Incremental probabilistic Latent Semantic Analysis for Video Retrieval

Ruben Fernandez-Beltran, Filiberto Pla

Institute of New Imaging Technologies, Universitat Jaume I, SPAIN

Abstract

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 work 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.

Keywords:
Content-Based Video Retrieval, Latent Topics, probabilistic Latent Semantic Analysis (pLSA), Relevance Feedback, Information Retrieval

1. Introduction

With the expansion of new technologies, video collections are increasingly larger and more complex, therefore one of the biggest current challenges is

Preprint submitted to Image and Vision Computing *December 21, 2014*

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, 2, 3, 4]. 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 [5] 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 [6, 7]. However, these measures tend not to work properly when the multimedia data is rather complicated [8]. Other ranking algorithms are based on inductive learning [9, 10] 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 [11, 12]. 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 [13]. However, the so-called semantic gap [14] 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 [8]. 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 [15]), spatio-temporal (Spatial Temporal Interest Points - STIP [16]) or audio (Mel Frequency Cepstral Coefficients - MFCC [17]). The standard procedure to encode all this low-level information in feature vectors is the visual Bag of Words (vBoW) [18]. 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 [19] 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 [8].

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 [20]. In general, topic models provide for automatically organizing, understanding, searching and summarizing large electronic archives [21]. 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) [22] and Latent Dirichlet Allocation (LDA) [23]. 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 [24] 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 work, 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 [25] 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 [26] presented a pLSA version to address the problem of on-line event detection and Wu et al [27] 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 work 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 [28], 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 [28] 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 [29, 30, 31] 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 [20] and (2) the cosine similarity function used in [32].

The rest of the work is organized as follows. In Section 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 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 over-fitting problem. In Section 4, the experimental setting is described as well as the empirical results obtained by the retrieval systems [20] and [32], including a comparison among pLSA, LDA, FSTM and IpLSA in terms of video retrieval performance using the Consumer Columbia Video database [33] and the collection TRECVID 2007 [34] . Finally, Section 5 discusses the results and Section 6 draws the main conclusions arising from this work.

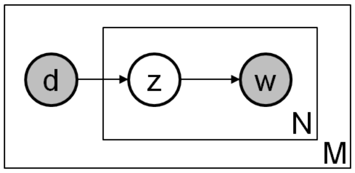


Figure 1: pLSA model: d represents the documents, z the topics (hidden variable) and w the words. M is the number of documents of the collection and N the number of words in the document d.

2. Background

Latent Semantic Analysis (LSA) [35] 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) [22] 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 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 [23] 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 [23] 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 [36] 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 [29] reveals that pLSA outperforms the performance of LDA for automatic essay grading tasks in a collection with less than 150 documents. In [31], the authors suggest that LDA does not have a competitive edge over pLSA especially for small training datasets and other authors [30] 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.

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 [22] in order to figure out the best way to extend the model efficiently.

The pLSA implementation of Hoffmann [22] 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:

$$C_{time}(pLSA) = O(\underbrace{I}_{Iters}(\underbrace{VMK}_{Estep}+\underbrace{VMK}_{Mstep}) = O(IVMK) \qquad (1)$$

$$C_{space}(pLSA) = O(\underbrace{VMK}_{p(z|d,w)}+\underbrace{VK}_{p(w|z)}+\underbrace{KM}_{p(z|d)}) = O(VMK) \qquad (2)$$

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 [37], and other ones use an evaluation function to decide the best number of topics [38]. 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 [39] 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.

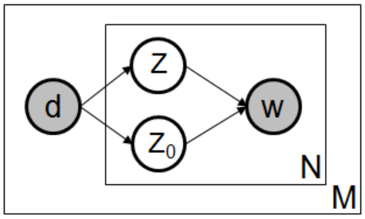


Figure 2: IpLSA model: d represents the new documents to add into the database, z_0 the initial topic structure of the previous stage, z the new extracted topics to describe the new documents and w the words. Eventually, N represents the number of words of the document d and M the number of new documents to add into the database.

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 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:

- 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.

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. (3) and later the log-likelihood Eq. (4) in terms of the joint probability distribution:

$$p(w,d,z) = p(w|z,z_0)p(z,z_0|d)p(d) \tag{3}$$

$$\mathcal{L} = \sum_w \sum_d n(w,d) \log p(w,d) \tag{4}$$

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:

$$E = \sum_{w}\sum_{d} n(w,d)(\mathcal{Z} + \mathcal{Z}_0) \tag{5}$$

$$\mathcal{Z} = \sum_{z} p(z|w,d)\log[p(w|z)p(z|d)p(d)] \tag{6}$$

$$\mathcal{Z}_0 = \sum_{z_0} p(z_0|w,d)\log[p(w|z_0)p(z_0|d)p(d)] \tag{7}$$

Introducing the normalization constraints of the parameters $p(z|d)$, $p(z_0|d)$ and $p(w|z)$ in expression (5) by inserting the appropriate Lagrange multipliers α and β:

$$H = E + \sum_{z}\alpha\left[1 - \sum_{w} p(w|z)\right] + \sum_{d}\beta\left[1 - \left(\sum_{z} p(z|d) + \sum_{z_0} p(z_0|d)\right)\right] \tag{8}$$

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

$$p(z|d) = \frac{\sum_{w} n(w,d)p(z|w,d)}{\sum_{z}\sum_{w} n(w,d)p(z|w,d) + \sum_{z_0}\sum_{w} n(w,d)p(z_0|w,d)} \tag{9}$$

$$p(z_0|d) = \frac{\sum_{w} n(w,d)p(z_0|w,d)}{\sum_{z}\sum_{w} n(w,d)p(z|w,d) + \sum_{z_0}\sum_{w} n(w,d)p(z_0|w,d)} \tag{10}$$

$$p(w|z) = \frac{\sum_{d} n(w,d)p(z|w,d)}{\sum_{w}\sum_{d} n(w,d)p(z|w,d)} \tag{11}$$

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:

$$p(z|w,d) = \frac{p(w,d,z)}{p(w,d)} = \frac{p(w|z)p(z|d)}{\sum\limits_{z} p(w|z)p(z|d) + \sum\limits_{z_0} p(w|z_0)p(z_0|d)} \tag{12}$$

$$p(z_0|w,d) = \frac{p(w,d,z_0)}{p(w,d)} = \frac{p(w|z_0)p(z_0|d)}{\sum\limits_{z} p(w|z)p(z|d) + \sum\limits_{z_0} p(w|z_0)p(z_0|d)} \tag{13}$$

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. (12) and (13)) and the M-step (Eqs. (9) and (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 (4)) between two consecutive iterations and a maximum of 1000 EM iterations.

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 [22], 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. (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.

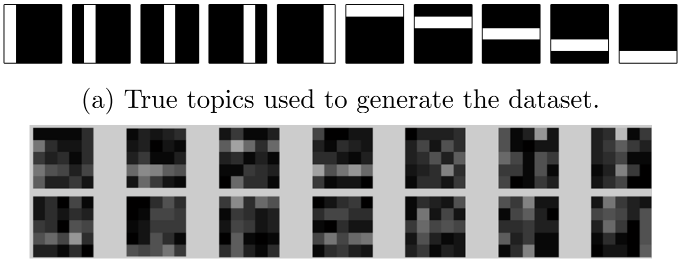
The incremental IpLSA model tries to reduce the over-fitting problem of the global pLSA usage in two ways: (1) using only the new documents d to extract the new set of topics z and (2) avoiding learning topics which have been extracted in the previous stage. The standard pLSA uses the documents of the 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.

4. Experiments

This section presents the experimental part of the work. First (Section 4.1), we use a synthetic dataset in order to highlight how the proposed method works. Subsequently, Section 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.

4.1. Toy Dataset

The toy dataset [40] 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 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 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 3: Toy Dataset.

Let us start by showing the behavioural differences between pLSA and IpLSA by means of Figure 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

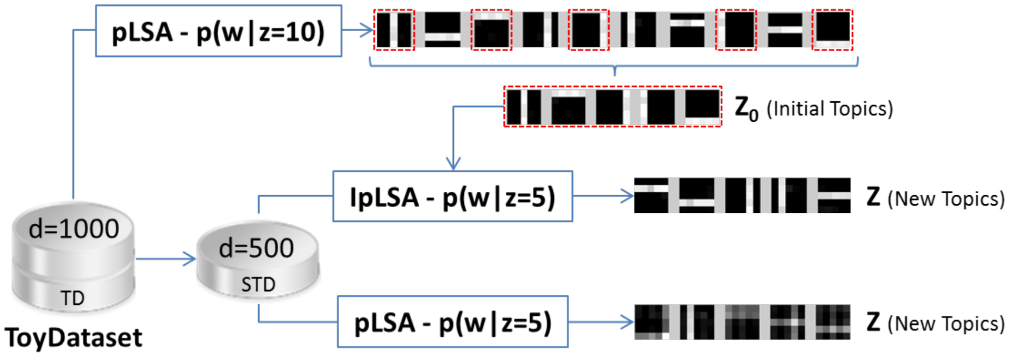


Figure 4: IpLSA vs pLSA.

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 4).

Another practical consideration is the difference between pLSA-based models and LDA. In figure 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 [29, 30, 31]. Therefore, despite the fact that LDA provides a more general framework than pLSA, in some applications in which we do not have too much information about

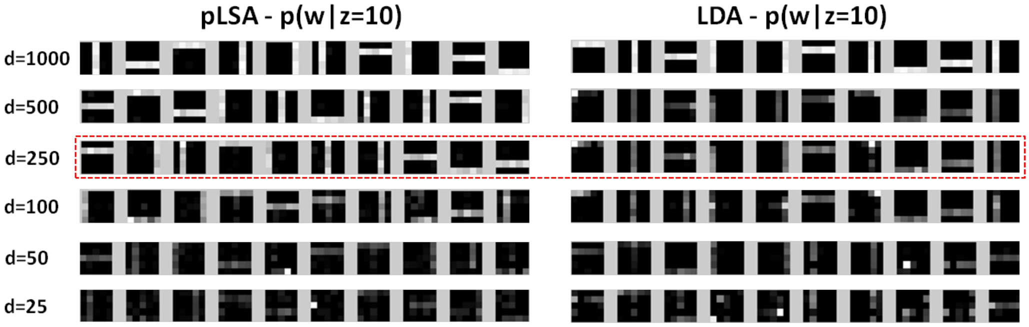


Figure 5: pLSA vs LDA.

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.

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.

4.2.1. Relevance Feedback simulations

In order to evaluate the effectiveness of the considered topic models for CBVR, we use the Relevance Feedback scheme proposed in [20] with two different ranking functions: the probabilistic ranking function presented in [20] and the cosine similarity function used in [32]. 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 1.

Table 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

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 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 [20] and the cosine similarity function used in [32]. 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 environ-

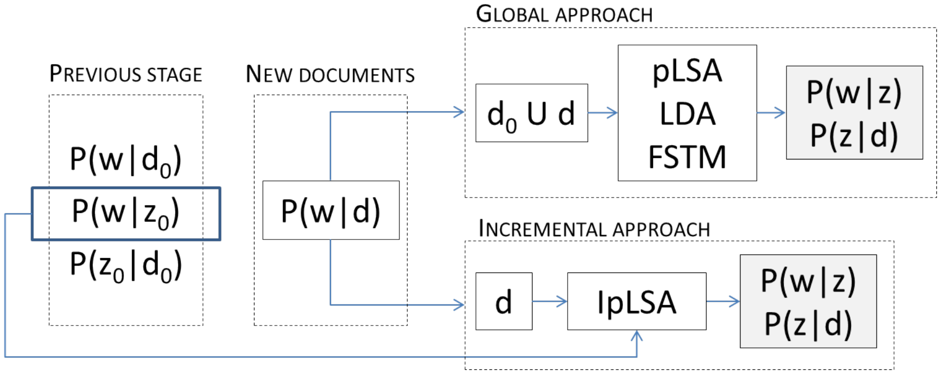


Figure 6: Stages used for the experiments.

ment. 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 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.

4.2.2. 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 [37, 38], 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.

4.2.3. Consumer Columbia Video (CCV) database

The Columbia Consumer Video (CCV) database [33] 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 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 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

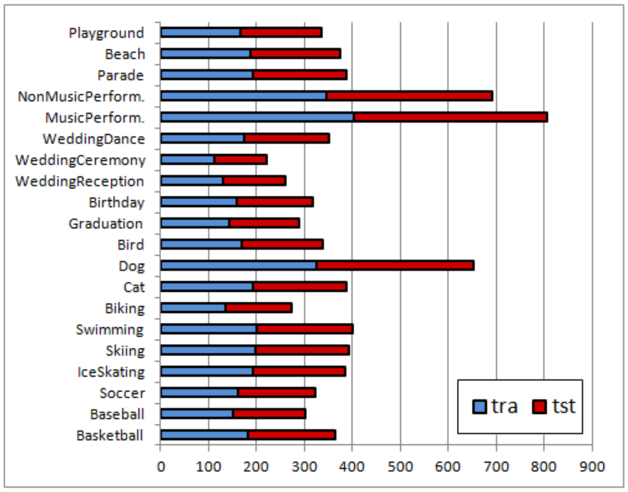


Figure 7: Samples per class of the CCV database.

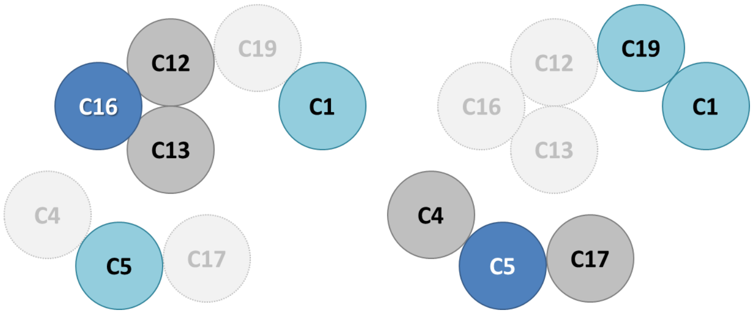
has been considered the initial set of samples d_0 and the test partition the new set of samples d to be retrieved.

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 8 shows a schematic representation of the distance among the centroid of the considered classes.



(a) Class C16.

(b) Class C5.

Figure 8: Scheme of the distance among the considered class centroids.

4.2.4. Video collection TRECVID 2007

The TRECVID 2007 dataset [34] 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 [41] 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 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.

4.2.5. 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.

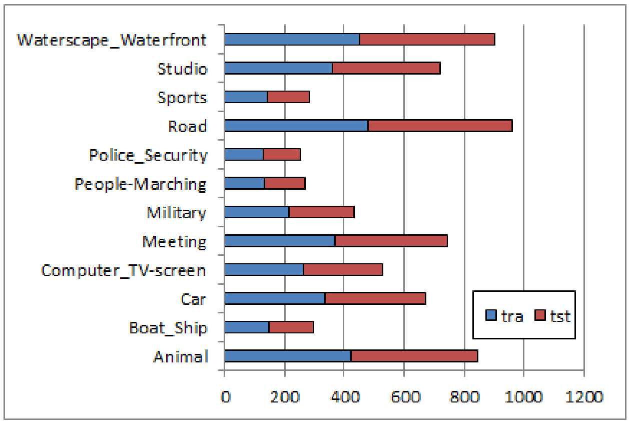


Figure 9: Subset of TRECVID 2007.

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 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 10, both videos 48.d

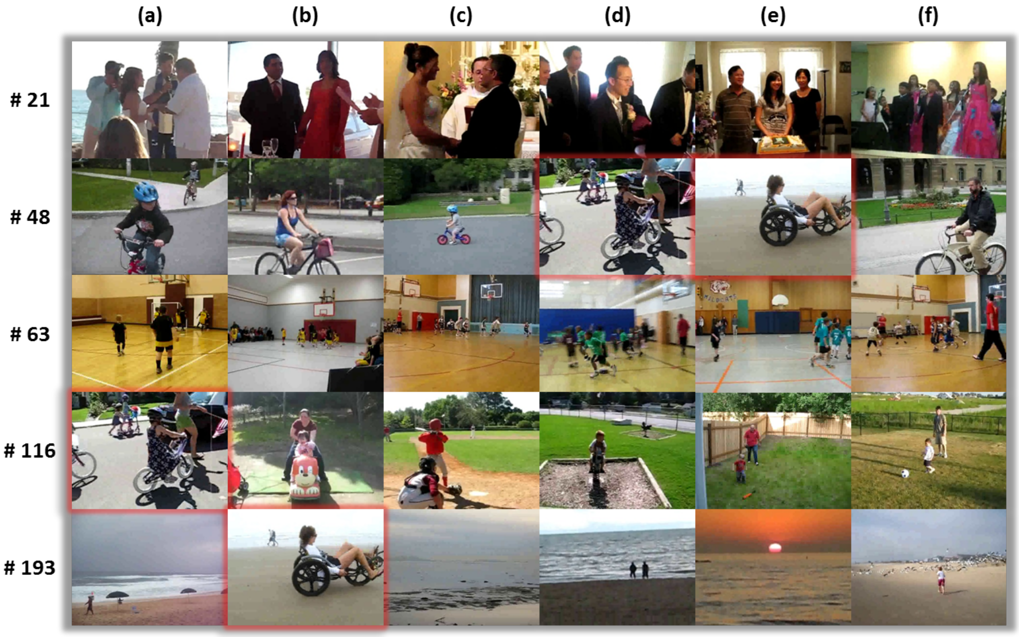


Figure 10: The six most probable documents of five topics from CCV.

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.

4.2.6. Results

Table 2 shows the abbreviation used for each partition as well as the details for the global approach, the incremental approach and the retrieval

Table 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 $d = 3455$

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 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

Table 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

of the models have been discussed in section 4.2.2. 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 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 4 contains the average precision of the experiments and Table 5 shows the F_1 measure calculated as $2(Precision * Recall)/(Precision + Recall)$.

Table 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: Video retrieval results: **$\mathbf{F_1}$ 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**

4.2.7. 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 [42] 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 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 6 shows the statistical differences among the used topic models with the LTR ranking function and Table 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 $\bullet$ indicates that the model in the row significantly outperforms the model in the column, and the symbol $\circ$ indicates that the model in the column significantly surpasses the model in the row.

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

Table 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				-				-		●		-		●		-

Table 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		●	●	-		●	●	-		●		-		●		-

discussion is presented.

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.

5.2. 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 over-fitting. 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.

5.3. 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 samples 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.

5.4. General issues

According to the results, we agree with [31] 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 [29] 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.

6. Conclusions and Future Work

This work 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 [20] and [32], 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. Related 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.

Acknowledgements

This work was partially supported by FPU-AP-2009-4435 from the Spanish Ministry of Education and PROMETEOII/2014/062 project from Generalitat Valenciana.

References

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

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

[3] L. Ying, Z. Dengsheng, L. Guojun, M. Wei-Ying, A survey of content-based image retrieval with high-level semantics, Pattern Recognition 40 (2007) 262–282.

[4] M. Cord, P. H. Gosselin, S. Philipp-Foliguet, Stochastic exploration and active learning for image retrieval, Image and Vision Computing 25 (2007) 14–23.

[5] G. Chechik, V. Sharma, U. Shalit, S. Bengio, Large scale online learning of image similarity through ranking, Journal of Machine Learning Research 11 (2010) 1109–1135.

[6] G.-H. Liu, Z.-Y. Li, L. Zhang, Y. Xu, Image retrieval based on micro-structure descriptor, Pattern Recognition 44 (2011) 2123–2133.

[7] M. Arevalillo-Herrez, F. J. Ferri, An improved distance-based relevance feedback strategy for image retrieval., Image Vision and Computing (2013).

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

[9] S. Tong, E. Chang, Support vector machine active learning for image

retrieval, in: ACM International Conference on Multimedia, pp. 107–118.

[10] K. Tieu, P. Viola, Boosting image retrieval, International Journal of Computer Vision 56 (2004) 17–36.

[11] D. Zhou, J. Weston, A. Gretton, O. Bousquet, B. Schölkopf, Ranking on data manifolds, in: Advances in Neural Information Processing Systems.

[12] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, Y. Pan, A multimedia retrieval framework based on semi-supervised ranking and relevance feedback, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (2012) 723–742.

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

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

[15] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.

[16] I. Laptev, On space-time interest points, International Journal of Computer Vision 64 (2005) 107–123.

[17] C. V. Cotton, D. P. W. Ellis, Audio fingerprinting to identify multiple videos of an event, in: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2386–2389.

[18] J. Sivic, A. Zisserman, Video Google: A text retrieval approach to object matching in videos, in: International Conference on Computer Vision, volume 2, pp. 1470–1477.

[19] H. Wang, C. Schmid, Action recognition with improved trajectories, in: IEEE International Conference on Computer Vision, pp. 3551–3558.

[20] R. Fernandez-Beltran, F. Pla, An interactive video retrieval approach based on latent topics, in: International Conference on Image Analysis and Processing, pp. 290–299.

[21] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (2012) 77–84.

[22] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Machine Learning 42 (2001) 177–196.

[23] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[24] C.-Y. Chiu, T.-H. Tsai, Y.-C. Liou, G.-W. Han, H.-S. Chang, Near-duplicate subsequence matching between the continuous stream and large video dataset, IEEE Transactions on Multimedia 16 (2014) 1952–1962.

[25] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: ACM International Conference on Machine Learning.

[26] T.-C. Chou, M. C. Chen, Using incremental plsi for threshold-resilient online event analysis, IEEE Transactions on Knowledge and Data Engineering 20 (2008) 289–299.

[27] H. Wu, Y. Wang, X. Cheng, Incremental probabilistic latent semantic analysis for automatic question recommendation, in: ACM conference on Recommender systems, ACM, 2008, pp. 99–106.

[28] K. Than, T. B. Ho, Fully sparse topic models, in: European Conference on Machine Learning.

[29] T. Kakkonen, N. Myller, E. Sutinen, J. Timonen, Comparison of dimension reduction methods for automated essay grading, Educational Technology & Society 11 (2008) 275–288.

[30] X. Yi, J. Allan, A comparative study of utilizing topic models for information retrieval, in: European Conference on IR Research on Advances in Information Retrieval.

[31] Y. Lu, Q. Mei, C. Zhai, Investigating task performance of probabilistic topic models: An empirical study of plsa and lda, Information Retrieval 14 (2011) 178–203.

[32] R. Zhang, Z. Zhang, Effective image retrieval based on hidden concept discovery in image database, IEEE Transactions on Image Processing 16 (2007) 562–572.

[33] Y. G. Jiang, G. Ye, S. F. Chang, D. Ellis, A. C. Loui, Consumer video understanding: a benchmark database and an evaluation of human and machine performance, in: ACM International Conference on Multimedia Retrieval.

[34] S. Ayache, G. Qunot, Trecvid 2007 collaborative annotation using active learning, in: Proceedings of the TRECVID 2007 Workshop.

[35] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of Machine Learning Research 41 (1990) 391–407.

[36] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, D. M. Blei, Reading tea leaves: How humans interpret topic models, in: Advances in Neural Information Processing Systems 22, 2009, pp. 288–296.

[37] Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei, Hierarchical dirichlet processes, Journal of the American Statistical Association 101 (2004).

[38] R. Arun, V. Suresh, C. E. Veni Madhavan, 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.

[39] R. Fernandez-Beltran, R. Montoliu, F. Pla, Vocabulary reduction in bow representing by topic modeling, in: Iberian Conference on Pattern Recognition and Image Analysis, pp. 648–655.

[40] T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences 101 (2004) 5228–5235.

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

[42] J. Urbano, M. Marrero, D. 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, pp. 925–928.