VIDEO CONTENT MODELING
WITH 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

In this paper we present a novel approach to fully automatic video content modeling. We introduce the concept of visual dictionary to describe visual video elements, called words, which appear through video sequences. Their co-occurrences in contexts, i.e. the main video entity to be indexed (frame, shot, scene, . . .), compose signatures usable for indexing and comparison. Latent Semantic Analysis (LSA) is naturally introduced to improve the robustness to noise and discover the latent semantic. This new representation along with its associated similarity measure, has many applications including indexing, retrieval, summarization or enhanced navigation, on single as well as multiple video sequences. Once the framework is presented, we investigate three methods to efficiently exploit the information provided by multiple features in order to improve the video analysis. Promising results were obtained on the object and frame retrieval tasks across a single video document.

1. INTRODUCTION

Multimedia documents are becoming very popular and are spreading over the entire world in many databases and the web. Unfortunately, this increasing amount of available information emphasizes the lack of organization of such contents and renders more difficult the usual tasks performed over text documents. Montaigne's remark "Mieux vaut une tête bien faite que bien pleine" [1] is up to date and many researchers are currently investigating methods to automatically analyze, organize, index and retrieve video information [1, 2]. This effort is further underlined by the emerging Mpeg-7 standard that provides a rich description tool of multimedia contents.

Video analysis research is divided in several fields. Much prior work has been conducted in temporal video segmentation [3]. In most cases shot segmentation tools are quite reliable whereas scene segmentation [4] algorithms still have to be proven effective. Another popular field is the automatic creation of video summaries that have raised the interest of many researchers [5, 6] while solutions to semantic analysis are only just emerging [7, 8].

In this article, we propose an original and flexible approach to automatic video content modeling while studying the ways to use multiple features (color, texture, . . .). The main idea is to decompose video sequences into contexts, like frames, shots, scenes or semantic concepts. Then, a context is described by words belonging to one or more dictionaries and the occurrence of words composes the signature of context. The relationships between words and contexts provide a very rich information captured and enhanced by Latent Semantic Analysis (LSA) in a reduced space, where a measure is derived to compare simultaneously both entities. This measure is then exploited for advanced video content analysis at the frame and object level. In particular we investigate the potential of using multiple dictionaries through three distinct methods, to improve the overall performance.

Latent Semantic Analysis has been proven effective for text document analysis, indexing and retrieval [9] and some extensions to audio and image features were proposed [10, 11]. Here, we propose to extend its application to video content modeling in order to reduce noise and enhance co-occurrence information.

The rest of the paper is organized as follows: The next section presents the framework in three parts, one related to the decomposition of video sequences and the definition of visual dictionary and context; the second to the analysis via LSA and the last to the exploitation of multiple dictionaries. Then, we present preliminary results to validate the framework through an initial application. Experiments

This research was supported by the EU project Spation

[1] Choose a guide with a well-made rather than a well-filled head

have been conducted on the frame and object retrieval tasks within single video documents. Applications are mainly enhanced navigation and automatic summary creation. Finally, we conclude by summarizing our findings and providing research directions.

2. FRAMEWORK PRESENTATION

The major problem tackled in image or video analysis tools is feature extraction since visual contents are extremely rich and various. In many cases, due to shadows, highlights, camera or object motions, deformations, ... in images, visual features described in a high dimensional space, tend to be extremely noisy. Despite the presence of noise, the repetitions contained in this huge amount of information can be used to extract important visual properties. We propose a statistical method that takes advantage of the information repetition, through co-occurrences, to partially eliminate the noise in a robust video content model.

Video sequences are decomposed into two kinds of categories. On one hand stand elementary units (pixels, regions or frames, ...) considered as words. They are mapped into one or more visual dictionaries that capture local similarities in video sequences. On the other hand stand word agglomerations assimilated to contexts such as frames, shots, scenes or semantic structures which are the main entities to index and compare. Occurrences of words in contexts define a set of raw context signatures forming the co-occurrence matrix word-context. The important relationships between words and contexts provide very rich information that can be used as it is (comparison of raw signatures) or further enhanced by LSA (comparison of transformed signatures).

2.1. Visual Dictionary and Word Association

In our model, video sequences are described by small entities, i.e. words, that compose the contexts on which operations are accomplished. Thus an initial stage consists in deciding what kind of words have to be extracted with respect to the desired type of context. Diverse combinations of word and context types are envisageable. One example that is used later, is the couple (frame-region, frame) that permits to analyze the frame content.

The key point of our approach is the modeling of video documents in words belonging to one or multiple visual dictionaries, i.e. sets of predefined words, to describe contexts. In fact, words are described by some noisy high dimensional features extracted from the video content. The dictionary is then introduced to identify similar words in video sequences. While matching two textual words is rather straightforward, it is more difficult to effectively compare visual features. Moreover, dictionaries naturally exist for text but it is not the case for multimedia contents and they have to be build. The creation of visual dictionaries is a challenging task often related to data-mining problems. Nevertheless it is not in the scope of this paper to discuss these techniques, the reader can refer to [12] for a comprehensive survey. We should just keep in mind that these partitioning operations are often sensitive to noise or outliers and partitions are suboptimal in most cases. Additionally, the choice of the dictionary size is far from obvious. One possible approach to build a dictionary is to describe words by some features and to cluster elements with the k-means algorithm. Finally, the resulting centroids define the dictionary.

We can summarize the video structure for one dictionary as follows. Let $\mathcal{F} = \Re^N$ be the feature space where elementary entities, i.e. words, of video sequences are modeled. A dictionary of size N, denoted D, is defined by a set of words $D = \{r_i^D \in \mathcal{F}, i \in [0..N]\}$ to which is associated a distance d between words. A word w matches a word of the dictionary r_i^D if and only if $i = arg \min_j \{d(w, r_j^D)\}$. Finally a context is described by its raw signature defined as a vector containing the occurrence of each word of the dictionary. It is clear that the dictionary must contain good representatives of the encountered words to efficiently represent the data cloud of raw features.

2.2. Latent Semantic Analysis

Once the video sequences are decomposed in context and words, we take advantage of the LSA properties to induce relationships between words and contexts depending on the co-occurrences of words in contexts. These inductions improve noise robustness from the dictionary while highlighting synonyms.

Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. The adequacy of LSA's reflection of human knowledge has been established in a variety of ways [13]. For example, its scores overlap those of humans on standard vocabulary and subject matter tests; it mimics human word sorting and category judgments; it simulates word-word and passage-word lexical priming data; and it accurately estimates passage coherence, learnability of passages by individual students, and the quality and quantity of knowledge contained in an essay.

The previous part has introduced the notion of words and contexts for video content, so that the LSA theory can be applied to video documents. The following gives an overview of the method. We construct the co-occurrence matrix of words in contexts (raw signatures). The Singular Value Decomposition (SVD) gives the transformation parameters to a new space were both words and contexts are mapped and comparable. The dimension of the transformed space is then reduced to enhanced words and contexts relationships. The number of factors k to keep is crucial and difficult to choose since we do not really want to reduce the dimension for compression but to create induction rules and improve the comparison task. This simplification provides a least squared approximation of the original matrix, therefore it can be seen as a filter that removes the noisy part of the co-occurrence matrix. A threshold has to be defined to effectively remove noise while keeping the integrity of word equivalences. Mathematical operations are finally conducted in the following manner:

- First the co-occurrence matrix is constructed:
  Let A of size M by N be the co-occurrence matrix of M words (defining a dictionary) into N contexts (representing the video sequence). Its value at cell (i, j) corresponds to the number of times the word i appears in the context j.

- Next, it is analyzed through LSA:
  The SVD decomposition gives $A = USV^t$ where

  $$UU^t = VV^t = I,\ L = \min(M, N)$$

  $$S \approx diag(\sigma_1, .., \sigma_L),\ \sigma_1 \geq \sigma_2 \geq ... \geq \sigma_L$$

  Then A is approximated by truncating U and V matrices to keep k factors in S corresponding to the highest singular values.

  $$\hat{A} = U_k S_k V_k^t \text{ with } S_k = diag(\sigma_1, .., \sigma_k)$$

- Finally, indexing of a context of A noted $c(j)$ and a new context q is realized as follows:

  $$p_{c(j)} = \text{row j of } VS$$

  $$p_q = q^t U_k$$

- And to retrieve the context q in a database containing indexed contexts p_j, the cosine measure m_c is used to compare elements.

  $$m_c(p_j, q) = \frac{p_q . p_j}{||p_q||.||p_j||}$$

  The most similar elements to the query are those with the highest value of m_c.

2.3. Multiple Dictionaries

Combining various features like color, texture, shape, . . . is a key step to improve video content analysis performance. Our framework handles this consideration in three ways presented below.

2.3.1. Basic method

A direct approach consists in combining features before the creation of a unique dictionary. However as we mentioned earlier, dictionaries are not obvious to create and this task is even more complex when the number of features is large.

To solve this problem, instead of merging different features into a single vector, we propose to construct independent dictionaries, for example based on color and then on texture, and conjointly use them to improve the overall performance of our framework. The next two methods merge the information of both dictionaries. In merged dictionaries method (MDM), LSA accomplishes this task when applied to merged co-occurrence matrices. In independent dictionaries method (IDM), each co-occurrence matrix is enhanced independently and the comparison measure is modified to take advantage of available features.

Let $\{D_i\}_{1 \leq i \leq n}$ be the set of available dictionaries.

2.3.2. Merged dictionaries

Contexts are now described by the concatenation of raw signatures. They are represented by a vector containing the occurrence of words of all available dictionaries. Occurrence values can be differently weighted in order to give more importance to some features. Finally, the LSA is applied to the co-occurrence matrix and comparison operations are conducted as previously explained. This method is denoted merged dictionaries.

2.3.3. Independent dictionaries

Dictionaries are used independently as presented in section 2.2. Thus for a query q and an indexed context p, we obtain n similarity measure values $m_c^i(p, q)$. We propose to compute a unique similarity value as a weighted sum:

$$m_c(p, q) = \sum_{i=1}^{i=n} \alpha_i m_c^i(p, q)$$

For instance $\alpha_i = 1$ and the optimal weighting will be the scope of a future work. This method is denoted independent dictionaries.

MDM takes full advantage of co-occurrence information whereas it is more rigid. Indeed the LSA has to be reconducted each time weights are modified or new features are added. This is not the case of IDM that has the main advantage of being adaptive.

MDM underlines the possibility to weight the co-occurrence matrix before applying LSA. This is a common and effective approach for improving full-text retrieval performance that consists on weighting the matrix values with global and local weights [14]. Usually global weights indicates the overall importance of a word while local weights indicates its importance in a specific context. Nonetheless this possibility will be the scope of another study and for now we restrict our evaluation to unit weights in the same way that $\alpha_i = 1$.

3. VIDEO NAVIGATION

We are interested in advanced navigation at the frame as well as sub-frame level. The user can interactively navigate through a video sequence by querying for similar frames or frames containing a selected group of regions. Therefore words are equivalent to frame regions and contexts are equivalent to frames as detailed in the next part. It allows creating a query with a set of words representing either an object, background or frame. The work presented here was lead on 1000 frames of Docon's Production donation to the mpeg-7 data collection, where 7 characters were manually selected to represent the set of four animated cartoons and evaluate the performances.

3.1. Dictionary Design

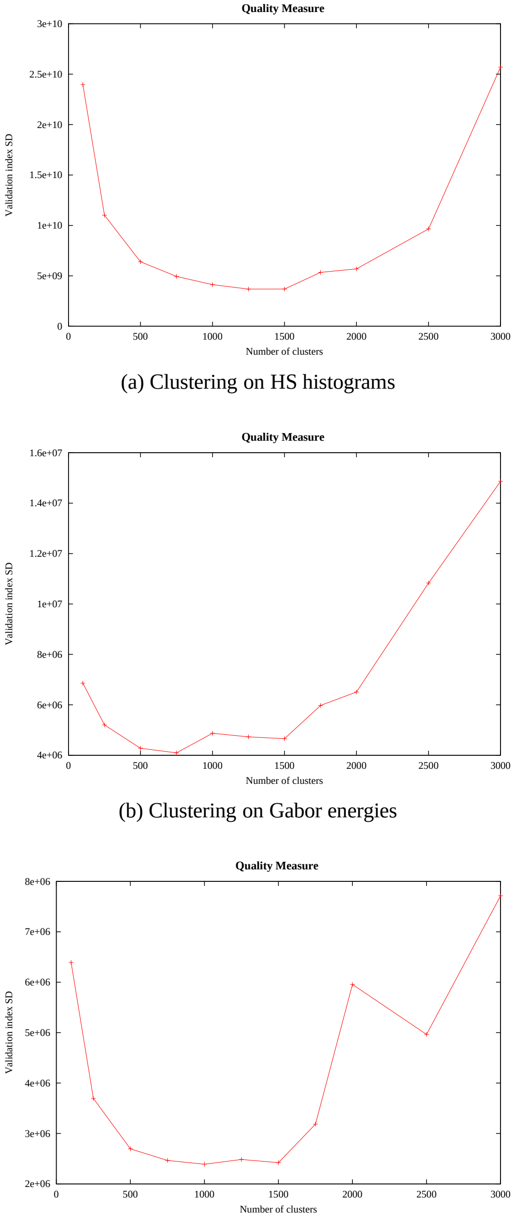
The generality of the framework offers many possibilities that may behave differently. We restricted our study to the specific case of frame and sub-frame queries inside a single video sequence for the purpose of enhanced navigation. One approach is to use histograms and directly compare them (in that case words are pixels and dictionary words are bins). However histograms are not adapted when the dimension of the feature space is high; furthermore spatial information is lost in the process. To overcome these drawbacks, we introduce a codebook, i.e. dictionary created with data mining techniques, computed on frame regions. Regions permits to capture the local information and the codebook permits to include many features of high dimension. A natural idea is to adapt regions to the frame content by using a segmentation algorithm. The choice of the algorithm is not fundamental as far as it provides homogeneous regions and favors over-segmentation (to avoid the exhausting of co-occurrence information).

3.2. Dictionaries

Video frames are automatically segmented thanks to the algorithm proposed by Felzenszwalb and Huttenlocher [15]. Then regions are modeled by two types of features proven effective in their category [16] for content-based image retrieval:

1. Color. It is described by a Hue (H) and Saturation (S) histogram with 8 bins for H and 4 for S,

2. Texture. We use 24 Gabor's filters at 4 scales and 6 orientations to capture the texture characteristics in frequency and direction. The feature vector is composed of the output energy of each filter.



(a) Clustering on HS histograms

(b) Clustering on Gabor energies

(c) Clustering on color and texture features

Fig. 1. Measure of clustering validity for three feature vectors, to choose the optimal number of clusters with respect to SD index in the range [100..3000]. The optimal number of clusters is the global minimum.

These features are used either independently or conjointly to form three visual dictionaries. Each of them results from the k-means algorithm that provides a set of k-centroids that define dictionaries. Frame regions are then associated to one element of each dictionary in order to build the appropriate co-occurrence matrix region-frame.

Since the clustering process is unsupervised, the final clusters require some kind of validation. This procedure has tackled difficult problems that can be qualitatively expressed as i. quality of clusters, ii. degree to which clustering scheme fits to a specific dataset, iii. optimal number of clusters in a partitioning. Many methods have been proposed in the literature [17] and we have opted for a recent and fast quality index (SD) presented in [18] to choose the optimal number of cluster for crisp clustering. The SD index measures the average scattering of clusters and the total separation between clusters. It has the property to result in a local minimum when computed in the range $[c_{min}, c_{max}]$ for the number of clusters that we considered. This minimum corresponds to an optimal number of clusters with respect to the SD index, to partition data. Experiments conducted by Halkidi in [18] shown that this optimal number is only slightly influenced by c_{max}. The figure 1 represents the evolution of the index with respect to the number of clusters when color features are used 1(a), then texture features 1(b) and finally color and texture features 1(c). Minima present respectively at 1250, 750 and 1000 are the optimal numbers of clusters that we retain with respect to SD index and the number of clusters range [100..3000].

3.3. Object and Frame Retrieval

We place our work in the particular context of object and frame retrieval across a video sequence for the purpose of enhanced video navigation. The user while looking at the video is able to select a part of the frame and ask to find similar objects in the video sequence. This situation is appropriate to evaluate our framework since it entirely relies on it. To evaluate the retrieval, we use the available object annotations: a retrieved frame is declared relevant if it contains the query object and in the case of frame retrieval at least one common object with the query. Then, we draw the average precision versus recall curves to illustrate the performances. These curves are obtained by computing the precision for all possible queries over an object at standard recall values, i.e. 0.1, 0.2, . . . , 1.0, and then taking the mean values. Average over all objects is then computed to give the overall performance.

Three types of experiments were conducted. The first set consists in evaluating the interest of LSA on individual dictionaries: color, texture and both. Then we evaluate the improvement obtained by combining color and texture dictionaries; on one hand by merging co-occurrences matrices with MDM, on the other hand by combining similarity measures with IDM. All experiments involves two types of queries, one over whole frames and the other over isolated objects.

3.3.1. Single dictionaries

Figures 2 illustrate the performance of our method in its basic form, i.e.: unique dictionary, for frame queries. They emphasize the effect of LSA compared to direct processing. We explored several values for k the number of factors kept in LSA and 25 is the optimal number for all cases. Figure 2(a) shows the evolution of precision and recall curves with respect to k. The behavior is almost the same for all experiments. 25 five factors give the best result, closely followed by 50 factors. Then starting at 200, performance are similar and close to direct processing.

HS histogram performs well while texture features alone give poor results. This is mainly due to the weak texture information contained in cartoons.

Figures 3 illustrate the performance of our method for object queries. They lead to the same conclusions while the effect of LSA is more visible. The optimal factor found is still 25 for object queries and figure 3(a) shows a similar behavior as 2(a) with respect to k.

Of course performances are less impressive, this is mainly due to the weak occurrence information available in small parts of frames. It reveals the need to include more information via multiple dictionaries.

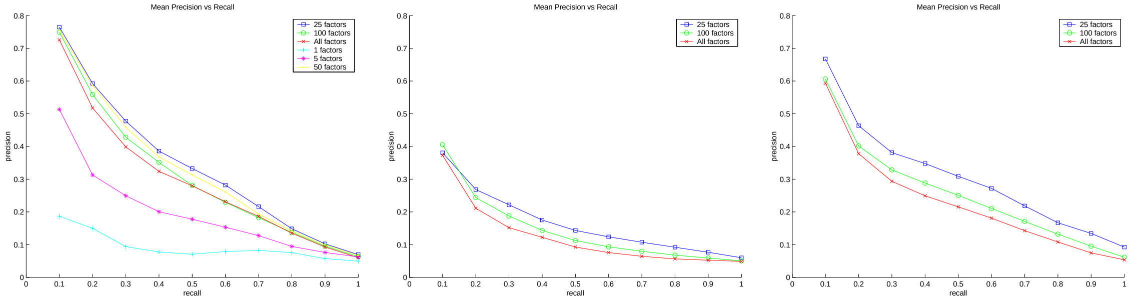
3.3.2. Twin dictionaries

Using multiple features at a later stage of the process slightly increases the performances as shown in figures 4 and 5. Both methods are comparable with a light improvement for MDM that entirely exploits co-occurrence information.

Figures 6 and 7 illustrate what is returned to the user for two queries on isolated part of a frame. One query is the shark present in the second part of the video, the other is the dog present in the last part of the video (fourth part).

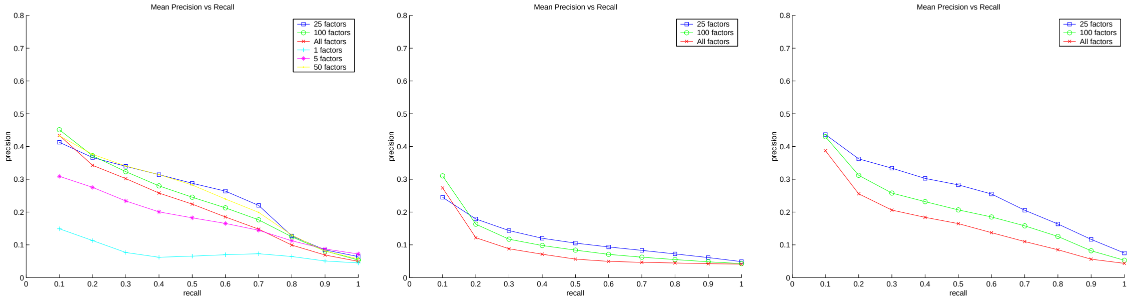
4. CONCLUSION

In this paper, we have presented a new model to represent video sequences through visual dictionaries and co-occurrences of words in contexts. Latent Semantic Analysis was then naturally included to improve robustness to the inherent noise present in data. It reduces the effect of noise while finding synonyms. This flexible model is a starting point to advanced video analysis tools, like shot detection, indexing and semantic analysis. We have shown interesting preliminary results to object and frame retrieval within a single video sequence that confirmed the potential of our novel model. In addition, we have tackled the problem of efficiently combining features through three solutions. Performances are summarized in figures 8. Frame retrieval performs very well when compared to object retrieval, which obtains only average results. This underlines the importance



(a) HS histograms dictionary (b) Gabor energies dictionary (c) Color and texture dictionary

Fig. 2. Performance evaluation of frame queries. Mean precision at standard recall values over all possible frame queries for 7 objects.



(a) HS histograms dictionary (b) Gabor energies dictionary (c) Color and texture dictionary

Fig. 3. Performance evaluation of object queries. Mean precision at standard recall values over all possible object queries for 7 objects.

of the co-occurrence information that is weak for isolated parts of the frame.

Future works will concern the enhancement and evaluation of the model for other applications. More specifically, investigations will focus on the creation and evaluation of optimal dictionaries. We will also envisage to integrate in the framework multi-modal dictionaries (visual, text and audio) to enrich the information available. Additionally, the co-occurrence analysis can be improved by pre-processing data as for text documents and/or using continuous values like a function of the distance to root words instead of just counting occurrences.

5. 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.

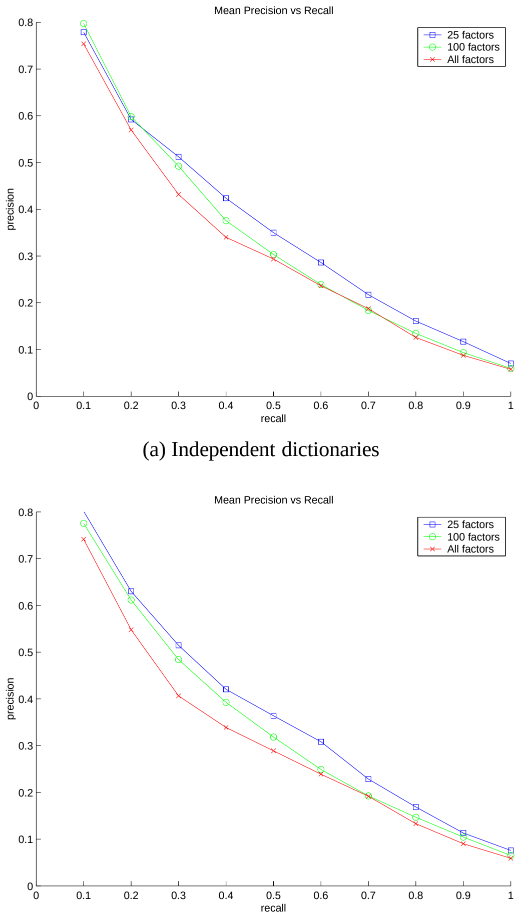
[3] Irena Koprinska and Sergio Carrato. Temporal video segmentation: A survey. *Signal Processing: Image Communication*, 16:451–460, 2001.

[4] Tong Lin and Hong-Jiang Zhang. Automatic video scene extraction by shot grouping. In *IEEE International Conference on Pattern Recognition*, volume 4, pages 39–42, 2000.

[5] Yu-Fei Ma, Lie Lu, and Hong-Jiang Zhang. A user attention model for video summarization. In *ACM Multimedia*, pages 533–542, December 2002.

[6] Yihong Gong and Xin Liu. Generating optimal video summaries. In *International Conference on Multimedia and Expo*, volume 3, pages 1559–1562, 2000.

[7] Shih-Fu Chang and Hari Sundaram. Structural and semantic analysis of video. In *International Conference on Multimedia and Expo*, 2000.



(a) Independent dictionaries

(b) Merged dictionaries

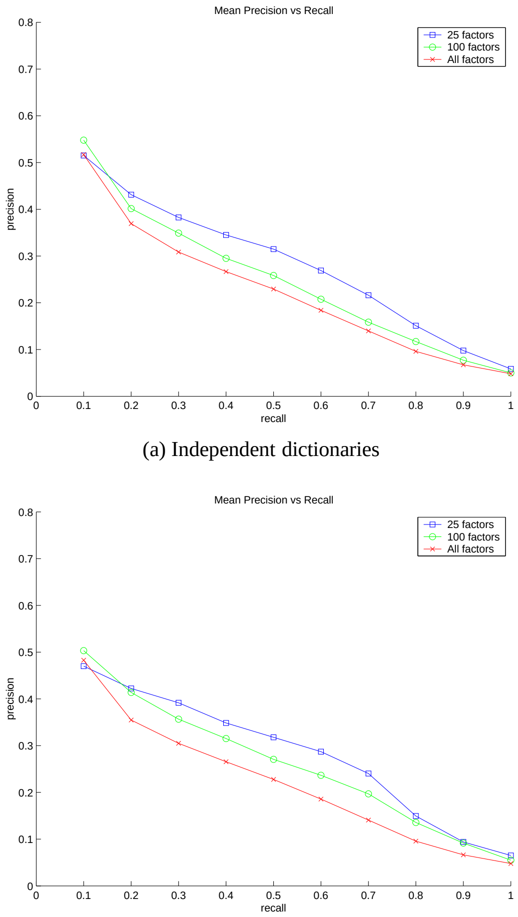
Fig. 4. Performance evaluation of frame queries. Mean precision at standard recall values over all possible frame queries for 7 objects.

[8] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. Semantic feature extraction using mpeg macro-block classification. In *The 11th Text REtrieval Conference (TREC)*, 2002.

[9] 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.

[10] 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.

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



(a) Independent dictionaries

(b) Merged dictionaries

Fig. 5. Performance evaluation of object queries. Mean precision at standard recall values over all possible object queries for 7 objects.

[12] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[13] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998.

[14] Susan T. Dumais. Improving the retrieval of information from external sources. *Behavior research methods, instruments and computers*, 23(2):229–236, 1991.

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

[16] 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.

[17] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[18] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001.

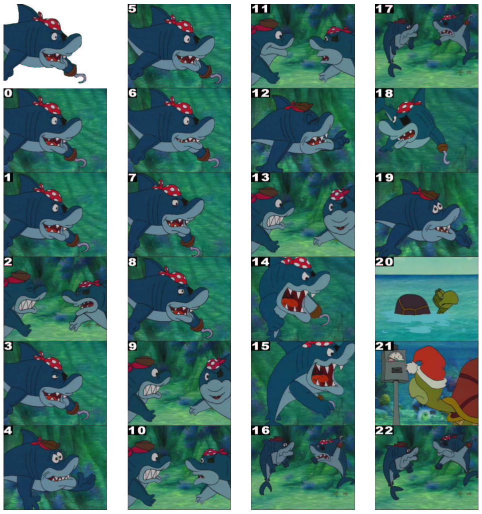


Fig. 6. First samples of retrieved images for one query on the shark. The first picture is the query.

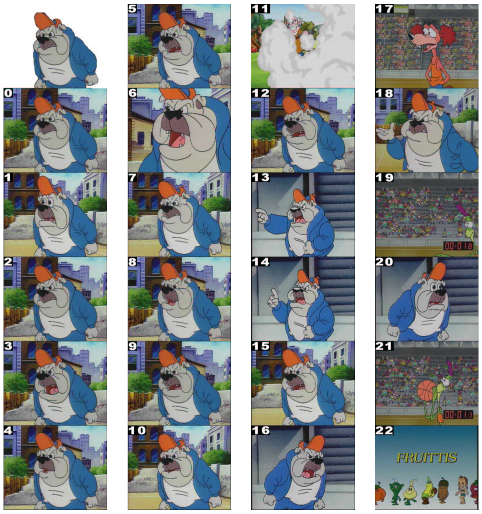
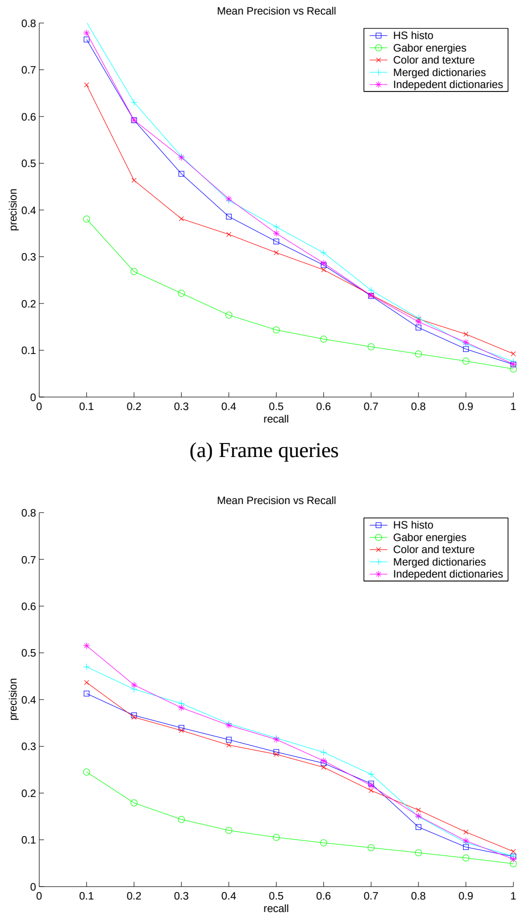


Fig. 7. First samples of retrieved images for one query on the dog. The first picture is the query.



(a) Frame queries

(b) Object queries

Fig. 8. Summary of the performances. Mean precision at standard recall values over all possible object queries for 7 objects with a LSA factor of 100.