LATENT SEMANTIC INDEXING FOR SEMANTIC CONTENT DETECTION OF VIDEO SHOTS

Fabrice Souvannavong, Bernard Merialdo and Benoît Huet

Departement Communications Multimédias
Institut Eurecom
2229 route des Crêtes
06904 Sophia-Antipolis - France
e-mail: {souvanna, merialdo, huet}@eurecom.fr

ABSTRACT

Low-level features are now becoming insufficient to build efficient content-based retrieval systems. The interest of users is not anymore to retrieve visually similar content, but they expect that retrieval systems 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 Indexing (LSI) was successfully introduced to efficiently index text documents. In this paper we propose to adapt this technique to efficiently represent the visual content of video shots for semantic content detection. Although we restrict our approach to visual features, it can be extended with minor changes to audio and motion features to build a multi-modal system. The semantic content is then detected thanks to two classifiers: k-nearest neighbors and neural network classifiers. Finally, in the experimental section we show the performances of each classifier and the performance gain obtained with LSI features compared to traditional features.

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 which aims at developing video content analysis and retrieval.

Currently, one of the main challenges in the field of image and video retrieval is to automatically bridge the gap from low-level visual features to the semantic content. Since three years, TREC [1] has been setting up a new track to encourage research and development in the domain of video content analysis, indexing and retrieval. In particular, one of the proposed task is the extraction of semantic features, like *people, indoors, news subject*, . . . , in video shots.

We propose a system to efficiently index visual features in order to extract the semantic content of video shots. The first step is conducted with an adaptation of Latent Semantic Indexing (LSI) to image or video content. LSI has been proven effective for text document analysis, indexing and retrieval [3]. Some extensions to audio and image features were then proposed [4, 5]. The adaptation we present models video shots in a similar way as text documents. Key frames of shots are described by the occurrence of a set of predefined region types. The underlying idea is that each region of an image carries a semantic information that influences the semantic content of the whole shot. In [6], 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 to detect its semantic content. The second step, i.e. the semantic analysis, is achieved thanks to the well-known k-nearest neighbors and neural network classifiers. The advantage of k-nearest neighbors classifiers resides in their independency with respect to data distribution, while neural network classifiers take advantage of label correlation in the context of multi-label classification.

The next section presents our adaptation of Latent Semantic Indexing to video shots. Next we present the k-nearest neighbors and neural network classifiers. Then we set up the experimental framework to discuss results and compare LSI to traditional features. Finally we conclude with a brief summary and future work.

2. LATENT SEMANTIC INDEXING

In the field of text document analysis, Latent Semantic Indexing (LSI) is a theory and method for extracting and representing the contextual meaning of words by statistical computations applied to a large corpus of text. The underlying idea is that the aggregate of all the 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 LSI's reflection of human knowledge has been established in a variety of ways [7]. 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

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

by individual students, and the quality and quantity of knowledge contained in an essay.

We propose to adapt this powerful tool to video content analysis. An image can be seen as a set of regions that individually contribute to the semantic content of the image. Moreover, an image composition is restricted by its semantic context. These two remarks highlight the high correlation between the semantic content and the composition of an image. And by definition, LSI allows to emphasize this mutual relationship. We present in the following our adaptation of LSI to video or image content.

For sake of simplicity, we assume that a video shot is well represented by its key frame. The LSI is then conducted in four steps. The first step consists in decomposing frames into regions. However image segmentation is a challenging task and a perfect solution does not exist for a general purpose. We assume that the segmentation provides homogeneous regions and favors over-segmentation (to avoid the exhausting of co-occurrence information). Thus, we expect that the occurrence information provided by the segmentation is robust enough to overcome segmentation variations. The second step classifies regions into region types in order to occupy a discrete space. The simplest solution to this problem is to use the k-means algorithm to vector quantize the region representation. We first model regions by two varieties of features proven effective in their category for content-based image retrieval [8] :

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.

Next, we build two dictionaries of region types, one quantifying color features and the second quantifying texture features. In the original approach presented in [9] a region is mapped to its nearest region type. However, when dealing with a large amount of video, this one-to-one mapping reveals inefficient. In order to add robustness to the clustering, we map a region to its k-nearest region types. Hence, the effect of quantification errors is diminished. A document d is thus described by $f(c)$ and $f(t)$ that are the occurrence vectors of color and texture region types. The third step is the Latent Semantic Indexing for each variety of features. LSI is obtained through a singular value decomposition of the occurrence matrix O. For a variety, O is defined as:

$$O_{i,j} = \text{occurrence of region type i in frame j}$$

The SVD gives the following factorization of O:

$$O = USV^t \tag{1}$$

$$\text{where } UU^t = VV^t = I \tag{2}$$

$$\text{and } S = diag(\sigma_1, .., \sigma_L) \tag{3}$$

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_L,\ L = \min(M, N) \tag{4}$$

In theory, removing smallest singular values $(\sigma_{L-k}, ..., \sigma_L)$ provides a least squared approximation of the original matrix O, therefore it can be seen as a simplification that reduces the noisy part of the co-occurrence matrix. 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, enhance region and image relationships and improve the performance of comparison tasks. Thus a threshold has to be defined to effectively remove noise while keeping the integrity of region equivalences. Any solution was proposed in the literature to solve this difficult issue. Empirically, we keep one third of eigenvalues. At last, we can demonstrate with some simple algebra that comparing two frames can be achieved by comparing their projection using the transformation matrix U_k such that:

$$\hat{O} = U_k S_k V_k \tag{5}$$

$$\text{and } S = diag(\sigma_1, .., \sigma_k) \tag{6}$$

We have separated varieties of features until the end of the process. Indeed in [9], we observed that merging features at the latest stage does not affect performances and has two advantages. First, features can easily be weighted again. Secondly, adding new features is very simple. Thus to compare two documents $d_1 = \{f_1(c), f_1(t)\}$ and $d_2 = \{f_2(c), f_2(t)\}$, we compute a similarity value, s_v, for each variety.

$$s_v(p_1, p_2) = cos(p_1, p_2) \tag{7}$$

$$p_1 = f_1(v)^t U_k^v,\ p_2 = f_2(v)^t U_k^v \tag{8}$$

The global similarity is then:

$$s_g(d_1, d_2) = \sum_{v \in \{c,t\}} \alpha_i s_v(d_1(v)^t U_k^v, d_2(v)^t U_k^v) \tag{9}$$

For instance $\alpha_i = 1$ and the optimal weighting will be the scope of a future work.

Now that we have an efficient representation of video shots and a similarity measure to compare them. We present the two solutions we have retained to attribute semantic labels to video shots.

3. SEMANTIC CLASSIFICATION

We define the classification problem as it follows. We have a database of video sequences, denoted D, whose shots have to be annotated. A shot is represented by a vector x taking values in X. Formally, the learning algorithm takes a set of training examples $L = \{(x_1, y_1), ..., (x_N, y_N)\}$ as input where $y_i = \{l_{i1}, ..., l_{iM}\} \in \{0, 1\}^M$ are the labels assigned to x_i. $l_{ij} = 1$ if the label j is present and 0 otherwise. It produces an hypothesis $f_L : X \mapsto \Re^M$ that in theory minimizes the generalization expected error:

$$E_L = \int_X E_{Y|X}[C(f_L(x), y)]P(x)dx \tag{10}$$

Where P(x) is the marginal distribution of x and $C : X, Y \mapsto \Re^+$ a predefined loss function.

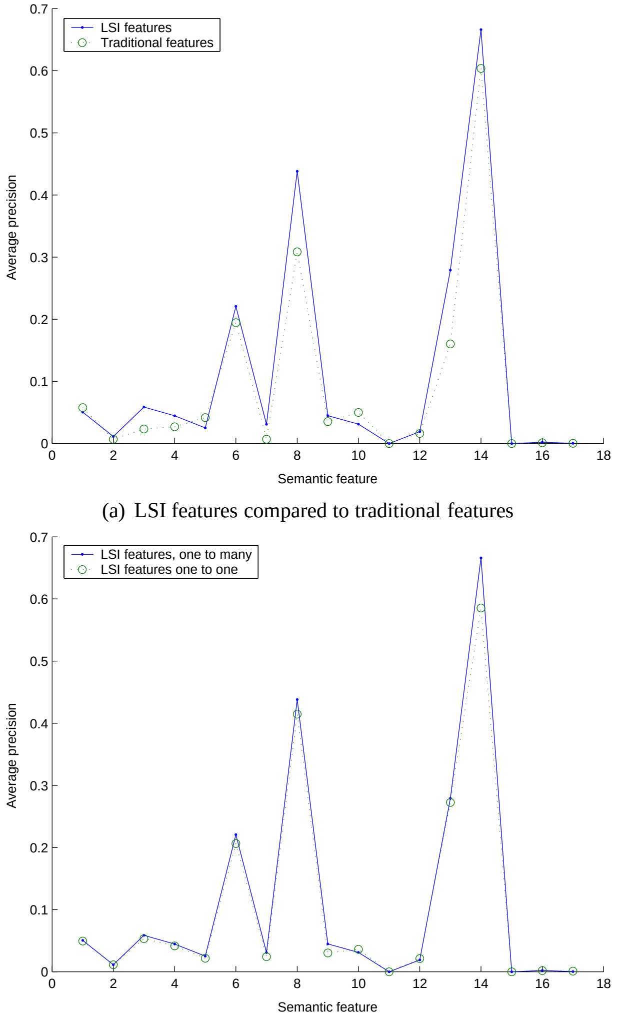
We propose two algorithms to achieve our classification task, one using k-nearest neighbors and the others based on a neural network composed of two hidden layers. On one hand we have selected the k-nearest neighbors classifier because no information about the distribution shape of the data is available. On the other hand, we expect the neural network classifier to learn the relationships between semantic concepts in order to improve their individual detection. We do not present these traditional algorithms. However for the evaluation presented in the next section, we assume that the output for each feature is equivalent to a detection score such that high values are more likely to indicate a high probability of detection. Thus we avoid the difficult task of threshold selection necessary to have a binary decision per semantic feature.

4. EXPERIMENTS

The classification task and the evaluation require annotated data. In June 2003, Video-TREC has launched a collaborative effort to annotate video sequences in order to build a labeled reference database. It 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 [10] was used for this time-consuming task. We use this huge annotated database to train classifiers and evaluate their performance. The evaluation is conducted like for Video-TREC. 17 features are involved: (1) Outdoors, (2) News-subject, (3) People, (4) Building, (5) Road, (6) Vegetation, (7) Animal, (8) Females-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 and (17) Madeleine Albright. For each feature, test documents are ordered with respect to their detection score value. Then the average precision at 2,000 documents is computed to characterize the performance of the system for each feature.

Figure (1) compares the performances obtained thanks to the k-nearest neighbors classifier. As we can see in figure (1(a)), LSI features give better performance for five semantic features while similar performances to traditional features are obtained for other semantic features. The gain we have by mapping regions to their k-nearest region types is weak, as shown in figure (1(b)), but encourages to a more thoroughly study of the problem.

Despite lower performances shown in figure (2), neural networks show a similar behavior. However the impact of LSI over performances is much more visible as we can see in figure (2(a)). A major gain obtained with LSI features over traditional features is observed for 50 percent of semantic features.



(a) LSI features compared to traditional features

(b) Impact of one to many mapping of regions

Fig. 1. Classification performance of Video-TREC features with the k-nearest neighbors classifier. Impact study of the one to many mapping and comparison of performances obtained with LSI and traditional features.

5. FUTURE WORK

We have presented Latent Semantic Indexing to efficiently model video contents. It gives an efficient representation of key-frame content. However the proposed adaptation relies on the creation of a codebook, operation that is often sub-optimal. To overcome this problem, we have introduced a method that improves noise robustness by matching a frame-region to its k-closest region types. We then used these LSI features to train two classifiers: k-nearest neighbors and neural networks classifiers. Finally classifier performances were compared and used to evaluate the gain obtained with LSI compared to traditional features. The LSI gain is significant for 30 percent of features and small for the remaining once.

Future works will take several directions. One disadvantage of Latent Semantic Indexing, as presented, is the lost of spatial information. Thus, efforts will be conducted to include spatial relationship between regions. On the other hand, we do not take advantage of the whole video content. New features, specific to video content analysis, will be included, such as object and camera motion, text and audio. Moreover a shot can be represented by all its frames instead of only its key-frame. Finally we expect to reduce the impact of segmentation variations by including a multi-level segmentation and representation of the visual content of shots.

6. 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*, 1998, vol. 8, pp. 602– 615.

[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 Interna-*

tional Conference on Image Processing, 1998, vol. 3, pp. 536–540.

[3] 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*, vol. 41, no. 6, pp. 391–407, 1990.

[4] Mikko Kurimo, "Indexing audio documents by using latent semantic analysis and som," in *Kohonen Maps*, Erkki Oja and Samuel Kaski, Eds., pp. 363–374. Elsevier, 1999.

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

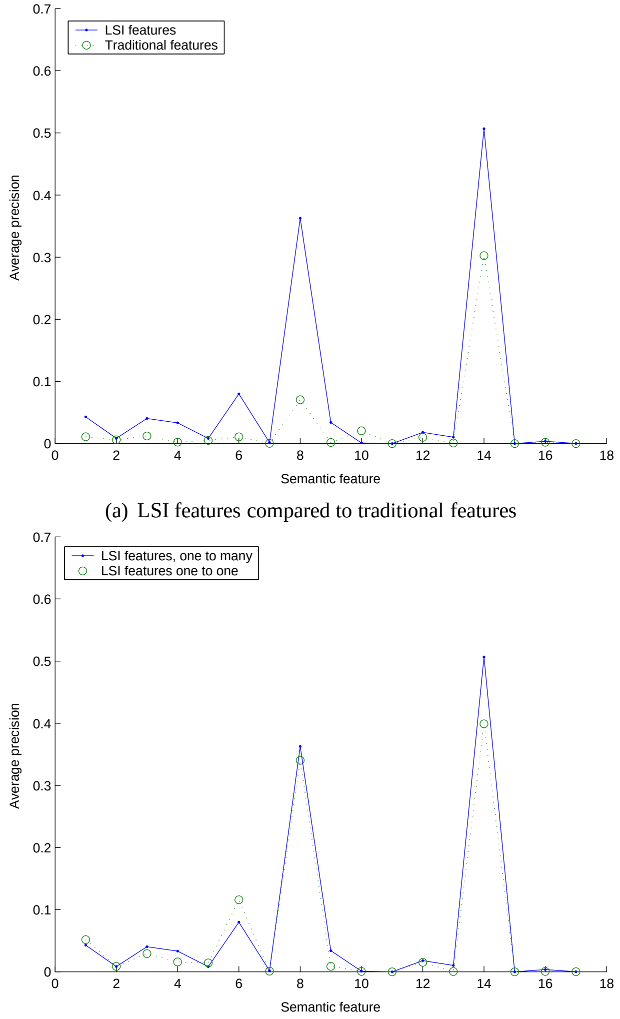
[6] Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *IEEE International Conference on Computer Vision*, 2002, pp. 97–112.

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

[8] 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*, 1998, vol. 1, pp. 253–257.

[9] Fabrice Souvannavong, Bernard Merialdo, and BenoˆHuet, "Video content modeling with latent semantic analysis," in *Third International Workshop on Content-Based Multimedia Indexing*, 2003.

[10] 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.



(a) LSI features compared to traditional features

(b) Impact of one to many mapping of regions

Fig. 2. Classification performance of Video-TREC features with the neural network classifier. Impact study of the one to many mapping and comparison of performances obtained with LSI and traditional features.