Universitat Jaume I

Escola Superior de Ciències i Tecnologia

Characterisation and adaptive learning in interactive video retrieval

Thesis presented by Rubén Fernández Beltrán
Supervised by Filiberto Pla Bañón
to obtain Doctor of Philosophy

4/2016

Departament de Llenguatges i Sistemes Informàtics

Abstract

Retrieving videos by content is a very challenging task because it involves a wide variety of fields. From low-level video descriptors to high-level visual understanding, Content-based Video Retrieval (CBVR) systems have to fill a huge semantic gap to provide users with those videos which satisfy their queries. Even though some of the state-of-the-art approaches have shown to be successful on reduced databases, the ongoing expansion of video collections demands new capabilities in CBVR. Retrieval systems are required to be more efficient to deal with this increasing amount of samples and more effective to cope with more complex query concepts. In this thesis, we explore how difficult this task is and how our contributions try to improve the current state-of-the-art.

In this work, we are interested in the use of latent topics to overcome the current limitations in CBVR. Despite the potential of topic models to uncover the hidden structure of a collection, they have traditionally been unable to provide a competitive advantage in CBVR because of the high computational cost of their algorithms and the complexity of the latent space in the visual domain. Throughout this thesis we focus on designing new models and tools based on topic models to take advantage of the latent space in CBVR. Specifically, we have worked in four different areas within the retrieval process: vocabulary reduction, encoding, modelling and ranking, being our most important contributions related to both modelling and ranking.

Initially, we present a novel approach to vocabulary reduction based on latent topics in order show how topic models are able to capture the more relevant words of a collection. Subsequently, a new encoding approach specially designed to Content-Based Retrieval tasks is proposed. In the modelling stage, we study how the use of different topic models affects video retrieval performance and present an incremental topic model to cope with incremental scenarios in an effective and efficient way. Regarding the ranking stage, we propose a new proba-

bilistic ranking function which is deduced from a supervised topic model to tackle the semantic gap between low-level features and high-level concepts through the patterns defined by topics. Finally, we conclude the work with observations on how this investigation has impacted the use of topic models in CBVR.

Acknowledgments

I would like to start by thanking my thesis supervisor Prof. Filiberto Pla for having given me the opportunity to reach this goal and, of course, for having shared with me his brilliant knowledge and ideas. Although obvious, I find it essential to remark that this thesis would have not been possible without him.

I would also like to thank all the people who have helped me to shape this work during these years. First, thanks to my colleagues at University Jaume I for your useful comments and suggestions. Thanks as well to my colleagues at Bristol University where I had a short but fruitful research stay. Last but not least, I would like to thank my family and friends for their unconditional love and support.

Finantial Support

This thesis has been funded by FPU-AP-2009-4435 from the Spanish Ministry of Education and partially supported by PROMETEO/2010/028 and PROMETEOII/2014/062 projects from Generalitat Valenciana, P1-1B2010-27 project from the Plan de Promoció de la Investigació UJI and ESP2013-48458-C4-3-P project from the Spanish Ministry of Economy and Competitiveness.

Sinopsis de la Tesis

Este capítulo tiene como objeto cumplir con la normativa de los estudios de doctorado regulados por el RD 99/2011 de la Universistat Jaume I, que establece los criterios necesarios para obtener la mención internacional en el título de Doctor. En este capítulo se proporcionará una visión global de la tesis en español así como su motivación, objetivos, contribuciones y conclusiones.

Introducción

La recuperación automática de vídeo a partir de su contenido consiste en la búsqueda automática de vídeos mediante el análisis de su propio contenido. En este tipo de sistemas, el usuario proporciona inicialmente una consulta, es decir, uno o más ejemplos del tipo de vídeo que pretende extraer de una determinada base de datos, y el sistema obtiene como salida aquellos vídeos de la base de datos que se corresponden con el concepto semántico asociado a dicha consulta.

Motivación

Ante la gran expansión que están experimentando las colecciones multimedia, existen todo tipo de aplicaciones donde la recuperación automática de vídeo resulta de gran utilidad. Por ejemplo, la ayuda al diagnóstco médico, la gestión de catálogos multimedia o incluso la prevención de delitos son algunas de las aplicaciones en las que recuperar videos automáticamente a través de su contenido puede resultar de gran ayuda. Sin embargo, esta tarea de recuperación implica la necesidad de tratar con uno de los sistemas más complejos que conocemos, el sistema de la comprensión visual humana. Desde el punto de vista de un computador, un vídeo no es más que un conjunto de píxeles en un determinado orden, sin embargo existe una gran diferencia entre el valor numérico de dichos píxeles

y el significado semántico que tienen para nosotros. Esta diferencia se conoce con el nombre de laguna semántica y es el principal problema que deben afrontar los sistemas de recuperación. Además, las bases de datos de vídeos son cada vez más grandes y más complejas cosa que agrava el problema y requiere que los sistemas de recuperación sean cada vez más eficientes y tengan mayor poder de generalización.

Objetivos

El objetivo principal de esta tesis consiste en utilizar eficazmente los modelos de tópicos latentes para afrontar el problema de la recuperación automática de vídeo. Concretamente, se pretende mejorar tanto a nivel de eficiencia como a nivel de precisión el actual estado del arte en materia de los sitemas de recuperación automática de vídeo. De una forma muy resumida, los modelos de tópicos latentes son un conjunto de herramientas estadísticas que permiten extraer los patrones generadores de una colección de datos. Tradicionalmente, este tipo de técnicas no han sido consideradas de gran utilidad para los sistemas de recuperación automática de vídeo debido a su alto coste computacional y a la propia complejidad del espacio de tópicos en el ámbito de la información visual.

A continuación, pasamos a listar los objetivos de la tesis en relación a la eficiencia y eficacia de los sistemas de recuperación:

Eficiencia: mejoras en el tiempo de ejecución

- Utilizar los tópicos latentes como herramienta de síntesis de datos.
- Mejorar la eficiencia del proceso de extración de tópicos.
- Mejorar la eficiencia del proceso de recuperación de vídeos.

Efectividad: mejoras en el la precisión

- Utilizar los tópicos latentes como una representación de datos de alto nivel.
- Aportar un nuevo enfoque para abordar el problema de la recuperación de vídeo desde el punto de vista de los tópicos latentes.

- Desarrollar una nueva función de recuperación especialmente diseñada para utilizar tópicos latentes y capaz de obtener una mejora en la precisión del proceso de recuperación.

Contribuciones

Las contribuciones de esta tesis se centran en la mejora del proceso de recuperación mediante el uso de los modelos de tópicos latentes. Concretamente, podemos destacar los siguientes puntos como aportaciones de nuestro trabajo:

- Mecanismo para reducir el numero de palabras de una colección mediante el uso de tópicos latentes.
- Nueva función de codificación basada en tópicos latentes especialmente diseñada para combatir la laguna semántica.
- Evaluación del rendimiento de diferentes modelos de tópicos para el problema concreto de la recuperación automática de vídeo.
- Nuevo modelo incremental de tópicos diseñado para mejorar la eficiencia del proceso de extracción de tópicos en un esquema de recuperación dinámico.
- Nuevo modelo supervisado de tópicos para afrontar el problema de recuperación como un problema de descubrimiento de clase mediante tópicos latentes.
- Nueva función de recuperación capaz de mejorar los sistemas de actuales tanto el tiempo de ejecución como en precisión del proceso de recuperación.

Conclusiones y Trabajo Futuro

La principal conclusión que podemos extraer del trabajo desarrollado es la importancia de los modelos de tópicos latentes para afrontar el problema de la laguna semántica en la recuperación automática de vídeo. Concretamente, este tipo de modelos son capaces de aportar una representación de los datos de más alto nivel que resulta de gran utilidad cuando la estructura de datos es inicialmente desconocida o cuando tenemos muy poca información acerca del objetivo. Precisamente, ésto es lo que ocurre en la recuperación automática de vídeo donde

normalmente tenemos que manejar conceptos semánticos complejos conociendo un número ejemplos positivos muy reducido.

A lo largo del trabajo, hemos podido comprobar cómo los tópicos latentes son útiles a diferentes niveles en el proceso de recuperación, desde la codificación de las muestras, hasta el modelado y la recuperación de vídeos. No obstante, conviene resaltar algunas de las limitaciones que presentan. La primera de ellas es su alto coste computacional que limita su aplicación en colecciones de vídeos con millones de muestras. Otra de sus limitaciones consiste en el hecho que requieren de la existencia de cierta brecha semántica, es decir, en aquellos casos en los que se pretende recuperar una propiedad bien caracterizada en el espacio de representación inicial los tópicos latentes no son capaces de proporcionar una mejora real en el sistema.

Finalmente, destacar algunos de los puntos que podrían extender el trabajo desarrolado en esta tesis:

- Desarrolar estrategias automáticas para escoger el número ideal de palabras y tópicos en una colección de datos.
- Utilizar técnicas de cuantización para poder llevar a cabo la tarea de extracción de tópicos de manera más eficiente.
- Extensión del modelo de recuperación propuesto mediante un enfoque a largo plazo así como con la inclusión de diferentes modalidades de datos.

Contents

Chapter 1

Introduction

Imagine you are searching through your family video collection in order to extract those video shots in which your grandparents appear. A Content-Based Video Retrieval (CBVR) system may help you to do it automatically just from a single picture of them. Now, imagine your TV was able to learn the kind of programs you like and was able to automatically suggest new TV shows with a similar content. Or even, it was able to summarize a program by showing only your favourite parts. For sure, all these functionalities will be available sooner or later thanks to CBVR research. Searching and retrieving relevant videos according to users' queries is one of the most popular fields in multimedia research as well as in real life applications. Aided searches over huge video collections are also possible via CBVR. You can start exploring a collection looking for general sport videos, then you focus on cycling and finally you end up focusing on videos of a certain cyclist just with a few mouse clicks without introducing any text term. There are plenty of scenarios in which CBVR may help. Multimedia catalogues, crime prevention, copyrights violation, medical diagnosis and many others applications make the CBVR research profitable to industry and government as well. The wide demand for such applications together with the potential of CBVR to cope with new challenges motivate the video retrieval research and therefore the work developed in this thesis.

Retrieving videos by content is a very challenging task because it involves dealing with one of the most complex processes, the human visual understanding. Even the most technologically advanced machine struggles at the task of making sense of what it sees. For a computer, a video shot is just a collection of pixels in a specific order but for us it is usually a scene full of meaning. A little child playing

football with his dad, a happy couple being on vacations or a man doing sport in the park are concepts easily understandable for us but not for a computer. There is still a huge gap between pixel values and the meaning they represent to us. Precisely, this issue is known by the research community as the semantic-gap challenge and filling this gap is the key to bring the computer vision technology towards the highest level of visual understanding. In this thesis, we explore how difficult this task is and how our contributions try to improve the current state of the art in the context of video retrieval.

1.1 Content-Based Video Retrieval

In general, Content-Based Video Retrieval (CBVR) is concerned about providing users with those videos which satisfy their queries by means of the video content analysis. The standard CBVR procedure involves three main components: (i) a query, containing a few video examples of the semantic concept that the user is looking for; (ii) a database, which is used to retrieve videos related to the query concept; and (iii) a ranking function, which sorts the database according to the relevance with respect to the user's query.

These three components are typically integrated with the user in a Relevance Feedback (RF) scheme to provide the most relevant videos through several feedback iterations. Fig. 1.1 shows the general RF retrieval scheme. At the initialisation stage (stage 0), the user introduces the query concept into the system by providing Q examples of the concept of interest. Then, the interactive process consists of the alternation of two stages through I feedback iterations. In the retrieval stage (stage 1), the system ranks the database according to the query and shows the S top items (scope) to the user. In the feedback stage (stage 2), the user checks the scope to select the correctly retrieved samples and finally the query is expanded with these new positive examples to carry out the next iteration.

The ranking function can be considered the kernel of the retrieval system because it is in charge of scoring the samples of the database according to the query. As a result, the nature of the ranking function and the nature of the video representation space where the ranking function works are two of the most important factors for the precision of a CBVR system.

In the literature, as we will see in following chapters, a wide variety of CBVR

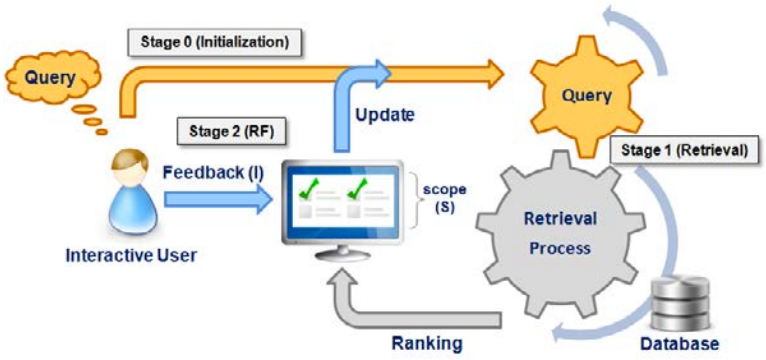


Figure 1.1: Relevance Feedback scheme. Q is the number of initial examples in the query, I the number of feedback iterations and S the number of top ranked samples.

systems have been presented using different kinds of ranking functions and video descriptors. Even though several of those approaches have shown to be successful when they are used on reduced databases with a small number of concepts, the huge expansion of video collections demands new capabilities in CBVR. Multimedia databases are getting bigger and more complex, as a result retrieval systems are required to be more efficient to deal with this increasing amount of samples and more effective to cope with more complex query concepts. In a sense, the visual variability of semantic concepts can be so high that current approaches are often not able to properly capture unconstrained queries over extensive collections.

1.2 Objectives and contributions of this Thesis

The general objective of this thesis is based on developing new retrieval models and tools to improve the current state-of-the-art in both efficiency and effectiveness aspects along the retrieval process. Specifically, we are interested in the use of latent topics to overcome the aforementioned limitations in CBVR.

Over the following chapters, more details about topic models will be provided where necessary but now let us introduce the general concept and motivation. Briefly, latent topics [9] are a kind of statistical models which provide methods to automatically understand and summarize large data collections by means of their hidden patterns. Despite the fact that these models have been successfully used within several fields, including text [13, 70], image [12, 49] and even video domain [73], their application to CBVR has been traditionally rather limited.

Although this thesis is focused on the video domain, some experiments have also been performed using text or image collections for initial validation of the topic models proposed and as illustrative examples of the applicability of the proposed techniques to other data that can be modelled as "documents of words".

In the literature, it is possible to find some attempts to use topic models in video retrieval [59, 60] but in all those works topic models are used just as a characterisation method over traditional classification-based retrieval frameworks what eventually makes topics unable to provide a competitive advantage because of the special nature of the latent space. The latent topic space tends to relate documents according to the patterns of the collection, therefore many of classic strategies, which assume that similar things have to be close in the representation space, do not work properly in the latent space. That is, traditional retrieval engines do not take into account topics' nature and precisely this fact limits the actual potential of topics to outperform the current state-of-the-art in CBVR. Even some works [6] advise against the general use of latent topics in retrieval tasks without considering this fact.

The main problem when retrieving samples by content is the semantic gap between low-level features and high-level concepts, and topics can be helpful to reduce this gap, especially in CBVR where the semantic gap is particularly important due to the higher complexity of the dynamic visual domain. In a sense, the hidden structure uncovered by topics represents a higher characterization level where samples are described according to their feature patterns instead of their low level features. As a result, this space enables semantic connections among different concepts through the patterns defined by topics and these relations can be very useful when dealing with a huge semantic gap like in the case of CBVR.

However, the aforementioned complex nature of the latent space together with the high computational burden of their algorithms make difficult the application of topics from a classic retrieval point of view and demand new retrieval models and tools to effectively take advantage of latent topics in CBVR. Precisely, the research developed in this thesis pursues a two-fold objective to improve the current state-of-the-art via topic models: (i) efficiency improvement in order to reduce the computational time along the retrieval process and (ii) effectiveness improvement in order to increase the final retrieval precision. The following two sections summarise the thesis objectives and their correspondent achievements focused on these two aspects.

Efficiency: improving computational time

1. **Topic models as a data summarising tool**: The first objective of this thesis is to show the utility of topic models to summarise data collections. Specifically, a new vocabulary reduction approach is proposed to highlight how topics are able to capture the most relevant words in a collection.

2. **Efficient topic extraction processes for CBVR**: The high computational cost of standard topic models hinders their application in large-scale scenarios like in the case of CBVR. Because of this, the second objective is to develop more efficient models specially designed to CBVR. In particular, a new topic model is presented which is able to reduce the computational burden by taking advantage of the incremental nature of the retrieval problem.

3. **Efficient topic-based ranking functions**: The huge expansion of video collections demands more efficient ranking functions for CBVR. The third objective is to design new ranking functions capable to obtain a competitive advantage in terms of query processing time. In this thesis, a new topic-based ranking function is presented which provides an efficiency improvement with respect to state-of-the-art functions.

Effectiveness: improving retrieval precision

4. **Topic models as a high-level data representation**: Topic models have been widely used in many areas and their potential to uncover hidden information supports the viability of using this kind of models to deal with the semantic gap in CBVR. However, there are not works in the literature studying how the use of different topic models affects video retrieval performance. The fourth objective is to analyse the differences among topic models within CBVR field. Precisely, this thesis presents a study on this line of work.

5. **Latent topics as a different point of view to tackle the retrieval problem**: Many of current approaches address the retrieval problem as a pseudo-classification problem with only two classes, relevant and not-relevant according to the query. However, this approach is often not useful with topics because of the complex nature of the latent space. The fifth

objective is to redefine the retrieval problem to take into account the latent space nature. Specifically, this thesis reformulates the classic retrieval approach into a class discovery problem by topic models to encapsulate the topic nature.

6. **Ranking functions based on probabilistic latent topic models**: The especially important semantic gap in CBVR often leads to a poor retrieval precision and therefore new ranking functions are required to improve the current retrieval performance. The sixth objective is to design a new ranking function following the rationale behind latent topics. In particular, this thesis presents a new topic-based ranking function which is able to to provide a precision improvement with respect to state-of-the-art methods.

1.3 Thesis overview

In section 1.1, we introduced the CBVR problem by showing how an interactive retrieval system works, nonetheless let us now define the retrieval problem from a more general point of view according to the objectives of the thesis. As it is shown in fig. 1.2, a CBVR system could be seen as a procedure which connects videos to users through five steps: (1) feature extraction, (2) encoding, (3) vocabulary reduction, (4) modelling and (5) ranking. Initially, a feature extraction process is performed to extract low level information[1] which is useful to codify video samples. Later, that low level information is encoded in visual words to represent each video sample as a single vector of words. As a third step, a vocabulary reduction process can be performed in order to reduce the computational complexity of subsequent steps. The modelling step consists in applying machine learning algorithms to deal with the semantic gap challenge by bringing the low level video representation to a higher semantic level. Finally, the most relevant videos according to users' queries are retrieved in the ranking step where a ranking function is applied over a specific video database. Note that this ranking step corresponds to the retrieval stage shown in fig. 1.1.

The final target of these five steps is to relate low level video information to high level user query concepts, in other words, how to connect the information provided by feature extraction methods with users' queries. Despite the fact that

[1] For instance, color information, gradient or optical flow.

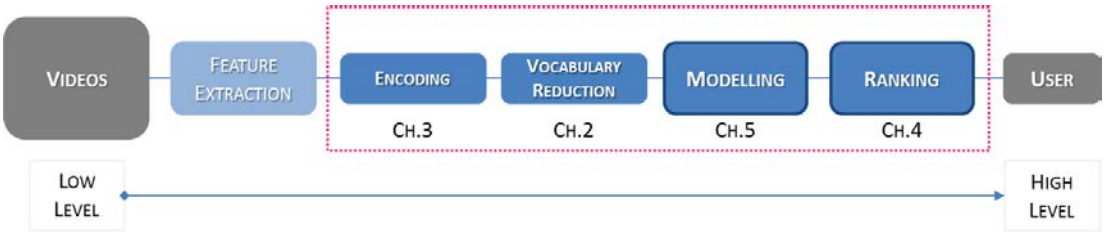


Figure 1.2: Thesis scheme.

feature extraction is an important part of CBVR, in this thesis we focus on the highest part of the scheme in fig. 1.2. Specifically, we are interested in the use of latent topics techniques from encoding to ranking in order to deal with the semantic gap challenge in the CBVR field.

The format of this doctoral dissertation is by compendium of publications, therefore it comprises the compilation of papers developed during the thesis period. From chapter 2 to 5, each unit contains specific published papers as contributions. Finally, we provide the global conclusions of the thesis in chapter 6. Note that chapters from 2 to 5 are self-contained, that is, they first introduce the problem to address while reviewing the state-of-the-art, then propose a methodology, and finally show the experimental validation and conclude the work.

Each chapter of the thesis is related to one specific step in fig. 1.2. That is, chapter 2 includes the work developed in the vocabulary reduction step, chapter 3 related to the encoding step, chapter 4 related to ranking and chapter 5 presents the work carried out regarding the modelling step. Now, let us provide an overview of the content presented in each chapter:

- **Chapter 2: Vocabulary Reduction**
  - This chapter presents a novel approach to vocabulary reduction based on latent topics. In particular, it is based on filtering words in the topic feature space instead of in the original word space. Experiments show how topic models are able to capture the more relevant words of a collection.
  - ⋆ [27] Ruben Fernandez-Beltran, Raul Montoliu, and Filiberto Pla. Vocabulary reduction in bow representing by topic modeling. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 648–655, 2013.

- **Chapter 3: Encoding**
  - In this chapter, a new encoding approach specially designed to Content-Based Retrieval tasks is presented. The novelty of the proposed encoding lies in both using hidden patterns as visual words and encoding the samples by accumulating the proportion of features over topics. Results show that the proposed technique is able to outperform the traditional visual Bag-of-Words when the retrieval task is performed in the latent topic space.
  - ⋆ [24] Ruben Fernandez-Beltran and Filiberto Pla. Latent topic encoding for content-based retrieval. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 640–648, 2015.
- **Chapter 4: Ranking**[2]
  - This chapter presents a novel Content-Based Video Retrieval approach in order to cope with the semantic gap challenge by means of latent topics. Firstly, a supervised topic model is proposed to transform the classic retrieval approach into a class discovery problem. Subsequently, a new probabilistic ranking function is deduced from that model to tackle the semantic gap between low-level features and high-level concepts. Finally, a short-term relevance feedback scheme is defined where queries can be initialised with samples from inside or outside the database. Several retrieval simulations have been carried out using three databases and seven different ranking functions to test the performance of the presented approach. Experiments revealed that the proposed ranking function is able to provide a competitive advantage within the content-based retrieval field.
  - ⋆ [23] Ruben Fernandez-Beltran and Filiberto Pla. An interactive video retrieval approach based on latent topics. In *International Conference on Image Analysis and Processing*, pages 290–299, 2013.
  - ⋆ [26] Ruben Fernandez-Beltran and Filiberto Pla. Latent topics-based relevance feedback for video retrieval. *Pattern Recognition*, 51:72–84, 2016.

[2]This chapter omits the content of the conference paper [23] because it is included in its journal version [26].

- **Chapter 5: Modelling**
  - The work presented in this chapter has a dual target: (1) it is aimed at studying how the use of different topic models (pLSA, LDA and FSTM) affects video retrieval performance; (2) a novel incremental topic model (IpLSA) is presented in order to cope with incremental scenarios in an effective and efficient way. A comprehensive comparison among these four topic models using two different retrieval systems and two reference benchmarking video databases is provided. Experiments revealed that pLSA is the best model in sparse conditions, LDA tend to outperform the rest of the models in a dense space and IpLSA is able to work properly in both cases.
  - ⋆ [25] Ruben Fernandez-Beltran and Filiberto Pla. Incremental probabilistic latent semantic analysis for video retrieval. *Image and Vision Computing*, 38:1–12, 2015.

Chapter 2

Vocabulary Reduction by Topic Modelling

Publication

[27] Ruben Fernandez-Beltran, Raul Montoliu, and Filiberto Pla. Vocabulary reduction in bow representing by topic modeling. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 648–655, 2013.

In this chapter, a new approach to vocabulary reduction is presented. It is based on filtering words in the topic feature space instead of in the original word space. The goal is to analyse the differences between the application of the Cumulative Count-based word filter in the Bag of Words feature space with respect to its application in the topic descriptions obtained by Latent Dirichlet Allocation. Three well-known text datasets (Reuters, WebKB and NewsGroup) have been used to show the performance of the proposed approach.

2.1 Introduction

In recent years the expansion of new technologies has produced a lot of data available for their study. This leads to more and larger data sets. One of the most common representation of this data is the Bag of Words model (BoW) [37, 55] in which a dataset is represented as an unordered collection of word frequencies. Large datasets have often an unmanageable vocabulary size and besides many of these words may be redundant, have no semantic meaning or

induce errors in classification tasks. One possible solution to address this issue is to use vocabulary reduction techniques which consists on reducing the number of words to characterize a specific set of samples. The objective of this reduction is to decrease the vocabulary size keeping the semantic meaning of the documents and allowing good performances in terms of accuracy and processing time of the target task.

In previous works [41], this vocabulary reduction has been typically performed using word filters in the original BoW representation space, however this chapter presents a new approach which makes that reduction in the topics space instead. In particular, the Cumulative Count-based word filter is used to test the performance differences between the vocabulary reduction in the original word feature space and in the topic feature space produced by LDA (Latent Dirichlet Allocation).

The rest of the chapter is organized as follows. Section 2.2 discusses the background of the work. In Section 2.3 the word filter used in this work is explained. An overview of the proposed method is outlined in Section 2.4. Section 2.5 presents and discusses the experimental results. Finally, Section 2.6 draws the main conclusions arisen from this piece of work and notes the future work.

2.2 Background

In this section, the well-known Bag of Words and topic modelling methods used to characterize samples are briefly discussed.

2.2.1 Bag of Words Model (BoW)

In the BoW model [37, 55], a dataset $D = \{d_1,d_2,...,d_N\}$ is a collection of N documents where each document $d_i = \{n(d_i,w_1),...,n(d_i,w_V)\}$ is represented by an histogram of word counts. The number of occurrences of the word j in the document i is represented by $n(d_i,w_j)$. The vocabulary $\omega = \{w_1,w_2,...,w_V\}$, contains the V different words appearing in the whole dataset. The probability of the j^{th} word given the i^{th} document $P(w_j|d_i)$ can be estimated as $P(w_j|d_i) = n(d_i,w_j)/n(d_i)$, where $n(d_i)$ is the number of words that appears in document i. In classification problems, a document i can be represented by a feature vector $P(\omega|d_i)$ composed by the concatenation of the $P(w_j|d_i)$ for all words in ω.

2.2.2 Latent Dirichlet Allocation (LDA)

In general, latent topics [9] are a type of statistical models for discovering the hidden patterns that occur in a data collection. One of the most popular algorithms is Latent Dirichlet Allocation (LDA) [8], which has a parameter K defined by the user to establish the number of topics of the model $Z = \{z_1,...,z_K\}$. The objectives of the LDA inference are the following:

1. Estimating the probability of the j^{th} word of the vocabulary ($j = 1,...,V$) given the k^{th} topic ($k = 1,...,K$), $P(w_j|z_k)$.
2. Estimating the probability of the k^{th} topic given the i^{th} document of the corpus, $P(z_k|d_i)$.

Likewise to the BoW approach, a feature vector can be built using the description of documents in the discovered topics $P(Z|D)$ rather than the initial BoW representation $P(\omega|D)$. In general, the topic representation space $P(Z|D)$ tend to be highly interconnected and often generates a precision drop in classification tasks [48]. However, $P(\omega|Z)$ (i.e. topic descriptions in words) is able to summarize a large dataset into a reduced set of topics ($K << N$) and it may be suitable to reduce the vocabulary ω using far fewer samples.

2.3 Word filter

Word filters allow us to reduce the vocabulary size by removing words that have not semantic meaning or are redundant according to a specific strategy. The aim of word filtering is to reduce the number of words minimizing the loss of semantic information. The original vocabulary $\omega = \{w_1,...,w_v\}$ contains all the words and the reduced vocabulary ω' contains only the selected words by the filter as it is shown in equation 2.1.

$$\omega' = \{w_j | filter(w_j) = 1, \forall j = 1..V\} \tag{2.1}$$

A word filter is a function that indicates for each word if it should be selected by the filter method. Equation 2.2 shows the filter function.

$$filter(w_j) = \begin{cases} 1 & if\ w_j\ is\ selected \\ 0 & otherwise \end{cases} \tag{2.2}$$

In rest of this section, the word filter used in this work is explained. The objective of this simple word filter is not improving the classification accuracy but is to show the plausibility of the proposed approach.

2.3.1 Cumulative Count-based word filter (f_{cc})

For text datasets, this word filter needs a pre-process step to clean non-semantic words (e.g. removing articles, prepositions, derivations, among other actions). f_{cc} word filter considers that the most used words (from the remaining words after performing the preprocessing step) in all documents (BoW filtering) or topics (proposed approach) are more characteristic to represent the documents. First, the size of the reduced vocabulary V' is defined. Then, the V' more used words are selected to be part of the reduced vocabulary. Equation 2.3 shows the filter function for f_{cc} filter.

$$f_{cc}(w_j) = \begin{cases} 1 & if\ w_j \in \omega_{max} \\ 0 & otherwise \end{cases} \tag{2.3}$$

The set ω_{max} is composed of the V' most used words in all the documents or topics (depending on where the word filter is applied) and is obtained as follows:

1. `for` w_j `in vocabulary` ω :
   `(BoW filtering):` $total(w_j) = \sum_{i=1}^{N} P(w_j \mid d_i)$
   `(LDA filtering):` $total(w_j) = \sum_{k=1}^{K} P(w_j \mid z_k)$
2. `Sort` $total$ `in descending order.`
3. ω_{max} `is the first` V' `words of` $total$.

2.4 Proposed Method

The proposed method for vocabulary reduction by topic modelling consists of 3 steps:

1. First, LDA procedure is applied to the original BoW problem (i.e. to the matrix $n(\omega|D)$ in order to obtain the probabilities $P(Z|D)$ (i.e. the description of the documents in the topic space) and $P(\omega|Z)$ (i.e. the description of the topics in the word space).

2. Second, a word filter is applied using as input the matrix defined by $P(\omega|Z)$ to select the reduced set of words ω'.

3. Third, the BoW model is recalculated using only the words in the set ω'.

Note that, the word filters are applied using $P(\omega|Z)$ instead of $P(\omega|D)$ and also note that the number of elements used to perform the vocabulary reduction is the number of topics K instead of the number of documents D. In this work, the number of topics has been set to the number of classes of the dataset ($K = |classes|$) and the vocabulary has been reduced in a 80% ($V' = 0.2V$).

2.5 Experiments

2.5.1 Datasets

In order to analyse the behaviour of the proposed method, three different text datasets have been tested: Reuters, WebKB and Newsgroup [13]. Table 2.1 shows the description of the datasets.

Table 2.1: Datasets description.

dataset	classes	words	documents
Reuters-21578-R8	8	17387	7674
WebKB	4	7770	4168
20 Newsgroup	20	70217	18821

For these text datasets the following pre-processing has been applied [61]:

1. Substitute TAB, NEWLINE and RETURN characters by SPACE.
2. Keep only letters (that is, turn punctuation, numbers, etc. into SPACES).
3. Turn all letters to lower-case.
4. Substitute multiple SPACES by a single SPACE.
5. The title/subject of each document is simply added in the beginning of the document's text.
6. Removing words that are less than 3 characters long.
7. Obtained from the previous step, by removing the 524 SMART stop-words. Some of them had already been removed, because they were shorter than 3 characters.

8. Apply Porter's Stemmer algorithm [69] to the remaining words.

The datasets are divided into 10 folds. These folds are balanced with respect to the classes and the number of samples i.e. all the folds have almost the same number of samples of each class.

2.5.2 Classifier: Support Vector Machine (SVM)

The well-known SVM classification technique [20] has been used for this work. It uses a kernel to transform the original data in a transformed space when a linear classifier is applied. Specifically, the LIBSVM package [30] has been used, which supports multi-class classification. The kernel used has been the Radial Basis Function (RBF). In SVM, it is not known beforehand which parameters are best for a given problem, consequently an estimation of the parameters must be done. For this purpose, a search on SVM has been performed using cross validation in the training set. Various values for the parameters have been tried and the one with the best score has been picked.

2.5.3 Baseline $P(\omega|D)$: Classification on original word feature space

We have first applied a Support Vector Machine (SVM) [20] classification technique using the document representation on the original word feature space of the three datasets, i.e. BoW model $P(\omega|D)$. First row of the Table 2.2 shows the baseline results for the three dataset.

2.5.4 Classification $P(\omega'_1|D)$: Vocabulary reduction on word feature space

f_{cc} word filter has been applied to the original word feature space in order to reduce the number of words and to obtain the reduced vocabulary ω'_1. After that, BoW model has been recalculated using this reduced vocabulary and a classification test has been performed using the SVM in the same way that it was used in the baseline. Second row of the Table 2.2 shows the classification results considering only the reduced vocabulary ω'_1. In this case, the accuracy obtained is almost the same than the baseline, but using a 80% reduced vocabulary.

Table 2.2: Mean and the standard deviation (in brackets) of the classification results (10-fold cross-validation). The best accuracy for each dataset is highlighted in bold.

	Reuters	WebKB	NewsGroup
$P(\omega\|D)$	97.35% (0.37)	90.73% (1.73)	**89.75% (0.73)**
$P(\omega'_1\|D)$	97.26% (0.50)	90.73% (1.39)	89.42% (0.74)
$P(\omega'_2\|D)$	**97.39% (0.42)**	**91.04% (1.75)**	89.38% (0.70)

2.5.5 Classification $P(\omega'_2|D)$: Vocabulary reduction on topic descriptions

The proposed approach (section 2.4) has been performed to obtain the reduced vocabulary ω'_2 using the word filter on topic descriptions $P(\omega|Z)$ obtained by LDA. Once again, BoW model have been recalculated using this reduced vocabulary and a classification test has been performed using the SVM in the same way that it was used in the baseline. LDA parameters have been set to $K = |classes|$, $\alpha = 50/K$, $\beta = 0.1$ and $ITERS_{max} = 100$. The classification results using the reduced vocabulary ω'_2 are showed in the third row of the Table 2.2. In this case, the accuracy obtained is slightly better than the previous two cases when Reuters and WebKb has been used and slightly worse when NewsGroup has been used.

2.5.6 Statistical tests: Wilcoxon Paired Signed Rank Test

In order to compare the results obtained, a Wilcoxon PSRT statistical test [74] has been used. This test is a nonparametric evaluation of paired differences and it allows us to find significant differences in the results. We have compared the classification results in pairs (ω vs ω'_1, ω vs ω'_2 and ω'_1 vs ω'_2) . Table 2.3 shows the test statistics. The statistics showed are $n.s.$ (not significant differences), $p < 0.05$ (significant difference with confidence 95%) and $p < 0.001$ (significant difference with confidence 99.9%).

2.5.7 Discussion

The first question that arises from the experiments is the importance of the vocabulary reduction. According to the results, vocabulary reduction is a useful way to

Table 2.3: Wilcoxon PSRT statistical test

Wilcoxon PSRT	Reuters	WebKB	NewsGroup
$P(\omega\|D)$ vs. $P(\omega'_1\|D)$	n.s.	n.s.	$p < 0.001$
$P(\omega\|D)$ vs. $P(\omega'_2\|D)$	n.s.	n.s.	$p < 0.001$
$P(\omega'_1\|D)$ vs. $P(\omega'_2\|D)$	$p < 0.05$	n.s.	n.s.

reduce the number of words maintaining the classification accuracy. Cumulative-count filter with a vocabulary reduction of the 80% obtains similar results or even slightly improve the classification accuracy with respect to baseline.

According to Wilcoxon statistical tests, the classification task using reduced vocabularies (ω'_1 and ω'_2) has not significant difference with respect to complete vocabulary ω for Reuters and WebKB dataset. Nevertheless, for NewsGroups dataset the classification accuracy is slightly worse, but this difference is lower than 0.5% in the worst case. This accuracy reduction can be accepted considering that the vocabulary reduction is a 80%. Comparing the use of the reduced vocabularies, ω'_1 and ω'_2 have no significant differences to classify for WebKB and NewsGroup datasets. However, for Reuters dataset the classification with ω'_2 (proposed approach) obtains a slight improvement of about 0.5%.

Vocabulary reduction by topic modelling can be an alternative way to reduce the vocabulary size because obtains similar results or slightly better than vocabulary reduction in word space. Also, proposed approach has a main advantage: the sample reduction. To reduce the vocabulary in word space (ω'_1) word filter has to be applied to all the documents of the datasets. However, proposed approach summarizes all the documents in $K = |classes|$ topics and only these topics are used to select the words of the reduced vocabulary (ω'_2).

2.6 Conclusions

In this chapter, a new approach to vocabulary reduction has been presented. This method uses topic models to summarize the samples of a dataset into K topics and it applies a word filter in topic descriptions to reduce the vocabulary in an alternative way. Several classification experiments have been made to compare BoW filtering with respect to proposed approach. A simple cumulative-count based filter (f_{cc}) has been used to test the performance of the proposed method

but other more advanced word filters could be used to improve the classification accuracy. According to the results, vocabulary reduction by topic modelling can effectively deal and slightly improve the vocabulary reduction in original word space for tested datasets. Future work is focused on using more advanced word filters to improve the accuracy classification and define more advanced and effective strategies to choose the number of topics and the size of the reduced vocabulary.

Chapter 3

Latent Topic Encoding for Content-based Retrieval

Publication

[24] Ruben Fernandez-Beltran and Filiberto Pla. Latent topic encoding for content-based retrieval. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 640–648, 2015.

This chapter presents a new encoding approach based on latent topics which is specially designed to Content-Based Retrieval tasks. The novelty of the proposed Latent Topic Encoding (LTE) lies in two points: (1) defining the visual vocabulary according to the hidden patterns discovered from the local descriptors; and (2) encoding each sample by accumulating the proportion of its local features over topics. Several retrieval simulations using two different databases have been carried out to test the performance of the proposed approach with respect to the standard visual Bag of Words (BoW). Results show that LTE encoding is able to outperform the traditional visual BoW when the retrieval task is performed in the latent topic space.

3.1 Introduction

The evolution of technology is leading to bigger multimedia databases and this fact makes the task of retrieving relevant data more complex. Content-Based Retrieval (CBR) is concerned about providing users with those images or videos

which satisfy their queries, that is, semantic concepts that users have in their minds and they are looking for. Over the last years, CBR has been widely addressed by the research community and many approaches have been developed [36, 51]. Despite all this research, the semantic gap [57] between computable low-level features and high-level concepts makes the CBR field still a challenge especially for huge and complex databases.

In general, a CBR system has three main components involved in the retrieval process: (1) a query, represented by a few examples of the concept the user is interested in; (2) a database, which is used to extract samples related to the query concept; and (3) a ranking function, which sorts the database according to the relevance to the query. These three components are usually integrated together with the user in an relevance feedback scheme [82] to provide the most relevant samples through several iterations.

The ranking function can be considered the kernel of the retrieval system because it is in charge of scoring the samples to perform the ranking, however there are more factors which affect to the retrieval performance. One of the most important ones is the encoding technique. The ranking function requires the query as well as the database encoded in feature vectors, that is, samples have to be represented in a specific space in which the ranking function works. The typical pipeline to obtain this space is made up of two steps: (1) extraction of local features (e.g. SIFT [43]); and (2) encoding the local features of each sample in a vector (e.g. histogram of quantized local features). In this chapter, we are going to focus just on the second step, the encoding techniques specially applied to the CBR problem.

In computer vision, the standard encoding procedure is the visual Bag of Words (BoW) [54]. This encoding approach starts by learning a visual vocabulary composed from the clustering of the local features of the training set. Then, each sample is represented in a single histogram of visual words by accumulating the number of local features into their closest clusters. The main drawback of this approach is the hard assignment of words, i.e. it selects the best representing visual word ignoring the relevance and relationship with other clusters. This fact generates an information loss which may be critical to deal with the semantic gap challenge in CBR. More recent advances replace the hard quantization of features with alternative encodings which are able to retain more information about the original features. There are mainly two trends in this field: (1) expressing features

as combination of visual words (co-occurrence models [37], soft quantization [47], local linear encoding [72]); and (2) recording the differences between the features and the visual words (Fisher encoding [46], super-vector encoding [83]).

Despite the fact that some of these methods have shown to obtain good results in classification challenges, the CBR problem has an utterly different nature. In a typical classification problem, we have a training set which is supposed to provide enough information about the classes we want to classify. However, in a retrieval problem the class to retrieve is a priori unknown, it is up to the user's query, and besides we only have few examples of this class, the user initialization and feedback is very limited. As a result, we have to deal with complex classes having very little information about the target. The vast majority of encoding methods obtain the visual vocabulary by clustering the local features and doing that each visual word represents a visible pattern of the data. Nevertheless, in an application like CBR the visible patterns of the data may not be enough to distinguish among unconstrained classes with little information about their structure. At this point, it may be useful to consider other kind of representation techniques beyond the traditional clustering processes are able to provide for the visual vocabulary. Specifically, one of the most suitable techniques for this purpose may be latent topics.

Topic models have been successfully used in many areas (e.g. video classification [11] or even CBR [23]) because they are able to extract hidden patterns from the data distribution and represent the data according these patterns as well. The typical way they have been used is based on reducing the dimensionality of the initial representation space commonly obtained by the standard visual BoW. This chapter presents a novel encoding method completely based on latent topics, which defines the visual vocabulary by means of hidden patterns and performs the quantization by accumulating the contribution of each topic to each local feature. We argue that our proposal provides a more suitable codification for CBR than the standard visual BoW, especially when the retrieval task is performed in the latent topic space.

The rest of the chapter is organized as follows. Section 3.2 presents the Latent Topic Encoding (LTE) method. In Section 5.4, the experimental setting is described as well as the retrieval results obtained by LTE and BoW over two different databases. Finally, Section 5.6 draws the main conclusions arisen from this piece of work and highlights some points as a future work.

3.2 Latent Topic Encoding

The characterization of image or video samples is based on: (1) a local descriptor; and (2) an encoding function. The most common local descriptor methods provide a different number of feature points per sample. Besides, the dimensionality of these feature points is usually very limited to represent the wide variety of features in the visual domain. As a result, the encoding function is in charge of increasing the dimensionality of the descriptor space and representing the whole database using the same visual vocabulary.

In the case of the standard visual BoW, the visual vocabulary is obtained by a clustering process, typically K-Means. Each visual word represents a group of feature points which are spatially close in the descriptor space. However, a distance function is not the best discriminative criteria in applications with a huge semantic gap [57], like in CBR. In those cases, the topology of the space is often not well defined according to the semantics of the data, in other words, samples related to the same query concept may not be close in the descriptor space. At the same time, the hard-assignment of feature points to visual words generates an information loss which might lead to a retrieval precision drop. The proposed LTE encoding method tries to cope with these problems by using latent topics.

Topic models are a suite of statistical algorithms which are able to uncover the hidden structure in document collections. Starting from a specific data matrix $P(\mathcal{W}|\mathcal{C})$ which describes a corpus of documents $\mathcal{C}$ in a particular space $\mathcal{W} \subset \mathbb{N}^n$, latent topic algorithms are able to obtain two matrices: (1) the description of topics in words $P(\mathcal{W}|\mathcal{Z})$ (2) and the description of documents in topics $P(\mathcal{Z}|\mathcal{C})$. The number of extracted topics (Z) is a parameter which has to be established in advance for the most common algorithms. Let us show how the proposed LTE encoding method takes advantage of topic models to define the visual vocabulary as well as to assign feature points to visual words in the CBR context.

In a CBR system, we start with a set of D image or video samples $\mathcal{D} = \{d_1,d_2,...,d_D\}$ in the database, Q query examples $\mathcal{Q} = \{q_1,...,q_Q\}$ to represent the concept the user wants to retrieve and a specific ranking function $\mathcal{R}$ which obtains a ranking $\mathcal{D}'$ of the database $\mathcal{D}$ given the query $\mathcal{Q}$. For this work, we are going to assume that $\mathcal{Q} \subset \mathcal{D}$, that is, queries are selected from the own database, nevertheless further improvements can be aimed at allowing the use of external queries. For each sample d_i,$1 <= i <= D$, a local descriptor algorithm (e.g. SIFT

[43]) is applied to obtain a set of P_i feature points $\mathcal{P}_i = \{p_{i1},p_{i2},...,p_{iP_i}\}$ for that specific sample d_i. We assume that $p_{ij} \in \mathbb{N}^n$,$1 <= i <= D$,$1 <= j <= P_i$. Note that, some local descriptors characterize directly the feature points in $\mathbb{N}^n$ (e.g. counting orientations of gradient) but for other descriptors which characterize the points in $\mathbb{R}^n$ a truncating process must be done.

The proposed LTE method is based on considering each feature point $p_{ij} \in \mathbb{N}^n$,$1 <= i <= D$,$1 <= j <= P_i$, a document of a topic model algorithm, specifically LDA [8] has been used for this work but any other topic model algorithm could be used instead. The visual vocabulary is defined as the set of topics extracted from the corpus containing all the feature points, therefore each visual word represents a hidden pattern in the descriptor space instead of a group of points such as in the BoW approach. The assignment of feature points to visual words is made by accumulating the topic proportion of the points of each sample, that is, the feature points expressed in topics are used to weight the contribution of each point to each visual word. In particular, the LTE method is made up of the following steps:

1. Build a corpus $\mathcal{C}$ with all the feature points, $\mathcal{C} = \bigcup_{i=1}^{D} \mathcal{P}_i$
2. Apply a latent topic algorithm (LDA) over $\mathcal{C}$ in order to discover Z topics,
3. Define the visual vocabulary as the extracted topics, $P(\mathcal{W}|z_k)$,$1 <= k <= Z$
4. Represent each sample in a single histogram as the accumulation of its topic vectors, $h'_i = \sum_{j=1}^{P_i} P(\mathcal{Z}|p_{ij})$
5. Normalize each histogram to obtain the final encoding, $h_i = h'_i/|h'_i|$

Note that, the number of topics Z has the same meaning than the number of clusters in the BoW case, therefore it has to be a number much higher than the dimensionality n of the local feature space obtained by the descriptor. In fact, one of the novelties of LTE is to use topic models to increase the dimensionality of a space rather than decreasing it.

3.3 Experiments

The experiments aim at comparing the proposed LTE encoding method with respect to the classical BoW for CBR tasks. Section 3.3.1 describes the two used

databases, Section 3.3.2 presents the performed retrieval simulations and Section 5.4.2 shows the obtained results.

3.3.1 Datasets

- **Abnormal Object Dataset (AOD)**: The AOD dataset [53] is a balanced image collection with 617 challenging objects over 6 categories (Aeroplane, Boat, Car, Motorbike, Sofa and Chair). A sofa with the appearance of a car or a motorbike which looks like a plane are some instances of AOD. These unusual images have been selected to make confusion among categories in order to increase the semantic gap between low-level features and concepts. To extract the local features, we have used the SIFT [43] descriptor and over these features both BoW and LTE encoding methods have been applied.

- **Columbia Consumer Video Database (CCV)**: The CCV databse [32] contains 9317 YouTube videos over 20 semantic categories, most of which are complex events, along with several objects and scenes. For the experiments, we have considered a subset (sCCV) with 6 random classes (Playground, Wedding Ceremony, Swimming, Skiing, Bird and IceSkating) and for each one we have selected 100 random samples. Regarding to the description method, the SIFT [43] algorithm has been applied to the middle frame to obtain the local features of the videos. As in the former dataset, BoW and LTE encoding functions have been used over these features.

3.3.2 Retrieval Simulations

For the experiments, we have used the retrieval scheme proposed in [23] which is based on Relevance Feedback (RF). In this RF scheme, a simulation has four main parameters: Q the number of samples of the initial query, S the number of top items examined by the user in each feedback iteration, I the number of feedback iterations and R the number of times that the random initialization of the query is repeated per class. According to these parameters, we propose four different retrieval scenarios using in all of them I=5 and R=100: (1) Q=1 S=20, (2) Q=2 S=20, (3) Q=1 S=40 and (4) Q=2 S=40.

The target of each simulation is directed to retrieve samples of a specific class of the dataset, but without using any class label information. The initial query is initialized with Q random samples of a single class c and then the simulation

process has to retrieve samples of that class through I feedback iterations using three different ranking functions: (1) euclidean distance (EC); (2) cosine similarity (CS); (3) and the ranking function proposed in [23] called Latent Topic Rank (LTR). In a nut shell, EC ranks the database according the minimum average euclidean distance to the query, CS according the minimum angle and LTR according to the maximum probability following the expression presented in [23]. At each iteration, the S top ranked items are inspected by a simulated user who marks the samples of the class c. These positive samples are computed as correctly retrieved samples and they are used to expand the query. Finally, this expanded query is triggered as a new query for the next iteration.

In the case of the BoW approach, we have used the K-Means clustering to build the visual vocabulary whereas the LDA topic model [8] has been applied for the LTE encoding. Our objective is to compare the retrieval performance of the proposed LTE encoding with the traditional visual BoW using the retrieval scheme presented in [23] and three different ranking functions: EC, CS and LTR. In addition, we are interested in testing how both BoW and LTE encoding methods perform in the latent topic space. That is, we are going to use BoW and LTE representations as a base to apply LDA in order to analyse the retrieval performance in the latent space depending on the used encoding method.

3.3.3 Results and Discussion

Figure 3.1 shows the results in six graphics organized in a 3×2 matrix. Each row is related to a different ranking function (EC, CS and LTR) and each column contains the retrieval results for a specific dataset (AOD and sCCV). Inside each graphic, we can see the average precision for each one of the 4 simulations using 8 different representations of the data. That is, each bar represents a retrieval experiment using a particular characterization of the database. Specifically, *vBoW_500* relates to the standard visual BoW with 500 clusters by k-means, *LTE_500* indicates the proposed LTE encoding method with 500 topics by LDA, *vBoW_1000* is the visual BoW with 1000 clusters and *LTE_1000* the LTE with 1000 topics. Besides, we have applied LDA with 200 topics over these 4 characterizations to test how the encoding method affects the retrieval performance in the topic space. Note that we have added the text *LDA_200* to the four last captions to indicate that these simulations are performed in the topic space obtained by LDA with 200 topics. In order to make clearer the comparison between LTE

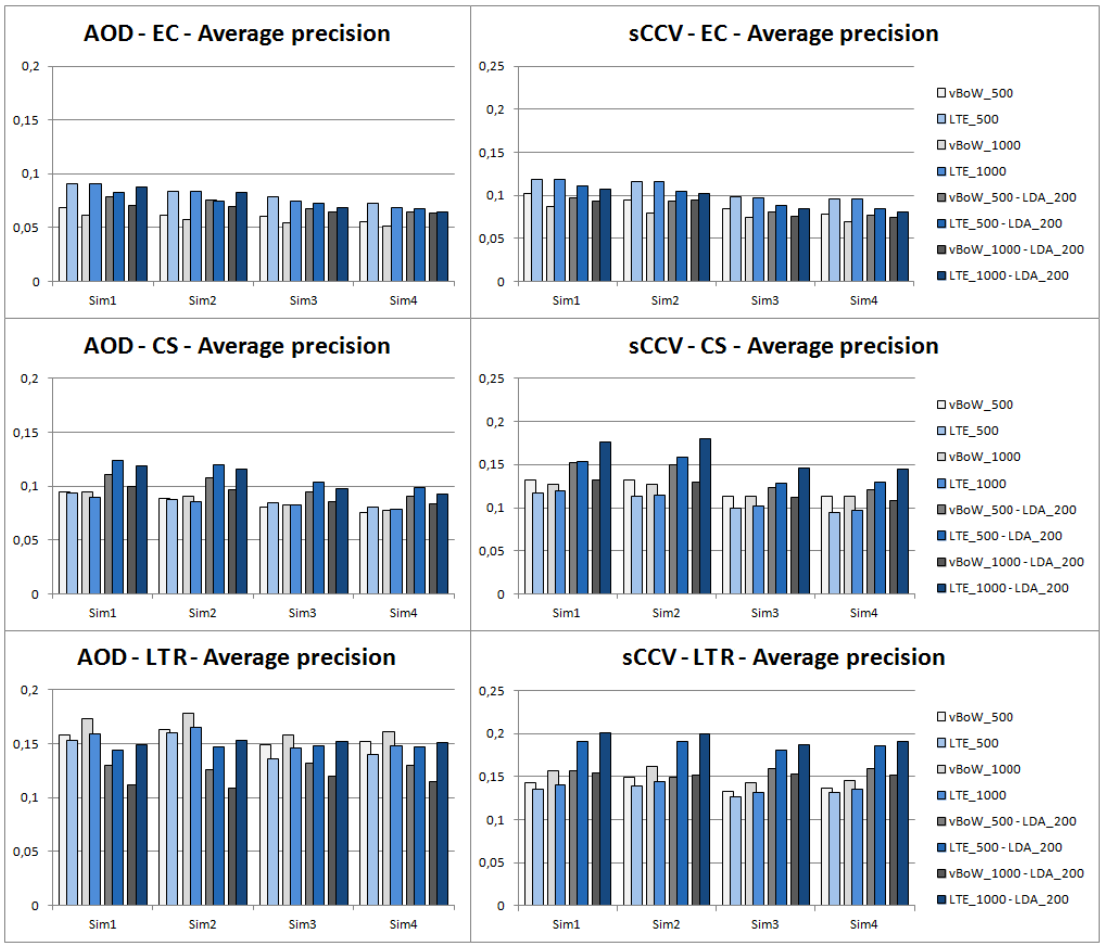


Figure 3.1: Average precision for the retrieval simulations. Ranking functions by rows: (1) euclidean distance (EC), (2) cosine similarity (CS) and (3) latent topic rank (LTR). Datasets by columns: (1) AOD and (2) sCCV.

and BoW, we show the results grouped in pairs of bars, one for the visual BoW approach blue (odd bars) and another for the LTE method (even bars) using in both cases the same vocabulary size. For each pair of bars, if the second bar is higher than the first one, the LTE encoding is outperforming the visual BoW codification in terms of average precision.

Having a look at Figure 3.1, the first noticeably point is the general low precision values obtained in the experiments. This fact shows how important the semantic gap is for these collections. Regarding to the ranking functions, the best average precision has been obtained by LTR and the worse by EC. In the case of the EC ranking, the LTE encoding outperforms the visual BoW approach in all the simulations. However, for the CS and LTR we observe a different pattern. For these two ranking functions, LTE is slightly worse than

visual BoW in the initial representation space (*vBoW_500* and *vBoW_1000*) but noticeable better in the latent topic space (*vBoW_500 - LDA_200* and *vBoW_1000 - LDA_200*). In general, we can see that the proposed LTE encoding function provides a competitive advantage over visual BoW when the retrieval task is performed in the topic space, that is, the retrieval function is used in the latent space obtained after applying LDA to the encoding produced by LTE.

3.4 Conclusions and Future Work

In this chapter, we have presented a new encoding method which defines the visual vocabulary according to the hidden patterns of the local descriptors and represents each sample as the accumulation of its local features represented in these topics. The novelty of the proposal lies on defining an encoding method completely based on latent topics, e.i. topics are used to define vocabulary as well as to make a soft encoding of the local features over topics. For the experiments, we have used the LDA model and SIFT descriptors but any other topic model or descriptor could be used. According to the retrieval results, we can highlight two main points: (1) LTE encoding is more effective than visual BoW for EC ranking and (2) LTE provides a competitive advantage for CS and LTR ranking functions when they are used in the topic space. That is, the proposed encoding method is more suitable than BoW approach in applications to manage samples in the topic space. LTE could be interpreted as extracting the topic structure twice from the descriptor space. The first topic extraction to encode the data and the second one to bring this encoding to a higher semantic level. Future work is focused on comparing the proposed LTE method with more advanced encoding functions and defining an automatic strategy to choose the size of the vocabulary.

Chapter 4

Latent Topics-based Relevance Feedback for Video Retrieval

Publications

[23] Ruben Fernandez-Beltran and Filiberto Pla. An interactive video retrieval approach based on latent topics. In *International Conference on Image Analysis and Processing*, pages 290–299, 2013.

[26] Ruben Fernandez-Beltran and Filiberto Pla. Latent topics-based relevance feedback for video retrieval. *Pattern Recognition*, 51:72–84, 2016.

This chapter presents a novel Content-Based Video Retrieval approach in order to cope with the semantic gap challenge by means of latent topics. Firstly, a supervised topic model is proposed to transform the classical retrieval approach into a class discovery problem. Subsequently, a new probabilistic ranking function is deduced from that model to tackle the semantic gap between low-level features and high-level concepts. Finally, a short-term relevance feedback scheme is defined where queries can be initialised with samples from inside or outside the database. Several retrieval simulations have been carried out using three databases and seven different ranking functions to test the performance of the presented approach. Experiments revealed that the proposed ranking function is able to provide a competitive advantage within the content-based retrieval field.

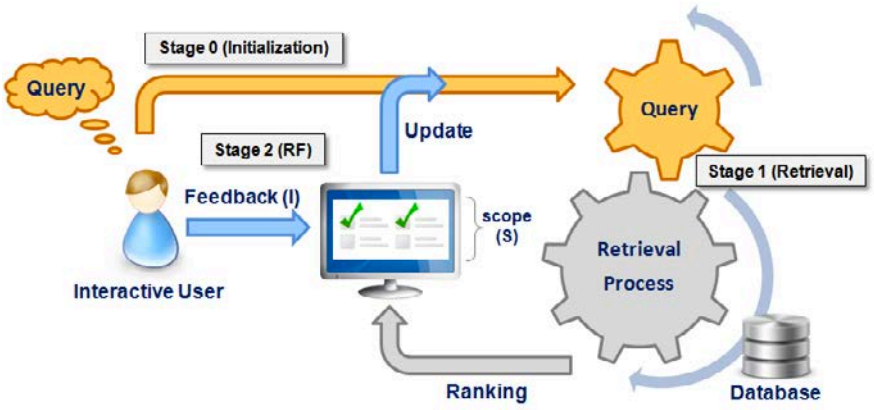


Figure 4.1: Relevance Feedback scheme. Q is the number of initial examples in the query, I the number of feedback iterations and S the number of top ranked samples.

4.1 Introduction

The low cost of image/video capture technology together with the increasing capacity of storage is producing a huge expansion of video collections. In this scenario, one of the most important challenges is how to retrieve users' relevant data from this vast amount of information. Content-Based Video Retrieval (CBVR) is concerned about providing users with those videos which satisfy their queries by means of the video content analysis. As a result, the CBVR field has become a very important research area and a wide variety of CBVR systems have been developed [2, 42, 56, 79]. The standard CBVR procedure involves three main components: (i) a query, containing a few video examples of the semantic concept that the user is looking for; (ii) a database, which is used to retrieve videos related to the query concept; and (iii) a ranking function, which sorts the database according to the relevance with respect to the user's query. These three components are typically integrated with the user in a Relevance Feedback (RF) scheme [15] to provide the most relevant videos through several feedback iterations.

Figure 4.1 shows the general RF scheme for retrieval. At the initialisation stage (stage 0), the user introduces the query concept into the system by providing Q examples of the concept of interest. Then, the interactive process consists of the alternation of two stages through I feedback iterations. In the retrieval stage (stage 1), the system ranks the database according to the query and shows the S top items (scope) to the user. In the feedback stage (stage 2), the user checks the

scope to select the correctly retrieved samples and finally the query is expanded with these new positive examples to carry out the next iteration. The ranking function can be considered the kernel of the retrieval system because it is in charge of scoring the samples of the database according to the query. As a result, the nature of the ranking function and the nature of the video representation space where the ranking function works are two of the most important factors for the precision of a CBVR system.

4.1.1 Ranking functions

One of the most common rankings in multimedia retrieval is the distance-based ranking. Such ranking is performed according to the minimum distance or maximum similarity in the video representation space. Several functions have been proposed in the content-based retrieval field. For instance, in [40] an image retrieval systems is presented which is based on an Euclidean ranking of micro-structure features that combine color, texture and shape. In other works, such as [4], the retrieval ranking is performed using combinations of similarity measures. Even, some authors [66] have combined several descriptors and distance measures to rank the database. Nevertheless, these kinds of functions tend to perform poorer when the query concepts to retrieve are rather complex [50].

Other ranking algorithms are based on inductive learning [64, 65] which typically use a bank of classifiers to represent the set of possible events to test. However, this approach usually leads to a constrained retrieval scheme where users are not allowed to search whatever they want. The CBVR problem itself has an unconstrained nature [31, 50] because the concept to retrieve is a priori unknown. Moreover, the performance of these methods highly depends on the used training data but in the CBVR application the initialisation and feedback are often too limited to provide a consistent training set.

Alternative ranking methods are based on transductive ranking. They use the own topology of the data to improve the output ranking. One of the most representative ones is Manifold Ranking (MR) [81] which ranks the data with respect to the intrinsic data distribution. In a more recent work [76], Yang et al. present a new transductive ranking function called Local Regression and Global Alignment (LRGA) to learn a robust Laplacian matrix which is able to slightly improve the performance of MR. The main drawback of these methods is their high computational cost because they require demanding matrix operations over

the retrieval process. Transductive ranking functions are usually applied in the original descriptor space, however other authors have used a different representation space to perform the ranking. In [80], Zhang et al. present an image retrieval system which computes the cosine similarity function in a topic space to rank the database. This work uses positive (checked) and negative (unchecked) samples in the interactive retrieval process, but managing negative samples adds an extra effort because users have to check false negatives in addition to true positives.

4.1.2 Video representation space

Ranking functions run in a specific representation space where videos are encoded in feature vectors according to the information provided by a descriptor. In the literature, different kinds of descriptors have been proposed using static information - Scale Invariant Feature Transform (SIFT) [43]), spatio-temporal - Spatial Temporal Interest Points (STIP) [35]) or audio - Mel Frequency Cepstral Coefficients (MFCC) [19]. The standard procedure to encode all this low-level information in feature vectors is the visual Bag of Words (vBoW) [54]. The vBoW quantisation starts by learning a visual vocabulary made up of the clustering of the local features. Then, each video is represented in a single histogram of visual words by accumulating the number of local features into their closest clusters. Authors usually refer to this quantised space as descriptor space although it is not the direct output of the descriptor functions. Some recent works have presented more advanced descriptors which are able to achieve better results for specific applications. Wang and Schmid [71] presented a video representation based on dense trajectories specially designed for action recognition which outperforms the most common motion-based descriptors. However, in unconstrained CBVR the type of concepts to deal with is so wide that simpler and non-specialised descriptors are commonly used [79].

4.1.3 Limitations of current approaches and topic models

Several of the aforementioned approaches have shown to be successful at retrieval tasks when they are used on reduced databases with a small number of concepts [58]. Nonetheless, the so-called semantic gap [57] between computable low-level features and query concepts is still a challenge for huge unconstrained video collections. The visual variability of semantic concepts is so high that often

current approaches are not able to capture properly unconstrained queries in extensive collections [79]. Therefore, new capabilities are required in CBVR to bring the video characterisation to a higher semantic level.

Although early research on topic models suggested that they may be used in video retrieval, it was not until recently that topic models were successfully applied to large unconstrained video collections [23]. In general, topic models can be used for automatically organising, understanding, searching and summarising large electronic archives [9]. For many years, topic models have not been considered useful in tasks where precision is important because traditional ranking functions tend to perform worse in the latent space than in the original characterisation space [6]. The latent topic space is usually a lower dimensionality representation where concepts and classes are more diffuse and besides it allows connections among different concepts through the patterns defined by topics. As a result, the most effective ranking functions in the original feature space are usually not useful in the topic space because this space has an utterly different nature. However, this fact does not mean the topics' lack of usefulness. In those applications in which the semantic gap is important, the retrieval precision in the original feature space tends to be very low and topic models can provide a competitive advantage by means of the hidden patterns that topics represent. It is the case of unconstrained CBVR. The difference between the low-level video features and the high-level query concepts can be so huge that the patterns defined by topics may be interpreted as a higher characterisation level and may help us to obtain a better retrieval performance. However, the most common ranking functions do not take into account the own nature of the topic space what eventually makes that many of them do not work properly in this representation.

4.1.4 Objectives and structure

The main objective of this chapter is to obtain an effective and efficient CBVR approach completely based on the rationale of latent topics in order to deal with the semantic gap challenge by means of the patterns defined by topics. First of all, the supervised Symmetric probabilistic Latent Semantic Analysis (sSpLSA) model is proposed to transform the classical retrieval approach into a class discovery problem what allows us to handle the user's searching concept as a mixture of hidden patterns. Subsequently, a new probabilistic ranking function is deduced from that model in order to estimate the probability that each sample of the

database belongs to the query class (searching concept). Finally, the proposed retrieval approach is defined allowing both internal and external queries. In this work, we have considered a short-term RF approach, that is, each searching process is independent from one another. However, further improvements could be aimed at developing a long-term approach where the system learns from previous searches as well.

This chapter extends our previous work [23] where the sSpLSA model was introduced to obtain an initial ranking function which had some limitations. One of those limitations was assuming that queries are only from inside the database. There are two different ways the user can initialise a query, selecting samples from the own database or by providing external ones. In the first case, the user explores the database and selects some samples containing the concept of interest. However, that is not always the case. When the database is really huge or the query concept is very rare, it could be rather difficult to find proper samples to initialise the query. In those cases, it is more effective to initialise the query with external samples as long as the user has some examples of what they are looking for. In the present chapter, the retrieval model is extended and the ranking function is revised using more realistic assumptions what leads to an improvement of the retrieval performance. In addition, this work extends the experimental part with a more comprehensive experimental setting, adding more relevant methods in the literature and using more databases.

The rest of the chapter is organised as follows: in Section 4.2, the proposed latent-topic retrieval model is presented including the definition of a new ranking function (Section 4.2.3) and a procedure (Section 4.2.4) to enable the use of external queries. Section 5.4 shows the retrieval experiments using three different databases: PAL [45], CCV [32] and TREVID [7]. Finally, Section 5.5 discusses the results and Section 5.6 draws the main conclusions arisen from the work.

4.2 Probabilistic latent topic retrieval model

4.2.1 Probabilistic topic models

In general, topic models are a kind of statistical graphical models which are able to uncover the hidden structure that pervade a data collection. Specifically, these methods take as an input a specific data probability matrix $P(W|D)$ which

describes a corpus of documents $D = \{d_1,...,d_N\}$ in a certain word space $W = \{w_1,..,w_M\}$ and obtain as an output two probability matrices, the description of K topics $Z = \{z_1,...,z_K\}$ in words $P(W|Z)$ and the description of documents in topics $P(Z|D)$. The majority of topic methods are in the families of two models, probabilistic Latent Semantic Analysis (pLSA) [29] and Latent Dirichlet Allocation (LDA) [8]. Both pLSA and LDA are a reference in topic modelling although there are significant differences between them. On the one hand, pLSA uses the documents of the collection as parameters of the model what makes pLSA a high spatial demanding model and generates topic over-fitting when too many parameters are considered. On the other hand, LDA tries to overcome pLSA drawbacks by using two Dirichlet distributions, one to model documents $P(Z|D) \sim Dir(\alpha)$ and another to model topics $P(W|Z) \sim Dir(\beta)$. Logically, these parameters α and β have to be estimated during the topic extraction process which adds an extra computational cost.

Despite the fact that the experimentation in [8] reveals that LDA is able to achieve lower perplexity than pLSA, it is not clear how the perplexity correlates with the performance in retrieval tasks. The same Blei in [14] concludes that pLSA often obtains a topic structure more correlated to the human judgement than LDA although the perplexity values suggest the opposite. In the standard LDA algorithm, the parameter estimation is performed by iterating over the document collection what produces that LDA requires a certain number of documents to adequately estimate its hyper-parameters. In an application like CBVR, the concept to retrieve is a priori unknown because it is up to the user. Besides, the initialisation and feedback are often very limited. As a result, it is usual to deal with complex concepts having very little information about them and in these circumstances pLSA is more accurate [44]. For these reasons, we have decided to use the pLSA model as the basis of our extended model for CBVR.

4.2.2 Supervised symmetric probabilistic Latent Semantic Analysis (sSpLSA)

The supervised Symmetric probabilistic Latent Semantic Analysis (sSpLSA) model (Figure 4.2) extends the unsupervised symetric pLSA [29] model by adding the observed random variable corresponding to class label y. In this case, the approach is directed to a similar scenario than the single-author topic model used

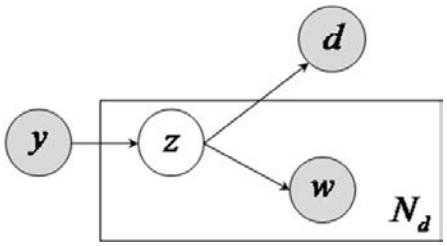


Figure 4.2: sSpLSA model. y is the class, z the topic (hidden variable), w the word, d the document and N_d the number of words of d.

by Fei-Fei and Perona [38] in the framework of a LDA-based model. The generative process of the sSpLSA model stems from the class probability distribution $p(y)$. In the model, classes y are expressed as topic mixtures of topics z according to parameters $p(z|y)$. Therefore, the process to generate a document d can be interpreted as follows:

- A class y is drawn for a document d from the probability distribution $p(y)$.
- For each one of the N_d words in the document d,
  - Given the document class y, a topic z is chosen according to conditional distribution $p(z|y)$ that expresses classes in topics.
  - Given the topic z chosen, a word w is drawn from the conditional distribution $p(w|z)$ that relates topics to words.
- Given the N_d topics drawn to extract the words, a document d is defined according to the class conditional distribution $p(d|z)$.

The sSpLSA model could be used to extract the topics of a data collection using information about class labels (y) like a regular supervised topic model but it is not the goal here. We aim at relating the sSpLSA general model (Fig. 4.2) to the RF retrieval scheme (Fig. 4.1) in order to obtain a probabilistic ranking function based on sSpLSA. For that purpose, we use the following notation: y' is the query class and represents the kind of videos the user wants to extract from the database and $D' = \{d'_1,...,d'_{N'}\}$ refers to the query set containing one or more positive examples of the query class. Note the difference between y and y'. The former (y) is related to the general concept of class label information used in the sSpLSA model and the latter (y') is the specific kind of videos the user wants to extract from the database in a specific retrieval session. Our objective is to sort

the database using as a score of the ranking the probability that each document d of the database belongs to the query class y', i.e. $p(y'|d)$. In the next section, we are going to deduce the ranking function of the proposed approach deriving this probability over the sSpLSA model.

4.2.3 Latent Topic Ranking (LTR)

Initially, we assume that a topic process has been carried out over the database in order to extract a specific number of topics (K) and to express the whole collection according to those extracted topics as $P(Z|D)$. For the topic extraction task, it can be used either a supervised model (sLDA...) or unsupervised (LDA, pLSA...) one. It should be noted that in the supervised case topics are extracted using some initial class label information y which does not have to be related to the concept y' (query class) that the user wants to retrieve in a particular session.

The proposed Latent Topic Ranking (LTR) function is aimed at providing a guess of the probability $p(y'|d)$ that each document d of the database belongs to the query class and using these probability values it performs the ranking at each retrieval iteration. According to the sSpLSA model (Fig. 4.2), this probability can be estimated from the present user's query by means of topic characterisations as follows. Let us express the conditional probability $p(y'|d)$ by marginalising over topics:

$$p(y'|d) = \frac{p(y',d)}{p(d)} = \frac{\sum_w \sum_z p(w,d,z,y')}{p(d)} = \frac{\sum_w \sum_z p(w|z)p(d|z)p(z|y')p(y')}{p(d)} \quad (4.1)$$

Where it has been assumed that the joint probability $p(w,d,z,y')$ is expressed according to the introduced sSpLSA model. Regarding the conditional topic probability of a given class $p(z|y')$, it can be estimated by marginalising over the query set $D' = \{d'_1,...\}$ as follows:

$$p(z|y') = \sum_{d'} p(z,d'|y') = \sum_{d'} \frac{p(z|d',y')p(d',y')}{p(y')} = \sum_{d'} \frac{p(z|d',y')p(y'|d')p(d')}{p(y')} \quad (4.2)$$

Inserting (4.2) in (4.1) we obtain

$$p(y'|d) = \frac{\sum_w \sum_z p(w|z)p(d|z) \sum_{d'} p(z|d',y')p(y'|d')p(d')}{p(d)} \tag{4.3}$$

The conditional probability $p(y'|d')$ represents the probability that a document of the query belongs to the query class which is always true, therefore $p(y'|d') = 1$. Moreover, assuming the normalisation constraint over topics $\sum_w p(w|z) = 1$, expression (4.3) can be simplified as follows:

$$p(y'|d) = \frac{\sum_z p(d|z) \sum_{d'} p(z|d',y')p(d')}{p(d)} \tag{4.4}$$

After multiplying and dividing by $p(z)$ and applying Bayes' rule $p(z|d) = p(d|z)p(z)/p(d)$ we obtain

$$p(y'|d) = \sum_z \frac{p(d|z)p(z)}{p(d)p(z)} \sum_{d'} p(z|d',y')p(d') = \sum_z \frac{p(z|d)}{p(z)} \sum_{d'} p(z|d',y')p(d') \tag{4.5}$$

Let us assume that the probability $p(d)$ of the documents of the database and the probability $p(d')$ of the documents of the query follows the same uniform distribution over the total number of documents of the database $|D|$, i.e. $p(d) = p(d') = 1/|D|$. This assumption implies that all the documents have the same prior probability independently of their number of words, features or even their relation with other samples. In the case of internal queries, it makes sense to use $1/|D|$ as an estimation of $p(d')$ because queries are selected from the own database. Besides, even in the case of external queries the number of samples from outside the database is so reduced compared with the number of documents in the database ($|D'| << |D|$) that the value $1/|D|$ is a good approximation to $1/(|D| + |D'|)$. Thus, $p(z)$ can be estimated by marginalising the documents d_i of the database and using Bayes' rule

$$p(z) = \sum_{d_i} p(z,d_i) = \sum_{d_i} p(z|d_i)p(d_i) \approx \frac{1}{|D|} \sum_{d_i} p(z|d_i) \tag{4.6}$$

Inserting (4.6) into (4.5), the probability $p(y'|d)$ can be expressed as

$$p(y'|d) \approx \sum_z \frac{p(z|d)}{\sum_{d_i} p(z|d_i)} \left[\sum_{d'} p(z|d',y') \right] \tag{4.7}$$

In the present work, we have considered a short-term RF scheme what means that each retrieval session is independent from one another. In other words, all the information we have about the query class y' is provided by the samples of the query set d', therefore $y' \approx d'$ and then $p(z|d',y') \approx p(z|d')$. As a result, the final expression to estimate the probability $p(y'|d)$ for the LTR function is as follows:

$$p(y'|d) \approx \sum_z \frac{p(z|d)}{\sum_{d_i} p(z|d_i)} \left[\sum_{d'} p(z|d') \right] \tag{4.8}$$

Expression (4.8) has two main factors. The left one is related to the document d of the database we want to rank and the right factor represents the query at a specific stage of the retrieval process. In the first factor, $p(z|d)$ is learned off-line using any latent topic algorithm, for instance pLSA or LDA. Then, $\sum_{d_i} p(z|d_i)$ can be precomputed off-line as well using all the documents of the database. In the second factor, $p(z|d')$ is the probability that a given query document d' belongs to the topic z.

The ranking process is made as follows. First of all, K topics are extracted from the database using some topic extraction method and subsequently each document d of the database and the initial query documents d' are represented in these topics as $p(z|d)$ and $p(z|d')$ respectively. Later, the database is sorted according to the probability that documents d belong to the query class y' using equation (4.8). Following the relevance feedback scheme, the S most likely samples (scope) are showed to the user who selects the P correctly retrieved samples. Then, these P samples are used as feedback to expand the query. At each iteration, the query is expanded with more positive examples and probabilities $p(y'|d)$ are recomputed to refine the ranking. In the end, the interactive process ends after I iterations when the user has retrieved enough samples.

Comparing the LTR function (4.8) with the version in [23], we can observe two main differences. On the one hand, in LTR documents are normalised by the global use of the topics in the collection, therefore the least used topics are able to generate a higher probability values. That is, the match with the query patterns is calculated by fostering the least used topics. On the other hand, expression (4.8) uses $p(z|d')$ (query expressed in topics) instead of $p(d'|z)$ (topics expressed in query documents). This allows to get rid of the simplification we made in the ranking process of [23] where we assumed that topics do not depend on queries to approximate $p(d'|z)$ as the transposed and normalised version of $p(z|d')$ what

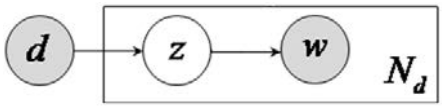


Figure 4.3: Graphical model representation of the aspect model in the asymmetric pLSA parametrization used by Hofmann in [29]. d is the document, z the topic (hidden variable), w the word and N_d the number of words of d.

is not a real premise.

Another important change is based on allowing the use of internal and external query samples. The off-line topic learning process obtains $P(W|Z)$ and $P(Z|D)$ from the database. Thus, when queries are inside the database, we already have the description of the query documents in topics. However, when queries are initialised with external samples, we have to use an estimation procedure to represent those external documents in the previously extracted topics. The following section shows the used procedure to represent external samples in a set of given topics.

4.2.4 Expectation Maximisation eStimator (EMS)

As it was mentioned earlier, regular topic algorithms such as pLSA and LDA are able to obtain from a data collection the description of topics in words as $P(W|Z)$ and the representation of the database in topics as $P(Z|D)$. However, in this work queries can be initialised with samples from outside the database and therefore the proposed approach requires an additional procedure to represent external query documents $D'_{out} = \{d'_{out_1},...\}$ in a given set of topics as $P(Z|D'_{out})$. Following the same notation than before, the upper-case letter represents the set and the lower-case an instance of that set.

We use the asymmetric version pLSA model (Fig. 4.3) to define the Expectation Maximization eStimator (EMS) procedure. Specifically, the parameter $p(z|d'_{out})$, which represents an external query document in a given set of topics, can be estimated following the pLSA model by maximizing the log-likelihood using the Expectation-Maximization (EM) algorithm. Let us define the joint distribution of the model Eq. (4.9) and the log-likelihood Eq. (4.10) in terms of the joint probability distribution

$$p(w,z,d'_{out}) = p(w|z)p(z|d'_{out})p(d'_{out}) \tag{4.9}$$

$$\mathcal{L} = \sum_{w} n(w,d'_{out})\log[p(w,d'_{out})] \tag{4.10}$$

Where $n(w,d'_{out})$ is the number of occurrences of the word w in the document d'_{out}. In order to maximize the log-likelihood by EM, the complete log-likelihood can be expressed using the latent variables z as

$$E = \sum_{w} n(w,d'_{out})\left(\sum_{z} p(z|w,d'_{out})\log[p(w|z)p(z|d'_{out})p(d'_{out})]\right) \tag{4.11}$$

Introducing in expression (4.11) the normalisation constraints of the parameter $p(z|d'_{out})$ by inserting the appropriate Lagrange multiplier β:

$$H = E + \beta\left[1 - \sum_{z} p(z|d'_{out})\right] \tag{4.12}$$

Taking the derivative with respect to $p(z|d'_{out})$, setting the expression equal to zero and solving the equation to isolate the parameter, the M-step of the EM algorithm is expressed as

$$p(z|d'_{out}) = \frac{\sum_{w} n(w,d'_{out})p(z|w,d'_{out})}{\sum_{z}\sum_{w} n(w,d'_{out})p(z|w,d'_{out})} \tag{4.13}$$

For the E-step, we need to estimate the parameter $p(z|w,d'_{out})$. Applying the Bayes' rule and the chain rule we obtain

$$p(z|w,d'_{out}) = \frac{p(w,z,d'_{out})}{p(w,d'_{out})} = \frac{p(w|z)p(z|d'_{out})}{\sum_{z} p(w|z)p(z|d'_{out})} \tag{4.14}$$

The EM process is performed as follows. First of all, the external query

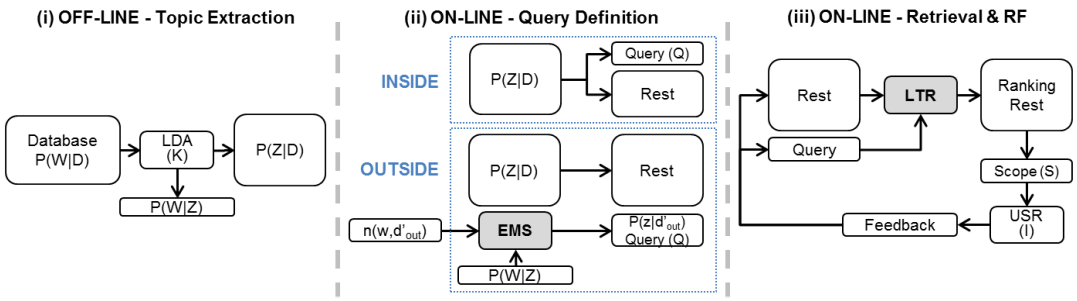


Figure 4.4: Proposed approach scheme. K (number of topics), Q (number of samples to initialise the query), I (number of feedback iterations) and S (number of top ranked samples).

document $n(w,d'_{out})$ and the set of previous topics $p(w|z)$ are loaded. Secondly, $p(z|d'_{out})$ is randomly initialised. Then, the E-step Eq. (4.14) and the M-step Eq. (4.13) are alternated until a convergence condition is reached. As default settings, we have used a threshold of 10^{-6} in the difference of the log-likelihood Eq. (4.10) between two consecutive iterations and a maximum of 1000 EM iterations to assure a fixed and sensible computational cost in the convergence process.

4.2.5 Latent Topic-based Relevance Feedback Framework

The proposed retrieval approach is made up of three main phases (Fig. 4.4): (i) off-line topic extraction, (ii) on-line query definition and (iii) on-line retrieval and relevance feedback. In the first phase (i), the LDA [8] algorithm is used over the collection in order to extract K topics as $P(W|Z)$ and to represent the samples of the database in those topics as $P(Z|D)$. Note that $P(W|D)$ represents the normalised word count of the documents of the collection. We have selected LDA instead of pLSA because the spatial cost of pLSA for the tested collections is unaffordable, however pLSA or any other topic model can be used in this phase instead. Once this off-line process has been carried out the system changes to the on-line mode which contains two more phases.

The phase (ii) on-line query definition is corresponded to the Stage 0 of the RF scheme showed in the Fig. 4.1. In this part, the user has two different alternatives to initialise the query. When queries are from inside the database, Q query samples are selected from the own collection as the initial query set and the rest of the samples make the *Rest* set which is the basis to perform the ranking. When queries are from outside, the EMS function is used to represent

the external query samples $n(w,d'_{out})$ in the previous K extracted topics $P(W|Z)$. Note that $n(w,d'_{out})$ represents the word count of the external query documents. Then, these external samples expressed in topics $p(z|d'_{out})$ make the *Query* set and the whole database $P(Z|D)$ is used as the *Rest* set. The external samples have to be represented in the same initial characterisation space W than the database used to extract the topics e.i. using the same descriptor.

Once the *Query* and *Rest* sets are initialised, the proposed approach changes to the phase (iii) on-line retrieval and relevance feedback which represents the stages 1 and 2 of Fig. 4.1. In this iterative stage, the LTR function uses both the *Query* and *Rest* sets to obtain a ranking of *Rest* by using equation (4.8). From this ranking, the S top samples are shown to the user who selects the positive samples to provide the feedback. These correctly retrieved samples are used to expand the *Query* and besides they are removed from the *Rest* set. In order to reduce the complexity of the interaction process, only positive feedback samples are used to expand the query. Finally, with the updated *Query* and *Rest* sets the next iteration is triggered. The number of total feedback iterations I depends on the user, that is, the user decides when the interaction ends.

4.3 Experiments

This section presents the experimental part of the chapter. Section 5.4.2 describes the kind of retrieval simulations performed in the experiments and the retrieval methods of the literature used to test the proposed approach. Subsequently, sections 4.3.2, 5.4.2 and 5.4.2 show the retrieval results for three different databases: PAL [45], CCV [32] and TRECVID 2007 [7].

4.3.1 Short-term Relevance Feedback simulations

A total of six different user interaction scenarios are defined to evaluate the effectiveness of the proposed approach with respect to seven different retrieval methods over three databases. We assume that each database used for the simulations is a pre-labelled collection, i.e. it is annotated according to a specific set of classes, and besides it is partitioned in two balanced halves, training and test.

Parameters of the simulations

Following the scheme of the proposed approach (Figure 4.4), the on-line stage has three main parameters: Q the number of samples of the initial query, S the number of top examined items and I the number of total iterations. The target of each simulation is directed to retrieve samples of a specific class, but without using any class label information. In other words, the query is initialised with Q samples of a single class c and the simulation process has to retrieve samples of that class through I feedback iterations. At each iteration, the S top ranked items are inspected by a simulated user who marks the samples of the class c (positive samples). These positive samples are computed as correctly retrieved samples and they are used to expand the query. Finally, the expanded query is triggered as a new query for the next iteration.

In this work, we assume a simulated-user reliability of a 100% in order to simplify, but some uncertainty could be introduced in the simulation process. This uncertainty could be introduced into the retrieval system in a soft way or in a more intense way. An example of the former case could be by limiting the number of feedback examples per iteration. That is, instead of selecting all the positive examples each feedback iteration just marking a few correctly retrieved samples. Note that, this is a quite common situation because real users do not often analyse the whole content of a screen. Another example of a more aggressive uncertainty could be by introducing some mistakes in the feedback process. This fact may produce a remarkable precision drop and its study would be interesting to test the stability of the different retrieval methods.

The experiments are divided in two kinds of simulations according to the initialisation of the query (Fig. 4.4): (a) when queries are from inside the database and (b) when queries are from outside the database. In the first case (a), the complete dataset is used to extract K topics by LDA and then for each class c of the database queries are initialised with Q random samples of the that class. This random initialisation is repeated R times in order to obtain an average value of the retrieval precision and an average computational time per query. Note that by complete dataset we mean the union of both partitions training and test because we assume that the dataset is initially divided into these two balanced partitions.

When queries are from outside the database (b), the training partition is used to extract K topics using LDA and the test set is represented according to those topics by means of the EMS function. Then, each sample of the test set is used

to trigger an external query and therefore the target is to retrieve videos from the training set which belong to the same class as the query sample of test set. Note that in this case there is no point in considering the parameter R because each test sample is a query itself, thus there is not random initialisation. Like in the previous case, the performance measures of the simulations are the average precision and average computational time per query.

Table 4.1: Parameters of the simulations for the experiments.

(a) INSIDE				**(b) OUTSIDE**			
Simulation	**Q**	**I**	**S**	**Simulation**	**Q**	**I**	**S**
1	1	5	20	1	1	5	20
2	2	5	20				
3	1	5	40	2	1	5	40
4	2	5	40				

Table 4.1 shows the six different simulations considered for the experiments. Those parameters have been set taking into account the user comfort in the retrieval process. For real users, it is not comfortable to initialise the query with many samples and for that reason we assume that the user only provides one or two examples, that is, $Q = \{1,2\}$. The number of feedback iterations is another important parameter. The retrieval systems require a certain number of iterations to be properly aided, but a high number of iterations affect negatively to the user's attention. Therefore, we consider $I = \{5\}$ to balance the efficacy of the retrieval system and the user's preferences.

Regarding the scope S, somehow this parameter is related to I. A bigger scope may reduce the number of feedback iterations, but it makes users to check more samples at each iteration which eventually affects to their comfort. As a result, we have chosen two reasonable values for the scope, $S = \{20,40\}$. Considering that the average number of videos which can be shown in a regular screen is around 20, that configuration simulates two different scenarios: one where the simulated users are inspecting only the first screen at each feedback iteration and another where they are inspecting two screens per iteration.

Other important parameters are R, the number of times the query is randomly initialised, and K, the number of extracted topics. Note that those parameters change from database to database, therefore they are not included in table 4.1 but in the tables with the results for each database in Section 5.4.2. The parameter R has been selected to perform a reasonable number of random initialisations of the query to obtain robust average values. Regarding the number of topics,

selecting the right number of topics is an open-ended issue, especially in the visual domain. In the literature, there are several approaches which try to tackle this problem but all of them require performing the topic extraction process several times which makes them impractical to be used in a real system. As a result, we have tested different number of topics according to the size of the databases to make the results consistent.

Retrieval methods for comparison

In order to evaluate the proposed approach, we have compared our method with seven different ranking functions. These functions have been selected because they are widely used in literature and they usually obtain a good performance in retrieval or classification tasks. In this work, we have used a short-term Relevance Feedback approach, thus simulations do not use training information of previous searches. Other important retrieval approaches need search examples as training set. Ranking SVM [33] is a powerful tool for optimising the similarity function of content-based retrieval systems, but it needs a reasonable training set to carry out the ranking. In the experimental comparison, we have only used retrieval methods suitable for a short-term Relevant Feedback scheme like the proposed approach. Specifically, we have considered distance-based and transductive-based ranking methods for the experiments.

The following distance/similarity ranking functions [22] have been tested: Euclidean distance (EC), symmetric Kullback-Leibler divergence (KL), Cosine similarity (CS), Hellinger distance (HL) and Bhattacharyya distance (BC). These functions have been used on the original BoW representation of the dataset $P(W|D)$ and besides on the topic space generated by LDA $P(Z|D)$ in order to compare the retrieval performance in both cases. The distance/similarity based ranking sorts the samples of the database according to the minimum distance or maximum similarity to the query. In the case that the query has more than one sample we have computed the arithmetic mean of the measure. Specifically, we have chosen this averaging strategy rather than a max pooling one because in CBVR the user's initialisation and feedback are too limited to take advantage of sub-sampling the query set.

Regarding the transductive learning, we have selected MR (Manifold Ranking [81]) and LRGA (Local Regression and Global Alignment [76]) as two of the most important retrieval algorithms. However, LRGA suffers from a high

computational cost when it is used over a large number of samples with high dimensionality. For this reason, we found that LRGA was not computationally affordable for the considered databases and therefore we have only tested MR ranking in both spaces $P(W|D)$ and $P(Z|D)$.

Additionally, we have tested another method in $P(Z|D)$. Zhang et al. [80] presented an image retrieval system which uses the cosine similarity function in the topic space to rank the database. It uses positive and negative samples in the Relevance Feedback (RF) process and summarises the query set in only one sample in the initial representation space $P(W|D)$. Then, this sample is represented as $p(z|d)$ according to the topics extracted from the database and eventually the cosine similarity is computed in the topic space to perform the ranking. For comparison purposes, we have adapted this approach to the framework used in this work in order to deal with only positive feedback. Algorithm 1 shows the Zhang (ZH) ranking function adapted for the experiments.

Algorithm 1: *RankingFunction* of Zhang simulations.

input: $QUERY$, $REST$

$size = |QUERY|$;
$pos = \frac{1}{size}\sum_{d \in QUERY} p(w|d)$;
Determine $p(z|pos)$ with EM [80];
for *video v in REST* **do**
  Compute *Cosine similarity* between $p(z|pos)$ and v;
end
Rank $REST$ according to maximum similarity;

4.3.2 Productive Ageing Lab (PAL) database

The Productive Ageing Lab (PAL) face collection [45] contains 573 colour images of size 640×480 pixels corresponding to 223 males and 350 females with ages ranging from 18 to 93. The dataset has been randomly split into two balanced partitions, one for training with 112 males and 175 females and another for test with 111 males and 175 females. As a characterisation of the data, we have used the images converted into grey levels, scaled to 16×13 pixels and vectorised. As a result, the original feature space $P(W|D)$ of this database contains 208 words.

We first use this dataset to find out easily the differences among the output rankings obtained by the tested retrieval methods. Gender recognition is a 2-class

problem with a wide intra-class variety, i.e. two very different faces could belong to the same gender, and this fact may easy the task to detect small ranking differences. Note that the gender recognition problem is extensively studied in the literature and the objective here is not to obtain a good accuracy but to compare the different output rankings.

4.3.3 Columbia Consumer Video (CCV) database

The Columbia Consumer Video (CCV) dataset [32] contains 9317 YouTube videos over 20 semantic categories, most of which are complex events, along with several objects and scenes. The authors of the database provide two balanced partitions, one for training with 4659 samples and another for test with 4658 samples. Besides, they provide three different video descriptors SIFT, STIP and MFCC. For the experiments, we have used the characterisation based on the SIFT descriptor which contains 5000 words. In particular, this codification is made up of the concatenation of five different parts: (1) the complete sample and (2)-(5) the division of the sample in a 2×2 grid. Each one of these parts is encoded using 1000 words as the concatenation of two different vocabularies: (a) BoW with 500 clusters over SIFT descriptor and Hessian-Affine detector and (b) BoW with 500 clusters over SIFT descriptor and DoG detector.

In this corpus, we have detected some samples with null descriptor content and others without annotation. In both cases these samples have been removed for the experiments. For the remaining ones, those samples labelled with more than one category have been replicated one for each class. As a result, we have considered a total of 7846 video samples, 3914 of training and 3932 of test, annotated in 20 classes as it is shown in Fig.5.7.

4.3.4 TRECVID 2007 database

The TRECVID 2007 collection [7] is made up of 47,548 video shots which are annotated according to 36 semantic concepts. These categories were selected in TRECVID 2007 evaluation and they include several objects as well as complex events and scenes. Regarding the description of the database, we have used a similar characterisation than in the case of CCV. Specifically, we have followed the suggestions of van de Sande et al. [68] about using opponent SIFT histograms when choosing a single descriptor and no prior knowledge about the dataset is

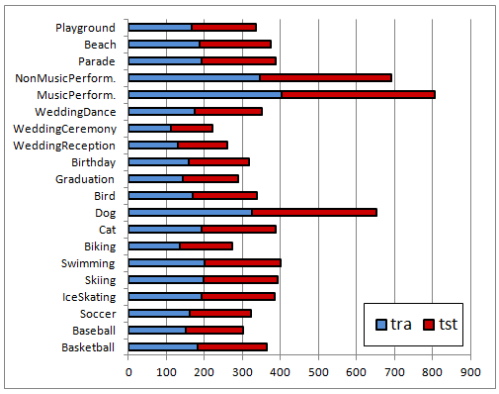


Figure 4.5: Samples per class of the CCV database.

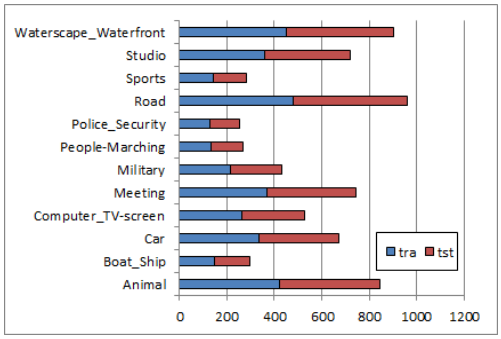


Figure 4.6: Samples per class of the considered subset of TRECVID 2007.

considered. The software provided by van de Sande has been applied to the middle frame of each shot and each sample has been encoded using a 3-level spatial pyramid codebook (1×1, 2×2 and 4×4) what makes a total of 2688 words per shot. In order to make affordable the computational cost of the topic extraction task, we have reduced the original database by selecting a balanced subset with a similar size to the CCV collection. Specifically, we have divided the whole collection in 10 balanced partitions. Later, we have removed the classes under 100 samples in any partition, resulting a total of 17 selected classes. Finally, we have chosen one random partition as a training set and another as a test. Figure 5.9 shows the considered subset of 8974 samples with 4487 for training and 4487 for test annotated in 17 classes.

4.3.5 Results

Tables 4.2, 4.3 and 4.4 present the retrieval result in terms of Average Precision (AP) and average computational Time per query in seconds (T) running in a

single processor Intel Xeon E5-2640. Each table corresponds to a particular database and the way they are organised is the following. In columns we have the six different simulations described in section 5.4.2, the first four (a) using internal queries and the last two (b) with external ones. The parameters of each simulation (R,Q,I,S) are indicated in the headings of the columns. In rows we have the different retrieval methods used for the experiments. In particular, there are three groups: LTR which contains the results of the proposed approach using several number of topics (K), $P(W|D)$ which has the results of six different ranking functions in the original characterisation space and $P(Z|D)$ contains the results of seven different ranking functions in the best topic space among the tested number of topics.

Related to the ranking functions, we use the following terminology: Euclidean distance (EC), symmetric Kullback-Leibler divergence (KL), Cosine similarity (CS), Hellinger distance (HL), Bhattacharyya distance (BC), Manifold ranking [81] (MF) and Zhang approach [80] (ZH).

Table 4.2: Retrieval result for PAL database: Average Precision (AP) and average seconds per query (T). For each group of ranking functions (in rows), the best AP value of each simulation is highlighted in bold and the best global value among all methods is underlined.

	METHOD		(a) INSIDE								(b) OUTSIDE			
			Sim1 R=100 I=5 Q=1 S=20		Sim2 R=100 I=5 Q=2 S=20		Sim3 R=100 I=5 Q=1 S=40		Sim4 R=100 I=5 Q=2 S=40		Sim1 R=1 I=5 Q=1 S=20		Sim2 R=1 I=5 Q=1 S=40	
			AP	T	AP	T	AP	T	AP	T	AP	T	AP	T
PAL DATABASE	LTR	K=20	0.5260	*0.00*	0.5342	*0.01*	0.4752	*0.01*	0.4872	*0.01*	0.4791	*0.01*	0.4062	*0.01*
		K=100	**0.5466**	*0.01*	0.5419	*0.01*	0.4983	*0.02*	0.5011	*0.03*	0.4774	*0.01*	0.4187	*0.01*
		K=200	0.5460	*0.03*	**0.5445**	*0.03*	**0.4998**	*0.05*	**0.5059**	*0.06*	**0.4985**	*0.01*	**0.4298**	*0.01*
	$P(W\|D)$	EC	**0.4562**	*0.02*	**0.4657**	*0.02*	**0.3954**	*0.04*	**0.3987**	*0.04*	**0.3973**	*0.01*	**0.3384**	*0.02*
		KL	0.4394	*1.60*	0.4400	*1.86*	0.3819	*3.15*	0.3828	*3.30*	0.3874	*0.65*	0.3273	*1.05*
		CS	0.4495	*0.04*	0.4530	*0.04*	0.3919	*0.07*	0.3898	*0.08*	0.3919	*0.02*	0.3358	*0.02*
		HL	0.4438	*0.23*	0.4487	*0.25*	0.3868	*0.42*	0.3897	*0.39*	0.3910	*0.10*	0.3305	*0.16*
		BC	0.4381	*0.11*	0.4338	*0.11*	0.3817	*0.18*	0.3826	*0.18*	0.3856	*0.04*	0.3275	*0.07*
		MF	0.3997	*0.12*	0.4169	*0.13*	0.3630	*0.11*	0.3715	*0.11*	0.3754	*0.05*	0.3305	*0.05*
	$P(Z=200\|D)$	EC	0.4134	*0.03*	0.3925	*0.03*	0.3306	*0.04*	0.3158	*0.04*	0.3089	*0.01*	0.2634	*0.01*
		KL	0.4898	*1.62*	0.5104	*1.91*	0.4197	*3.12*	0.4256	*1.82*	0.4195	*0.69*	0.3479	*1.06*
		CS	**0.5462**	*0.03*	**0.5834**	*0.04*	**0.4699**	*0.06*	**0.4941**	*0.06*	**0.4656**	*0.02*	**0.3914**	*0.03*
		HL	0.5102	*0.23*	0.5349	*0.26*	0.4344	*0.40*	0.4446	*0.39*	0.4365	*0.10*	0.3659	*0.17*
		BC	0.4978	*0.12*	0.5186	*0.14*	0.4194	*0.22*	0.4275	*0.22*	0.4209	*0.04*	0.3557	*0.07*
		MF	0.4032	*0.11*	0.4312	*0.11*	0.3445	*0.11*	0.3630	*0.11*	0.3895	*0.05*	0.3405	*0.05*
		ZH	0.3588	*0.03*	0.3582	*0.03*	0.3181	*0.03*	0.3173	*0.03*	0.4034	*1.34*	0.3344	*1.31*

Table 4.3: Retrieval result for CCV database: Average Precision (AP) and average seconds per query (T). For each group of ranking functions (in rows), the best AP value of each simulation is highlighted in bold and the best global value among all methods is underlined.

CCV DATABASE	METHOD		(a) INSIDE								(b) OUTSIDE			
			Sim1 R=500 I=5 Q=1 S=20		Sim2 R=500 I=5 Q=2 S=20		Sim3 R=500 I=5 Q=1 S=40		Sim4 R=500 I=5 Q=2 S=40		Sim1 R=1 I=5 Q=1 S=20		Sim2 R=1 I=5 Q=1 S=40	
			AP	T	AP	T	AP	T	AP	T	AP	T	AP	T
	LTR	K=100	0.1154	*0.07*	0.1313	*0.09*	0.1136	*0.13*	0.1275	*0.16*	0.1150	*0.02*	0.1111	*0.04*
		K=500	0.1496	*0.39*	0.1694	*0.42*	0.1529	*0.54*	0.1686	*0.65*	0.1473	*0.14*	0.1465	*0.23*
		K=1000	0.1597	*0.50*	0.1810	*0.58*	0.1716	*0.86*	0.1886	*0.99*	0.1664	*0.25*	0.1667	*0.41*
		K=1500	0.1860	*1.36*	0.2121	*1.25*	0.1935	*1.82*	0.2137	*2.07*	0.1793	0.47	0.1782	0.81
		K=2000	**0.1952**	*1.23*	**0.2198**	*1.38*	**0.1974**	*2.11*	**0.2163**	*2.37*	**0.1837**	0.53	**0.1824**	0.93
	$P(W\|D)$	EC	0.0964	*3.23*	0.0927	*3.32*	0.0786	*4.35*	0.0766	*4.74*	0.0782	*1.19*	0.0650	*1.87*
		KL	0.0708	*120*	0.0575	*108*	0.0688	*208*	0.0617	*215*	0.0840	*76.6*	0.0697	*124*
		CS	0.1111	*3.97*	0.1105	*4.52*	0.0924	*7.27*	0.0921	*9.54*	0.0922	*1.58*	0.0769	*2.50*
		HL	0.1001	*22.4*	0.0974	*28.3*	0.0827	*39.6*	0.0820	*40.8*	0.0859	*14.8*	0.0693	*22.9*
		BC	0.1005	*12.7*	0.0932	*13.7*	0.0821	*20.3*	0.0774	*21.3*	0.0838	*7.04*	0.0681	*10.8*
		MF	**0.1293**	*103*	**0.1390**	*103*	**0.1059**	*103*	**0.1121**	*103*	**0.1007**	*33.6*	**0.0796**	*33.6*
	$P(Z=2000\|D)$	EC	0.0710	*0.84*	0.0645	*1.03*	0.0556	*1.57*	0.0500	*1.53*	0.0433	*0.28*	0.0333	*0.42*
		KL	0.1423	*78.9*	0.1352	*85.6*	0.1165	*126*	0.1107	*132*	0.1078	*26.9*	0.0866	*43.5*
		CS	**0.2040**	*2.52*	**0.2344**	*2.76*	**0.1789**	*3.94*	**0.1978**	*5.17*	**0.1605**	*0.78*	**0.1386**	*1.31*
		HL	0.1748	*13.1*	0.1848	*14.7*	0.1427	*21.0*	0.1477	*23.6*	0.1426	*5.07*	0.1110	*8.24*
		BC	0.1684	*6.46*	0.1703	*7.20*	0.1380	*10.4*	0.1392	*11.7*	0.1361	*2.43*	0.1064	*3.89*
		MF	0.1059	*22.3*	0.1242	*22.3*	0.0776	*22.3*	0.0889	*22.3*	0.0706	*8.95*	0.0511	*8.95*
		ZH	0.1518	*539*	0.1707	*538*	0.1248	*539*	0.1371	*539*	0.1285	*450*	0.1028	*449*

4.4 Discussion

The first noteworthy point is the remarkable precision gains provided by topic models in the performed retrieval simulations. Comparing the best average precision value obtained in the original BoW space $P(W|D)$ with the best value among the seven ranking functions tested in the topic space $P(Z|D)$, we observe that in the topic space the precision is increased on average a 20.35% for the PAL database, 67.14% in the case of CCV and 21.10% for TRECVID. These significant precision gains support our statement that the hidden patterns provided by topic models are useful to fill the semantic gap in CBVR. Topic models have shown to help in many areas, such us text categorisation or image recognition, but in tasks where precision is important, like in CBVR, they have been traditionally considered useless. Some authors have this belief because the best ranking functions in the original BoW space tend not to work properly in the latent space. As we can see in the results, the best ranking functions in the original BoW space are EC (for PAL) and MF (for CCV and TREVID) but these two functions are often two of the worse in the latent space. However, the CS function is able to obtain a real precision improvement in the topic space. In fact, CS is the unique function

Table 4.4: Retrieval result for TRECVID database: Average Precision (AP) and average seconds per query (T). For each group of ranking functions (in rows), the best AP value of each simulation is highlighted in bold and the best global value among all methods is underlined.

TRECVID DATABASE	METHOD		(a) INSIDE								(b) OUTSIDE			
			Sim1 R=500 I=5 Q=1 S=20		Sim2 R=500 I=5 Q=2 S=20		Sim3 R=500 I=5 Q=1 S=40		Sim4 R=500 I=5 Q=2 S=40		Sim1 R=1 I=5 Q=1 S=20		Sim2 R=1 I=5 Q=1 S=40	
			AP	T	AP	T	AP	T	AP	T	AP	T	AP	T
	LTR	K=100	0.0920	*0.08*	0.0972	*0.09*	0.0951	*0.10*	0.0996	*0.12*	0.0867	*0.02*	0.0828	*0.03*
		K=500	0.1298	*0.22*	0.1349	*0.23*	0.1366	*0.28*	0.1375	*0.30*	0.1279	*0.11*	0.1319	*0.16*
		K=1000	0.1482	*0.43*	0.1529	*0.45*	0.1626	*0.56*	0.1666	*0.60*	0.1302	*0.23*	0.1389	*0.35*
		K=1500	0.1547	*0.64*	0.1538	*0.69*	0.1659	*0.85*	0.1676	*0.91*	0.1331	*0.35*	0.1418	*0.53*
		K=2000	**0.1553**	*0.86*	**0.1595**	*0.93*	**0.1698**	*1.23*	**0.1740**	*1.21*	**0.1354**	*0.47*	**0.1435**	*0.70*
	$P(W\|D)$	EC	0.1100	*0.73*	0.1091	*0.97*	0.1080	*1.21*	0.1066	*1.42*	0.0692	*0.58*	0.0663	*1.02*
		KL	0.1214	*36.6*	0.1171	*45.9*	0.1179	*59.8*	**0.1162**	*70.5*	0.0717	*34.3*	0.0695	*50.6*
		CS	0.1177	*0.97*	0.1164	*1.23*	0.1130	*1.58*	0.1103	*1.83*	0.0670	*0.71*	0.0654	*1.25*
		HL	0.1156	*6.86*	0.1111	*7.41*	0.1144	*9.75*	0.1092	*11.2*	0.0738	*4.93*	0.0710	*8.90*
		BC	0.1141	*2.93*	0.1104	*3.67*	0.1139	*4.75*	0.1083	*5.39*	0.0733	*2.41*	0.0709	*4.66*
		MF	**0.1521**	*47.7*	**0.1365**	*47.5*	**0.1332**	*47.6*	0.1158	*47.9*	**0.1046**	*3.72*	**0.0791**	*3.72*
	$P(Z = 2000\|D)$	EC	0.1001	*0.52*	0.0962	*0.67*	0.0957	*0.82*	0.0906	*0.97*	0.0979	*0.60*	0.0914	*0.80*
		KL	0.1272	*32.7*	0.1248	*40.5*	0.1233	*53.7*	0.1204	*62.1*	0.1069	*34.3*	0.1014	*50.6*
		CS	**0.1523**	*0.72*	**0.1603**	*0.97*	**0.1547**	*1.22*	**0.1556**	*1.45*	**0.1278**	*0.62*	**0.1253**	*1.10*
		HL	0.1322	*4.95*	0.1313	*6.14*	0.1261	*7.95*	0.1252	*9.32*	0.1132	*4.34*	0.1066	*7.81*
		BC	0.1298	*2.48*	0.1293	*3.12*	0.1247	*3.97*	0.1234	*4.63*	0.1116	*3.72*	0.1057	*3.84*
		MF	0.1441	*37.1*	0.1229	*37.1*	0.1320	*37.1*	0.1155	*37*	0.0929	*2.87*	0.0556	*2.87*
		ZH	0.1161	*266*	0.1276	*280*	0.1069	*275*	0.1148	*267*	0.0976	*305*	0.0840	*314*

which has shown to be effective in the latent space among the tested retrieval methods of the literature.

Another noticeable question related to the latent space is the adequate number of topics. We have tested several values for each database and the best precision results are obtained using the highest numbers, that is, 200 topics for PAL database and 2000 for both CCV and TRECVID. However, for PAL and TRECVID the precision improvement using these values is quite slight compared with the results obtained with 100 and 1500 respectively. This indicates that, depending on the database and the kind of queries, increasing the number of topics reaches a point in which it does not provide an actual improvement. Selecting the appropriate number of topics is an important question and still remains an open-ended issue in the literature. Even though the number of topics may significantly affect to the performance of a system, for this kind of application it is more important to have enough topics to obtain a fine granularity of patterns to describe queries than to find out exactly the optimum number. Somehow, it is similar to the case of finding the optimum number of clusters in a classification problem. As long as you have enough clusters to represent the classes it is not so

important if some classes are represented with more than exactly one cluster.

In general, the results show a similar trend in the three tested databases. The LTR function achieves the best retrieval precision on average compared with the best methods in both the original BoW space $P(W|D)$ and the latent space $P(Z|D)$. In the PAL database, LTR outperforms EC-pwd in a 23.38% and CS-pzd in a 2.52%. For CCV, the precision gain of LTR is a 79.19% over MF-pwd and a 7.22% over CS-pzd. In the case of TRECVID, LTR increases the precision of MF-pwd in a 29.58% and the precision of CS-pzd in a 7.00%.

Related to the parameters of the simulations, we observe the that average precision tend to increase using a bigger Q (number of samples to initialise queries) and it drops with a larger value of S (scope). The rationale behind this is the following: on the one hand, initialising queries with more samples provides more information about the concept of interest and then the retrieval systems is more effective. On the other hand, considering a larger scope makes the retrieval system use more samples which are less likely to belong to the query class what eventually generates a precision drop.

According to the results, the proposed LTR shows a good robustness regarding the parameters Q and S of the simulations. Focusing on CCV and TRECVID, the proposed LTR method obtains a similar precision gain to the best tested function (CS-pzd) when the parameter Q increases (Sim2-inside and Sim4-inside). In addition, LTR is able to reduce the precision drop compared with CS-pzd when the parameter S is increased (Sim3-inside, Sim4-inside and Sim2-outside).

The proposed LTR function is able to outperform the tested methods in the original BoW space and in the latent space with the exception of CS-pzd. For that reason, we discuss in more detail the differences between LTR and CS to highlight the advantages of the proposed approach.

4.4.1 Cosine Similarity (CS) vs. Latent Topic Ranking (LTR)

The CS function uses the cosine of the angle between two samples as a similarity measure. That is, the most similar documents to the query are those which have the lowest angle with respect to the query (angular similarity). In the case the query has more than one document, we have used the average cosine similarity value. Equation (4.15) shows the CS function where d represents a document of

the database and D' the query set.

$$Sim(d,D') = \frac{\sum_{d' \in D'} \cos(\theta_{d,d'})}{|D'|} = \frac{\sum_{d' \in D'} \frac{d \cdot d'}{|d||d'|}}{|D'|} \tag{4.15}$$

The proposed LTR function provides a probabilistic approach to discover the most likely samples according to the query. As it is shown in equation (4.16), this function can be interpreted as a weighted scalar product between the document d and a summary of the query set in a single document. The topics weights are computed as the inverse of the prior of each topic in the database. Therefore, the least used topics generate higher probability values, that is, the LTR function is a weighted scalar product which fosters the least used topics in the database. Intuitively, this makes sense because the least used patterns may allow us to discriminate better among samples for complex query concepts.

$$LTR(d,D') = \overbrace{\sum_z \underbrace{\left(\frac{1}{\sum_{d_i} p(z|d_i)}\right)}_{\text{topic weight}} \underbrace{(p(z|d))}_{\text{document}} \underbrace{\left(\sum_{d' \in D'} p(z|d')\right)}_{\text{query summary}}}^{\text{weighted scalar product}} \tag{4.16}$$

At the same time, the scalar product (dot product) between two vectors is directly proportional to the projection of the first on the second vector. That is, $(d \cdot d') = |d'|\ Projection_{d-d'}$. Logically, cosine similarity and LTR function have some similar features because the less angle often implies the more projection and then the more scalar product value. However, there is a main difference which enables the LTR function to overcome the cosine similarity retrieval precision. The weighting scheme gives more flexibility to the LTR function in order to deal with the semantic gap challenge.

Comparing both behaviours, the LTR function selects the documents within a margin of larger weighted scalar projection by fostering the least used topics. In a real application, this produces a top ranking with more variety of documents and then the user's feedback is able to provide more relevant information about the scope of the query concept. Let us see it through an example of gender retrieval using the PAL database. We are going to use an initial query of an elderly woman to compare the LTR and CS rankings at the first ranking iteration. Assuming queries from inside the PAL database, Figure 4.7 shows the differences between

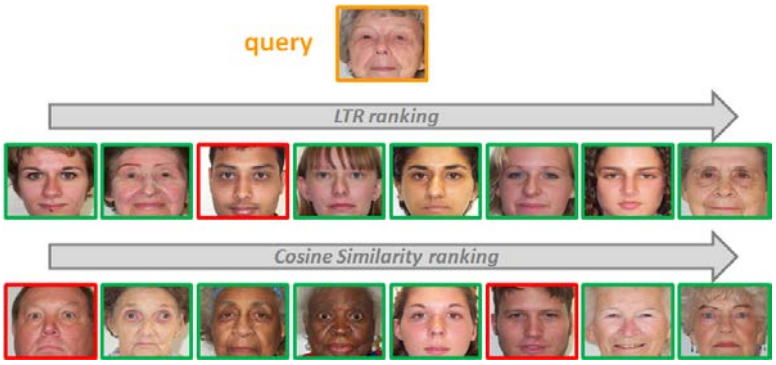


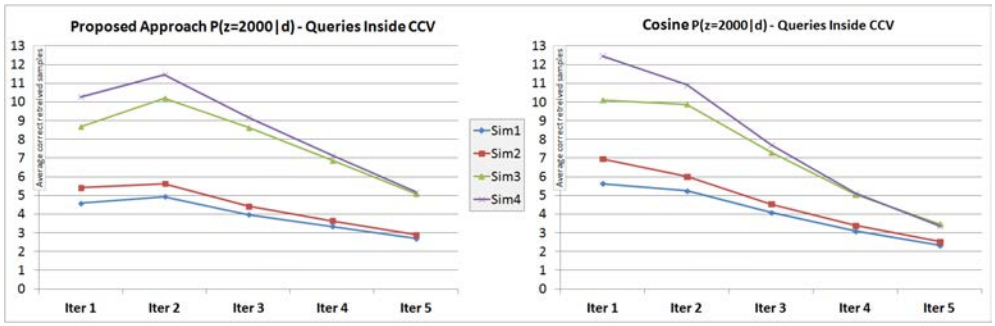
Figure 4.7: Example of gender retrieval. Cosine and LTR rankings at the first iteration with a scope of 20. The figure only shows the images which are different in both 20-top rankings (8 pictures). The images are sorted from left to right according to the original 20-top ranking order.

both 20-top rankings at the first iteration. The shared images have been removed to highlight the differences between the two ranking functions.

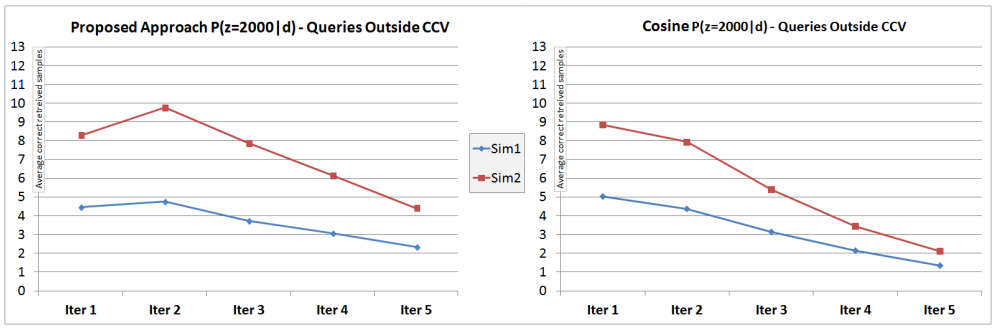
First of all, it should be noted the relationship between LTR and CS. A total of 12 images are the same in both 20-top rankings at the first iteration. Specifically, 8 white old women, 2 black women, one black man and another white man are shared by both rankings. This fact clearly shows the aforementioned relation between projection (LTR) and angular similarity (CS). However, we can appreciate a very important difference between the not overlapped images of both rankings. The cosine similarity function tends to retrieve samples of older women (the initial query) whereas LTR first retrieves women with different appearances. That is, the LTR function provides a broader kind of women images, thus the proposed approach is able to obtain a broader and more meaningful feedback about the query class.

As it was introduced in Section 5.1, the main problem in CBVR is the semantic gap challenge i.e. the difference between the user's understanding and the data representation. In CBVR, the same video sample can be related to very different concepts (queries) and the only way we have to distinguish among them is by the user's feedback. Therefore, enriching the query with a wide variety of positive examples in the feedback is a key factor to deal with unconstrained concepts.

Figure 4.8 shows the number of correctly retrieved videos per ranking iteration for the experiments using the CCV database. In both internal and external queries, we can see how the CS function archives the best performance at the first iteration and then the precision decreases in the subsequent iterations. However,



(a) Simulations when queries are from inside the CCV database.



(b) Simulations when queries are from outside the CCV database.

Figure 4.8: Proposed approach (LTR) vs Cosine Similarity (CS). Average of correctly retrieved samples per iteration for the performed video retrieval simulations using the CCV database.

LTR obtains the best performance at the second iteration after the first user's feedback and for the following iterations the precision drop is smoother than in the case of CS. This example shows that the feedback extracted from the LTR ranking contains a more useful information of the query class. Even though the number of positive samples is lower at the first iteration, the fostering of the least used topics made by LTR generates a user's feedback more meaningful because it includes samples with a broader variety of topics related to the query. Eventually, this variety of hidden patterns allows users to describe better the concept of interest though the feedback they provide.

4.4.2 Computational complexity issues

Regarding the computational burden, the results show a high performance of the LTR function with respect to the best tested methods in both $P(W|D)$ and $P(Z|D)$ spaces. LTR can process documents faster than the methods tested in the original BoW space $P(W|D)$ because the proposed function performs the ranking

in the topic-model space $P(Z|D)$ and this space has usually a lower dimensionality than the former. For instance, in the CCV simulations the original feature space with 5000 words is reduced by LDA to a topic space with 2000 topics what means a 60% dimensionality reduction. Comparing LTR with the methods tested in $P(Z|D)$, the proposed function is able to obtain a good computational time as well. Despite the fact that EC-pzd is more efficient than LTR, the precision of the Euclidean distance in the latent space is so poor that the single competitor of LTR is CS-pzd.

According to the results, LTR tends to outperform the computational time obtained by CS-pzd. The proposed LTR function (Eq. (4.16)) summarises the whole query set in a single document (query summary), then a single scalar product is performed for each sample to be ranked. That is, the cost of obtaining the score of a document is $O(|D'|K + K) = O(|D'|K)$ where $|D'|$ represents the size of the query set at a specific time moment and K the number of topics. Note that topic weights ($\sum_{d_i} p(z|d_i)$) in Eq. (4.16) are computed off-line. The CS function (Eq. (4.15)) uses the average cosine value for all the documents of the query, therefore it needs to compute $|D'|$ scalar products, two magnitudes and a query cardinality per document to rank. That makes a total cost of $O(3|D'|K + |D'|) = O(|D'|K)$. The asymptotic cost of both functions is the same but in practice LTR is able to achieve a better computational time because of the multiplicative constants.

Table 4.5: Computational time of LDA and EMS for the CCV database.

LDA (default parameters)				**EMS (default parameters)**			
	K	**Time**	**RAM**	**CCV - tst**	**AVG Time per Doc**	**EM Iters**	**MAX Iters**
CCV (tra + tst)	100	2 days	0.40 GB	$P(z = 100\|d')$	0.63 sec	475.55	2.23%
	500	8 days	1.20 GB	$P(z = 500\|d')$	3.68 sec	548.83	4.34%
	1000	15.5 days	2.20 GB	$P(z = 1000\|d')$	9.05 sec	662.76	5.49%
	1500	23 days	3.20 GB	$P(z = 1500\|d')$	15.24 sec	668.72	6.26%
	2000	30.5 days	4.20 GB	$P(z = 2000\|d')$	22.31 sec	711.14	6.94%

The average computational time per query shown in the results corresponds to the cost of the ranking function itself, that is, the stage (iii) of Fig. 4.4. However, the RF scheme contains two more procedures that we should take into account: LDA in stage (i) and EMS in (ii). Table 4.5 presents the computational time of both procedures for the CCV database. In the case of LDA, we use a parallel version running in 24 Intel Xeon E5-2640 processors and in the case of EMS a single processor Intel Xeon E5-2640. As we can see, the topic extraction task is a very time-consuming process. Although LDA runs off-line, its cost may limit its usage in much larger databases. However, the proposed LTR function is in-

dependent of the topic extraction algorithm, therefore further improved methods could be used instead of LDA.

Related to the EMS function, the more topics the more costly the process. In the case of CCV, the average time to represent an external query document in 2000 topics is over 20 seconds what seems noticeable higher compared with the costs of the ranking functions. However, this cost has to be taken as a pre-processing step although it is part of the ON-LINE - Query Definition stage. In the case of external queries, a pre-processing step is always required to represent those external samples in the same way the database was encoded in visual Bag of Words. Finally, note that these computational disadvantages of LDA and EMS are not exclusive for the proposed LTR function but for all the retrieval methods running in latent topic spaces.

In addition to computational time, Table 4.5 shows the convergence average values of the EMS function for the CCV database. As we can see the average number of EM iterations per document is below the considered default limit of 1000 and besides there is a small percentage of documents which actually reach this limit in the convergence process.

4.4.3 Limitations of the proposed approach

Although the presented LTR function has shown to outperform the rest of the tested methods, there are two points which have to be taken into account: (i) the topic extraction cost and (ii) the patterns diversity provided by LTR. Related to the first point, current topic extraction algorithms are still very costly and more research in that field is required to enable processing video collections with millions of videos. This is is not a limitation specifically of LTR but it is a drawback of all the ranking functions working in the latent topic space. However, the proposed approach has been designed isolating the off-line topic extraction process from the on-line retrieval task. This makes that further improvements on the topic extraction methods can be directly used by replacing LDA to extract the topics. The second point to be considered is the patterns diversity provided by LTR. The proposed retrieval approach has been designed assuming a wide semantic gap to deal with by means of a RF scheme, that is the typical situation in CBVR. As we have shown, the topic diversity provided by LTR at the top-ranking is able to provide a competitive advantage because it may obtain a more informative feedback. However, this diversity is only useful when there is feedback

itself because the user discards those samples related to useless patterns. That limits the effectiveness of LTR in those situations where there is not feedback at all. As we can see in Fig. 4.8, the precision gains of LTR over CS are obtained after the first user's feedback. That is, CS obtains a better precision than LTR at the first ranking iteration where there is not feedback just the query initialisation.

4.5 Conclusions

In this chapter, we have presented a novel interactive retrieval approach addressing the retrieval problem as a class discovery problem using latent topics. The sSpLSA model has been introduced to deduce the LTR probabilistic ranking function and the EMS procedure has been defined to enable external queries. Later, we have defined the proposed retrieval framework based on short-term relevance feedback. Finally, several retrieval simulations have been performed using three different datasets (PAL, CCV and TRECVID) and several of the most relevant retrieval methods in the literature.

One of the main conclusions that arises from the chapter is the importance of topic models to deal with the semantic gap in CBVR. Although topic models have shown to be helpful in many areas, they have not been traditionally considered useful in CBVR because of the special nature of the latent space. However, this work shows that (i) the hidden patterns defined by topics can be effectively used in video retrieval tasks and (ii) the proposed LTR ranking function is able to outperform the rest of the tested functions mainly because it has the same probabilistic nature than topic models.

The results of the chapter provide evidences about the viability of the proposed approach in terms of effectiveness and efficiency to deal with the semantic gap challenge in the CBVR field. In this domain, two users could provide the same query initialisation but referring to two different query concepts because each one is focusing on a different aspect. As a result, the feedback quality is an essential issue to find out about the query concept. As we have shown, the proposed LTR function promotes the least used topics and then it enriches the top-ranking with a more variety of related hidden patterns what eventually produces a more meaningful feedback.

Although results are encouraging, much more progress is needed to really address the semantic gap problem which involves several fields, from low level

descriptors to high level understanding and user interaction. Specifically, further work is directed to extend the work in the following directions:

- Automatic strategies to set the most appropriate number of topics.
- Extension of the retrieval model to a long-term RF approach.
- Reduction of the computational time of the topic extraction task by applying quantisation methods in the initial object (video) space.

Chapter 5

Incremental probabilistic Semantic Analysis for Video Retrieval

Publication

[25] Ruben Fernandez-Beltran and Filiberto Pla. Incremental probabilistic latent semantic analysis for video retrieval. *Image and Vision Computing*, 38:1–12, 2015.

Recent research trends in content-based video retrieval have shown topic models as an effective tool to deal with the semantic gap challenge. In this scenario, this chapter has a dual target: (1) it is aimed at studying how the use of different topic models (pLSA, LDA and FSTM) affects video retrieval performance; (2) a novel incremental topic model (IpLSA) is presented in order to cope with incremental scenarios in an effective and efficient way. A comprehensive comparison among these four topic models using two different retrieval systems and two reference benchmarking video databases is provided. Experiments revealed that pLSA is the best model in sparse conditions, LDA tend to outperform the rest of the models in a dense space and IpLSA is able to work properly in both cases.

5.1 Introduction

With the expansion of new technologies, video collections are increasingly larger and more complex, therefore one of the biggest current challenges is how to retrieve users' relevant data from this huge amount of information. The Content-Based Video Retrieval (CBVR) problem is concerned about how to provide users with videos which satisfy their queries by means of video content analysis. Over the past years, CBVR has become a very important research field and several CBVR systems have been developed [1, 18, 36, 78]. In general, a CBVR system has three main components involved in the retrieval process: (1) a query, represented by a few video examples of the semantic concept the user is looking for; (2) a database, which is used to extract videos related to the query concept; and (3) a ranking function, which sorts the database according to the relevance to the query. These three components are usually integrated together with the user in a Relevance Feedback (RF) scheme [15] to provide the most relevant videos through several feedback iterations.

One of the most used rankings in multimedia retrieval is distance-based ranking. Such ranking is performed according to the minimum distance or maximum similarity to the query in the video representation space [3, 39]. However, these measures tend not to work properly when the multimedia data is rather complicated [52]. Other ranking algorithms are based on inductive learning [64, 65] which typically use a bank of classifiers to represent the set of possible events to test. Nevertheless, the performance of this approach heavily depends on the training data what limits its usage in unconstrained retrieval applications. Alternative ranking methods are based on transductive ranking which use the topology of the data distribution to improve the output ranking [76, 81]. The main drawback of these functions is their high computational cost because they need to carry out demanding matrix operations over the retrieval process.

Several of these approaches have shown to be successful at retrieval tasks when they are used on reduced databases with a small number of concepts [58]. However, the so-called semantic gap [57] between computable low-level features and query concepts is still a challenge for large unconstrained video collections. The visual variability of unconstrained queries is so high that current approaches often do not adequately scale semantic concepts [52]. As a result, new capabilities are required in CBVR to bring the video characterization to a higher semantic level.

Ranking functions work in a specific representation space where videos are encoded in feature vectors according to the information provided by a descriptor. Different types of descriptors have been developed using static information (Scale Invariant Feature Transform - SIFT [43]), spatio-temporal (Spatial Temporal Interest Points - STIP [35]) or audio (Mel Frequency Cepstral Coefficients - MFCC [19]). The standard procedure to encode all this low-level information in feature vectors is the visual Bag of Words (vBoW) [54]. The vBoW quantization starts by learning a visual vocabulary made up of the clustering of the local features. Then, each video is represented in a single histogram of visual words by accumulating the number of local features into their closest clusters. In the literature, it is quite common to see how authors refer to this quantized space as descriptor space although it is not the direct output of the descriptor functions

Some recent works have presented more advanced descriptors which are able to achieve better results for a specific sort of applications. For example, in [71] Wang and Schmid presented a video representation based on dense trajectories specially designed for action recognition which outperforms the most common motion-based descriptors. However, in unconstrained CBVR the type of concepts to deal with is so wide that simpler and non-specialised descriptors are commonly used [52].

Although early research on topic models suggested that they may be used in video retrieval, it was not until recently that topic models were successfully applied to large unconstrained video collections [23]. In general, topic models provide for automatically organizing, understanding, searching and summarizing large electronic archives [9]. For many years, topic models have not been considered useful in tasks where precision is important because traditional ranking functions tend to perform worse in the latent space than in the original characterisation space. The latent topic space is usually a lower dimensionality representation where concepts and classes are more diffuse and besides it allows connections among different concepts through patterns defined by topics. As a result, the most effective ranking functions in the original feature space are usually not useful in the topic space because this space has an utterly different nature.

However, this fact does not mean the topics' lack of usefulness. In those applications in which the semantic gap is important, the retrieval precision in the original feature space tend to be very low and topic models can provide a competitive advantage by means of hidden patterns which may be interpreted

as a higher characterization level. It is the case of unconstrained CBVR, where the difference between the low-level characterization of the videos and the query concepts that users can manage is so huge that topic models can help us to obtain a better performance in retrieval tasks.

The majority of the topic methods are in the families of two reference models: probabilistic Latent Semantic Analysis (pLSA) [29] and Latent Dirichlet Allocation (LDA) [8]. These two algorithms and other topic models are typically used by retrieval systems in three steps: (1) Extract the hidden patterns (topics) that pervade the data collection; (2) Annotate the documents according to these topics; (3) Use these annotations to rank the documents according to users' queries. The topic extracting process has shown to be affordable when it is carried out in moderate size databases with a limited number of concepts. However, current video collections tend to be very large and besides they grow day by day with a wide range of concepts. For these incremental databases, topic extraction algorithms such as pLSA and LDA, have a computational burden too heavy to recompute topics each time the databases increase their size with new samples. In other kinds of applications, some authors [16] have shown the advantages of considering an incremental scenario to manage large video collections in an efficient way, therefore this scheme may help us to improve the topic extraction task. In this chapter, we are interested in exploring whether video retrieval performance is affected by the use of different topic models and how video retrieval systems based on topic models are able to efficiently manage these incremental databases.

In the literature, several alternative models have been proposed in order to improve the computational efficiency of the topic extraction process. Some authors have proposed dynamic models which are able to adapt topic structure over time. One of the most representative ones is presented in [10] where Blei and Lafferty developed the Dynamic Topic Model which can capture the evolution of topics in a sequentially organized corpus of documents. Other authors have developed window-based models where the database is considered a temporal flow in which old documents are removed as new documents are introduced. For instance, Tzu-Chuan et al [17] presented a pLSA version to address the problem of on-line event detection and Wu et al [75] developed a pLSA extension for automatic question recommendation. In general, these models follow the same idea than that of dynamic models but allow the management of new words in documents. Dynamic models as well as window-based models use the concept *incremental* in the sense

of changing word distribution of topics over time, that is, they maintain the number of topics fixed and adapt these topics to the new samples. However, in an incremental retrieval environment the new samples may require additional topics to capture new patterns for retrieving these new samples. This fact makes these models unsuitable for an incremental retrieval scenario and in this chapter we use the concept *incremental* in the sense of extending the number of topics by adding new patterns.

Traditional topic models assume that topics have a non-zero contribution to generate documents and this leads to a dense representation with a high computational complexity. Other authors have proposed more efficient approaches which assume sparse topic proportions in documents. In [63], Khoat and Bao presented the Full Sparse Topic Model (FSTM) which is able to reduce significantly the computational burden with respect to pLSA and LDA. Although experimental results in [63] are encouraging, there are not works in the literature which have tested the performance of FSTM in a video retrieval system based on latent topics.

In this scenario, the presented work has a dual target. On the one hand, we pretend to study the performance of pLSA, LDA and FSTM models for the unconstrained video retrieval problem. On the other hand, we present an extension of the pLSA model in order to enable CBVR systems based on latent topics to handle incremental collections in an effective and efficient way. Some works [34, 44, 77] have already explored topic performance but always related to text or image retrieval, in this case we would like to test if the same behaviour can be observed in an unconstrained video retrieval system. In particular, we are going to use as a testing protocol two different retrieval systems based on latent topics: (1) the retrieval method proposed in [23] and (2) the cosine similarity function used in [80].

The rest of the chapter is organized as follows. In Section 5.2, a short review about topic models is provided mainly focused on pLSA and the reasons to extend this model rather than any other. Section 5.3 presents the Incremental probabilistic Latent Semantic Analysis (IpLSA) model which is an extension of pLSA in order to reduce computational complexity and to deal with the overfitting problem. In Section 5.4, the experimental setting is described as well as the empirical results obtained by the retrieval systems [23] and [80], including a comparison among pLSA, LDA, FSTM and IpLSA in terms of video retrieval

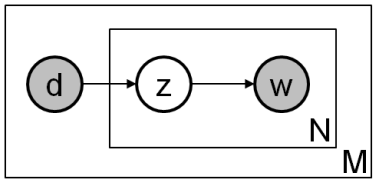


Figure 5.1: pLSA model: d represents the documents, z the topics (hidden variable) and w the words. M is the number of documents of the collection and N the number of words in the document d.

performance using the Consumer Columbia Video database [32] and the collection TRECVID 2007 [7] . Finally, Section 5.5 discusses the results and Section 5.6 draws the main conclusions arising from this chapter.

5.2 Background

Latent Semantic Analysis (LSA) [21] was one of the starting points for a group of techniques aimed at mapping the original high dimensional representation of data into a reduced representation, the so-called latent semantic space, where it is supposed that objects (documents, speech, images, videos ...) will represent semantic relationships among them. LSA had an algebraic interpretation of the latent semantic space, using a Singular Value Decomposition (SVD) approach to find such a representation. Probabilistic Latent Semantic Analysis (pLSA) [29] was later introduced by Hofmann, which is based on a statistical approach, defining a semi-generative data model and introducing a latent context variable associated with the different word polysemy occurrences. In pLSA (Figure 5.1), each document d is modelled as a mixture of topics z. The generative process is made as follows: (1) Select a document d with probability $p(d)$; (2) Pick a latent class z with probability $p(z|d)$; (3) Generate a word w with probability $p(w|z)$.

Statistical topic models have become an important data analysis tool, and pLSA has been developed in more general frameworks. Blei et al. introduced the Latent Dirichlet Allocation (LDA) model [8] which represents documents as a multinomial of topic mixtures generated by a Dirichlet prior. Both pLSA and LDA are a reference in topic modelling although there are significant differences between them. On the one hand, pLSA uses the documents of the collection as parameters of what makes the model pLSA a highly spatial demanding model and generates topic over-fitting when too many parameters are considered. On the other hand, LDA tries to overcome pLSA drawbacks by using two Dirichlet

distributions, one to model documents $p(z|d) \sim Dir(\alpha)$ and another to model topics $p(w|z) \sim Dir(\beta)$. Logically, these parameters α and β have to be estimated during the topic extraction process which adds an extra computational burden.

Although the experimentation in [8] shows that LDA is able to achieve lower perplexity than pLSA, it is not clear how the perplexity correlates with the performance in retrieval tasks and other kind of applications. The same Blei [14] concludes that pLSA often obtains a topic structure more correlated to the human judgement than LDA, even though the perplexity values suggest the opposite. The work presented in [34] reveals that pLSA outperforms the performance of LDA for automatic essay grading tasks in a collection with less than 150 documents. In [44], the authors suggest that LDA does not have a competitive edge over pLSA especially for small training datasets and other authors [77] conclude that more elaborated topic models provide no additional gains in retrieval tasks.

As a result, it seems that the pLSA scheme may enable to adapt the topics to the data distribution better when few samples are available according to the complexity of the problem. In the standard LDA algorithm, the parameter estimation is carried out by maximizing the marginal log-likelihood of the data using a tractable lower bound. In practice, this estimation is performed by iterating over the document collection what produces that LDA requires a certain number of documents to adequately estimate its hyper-parameters. In an application like CBVR, the concept to retrieve is a priori unknown because it is up to the user and besides the initialization and feedback are often very limited. Then, it is usual to deal with complex concepts having very little information about them. For these reasons, we have decided to extend the pLSA model as the basis of our incremental model for CBVR.

5.2.1 Computational complexity issues

One of the most important drawbacks of topic models is the computational complexity of their algorithms. In this section, we are going to have a look at the computational cost of the original pLSA algorithm [29] in order to figure out the best way to extend the model efficiently.

The pLSA implementation of Hoffmann [29] uses the Expectation Maximization (EM) algorithm. EM alternates into two steps: E-step (expectation) where the posterior probability of topics (z) given documents (d) and words (w) $p(z|d,w)$ is calculated, and M (maximization) which maximizes the complete log-likelihood

that depends on the posterior computed in the E-step. Therefore, the complexity of the standard pLSA algorithm is the following:

$$C_{time}(pLSA) = O(\underbrace{I}_{Iters}(\underbrace{VMK}_{Estep}+\underbrace{VMK}_{Mstep}) = O(IVMK) \tag{5.1}$$

$$C_{space}(pLSA) = O(\underbrace{VMK}_{p(z|d,w)}+\underbrace{VK}_{p(w|z)}+\underbrace{KM}_{p(z|d)}) = O(VMK) \tag{5.2}$$

where I is the maximum EM iterations, V the size of the vocabulary, M the number of documents and K the number of topics. According to these expressions, we can improve the computational complexity of the model by reducing any of these variables, but we have to analyse the best option according to our aims.

The maximum number of EM iterations (I) is a pre-fixed value which is typically set at 1000 by default and a lower value may produce a worse convergence of the algorithm, then taking a lower value does not seem to be a good alternative. Another possibility of reducing the complexity of pLSA could be by reducing the number of topics K. Choosing the right number of topics is a critical question in topic modelling and there are several works which deal with this problem. Some approaches are based on non-parametric topic models, such as the case of the Hierarchical Dirichlet Processes [62], and other ones use an evaluation function to decide the best number of topics [5]. However, all of them require performing the topic extraction process several times and therefore they are not practical in improving the efficiency of the topic extraction process. In order to simplify, we are going to assume that the number of topics K is set manually following a specific criterion, for instance a percentage of the total number of documents M.

Reducing the number of words of the vocabulary could be another option to improve the efficiency of the pLSA model. In fact, we explored vocabulary reduction in a previous work [27] where we used the LDA model to reduce the vocabulary size and that reduction allowed us to carry out the topic extraction process faster. However, reducing the vocabulary may not be enough especially when the number of documents increases dramatically.

With a huge number of documents, the pLSA model has two main drawbacks: the high spatial complexity and the over-fitting problem. By reducing the number of documents to extract the topics, we can try to cope with these two issues at the same time. On the one hand, the less documents the less parameters, and

then the less spatial complexity. On the other hand, by using less parameters the model is supposed to avoid part of the over-fitting produced when all the documents of the collection are considered parameters. Note that the pLSA-based models always have over-fitting because documents are parameters of the model, but using less parameters may allow us to avoid part of it.

Therefore, reducing the number of documents seems to be the best option to improve the efficiency and to obtain a better performance of a pLSA-based model. In an incremental environment, a CBVR system based on latent topics starts from an initial stage where it has a set of initial M_0 documents expressed as $p(w|d_0)$, a set of initial Z_0 topics $p(w|z_0)$ and the description of the documents in these topics $p(z_0|d_0)$. For the next stage, a set of M new documents $p(w|d)$ arrives into the database and topics must be recomputed to take into account the new data distribution. Normally, the amount of new samples will be quite lower than the number of samples of the previous stage ($M_0 << M$), therefore if the initial topics could be expanded using only the new documents the process would reach a great efficiency improvement. Precisely, the proposed incremental model follows that idea.

5.3 Incremental probabilistic Latent Semantic Analysis (IpLSA)

At a given stage of the retrieval process, an incremental database has three main components: a set of previous documents d_0, a set of topics z_0 extracted from the previous documents and a set of new documents d to extend the database. The goal of the proposed incremental model is to extract a new set of topics z using only the new documents d but taking into account the initial topics z_0 in order to extract only new patterns. In the end, these new documents will be represented using a combination of previous topics z_0 and new topics z.

The IpLSA model (Figure 5.2) extends the pLSA model by adding the random variable corresponding to topics z_0 of the previous stage. The generative process of the IpLSA model stems from the document probability distribution $p(d)$ of the new documents. In the model, documents $d;d = 1,...,M$ are expressed as topic mixtures of previous topics $z_0;z_0 = 1,...,Z_0$ and new topics $z;z = 1,...,Z$, according to parameters $p(z_0,z|d)$. Therefore, the process to generate a document d can be interpreted as follows:

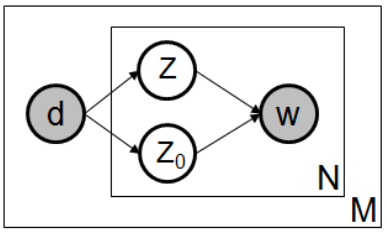


Figure 5.2: IpLSA model: d represents the new documents to add into the database, z_0 the initial topic structure of the previous stage, z the new extracted topics to describe the new documents and w the words. Eventually, N represents the number of words of the document d and M the number of new documents to add into the database.

- A document d is chosen from $p(d)$ probability distribution.
- For each one of the N words in the document d,
  - A topic pair (z_0,z) is chosen according to conditional distribution $p(z_0,z|d)$ that expresses documents in the previous topics z_0 and the new ones z.
  - A word w is chosen according to the conditional distribution $p(w|z_0,z)$ which expresses the set of previous and new topics in words.

5.3.1 Formulation by EM

The parameters $p(w|z)$, $p(z|d)$ and $p(z_0|d)$ of the IpLSA model can be estimated by maximizing the log-likelihood using an Expectation-Maximization (EM) algorithm. In particular, let us define first the joint distribution of the model Eq. (5.3) and later the log-likelihood Eq. (5.4) in terms of the joint probability distribution:

$$p(w,d,z) = p(w|z,z_0)p(z,z_0|d)p(d) \tag{5.3}$$

$$\mathcal{L} = \sum_{w}\sum_{d} n(w,d)\log p(w,d) \tag{5.4}$$

where $n(w,d)$ is the number of occurrences of the word w in the document d. In order to maximize the log-likelihood by EM, the complete log-likelihood can be expressed using the latent variables z and z_0 as:

$$E = \sum_{w}\sum_{d} n(w,d)(\mathcal{Z} + \mathcal{Z}_0) \tag{5.5}$$

$$\mathcal{Z} = \sum_{z} p(z|w,d)\log[p(w|z)p(z|d)p(d)] \tag{5.6}$$

$$\mathcal{Z}_0 = \sum_{z_0} p(z_0|w,d)\log[p(w|z_0)p(z_0|d)p(d)] \tag{5.7}$$

Introducing the normalization constraints of the parameters $p(z|d)$, $p(z_0|d)$ and $p(w|z)$ in expression (5.5) by inserting the appropriate Lagrange multipliers α and β:

$$H = E + \sum_{z}\alpha\left[1 - \sum_{w} p(w|z)\right] + \sum_{d}\beta\left[1 - \left(\sum_{z} p(z|d) + \sum_{z_0} p(z_0|d)\right)\right] \tag{5.8}$$

Taking derivatives with respect to the parameters, setting them equal to zero and solving the equations to isolate each parameter:

$$p(z|d) = \frac{\sum_{w} n(w,d)p(z|w,d)}{\sum_{z}\sum_{w} n(w,d)p(z|w,d) + \sum_{z_0}\sum_{w} n(w,d)p(z_0|w,d)} \tag{5.9}$$

$$p(z_0|d) = \frac{\sum_{w} n(w,d)p(z_0|w,d)}{\sum_{z}\sum_{w} n(w,d)p(z|w,d) + \sum_{z_0}\sum_{w} n(w,d)p(z_0|w,d)} \tag{5.10}$$

$$p(w|z) = \frac{\sum_{d} n(w,d)p(z|w,d)}{\sum_{w}\sum_{d} n(w,d)p(z|w,d)} \tag{5.11}$$

For the E-step, we need to estimate the parameters $p(z|w,d)$ and $p(z_0|w,d)$. Applying the Bayes' rule and the chain rule, we obtain:

$$p(z|w,d) = \frac{p(w,d,z)}{p(w,d)} = \frac{p(w|z)p(z|d)}{\sum_{z} p(w|z)p(z|d) + \sum_{z_0} p(w|z_0)p(z_0|d)} \tag{5.12}$$

$$p(z_0|w,d) = \frac{p(w,d,z_0)}{p(w,d)} = \frac{p(w|z_0)p(z_0|d)}{\sum_{z} p(w|z)p(z|d) + \sum_{z_0} p(w|z_0)p(z_0|d)} \tag{5.13}$$

The EM process is performed as follows. First of all, the set of new documents $p(w|d)$ and the set of previous topics $p(w|z_0)$ are loaded. Secondly, $p(w|z)$, $p(z|d)$ and $p(z_0|d)$ are randomly initialized. Then, the E-step (Eqs. (5.12) and (5.13)) and the M-step (Eqs. (5.9) and (5.10)) are alternated until a convergence condition is reached. As default settings to converge, we have used a threshold of 10^{-6} in the difference of the log-likelihood (equation (5.4)) between two consecutive iterations and a maximum of 1000 EM iterations.

5.3.2 Relation between IpLSA and pLSA

The proposed IpLSA model has a similar basis to pLSA, however IpLSA provides some novelties which may be interesting for incremental CBVR. In [29], Hofmann proposed a folding-in strategy to estimate the representation of new documents given a set of topics. Mainly, this strategy fixes the parameter $p(w|z)$ of the EM formulation in order to estimate only $p(z|d)$. The proposed IpLSA model follows a similar idea but was used in a different manner. Specifically, IpLSA makes a kind of combination of folding-in from previous topics and a regular pLSA for new topics at the same time. In contrast to pLSA, the proposed model manages the initial topics z_0 and the new ones z simultaneously, which enables the connection between previous and new patterns via the Lagrange multiplier β in Eq. (5.12). This connection is aimed at fostering the unseen patterns of the data in order to avoid extracting redundant topics. In other words, the proposed model allows us to learn only new patterns from the data, it does not matter if these patterns are refining a previous concept or they are related to a completely new one. The standard pLSA model does not have the capability to take into account knowledge of previous stages, however IpLSA takes advantage of incremental scenarios to reduce the number of parameters of the model and to extract only new patterns.

The incremental IpLSA model tries to reduce the over-fitting problem of the

global pLSA usage in two ways: (1) using only the new documents d to extract the new set of topics z and (2) avoiding learning topics which have been extracted in the previous stage. The standard pLSA uses the documents of the collection as parameters of the model, as a result the model may over-fit when too many parameters are considered. Assuming an incremental scenario, IpLSA extract the new topics only using the set of new documents, therefore the incremental IpLSA uses less parameters than the global pLSA and then it is avoiding part of the over-fitting produced in the global pLSA approach.

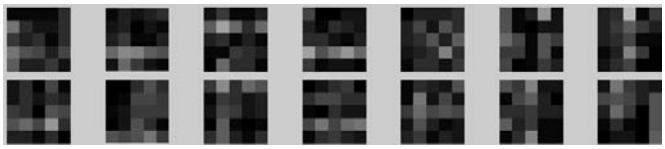
5.4 Experiments

This section presents the experimental part of the chapter. First (Section 5.4.1), we use a synthetic dataset in order to highlight how the proposed method works. Subsequently, Section 5.4.2 shows the performances of the IpLSA, pLSA, LDA and FSTM models specially applied to CBVR using two different video databases and several configurations.

5.4.1 Toy Dataset

The toy dataset [28] consists of 1000 gray level images with a size of 5×5 pixels. The samples have been generated synthetically according to the LDA model from a set of 10 topics (Figure 5.3a) which are distributed over each row and column. The vocabulary is a collection of 25 pixels in the images and the value of a pixel is the number of occurrences of a word in the document. Figure 5.3b shows some examples of the generated images. Note that words tend to co-occur along the same row or column.

(a) True topics used to generate the dataset.



(b) Some random images.

Figure 5.3: Toy Dataset.

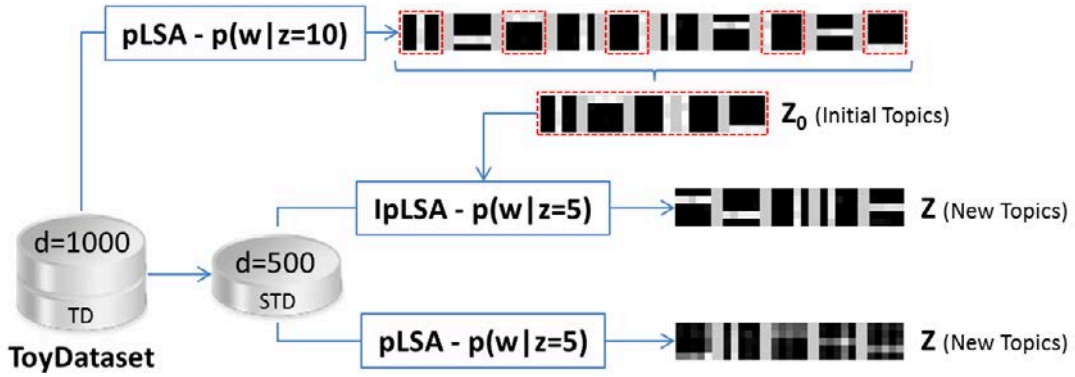


Figure 5.4: IpLSA vs pLSA.

Let us start by showing the behavioural differences between pLSA and IpLSA by means of Figure 5.4. We have used the following notation: TD_{1000} for the whole toy dataset made up of 1000 images and STD_{500} for a random subset of 500 samples. Extracting 10 topics over TD_{1000} by pLSA, we can obtain the topics which have generated the data (true topics). Note that these topics are completely precise and clean patterns. However, if we extract 5 topics by pLSA over STD_{500} we can observe that the obtained topics are a kind of combination of the true topics because the number of extracted topics is not adapted to the real number of patterns of the data. The idea with IpLSA is to avoid extracting topics which have been extracted in a previous stage. For example, if we think in an incremental scenario in which we have the initial topics z_0 and the set of new documents STD_{500}, IpLSA is able to extract only those new patterns which are not contained in z_0 (see Figure 5.4).

Another practical consideration is the difference between pLSA-based models and LDA. In figure 5.5, we can see the result of extracting 10 topics by pLSA and LDA over six subsets of the toy dataset. Each subset contains a different number of random images, from 25 samples to 1000. As we use more samples to extract the topics, we can see how pLSA is obtaining more precise topics, in particular with 250 documents pLSA obtains quite clearly the true topics. However, with LDA we can see that 250 samples are not enough to obtain a clear topics because with this number of samples the Dirichlet parameters are not well estimated yet. In this case, LDA requires 1000 documents to fit the parameters of the Dirichlet distributions. This fact has been reported in some previous works such as in [34, 44, 77]. Therefore, despite the fact that LDA provides a more general framework than pLSA, in some applications in which we do not have too much

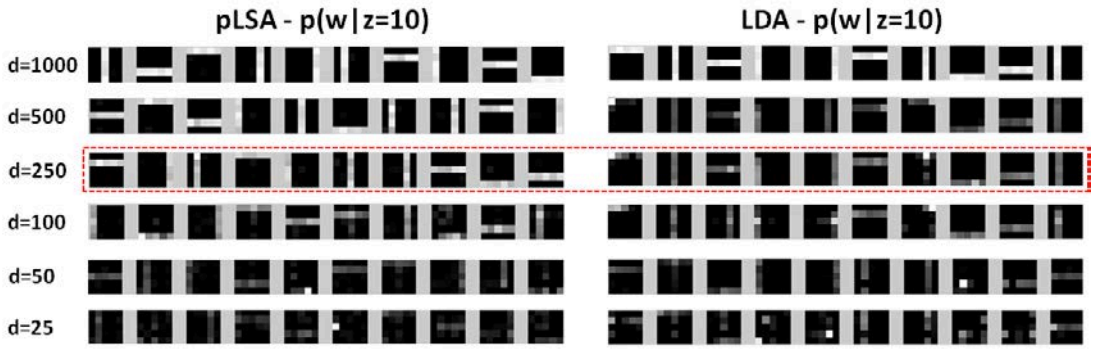


Figure 5.5: pLSA vs LDA.

information about the structure of the data, pLSA is able to extract the topics more accurately than LDA because it does not need any parameter estimation. In CBVR, we usually have to deal with complex query concepts having a few examples of this concept, therefore we think it makes sense to base our extension on pLSA rather than LDA for this kind of application.

5.4.2 Content-Based Video Retrieval

This section contains the experimental settings and the obtained results of IpLSA, pLSA, LDA and FSTM specially applied to the video retrieval problem using two different video databases.

Relevance Feedback simulations

In order to evaluate the effectiveness of the considered topic models for CBVR, we use the Relevance Feedback scheme proposed in [23] with two different ranking functions: the probabilistic ranking function presented in [23] and the cosine similarity function used in [80]. In that RF scheme, a simulation has four main parameters: Q the number of samples of the initial query, S the number of top examined items in each feedback iteration, I the number of total iterations and R the number of times which is the repeated random initialization of the query. According to these parameters, we propose the retrieval scenarios shown in Table 5.1.

Starting from a specific labelled retrieving set, the target of each simulation is directed to retrieve samples of a specific class but without using any class label information. The initial query is initialized with Q samples of a single class c

Table 5.1: Scenarios for the retrieval simulations.

Scenario	R	Q	I	S
1	100	1	5	20
2	100	2	5	20
3	100	1	5	40
4	100	2	5	40

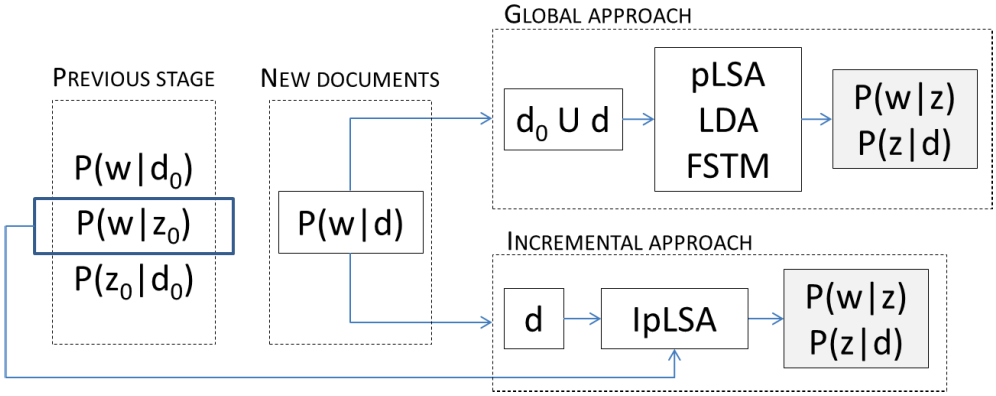


Figure 5.6: Stages used for the experiments.

and then the simulation process has to retrieve samples of that class through I feedback iterations using the Latent Topic Ranking (LTR) function proposed in [23] and the cosine similarity function used in [80]. At each iteration, the S top ranked items are inspected by a simulated user who marks the samples of the class c (positive samples). These positive samples are computed as correctly retrieved samples and they are used to expand the query. Finally, this expanded query is triggered as a new query with more examples for the next iteration.

Our objective is to compare the retrieval performance and the computational time among pLSA, LDA, FSTM and IpLSA in an incremental environment. The database starts from a previous stage when it has a set of initial documents $p(w|d_0)$, a initial set of topics $p(w|z_0)$ and the representation of the initial documents in the initial topics $p(z_0|d_0)$. Then a set of new documents $p(w|d)$ arrives into the database and topics have to be recomputed in order to retrieve these new samples. In this incremental scheme, we are going to compare the global approach using pLSA, LDA and FSTM with the incremental one using IpLSA.

Figure 5.6 shows the two tested alternatives. On the one hand, the global approach uses the union of previous and new samples to extract a new set of

topics and to represent the new samples in these topics. On the other hand, the incremental approach takes advantage of the initial topics in order not to process the previous documents.

Parameters of the models

Number of topics: In this work, we have set the number of topics to a percentage of the number of samples used to extract them. In particular, we have considered 10% of samples as the number of topics, except for collections bigger than 6,000 documents where we have taken 100 topics for each 3,000 samples. This may not be the best scenario but it allows us to perform the topic extraction task in an affordable time and space and besides that, it allows us to compare all the topic models in the same conditions in this incremental scenario. Choosing the right number of topics is an open ended question in the literature, especially for the visual domain. Despite the fact there are some approaches which try to tackle this problem [5, 62], all of them require performing the topic extraction process several times which eventually makes it impractical to use them in an interactive video retrieval system with a relatively large database.

Corvengence parameters: For all the tested models, we have used the original implementation of the authors with a threshold of 10^{-6} in the difference of the log-likelihood between two consecutive iterations and a maximum of 1,000 EM iterations. For the rest of the parameters, we have used the default settings with automatic estimation of the Dirichlet hyper-parameters for LDA and FSTM. The default settings are not always the optimal configuration for a particular dataset, but there are several reasons to use those configurations. First of all, the topic model algorithms are too costly to perform the extracting process multiple times using several settings. Second, the CBVR problem is not a classical classification problem in which we can use a partition of the training set to validate those parameters. In this case, the query itself defines the target and the test of the retrieval process. Finally, using the same convergence configuration makes the result comparable although it may be not optimal.

Consumer Columbia Video (CCV) database

The Columbia Consumer Video (CCV) database [32] contains 9,317 YouTube videos over 20 semantic categories, most of which are complex events, along with several objects and scenes. The authors of the database provide three different

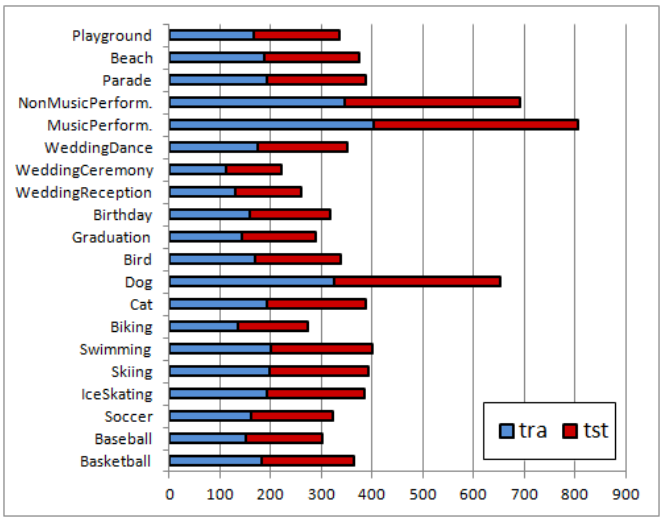


Figure 5.7: Samples per class of the CCV database.

characterizations for the videos of the collection: (1) based on SIFT descriptors (static info); (2) STIP (dynamic info); (3) and MFCC (audio). According to the classification accuracy reported by the authors, the SIFT descriptor achieves the best accuracy and a combination of all of them does not improve the performance in a significant way. Besides, the concatenation of all the descriptors produces a remarkable dimensionality increase which leads to an increase of the computational burden of the topic extraction task. Taking these reasons into account, we have decided to use the characterization based on the SIFT descriptors in order to simplify the testing of the proposed approach. However, further improvements could be aimed at considering multiple information channels. The vocabulary of the SIFT characterization was defined as a Bag of Words (BoW) model from 500 clusters on SIFT descriptors over Hessian-Affine and DoG feature points extracted over the entire and 2×2 image blocks, which makes a total of 5,000 words. From this corpus, we have eliminated samples with null descriptor information or with no annotation. For the remaining ones, samples labelled with more than one category have been replicated one for each class. Eventually, we have considered a total of 7,846 video samples annotated in 20 classes (Figure 5.7). We have used the same training and test partitions provided by the authors of the dataset which makes a total of 3,914 samples for training and 3,932 for test. Regarding the incremental scheme, the training partition has been considered the initial set of samples d_0 and the test partition the new set of samples d to be retrieved.

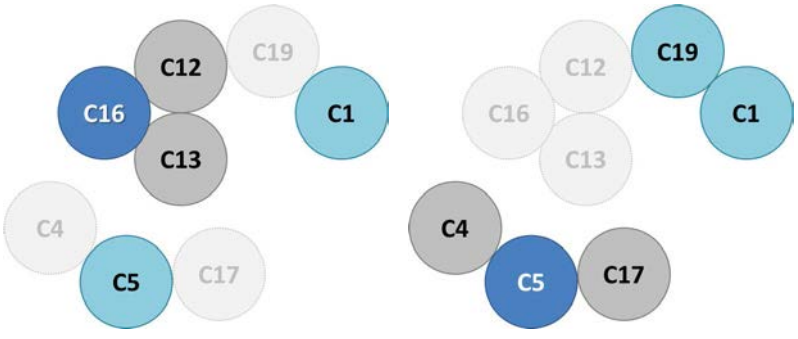
In addition to the entire dataset, we have considered four additional partitions

with 1,000 samples to allow us to analyse slight differences between the considered models. The goal is to test the performance of the models depending on the topology of the data with an affordable cost of the topic extraction process.

For the first partition (C16C12C10), we have selected the class NonMusicPerformance (C16) and its two nearest classes, WeddingReception (C12) and Graduation (C10). That is, C12 and C10 are those classes whose centroids have less euclidean distance to the centroid of C16 in the initial BoW representation using SIFT descriptors. For the incremental scheme, we have considered the class C16 as the initial set of samples d_0 and the rest of the two classes as the new set of samples d. With this partition, we pretend to simulate a situation when the new samples are similar to the initial ones but belonging to utterly different query concepts.

In the second partition (C16C1C5), we have selected class C16 and its two furthest classes, Baseball (C1) and Swimming (C5). In this case, we have considered class C16 as the initial set of samples (d_0) and classes C1 and C5 as the set of new samples (d). This partition tries to simulate a case where the new samples are quite different with respect to the initial ones and they are related to different query concepts as well.

In the case of the third (C5C17C4) and fourth (C5C1C19) partitions, we have considered class Swimming (C5) as the initial set of samples d_0 and the two nearest classes (C17 Parade and C4 Skiing) as the set of new samples d and two further ones (Baseball (C1) and Playground (C19)). With these partitions, we want to test the same configuration as before but using a different initial class. Figure 5.8 shows a schematic representation of the distance among the centroid of the considered classes.



(a) Class C16. (b) Class C5.

Figure 5.8: Scheme of the distance among the considered class centroids.

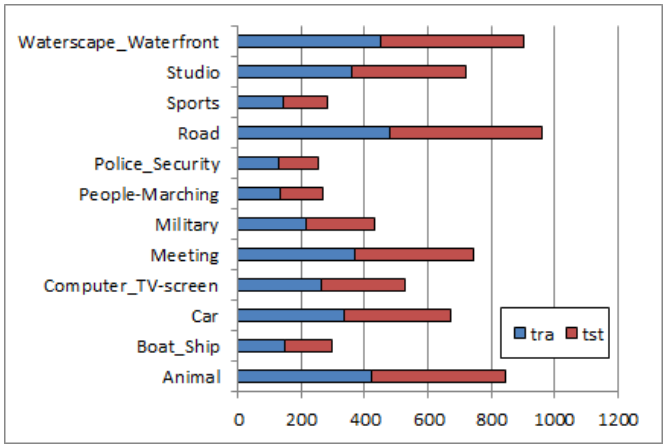


Figure 5.9: Subset of TRECVID 2007.

Video collection TRECVID 2007

The TRECVID 2007 dataset [7] is made up of 47,548 video shots which are annotated according to 36 semantic concepts. These categories were selected in TRECVID 2007 evaluation and they include several objects as well as complex events and scenes. Regarding the description of the database, we have used a characterization similar to that in the case of CCV. In particular, we have followed the suggestions of van de Sande et al. of using opponent SIFT histograms [68] when choosing a single descriptor and no prior knowledge about the dataset is considered. The software provided by van de Sande has been applied to the middle frame of each shot and each sample has been encoded using a 3-level spatial pyramid codebook (1×1, 2×2 and 4×4) that makes a total of 2,688 words per shot. In order to make affordable the computational cost of the topic extraction task, we have reduced the original database by selecting 12 of the 36 classes of the collection. Specifically, we have chosen those classes with a number of samples between 200 and 1,000 which makes a total of 6,906. Besides, these samples have been divided into two balanced partitions, one for training with 3,451 shots and another for testing with 3,455 (Figure 5.9). For the incremental scheme, the training partition has been considered the initial set of samples d_0 and the test partition the new set of samples d to be retrieved.

Visual information of topics for CBVR

Different from the text domain, the standard visual description methods generate a vocabulary so complex that their words are not easily interpretable in a visual

way. As a result, the direct visualization of topics is not helpful to understand the advantages of latent topics in the video retrieval domain. However, given the representation of documents in topics $p(z|d)$ those documents which are more probable to belong to a specific topic are somehow describing the kind of information that this topic is encapsulating and may help us to understand why topics can be useful for CBVR.

Considering the complete CCV database, we have used pLSA to extract 200 topics and to represent the whole collection in those topics. Using the representation $p(z|d)$, we have selected the six most probable documents per topic and five examples of these topics are shown in Figure 5.10. According to this figure, topic 21 tends to appear in videos related to the concept of ceremony, topic 48 refers to people riding a bike, topic 63 clearly shows videos of basketball games, topic 116 seems to represent videos of children playing with adults and topic 193 contains videos related to beach scene.

In general, it seems that topics tend to represent related patterns such us those in the text domain, but the issue that makes topic modelling suitable for CBVR is the capability to connect different kinds of samples through the concepts defined by topics. As we can see in Figure 5.10, both videos 48.d and 48.e have a high proportion of topic 48 because they are strongly related through the concept of "riding a bike", but at the same time those videos have a high proportion of topics 116 "children playing" and 193 "beach" respectively. This fact allows the video retrieval system to connect 48.d and 48.e with other videos through two different topics depending on the feedback provided by the user. In CBVR, these kinds of connections are very important because the query concept is completely unconstrained and videos can be related to several semantic concepts simultaneously.

Results

Table 5.2 shows the abbreviation used for each partition as well as the details for the global approach, the incremental approach and the retrieval set used in each case.

Using these partitions, we have compared the global use of pLSA, LDA and FSTM with the incremental IpLSA in terms of average precision, F_1 score and computational cost of the topic extraction algorithm. In all the cases, the retrieval simulation intends to retrieve the new set of samples d, that is, given a random

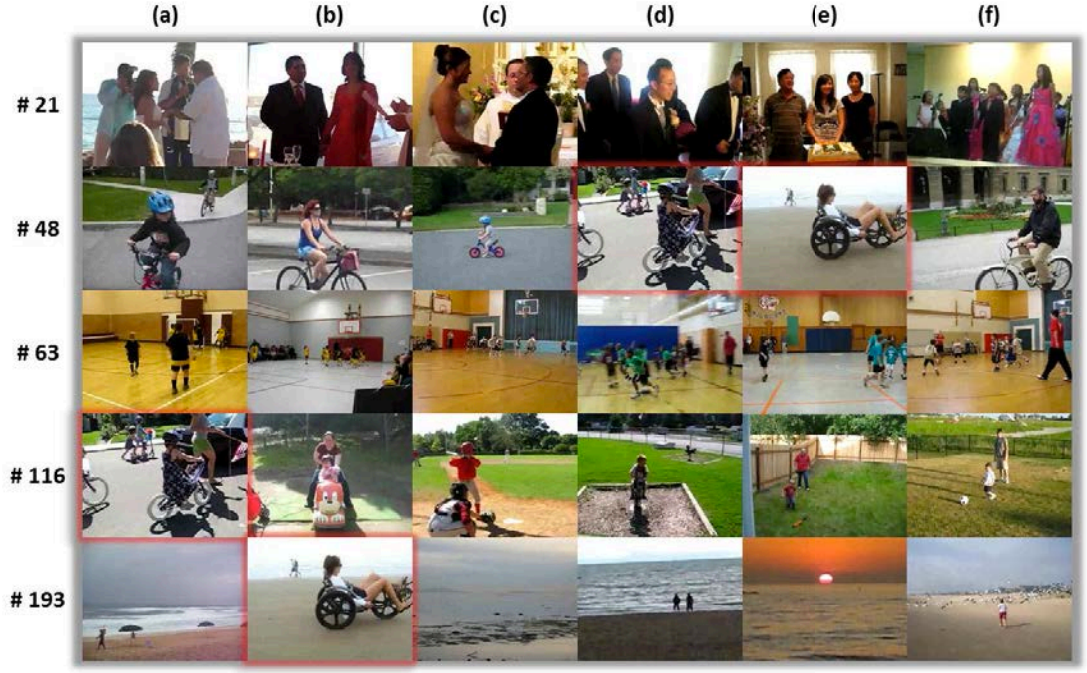


Figure 5.10: The six most probable documents of five topics from CCV.

Table 5.2: Partitions used for the video retrieval simulations.

	Global Scenario		*Incremental Scenario*			*Retrieval Set*	
	Name	**Partition**	**Name**	**Previous Stage**	**New Samples**	**Name**	**Partition**
CCV	A	C16C12C13 $d_0 \cup d = 1239$	A'	C16 $d_0 = 692$ $z_0 = 70$ (pLSA)	C12C13 $d = 547$	R_A	C12C13 $d = 547$
	B	C16C1C5 $d_0 \cup d = 1394$	B'	C16 $d_0 = 692$ $z_0 = 70$ (pLSA)	C1C5 $d = 702$	R_B	C1C5 $d = 702$
	C	C5C17C4 $d_0 \cup d = 1180$	C'	C5 $d_0 = 401$ $z_0 = 40$ (pLSA)	C17C4 $d = 779$	R_C	C17C4 $d = 779$
	D	C5C19C1 $d_0 \cup d = 1036$	D'	C5 $d_0 = 401$ $z_0 = 40$ (pLSA)	C19C1 $d = 635$	R_D	C19C1 $d = 635$
	E	TRA-TST $d_0 \cup d = 7846$	E'	TRA $d_0 = 3914$ $z_0 = 100$ (pLSA)	TST $d = 3932$	R_E	TST $d = 3923$
TRECVID	F	TRA-TST $d_0 \cup d = 6906$	F'	TRA $d_0 = 3451$ $z_0 = 100$ (pLSA)	TST $d = 3455$	R_F	TST $d = 3455$

Table 5.3: Computational cost of the topic extraction process (Intel Xeon E5-2640).

	A			**A′**	**B**			**B′**	**C**			**C′**
Model	pLSA	LDA	FSTM	IpLSA	pLSA	LDA	FSTM	IpLSA	pLSA	LDA	FSTM	IpLSA
NumTopics	130	130	130	60	140	140	140	70	120	120	120	80
Time (h)	20	43	3	8	24	49	4	12	20	34	3	12
Mem (MB)	3,101	182	81	1,374	3,754	196	88	1,896	2,728	148	73	1,805
	D			**D′**	**E**			**E′**	**F**			**F′**
Model	pLSA	LDA	FSTM	IpLSA	pLSA	LDA	FSTM	IpLSA	pLSA	LDA	FSTM	IpLSA
NumTopics	110	110	110	70	200	200	200	100	200	200	200	100
Time (h)	16	29	2	9	259	518	18	128	113	309	10	52
Mem (MB)	2,198	135	65	1,351	30,092	697	434	15,090	14,243	398	210	7,132

query from d the simulation pretends to retrieve the rest of the samples of d which belong to the same class than the query. The parameters of the models have been discussed in section 5.4.2. Note that the number of topics has been fixed depending on the number of samples used to extract the topics, that is, $d_0 \cup d$ for the global approach and d for the incremental one. In the incremental approach, it has been assumed that the pLSA model is used to obtain the topics of the previous stage (documents d_0) but any other model could be considered. Taking into account these previous topics, the IpLSA model only needs the new documents d to extract the topics, as a result the number of topics for the IpLSA is substantially lower than that in the global approach. Table 5.3 shows the computational efficiency of the topic extraction process for the considered models (temporal complexity in hours running in an Intel Xeon E5-2640 processor and spatial complexity in MB of RAM). Table 5.4 contains the average precision of the experiments and Table 5.5 shows the F_1 measure calculated as $2(Precision * Recall)/(Precision + Recall)$.

Statistical tests

In order to ease the comparison, Wilcoxon's signed rank test has been applied to show whether statistical differences exist among the video retrieval performances of the considered topic models. Despite some previous works advocated for the discontinuation of this statistical test, other recent papers like [67] conclude that Wilcoxon's test is able to provide more robust significance levels in information retrieval and for that reason we have decided to use it.

Wilcoxon's signed rank test provides pairwise comparisons, so statistical differences between each pair of topic models can be found. This statistical test is based on a null hypothesis which assumes statistical equality. In this case, it is

Table 5.4: Video retrieval results: **Average Precision**. For each simulation of each partition the best result is highlighted in bold.

			Latent Topics Rank				*Cosine Similarity*			
Partition	**TM**	**Retr. Set**	**Sim1**	**Sim2**	**Sim3**	**Sim4**	**Sim1**	**Sim2**	**Sim3**	**Sim4**
A	pLSA	R_A	**0.67**	**0.69**	0.59	0.59	0.48	0.48	0.40	0.39
	LDA		0.63	0.66	0.56	0.58	0.48	0.49	0.41	0.41
	FSTM		0.47	0.45	0.47	0.47	0.47	0.50	**0.46**	**0.46**
A'	IpLSA		0.66	0.67	**0.59**	**0.61**	**0.51**	**0.52**	0.44	0.45
B	pLSA	R_B	0.70	0.73	0.67	0.69	0.63	0.65	0.56	0.57
	LDA		0.72	0.74	0.67	0.69	0.63	0.66	0.56	0.58
	FSTM		0.60	0.65	0.60	0.62	0.58	0.61	0.53	0.53
B'	IpLSA		**0.74**	**0.76**	**0.70**	**0.72**	**0.65**	**0.68**	**0.61**	**0.62**
C	pLSA	R_C	**0.93**	**0.94**	**0.93**	**0.94**	0.94	0.97	0.92	0.94
	LDA		0.92	0.93	0.93	0.94	0.94	0.96	0.92	0.94
	FSTM		0.87	0.88	0.88	0.88	0.91	0.92	0.91	0.92
C'	IpLSA		0.92	0.94	0.93	0.94	**0.95**	**0.97**	**0.94**	**0.96**
D	pLSA	R_D	**0.62**	0.65	0.57	0.59	0.54	0.55	0.48	0.48
	LDA		0.62	0.65	0.57	0.59	0.56	0.59	0.50	0.52
	FSTM		0.54	0.56	0.54	0.56	**0.58**	**0.62**	**0.56**	**0.58**
D'	IpLSA		0.62	**0.67**	**0.58**	**0.61**	0.56	0.59	0.50	0.52
E	pLSA	R_E	0.10	0.12	0.10	0.11	0.14	0.15	0.11	0.12
	LDA		0.09	0.11	0.09	0.10	0.12	0.13	0.10	0.11
	FSTM		0.08	0.10	0.08	0.10	0.07	0.10	0.07	0.08
E'	IpLSA		**0.11**	**0.13**	**0.11**	**0.12**	**0.14**	**0.17**	**0.12**	**0.14**
F	pLSA	R_F	**0.39**	**0.39**	0.35	0.34	**0.36**	**0.37**	0.29	0.29
	LDA		0.35	0.35	0.31	0.31	0.29	0.30	0.26	0.27
	FSTM		0.26	0.27	0.28	0.27	0.30	0.30	0.30	0.29
F'	IpLSA		0.34	0.36	**0.35**	**0.35**	0.35	0.35	**0.30**	**0.31**

Table 5.5: Video retrieval results: **F_1 Score**. For each simulation of each partition the best result is highlighted in bold.

			Latent Topics Rank				*Cosine Similarity*			
Partition	**TM**	**Retr. Set**	**Sim1**	**Sim2**	**Sim3**	**Sim4**	**Sim1**	**Sim2**	**Sim3**	**Sim4**
A	pLSA	R_A	**0.36**	**0.37**	0.50	0.50	0.26	0.26	0.34	0.33
	LDA		0.34	0.35	0.47	0.49	0.26	0.26	0.35	0.34
	FSTM		0.25	0.24	0.40	0.40	0.25	0.27	**0.39**	**0.39**
A'	IpLSA		0.35	0.36	**0.50**	**0.51**	**0.27**	**0.28**	0.37	0.38
B	pLSA	R_B	0.31	0.32	0.49	0.50	0.27	0.27	0.39	0.40
	LDA		0.32	0.33	0.49	0.51	0.27	0.28	0.39	0.41
	FSTM		0.27	0.29	0.43	0.44	0.25	0.26	0.37	0.38
B'	IpLSA		**0.33**	**0.34**	**0.51**	**0.53**	**0.28**	**0.29**	**0.43**	**0.44**
C	pLSA	R_C	**0.38**	**0.38**	**0.63**	**0.64**	0.38	0.39	0.62	0.64
	LDA		0.37	0.38	0.63	0.63	0.38	0.39	0.62	0.64
	FSTM		0.35	0.35	0.59	0.60	0.37	0.37	0.62	0.63
C'	IpLSA		0.37	0.38	0.63	0.64	**0.39**	**0.40**	**0.64**	**0.65**
D	pLSA	R_D	**0.30**	0.31	0.44	0.45	0.25	0.26	0.36	0.36
	LDA		0.29	0.31	0.44	0.45	0.26	0.28	0.38	0.39
	FSTM		0.26	0.26	0.41	0.43	**0.27**	**0.29**	**0.43**	**0.44**
D'	IpLSA		0.29	**0.32**	**0.44**	**0.47**	0.26	0.28	0.38	0.39
E	pLSA	R_E	0.07	0.08	0.10	0.11	0.09	0.09	0.11	0.11
	LDA		0.06	0.07	0.09	0.10	0.09	0.10	0.12	0.13
	FSTM		0.05	0.07	0.08	0.10	0.05	0.07	0.07	0.09
E'	IpLSA		**0.07**	**0.08**	**0.11**	**0.12**	**0.10**	**0.11**	**0.12**	**0.14**
F	pLSA	R_F	**0.18**	**0.18**	0.26	0.25	**0.16**	**0.17**	0.22	0.22
	LDA		0.16	0.16	0.23	0.23	0.13	0.14	0.20	0.20
	FSTM		0.12	0.12	0.21	0.20	0.14	0.14	0.23	0.22
F'	IpLSA		0.16	0.16	**0.26**	**0.26**	0.16	0.16	**0.23**	**0.23**

Table 5.6: Summary of Wilcoxon's statistic test applied over video retrieval precision values for all pairs of topic models using the **LTR ranking function**.

		Simulation 1				*Simulation 2*				*Simulation 3*				*Simulation 4*			
		pLSA	LDA	FSTM	IpLSA	pLSA	LDA	FSTM	IpLSA	pLSA	LDA	FSTM	IpLSA	pLSA	LDA	FSTM	IpLSA
A	pLSA	-	•	•		-		•		-	•	•		-		•	○
	LDA		-	•	○		-	•			-	•	○		-	•	○
	FSTM	○	○	-	○	○	○	-	○	○	○	-	○	○	○	-	○
A′	IpLSA		•	•	-			•	-		•	•	-		•	•	-
B	pLSA	-		•	○	-		•	○	-		•	○	-		•	○
	LDA		-	•	○		-	•			-	•	○		-	•	○
	FSTM	○	○	-	○		○	-	○	○	○	-	○	○	○	-	○
B′	IpLSA	•	•	•	-	•		•	-	•		•	-	•		•	-
C	pLSA	-	•	•	•	-	•	•	•	-		•		-	•	•	
	LDA		-	•		○	-	•			-	•		○	-	•	
	FSTM	○	○	-	○	○	○	-	○	○	○	-	○	○	○	-	○
C′	IpLSA	○		•	-			•	-			•	-			•	-
D	pLSA	-		•		-		•	○	-				-			○
	LDA		-				-	•	○		-				-		○
	FSTM			-				-	○			-	○			-	○
D′	IpLSA				-				-				-				-
E	pLSA	-	•	•	•	-	•	•	•	-	•	•		-	•	•	
	LDA	○	-	•		○	-	•		○	-	•	○	○	-	•	○
	FSTM	○	○	-	○	○	○	-	○	○		-	○	○	○	-	○
E′	IpLSA	○		•	-	○		•	-		•	•	-			•	-
F	pLSA	-	•			-	•			-	•	•		-	•		
	LDA		-		○		-			○	-		○	○	-		○
	FSTM			-				-				-	○			-	
F′	IpLSA				-				-		•		-		•		-

assumed certain that all topic models perform equally for the video retrieval task and evidence is searched for in the data to reject it. Table 5.6 shows the statistical differences among the used topic models with the LTR ranking function and Table 5.7 the differences using the cosine similarity function. In both tables, a summary of Wilcoxon's statistic test applied over the video retrieval precision values for all pairs of topic models is shown. Above the main diagonal with a 90% confidence level and below it with 95%. The symbol • indicates that the model in the row significantly outperforms the model in the column, and the symbol ○ indicates that the model in the column significantly surpasses the model in the row.

5.5 Discussion

This section contains a discussion about the obtained results. Initially, we discuss the results focused on each kind of partition and later a global discussion is presented.

Table 5.7: Summary of Wilcoxon's statistic test applied over video retrieval precision values for all pairs of topic models using the **cosine ranking**.

		Simulation 1				*Simulation 2*				*Simulation 3*				*Simulation 4*			
		pLSA	LDA	FSTM	IpLSA	pLSA	LDA	FSTM	IpLSA	pLSA	LDA	FSTM	IpLSA	pLSA	LDA	FSTM	IpLSA
A	pLSA	-				-			○	-		○	○	-		○	○
	LDA		-				-		○		-	○	○		-	○	○
	FSTM			-				-				-	●			-	
A′	IpLSA				-	●			-	●			-	●			-
B	pLSA	-				-				-				-			○
	LDA		-	●	○		-		○		-		○		-		○
	FSTM			-	○			-				-	○			-	○
B′	IpLSA			●	-				-		●	●	-		●	●	-
C	pLSA	-				-		●		-				-			
	LDA		-	●	○		-	●	○		-		○		-	●	○
	FSTM			-	○		○	-	○			-	○			-	○
C′	IpLSA		●	●	-			●	-		●	●	-		●	●	-
D	pLSA	-	○		○	-	○	○	○	-	○	○	○	-	○	○	○
	LDA	●	-			●	-			●	-			●	-		
	FSTM			-				-				-				-	●
D′	IpLSA				-				-				-	●			-
E	pLSA	-	●	●	○	-	●	●	○	-	●	●	○	-	●	●	○
	LDA	○	-	●	○	○	-	●	○	○	-	●	○	○	-	●	○
	FSTM	○	○	-	○	○	○	-	○	○	○	-	○	○		-	○
E′	IpLSA		●	●	-		●	●	-		●	●	-	●	●	●	-
F	pLSA	-	●	●		-	●	●		-	●			-			
	LDA	○	-		○	○	-		○		-	○	○		-	○	○
	FSTM	○		-	○	○		-	○		●	-				-	
F′	IpLSA		●	●	-		●	●	-		●		-		●		-

5.5.1 Unbalanced nearest partitions (A and C)

In the case of unbalanced nearest partitions (A and C), the set of new samples d is very close to the initial set d_0 despite the fact that d contains two new video classes to be retrieved. Although there are slight differences between the performance of both ranking functions, pLSA-based models tend to obtain the best average precision. Statistical tests support these results especially with a confidence level of 95%. In general, there are no statistical differences between pLSA and IpLSA, besides both models are able to outperform LDA and FSTM in many cases.

In these kinds of partitions, the new classes to retrieve are rather confusing what forces topics to be very adjusted to the data distribution in order to distinguish slight differences over patterns. LDA seems to not have enough samples to adequately estimate the Dirichlet parameters for these fuzzy concepts whereas pLSA-based models are taking advantage of using their own documents as parameters.

In terms of computational efficiency, FSTM shows an impressive performance but its sparse assumptions seem inadequate especially for the LTR ranking. For the rest of the models, IpLSA obtains an important time reduction with respect to pLSA and LDA, but in terms of space LDA is able to obtain a high efficiency.

This memory reduction is produced by the fact that LDA uses an external Dirichlet distribution rather than using its own documents as parameters as is in the case of pLSA-based models. However, the parameter estimation for this external distribution is making the topic extraction process much slower. Comparing the two pLSA-based models, IpLSA obtains a noticeable spatial improvement with respect to pLSA because it only uses the new documents to obtain the new topics and as a result it stores much less documents during the topic extraction process.

5.5.2 Unbalanced furthest partitions (B and D)

For these partitions (B and D), the new set of documents d pretends to be quite different from the initial set of samples d_0 in order to capture new patterns. In this case, the results show that IpLSA outperforms many of the models. According to the statistical tests, these improvements are particularly important for the LTR ranking with a confidence level of 90%.

Now, we can observe how LDA tends to perform better than pLSA because the classes to retrieve are quite separated and dense enough to enable LDA to estimate the Dirichlet parameters properly whereas pLSA may produce overfitting. Related to the incremental scheme, IpLSA is able to obtain a better result than the global use of LDA because IpLSA is focused on detecting unseen patterns and then it can take advantage of partitions where the new set of samples contains a clearly new patterns.

Regarding the computational complexity, we can observe the same behaviour as that in the previous section, because the complexity of the topic extraction process is proportional to the number of documents, words and topics, and these variables are similar to the previous partitions. FSTM is much more efficient than the rest of the models. IpLSA is faster than LDA but it has a bigger spatial complexity and pLSA is quite worse than IpLSA in terms of time and space.

5.5.3 Complete collections (E and F)

These partitions (E and F) try to reproduce a situation in which the new documents d are not introducing a very different new topics but refining the previous ones. In general, the average precision has significantly fallen because now we are trying to retrieve much more concepts than before and besides the amount of topics is quite limited. We have extracted only 100 topics for each 3,000 sam-

ples in order to make the extraction process affordable. However, the ranking functions may require more topics to distinguish better among all the classes because of data complexity. IpLSA has obtained the best average precision for CCV and both pLSA and IpLSA for TRECVID. The statistical tests show that the pLSA-based models tend to outperform the rest of the models.

In this case, we would have expected a better performance of LDA because topics have been extracted using much more samples than those in the previous partitions. However, LDA has obtained a worse result than both the pLSA and IpLSA models. The semantic gap of the characterization together with the high number of classes to retrieve may produce this low performance of LDA. The fact of considering a relatively high amount of classes with a huge semantic gap is generating a sort of complex space where some concepts are not well defined, and in this circumstance pLSA-based models are able to adapt the topic structure using documents lesser than those of LDA.

Related to the efficiency of the models, LDA is by far the worse model in terms of time and pLSA in terms of space. The topic extraction task by LDA takes over 2 times more computational time than pLSA, 5 times more than IpLSA and 10 times more than FSTM. On the other hand, the memory usage of pLSA is over the double that of IpLSA, 10 times more space than that of LDA and more than 20 times than that of FSTM.

5.5.4 General issues

According to the results, we agree with [44] to conclude that LDA is able to outperform pLSA for the video retrieval field as well, when the partition used to extract the topics is quite unambiguous and dense like in partitions B and D. In these circumstances, the retrieval system needs a general fine-granularity representation which can be provided better by LDA due to the fact that pLSA tends to over-fit whereas LDA is able to estimate the Dirichlet parameters properly. However, pLSA-based models have shown to be more effective in fuzzy conditions where concepts are not described with enough documents. As a result, we agree with [34] by saying that pLSA-based models are able to outperform the LDA model because the use of the documents as parameters allows the topics to fit better to a sparse data distribution.

Regarding the proposed incremental model, IpLSA has shown to be effective in both cases. On the one hand, when pLSA tends to over-fit the incremental

model IpLSA is able to work properly by avoiding learning repetitive patterns and reducing the computational cost. On the other hand, IpLSA takes advantage of considering the document parameters of the model when LDA does not have enough documents to adequately estimate the Dirichlet parameters. In general, pLSA has shown to be effective for CBVR although the over-fitting problem but the proposed incremental model is able to obtain some improvements over pLSA in terms of precision and cost.

In relation to computational complexity, FSTM has shown an impressive computational performance but unfortunately in many cases its results are not good enough for unconstrained video retrieval. According to the results, the FSTM model is clearly outperformed by the rest of the tested models for the LTR function and in many cases for the cosine similarity function. In unconstrained video retrieval, it is usual to have to manage very complex concepts without having enough samples to describe them properly. In this kind of application, a dense contribution of topics as in the case of pLSA or LDA has proved to be more effective. For the rest of the tested methods, LDA has obtained the best spatial performance and IpLSA the best computational time.

5.6 Conclusions and Future Work

This chapter has presented an incremental extension of the pLSA model in order to enable video retrieval systems based on latent topics to deal with incremental databases in an effective way as well as an experimental study on the performance of different topic models for the video retrieval problem.

Using the video retrieval systems presented in [23] and [80], four retrieval scenarios have been simulated using two different databases and four topic extraction algorithms. From the results, we can draw three main trends in CBVR: (1) LDA is able to outperform pLSA in unambiguous and dense conditions; (2) pLSA-based models performs better in fuzzy and sparse distributions; (3) IpLSA is able to obtain good results in both cases using an incremental approach. In general, the IpLSA model has shown to be more effective in dealing with incremental databases than the rest of the tested global methods. In terms of video retrieval precision, the IpLSA model is able to outperform pLSA and LDA when these two models obtain the lowest performance. Moreover, when they achieved the highest precision, IpLSA was able to work without statistical differences. Re-

lated to the computational complexity, the results have shown that IpLSA is able to significantly reduce the time of the LDA/pLSA models and the space of the pLSA as well.

Although the results are encouraging, much more progress is needed to really address the efficiency problems of the topic extraction methods for video retrieval. Thus, further work is directed to extend the work in the following directions:

- Automatic strategies to choose the number of new topics at each iteration of the incremental scheme.
- Extension of the model to allow the use of multi-modal data from multiple channels.
- Reduction of the over-fitting in pLSA-based models by applying quantization techniques over the samples used to extract the topics.

Chapter 6

Global discussion and conclusions

At the beginning of this thesis, we were concerned about how to use latent topics to deal with the semantic gap challenge in CBVR, and along the different chapters we have shown the potential of topics models at different levels. Thus, the contributions of this thesis cover four of the stages of the video retrieval process:

(i) Encoding (chapter 3). Related to the encoding level, the presented topic-based method (LTE) provides a competitive performance especially for those retrieval methods used in the latent space. The definition of the visual vocabulary according to the hidden patterns together with the soft encoding of the local features over topics generates a more suitable encoding than the regular BoW approach to work in the latent topic space.

(ii) Vocabulary reduction (chapter 2). The second level in which topic models has shown to be effective is the vocabulary reduction stage. Specifically, applying word filters over the uncovered topics instead of over the own documents can effectively be used to reduce the vocabulary of a collection leading to a performance improvement. Our approach takes advantage of the fact that topic models are able to summarise the semantics of a collection in a reduced set of topics.

(iii) Modelling (chapter 5). The use of latent topics in the CBVR modelling phase has traditionally been rather unusual because of the complexity of the visual domain. However, our work shows that topic models are able to provide a competitive advantage to deal with the semantic gap challenge in video retrieval. In particular, our contribution in the modelling stage is twofold: (a) studying how the use of different topic models affects to the

video retrieval performance and (b) presenting a novel incremental topic model (IpLSA) to cope with incremental retrieval scenarios in an effective way. A noteworthy conclusion regarding the use of different topic models is based on the fact that pLSA-based models may be able to estimate the actual patterns of a collection using less samples than LDA and this may be useful in CBVR where it is common to deal with complex query concepts having very little information about them.

(iv) Ranking (chapter 4). Regarding the ranking stage, this level can be considered the most important one because eventually it is in charge of selecting the samples which are extracted over the retrieval process. Until now, topic models were used as a mere alternative characterisation for traditional ranking functions based on distances, similarities or even classifiers. However, many of these functions tend to perform worse in the latent space than in the original characterisation space. Precisely, this fact has made that topic models have not been considered useful for many years in tasks where precision is important like CBVR. Our contribution in this area consists in turning the classic retrieval approach into a class discovery problem via topic models to perform the ranking task according to the nature of the latent space. The proposed ranking function (LTR) is able to provide a competitive advantage to cope with the semantic gap because it is deduced following the same probabilistic nature than topic models unlike the state-of-the-art ranking functions.

From the work developed in this thesis, we can state that topic models can be very useful at different levels in CBVR, being its main advantages:

- In a sense, topic models can be helpful to analyse a data collection as well as to provide a higher characterisation level when the data structure is a priori unknown or there is only little information about the target. Precisely, this is the typical case in CBVR. In this kind of application, we usually have to deal with complex query concepts having very little information about them because the number of examples in the query initialisation and feedback are usually very limited. Besides, the semantic meaning of a specific video may depend on the user bias what eventually makes more difficult to find about the concept of interest. A retrieval system has to tackle these problems by providing more flexibility in the output ranking and topic models are

excellent tools for this purpose. The use of topic models allow video retrieval systems to make connections among patterns defined by topics, that is, even two very different videos may be related as long as they share some of the topics extracted from the data collection.

- However, the direct use of latent topics is often unsuccessful because of the special nature of the latent space. As part of the computer vision community, we all are used to isolating classes to obtain better decision boundaries but topic models work for just the opposite. The latent space tends to mix classes and stablishes links between objects (text documents, videos...) according to the patterns of the collection, therefore many of classic strategies, which assume that similar things have to be close in the representation space, do not work properly in the latent space. In fact, one of our main contributions is based on highlighting that point and providing an effective retrieval scheme which takes into account the own topic nature.

Despite the aforementioned advantages, topic models have some limitations which have to be considered:

- The first one is related to the fact that the connections among patterns provided by topic models are only useful assuming a specific semantic gap. That is, the bigger the semantic difference between the data representation and the final target, the more effective the topic models tend to be, in fact, topic models may be counter-productive without this semantic gap. If the initial data representation space is able to capture the properties we are looking for in the data, representing the data according to its hidden patterns is going to mix those properties and eventually is not going to provide any advantage with respect to the original space.

- The second limitation of topic models is based on the high computational cost of their algorithms. Extracting the topics of a collection is a very demanding process whose cost highly depends on the number topics and the number of samples in the collection. Even though some models, such as the proposed IpLSA model, try to reduce this computational cost making some assumptions, the computational burden of the process remains still a challenge for collections with millions of samples.

6.1 Future Work

Finally, we provide some possible lines of work for the extension of the methods proposed in this thesis. Somehow, they summarise those already introduced in the closing sections of the different chapters:

- Assessing and developing automatic strategies to choose the ideal vocabulary size in a collection and the more appropriate number of topics.

- Applying quantisation techniques to reduce the number of documents used to extract the topics and therefore reducing the computational cost of this task.

- Extending the proposed LTR ranking function to a long-term relevance feedback approach. A possible way to do this could be by integrating the sSpLSA model in an incremental scenario and deducing the new ranking function according to those new constraints. In a sense, it would be a kind of fusion between sSpLSA and IpLSA but focused on query classes instead of topics.

- Developing a multi-modal extension of the proposed retrieval model. In this particular line of work, there are two different options: (i) use our models as they are and try to fuse the results obtained by different modalities in a final step, and (ii) extend our models to manage multiple vocabularies in order to obtain a single result which takes into account the different input channels.

- Providing an end-to-end video retrieval approach. Inspired by the deep learning research, this extension is related to avoid using any encoding stage as a base of our topic-based video retrieval approach. In some way, the proposed LTE technique tries to use this idea but to obtain just the encoding stage. The idea would be to use a multiple layer topic model to extract the topics directly from the raw data without using any kind of pre-computed Bag-of-Words. However, some questions like the definition of words from the raw data should be discussed before.

6.2 List of Publications

Journals

- [25] Ruben Fernandez-Beltran and Filiberto Pla. Incremental probabilistic latent semantic analysis for video retrieval. *Image and Vision Computing*, 38:1–12, 2015.

- [26] Ruben Fernandez-Beltran and Filiberto Pla. Latent topics-based relevance feedback for video retrieval. *Pattern Recognition*, 51:72–84, 2016.

International Conferences

- [27] Ruben Fernandez-Beltran, Raul Montoliu, and Filiberto Pla. Vocabulary reduction in bow representing by topic modeling. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 648–655, 2013.

- [23] Ruben Fernandez-Beltran and Filiberto Pla. An interactive video retrieval approach based on latent topics. In *International Conference on Image Analysis and Processing*, pages 290–299, 2013.

- [24] Ruben Fernandez-Beltran and Filiberto Pla. Latent topic encoding for content-based retrieval. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 640–648, 2015.

Bibliography

[1] S Antani. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition*, 35(4):945–965, 2002.

[2] S Antani. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition*, 35(4):945–965, 2002.

[3] Miguel Arevalillo-Herrez and Francesc J. Ferri. An improved distance-based relevance feedback strategy for image retrieval. *Image Vision and Computing*, 2013.

[4] Miguel Arevalillo-Herrez, Francesc J. Ferri, and Juan Domingo. A naive relevance feedback model for content-based image retrieval using multiple similarity measures. *Pattern Recognition*, 43(3):619–629, 2010.

[5] R. Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, 2010.

[6] Avinash Atreya and Charles Elkan. Latent semantic indexing (lsi) fails for trec collections. *SIGKDD Explor. Newsl.*, 12(2):5–10, 2011.

[7] Stephane Ayache and Georges Qunot. Trecvid 2007 collaborative annotation using active learning. In *Proceedings of the TRECVID 2007 Workshop*, 2007.

[8] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003.

[9] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55 (4):77–84, 2012.

[10] David M. Blei and John D. Lafferty. Dynamic topic models. In *ACM International Conference on Machine Learning*, 2006.

[11] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via pLSA. In *European Conference on Computer Vision*, pages 517–530, 2006.

[12] A. Bosch, A. Zisserman, and X. Muoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

[13] Ana Cardoso-Cachopo and Arlindo Oliveira. Combining lsi with other classifiers to improve accuracy of single-label text categorization. In *European Workshop on Latent Semantic Analysis in Technology Enhanced Learning*, 2007.

[14] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22*, pages 288–296. 2009.

[15] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11:1109–1135, 2010.

[16] Chih-Yi Chiu, Tsung-Han Tsai, Yu-Cyuan Liou, Guei-Wun Han, and Hung-Shuo Chang. Near-duplicate subsequence matching between the continuous stream and large video dataset. *IEEE Transactions on Multimedia*, 16(7): 1952–1962, 2014.

[17] Tzu-Chuan Chou and Meng Chang Chen. Using incremental plsi for threshold-resilient online event analysis. *IEEE Transactions on Knowledge and Data Engineering*, 20(3):289–299, 2008.

[18] Matthieu Cord, Philippe H. Gosselin, and Sylvie Philipp-Foliguet. Stochastic exploration and active learning for image retrieval. *Image and Vision Computing*, 25(1):14–23, 2007.

[19] Courtenay V. Cotton and Daniel P. W. Ellis. Audio fingerprinting to identify multiple videos of an event. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2386–2389, 2010.

[20] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, 2000. ISBN 0-521-78019-5.

[21] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of Machine Learning Research*, 41:391–407, 1990.

[22] Michel Marie Deza and Elena Deza. *Encyclopedia of Distances*. Springer Berlin Heidelberg, 2009. ISBN 3642002331.

[23] Ruben Fernandez-Beltran and Filiberto Pla. An interactive video retrieval approach based on latent topics. In *International Conference on Image Analysis and Processing*, pages 290–299, 2013.

[24] Ruben Fernandez-Beltran and Filiberto Pla. Latent topic encoding for content-based retrieval. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 640–648, 2015.

[25] Ruben Fernandez-Beltran and Filiberto Pla. Incremental probabilistic latent semantic analysis for video retrieval. *Image and Vision Computing*, 38:1–12, 2015.

[26] Ruben Fernandez-Beltran and Filiberto Pla. Latent topics-based relevance feedback for video retrieval. *Pattern Recognition*, 51:72–84, 2016.

[27] Ruben Fernandez-Beltran, Raul Montoliu, and Filiberto Pla. Vocabulary reduction in bow representing by topic modeling. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 648–655, 2013.

[28] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, 2004.

[29] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.

[30] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Network*, 13(2): 415–425, 2002.

[31] Y. Jiang, S. Bhattacharya, S. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, 2013.

[32] Yu G. Jiang, Guangnan Ye, Shih F. Chang, Daniel Ellis, and Alexander C. Loui. Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In *ACM International Conference on Multimedia Retrieval*, 2011.

[33] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142. ACM, 2002. ISBN 1-58113-567-X.

[34] Tuomo Kakkonen, Niko Myller, Erkki Sutinen, and Jari Timonen. Comparison of dimension reduction methods for automated essay grading. *Educational Technology & Society*, 11(3):275–288, 2008.

[35] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[36] M.S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, 2006.

[37] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[38] Fei-Fei Li and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[39] Guang-Hai Liu, Zuo-Yong Li, Lei Zhang, and Yong Xu. Image retrieval based on micro-structure descriptor. *Pattern Recognition*, 44(9):2123–2133, 2011.

[40] Guang-Hai Liu, Zuo-Yong Li, Lei Zhang, and Yong Xu. Image retrieval based on micro-structure descriptor. *Pattern Recognition*, 44(9):2123–2133, 2011.

[41] Huan Liu and Hiroshi Motoda. *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 2007. ISBN 1584888784.

[42] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[43] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[44] Yue Lu, Qiaozhu Mei, and Chengxiang Zhai. Investigating task performance of probabilistic topic models: An empirical study of plsa and lda. *Information Retrieval*, 14(2):178–203, 2011.

[45] Meredith Minear and DeniseC. Park. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, and Computers*, 36(4): 630–633, 2004.

[46] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, pages 143–156, 2010.

[47] J. Philbin, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[48] Pedro Quelhas, Florent Monay, Jean-Marc Odobez, Daniel Gatica-Perez, and Tinne Tuytelaars. A thousand words in a scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[49] N. Rasiwasia and N. Vasconcelos. Latent dirichlet allocation models for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2665–2679, 2013.

[50] W. Ren, S. Singh, M. Singh, and Y. S. Zhu. State-of-the-art on spatio-temporal information-based video retrieval. *Pattern Recognition*, 42(2):267–282, February 2009.

[51] W. Ren, S. Singh, M. Singh, and Y. S. Zhu. State-of-the-art on spatio-temporal information-based video retrieval. *Pattern Recognition*, 42(2):267–282, 2009.

[52] W. Ren, S. Singh, M. Singh, and Y. S. Zhu. State-of-the-art on spatio-temporal information-based video retrieval. *Pattern Recognition*, 42(2):267–282, 2009.

[53] B. Saleh, A. Farhadi, and A. Elgammal. Object-centric anomaly detection by attribute-based reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–794, 2013.

[54] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, volume 2, pages 1470–1477, 2003.

[55] Josef Sivic and Andrew Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):591–606, 2009.

[56] A. F. Smeaton. Techniques used and open challenges to the analysis, indexing and retrieval of digital video. *Information Systems*, 32(4):545–559, 2006.

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

[58] C. Snoek, M. Worring, J. Gemert, J. Geusebroek, and A. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. *ACM International Conference on Multimedia*, 2006.

[59] Fabrice Souvannavong, Lukas Hohl, Bernard Merialdo, and Benoit Huet. Enhancing latent semantic analysis video object retrieval with structural information. In *IEEE International Conference on Image Processing*, 2004.

[60] Fabrice Souvannavong, Lukas Hohl, Bernard Merialdo, and Benoit Huet. Latent semantic analysis for an effective region-based video shot retrieval system. In *ACM International Workshop on Multimedia Information Retrieval*, 2004.

[61] Karen Sparck Jones and Peter Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann Publishers Inc., 1997. ISBN 1-55860-454-5.

[62] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2004.

[63] Khoat Than and Tu Bao Ho. Fully sparse topic models. In *European Conference on Machine Learning*, 2012.

[64] Kinh Tieu and Paul Viola. Boosting image retrieval. *International Journal of Computer Vision*, 56(1-2):17–36, 2004.

[65] Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *ACM International Conference on Multimedia*, pages 107–118, 2001.

[66] Ricardo da S. Torres, Alexandre X. Falcão, Marcos A. Gonçalves, João P. Papa, Baoping Zhang, Weiguo Fan, and Edward A. Fox. A genetic programming framework for content-based image retrieval. *Pattern Recognition*, 42 (2):283–292, 2009.

[67] Julián Urbano, Mónica Marrero, and Diego Martín. A comparison of the optimality of statistical significance tests for information retrieval evaluation. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 925–928, 2013.

[68] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[69] C.J. van Rijsbergen, S.E. Robertson, and M.F. Porter. New models in probabilistic information retrieval. 1980.

[70] Duc-Thuan Vo and Cheol-Young Ock. Learning to classify short text from scientific documents using topic models with various types of knowledge. *Expert Systems with Applications*, 42(3):1684–1698, 2015.

[71] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[72] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[73] J. Wang, W. Fu, H. Lu, and S. Ma. Bilayer sparse topic model for scene analysis in imbalanced surveillance videos. *IEEE Transactions on Image Processing*, 23(12):5198–5208, 2014.

[74] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 1945.

[75] Hu Wu, Yongji Wang, and Xiang Cheng. Incremental probabilistic latent semantic analysis for automatic question recommendation. In *ACM conference on Recommender systems*, pages 99–106. ACM, 2008.

[76] Yi Yang, Feiping Nie, Dong Xu, Jiebo Luo, Yueting Zhuang, and Yunhe Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2-3):723–742, 2012.

[77] Xing Yi and James Allan. A comparative study of utilizing topic models for information retrieval. In *European Conference on IR Research on Advances in Information Retrieval*, 2009.

[78] Liu Ying, Zhang Dengsheng, Lu Guojun, and Ma Wei-Ying. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[79] Dengsheng Zhang, Md. Monirul Islam, and Guojun Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346–362, 2012.

[80] R. Zhang and Z. Zhang. Effective image retrieval based on hidden concept discovery in image database. *IEEE Transactions on Image Processing*, 16: 562–572, 2007.

[81] Dengyong Zhou, Jason Weston, Arthur Gretton, Olivier Bousquet, and Bernhard Schölkopf. Ranking on data manifolds. In *Advances in Neural Information Processing Systems*, 2004.

[82] X. Zhou and T. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8(6):536–544, 2003.

[83] X. Zhou, K. Yu, T. Zhang, and T. Huang. Image classification using supervector coding of local image descriptors. In *European Conference on Computer Vision*, pages 141–154, 2010.