# Latent Semantic Analysis For An Effective Region-Based Video Shot Retrieval System

Fabrice Souvannavong, Bernard Merialdo and Benoît Huet
Institut Eurécom
Département Communications Multimédias
2229, route des crêtes
06904 Sophia-Antipolis - France
(Fabrice.Souvannavong, Bernard.Merialdo, Benoit.Huet)@eurecom.fr

## ABSTRACT

We present a complete and efficient framework for video shot indexing and retrieval. Video shots are described by their key-frame, themselves described by their regions. Region-based approaches suffer from the complexity of segmentation and comparison tasks. A compact region-based shot representation is usually obtained thanks to vector-quantization method. We thus introduce LSA to reduce the noise inherent to the segmentation and the quantization processes. Then to better capture the content of video shots, we propose two original methods. The first takes advantage of a multi-scale segmentation of frames while the second uses multiple frames to represent a shot. Both approaches require more computation time during the pre-processing but not for indexing and comparison tasks. Indeed the extra information is included in the original signatures of shots. Finally we introduce a relevance feedback loop to optimize the search and propose a new method to optimize the effect of LSA. In the experimental section, we make an evaluation of latent semantic analysis and proposed approaches on two problems, namely object retrieval and semantic content estimation.

## Categories and Subject Descriptors

H.3.1 [**Information storage and retrieval**]: Content analysis and indexing—*Indexing methods*; H.3.3 [**Information storage and retrieval**]: Information search and retrieval—*Relevance feedback*

## General Terms

Algorithms, Design

## Keywords

Video analysis, Region Similarity, Region Clustering, Latent Semantic Analysis, Region-Based Video Retrieval

*MIR'04,* October 15–16, 2004, New York, New York, USA.
Copyright 2004 ACM 1-58113-940-3/04/0010 ...$5.00.

## 1. INTRODUCTION

The growth of numerical storage facilities enables large quantities of documents to be archived in huge databases or to be 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 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, 3, 4]. On one hand this effort is further stressed by the emerging MPEG-7 standard that provides a rich and common description tool of multimedia contents. On the other hand it is encouraged by Video-TREC [1] which aims at evaluating state of the art developments in video content analysis and retrieval tools.

We propose a region-based system to efficiently index visual features of video shots. Contrasting to traditional approaches which compute global features, the region-based methods extract features of the segmented frames and perform comparisons at the granularity of the region. The main objective is to keep the local information in a way that reflects the human perception of the content [5, 6]. In order to keep both reasonable computation complexity and storage requirements, region features are usually quantized, thus allowing a compact frame representation. Unfortunately, region-based methods are sensitive to the content, the segmentation and the quantization. We thus introduce latent semantic indexing to reduce the side effects of the segmentation and quantization. Furthermore to capture the content of video shots more finely, we propose to add the information present at multiple scales in key-frames or in multiple frames of shots. Next we include a relevance feedback loop on the search process to create an optimal query.

The first step is conducted with an adaptation of Latent Semantic Analysis (LSA) to image or video content. LSA has been proven effective for text document analysis, indexing and retrieval [7]. Some extensions to audio and image features were proposed in the literature [8, 9]. The adaptation we present models video shots by a count vector in a similar way as for text documents. Key frames of shots are described by the occurrence of a set of predefined *visual terms*. *Visual terms* are based on a perceptual segmentation

[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

of images as opposed to the method proposed in [10] which uses a type aggregate map of images. The underlying idea is that each region of an image carries a semantic information that influences the semantic content of the whole shot. In [11], the authors propose a statistical model to map image regions to keywords in order to annotate the complete image. In this paper, we study the occurrence of regions in many shots to build efficient signatures of shots. Obtained signatures contain the most informative part of each shot that is used for indexing or to detect its semantic content. The first contribution of the paper is then the method used to improve video shot signatures. We enrich basic shot signatures by either including the content of key-frames at multiple segmentation levels or including the content of multiple frames composing shots. Proposed methods have the advantage to improve the representation of the content of shots without altering computation and storage needs. Following the idea of improving the shot signature, the second contribution is the introduction of relevance feedback loop used to create an optimal query and to optimize LSA with respect to the query.

The paper is organized as follows: Section 2 explains the process to construct a compact representation of video shots. Then we present the adaptation of latent semantic indexing to improve shot indexing and retrieval. Section 4 presents two methods to improve the representation of shots in the context of LSI. Following the same idea we present a relevance feedback loop to boost the performance of the system. Next, section 5 describes experiments to evaluate the different aspects of the framework in the context of object retrieval and semantic content estimation. Results of our proposed methods on two distinct tasks (information retrieval and semantic classification) are presented and discussed. Finally we conclude with a brief summary and future work.

# 2. REGION-BASED SHOT REPRESENTATION

In the proposed framework, shots are represented by their key frame that carries the most relevant information of the shot content. This allows to reduce computational efforts. The presented method is thus applicable for video shots as well as images indexing. Frames are segmented into homogeneous regions that are clustered in a small number of groups with respect to their low-level features. A frame can thus be represented as a count vector in a similar way to a text document. This compact representation not only reduces the storage requirements, it also emphasizes co-occurrences of regions. Furthermore region-based approaches attempt to overcome the drawback of global features by representing images at object-level, which is intended to be close to the human perception model [5].

## 2.1 Frame segmentation

Frames are automatically segmented thanks to the algorithm proposed by Felzenszwalb and Huttenlocher [12]. The important advantage of the method is its ability to preserve details in low-variability images while ignoring details in high-variability images. Moreover the algorithm is fast enough to deal with a large number of frames.

## 2.2 Region features

Regions are modeled by two types of features proven effective in their category [13] for content-based image retrieval:

- The color feature is described by a hue, saturation and value histogram with 4 bins for each channel,
- We use 24 Gabor's filters at 4 scales and 6 orientations to capture the texture characteristics in frequency and direction. The texture feature vector is composed of the output energy of each filter.

These visual features are then processed independently for two reasons. Firstly, combining features increases the variability of the data rendering more difficult the quantization task that follows. Secondly features can be more efficiently combined at the end with respect to the task. Different metrics can then be used or different weights can be assigned to different features by users, learning algorithms or a relevance feedback loop.

The remainder of this section and the next section deal with a single feature model to index shots. However features can be combined at this stage and the presented methods remain valid.

## 2.3 Quantization

This operation consists in gathering regions having a similar content with respect to low-level features. The objective is then to have a compact representation of the content without sacrificing much accuracy. For this purpose, the k-means algorithm is used. We call *visual terms* the representative regions obtained from the clustering and *visual dictionary* the set of *visual terms*. For each region of a frame, its closest visual term is identified and the corresponding index is stored discarding original features.

Unfortunately the quantization process can imply prejudicial approximations. This is further stressed with the k-means algorithm that is very sensitive to initial clusters. In [14], we proposed to map a region to its k-nearest *visual terms*, with k depending on the distance between the region and the closest *visual term*. Thus regions that are bordering on several groups are identified to all corresponding *visual terms*. In that case we associate to each index a probability of being a member of the group.

This one to many mapping also allows to reduce the sensitivity to the *visual dictionary* size. Indeed a major parameter of clustering algorithms is the final number of clusters. For now it is empirically selected based on previous experiments and observations.

# 3. LATENT SEMANTIC INDEXING

Latent Semantic Analysis (LSA) has been proven efficient for text document analysis and indexing. As opposed 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 [15] an adaptation of LSA to model the visual content of a video sequence for object retrieval.

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. The noise is generated by multiple sources from the visual acquisition system to the segmentation and clustering processes. Latent Semantic Analysis is a solution to remove some of 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 common approach is to use the singular value decomposition (SVD) of the occurrence matrix of features in shots to achieve this task.

Shots are represented by the count vector of *visual terms* that describes the content of their regions. Let now denote $q$ this feature vector. The singular value decomposition of the occurrence matrix C of visual terms in video 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 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 allow to find the appropriate factor number.

## 3.1 Video shot retrieval

Finally shots are directly compared in the singular space. Let $f_q = (f_{i,k_i})_{1 \le i \le n}$ be the representation of a shot with different features such as color and texture. $f_{i,k_i}$ is the feature vector of i projected on the singular space of i whose size is $k_i$. We compute the weighted sum of cosine values over each feature. Thus the similarity value between q1 and q2 is,

$$sim(q, q') = \sum_i w_i \cos(f_{i,k_i}, f'_{i,k_i}) \tag{4}$$

This formulation is interesting since it does not only allow to dynamically select the weights between features but also to select the projection size.

# 4. OPTIMIZING SHOT REPRESENTATION

Our method relies on the LSA to discover *visual terms* equivalences with respect to their occurrence in video shots. However region-based approaches suffer from their dependence to the segmentation and the clustering. On one hand similar objects and backgrounds are not segmented in the same way from one frame to another. On the other hand the clustering can amplify segmentation variations. We now focus our interest on improving region-based signatures to reduce side-effects of the segmentation and the clustering. We propose two approaches. The first is a multi-scale view of frames and the other uses multiple key-frames per shot. We then introduce a relevance feedback system that fits in our framework.

## 4.1 Multi-scale approach

A perfect and unique segmentation does not exist. For a given segmentation algorithm, parameters have to be tuned to find the closest answer to user's expectations which differ from one frame to another. Indeed, an object or background can appear on a frame at different scales or under different illumination conditions, altering the outputs of segmentation algorithms. To deal with such drawbacks, the common solution is to segment images at different levels and use outputs as independent representations of images.

Several methods have been presented in the literature to combine multi-resolution images [16, 17]. However we want to avoid multiplying the number of vectors to describe a frame and propose to represent the content of a frame at different segmentation levels in a unique vector: the occurrence of *visual terms* in a frame at different level. This approach is consistent with the cosine measure that favors common regions and is indifferent to misses. Given a query, frames having same regions at all scales are favored over frames having same regions at some scales which in turn are favored over frames having just few common regions.

In addition to allow multi-scale content matching, this approach is an efficient way to deal with segmentation variations without adding computation and storage requirements after features are extracted.

## 4.2 Multiple key frames

Following the idea of adding robustness to the representation of shots, we propose to use more than one frame to represent their content. Indeed due to motion and temporal segmentation algorithm, the key frame might not represent well the shot nor be the only representative of the shot. Furthermore, region-based representation of a frame is very sensitive to its segmentation as we have seen in the previous subsection.

Using multiple frames is a solution to capture most of the shot content and adding robustness to segmentation variations. Solutions have been proposed for spatio-temporal segmentation into regions [18, 19]. However we want to avoid the extra computation cost of spatio-temporal segmentation and multiplying the number of vectors to describe a shot, thus as for the multi-scale approach, a shot is described by the occurrence of *visual terms* in different segmented frames composing the shot. For simplicity, we have selected the left and right frames surrounding the key-frame at a distance of half a second.

## 4.3 Relevance feedback

The Rocchio relevance feedback algorithm [20] is one of the most popular and widely applied learning method in information retrieval. When documents are ranked with respect to their similarity to the query, the ideal query should favor all relevant documents while discarding non-relevant documents. Rocchio proposed to maximize the mean similarity to positive samples ($d \in P$) minus the mean similarity to negative samples ($d \in N$). This results in the optimal query Q' defined as:

$$Q' = aQ + b\sum_P d - c\sum_N d \tag{5}$$

$a, b$ and $c$ are Rocchio's weights that can have the following values:

$$a = 1, b = \frac{1}{Card(P)}, c = \frac{1}{Card(N)}$$
$$a = b = c = 1$$
$$c = 0$$

This formulation reminds the proposed approaches consisting in including multiple scales and frames to model the shot content. It naturally fits in our representation of a shot. Thus in an information retrieval framework, the user can update the query with respect to correct and incorrect retrieved frames. The resulting query is then a weighted sum of first retrieved frames like in equation (5). We also propose to update the query with only the most relevant information. For this purpose, the user selects positive regions in frames and the query is updated only with these regions.

Finally, we suggest to dynamically select the projection size in the singular space of LSA. First the optimal query is computed, then the similarity between the new query and positive samples is incrementally computed. The incremental algorithm is based on the relation between the cosine functions when the projection sizes are k and (k+1):

$$\cos^k(u,v) = \frac{\sum_{i=1}^{i=k} u_i v_i}{\sqrt{\sum_{i=1}^{i=k} u_i^2 \sum_{i=1}^{i=k} v_i^2}}$$

$$\cos^{k+1}(u,v) = \frac{\sum_{i=1}^{i=k} u_i v_i + u_{k+1} v_{k+1}}{\sqrt{(\sum_{i=1}^{i=k} u_i^2 + u_{k+1}^2)(\sum_{i=1}^{i=k} v_i^2 + v_{k+1}^2)}}$$

Independently updating the numerator and denominators at each iteration allows to efficiently compute the cosine values between two vectors at successive projection sizes. At each iteration, we evaluate the quality of the current projection size to find the most appropriate. We propose to maximize either the mean precision value that evaluates the retrieval order of frames or Rocchio's function that is the mean similarity to positive samples minus the mean similarity to negative samples. The second solution suffers from the fact that the cosine value usually decreases when the projection size increases. Thus to be able to evaluate the performance between different factors, Rocchio's function is normalized using a min-max method.

# 5. EXPERIMENTS

Our proposed approaches to model a video shot thanks to latent semantic indexing are 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 on full frames. Indeed, it would be interesting to perform the evaluation of both tasks on the same dataset, and more particularly the Video-TREC one. However, the ground truth available for Video-TREC does not feature object level annotations. Therefore, and in order to minimize the annotation effort we opted for cartoon videos.

## 5.1 Object retrieval

The object retrieval evaluation is conducted on Docon's production donation to the MPEG-7 dataset. First the video sequence is sub-sampled by keeping one frame per second. Selected frames are then segmented into regions [12] described by a 32 bins HS histogram (the value is omitted in these experiments since it is not really relevant for cartoons). A ground truth has been manually established to measure the performance and 5 different objects were selected and annotated in 950 frames, see figure (5.1) for an illustration. 17 to 108 queries are possible per object with a total of 330 queries. The average mean precision is computed to have an overview of the performances at different projection factors that determines projection sizes, i.e. the number of factors kept by the LSA.

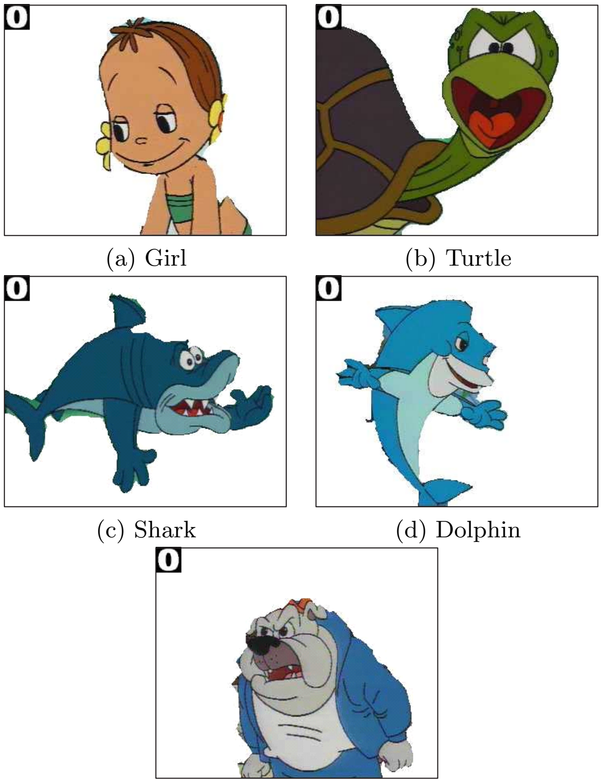
To determine the size of the *visual dictionary*, different numbers of clusters were tested. In parallel, different projection sizes are used to study the capacity of LSA. Figure (2(a)) shows the performance of the system in its simplest aspect, i.e. one key frame per shot at a given segmentation level. Drawn curves were selected for the optimal number of clusters at each segmentation level. We observe that in the best cases finer segmentation leads to higher performance but also to an higher *visual dictionary* size. Further experiments reveal that decreasing again the granularity of the segmentation was not providing significant improvements while requiring more clusters. Finally the average mean precision is boosted with LSA whatever the segmentation level. An improvement of about 50% is realized when using 10% of features.

Figures (2(b),2(c)) show the performance gain that can be obtained by improving the shot representation. The multi-scale approach using all available scales (three levels) achieves the best average mean precision value with a major gain of about 8%. The multiple key-frames method does not significantly improves the retrieval performance, but more stability is gained with respect to the projection factor. In both situations, the impact of LSA remains important.

Finally, the relevance feedback loop boosts performances, see figure (5.1). And we reach an average mean precision of more than 0.75 in the best case. The effect of LSA is strangely baneful: performances decrease when the projection size decrease. The explanation is that the construction of the optimal query realizes a part of LSA's task. However performances are still good when the projection size is reduced by a factor of 15%. In that case, LSA has also the important advantage of reducing the dimension of the feature vector, and thus decreasing the speed of retrieval operations.

By using only positive samples we obtain better results. This suggests to classify retrieved shots in three categories: negative, positive and indifferent samples in future works. The method can also be extended at the object level. In that case the user selects the relevant part of the frame and only the information of selected regions is injected in the loop. This method does not perform as well as previous ones. It reveals the importance of the context in the retrieval process for the video involved.

Several experiments were conducted to dynamically identify the optimal projection size with relevance feedback. The idea is to optimize the mean precision value (the importance is put on the order of retrieved elements) or Rocchio's function (the importance is put on similarity values) over the first retrieved frames judged by the user. Unfortunately as we have seen, the influence of the projection size is reduced when a relevance feedback loop is involved and no real improvement was observed. Moreover the mean precision criteria suffers from the lack of data to be efficient and thus



(a) Girl (b) Turtle

(c) Shark (d) Dolphin

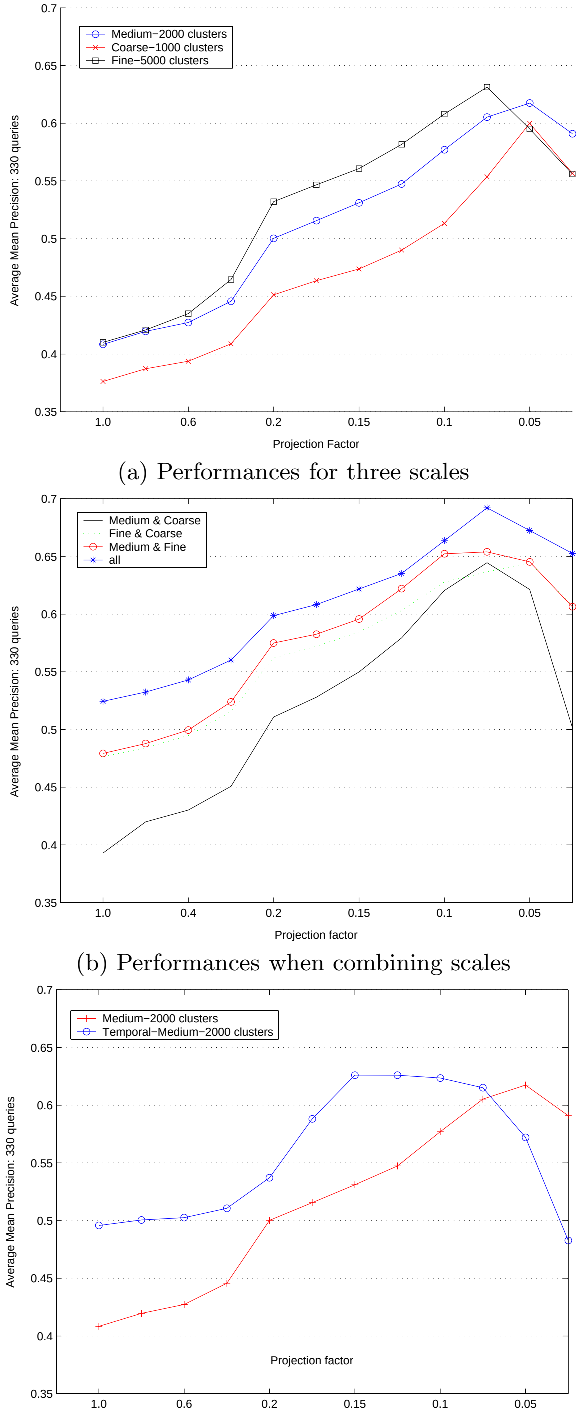
(e) Dog

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

performs lower. However using Rocchio's function criteria, we observe that the stability with respect to the initial factor size is very good, see figure (5.1). A first search can thus be quickly conducted with a small number of factors and then another search can be ran with the appropriate projection factor on the optimal query.

## 5.2 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. The evaluation requires annotated data. In June 2003, Video-TREC has launched a collaborative effort to annotate video sequences in order to build a labeled reference database. The database is composed of about 63 hours of news videos that are segmented into shots. These shots were annotated with items in a list of 133 labels which root concepts are the event taking place, the context of the scene and objects involved. The tool described in [21] was used for this time-consuming task. We have selected 10 features among those items to evaluate the performances of proposed approaches: *Sport Event, Cartoon, Weather News, Studio Settings, Nature Vegetation, Cityscape, Animal, Face, People* and *Transportation*. Simple and complex semantic features were retained to evaluate our system. Then for computation and time requirements we decided to use 12,000 shots for the training set and 3,000 for the test set. For each feature, test shots are ordered with respect to their detection score value. Next the average precision at 2,000 shots is computed to characterize the performance of the system for each feature.

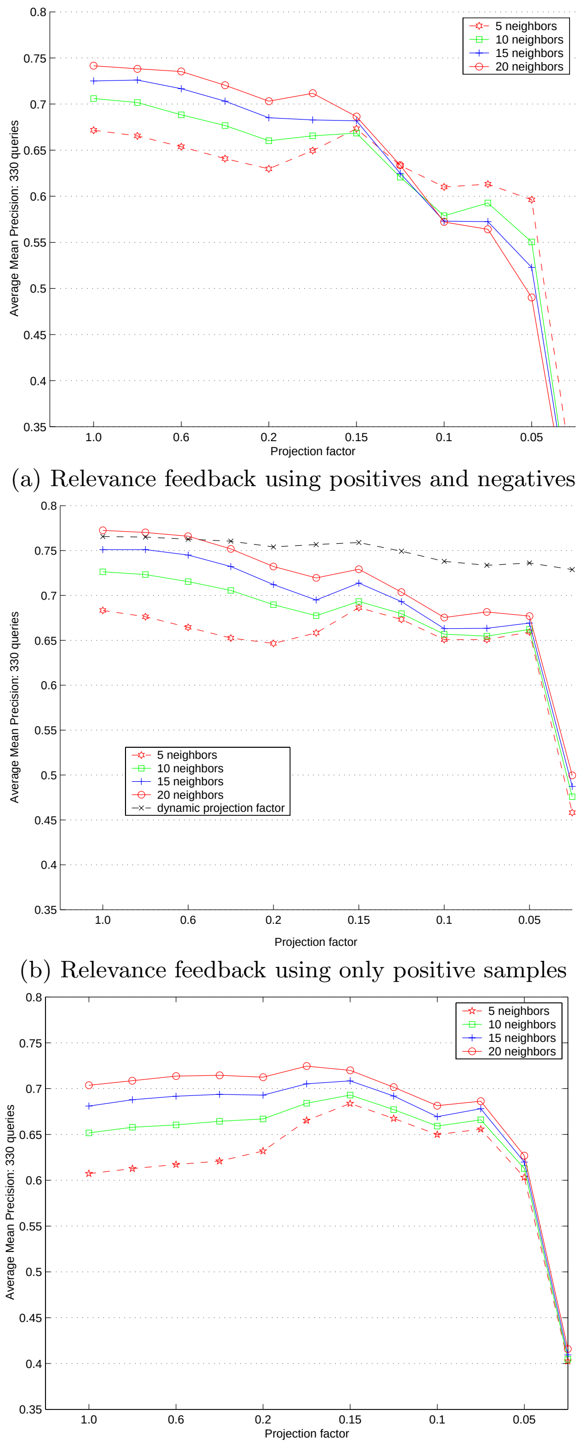


(a) Performances for three scales

(b) Performances when combining scales

(c) Performances when using multiple frames for one video shot

**Figure 2: These three figures compare the performances of latent semantic indexing with respect to input vectors on the task of object retrieval.** *Vectors are build either from key-frames either at a given scale or many scales, or from two frames in a same shot. We observe a gain of 15% from the best single scale approach to all scale approach. Using multiple frames reduces the sensibility to the projection factor. These figures highlight the impact of LSA that allows to boost performances.*



(a) Relevance feedback using positives and negatives

(b) Relevance feedback using only positive samples

(c) Relevance feedback on selected objects

**Figure 3: Relevance feedback impact on object queries.**
*On one hand, we observe that using only positive samples gives the best performances. On the other hand, contrary to direct retrieval, relevance feedback gives the highest results with highest projection size, i.e. where the impact of LSA is reduced. The third figure shows that the context of objects has its importance since relevance feedback on selected objects in returned frames has not a positive impact.*

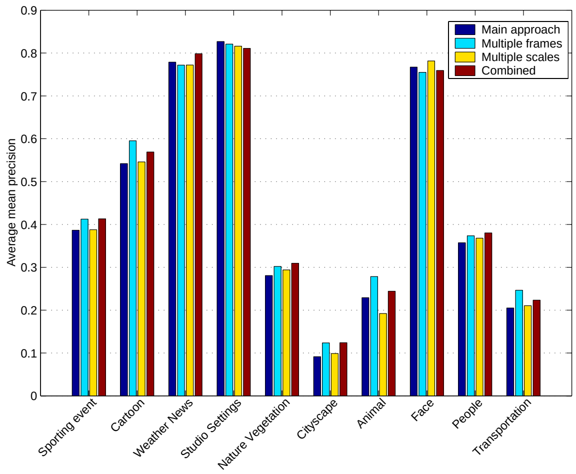


**Figure 4: Semantic classification of video shots.**
*Proposed approaches allow to increase classification performances. The importance of the gain is different from one feature to another. Multi-frame and combining the multi-scale and multi-frame approaches give the best performances in average; the first requiring less pre-processing.*

We have proposed in [22, 14] several approaches to estimate shot semantic features and compute their detection score. The k-nearest neighbors classifier on LSA features gave better performances for the semantic classification task than Gaussian mixture models and neural nets. We thus use the k-nearest neighbors classifier to estimate the semantic content of shots.

For this difficult task, two dictionaries are used: one containing color terms through 64 bins HSV histograms and the other containing texture terms through 24 Gabor's energies. Similarity measures are independently computed for each feature type and then combined as follows:

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

For simplicity $w_c = w_t = 1$ knowing that the appropriate selection of weights can be included in a training algorithm [23].

Let $N_s$ be the neighborhood of a shot s in the training set L, i.e. the k-nearest neighbors of s in the training set, and $y_i \in \{0,1\}^l$ the semantic value of the neighbor i. The detection score is a vector defined as:

$$d_L(s) = \sum_{N_s} sim(s,n_i) * y_{n_i} \qquad (6)$$

We experimented several forms of the estimator: we normalized by $\sum_{N_s} sim(s,n_i)$ or used $y_i \in \{-1,1\}^l$ and equation (6) gave the best performances. Indeed we are computing a detection score to order shots. The lack of normalization favors shots that have really close neighbors, and thus shots for which the estimation is the most reliable.

Figure (5.2) shows the performances of the system for the different approaches. By experiment twenty neighbors reveals to be a good neighborhood size for the estimator. Two segmentation levels are used in the multi-scale approach and the best projection size is selected for each semantic feature and each approach. Performance gains are not as high as for the retrieval task. Their importance is different from one feature to another. Multi-frame and combining the multi-scale and multi-frame approaches give the best performances in average.

## 6. CONCLUSION AND FUTURE WORK

We presented a complete system for both video shot and video objects retrieval which includes a relevance feedback loop. We extented further the use the introduced indexing technique to estimate the semantic content of video shots based on the semantic contained in similar shots. The key of our method is the introduction of latent semantic indexing to the region-based representation of the video content. Moreover, we proposed two solutions to deal with the noise generated by both the segmentation and the clustering while capturing the visual content of shots with an improved accuracy. For this purpose we used multi-scale segmentation of key frames and we included multiple frames in the representation of video shots. As opposed to common approaches that would increase the number of parameters to describe the content, we fitted the extra information within the existent signature. Following a similar approach, we proposed a relevance feedback loop to boost the retrieval performance of the system. Additionally, we included an optimization of LSA in the loop by selecting the appropriate projection size.

The proposed methods perform well for the task of information retrieval. A significant performance gain is observed when using the multi-scale approach while we obtain more stability with respect to the projection size with the multiple key-frames approach. As expected the relevance feedback loop boosts performances. Moreover, the dynamic selection of the projection size in the relevance feedback loop allows to make a rapid initial query with a small projection size without altering the final performance of the retrieval. On the task of semantic content estimation, improvements are lower but present still.

Future work will concern the study of the relevance feedback loop as well as estimators of the semantic content to include automatic feature weighting. We will investigate ways to include motion and region relationship in our system. We also envisage to include more features like salient visual points and text information. Finally more effort will be focused on the complex task of semantic content detection.

## 7. ACKNOWLEDGMENTS

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

## 8. REFERENCES

[1] Shih-Fu Chang, W. Chen, H.J. Meng, H. Sundaram, and Di Zhong. A fully automated content-based video search engine supporting spatiotemporal queries. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 8, pages 602– 615, 1998.

[2] M.R. Naphade, T. Kristjansson, B. Frey, and T.S. Huang. Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval. In *IEEE International Conference on Image Processing*, volume 3, pages 536–540, 1998.

[3] Howard Wactlar, Takeo Kanade, Michael A. Smith, and Scott M. Stevens. Intelligent access to digital video: The informedia project. *IEEE Computer*, 29(5), 1996.

[4] E. Ardizzone and M. La Cascia. Automatic video database indexing and retrieval. *Multimedia Tools Applications*, 4(1):29–56, 1997.

[5] Chad Carson, Megan Thomas, and Serge Belongie. Blobworld: A system for region-based image indexing and retrieval. In *Third internation conference on visual information systems*, 1999.

[6] Feng Jing, Mingling Li, Hong-Jiang Zhang, and Bo Zhang. An effective region-based image retrieval framework. In *ACM Multimedia*, 2002.

[7] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[8] Mikko Kurimo. Indexing audio documents by using latent semantic analysis and som. In Erkki Oja and Samuel Kaski, editors, *Kohonen Maps*, pages 363–374. Elsevier, 1999.

[9] Rong Zhao and William I Grosky. From features to semantics: Some preliminary results. In *International Conference on Multimedia and Expo*, 2000.

[10] Joo-Hwee Lim. Learning visual keywords for content-based retrieval. In *IEEE International Conference on Multimedia Computing and Systems*, volume 2, pages 169–173, 1999.

[11] Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *IEEE International Conference on Computer Vision*, pages 97–112, 2002.

[12] P. Felzenszwalb and D. Huttenlocher. Efficiently computing a good segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–104, 1998.

[13] Wei-Ying Ma and Hong Jiang Zhang. Benchmarking of image features for content-based image retrieval. In *Thirty-second Asilomar Conference on Signals, System and Computers*, volume 1, pages 253–257, 1998.

[14] Fabrice Souvannavong, Bernard Merialdo, and Benoît Huet. Latent semantic analysis for semantic content detection of video shots. In *International Conference on Multimedia and Expo*, 2004.

[15] Fabrice Souvannavong, Bernard Merialdo, and Benoît Huet. Video content modeling with latent semantic analysis. In *Third International Workshop on Content-Based Multimedia Indexing*, 2003.

[16] M. Mirmehdi and R. Perissamy. Perceptual image indexing and retrieval. *Journal of Visual Communication and Image Representation*, 13(4):460–475, December 2002.

[17] Charles E. Jacob, Adam Finkelstein, and David H. Salesin. Fast multiresolution image querying. In *International conference on computer graphics and iteractive techniques*, pages 277–286, 1995.

[18] Fabrice Moscheni, Sushil Bhattacharjee, and MuratKunt. Spatio-temporal segmentation based on region merging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:897–915, 1998.

[19] Daniel DeMenthon. Spatio-temporal segmentation of video by hierarchical mean shift analysis. In *Workshop on Statistical Methods in Video Processing*, 2002.

[20] J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, 1971.

[21] Ching-Yung Lin, Belle L. Tseng, and John R. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *Proceedings of the TRECVID 2003 Workshop*, 2003.

[22] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. Latent semantic indexing for video content modeling and analysis. In *The 12th Text REtrieval Conference (TREC)*, 2003.

[23] K. Kira and L. Rendell. A practical approach to feature selection. In *Proceedings of the 9 International Conference on Machine Learning*, pages 249–256, 1992.