Multi-level Fusion for Semantic Video Content Indexing and Retrieval

Rachid Benmokhtar and Benoit Huet

Département Communications Multimédias
Institut Eurécom
2229, route des crêtes
06904 Sophia-Antipolis - France
(Rachid.Benmokhtar, Benoit.Huet)@eurecom.fr

Abstract. In this paper, we present the results of our work on the analysis of an automatic semantic video content indexing and retrieval system based on fusing various low level visual and edges descriptors. Global MPEG-7 features, extracted from video shots, are described via IVSM signature (Image Vector Space Model) in order to have a compact description of the content. Both static and dynamic feature fusion are introduced to obtain effective signatures. Support Vector Machines (SVMs) are employed to perform classification (One classifier per feature). The task of the classifiers is to detect the video semantic content. Then, classifier outputs are fused using a neural network based on evidence theory (NN-ET) in order to provide a decision on the content of each shot. The experimental results are conducted in the framework of the TrecVid feature extraction task.

1 Introduction

To respond to the increase in audiovisual information, various methods for indexing, classification and fusion have emerged. The need to analyse the content has appeared to facilitate understanding and contribute to a better automatic video content indexing and retrieval.

The retrieval of complex semantic concepts requires the analysis of many features per modalities. The task consisting of combining of all these different parameters is far from trivial. The fusion mechanism can take place at different levels of the classification process. Generally, it is either applied on signatures (feature fusion) or on classifier outputs (classifier fusion).

This paper presents our research conducted toward a semantic video content indexing and retrieval system aimed at the TrecVid high level feature detection task. It starts with a description of our automatic system architecture. We distinguish four steps: feature extraction, feature fusion, classification and classifier fusion. The overall processing chain of our system is presented in figure 1. The feature extraction step consists in creating a set of global MPEG-7 low level descriptors (based on color, texture and edges). Two feature fusion approaches are used: Static and dynamic. The static approach is based on simple operators while the dynamic approach consist in reducing the data dimensionality using Principal Component Analysis (PCA). Both are implemented and evaluated with the aim to obtain effective signature for each shot. The classification

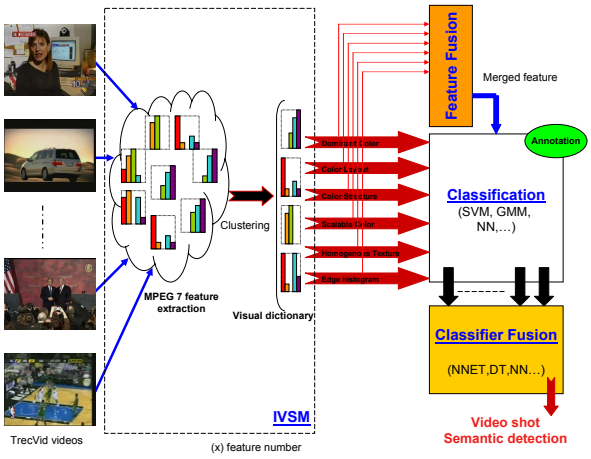

Fig. 1. General framework of the application.

step is used to estimate the video semantic content. Support Vector Machine (SVMs) are employed. In the final stage of our system, fusion of classifier outputs is performed thanks to a neural network based on evidence theory (NN-ET).

The experimental results presented in this paper are conducted on the TrecVid collection, varying the automatic generation techniques and combination strategies. Finally, we examine the outcomes of our experiments, detail our continuing work on how dynamic feature fusion could be used to complement, rather than replace existing approaches.

2 System Architecture

The MPEG-7 standard defines a comprehensive, standardized set of audiovisual description tools for still images as well as movies. The aim of the standard is to facilitate quality access to content, which implies efficient storage, identification, filtering, searching and retrieval of media [1]. We have used the following still image features:

- **Dominant color (DC)** represents the most dominant colors,
- **Color layout (CL)** specifies a spatial distribution of colors. The image is divided into (8x8) blocks and the dominant colors are solved for each block in the YCbCr color system. Discrete Cosine Transform is applied to the dominant colors in each channel and the DCT coefficients are used as a descriptor.
- **Color structure (CS)** slides a structuring element over the image, the numbers of positions where the element contains each particular color is recorded and used as a descriptor.
- **Scalable color (SC)** is a 256-bin color histogram in HSV color space, which is encoded by a Haar transform.
- **Edge histogram (EH)** calculates the amount of vertical, horizontal, 45 degree, 135 degree and non-directional edges in 16 sub-images of the picture.

– **Homogeneous texture (HT)** descriptor filters the image with a bank of orientation and scale tuned filters that are modeled using Gabor functions. The first and second moments of the energy in the frequency domain in the corresponding sub-bands are then used as the components of the texture descriptor.

Fig. 2. Example of key-frames illustrating three semantic concepts (*Person, airplane and car*).

The obtained vectors over the complete database are clustered to find the N most representative elements. The clustering algorithm used in our experiments is the well-known k-means with the Euclidean distance. Representative elements are then used as visual keywords to describe video shot content. Then, the occurrence vector of the visual keywords in the shots are built and this vector is called the IVSM signature (Image Vector Space Model). The number of visual terms used in our experiments is 70.

2.1 Static feature fusion

In this work, experiments describe an automatic detection of semantic concepts. Four color descriptors, edges histogram and homogenous texture descriptor are extracted from the visual content of a video shot. The main objective of feature fusion step is to reduce redundancy, uncertainty and ambiguity of signatures, in order to obtain a complete information of better quality, for take better decision and act.

Concatenation of features In the first fusion strategy, all global MPEG-7 descriptors are merged into a unique vector, that is called *merged fusion*(D_{merged}) as follow :

$$D_{merged} = [DC|CL|CS|SC|EH|HT] \tag{1}$$

All descriptors must have more or less the same numerical values to avoid scale effects [2].

Average of features This approach builds an average of the different descriptors. It requires no compilation of data, a simple normalization step is required before data can be added. It is interesting to give a weight or confidence level to each of the descriptors.

This method is commonly used, in particular in the automatic video concepts detection of the TrecVid project [3], where we observe the good contribution of the fusion operators as *Min* and average.

2.2 Dynamic feature fusion using PCA

Many techniques for dimensionality reduction have been proposed in the literature. However, Principal Component Analysis (PCA), latent semantic analysis (LSA) [4] and recently Independent Component Analysis (ICA) are the most frequently used. PCA extracts the features as the projections on the principal subspace whose basis vectors correspond to the maximum variance directions in the original space, while discarding the complementary subspace as a noise subspace. In some cases, PCA can obtain satisfactory performance. However, no theory can prove the complementary subspace is useless for recognition, and, on the contrary, experiments show that using the complementary subspace properly may improve recognition performance [5,6]. In our work, the dimension $m \in [10, 450]$ evolve per step of 20.

2.3 Support Vector Machines Classification

SVMs were widely used in the past ten years and they have been proved efficient in many classification applications. They have the property to allow a non linear separation of classes with very good generalization capacities. They were first introduced by Vapnik [7] for the text recognition task. The main idea is similar to the concept of a neuron: separate classes with a hyperplane. However, samples are indirectly mapped into a high dimensional space thanks to a kernel function. To this end, the selected kernel denoted $\mathcal{K}(.)$ is a radial basis function which normalization parameter σ is chosen depending on the performance obtained on a validation set. The radial basis kernel is chosen for his good classification results comparing to polynomial and sigmoidal kernels [4].

$$\mathcal{K}_1(x, y) = \exp\left(\frac{-||x - y||^2}{\sigma}\right) \tag{2}$$

2.4 Classifier fusion : Neural network based on evidence theory (NN-ET)

Classifier fusion is a necessary step to efficiently classify the video semantic content from multiple cues. For this aim, an improved version of RBF neural network based on evidence theory [8] witch we call NN-ET is used [9], with one input layer L_{input}, two hidden layers L_2 and L_3 and one output layer L_{output} (figure 3). Each layer corresponds to one step of the procedure described in following:

1. **Layer L_1:** Contains N units (prototypes). It is identical to the RBF network input layer with an exponential activation function ϕ and d a distance computed using training data. $\alpha \in [0, 1]$ is a weakening parameter associated to prototype i, where $\epsilon = 0$ at the initialization stage [9]:

$$\begin{cases} s^i = \alpha^i \phi(d^i) \\ \phi(d^i) = \exp\left(-\gamma^i (d^i)^2\right) \\ \alpha^i = \frac{1}{1+\exp\left(-\epsilon^i\right)} \end{cases} \tag{3}$$

   where γ^i is a positive parameter defining the receptive field size of prototype $i = \{1, ..., N\}$.

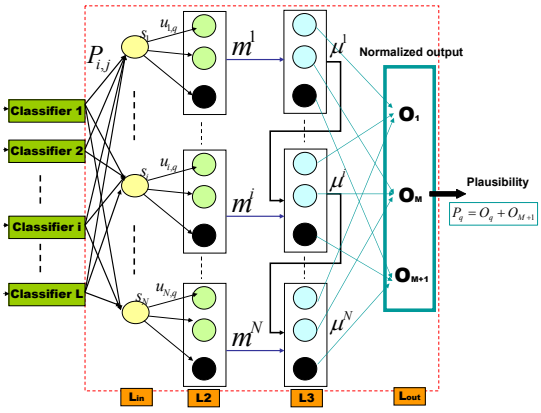

Fig. 3. Neural network based on evidence theory (NN-ET) classifier fusion structure

2. **Layer** L_2**:** Computes the belief masses m^i (Equ. 4) associated to each prototype. It is composed of N modules of $M+1$ units each (Equ. 5). The units of module i are connected to neuron i of the previous layer. Knowing that each image can belong to only one class (annotation clauses), we write:

$$\begin{cases} m^i(\{w_q\}) = \alpha^i u_q^i \phi(d^i) \\ m^i(\{\Omega\}) = 1 - \sum_{q=1}^{M} m^i(\{w_q\}) \end{cases} \tag{4}$$

hence,

$$\begin{aligned} m^i &= (m^i(\{w_1\}), ..., m^i(\{w_{M+1}\})) \\ &= (u_1^i s^i, ..., u_M^i s^i, 1 - s^i) \end{aligned} \tag{5}$$

where u_q^i is the membership degree to each class w_q, q classe index $q = \{1, ..., M\}$

3. **Layer** L_3**:** The Dempster-Shafer combination rule combines N different mass functions in one single mass. It's given by the following conjunctive combination:

$$m(A) = (m^1 \oplus ... \oplus m^N) = \sum_{B_1 \bigcap ... \bigcap B_N = A} \prod_{i=1}^{N} m^i(B_i) \tag{6}$$

The N mass function m^i are composed of N modules of $M+1$ units. The activations vector of modules i is defined as $\vec{\mu^i}$.

$$\begin{cases} \mu^i = \bigcap_{k=1}^{i} m^k = \mu^{i-1} \bigcap m^i \\ \mu^1 = m^1 \end{cases} \tag{7}$$

The activation vectors for $i = \{2, ..., N\}$ can be recursively computed using the following formula:

$$\begin{cases} \mu_j^i = \mu_j^{i-1} m_j^i + \mu_j^{i-1} m_{M+1}^i + \mu_{M+1}^{i-1} m_j^i \\ \mu_{M+1}^i = \mu_{M+1}^{i-1} m_{M+1}^i \end{cases} \tag{8}$$

4. **Ouput Layer:** We build the normalized output O defined as:

$$O_j = \frac{\sum_{i=1}^{N} \mu_j^i}{\sum_{i=1}^{N} \sum_{j=1}^{M+1} \mu_j^i} \tag{9}$$

The different parameters (Δu, $\Delta \gamma$, $\Delta \alpha$, ΔP, Δs) can be determined by gradient descent of output error for an input pattern x (more explanations see [9]). Finally, we compute the maximum of P_q (i.e the plausibility of each class w_q) as follow:

$$P_q = O_q + O_{M+1} \tag{10}$$

3 Experiments

Experiments are conducted on TrecVid videos [10]. The main goal of TrecVid is to promote progress in content-based retrieval from digital video via open, metrics-based evaluation. TrecVid is a laboratory-style evaluation that attempts to model real world situations or significant component tasks involved in such situations. It will test on *news reports, science news, documentaries, educational programming, and archival video*, to see how well the technologies apply to new sorts of data. In this work, about 5 hours of video (4000 shots) are used to train the feature extraction system and 1 hour (800 shots) are used for evaluation purpose. The training set is divided into two subsets in order to train both classifiers and subsequently determine through learning the fusion parameters. Detection performance was measured using the standard precision and recall metrics. We are interested by the precision to have a measure of the ability of a system to present only relevant shots. Average precision is given in follow:

$$\text{AP} = \frac{\left(\frac{\text{number of relevent shots retrieved}}{\text{total number of shots retrieved}}\right)}{\text{total number of relevent shots}} \tag{11}$$

The feature extraction task consists in retrieving video shots expressing one of the following six concepts (Sports, outdoor, building, mountain, waterscape, maps) among 36 proposed semantic concepts. Table 1 provides some insight about the composition in terms of our selected semantic concepts.

We start the experimentations with the description of three following system configurations:

- **System 1:** System without feature fusion step (see figure 1, we build one SVM classifier per feature, then fuse all classifier outputs by NN-ET);
- **System 2:** System with concatenation static feature fusion (we add in the system 1, the merged feature by concatenation to be classified);
- **System 3:** System with average static feature fusion (we add in the system 1, the merged feature by average to be classified).

Id	Concepts	*test*	*train*
1	Sports	19	86
2	Outdoor	260	512
3	Building	90	205
4	Mountain	12	45
5	Waterscape	23	108
6	Maps	13	29

Table 1. key-frames distribution of the video key-frames in the various sets by semantic concepts. The relative quantity of every class is clarified to give an idea of the lower border of the performances to be obtained

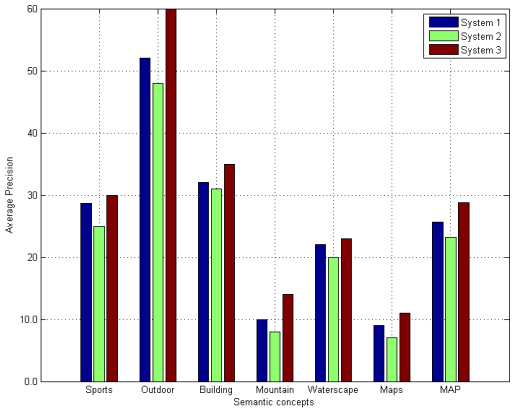

Fig. 4. Comparison of system configurations results

Figure 4 shows average precision results for the three distinct experiences. It can be seen that the *system 3* improves the average precision for all semantic concepts and obtains better scores (4% improvement on MAP) comparing to *system 1*. Contrary to the *system 2*, that decreases precision.

The average precision $AP \in [8, 60\%]$, for exemple the semantic concept (*outdoor*) obtain an average precision of 60%. This is can be explained by the high number of positive samples in the test set. Here, almost all positive samples are retrieved in the 100 first video shots returned by systems.

For semantic concepts (*mountain,maps*), the *system 3* obtain $(14\%, 11\%)$, which can be explained per the low number of positive samples in the training and test sets.

On average, the MAP oscillates around 29% using average feature fusion step, which represents a good performance considering the video shots annotation complexity.

Figure 5 shows the variation of average precision results using PCA feature fusion *vs* dimension for each concept. The dimension $dim \in [10, 450]$ per step of 20. The system with PCA step improve the precision of all concepts. The best $MAP = 36.32$

is obtained using $dim = 410$. In the test set, we have several monochrome video shots. We notice that the descriptors are highly redundant. This is not very surprising, because four of six investigated MPEG-7 descriptors are colour descriptors.

So, smaller dimensions $dim \in [10, 170]$ lead to loss of information and explains the bad performances with the small dimensions, and a high dimension raises the calculation time problem and also the relevance of our signatures. Observe that the stable dimension interval is $dim \in [190, 450]$.

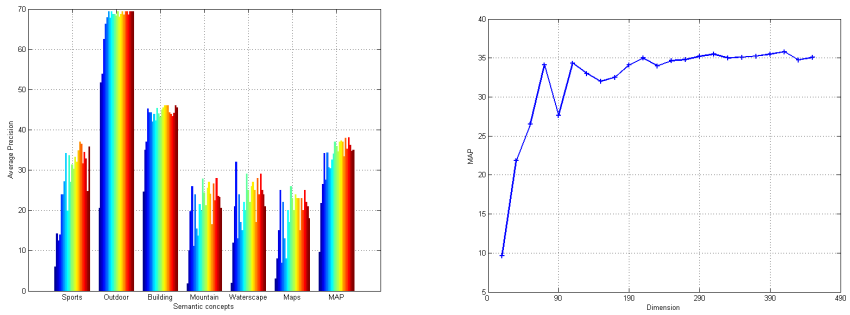

Fig. 5. NN-ET results using PCA feature fusion step. Mean Average Precision for PCA feature fusion, from 10 to 450 dimension by step of 20.

Systems	**Without Feature Fusion**	**Average Feature Fusion**	**PCA Feature Fusion**
MAP	25.61%	29.16%	36.32%

Table 2. Mean Average Precision (MAP) for different systems.

Finally, the table 2 summarizes the mean average precision (MAP) for different systems. We notice that PCA feature fusion system obtain superior results to those obtained by the static feature fusion for all semantic concepts and to system without feature fusion step. This shows the importance of feature fusion.

4 Conclusion

In this paper, both static and dynamic feature fusion approaches have been evaluated. Six global MPEG-7 visual descriptors are being employed for this difficult task. The aim of this feature fusion step is to provide a compact and effective representation for an SVM classifier which is trained to solve the challenging task of video content detection. A further classifier fusion step, featuring a neural network based on evidence

theory, is also employed within the proposed system in order to combine in the most effective way the output of the SVMs.

We have demonstrated through empirical testing the potential of feature fusion, to be exploited in video shots retrieval. Our model, achieves respectable performance, particularly, for certain semantic concepts like outdoor and building, when the variety of the quality of features used is considered.

Of course, these tests (Small data) do not guarantee the usability of global MPEG-7 descriptors in the general case, but they imply that global MPEG-7 descriptors are worth experimenting with. We start to investigate the effect of TrecVid'07 data in our system (50 hours of videos) with new semantic concepts like (Person, face, car, explosion,...).

We believe that the dynamic feature fusion of different MPEG-7 descriptors based on dimensionnality reduction has a positive impact on our system. Other statistical approaches, such as LDA and SOM are under investigation. In parrallel, we are extending the global MPEG-7 descriptors used in this paper with local MPEG-7 descriptor in order to enrich the low level representation and to study the effect of their addition to a fusion system like ours.

Acknowledgement

The work presented here is supported by the European Commission under contract FP6-027026-K-SPACE. This work is the view of the authors but not necessarily the view of the community.

References

1. M. Mottaleb and S. Krishnamachari, "Multimedia descriptions based on MPEG-7: Extraction and applications," *Proceeding of IEEE Multimedia*, vol. 6, pp. 459–468, 2004.
2. E. Spyrou, H. Leborgne, T. Mailis, E. Cooke, Y. Avrithis, and N. O'Connor, "Fusing MPEG-7 visual descriptors for image classification," *Proceedings of ICANN*, vol. 3697, pp. 847–852, 2005.
3. M. Rautiainen and T. Seppanen, "Comparison of visual features and fusion techniques in automatic detection of concepts from news video based on gabor filters," *Proceeding of ICME*, pp. 932–935, 2005.
4. F. Souvannavong, B. Merialdo, and B. Huet, "Latent semantic analysis for an effective region based video shot retrieval system," *Proceedings of ACM MIR*, pp. 243–250, 2004.
5. I. Jolliffe, "Principle component analysis," *Springer-Verlag*, 1986.
6. W. Zhang, S. Shan, W. Gao, Y. Chang, B. Cao, and P. Yang, "Information fusion in face identification," *Proceedings of IEEE ICPR*, vol. 3, pp. 950–953, 2004.
7. V. Vapnik, "The nature of statistical learning theory," *Springer*, 1995.
8. G. Shafer, "A mathematical theory of evidence," *Princeton University Press*, 1976.
9. R. Benmokhtar and B. Huet, "Neural network combining classifier based on Dempster-Shafer theory," *Proceedings of Multimedia MMM*, vol. 4351, pp. 196–205, 2007.
10. TrecVid, "Digital video retrieval at NIST," *http://www-nlpir.nist.gov/projects/trecvid/*.