Improved Video Content Indexing
By Multiple Latent Semantic Analysis

Fabrice Souvannavong, Bernard Merialdo, and Benoît Huet [*]

Département Communications Multimédias
Institut Eurécom
2229, route des crêtes
06904 Sophia-Antipolis - France
(souvanna, merialdo, huet)@eurecom.fr

Abstract. Low-level features are now becoming insufficient to build efficient content-based retrieval systems. Users are not interested any longer in retrieving visually similar content, but they expect retrieval systems to also find documents with similar semantic content. Bridging the gap between low-level features and semantic content is a challenging task necessary for future retrieval systems. Latent Semantic Analysis (LSA) was successfully introduced to efficiently index text documents by detecting synonyms and the polysemy of words. We have successfully proposed an adaptation of LSA to model video content for object retrieval and semantic content estimation. Following this idea we now present a new model composed of multiple LSA's (M-LSA) to better represent the video content. In the experimental section, we make a comparison of LSA and M-LSA on two problems, namely object retrieval and semantic content estimation.

1 Introduction

Because of the growth of numerical storage facilities, many documents are now archived in huge databases or extensively shared over the Internet. The advantage of such mass storage is undeniable. However the challenging tasks of multimedia content indexing and retrieval remain unsolved without the expensive human intervention to archive and annotate contents. Many researchers are currently investigating methods to automatically analyze, organize, index and retrieve video information [1, 2]. This effort is further stressed by the emerging MPEG-7 standard that provides a rich and common description tool of multimedia contents. It is also encouraged by Video-TREC [1] which aims at developing video content analysis and retrieval.

One of the major task is to bridge the gap between low-level features and the semantic content. To address this problem we propose a new robust method to index video shots. Based on our previous work on Latent Semantic Analysis (LSA) for object retrieval in [3] and semantic content estimation in [4], we present a new model that uses

[*] This research was supported by the EU project GMF4iTV under the IST-programme (IST-2001-34861)

[1] Text REtrieval Conference. Its purpose is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation. http://trec.nist.gov

Multiple Latent Semantic Analysis. LSA has been proven effective for text document analysis, indexing and retrieval [5]. The key idea is to map high dimensional count vectors to a lower dimensional space so-called latent semantic space. Some extensions to audio and image features were then proposed [6, 7]. In 1999, a probabilistic framework, called PLSA, was introduced for text document indexing in [8]. Then authors of [9] have recently made a comparison of both methods, i.e. LSA and PLSA, for image auto-annotation. They conclude that classic LSA model defined on a very basic image representation performs as well as much more complex state-of-the-art methods and outperforms PLSA. In this paper, we propose a new method that relies on Multiple Latent Semantic Analysis. The underlying idea is to group shots in order to better detect the latent semantic that locally resides in groups and that might be covered by a global approach.

The first part briefly presents the adaptation of LSA to video content modeling. Next we present Multiple Latent Semantic Indexing. Then experimental results are presented and commented to finish with the conclusion and future work.

2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) has been proven efficient for text document analysis and indexing. Contrary to early information retrieval approaches that used exact keyword matching techniques, it relies on the automatic discovery of synonyms and the polysemy of words to identify similar documents. We proposed in [3] an adaptation of LSA to model the visual content of a video sequence for object retrieval. We summarize the proposed solution in the following before presenting our new approach.

Let $V = \{S_i\}_{1<i<N}$ be a sequence of shots representing the video. Usually many shots contain the same information but expressed with some inherent visual changes and noise. Latent Semantic Analysis is a solution to remove the noise and find equivalences of the visual content to improve shot matching. It relies on the occurrence information of some features in different situations to discover synonyms and the polysemy of features. A classical approach is to use the singular value decomposition (SVD) of the occurrence matrix of features in shots to achieve this task. The content of shot i is described by a raw feature vector r_i, such as color histogram, gabor's energies, motion, ... However such feature vectors suffer from the loss of spatial information by keeping only global features. To overcome this problem, a more appropriate signature is used. First of all, frames composing shots are segmented into homogeneous regions. Similar regions, described by raw feature vectors, are then clustered in groups where they are finally mapped. The representative of each cluster is then called a visual term while the set of clusters is called the dictionary. Shots are now represented by the count of visual terms that describes the content of their regions. Let now denote q this new feature vector. The singular value decomposition of the occurrence matrix C of visual terms in shots gives:

$$C = UDV^t \quad \text{where} \quad U^tU = V^tV = I \tag{1}$$

With some simple linear algebra we can show that a shot (with a feature vector q) is indexed by p such that:

$$p = U^t q \tag{2}$$

U^t is then the transformation matrix to the latent space. The SVD allows to discover the latent semantic by keeping only the L highest singular values of the matrix D and the corresponding left and right singular vectors of U and V. Thus,

$$\hat{C} = U_L D_L C_L^t \quad and \quad p = U_L^t q \tag{3}$$

The latent space of size L is now ready for improved shot comparison thanks to the cosine measure. The number of singular values kept drives the LSA performance. On one hand if too many factors are kept, the noise will remain and the detection of synonyms and the polysemy of visual terms will fail. On the other hand if too few factors are kept, important information will be lost degrading performances. Unfortunately no solution has yet been found and only experiments allows to find the appropriate factor number.

In [4], we also noticed that the creation of a visual dictionary has a major disadvantage when dealing with many videos: it introduces differences between regions, i.e. due to the mapping, that might be too important. To diminish this side effect of the mapping, regions are mapped to their k-closest visual items and this conducted to a better performance.

3 Multiple Latent Semantic Analysis

Visual content carries an extremely rich information and LSA through SVD is a simple linear approach to model this diversity. We propose to introduce Multiple Latent Semantic Analysis (M-LSA) to describe the content and *locally* find its “latent semantic”. M-LSA involves two steps. Firstly we find K homogeneous partitions P_k in the feature space with respect to training shots. We then apply classical LSA to model each area and find K latent spaces. Secondly we index shots with respect to this decomposition and modeling of the feature space. We now thoroughly describe the method.

3.1 Local Latent Semantic Analysis

In order to improve the effect of SVD which is a linear transformation, we propose to apply the SVD locally in homogeneous partitions P_k of the feature space. This operation aims at detecting singular directions more accurately in local areas of the feature space. Thus we expect the LSA to be locally more efficient. Training shots allow to construct an efficient partition with respect to the content and one way to proceed is to use k-means algorithm on training shots.

Once the feature space is partitioned, we construct matrices C_k that contain the occurrences of visual terms in shots belonging to the partition k. We then apply classical LSA to model each area. Thus,

$$C_k = U_k D_k V_k^t \tag{4}$$

In this situation where k models are constructed, it is difficult to select the appropriate number of factors L_k kept per model. Empirically, we select a single value l, the selection coefficient, that gives the percentage of factor involved in each model to make the projection. Finally,

$$\hat{C_k} = U_{L,k} D_{L,k} V_{L,k}^t \tag{5}$$

$$where \quad L = l \times \min(\text{number of shots, number of visual terms})$$

In the following, $U_{L,k}$ is denoted $U_k(l)$ for convenience.

3.2 Indexing With Local LSA

From the presented decomposition and modeling of the feature space, we derive a new representation of shots. A direct approach is to index shots with respect to the partition where they are located. A shot signature is then composed of a partition number and its projection in the associated left singular space. Let $q \in P_k$ and $p = U_k^t(l)q$

$$sim(q,q') = \begin{cases} \cos(p,p') \text{ if } q' \in P_k & (p' = U_k^t(l)q') \\ -1 \qquad\quad \text{else} & \end{cases} \tag{6}$$

Unfortunately, shots can not be compared between partitions but only intra-partition comparisons are possible. This drawback becomes particularly important when looking for an object, i.e. only a subpart of the shot. In that case the query is not well classified in partitions that are homogeneous only with respect to complete shots. This suggests to project shots in all partitions, compare shots in each partition and then combine similarity measures to form a single-valued similarity measure. This second approach suffers from the fact that the projection errors can be high whereas the projected shots are close. In this case even if projected shots are similar, we can not guarantee that shots are similar. To take into account all these parameters, we derive a similarity measure of the form:

$$sim(q,q') = \max_i \{ (\cos(q,\hat{q}_i)\cos(q',\hat{q_i'}))^2 \cos(p_i,p_i') \} \tag{7}$$

$$\text{where} \quad p_i = U_i^t(l)q \quad \text{and} \quad \hat{q}_i = U_i(l)p$$
$$\text{and} \quad p_i' = U_i^t(l)q' \quad \text{and} \quad \hat{q_i'} = U_i(l)p'$$

The first two cosine functions measure the similarity between shots and their reconstruction when using the local model i. Cosine functions are used to obtain normalized values whatever values are taken by the selection coefficient. We then take the square of the product to diminish the impact of projection errors. The third cosine function measure the similarity between projected shots with the local model i. This similarity measure has the property to favor similar shots in a partition where they are both well projected. However, indexed shots need the value $\cos(q,\hat{q}_i)$ and the vector p_i, thus the selection coefficient can not be easily changed without computing again the value of the cosine. In future work we are going to evaluate a measure of the form:

$$sim(q,q') = \max_i \frac{\cos(p_i, p'_i)}{\frac{\|p_i^{\perp}\|}{\|p_i^{\perp}+p_i\|} \frac{\|p_i'^{\perp}\|}{\|p_i'^{\perp}+p'_i\|}} \tag{8}$$

4 Experiments

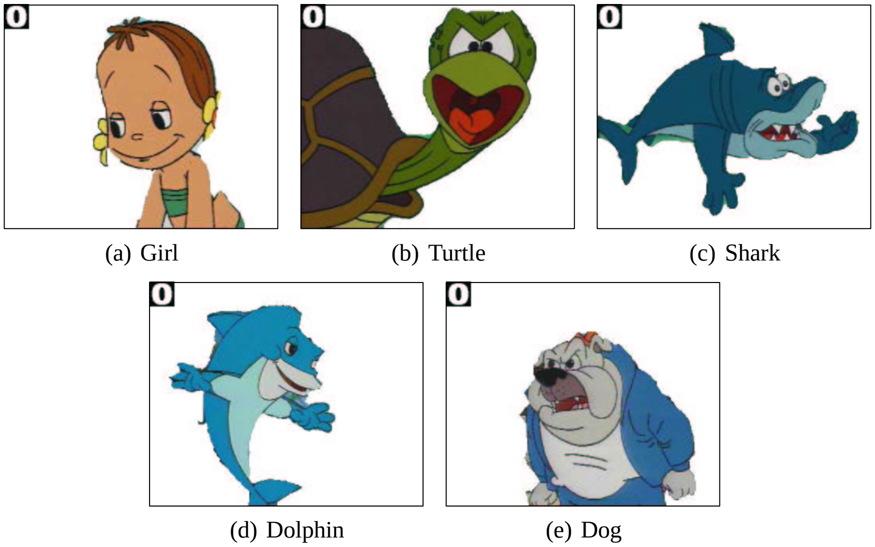
This new Multi Latent Semantic Analysis approach is evaluated on two different tasks. First the system performance is measured in the framework of object retrieval on a short set of cartoons (approximatively 10 minutes) from the MPEG-7 data set. Then, its is evaluated in the context of Video-TREC feature extraction.

4.1 Object Retrieval

The object retrieval evaluation is conducted on Docon's production donation to the MPEG-7 dataset. First the video sequence is subsampled by keeping one frame per second. Selected frames are then segmented into regions ([10]) described by a 32 bins HS histogram. To measure the performance a ground truth has been established and 5 different objects were selected and annotated in 950 frames, see figure (4.1) for an illustration. 17 to 108 queries are possible per object with a total of 245 queries. The mean precision is computed to have a global overview per object in figure (2(a)). The partition size is 2, i.e. two local LSA's. And the curves are the result of extensive experiments conducted to select the best number of factors for each model. A selection coefficient of 5.2% was kept for LSA method, yielding to a latent space of 39 features. A selection coefficient of 4% was kept for M-LSA method, yielding to two latent spaces of size 17 and 21. Thus, we have indexing signatures of reasonable and similar size in both cases. The first plot reveals the interest of M-LSA which outperforms LSA on shark, dolphin and dog objects. Figure (2(b)) shows the evolution of the mean precision over all possible queries with respect to standard recall values. M-LSA improves the stability of the IR system. However performances are under our expectation. This might be due to the video length that is too short. Indeed the dictionary computed from all regions with the k-means clustering has a size of 750. There are less shots than visual terms in partitions. Latent spaces have not enough samples to correctly remove noise and discover synonyms and we expect more improvements when enough data are available to train transformations to local latent spaces.

4.2 Video-TREC Feature Extraction

Our system is also evaluated in the context of Video-TREC. One task is to detect the semantic content of video shots. 17 features were proposed: (1) Outdoors, (2) News-subject, (3) People, (4) Building, (5) Road, (6) Vegetation, (7) Animal, (8) Female-speech, (9) Car-truck-bus, (10) Aircraft, (11) News-subject-monologue, (12) Non studio-settings, (13) Sporting-event, (14) Weather, (15) Zoom-in, (16) Physical-violence



(a) Girl (b) Turtle (c) Shark

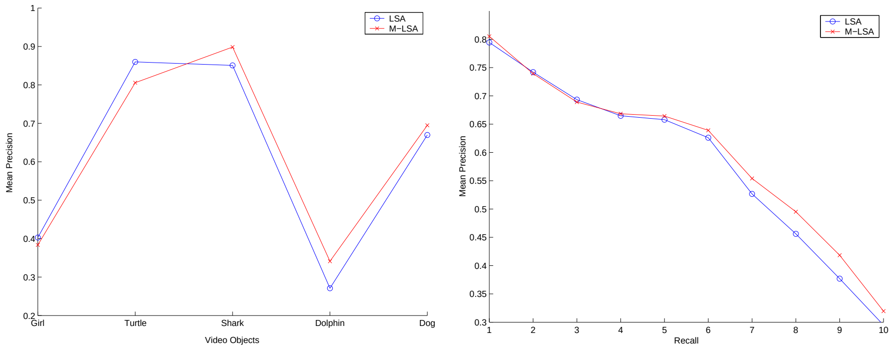
(d) Dolphin (e) Dog

Fig. 1. An illustration of the five selected and annotated objects in Docon's production cartoons.

and (17) Madeleine Albright. For each feature, 30.000 test shots are ordered with respect to their detection score value. Then the average precision at 2,000 shots is computed to characterize the performance of the system for each feature. We have proposed in [4] a simple approach using k-nearest neighbors on LSA features to estimate shot semantic features and compute their detection score. The "training" set is constituted of 44.000 shots and 17.000 were used to build the latent space. For this difficult task, two dictionaries are used: one containing color terms through 32 bins HS histograms and the other containing texture terms through 24 gabor's energies. Shots are reduced to their key-frame to save computation efforts. Indeed the segmentation process is very time consuming and untractable for all frames of the database. Visual terms are most representatives regions present in all key-frames and the signature of a shot is simply the count of visual terms where its key-frame regions are mapped. Similarity measures are independently computed for each feature type and then combined as follows:

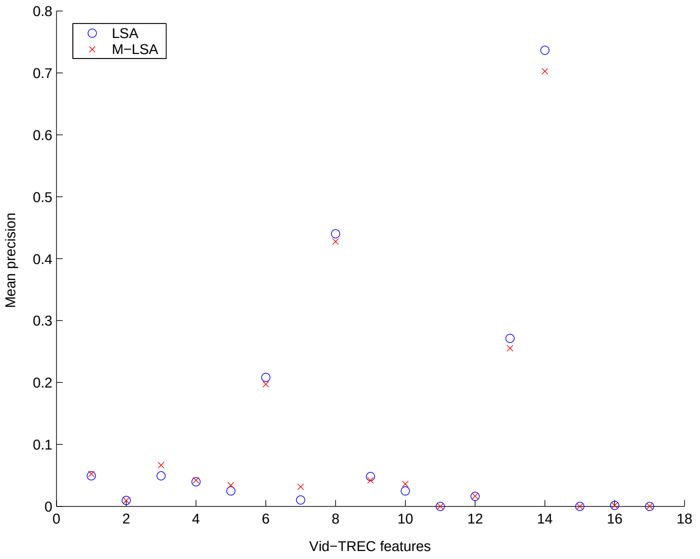
$$sim(q,q') = w_c \times sim_{color}(q,q') + w_t \times sim_{texture}(q,q') \qquad (9)$$

For simplicity $w_c = w_t = 1$ knowing that the appropriate selection of weights can be included in a training algorithm. The figure (3(a)) compares performances of the proposed M-LSA and LSA approaches. Given the volume of data to process, the tuning of parameters is very time consuming. Thus the selection coefficient l is empirically set to 10%.



(a) Object Retrieval (b) Mean Performance

Fig. 2. Object retrieval performance evaluation. The first figure shows individual performances of the system for each object, while the second curve is the mean precision curve for standard recall values.



(a) Estimation of Video-TREC features

Fig. 3. M-LSA compared to LSA for the difficult problem of semantic content analysis.

5 Conclusion and Future Work

Our previous work on Latent Semantic Analysis revealed the high potential of this simple method for object retrieval and semantic content estimation. In this paper we have presented a new approach to model video content with Latent Semantic Analysis. In particular we introduced multiple latent spaces to better represent the content. The feature space defined by video shots is decomposed into partitions where LSA's models are defined. This new representation of the content in multiple latent spaces rises the problem of indexing. We have proposed a method to index and compare video shots in this framework by taking into account shot similarities and projection errors. The method is then evaluated on object retrieval and semantic content estimation problems. A slight improvement is observed for the task of object retrieval, but results are more lukewarm when dealing with semantic content estimation.

Future work will concern the study of methods to improve the effectiveness of the similarity measure and to select factors in a more appropriate way. We are also interested in looking to probabilistic approaches to build mixture of models that is a very interesting extension to the proposed method. On the other hand, efforts will be provided to construct more sophisticated shot signatures and include more raw features such that motion, audio and text.

References

1. Chang, S.F., Chen, W., Meng, H., Sundaram, H., Zhong, D.: A fully automated content-based video search engine supporting spatiotemporal queries. In: IEEE Transactions on Circuits and Systems for Video Technology. Volume 8. (1998) 602– 615
2. Naphade, M., Kristjansson, T., Frey, B., Huang, T.: Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval. In: IEEE International Conference on Image Processing. Volume 3. (1998) 536–540
3. Souvannavong, F., Merialdo, B., Huet, B.: Video content modeling with latent semantic analysis. In: Third International Workshop on Content-Based Multimedia Indexing. (2003)
4. Souvannavong, F., Merialdo, B., Huet, B.: Latent semantic indexing for video content modeling and analysis. In: The 12th Text REtrieval Conference (TREC). (2003)
5. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science **41** (1990) 391–407
6. Kurimo, M.: Indexing audio documents by using latent semantic analysis and som. In Oja, E., Kaski, S., eds.: Kohonen Maps. Elsevier (1999) 363–374
7. Zhao, R., Grosky, W.I.: From features to semantics: Some preliminary results. In: International Conference on Multimedia and Expo. (2000)
8. Hofmann, T.: Probabilistic latent semantic indexing. In: ACM SIGIR. (1999)
9. Monay, F., Gatica-Perez, D.: On image auto-annotation with latent space models. In: ACM Multimedia. (2003) 275–278
10. Felzenszwalb, P., Huttenlocher, D.: Efficiently computing a good segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. (1998) 98–104