Dirichlet Allocation in the classification task. In a posterior work, authors explored a similar approach but with Deep Boltzmann Machines obtaining similar results [75, 76].

Cross-modal retrieval is a related multimodal retrieval task. He et al. proposed a model for bidirectional representation learning using convolutions for image and text, through the use of cosine similarity between image and text samples learning similarities between positive samples and differences from negative samples, this algorithm shows better performance than traditional cross-modal retrieval methods like CCA and KCCA [34]. Similarly, Feng et al. developed a correspondence autoencoder (Corr-AE) where hidden representations of two unimodal autoencoders are correlated, combining representation learning and correlation learning in a single process, they also extend the basic model to models (Corr-Cross-AE) and (Corr-Full-AE) where cross reconstruction was incorporated, in other related work this author also proposed the correspondence restricted Boltzmann Machine model (Corr-RBM) with similar performance and approach, in posterior work Peng et al. introduced cross-media multiple deep network (CMDN) which learn a combined representation from hierarchical learning where intra-modality and inter-modality are learned in the first stages and then are combined to produce a multimodal score. [23, 22, 64].

Another popular approach for cross-modal retrieval is based in hash codes, this is a binary codification of documents for each modality the retrieval process is executed measuring the similarity of hashing codes from query and target documents using Hamming distance, Ma et al. proposed a global and local semantics-preserving based deep hashing for cross-modal retrieval in this work author proposed a deep learning architecture where hamming codes and features are learned along within model architectures using local semantic preserving structure and regularization objective functions where the hamming distance function between hash codes are minimized [55]. Li et al. combined self-supervised learning and generative adversarial neural networks (GAN) into a model denominated self-supervised adversarial hashing (SSAH), this model outperformed the state of the art in the cross-modal retrieval task, this model uses semantic labels, text and visual data to learn semantically related hashing codes [49]. Wu et al. also employed a generative approach as Li did but with the advantage that using cycle-consistent loss this model can learn multimodal hash codes without explicit pairing samples [86]. Wang et al. introduced a model able to learn compact hash codes imposing an orthogonal regularization for codes, this regularization reduces the redundancy of information inside hash codes as other hashing methods this algorithm learn features and hash codes from raw data [84].

3.3 Kernel methods

Kernel methods are very popular due to its strong mathematical foundation, that has attracted the attention of mathematicians, statisticians and computer scientists. Kernel methods can perform a mapping from original space known as *input space* to a Hilbert space, known as feature space, where a set of learning algorithms can be applied over transformed data. An advantage of these methods is their ability to perform non-linear mappings without an explicit calculation of transformation, this is achieved using the *Kernel trick*, where applying the kernel function to input data can induce implicitly a non-linear mapping [72].

3.3.1 Large scale approaches

The huge amount of multimodal data generated requires large scale algorithms and methods to process them. Several large scale approaches have been proposed in the literature to deal with large scale data. First, we found the online learning approach where the update

process in the training phase is executed in mini-batches or random subsets of the training set [70, 9, 7, 8], this combined with the Stochastic Gradient Descent optimization method where the solution space is explored finding an optimal value based on information obtained from each mini-batch [6, 39].
In kernel methods, the large scale approaches include the Budget method where the kernel matrix is calculated against a random subset of the training set called a budget that contains a reduced number of samples [85, 13, 21]. Another popular approaches are based on spectral decomposition. In this category we found the Random Fourier Features [66, 14] and the Nystrom methods where the complete kernel matrix is approximated by a low-rank matrix [25, 24, 43, 88].

3.4 Multimodal information retrieval systems

Multimodal information retrieval systems have attracted the attention of researchers and companies for a long time. One of the best-known examples of these systems, we found IBM's content-based image retrieval system QBIC (Querying Image by content using Colour, Texture and Shape) [62]. Although, this system was designed to be used to retrieve images their internal design implies data fusion techniques as used in multimodal retrieval systems because Colour descriptions, Texture and shape act like multimodal representations of the same phenomena, in this case, the image is the natural phenomena and it can be represented with several descriptors sensible to different aspects of the image. QBIC also allowed adding textual information attached to images in the database describing the whole image or a certain region or object within it. However, in this system text data was not used to execute queries or complement image features.

Medical domain is an interesting field where multimodal information retrieval systems can be employed to improve medical practice, patient attention and treatment, reduce costs and times. Müller et al. [60] present a survey where different information retrieval systems are described, despite this work is focused to Content-Based Image retrieval the author draws the attention to the potential of improvement that can be reached combining textual features with visual features and patient records providing contextual information that potentially can improve system performance and enrich image representations.

In 2010 Bozzon and Fraternali [10] present a general architecture for multimodal information retrieval systems where content providers and developers applied several techniques to each modality to get a high-level semantic representation for each modality, understanding this process as the association of image content with high-level semantic concepts such as "sky", "city", etc. To perform a semantic association, the three-stage process was proposed composed by a stage of transformation where preprocessing is performed, the next stage corresponds to a feature extraction operation where low-level representations are generated for each modality and at the final stage, high-level semantics was introduced through classification mechanism, where a low-level representation for data is used to predict high-level concepts and with these high-level concepts with can index and query multimodal data.

Ghosh et al. [26] publish a review where seven multimodal retrieval systems where compared in performance and feature and semantic gaps. Five of these systems retrieve figures from biomedical papers databases using text query and information from captions and full-text content, the remaining two systems IRMA and iMedLine allow text and images queries, using features such as colour, shape and texture to represent image content. Despite, advances showed in these multimodal retrieval systems there is still a huge semantic gap due to employed of global image descriptors that do not take into account region of interest, objects in images and in general ignore semantic content within an image.

Liu et al. [53] applied hashing methods over graphs manifolds to perform related semantic retrieval. This work describes a scalable unsupervised graph-based hashing method to retrieve semantic similar neighbours, this method reach scalability applying the budget approach to approximate the adjacency matrix by a low-rank matrix reducing the training time from quadratic to linear. This method was applied to MNIST(70k) and NUS-WIDE(270k) dataset outperforming other hashing algorithms.

Liu et al.[52] presented a method based on supervised kernel methods. Authors show the method KSH (Kernel-Based Supervised Hashing) the aim of this method is to construct a hash code representation for queries and targets employing kernels to induce non-linear ability to deal with non-linear separable samples, The supervision is used to learn the hash codes according to labels using a greedy algorithm to train the system. The time complexity of the method is bounded by the following function:

$$O((nm + l^2m + m^2l + m^3)r)$$

where n corresponds to the number of samples in the training set, m corresponds to a randomly selected subset of training samples that conform the budget and are used to construct a reduced version of the kernel matrix, l corresponds the label space size conformed for the number of different labels in the train set, and r is the size of hash code in bits. The proposed method was applied to two datasets CIFAR-10 and Tiny-1M, outperforming the previous hashing approaches.

Mourao et al. [59] proposed a system with rank fusion method for multimodal medical information retrieval. The proposed system NovaMedSearch uses best representations available for text and image and allow query expansion using a medical thesaurus, after retrieving data for each modality they can be combined using a ranking fusion approach, this method was tested on ImageCLEF 2013, the experiments showed that in general best performance is achieved when Inverse Square Ranking (ISR) was employed.

Zhang et al.[92] proposed a multimodal re-ranking strategy, where a ranking is constructed for each modality and the final ranking is the combination of the individual rankings using graph fusion approach for heterogeneous and it was tested on 120 breast tissue images from 40 patients achieving 80% accuracy.

More recently Li et al. [50] published a survey on large scale image retrieval systems. In this comprehensive work, several dominant techniques and approaches are described showing the evolution and future directions. Although in this work is notorious the incursion of deep learning as feature extractor, feature encoding and indexing method with great success worth it note that hand-crafted features are still very popular due to its specificity and explainability. Also is important to highlight the more popular methods described many of then are based on re-ranking, hashing, deep learning and graph approaches.

3.5 Conclusion

Properties in multimodal data have been exploited using different approaches some of them employ statistical properties in data to establish relationships between different modal representations, other approaches learn common representations using mapping functions or projections from input spaces to semantic spaces. Many of the first approaches show limitations related to scalability, to overcome these limitations many methods based on low-rank approximation, stochastic gradient descent, and online learning have been proposed. Despite research developed in this area, multimodal information retrieval is still an open issue that requires more research and novelty proposals to solve the challenges that this topic poses.

4 Multimodal information retrieval System for medical case retrieval

This chapter describes a multimodal information retrieval system based on a supervised multimodal kernel semantic embedding model. This system was tested on multimodal medical case dataset composed of histopathology images and text clinical reports showing outstanding results for the multimodal information retrieval task with a Mean Average Precision value of 0.6263. This work was published in Proceedings from 14th International Symposium on Medical Information Processing and Analysis [15].

4.1 Multimodal retrieval system

Complete retrieval pipeline is shown in Figure **4-1**. This pipeline can be described as a process with four phases. Phase 1 performs feature extraction for visual and textual data; phase 2 build a semantic index using the Multimodal Kernel Latent Semantic Embedding for Cross-Modal retrieval (MKSE-CM) algorithm proposed by Vanegas [81]; phase 3 produces a ranking for each modality and then these rankings are combined using late fusion approach to produce a multimodal ranking; in final phase results are shown. These phases are shown in Figure **4-1** with doted color lines.

4.1.1 Feature extraction

The feature extraction phase process the query provided by the user, which consists in a whole slide image (WSI), corresponding to a histopathology sample, along with a short text report describing the pathologist's findings. Each modality is processed independently. For images, the WSI is divided in patches, then each patch is represented as a feature vector using transfer learning from a pretrained convolutional neural network. For the text modality, each report is projected to an embedding using Doc2Vec [46]. The textual and visual features are the input to the semantic indexing system, MKSE-CM, described in the next subsection. MKSE-CM is a supervised algorithm, thus the Gleason score of each WSI was used as the corresponding label.

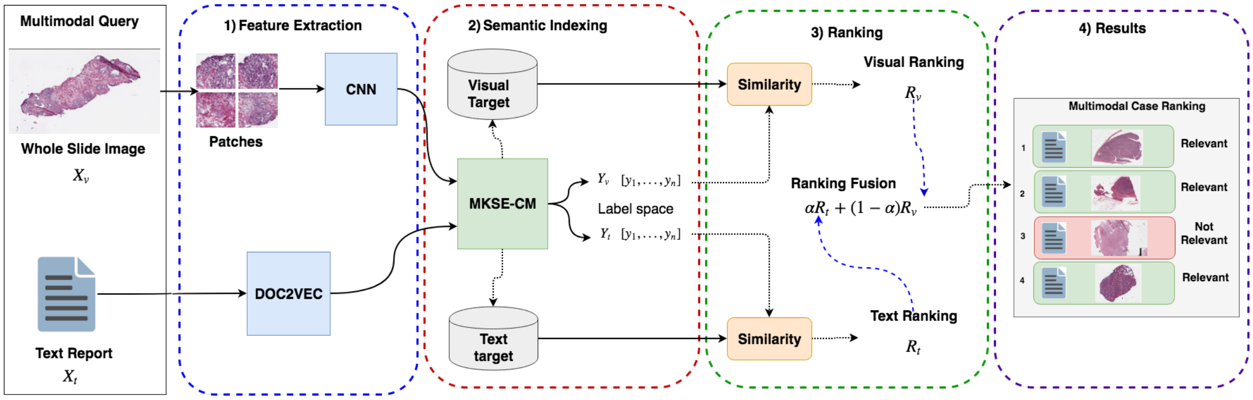

Figure 4-1: Complete retrieval process in the proposed system.

4.1.2 Latent semantic indexing

For representing the multimodal pairs that correspond to cases with visual and textual information we used a multimodal embedding method: Multimodal Kernel Semantic Embedding for Cross-Modal retrieval (MKSE-CM) proposed by Vanegas [81]. MKSE-CM combines a deep learning architecture and kernel methods within a modular approach. The main modules of MKSE-CM are: i) kernelized latent semantic embedding, ii) semantic alignment, iii) label prediction and iv) cross-alignment reconstruction. Figure **4-2** gives an overall view of the method. The next subsections describe the different phases.

Kernelized latent semantic embedding

The input image and text are mapped to an embedding latent space using a non-linear matrix factorization approach [82] based on kernels. A drawback of kernel methods is their poor scalability due to the need of calculating a full kernel matrix of the training samples. A strategy to mitigate this effect is to use a "budget"[19], which is a subset of the training samples. As shown in Figure **4-2**, the kernel of the input samples against the budget is calculated in the kernel projection layer. Visual and textual kernels are embedded in low-dimensional representations, h_v and h_t.

Semantic Alignment

The latent representation for visual, h_v, and textual content, h_t, are aligned by enforcing a similarity constraint. If the cosine similarity of h_v and h_t is small, it is penalized in the loss function. This promotes an alignment between the latent representations of both modalities.

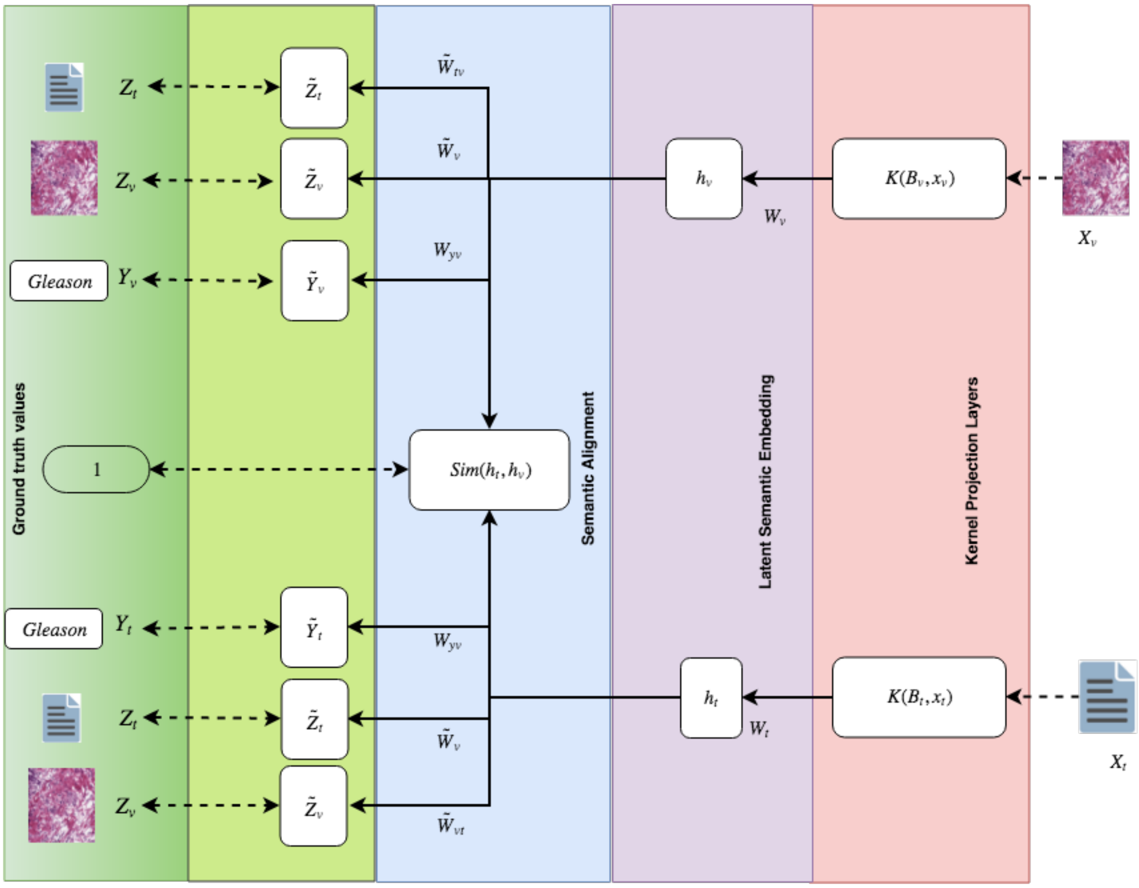

Figure 4-2: Multimodal Kernel Embedding for Cross-modal retrieval (MKSE-CM) architecture.

Label prediction

Including rich semantic information from labels, the model can learn a better latent semantic embedding. Furthermore, this label information also helps to align the embeddings and can be used to represent new data in a label semantic space Y [16].

Cross-alignment reconstruction

An additional mechanism to promote the alignment of the latent representations of both modalities is a cross-reconstruction strategy. The main idea of this strategy is that the model must learn to reconstruct one modality from the latent representation of the other.

4.1.3 Ranking

In this retrieval setup queries are taken from the testing set and the target elements are taken from the training set. The relevance criteria is given by Gleason score. Visual (R_v) and textual (R_t) rankings are combined with late fusion strategy based on a convex combination of unimodal rankings, as it is shown in equation 4-1, where R_m stands for multimodal ranking.

$$R_m = \alpha * R_v + (\alpha - 1) * R_t \tag{4-1}$$

The multimodal ranking (R_m) is used to retrieve results and evaluate performance of the system.

4.1.4 Complexity Analysis

This section presents complexity analysis for MKSE-CM model.

Space complexity

According to Arora et al. and Vavasis [3, 83] space complexity for Non-Negative Matrix Factorization can be constrained by the following function:

$$O(mnp) \tag{4-2}$$

where m is the number of rows in input matrix, n is the number of columns in input matrix, p is the dimension of latent semantic space, in our case we have two input matrices X_t a matrix $m \times n_t$ and a matrix X_v a matrix $m \times n_v$, in this case n_t and n_v correspond to textual and visual features and in general we have $n_t \neq n_v$. However, when we introduce kernel computation with a randomly selected subset of k training samples employed budget method, we got two kernel matrices K_v and K_t each one with size $m \times k$ according to equation 4-2 space complexity required to store these kernel matrices matrices and latent space embeddings can be calculated as:

$$O(2mkp) \approx O(mkp) \tag{4-3}$$

Since $k \ll n$ we get a huge reduction on space complexity applying learning in a budget method. Additional to previous reduction our model can be trained in mini-batches this means that our model can be trained with stochastic gradient descent algorithms on mini-batches of few samples, this lead to following spatial complexity upper bound:

$$O(\eta kp) \tag{4-4}$$

Final space complexity upper bound for MKSE-CM model is shown in equation 4-4 where η is the number of samples in each mini-batch, k the budget size, and p the size of latent semantic embedding space.

Time Complexity

According to Arora et al., Gillis et al. and Vavasis [3, 27, 83] exact solution to NMF problem is a Np-hard problem with the following upper bound function:

$$(mn)^{O(p^2)} \tag{4-5}$$

where m is the number of rows in input matrix, n is the number of columns in input matrix and p corresponds to latent semantic embedding space size.
In Gillis and Vavasis works an online learning setup to find an approximate solution for NMF problem is presented, this solution can be computed in the quadratic upper room:

$$O(n^2m + \eta(nm + np^2)) \tag{4-6}$$

In equation 4-6 n is the number of columns in input matrix, m is the number of rows in input matrix, p is the size of latent semantic space and η is the number of mini-batch samples per epoch.
Extending equation 4-6 to apply learning in a budget and kernel methods we have the following equation:

$$O(2k^2m + 2\eta(km + kp^2)) \tag{4-7}$$

when k is the budget size, the number of training samples randomly selected and used to computed a reduced kernel matrices for each modality, m is the number of training samples, p the size of latent semantic embedding.
As k is fixed value we can factorize it from equation 4-7, we get:

$$O(2k(km + \eta(m + p^2))) \tag{4-8}$$

Removing constant factor $2k$ for upper bound equation 4-9 we have the final upper bound for MKSE-CM:

$$O(km + \eta(m + p^2)) \tag{4-9}$$

from equation 4-9 as p is a fixed hyper-parameter value usually $p \ll n$ we can conclude that MKSE-CM algorithm complexity is linear on number of samples m.

4.2 Experimental Setup

The goal of the experimental evaluation is to measure the effectiveness of the proposed multimodal retrieval system to find a ranking of relevant clinical cases given a multimodal query. Clinical cases are composed of histology images and the pathologist's text report.

4.2.1 Dataset Description

TCGA-PRAD is a publicly available dataset from the Cancer Genome Atlas project [1] the one that contains a total of 500 cases of prostate adenocarcinoma. We use a subset of TCGA-PRAD similar to proposed by Jimenez-del-Toro et al. [79], with a total of 235 cases. Each case is composed of a Whole-Slide-Image (WSI) and a pathologist's report. From the total cases, 141 were taken as the train set, 48 as validation set and 46 as the test set. The Gleason score for each WSI was extracted manually from reports, thus each case is constituted by WSI, textual pathologist's report and Gleason score. The Gleason score is used for train the MKSE-CM and for construct ground truth for retrieval evaluation.

4.2.2 Feature Extraction

The next subsections describe the feature extraction process for image and text modalities.

Whole-slide Image Representation

The feature extraction process for WSI begins with sampling image in 5000 random patches with the same size. Then, the patches are passed through a Convolutional Neural Network (CNN), fined tuned to predict Gleason score, as is described by Jimenez-del-Toro et al. [79], the CNN employed for feature extraction was based on the GoogleNet architecture whose last layer was modified to predict high and low Gleason score. The feature representation for each patch was taken from the layer previews to softmax, which has 1024 neurons, thus each patch is represented by a 1024-length vector with negative and positive values.

Text Report Representation

The feature extraction process for pathologist's text report uses Doc2vec model [46], appropriate for documents with a variable length. The model was trained with an embedding dimension of 100. The text representation consists of embedding vector predicted from Doc2vec, thus each text report associated to one WSI is represented as a feature vector of length 100.

The MKSE-CM is a supervised model, hence the Gleason score grade associated with each WSI was used as label supervision.

[1] Available in https://portal.gdc.cancer.gov/.

4.2.3 Baseline description

The baseline is reported by Jimenez-del-Toro et al. [79], they proposed a multimodal information retrieval system for case retrieval based on late fusion. The information retrieval system was tested on TCGA-PRAD dataset.

The baseline system starts with feature extraction process for textual and image data, similar to the described in section 4.2.2. Then, a ranking is constructed for each modality measuring the similarity between query and target features. Finally, the two unimodal rankings are combined using late fusion to produce the multimodal ranking.

4.2.4 Performance Evaluation

The relevance criteria used to determine if a clinical case is relevant for a given query is based on the Gleason score: a clinical case is relevant if it shares the Gleason score value with the query.

To measure the effectiveness of the proposed information retrieval system, we employ a set of commonly used measures in the literature. The most common measure reported in information retrieval works is Mean Average Precision (MAP). Precision is another measure of quality for evaluated information retrieval systems. Precision measures the proportion of relevant elements in the retrieved set. Precision measure does not take into account order in result set, to provide additional information about the quality of result set precision is measure at top 10 and 30.

Preferential-based (Bpref) is a performance measure that takes into account the relevant elements retrieved first than non-relevant.

All performance measures mentioned in this section are described by a real value between 0 and 1, best performance is reached when measure value is close to 1.

4.2.5 Experimental Evaluation

Experimental execution starts with feature extraction process for images and text as described in section 4.2.2.

The system requires multimodal cases, image-text pairs, for training phase. In this phase, the system requires supervision, which is provided by the Gleason score associated with each WSI.

MKSE-CM is trained at patch level: each patch was associated with their concerned text report and Gleason score.

Since the model was trained with patches but retrieval task is evaluated at WSI level [79], we construct the WSI ranking from patches predictions, calculating cosine similarity between query patches and target patches and then aggregate patch similarities into global WSI similarity measure. With global WSI similarities we construct the image ranking, similar process is applied to construct textual ranking from text predictions that accompanied image patches.

Results are evaluated using the TREC-eval tool developed for evaluated results in the Text REtrieval Conference (TREC), the tool was provided with the multimodal Ranking prediction and the ground truth file. Ground truth is constructed under the relevance criteria that defined two cases are relevant if they share the same Gleason score. The tool produces a result file with metrics such as MAP, GM-MAP, precision at 10, precision at 30, etc.

4.2.6 Hyperparameter selection

Hyperparameter selection was performed using a random search over 25 configurations, each configuration was tested on validation set five times, the configuration with the highest mean MAP on validation was evaluated on the test set. The model output is a 5-dimensional probabilistic vector that describes the sample's membership to a certain Gleason score from 6 to 10. This vector space is called label semantic space ($\mathbf{Y}$).

4.3 Results and Discussion

As the baseline, we used the approach presented by Jimenez et al. [79], based on late fusion. Their experimental setup starts with a feature extraction process for each modality as is described in subsection 4.2.2, then a ranking is constructed for each modality, and finally the two rankings are combined using late fusion and convex combination of the unimodal rankings as is described in equation 4-1, to produce multimodal ranking. Using unimodal and multimodal rankings, they evaluate retrieval performance of each approach. That results are shown in the top of the table **4-1**. Their reported results demonstrated the advantage of multimodal retrieval method, based on late fusion, over unimodal retrieval approaches.

In top of the table **4-1** *Image retrieval* reported the unimodal retrieval results using CNN visual features, *Text retrieval* reported the unimodal retrieval results using Doc2vec features and finally *Multimodal* stands for multimodal ranking construction based on late fusion of image and text rankings, these results constitute our baseline[79].

Our approach is based on kernelized latent semantic embedding, results obtained in label semantic space ($\mathbf{Y}$) are reported in the lower part of the table **4-1** we report our results for linear, cosine and Radial Basis Function (RBF) kernels, each result also has a budget size described in the table with capital B letter and the budget size employed.

Results in table **4-1** show that our method has better performance than baseline when the budget size is enough large, this is very interesting result since the budget was selected randomly from train patches whose total number is approximately 289.000, with only 10.000 we get competitive results. For the other kernels, we use short budget size because as the budget size increases the size the computational complexity and space complexity grow exponentially with and strong impact on performance when complex kernels are used such as RBF.

Retrieval Method	MAP	GM-MAP	bpref	P@10	P@30
Image retrieval [79]	0.5113	0.3921	0.4706	0.4500	0.4600
Text retrieval [79]	0.4092	0.3561	0.31116	0.4913	0.3775
Multimodal [79]	0.5404	0.4196	0.4890	0.5217	0.4884
MKSE-CM linear Kernel B=10k	**0.6263**	**0.4843**	**0.6425**	0.5667	**0.6326**
MKSE-CM cosine Kernel B=4k	0.5044	0.3838	0.4301	**0.5979**	0.5590
MKSE-CM RBF Kernel B=4k	0.4835	0.3675	0.4002	**0.5979**	0.5090

Table 4-1: Evaluation results for baseline and our method.

4.4 Conclusions

In this chapter, we present a multimodal retrieval system based on online kernel matrix factorization, cross-alignment, semantic alignment, supervised label prediction, and late fusion. Results show that the proposed strategy has advantages over unimodal retrieval methods, taking advantage of rich multimodal information and label supervision. We projected the original representations to a rich semantic label space (Y) that is compact, low rank and semantically significant because vector values in that space are related with Gleason scores of cases. Additionally, our approach has scalability advantages thanks to the application of a learning in a budget strategy, which uses only a small subset of the train set, improving train speed and limiting the memory usage to the size of the selected budget.

We found a strong correlation between budget size and algorithm accuracy, as it is shown in results table **4-1**, the bigger the budget proportion respect to train set the better accuracy is obtained, but at the same time, as the budget is increased the algorithm performance decays.

Results from table **4-1** show the advantage of multimodal approaches over unimodal in the information retrieval task applied to histopathology image analysis. Multimodality reduces the semantic gap in medical image analysis and constitutes a suitable method to construct a support decision tool to help pathologists in their work, retrieving cases and ranked images.

5 Multimodal Kernelized Latent Semantic Regression

This chapter introduces the multimodal kernelized latent semantic regression model. This model performs a supervised non-linear latent semantic embedding using kernel methods and matrix factorization techniques. The model was used to execute the indexing stage within a multimodal information retrieval system.

5.1 Introduction

Multimodal information retrieval is a challenging task where queries and target documents are expressed as a combination of one or more data representations, e.g. text, images, audio, etc. Dealing with this variety of representations poses different challenges: how to build algorithms that successfully combine media information to produce accurate information retrieval results, how to combine heterogeneous information to produce a consistent representation, how to determine relevance between queries and target documents when they are represented in different representation spaces.

The multimodal information is obtained from the measurement of the same phenomena through different sensors and scales [44], this implies that different phenomena's characteristics are captured and represented using different modalities. Each modality contains common and complementary information about the phenomena under analysis, the additional information provided by each modality can be used to improve information retrieval performance [93]. In clinical settings, valuable multimodal information is stored but not fully exploited, for example in PACS systems were radiology departments store radiology images, they are usually accompanied by text data in the form of clinical records and radiology reports. Also in computational pathology laboratories, digital slides repositories store whole slide images together with pathology reports which are valuable resource for helping physicians to make informed decisions about rare cases by retrieving similar cases[60].

Notably, the computational pathology setting poses additional challenges, coming from the nature of the modalities, where images can consist of up to 800.000^2 pixels and pathologist write reports in natural language with variable lengths and specialized medical terminology. Therefore, designing multimodal strategies that models and fully exploits the correlations of these complex high dimensional spaces is imperative to simplify and improve clinicians workflows [78].

In this paper, we propose a new information retrieval system based on a method that learns a semantic embedding of multimodal data. This semantic embedding is used by the retrieval system as a high-level representation to calculate a multimodal ranking of medical cases. The semantic embedding is built using the Multimodal Kernelized Latent Semantic Regression method (MKLSR) [82], which performs a kernelized matrix factorization as an embedding strategy and an information fusion technique applied to learn a multimodal latent semantic embedding. We evaluated the proposed approach over a multimodal histopathology dataset that contains documents including text and image modalities. The results show that this approach is successful in exploiting the multimodal content to find relevant documents.

5.2 Multimodal Kernelized Latent Semantic Regression

Multimodal Kernelized Latent Semantic Regression is a complex model that integrates several processes, for an easier analysis and understanding of the model, it was divided into the following parts: 1) Kernel projection; 2) Latent semantic embedding; 3) Semantic alignment; 4) Multimodal latent semantic embedding; 5) Regression. These parts are shown in figure **5-1** inside dotted-line boxes with a different color for each one.
In the following sections, each model part is described, and the whole model architecture is analyzed.

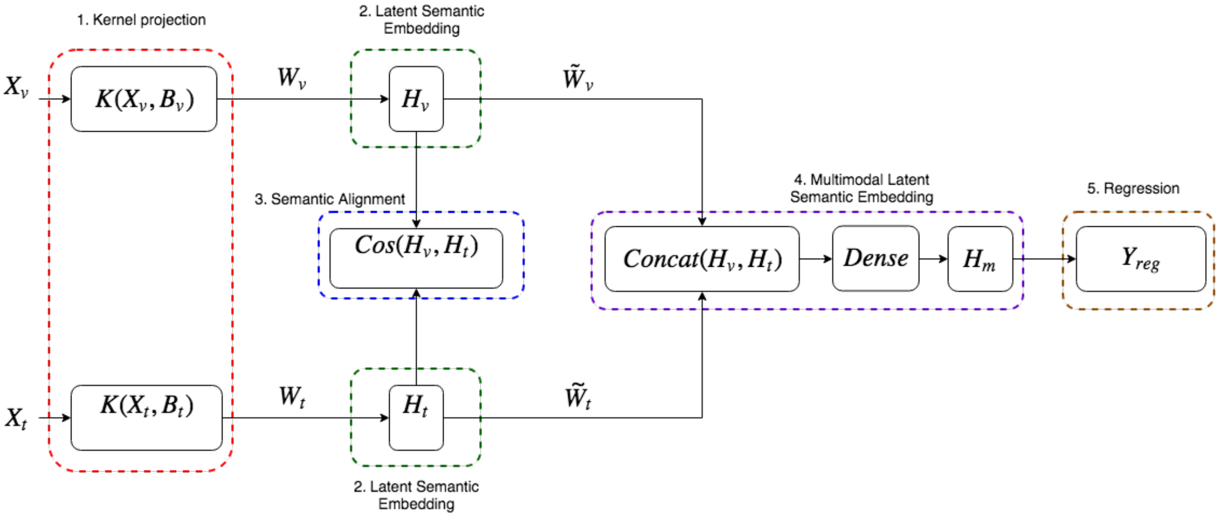

Figure 5-1: MKLSR model architecture.

5.2.1 Kernel projection

MKLSR model takes as input text feature vectors $X_t \in \mathbb{R}^{d \times n}$ with d samples of n features and visual feature vectors $X_v \in \mathbb{R}^{d \times m}$ with d samples of m features. These feature vectors are embedded into a high-dimensional Hilbert space using non-linear projection function.

5.2.2 Latent semantic embedding

A non-linear latent semantic embedding for each modality can be learned using online kernelized matrix factorization algorithm proposed by Vanegas [81]. According to this formulation find a latent semantic embedding in feature space for the i-th sample x^i is equivalent to minimize the following loss function:

$$\begin{aligned}\underset{W,\tilde{W}}{\mathtt{min}} J^i_\phi\left(W,\tilde{W}\right) &= \tfrac{\alpha}{2}\Big(K\left(x^i,x^i\right) - 2K\left(x^i,B\right)\tilde{W}WK\left(B,x^i\right) \\ &+ K\left(B,x^i\right)^T W^T\tilde{W}^T K\left(B,B\right)\tilde{W}WK\left(B,x^i\right)\Big)\end{aligned} \tag{5-1}$$

where $K(\cdot,\cdot)$ is a kernel function that implicitly compute inner product of entries in feature space, B is the Budget matrix, a matrix with b samples, randomly selected from train data with d samples that satisfy the condition $b << d$. The optimization process minimize the loss function updating the values of projection matrices W and $\tilde{W}$ with stochastic gradient descent optimizer.

Applying equation 5-1 to textual and visual modalities we have the following equations:

$$\begin{aligned}\underset{W_v,\tilde{W_v}}{\mathtt{min}}\ J^i_{\phi v}\left(W_v,\tilde{W_v}\right) &= \tfrac{\alpha}{2}\Big(K\left(x^i_v,x^i_v\right) - 2K\left(x^i_v,B_v\right)\tilde{W_v}W_vK\left(B_v,x^i_v\right) \\ &+ K\left(B_v,x^i_v\right)^T W_v^T\tilde{W_v}^T K\left(B_v,B_v\right)\tilde{W_v}W_vK\left(B_v,x^i_v\right)\Big)\end{aligned} \tag{5-2}$$

$$\begin{aligned}\underset{W_t,\tilde{W_t}}{\mathtt{min}}\ J^i_{\phi t}\left(W_t,\tilde{W_t}\right) &= \tfrac{\alpha}{2}\Big(K\left(x^i_t,x^i_t\right) - 2K\left(x^i_t,B_t\right)\tilde{W_t}W_tK\left(B_t,x^i_t\right) \\ &+ K\left(B_t,x^i_t\right)^T W_t^T\tilde{W_t}^T K\left(B_t,B_t\right)\tilde{W_t}W_tK\left(B_t,x^i_t\right)\Big)\end{aligned} \tag{5-3}$$

5.2.3 Semantic alignment

As was mentioned in chapter 2 latent representation of data samples are found in encoding matrix H. In MKLSR the non-linear latent semantic embedding is performed using equation 5-1 in each modality. Latent representations in H matrix are implicitly defined in the equation 5-1, making explicit for i-th sample in each modality we obtain the following equations [81]:

$$h^i_v = W_vK(B_v,x^i_v) \tag{5-4}$$

$$h^i_t = W_tK(B_t,x^i_t) \tag{5-5}$$

Equations 5-4 and 5-5 correspond to two independent latent semantic spaces with same size but unrelated. To take advantage of the additional information in multimodal data, an alignment process between these two spaces is introduced. The alignment process enforces the semantic similarity between latent spaces, using a cosine similarity function between them. The following equation describes the optimization problem to align latent semantic spaces:

$$\min_{W,\tilde{W}} J^{i}_{\phi sim} = (cos\_sim(h^{i}_{v}, h^{i}_{t}) - 1)^{2} \tag{5-6}$$

Minimizing equation 5-6 is equivalent to align latent semantic spaces H_v and H_t.

5.2.4 Multimodal latent semantic embedding

Latent semantic spaces H_v and H_t described in the previous section are per-modal spaces. Despite these spaces are alignment and comparison between them are possible, they are not multimodal spaces.
To construct a multimodal space, we combine uni-modal embeddings H_v and H_t concatenating them. Then, concatenated spaces are passed through a dense neural network layer with relu activation function and f neurons, to introduce an additional abstraction level the output of the dense layer is the multimodal latent semantic embedding that encodes information from previous embeddings into a unique latent semantic space H_m.

5.2.5 Regression

Regression is the final module in MKLSR algorithm, regression module takes the multimodal latent semantic embedding representation as its input and produces a real number output value $Y_reg \in \mathbb{R}$ as its output. The loss function for regression module is defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Y_{true} - Y_{reg}| \tag{5-7}$$

where Y_{true} is the ground truth real value and Y_{reg} is the output of MKLSR model, n is the total number of train samples. MAE stands for multimodal Absolute Error a common loss in regression problem.

5.2.6 MKLSR loss function

MKLSR architecture is shown in figure **5-1**, dotted-line boxes highlights the modular construction of the MKLSR, arrows in the figure show the data flow within model in a feed-forward phase and solid boxes represents embeddings, operations or layers.
In previous sections MKLSR modules were described, in order to earn a better understanding of this model, module losses was combined into the complete loss function for the model, described as follows:

$$
\begin{aligned}
\min_{W} J^i(W) &= \alpha_1 J^i_{\phi t} + \alpha_2 J^i_{\phi v} \\
&+ \alpha_3 |y^i_{true} - y^i_{reg}| \\
&+ \alpha_4 (\mathbf{cos\_sim}(h^i_v, h^i_t) - 1)^2 \\
&+ \lambda_{1..5} \left\| W \right\|_2^2
\end{aligned}
\tag{5-8}
$$

where $J^i_{\phi t}$ and $J^i_{\phi v}$ are the loss function for latent semantic embedding in textual and visual data as was defined in equations 5-3 and 5-2. $W = (W_t, \tilde{W}_t, W_v, \tilde{W}_v, W_m)$ are the model's parameters to be learned during training phase. $\alpha_{1,\ldots,4}$ controls relative importance of module losses within the total loss function, $\lambda_{1,\ldots,5}$ controls the effect of regularization terms added to avoid overfitting. MKLSR algorithm is described in algorithm pseudo-code 1.

```
MKLSR ($\boldsymbol{S} = \langle \boldsymbol{X_v}, \boldsymbol{X_t}, \boldsymbol{Y_{true}} \rangle$);
Inputs           : $\boldsymbol{X_v}$: visual features representation,
                   $\boldsymbol{X_t}$: textual features representation,
                   $\boldsymbol{Y_{true}}$, label ground truth values
Outputs          : $\mathcal{M}_{th} : \mathcal{X}_v \to \mathcal{H}_t$;
                   $\mathcal{M}_{vh} : \mathcal{X}_t \to \mathcal{H}_v$:
                   $\mathcal{M}_{hm} :< \mathcal{X}_t, \mathcal{X}_v > \to \mathcal{H}_m$:
                   $\mathcal{M}_{hmy} : \mathcal{H}_m \to \mathcal{Y}_{reg}$;
Hyperparameters: $b$: budget size, $l$: semantic space dimension $f$: multimodal
                   semantic space dimension
/* Budget construction                                                    */
 $\boldsymbol{B} = \langle \boldsymbol{B_v}, \boldsymbol{B_t} \rangle \subset \langle \boldsymbol{X_v}, \boldsymbol{X_t} \rangle \;\; : \;\; |\boldsymbol{B}| = b \ll n$
/* learn mapping functions to the semantic spaces $\mathcal{H}$ and $\mathcal{Y}$             */
$\mathcal{M}_{th}, \mathcal{M}_{vh}, \mathcal{M}_{hm}, \mathcal{M}_{hmy} = MKLSR(\boldsymbol{S}, \boldsymbol{B}, l, f)$;
return $\mathcal{M}_{th}, \mathcal{M}_{vh}, \mathcal{M}_{hty}, \mathcal{M}_{hvy}$;
```

Algorithm 1: Multimodal Kernelized Latent Semantic Regression (MKLSR)

5.2.7 MKLSR prediction

Once MKLSR model has been trained, input data can be projected to a latent and label spaces passing it through model. This model can project input data into three different spaces:

- **Latent semantic space per modality:** Visual and textual data are projected to H_v (latent semantic space for visual data) and H_t (latent semantic space for text data), this is, each modality is projected individually to its corresponding space. Latent spaces H_t and H_v are real value spaces with size equal to the number of latent topics

l.

For text data the mapping prediction is defined as follows:

$$x_t \in \mathbb{R}^{d\times n} \xrightarrow{\mathcal{M}_{th}} h_t \in \mathbb{R}^{d\times l} \tag{5-9}$$

For text data the mapping prediction is defined as follows:

$$x_v \in \mathbb{R}^{d\times m} \xrightarrow{\mathcal{M}_{vh}} h_v \in \mathbb{R}^{d\times l} \tag{5-10}$$

where d is the number of data samples to be projected, n is the number of input features for text data, m is the number of input features for visual data and l is the number of latent topics for these spaces.

- **Multimodal latent semantic space:** Multimodal latent semantic space $H_m \in \mathbb{R}^{d\times f}$ is a common space that combines H_t and H_v. Projection to these space can be define as follows:

$$(h_t, h_v) \xrightarrow{\mathcal{M}_{hm}} h_m \in \mathbb{R}^{d\times f} \tag{5-11}$$

where f is the dimension of multimodal latent semantic space.

- **Label space:** This space is a real space $Y \in \mathbb{R}$ that predict a continue value as is define as follows:

$$h_m \xrightarrow{M_{hmy}} y_{reg} \in \mathbb{R} \tag{5-12}$$

5.2.8 MKLSR in multimodal retrieval process

In the previous section, three different outputs of the MKLSR model were described. This section explains how these outputs can be used to perform an information retrieval task.

- **Latent semantic space per modality:** This output produces two latent representations h_t and h_v, with the same dimension size. This characteristic allows to apply similarity measures between them. There are two possible configurations that can be used with these outputs. First to perform cross-modal retrieval, in this setup the query is represented using one semantic space representation and in the retrieval phase, it is compared with target database elements represented in the other modality. The second setup executes multimodal information retrieval with late fusion, in this case, the multimodal ranking is constructed combining visual and textual rankings, generated comparing query latent representation with target elements one for each modality and then these results are combined using any of the available late fusion techniques.

- **Multimodal latent semantic space:** With this representation performs multimodal information retrieval is easy we need only compute the similarity between the multimodal representation of query and target database elements and construct a ranking using this information. Under this setup, all process is multimodal for this reason queries and database target are represented with both modalities.

- **Label space:** This space is a real number prediction that encodes specific domain information about Gleason scores used in this experiment, these scores are real values and follow an ordered similarity, this is, for example, a sample with Gleason score 6 is more similar to a sample with Gleason score 7 than a sample with Gleason score 10. To perform a multimodal information retrieval in this space we only need to calculate the absolute difference between query label value and target elements and rank items based on that value.

Algorithm complexity

Time and space complexities for this algorithm are identical to those described in subsection 4.1.4 and can be summarized in equation 4-4 for space complexity and equation 4-9 for time complexity, where we find that this algorithm is linear in number of samples, which is an scalable and competitive complexity.

5.2.9 Implementation details

MKLSR was implemented as a computational graph taking advantage of deep learning programming frameworks like Keras [30] or Tensorflow [1] that also allow using GPU hardware to improve training speed. To take full advantage of scalability features of deep learning frameworks the model was reformulated as deep learning model especially the latent semantic embedding module described in equation 2-3 was reformulated as an auto-encoder for each modality thus the model was trained using adaptive stochastic gradient descent optimizer Rmsprop on Keras framework.

5.3 Experimental evaluation

5.3.1 Dataset description

TCGA-PRAD is a prostate adenocarcinoma multimodal dataset composed of clinical cases with tissue samples captured in high-resolution image *Whole slide image* (WSI) and pathologist's text annotations from these annotations the global Gleason score for each case was extracted. The Gleason score values are between 6 and 10, and measure the aggressiveness of prostate cancer; Gleason score values 6 and 7 are lower score and values 8,9 and 10 are high score.

TCGA-PRAD contains 500 cases in total but in order to balance the samples for each Gleason score and preserve same experimental setup of baseline [79] only 141 cases were used for this experiment.

5.3.2 Feature extraction

The feature extraction process for each modality is presented in the following subsections:

Whole-slide image feature representation

The feature extraction process begins with division of Whole-slide images in patches, then each patch is processed by a fine-tuned convolutional neural network modified to classify image patches into high or low Gleason score, then 2000 patches with higher predicted values was selected to represent WSI [77]. After previous process each WSI is represented by 2000 feature vectors, with the aim of obtain compact representations for WSI, features vectors per WSI are combined using bag of visual words (BoVW) model [63].
Finally after exploring different vocabulary sizes for BoVW model the WSI was represented by a histogram of 300 visual words.

Text report representation

Text reports in this dataset are very short in many cases consist of some abbreviations and a Gleason score and are of variable length, to extract meaning from this modality we employ a document embedding algorithm doc2vec[46] with windows size of 5.

5.3.3 Baseline description

The baseline for this experiment is composed of two works, first ”deep multimodal case-based retrieval for large histopathology datasets”, in this work Jimenez-del-Toro et al. [79] presents a late fusion approach to perform multimodal information retrieval over the prostate dataset described in previous sections. In this work, the WSI is divided into around 5000 patches then each one was classified using a fine-tuned deep neural network trained to classify low and high Gleason scores, then around 2000 patches with higher predicted values are used to represent the WSI image. The text modality is represented using a document embedding doc2vec with a window size of 1 and Gleason score is utilized as a label for all case. A multimodal ranking is constructed using a convex combination of unimodal rankings. The ground truth relevance criteria are produced comparing the Gleason score of query and target database element if these two share a label, they are relevant.
The second baseline is based on work published by Contreras et al. [15] where they used the same features representation and relevance judgment with an MKSE-CM algorithm as indexing method.

5.3.4 Experimental setup

As was described in previous sections MKLSR model can projected data from input space to three different spaces each of one has its characteristics and procedures next the multimodal information retrieval process for each space projection is described:

- **Latent semantic space per modality:** To perform the multimodal information retrieval task in this space a ranking was constructed for each modality comparing H representations of query and target database, then the unimodal rankings R_v and R_t was joined using a convex combination to produce multimodal ranking and retrieval results.
- **Multimodal latent semantic space:** To perform a multimodal retrieval task in this space a similarity function is applied between query and target database elements, using the similarity values the multimodal ranking is constructed.
- **Label space:** The output of label space is a real number value, to perform the multimodal retrieval in this spaces a distance metric is computed between query value and target database elements, using this metric the ranking is constructed.

5.4 Results and discussion

As was mentioned in previous sections the MKLSR model can project data into three different spaces. Table **5-1** shown MAP results obtained after performing a multimodal information retrieval on latent semantic spaces H_v and H_t applying different kernel combinations for visual and textual data and three similarity functions, cosine similarity, euclidean distance and centered correlation distance. The best results in this space was obtained when a RBF kernel was applied to both modalities, similar results was obtained with the other kernel except when linear kernel was applied to one or both modalities, this result shows the ability of kernels to encode non-linear relationships.

In table **5-2** shown MAP results for different kernels in the multimodal latent semantic space H_m. Results in this table show a performance improvement compare with results in the table **5-1**.

The results shown in the table **5-3** show an improvement with respect to the shown in previous spaces, this is due to this space is contains the highest semantic content and is directly related with relevance judgment employed for this algorithm.

Results shown in tables **5-4** compare the proposed MKLSR model against baseline results with different metrics, in this table only the best results in each approach are included. The results in this table show that in MAP MKLSE surpasses the performance of MKLSE-CM model but in general these two algorithms have a some similar behavior because they are based on the same principle but implemented in different ways.

Kernel	$\mathbf{H_{euc}}$	$\mathbf{H_{cos}}$	$\mathbf{H_{corr}}$
Linear-Linear	0.319	0.342	0.341
Cos-Cos	0.547	0.552	0.552
RBF - RBF	**0.566**	**0.583**	**0.583**
Chi-Cos	0.554	0.568	0.567
Chi-Linear	0.444	0.405	0.399
Chi-RBF	0.590	0.543	0.543
RBF-Cos	0.514	0.551	0.551
RBF-Linear	0.347	0.364	0.365

Table 5-1: MAP results for multimodal information retrieval in latent semantic spaces H_v and H_t. With different similarity measures euc: Euclidean distance, cos: Cosine similarity corr: centered correlation.

Kernel	$\mathbf{H_{euc}}$	$\mathbf{H_{cos}}$	$\mathbf{H_{corr}}$
Linear-Linear	0.521	0.482	0.472
Cos-Cos	0.582	0.563	0.564
RBF - RBF	**0.612**	0.587	0.589
Chi-Cos	0.581	0.586	0.582
Chi-Linear	0.508	0.437	0.438
Chi-RBF	0.611	0.550	0.551
RBF-Cos	0.584	**0.601**	**0.607**
RBF-Linear	0.456	0.500	0.453

Table 5-2: MAP results for multimodal information retrieval in multimodal latent semantic space H_m. With different similarity measures euc: Euclidean distance, cos: Cosine similarity corr: centered correlation.

Kernel	$\mathbf{y_{mae}}$	$\mathbf{Y_{cos}}$	$\mathbf{Y_{corr}}$	$\mathbf{Y_{euc}}$
Linear-Linear	0.613	0.597	0.597	0.597
Cos-Cos	**0.655**	0.617	0.617	0.617
RBF - RBF	0.651	0.628	0.628	0.616
Chi-Cos	0.624	0.609	0.609	0.616
Chi-Linear	0.642	0.616	0.616	0.616
Chi-RBF	0.653	**0.635**	**0.635**	**0.635**
RBF-Cos	0.633	0.630	0.626	0.626
RBF-Linear	0.589	0.565	0.565	0.565

Table 5-3: MAP results for multimodal information retrieval in label space Y_reg. With different similarity measures euc: Euclidean distance, cos: Cosine similarity corr: centered correlation, mae: mean absolute error.

Retrieval Method	**MAP**	**GM-MAP**	**bpref**	**P@10**	**P@30**
Image retrieval [79]	0.5113	0.3921	0.4706	0.4500	0.4600
Text retrieval [79]	0.4092	0.3561	0.31116	0.4913	0.3775
Multimodal [79]	0.5404	0.4196	0.4890	0.5217	0.4884
MKSE-CM [15]	0.6263	**0.4843**	**0.6425**	0.5667	**0.6326**
MKLSR	**0.6552**	0.4536	0.5952	**0.5696**	0.5609

Table 5-4: Evaluation results against baseline.

5.5 Conclusions

In this chapter MKLSR model was presented, this model can projected data from input space to a three different spaces that have shown according to results in tables **5-1**, **5-2**, **5-3** a competitive performance in the multimodal information retrieval task. MKLSR model shows a similar performance in the task as MKSE-CM model employed for same task in chapter 4, this is due to they common concept design. The best results in the multimodal information retrieval task was obtained in both algorithms MKLSR and MKSE-CM when the retrieval was executed over label space (Y), this is due to label space incorporates a greater amount of semantic information provided by labels in dataset, in this case Gleason scores.

6 Conclusions and Future Work

In this chapter principal conclusions and remarks obtained from this research work are exposed.

6.1 Conclusions

6.1.1 Kernel functions

Kernel function selection has a huge impact on the model performance, as was shown in the results of experiments in chapters 4 and 5. For this reason, kernel function selection is a crucial hyper-parameter for models. Despite experimentation, we cannot determine a priori what kind of kernel is the most appropriate to a certain data modality or application; thus kernel function selection is an empirical process where several kernels should be tested for each modality.

6.1.2 Model semantic relationships

Feature extraction process and kernel function selection are critical to obtaining the best performance of these methods. However, this is only the first step, in the process to ensure the learning semantic relationships between low-level feature, kernel representation and high-level semantic concepts we employed the supervision process with labels as a codified representation of high-level concepts. Use supervised setup improve embedding quality and whole model performance in retrieval and classification tasks.

6.1.3 Joint Representation

Find the best way to produce a joint representation of heterogenous multimodal data is one of the main objectives of this work. From experimentation executed on different configurations, we can conclude that the best way to produce a joint data representation from multimodal data is through intermediate fusion approaches, these techniques can flexibly combine multimodal data representations with weights learned as the trained process. For instance, in the method presented at chapter 5 a joint representation is constructed combining latent semantic embeddings for each modality with a dense neural network that learnt weights for each component on the concatenated multimodal vector. Intermediate fusion techniques show better performance than late or early fusion.

6.1.4 Non-linear mappings

Through *kernel trick* a non-linear mapping to a high dimensional Hilbert space can be induced without explicit data projection. Selecting different kernel functions different linear or non-linear mappings can be induced. Kernel trick is pretty useful to deal with non-linear separable data in a transparent form as if it were linear-separable data.

6.1.5 Supervised learning to narrow the semantic gap

One of the most complicated challenges issues in machine learning is reducing the semantic gap, the difference between data representation and meaning. Despite, the difficulty of this issue, we found that supervised learning methods and inclusion of labels within the training stage reduce the semantic gap and improved learnt latent representations. An additional advantage of supervised learning setup is that models can predict labels, being able to also perform classification or regression tasks.

6.1.6 How to design a large scale multimodal retrieval systems

To develop a large scale information retrieval systems we need to ensure that selected algorithms and indexing techniques have space and time worst-case complexities bounded by the lowest complexity functions as is possible, in an ideal case bound functions are constant or linear in the number of samples. Usually, highly scale systems exceeded hardware capacities, for this reason, online learning techniques play an essential role such as stochastic gradient descent and train on mini-batch, with these kinds of techniques only a small subset of whole training data was loaded in memory each time. Another problem to consider is the *curse of dimensionality* both in training and indexing phase, this issue can be treated with several techniques, for instance, learning in a budget or dimensionality reduction techniques, to reduce computing time kernel matrices can be precomputed and load it for each min-batch.

In the inference stage, these systems need to fulfil several non-functional requirements such as high availability, high performance, high throughput and fault tolerance. To successfully fulfil user's requirements distributed systems for serving and training will be highly recommendable.

6.1.7 Impact of budget size and selection method

The budget size has a huge impact on the effectiveness and efficiency of the proposed models. There is a trade-off between efficiency and effectiveness, because as the budget size increases the effectiveness and accuracy of the model also increases, but on the other side, as the budget size is increased the efficiency and performance of the model decreases.

The technique employed to select budget samples also have a huge impact on this model, at this moment best performance was achieved when budget samples were selected randomly, other methods such as using clustering centroid as budget's samples, or sorted selection had got poor results.

6.1.8 Similarity function

The functions applied to determine the similarity between the query and target database representations have a quite impact on the multimodal information retrieval process. In this thesis, we found that some distance function performs better than others in many cases, this is the case of the cosine similarity and centered correlation functions that take advantage of the geometric and the statistics properties of vector spaces with inner product like feature spaces inducted by kernel functions.

6.1.9 Latent semantic embeddings

Label space retrieval

The previous results show that the best performance in multimodal information retrieval are obtained on label semantic spaces Y, this is due to these spaces incorporate a rich semantic and domain information through class labels.

Multimodal latent space retrieval

The proposed algorithm MKLSR can produce two different latent representations. In the first representation, the model learns a latent representation for each modality (H_v, H_t) and then aligns them using cosine similarity. In the second representation, the previously learned representations (H_v, H_t) are concatenated and passed through a dense layer with relu activation producing a new multimodal latent representation H_m. Comparing the results obtained by these two representations, we can conclude that the multimodal representation H_m outperforms the results obtained by (H_v, H_t) representations, but at the same time H_m is less flexible because it always requires the presence of textual and visual data in query and database to work in the correct form.

6.2 Future Work

The future work and other research directions are described in this section.

6.2.1 Budget selection strategies

The budget has a huge impact on the effectiveness and efficiency of multimodal information retrieval systems. Currently, the size of the budget is selected using a random search process and the composition of the budget is chosen in a random way. These strategies do not follow any systematic approach to find the best budget size and the best elements for budget construction.

In large scale problems, there is a redundancy assumption that states that the information in a dataset can be characterized and represented using a small portion of the total data. Following this assumption, the proposed research direction is to design a systematic approach to select elements for the budget in such a way that the selected elements preserve dataset information and structure without redundancy thus producing a representative budget with the smallest size.

6.2.2 Deep learning integration

This model incorporates some elements of deep learning elements like layer organization but it is not a deep learning model, to take advantage of deep learning models, structure and its ability to learn composed abstract representations from low-level features. A future research direction can be oriented to unify this model with deep learning models and produce an end to end model that incorporates deep learning representation learning ability, non-linear kernel, and latent semantic embedding strategies to perform a multimodal latent semantic spaces.

6.2.3 Semi-supervised non-linear latent semantic embedding

Algorithms presented in this thesis MKLSR and MKSE are supervised methods. These algorithms can be extended to learn in a semi-supervised setup. The semi-supervised learning approach provides to the algorithm with greater flexibility and capacity to learn from the label and unlabeled data. In real work applications, it is common to find partially labeled datasets.

Bibliography

[1] Abadi, Martín: *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* 2015. – Software available from tensorflow.org

[2] Arevalo, John ; Solorio, Thamar ; Montes-y Gómez, Manuel ; González, Fabio A.: Gated multimodal units for information fusion. In: *arXiv preprint arXiv:1702.01992* (2017)

[3] Arora, Sanjeev ; Ge, Rong ; Kannan, Ravi ; Moitra, Ankur: Computing a nonnegative matrix factorization—Provably. In: *SIAM Journal on Computing* 45 (2016), Nr. 4, S. 1582–1611

[4] In: Beitzel, Steven M. ; Jensen, Eric C. ; Frieder, Ophir: *GMAP.* Boston, MA : Springer US, 2009, S. 1256. – ISBN 978–0–387–39940–9

[5] In: Benesty, Jacob ; Chen, Jingdong ; Huang, Yiteng ; Cohen, Israel: *Pearson Correlation Coefficient.* Berlin, Heidelberg : Springer Berlin Heidelberg, 2009, S. 1–4. – ISBN 978–3–642–00296–0

[6] Bottou, Léon: Large-scale machine learning with stochastic gradient descent. In: *Proceedings of COMPSTAT'2010.* Springer, 2010, S. 177–186

[7] Bottou, Léon ; Cun, Yann L.: Large scale online learning. In: *Advances in neural information processing systems*, 2004, S. 217–224

[8] Bottou, Léon ; Le Cun, Yann: On-line learning for very large data sets. In: *Applied stochastic models in business and industry* 21 (2005), Nr. 2, S. 137–151

[9] Bottou, Léon ; Murata, Noboru: Stochastic approximations and efficient learning. In: *The Handbook of Brain Theory and Neural Networks, Second edition,. The MIT Press, Cambridge, MA* (2002)

[10] Bozzon, Alessandro ; Fraternali, Piero: Multimedia and multimodal information retrieval. In: *Search Computing.* Springer, 2010, S. 135–155

[11] Bruno, Eric ; Marchand-Maillet, Stephane: Multiview clustering: a late fusion approach using latent models. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, S. 736–737

[12] Caicedo, Juan C. ; González, Fabio A.: Online Matrix Factorization for Multimodal Image Retrieval. (2012), S. 340–347

[13] Cavallanti, Giovanni ; Cesa-Bianchi, Nicolo ; Gentile, Claudio: Tracking the best hyperplane with a simple budget perceptron. In: *Machine Learning* 69 (2007), Nr. 2-3, S. 143–167

[14] Chitta, Radha ; Jin, Rong ; Jain, Anil K.: Efficient kernel clustering using random fourier features. In: *2012 IEEE 12th International Conference on Data Mining* IEEE, 2012, S. 161–170

[15] Contreras, Victor H. ; Lara, Juan S. ; Perdomo, Oscar J. ; González, Fabio A.: Supervised online matrix factorization for histopathological multimodal retrieval. In: *14th International Symposium on Medical Information Processing and Analysis* Bd. 10975 International Society for Optics and Photonics, 2018, S. 109750Y

[16] Costa Pereira, Jose ; Coviello, Emanuele: On the role of correlation and abstraction in cross-modal multimedia retrieval. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (2014), Nr. 3, S. 521–535. – ISBN 0162–8828 VO – 36

[17] De Lathauwer, Lieven ; De Moor, Bart ; Vandewalle, Joos: A multilinear singular value decomposition. In: *SIAM journal on Matrix Analysis and Applications* 21 (2000), Nr. 4, S. 1253–1278

[18] Deerwester, Scott ; Dumais, Susan T. ; Furnas, George W. ; Landauer, Thomas K. ; Harshman, Richard: Indexing by latent semantic analysis. In: *Journal of the American society for information science* 41 (1990), Nr. 6, S. 391–407

[19] Dekel, Ofer ; Shalev-Shwartz, Shai ; Singer, Yoram: The Forgetron: A kernel-based perceptron on a fixed budget. In: *Advances in neural information processing systems*, 2006, S. 259–266

[20] Depeursinge, Adrien ; Müller, Henning: Fusion techniques for combining textual and visual information retrieval. In: *ImageCLEF*. Springer, 2010, S. 95–114

[21] Ebert, Sandra ; Fritz, Mario ; Schiele, Bernt: Semi-supervised learning on a budget: scaling up to large datasets. In: *Asian Conference on Computer Vision* Springer, 2012, S. 232–245

[22] Feng, Fangxiang ; Li, Ruifan ; Wang, Xiaojie: Deep correspondence restricted Boltzmann machine for cross-modal retrieval. In: *Neurocomputing* 154 (2015), S. 50–60

[23] Feng, Fangxiang ; Wang, Xiaojie ; Li, Ruifan: Cross-modal retrieval with correspondence autoencoder. In: *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, S. 7–16

[24] Fowlkes, Charless ; Belongie, Serge ; Chung, Fan ; Malik, Jitendra: Spectral grouping using the Nystrom method. In: *IEEE transactions on pattern analysis and machine intelligence* 26 (2004), Nr. 2, S. 214–225

[25] Fowlkes, Charless ; Belongie, Serge ; Malik, Jitendra: Efficient spatiotemporal grouping using the nystrom method. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* Bd. 1 IEEE, 2001, S. I–I

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

[27] Gillis, Nicolas: Introduction to nonnegative matrix factorization. In: *arXiv preprint arXiv:1703.00663* (2017)

[28] González, Fabio A. ; Caicedo, Juan C. ; Nasraoui, Olfa ; Ben-Abdallah, Jaafar: NMF-based multimodal image indexing for querying by visual example. In: *CIVR* ACM, 2010, S. 366–373

[29] Gower, J C.: Properties of Euclidean and non-Euclidean distance matrices. In: *Linear Algebra and its Applications* 67 (1985), S. 81–97. – ISSN 0024–3795

[30] Gulli, Antonio ; Pal, Sujit: *Deep Learning with Keras*. Packt Publishing Ltd, 2017

[31] Gunes, Hatice ; Piccardi, Massimo: Affect recognition from face and body: early fusion vs. late fusion. In: *2005 IEEE international conference on systems, man and cybernetics* Bd. 4 IEEE, 2005, S. 3437–3443

[32] Gupta, Amarnath ; Jain, Ramesh: Visual information retrieval. In: *Communications of the ACM* 40 (1997), Nr. 5, S. 70–78

[33] Han, Jiawei ; Kamber, Micheline ; Pei, Jian: 2 - Getting to Know Your Data. In: Han, Jiawei (Hrsg.) ; Kamber, Micheline (Hrsg.) ; Pei, Jian (Hrsg.): *Data Mining (Third Edition)*. Third Edit. Boston : Morgan Kaufmann, 2012 (The Morgan Kaufmann Series in Data Management Systems). – ISBN 978–0–12–381479–1, S. 39–82

[34] He, Jianfeng ; Ma, Bingpeng ; Wang, Shuhui ; Liu, Yugui ; Huang, Qingming: Cross-modal Retrieval by Real Label Partial Least Squares. In: *Proceedings of the 2016 ACM on Multimedia Conference* ACM, 2016, S. 227–231

[35] Hernando, Antonio ; Bobadilla, Jesús ; Ortega, Fernando: A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model. In: *Knowledge-Based Systems* 97 (2016), S. 188–202

[36] Hofmann, Thomas: Probabilistic latent semantic analysis. In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* Morgan Kaufmann Publishers Inc., 1999, S. 289–296

[37] Hofmann, Thomas: Unsupervised learning by probabilistic latent semantic analysis. In: *Machine learning* 42 (2001), Nr. 1-2, S. 177–196

[38] Hoyer, Patrik O.: Non-negative matrix factorization with sparseness constraints. In: *Journal of machine learning research* 5 (2004), Nr. Nov, S. 1457–1469

[39] Johnson, Rie ; Zhang, Tong: Accelerating stochastic gradient descent using predictive variance reduction. In: *Advances in neural information processing systems*, 2013, S. 315–323

[40] Kludas, Jana ; Bruno, Eric ; Marchand-Maillet, Stephane: Information fusion in multimedia information retrieval. In: *International Workshop on Adaptive Multimedia Retrieval* Springer, 2007, S. 147–159

[41] Kolda, Tamara G. ; O'leary, Dianne P.: A semidiscrete matrix decomposition for latent semantic indexing information retrieval. In: *ACM Transactions on Information Systems (TOIS)* 16 (1998), Nr. 4, S. 322–346

[42] Korenius, Tuomo ; Laurikkala, Jorma ; Juhola, Martti: On principal component analysis, cosine and Euclidean measures in information retrieval. In: *Information Sciences* 177 (2007), Nr. 22, S. 4893–4905

[43] Kumar, Sanjiv ; Mohri, Mehryar ; Talwalkar, Ameet: Ensemble nystrom method. In: *Advances in Neural Information Processing Systems*, 2009, S. 1060–1068

[44] Lahat, Dana ; Adali, Tülay ; Jutten, Christian: Multimodal data fusion: an overview of methods, challenges, and prospects. In: *Proceedings of the IEEE* 103 (2015), Nr. 9, S. 1449–1477

[45] Landauer, Thomas K. ; Foltz, Peter W. ; Laham, Darrell: An introduction to latent semantic analysis. In: *Discourse processes* 25 (1998), Nr. 2-3, S. 259–284

[46] Lau, Jey H. ; Baldwin, Timothy: An empirical evaluation of doc2vec with practical insights into document embedding generation. In: *arXiv preprint arXiv:1607.05368* (2016)

[47] Lee, Daniel D. ; Seung, H S.: Learning the parts of objects by non-negative matrix factorization. In: *Nature* 401 (1999), Nr. 6755, S. 788–791

[48] Lee, Daniel D. ; Seung, H S.: Algorithms for non-negative matrix factorization. In: *Advances in neural information processing systems*, 2001, S. 556–562

[49] Li, Chao ; Deng, Cheng ; Li, Ning ; Liu, Wei ; Gao, Xinbo ; Tao, Dacheng: Self-supervised adversarial hashing networks for cross-modal retrieval. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, S. 4242–4251

[50] Li, Zhongyu ; Zhang, Xiaofan ; Müller, Henning ; Zhang, Shaoting: Large-scale retrieval for medical image analytics: A comprehensive review. In: *Medical image analysis* 43 (2018), S. 66–84

[51] Liu, Dong ; Lai, Kuan-Ting ; Ye, Guangnan ; Chen, Ming-Syan ; Chang, Shih-Fu: Sample-specific late fusion for visual category recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, S. 803–810

[52] Liu, Wei ; Wang, Jun ; Ji, Rongrong ; Jiang, Yu-Gang ; Chang, Shih-Fu: Supervised hashing with kernels. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* IEEE, 2012, S. 2074–2081

[53] Liu, Wei ; Wang, Jun ; Kumar, Sanjiv ; Chang, Shih-Fu: Hashing with graphs. In: *Proceedings of the 28th international conference on machine learning (ICML-11)* Citeseer, 2011, S. 1–8

[54] Liu, Xinwang ; Zhu, Xinzhong ; Li, Miaomiao ; Wang, Lei ; Tang, Chang ; Yin, Jianping ; Shen, Dinggang ; Wang, Huaimin ; Gao, Wen: Late fusion incomplete multi-view clustering. In: *IEEE transactions on pattern analysis and machine intelligence* 41 (2018), Nr. 10, S. 2410–2423

[55] Ma, Lei ; Li, Hongliang ; Meng, Fanman ; Wu, Qingbo ; Ngan, King N.: Global and local semantics-preserving based deep hashing for cross-modal retrieval. In: *Neurocomputing* 312 (2018), S. 49–62

[56] Manning, Christopher ; Raghavan, Prabhakar ; Schütze, Hinrich: Introduction to information retrieval. In: *Natural Language Engineering* 16 (2010), Nr. 1, S. 100–103

[57] Manning, Christopher D. ; Raghavan, Prabhakar ; Schütze, Hinrich: Ch. 1 - Boolean retrieval. In: *Introduction to Information Retrieval* (2009), Nr. c, S. 1–18. – ISBN 0521865719

[58] Morvant, Emilie ; Habrard, Amaury ; Ayache, Stéphane: Majority vote of diverse classifiers for late fusion. In: *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* Springer, 2014, S. 153–162

[59] Mourão, André ; Martins, Flávio ; Magalhães, João: Multimodal medical information retrieval with unsupervised rank fusion. In: *Computerized Medical Imaging and Graphics* 39 (2015), S. 35–45

[60] Müller, Henning ; Michoux, Nicolas ; Bandon, David ; Geissbuhler, Antoine: A review of content-based image retrieval systems in medical applications - Clinical benefits and future directions. In: *International Journal of Medical Informatics* 73 (2004), Nr. 1, S. 1–23. – ISBN 1386–5056

[61] Müller, Henning ; Michoux, Nicolas ; Bandon, David ; Geissbuhler, Antoine: A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. In: *International journal of medical informatics* 73 (2004), Nr. 1, S. 1–23

[62] Niblack, Carlton W. ; Barber, Ron ; Equitz, Will ; Flickner, Myron D. ; Glasman, Eduardo H. ; Petkovic, Dragutin ; Yanker, Peter ; Faloutsos, Christos ; Taubin, Gabriel: QBIC project: querying images by content, using color, texture, and shape. In: *Storage and retrieval for image and video databases* Bd. 1908 International Society for Optics and Photonics, 1993, S. 173–187

[63] Peng, Xiaojiang ; Wang, Limin ; Wang, Xingxing ; Qiao, Yu: Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. In: *Computer Vision and Image Understanding* 150 (2016), S. 109–125

[64] Peng, Yuxin ; Huang, Xin ; Qi, Jinwei: Cross-media shared representation by hierarchical learning with multiple deep networks. In: *IJCAI*, 2016, S. 3846–3853

[65] Pereira, Jose C. ; Coviello, Emanuele ; Doyle, Gabriel ; Rasiwasia, Nikhil ; Lanckriet, Gert R. ; Levy, Roger ; Vasconcelos, Nuno: On the role of correlation and abstraction in cross-modal multimedia retrieval. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (2014), Nr. 3, S. 521–535

[66] Rahimi, Ali ; Recht, Benjamin: Random features for large-scale kernel machines. In: *Advances in neural information processing systems*, 2008, S. 1177–1184

[67] Rasiwasia, Nikhil ; Costa Pereira, Jose ; Coviello, Emanuele ; Doyle, Gabriel: A new approach to cross-modal multimedia retrieval. In: *Mm* (2010), S. 251–260. ISBN 9781605589336

[68] Rastegar, Sarah ; Soleymani, Mahdieh ; Rabiee, Hamid R. ; Mohsen Shojaee, Seyed: Mdl-cw: A multimodal deep learning framework with cross weights. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, S. 2601–2609

[69] Schölkopf, Bernhard ; Burges, Christopher J. ; Smola, Alexander J. [u. a.]: *Advances in kernel methods: support vector learning*. MIT press, 1999

[70] Shalev-Shwartz, Shai [u. a.]: Online learning and online convex optimization. In: *Foundations and Trends® in Machine Learning* 4 (2012), Nr. 2, S. 107–194

[71] Shao, Jie ; Wang, Leiquan ; Zhao, Zhicheng ; Cai, Anni [u. a.]: Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval. In: *Neurocomputing* 214 (2016), S. 618–628

[72] Shawe-Taylor, John ; Cristianini, Nello [u. a.]: *Kernel methods for pattern analysis.* Cambridge university press, 2004

[73] Singhal, Amit [u. a.]: Modern information retrieval: A brief overview. In: *IEEE Data Eng. Bull.* 24 (2001), Nr. 4, S. 35–43

[74] Snoek, Cees G. ; Worring, Marcel ; Smeulders, Arnold W.: Early versus late fusion in semantic video analysis. In: *Proceedings of the 13th annual ACM international conference on Multimedia* ACM, 2005, S. 399–402

[75] Srivastava, Nitish ; Salakhutdinov, Ruslan: Learning representations for multimodal data with deep belief nets. In: *International conference on machine learning workshop* Bd. 79, 2012

[76] Srivastava, Nitish ; Salakhutdinov, Russ R.: Multimodal learning with deep boltzmann machines. In: *Advances in neural information processing systems*, 2012, S. 2222–2230

[77] del Toro, Oscar J. ; Atzori, Manfredo ; Otálora, Sebastian ; Andersson, Mats ; Eurén, Kristian ; Hedlund, Martin ; Rönnquist, Peter ; Müller, Henning: Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score. In: *Medical Imaging 2017: Digital Pathology* Bd. 10140 International Society for Optics and Photonics, 2017, S. 101400O

[78] Jimenez-del Toro, Oscar ; Otálora, Sebastian ; Andersson, Mats ; Eurén, Kristian ; Hedlund, Martin ; Rousson, Mikael ; Müller, Henning ; Atzori, Manfredo: Analysis of histopathology images: From traditional machine learning to deep learning. In: *Biomedical Texture Analysis.* Elsevier, 2018, S. 281–314

[79] Jimenez-del Toro, Oscar ; Otálora, Sebastian ; Atzori, Manfredo ; Müller, Henning: Deep Multimodal Case–Based Retrieval forLarge Histopathology Datasets. In: Wu, Guorong (Hrsg.) ; Munsell, Brent C. (Hrsg.) ; Zhan, Yiqiang (Hrsg.) ; Bai, Wenjia (Hrsg.) ; Sanroma, Gerard (Hrsg.) ; Coupé, Pierrick (Hrsg.): *Patch-Based Techniques in Medical Imaging.* Cham : Springer International Publishing, 2017. – ISBN 978–3–319–67434–6, S. 149–157

[80] TREC: trec eval Evaluation Report. – Forschungsbericht

[81] Vanegas, Jorge A.: *Large-scale non-linear multimodal semantic embedding.* Junio 2018. – Doctor en Ingeniería. Línea de investigación: Ciencias de la computación.

[82] Vanegas, Jorge A. ; Escalante, Hugo J. ; González, Fabio A.: Semi-supervised Online Kernel Semantic Embedding for Multi-label Annotation. In: Mendoza, Marcelo (Hrsg.) ; Velastín, Sergio (Hrsg.): *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications.* Cham : Springer International Publishing, 2018. – ISBN 978–3–319–75193–1, S. 693–701

[83] Vavasis, Stephen A.: On the complexity of nonnegative matrix factorization. In: *SIAM Journal on Optimization* 20 (2010), Nr. 3, S. 1364–1377

[84] Wang, Daixin ; Cui, Peng ; Ou, Mingdong ; Zhu, Wenwu: Learning compact hash codes for multimodal representations using orthogonal deep structure. In: *IEEE Transactions on Multimedia* 17 (2015), Nr. 9, S. 1404–1416

[85] Wang, Zhuang ; Vucetic, Slobodan: Twin vector machines for online learning on a budget. In: *Proceedings of the 2009 SIAM International Conference on Data Mining* SIAM, 2009, S. 906–917

[86] Wu, Lin ; Wang, Yang ; Shao, Ling: Cycle-consistent deep generative hashing for cross-modal retrieval. In: *IEEE Transactions on Image Processing* 28 (2018), Nr. 4, S. 1602–1612

[87] Xu, Wei ; Liu, Xin ; Gong, Yihong: Document clustering based on non-negative matrix factorization. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* ACM, 2003, S. 267–273

[88] Yang, Tianbao ; Li, Yu-Feng ; Mahdavi, Mehrdad ; Jin, Rong ; Zhou, Zhi-Hua: Nyström method vs random fourier features: A theoretical and empirical comparison. In: *Advances in neural information processing systems*, 2012, S. 476–484

[89] Ye, Guangnan ; Liu, Dong ; Jhuo, I-Hong ; Chang, Shih-Fu: Robust late fusion with rank minimization. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition* IEEE, 2012, S. 3021–3028

[90] In: Zhang, Ethan ; Zhang, Yi: *Average Precision.* Boston, MA : Springer US, 2009, S. 192–193. – ISBN 978–0–387–39940–9

[91] Zhang, Jian ; Peng, Yuxin ; Yuan, Mingkuan: Unsupervised generative adversarial cross-modal hashing. In: *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018

[92] Zhang, Xiaofan ; Dou, Hang ; Ju, Tao ; Xu, Jun ; Zhang, Shaoting: Fusing heterogeneous features from stacked sparse autoencoder for histopathological image analysis. In: *IEEE journal of biomedical and health informatics* 20 (2016), Nr. 5, S. 1377–1383

[93] Zhang, Zhong ; Qin, Zhili ; Li, Peiyan ; Yang, Qinli ; Shao, Junming: Multi-view Discriminative Learning via Joint Non-negative Matrix Factorization. In: *International Conference on Database Systems for Advanced Applications* Springer, 2018, S. 542–557

[94] Zheng, Liang ; Wang, Shengjin ; Tian, Lu ; He, Fei ; Liu, Ziqiong ; Tian, Qi: Query-adaptive late fusion for image search and person re-identification. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, S. 1741–1750

[95] Zong, Linlin ; Zhang, Xianchao ; Zhao, Long ; Yu, Hong ; Zhao, Qianli: Multi-view clustering via multi-manifold regularized non-negative matrix factorization. In: *Neural Networks* 88 (2017), S. 74–89