An Interactive Video Retrieval Approach Based on Latent Topics*

Rubén Fernández-Beltran and Filiberto Pla

Institute of New Imaging Technology, Jaume I University, Castellón, Spain
{rufernan,pla}@uji.es

Abstract. The huge collections of unconstrained videos have amplified the so-called semantic gap for content-based video retrieval. Therefore, new efficient approaches with higher generalisation power are needed. In this work, we present an interactive video retrieval approach based on latent topics to cope with the semantic gap in an efficient way. A supervised Symmetric extension of probabilistic Latent Semantic Analysis model is presented (sSpLSA). Then, this model is adapted to an on-line interactive information retrieval problem and it is applied to a video retrieval framework based on explicit short-term Relevance Feedback (RF) where queries are inside the database. Finally, several retrieval simulations using the Consumer Columbia Video (CCV) database are performed to compare the proposed approach with a distance-based RF baseline.

Keywords: Content-based video retrieval, relevance feedback, latent topics.

1 Introduction

In recent years the great expansion of video collections has boosted the video media as the biggest resource for information exchange on Internet. In this situation, one of the most important challenges is how to retrieve user relevant data from this huge amount of information in an effective way. Traditional search engines retrieved videos using only textual information. This limits the capabilities of the retrieval process due to the fact that they were devoid of the capacity to understand media contents. Over the last years, Content-Based Video Retrieval (CBVR) has become a very important area of research and several content-based video retrieval systems have been developed [1]. Recent works deal with video content at various levels: low-level descriptors and concepts detector [14]. However, recent results have shown that this approach does not scale adequately when the number of trained concepts increases [16]. Therefore, this information is not enough for discriminating across different multimedia content when unconstrained videos are considered [11].

* This work was partially supported by FPU-AP-2009-4435 from the Spanish Ministry of Education, PROMETEO/2010/028 project from Generalitat Valenciana and P1-1B2010-27 project from the Plan de Promoció de la Investigació UJI.

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 290–299, 2013.

The main problem in content-based video retrieval is the so-called semantic gap between computable low-level features and semantic concepts that users want to retrieve [15]. Most of current approaches use classifiers to attempt filling the semantic gap. Thus, they use features without semantic meaning to represent semantic concepts. For this reason, current approaches in video retrieval are only reliable under a specific domain (constrained videos) and they also need an extensive computation. One of the promising research directions that might improve video retrieval performance is based on using other kind of representation methods beyond conventional bag-of-words (BoW). In this field, topic models based techniques [3] can be used to characterize the samples in a higher level of semantic significance. Topic models have been used obtaining satisfactory results in many areas, such as text categorization [6], image recognition [13] or video classification [5]. Two of the most used algorithms are probabilistic Latent Semantic Analysis (pLSA) [9] and Latent Dirichlet Allocation (LDA) [4].

The presented work is focused on providing an interactive video retrieval approach based on latent topics, in order to cope with the semantic gap challenge using topic model representations. In this work, we propose a new point of view for the retrieval process. The problem becomes in a class discovery problem using a supervised latent topic model. The underlying goal is twofold, on the one hand, the use of latent topic representation is intended to express the data in a characterization space that is semantically nearer to the user's concepts. On the other hand, the use of the probabilistic ranking approach to be developed based on topic model representations is intended to be computationally effective for class discovering tasks in large scale information retrieval systems.

The rest of the paper is organized as follows. Section 2 discusses the background of the work. In Section 3, we present a supervised Symmetric extension of pLSA model (sSpLSA) in order to relate class labels and topic model characterisation. Subsequently, we adapt this supervised model to an on-line interactive information retrieval system where the problem becomes in a new class discovery problem by using the feedback provided by the user. Later, we apply this approach to a video retrieval framework based on explicit short-term Relevance Feedback (Section 4). In Section 5, we perform several experiments using one of the most challenging dataset for unconstrained video retrieval (Consumer Columbia Video database) and we compare proposed approach with respect a distance-based baseline. Finally, Section 6 draws the main conclusions arisen from this work and notes the future work.

2 Related Work: Content-Based Video Retrieval

Popular video retrieval approaches [1] have been focused on shots retrieval, that is, short pieces of video with homogeneous content. Each shot is summarized in a feature vector (descriptor) using combinations of low-level feature values or concepts. Then, these descriptors are used to rank the database and return the most relevant videos according to user's query. Therefore, two issues are essential for video retrieval performance: the descriptor and the ranking algorithm.

Regarding to the ranking algorithms, there are several methods to rank the database samples according to their relevance to the query. One of the most frequently used in multimedia retrieval has been distance-based ranking (e.g. [10]). However, the measures of distance or similarity are not able to work properly when the multimedia data are too complicated. Other ranking algorithms are based on inductive learning which typically uses a bank of classifiers to represent the set of possible events to test [14]. But, the performance of this approach depends on the training data and also the number of concepts to retrieve are very constrained for general applications. An alternative ranking methods are based on transductive ranking. That is, they use the data distribution of all the samples in order to improve the output ranking in the retrieval process. One of the most representative is Manifold Ranking (MR) [17] which rank the data with respect to the intrinsic data distribution.

Several of these approaches have shown to be successful in video retrieval process when they are used on edited videos with a specific number of concepts [2]. However, with the popularity of hand-held video recording devices, a huge amount of videos are captured by non-professional users under unconstrained conditions [11]. For unconstrained videos, the visual representation variability of a concept is so high that these approaches do not scale adequately when the number of trained concepts increases [16]. This amplifies the semantic gap challenge and demands for new capabilities in video retrieval with a higher generalization power. Moreover, video collections contain increasingly samples, therefore new efficient approaches are needed to deal with this amount of data in real life applications.

In order to bridge the semantic gap Relevance Feedback (RF) techniques can be used where the user collaborates with the retrieval system. The RF can be defined inside of the on-line learning paradigm [7]. The system learns from the user feedback to update its internal representation for the user preferences and the user interacts with the system until a prefixed number of relevant items are at the top of the system output. The feedback can be explicit where users are asked to assess the relevance of the videos, or implicit where the system is able to extract indicators of relevance according to the user's interaction (e.g. playing a video can be interpreted as implicit indicator of relevance because user is interested in corresponding sample). In addition, RF can be divided into two groups, short-term and long-term. Short-term only considers the information provided by the current user, and long-term also uses the feedback of previous users.

3 An Interactive Information Retrieval Approach Based on sSpLSA

Extracting hidden information in a data set to map it into a feature space that may fill the gap between sensory representation and higher level understanding is a key factor in many data analysis applications. In order to develop a probabilistic approach for video retrieval based on latent topic models, the model

chosen to be explored is based on the unsupervised pLSA model [9], a relaxed version of LDA model [4], which will be extended to a supervised model by adding the observed random variable corresponding to class labels (y). In this case, the approach is directed to a similar scenario than the single author topic model, used by Fei-Fei and Perona [8] in the framework of a LDA-based model. The here proposed model is a supervised symmetric pLSA (sSpLSA) (figure 1). This model corresponds to an extension on the symmetric unsupervised pLSA parameterization introduced by Hoffman [9]. This supervised extension allows us to handle the query class, which is unknown, in order to develop a process to estimate the probability that samples belong to this unknown class.

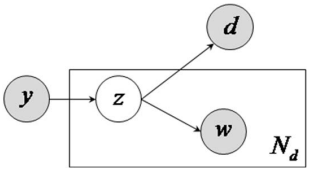

Fig. 1. Graphical representation of the sSpLSA model. y is the class, z the topic (hidden variable), w the word, d the document and N_d the number of words of d.

Based on the sSpLSA model, the next step is to extend it to an interactive on-line framework approach for information retrieval. An on-line user interactive problem can be formulated as follows: a user query at a given stage of the interactive process is represented by a query class y' of documents the user has in mind and he/she is looking for, and a set of query documents d' that represents instances of documents belonging to the query class. The query document set $d' = 1,...,D'$ is a dynamic set that changes during the interactive process, according to the positive documents the user provides as feedback from the iterations in the searching process.

Let us suppose that a learning process has been previously carried out with the documents d in the database, which will be expressed in a set of latent topics as $p(z|d)$. The latent topics may have been learned either in a supervised or unsupervised way. In the case of a supervised learning, this may have been done according to a pre-defined set of classes or concepts, different from the query class at each interactive user process, which is unknown.

The goal at each iteration in the interactive process is to provide a guess about the probability that a document of the database belongs to the query class, that is, we look for a estimation of the class conditional probability $p(y'|d)$. This probability will allow us to establish a retrieval ranking for the documents in the database. According to the sSpLSA model (figure 1), this probability could be estimated from the present user's query as follows. Let us express the conditional probability $p(y'|d)$ by marginalizing over topics, that is:

$$p(y'|d) = \frac{p(y',d)}{p(d)} = \frac{\sum_w \sum_z p(w,d,z,y')}{p(d)} = \frac{\sum_w \sum_z p(w|z)p(d|z)p(z|y')p(y')}{p(d)} \quad (1)$$

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 in function of the parameters of the model as follows by marginalizing over the query set $d' = 1,...,D'$:

$$\begin{aligned} p(z|y') &= \sum_{d'} p(z,d'|y') = \sum_{d'} p(z|d',y')p(d'|y') \\ &\approx \sum_{d'} p(z|d')p(d'|y') = \sum_{d'} \frac{p(z|d')p(y'|d')p(d')}{p(y')} \end{aligned} \quad (2)$$

Topics have been calculated in a previous step using the documents of the database, and therefore we can consider that topic descriptions do not depend on the new class y' (query class) in order to estimate the conditional probability $p(z|d',y') = p(z|d')$. Inserting (2) in (1), applying Bayes rule and assuming the normalization constraint $\sum_w p(w|z) = 1$, the expression to estimate the class conditional probability $p(y'|d)$ can be expressed as follows:

$$p(y'|d) \approx \sum_z p(z|d) \left[\sum_{d'} p(d'|z) \right] \quad (3)$$

In this expression, $p(z|d)$ stands for the parameters of the database, and $p(d'|z)$ is the probability that a given topic z belongs to the query document d', which could be estimated from the same learning model used to estimate the database document description in topics, for instance from a maximum log-likelihood approach. Expression (3) would allow us to infer the class conditional probabilities $p(y'|d)$ in a simple and fast way from the document database described in topics. It has the advantage of being very efficient computationally and easy to update for subsequent iterations in the interactive process that dynamically changes the query documents set $d' = 1,...,D'$ that defines the user's query class y'. In addition, latent topics would allow a representation nearer to the user's semantic understanding.

The ranking process is made as follows. First of all, z topics are extracted from the database using some topic extraction method (like pLSA [10], LDA [11] or any other topic model algorithm). Then, each document d and also the query documents d' are represented according these topics. To represent the samples in topics given a set of topics, we have used a maximum log-likelihood approach as in [9]. Later, the database is sorted according the probability that a sample belongs to the query class using expression (3). This equation has two terms: the document d (which is expressed in z topics) and the query d' (sum of the probabilities of the query documents given the z topics, obtained by the maximum log-likelihood approach).

4 Application to Video Retrieval Framework

In the previous section, a model for interactive information retrieval has been introduced which can be applied to several information retrieval problems. The goal in this work is the application for unconstrained video retrieval. Specifically, the main objective is to adapt the general model to an interactive video retrieval framework based on explicit short-term relevance feedback. We are going to model a simplified user interaction scenario (simulation) to evaluate the effectiveness of proposed approach. The relevance feedback simulation can be divided into two parts: Firstly, an initial search query is triggered and a simulated user provides feedback on retrieved results. Secondly, the initial query is expanded with the items extracted from the feedback and a new query is triggered. The grade of feedback quality depends on the simulated user reliability. Therefore, the video retrieval framework based on sSpLSA has to include the following functional requirements:

- *Video representation*: The video collection has to be described in latent topics $p(z|d)$. First of all, the videos are represented in low level features which include static or spatio-temporal information. From this low level representation, K latent topics are extracted and each video sample is represented according these topics.
- *Query initialization*: The system has to be initialized with a query which contains Q video samples of the dataset. This set of videos defines the concept that user wants to retrieve (target) and it has to be expressed according the K topics extracted from the database.
- *Feedback information*: The video retrieval process is an iterative method of I iterations. In each one, the system shows a video ranking with S retrieved videos (scope). From these plausible videos, user selects the P positive samples according to query concept (target).
- *Propagate feedback*: The system uses the positive samples to enlarge the query $(Q + P)$. Therefore, the new query has more video samples and the system is able to refine the retrieved videos for the next iteration.

The considered target for the simulation has been each video class of the database (each one of the C classes), that is, the query is initialized with Q random videos of one class and the simulation has to be able to retrieve videos of this class. Queries can be initialized using samples inside the database or external samples which are not in the database. The only requirement is that they have to be expressed using the topics extracted from the database. In this simulation, queries are initialized with samples from the database. Moreover, this process is repeated R times per class to obtain relevant statistical results. The simulation video retrieval process is presented in Algorithm 1.

It should be noted that the S most probable videos would be inspected and checked by the user in a real video retrieval system. However, this process is automatically made in the proposed simulation which assumes user reliability of 100%.

Algorithm 1. Video Retrieval Simulation for $DATASET$

Require: C=classes, R=repetitions, Q=initSize, I=iterations, S=scope.

for class c in C **do**
  for repetition r in R **do**
    Initialize $QUERY$ with Q random samples of the class c
    $REST = DATASET - QUERY$
    for iteration i in I **do**
      for video v in $REST$ **do**
        $p(y'|v) = \sum_z p(z|v) \sum_{d'} p(d'|z)$
      end for
      Rank $REST$ in descending order of probability
      P = Videos which belong to class c from the S top ranking
      Enlarge $QUERY$ adding P
      Update $REST$ subtracting P
    end for
  end for
end for

5 Experiments

5.1 CCV DataSet

The video dataset selected for the experiments is Columbia Consumer Video Database (CCV) [12]. This recent video collection is a benchmark for consumer video analysis. It contains 9.317 YouTube videos over 20 semantic categories. The total number of video hours is 210 and the average length of the videos is 80 seconds. Also, different video descriptors are available to run experiments with CCV dataset (SIFT, STIP and MFCC). According to the classification accuracy for the CCV database [12], SIFT descriptor achieves better average precision than STIP and MFCC. Furthermore, the combination of them does not improve the accuracy in a significant way. For this reason, we have decided to use the SIFT descriptors provided by the authors of the dataset as a preliminary experiment to test the proposed approach. The vocabulary 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 2x2 image blocks, which makes a total of 5000 words.

In this corpus, there are samples with null content that are been removed for the experiments. Furthermore, the samples with no annotation have been eliminated too. For the remaining samples, samples labelled with more than one category have been replicated one for each class. Therefore, we have considered a total of 7.846 video samples annotated in 20 classes.

5.2 Validation Set Up

As it has been mentioned in Section 4, the simulation of video retrieval framework has several functional requirements with parameters. These parameters have to

be defined and established for the experiments. We have chosen the parameters of the simulation thinking about a real user, that is, the size of the initial query, the number of retrieval iterations and the number of selected videos for propagation (scope) have to be in an appropriate range for user comfort in a real video retrieval system. Therefore, the selected configuration for the simulation process according to the functional requirements are the following:

- *Video representation*: LDA [4] procedure has been applied to the SIFT descriptors provided by the CCV authors [12], in order to express the database in a set of latent topics as $P(Z|D)$. Several values of the number of topics have been considered for the experiments: $K = \{100,200,300,400,500\}$.
- *Query initialization*: For each simulation, an initial query has been initialized with $Q = \{1,2\}$ number of random video samples of the same class of the CCV database. These values have been selected due to the fact that users do not usually select more than one or two videos to initialize a search.
- *Feedback information*: A maximum of $I = \{5\}$ retrieving iterations has been executed using relevance feedback, which is acceptable to the user.
- *Propagate feedback*: In the query updating process, the number of the top selected items has been established to $S = \{20,40\}$. We have considered that user only examines one or two screens of videos to find items of the query class considering a screen size of 20 videos. Moreover, we have assumed that user marks all the positive videos of the query class in the inspected set.

The queries have been initialized with random samples of the same class and this process has been repeated $R = \{500\}$ times. To evaluate the retrieval results the precision over the five iterations have been calculated.

5.3 Relevance Feedback Based on Distance Ranking

In order to evaluate proposed approach we have implemented and executed a baseline method for simulation video retrieval. This baseline algorithm has a RF structure similar to the simulation showed in Section 4, but with two main differences. On the one hand, it uses the original BoW representation of the dataset (SIFT features). On the other hand, the database is sorted by the minimum Euclidean distance to the query (when the query has more than one sample, this distance is the average).

5.4 Experimental Results and Discussion

According to the validation setup (Section 5.2), four main configurations have been used to run the experiments. For these scenarios, several values of number of topics have been used. In order to evaluate the results, two metrics have been used: overall mean precision and execution time. Figure 2 shows the results for each scenario considering the baseline and the proposed approach.

Over the different scenarios, results show that the best retrieval performance has been obtained using 500 topics. It should be noted that the size of the

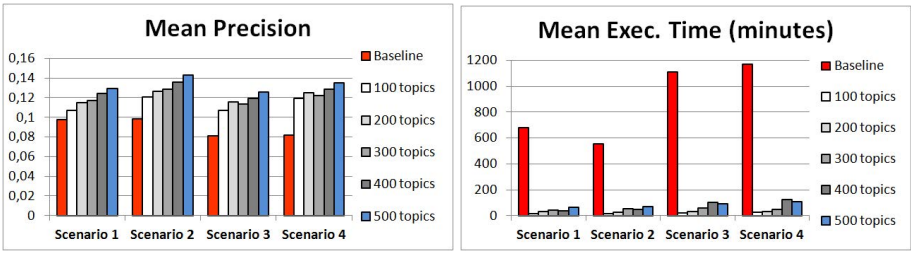

Fig. 2. Results for CCV database. Scenario 1 {Q=1, S=20}, Scenario 2 {Q=2, S=20}, Scenario 3 {Q=1, S=40} and Scenario 4 {Q=2, S=40}. Mean precision (right) and execution time in minutes (left).

initial query Q and the size of the scope S are important factors to the retrieval performance. With a bigger initial query the mean precision rises, but with a larger scope the precision drops. The best precision has been obtained in scenario 2 and the best mean of retrieved videos in scenario 4. Comparing the proposed approach with the baseline, the first one obtains a greater improvement with a bigger initial query and lesser loss with a larger scope.

Regarding to the computational time, results show a remarkable performance of the proposed topic-based approach. Note that the topic-model procedure is applied over the dataset as an off-line processing model, therefore it does not do any extra processing on queries. Furthermore, topic-model representation is able to significantly reduce the dimensionality of the samples. In other words, LDA model is estimated off-line once, and then the proposed approach can process queries much faster than baseline with a greater precision improvement.

6 Conclusions and Future Work

In this work, we have presented an alternative approach for interactive on-line video retrieval tasks based on latent topic models. The supervised latent topic model sSpLSA have been presented and adapted to a video retrieval framework based on relevance feedback. Several experiments have been done using a simulation-based methodology and a distance-based RF algorithm as baseline. The results provide evidences about the viability of proposed approach. The main conclusion that arises from the work is the importance of topic models to attempt filling the semantic gap for video retrieval. Topic models are able to extract hidden patterns of a data set and these patterns can be used to provide a higher level understanding. Besides, the proposed approach shows an important execution time reduction with respect the distance-based baseline. Although results are encouraging, much more experimental evidence is needed to really assess the properties and quality of the proposed approach. In particular, further work is directed to compare its performance with other recent information retrieval approaches and to implement a testing protocol in order to assess proposed approach using a user-based methodology.

References

1. Trecvid, http://trecvid.nist.gov/
2. The challenge problem for automated detection of 101 semantic concepts in multimedia (2006)
3. Blei, D.: Probabilistic topic models. Communications of the ACM 55(4), 77–84 (2012)
4. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research 3(4-5), 993–1022 (2003)
5. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via 2009. In: European Conference on Computer Vision (2009)
6. Brants, T., Chen, F., Tsochantaridis, I.: Topic-based document segmentation with probabilistic latent semantic analysis. In: International Conference on Information and Knowledge Management (2002)
7. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. Journal of Machine Learning Research 11, 1109–1135 (2010)
8. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR (2005)
9. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42(1-2), 177–196 (2001)
10. Huang, J., Kumar, S., Mitra, M., Zhu, W., Zabih, R.: Image indexing using color correlograms. In: IEEE Int. Conf. Computer Vision and Pattern Recognition (1997)
11. Jiang, Y., Bhattacharya, S., Chang, S., Shah, M.: High-level event recognition in unconstrained videos. International Journal of Multimedia Information Retrieval 2, 73–101 (2013)
12. Jiang, Y., Ye, G., Chang, S., Ellis, D., Loui, A.: Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: Proceedings ACM International Conference on Multimedia Retrieval, ICMR (2011)
13. Monay, F., Gatica-Perez, D.: On image auto-annotation with latent space models. In: ACM International Conference on Multimedia (2003)
14. Ren, W., Singh, S., Singh, M., Zhu, Y.: State-of-the-art on spatio-temporal information-based video retrieval. Pattern Recognition 42(2), 267–282 (2009)
15. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
16. Snoek, C., Worring, M.: Concept-based video retrieval. Foundations and Trends in Information Retrieval 4(2), 215–322 (2009)
17. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B.: Ranking on data manifolds. In: Advances in Neural Information Processing Systems. NIPS. MIT Press (2004)