# LOW-LEVEL FEATURE FUSION MODELS FOR SOCCER SCENE CLASSIFICATION

*Rachid Benmokhtar and Benoit Huet*[*]

Institut Eurécom - Département Multimédia
06904 Sophia Antipolis - France

*Sid-Ahmed Berrani*

Orange Labs-France Telecom
35510 Cesson - Sevigné - France

## ABSTRACT

This paper presents an automatic semantic concept extraction method which employs low level visual feature fusion. Both static and dynamic feature fusion approaches are studied and evaluated. The main contributions of this paper are: A novel dynamic feature fusion approach inspired from coding is proposed to create compact yet rich signatures; A statistical study of descriptors with and without fusion. To validate and evaluate our approach, we have conducted a set experiments on the classification of soccer video shots. These experiments show, in particular, that the feature fusion step of our system increases the classification rate of 17% comparing to a system without feature fusion.

## 1. INTRODUCTION

The emergence of multimedia technology coupled with the ever expanding image and video collections on the World Wide Web have attracted significant research efforts in providing tools for effective retrieval and management of visual informations. Many application domains making use of video data are available: Security, digital library, interactive TV, etc... Many of those rely on video content analysis and in particular video shot classification [1, 2].

Retrieving complex semantic concepts from video shots requires to finely analyze the video shot content and to extract a set of features best describing the content. Fusing these features toward an effective classification is however far from being trivial. The fusion mechanism can take place at different levels of the classification process. Generally, it is either applied on extracted features (Feature or early fusion) or on classifier outputs (Classifier or late fusion). The main objective of this paper is to show the importance and the role of fusion particularly at the feature fusion level. This study extends the architecture for video shot classification of [3] with a novel method for feature fusion which we call coder neural network. The general architecture of our semantic video content indexing and retrieval system is depicted in Figure 1. The overall chain can be divided into four parts: (1) Feature extraction, (2) feature fusion, (3) classification and (4) classifier fusion. The feature extraction step consists in extracting a set of low level features based on color, texture and edges. Two feature fusion approaches are used: A static approach and a dynamic one. The objective is to compute a compact and effective signature in order to describe key-frames of video shots. The static approach is based on average and concatenation operators while the dynamic approach uses coder neural network and Principal Component Analysis (PCA). The SVM classification step is used to feed the fusion layer, performed thanks to a neural network based on evidence theory (NNET) which labels video shots.

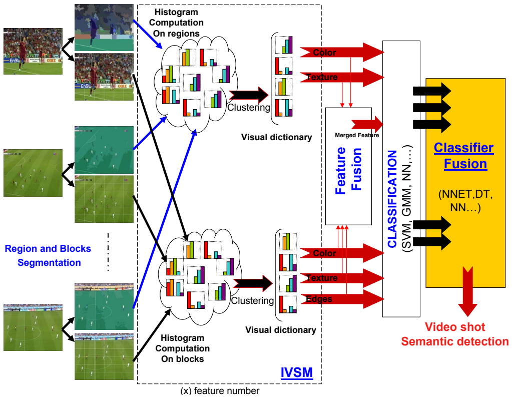


**Fig. 1**. General framework of the application.

In order to validate and evaluate our method, we have conducted a set of experiments on the classification of soccer footages because of its importance and commercial potential in several applications [4].

This paper is organized as follows. Section 2 describes our system architecture. We briefly discuss feature extraction as the basis for our system and we present our approaches for fusing the low level features. Section 3 evaluates the performance of the proposed system. Statistical studies of descriptors and fusion results are presented in Section 4. Finally, section 5 summarizes the main results and future research directions.

[*] The work presented here is supported by Orange-France Telecom R&D under research contract CRE 46134752.

## 2. SYSTEM ARCHITECTURE

This section describes the workflow of our system process.

### 2.1. Feature extraction

Key-frames are segmented using two techniques: Region and block segmentations. The first technique segments the image into homogeneous regions thanks to the graph-based image segmentation algorithm described in [5]. The second technique divides the image into a set of non-overlapping sub-images. Segments are represented using four visual features: RGB, HSV, energies of Gabor's filters, and edges histogram descriptor. Then, to reduce computation complexity and storage requirements, region and block features are quantized and video key-frames are represented using IVSM signatures (Image Vector Space Model) [2].

### 2.2. Feature fusion

The objective of feature fusion is to reduce the redundancy, the uncertainty and the ambiguity of signatures. Under this conditions, the fused feature should enable better classification performance. In our work, two approaches are compared, among them a novel one based on neural network.

*2.2.1.* **Static Approach:** This approach is based on simple operators such as **concatenation** and **average**. The descriptors are merged into a unique vector. It does not require any compilation of feature vectors. The average operator has need a simple sum of the IVSM region numbers for each key-frame. Although, it may be interesting to give a weight or a confidence level to each descriptor.

*2.2.2.* **Dynamic Approach:** It is based on dimensionality reduction, where the purpose is to map data onto low dimensional space, improving visualization. Several methods are proposed for the various learning problems and data mining, such as PCA, LDA, ICA, NMF [6, 7]. PCA is derived from eigenvectors (the principle components) corresponding to the largest eigenvalues of the covariance matrix for data of all classes. It seeks to optimally represent the data in terms of minimum mean square error between representation and the original data.

We propose to perform feature fusion in a novel way, using a trainable encoding scheme in order to reduce the dimension of the original data vector (Figure 2). The coder presents three layers (One input layer **X**, one hidden layer **U**, and one output layer **Y**). The particularity of this coder, is in the desired output **(Y=X)**. The system is trained by back-propagation and the number of neurons in the hidden layer defines the dimension of the merged feature. Unlike PCA, the coder makes a purely weighted additive representation, providing a compact learned feature vectors **U**. To analyze the power of the merged feature obtained by PCA and coder, a statistical study of descriptors before and after feature fusion is presented in Section 4.

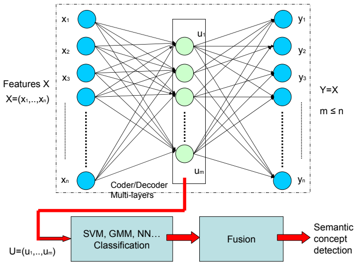


**Fig. 2**. Coder/decoder multi-layers perceptron scheme.

### 2.3. Classifier fusion (NNET)

In this part, we briefly describe our recently proposed neural network based on evidence theory (NNET) to address classifier fusion (Figure 3) [3].

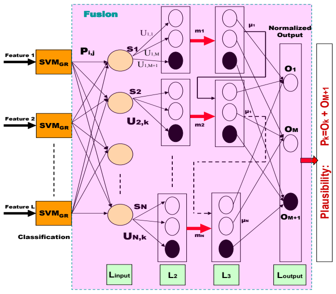


**Fig. 3**. NNET classifier fusion structure.

*1.* **Layer $L_{input}$:** Contains $N$ units. Identical to the RBF network input layer with an exponential activation function $\phi$. $d$: distance computed using training data. $\alpha \in [0,1]$ is a weakening parameter associated to unit $i$.

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

*2.* **Layer $L_2$:** Computes the belief masses $m^i$ (Equ. 2) associated to each unit. The units of module $i$ are connected to neuron $i$ of the previous layer.

$$\begin{cases} m^i(\{w_q\}) = \alpha^i u_q^i \phi(d^i) \\ m^i(\Omega) = 1 - \alpha^i \phi(d^i) \end{cases} \tag{2}$$

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

*3.* **Layer $L_3$:** The Dempster-Shafer combination rule combines $N$ different mass functions in one single mass. It is given by the conjunctive combination (Eq. 3):

$$m(A) = (m^1 \oplus ... \oplus m^N) = \sum_{B_1 \bigcap ... \bigcap B_N = A} \prod_{i=1}^{N} m^i(B_i) \tag{3}$$

The activation vector of modules $i$ is defined as $\overrightarrow{\mu^i}$. The activation vectors can be recursively computed using:

$$\begin{cases} \mu^1 = m^1 \\ \mu_j^i = \mu_j^{i-1} m_j^i + \mu_j^{i-1} m_{M+1}^i + \mu_{M+1}^{i-1} m_j^i \\ \mu_{M+1}^i = \mu_{M+1}^{i-1} m_{M+1}^i \end{cases} \tag{4}$$

*4.* **Layer $L_{output}$:** In [8], the output is directly obtained by $O_j = \mu_j^N$. The experiments show that this output is very sensitive to the number of prototype, where a small modification in the number can change the classifier fusion behavior. To resolve this problem, we use normalized output (Eq. 5). Here, the output is computed taking into account the activation vectors of all prototypes to decrease the effect of an eventual bad behavior of prototype in the mass computation.

$$O_j = \frac{\sum_{i=1}^{N} \mu_j^i}{\sum_{i=1}^{N} \sum_{j=1}^{M+1} \mu_j^i} \tag{5}$$

$$P_q = O_q + O_{M+1} \tag{6}$$

The different parameters ($\Delta u$, $\Delta\gamma$, $\Delta\alpha$, $\Delta P$, $\Delta s$) can be determined by gradient descent of output error for an input pattern $x$. Finally, the maximum of plausibility $P_q$ of each class $w_q$ is computed.

## 3. EXPERIMENTS

Experiments have been conducted on soccer videos. About 2256 key-frames have been used to train the feature extraction system and 1129 key-frames for evaluation. Classification task consists in retrieving key-frames expressing one of the 12 considered semantic concepts (Table 1). Performance is measured using the standard precision vs recall metrics. In order to assess the contribution of feature fusion, we designed five different systems (Table 2).

| Id | Concepts | *Test* | *Train* | *Total* |
|---|---|---|---|---|
| 1 | close-up action | 200 | 617 | 817 |
| 2 | game stop | 81 | 76 | 157 |
| 3 | goal camera | 5 | 13 | 18 |
| 4 | lateral camera | 50 | 92 | 142 |
| 5 | global center view | 217 | 507 | 724 |
| 6 | global rear view | 13 | 6 | 19 |
| 7 | global right view | 21 | 142 | 163 |
| 8 | global left view | 144 | 208 | 352 |
| 9 | zoom on public | 139 | 94 | 233 |
| 10 | zoom on player | 156 | 246 | 402 |
| 11 | aerial view | 15 | 15 | 30 |
|  | Others | 75 | 238 | 313 |

**Table 1**. Key-frames distribution of the video key-frames in the various sets by semantic concepts.

Figure 4.1 shows the average precision (AP) results for the five experiences. For concepts (*3,7*), all systems obtain high detection rate (100%, 86%) respectively. It is explained by the low number of positive samples in the test set (See Table 1). Here, almost all positive samples are retrieved in the 50 first key-frames returned by systems.

| Id | System |
|---|---|
| 1 | System without feature fusion step (See Figure 1). |
| 2 | System with a concatenation feature fusion approach. |
| 3 | System with an average feature fusion approach. |
| 4 | System with PCA feature fusion approach. |
| 5 | System with coder neural-network feature fusion approach. |

**Table 2**. Experiment systems.

For concept (*6*), *systems* have obtained bad detection rate, where the best one is given by the system 3 with $3\%$. It is due to two reasons: The first is the low number of positive samples in the training set and the second is due to the strong correlation between the global concepts (5,6,7,8). The MAP oscillates around $39.7\%$ using *system 3*, which represents a good performance considering the annotation complexity of the images under consideration.

We can notice that *System 5* improves the rest of concepts detection and obtains $17\%$ MAP improvement compared to *system 1* (Witout feature fusion level). The best MAP is $49,70\%$ using $dim = 390$.

*System 4* obtains the same results as *system 2* for concepts (2,5,8,10) and improves concept (11). The best MAP of *System 4* is $44,50\%$ using $dim = 270$.

Figure 4.2 shows MAP results for systems 4 and 5 *vs* dimension. The dimension $dim \in [10, 450]$ per step of 20. Smaller dimensions lead to loss of information which explains the poor performances. In higher dimensions signatures are very sparse and computation time is unnecessarily high. The best result is given by $dim \in [70, 400]$.

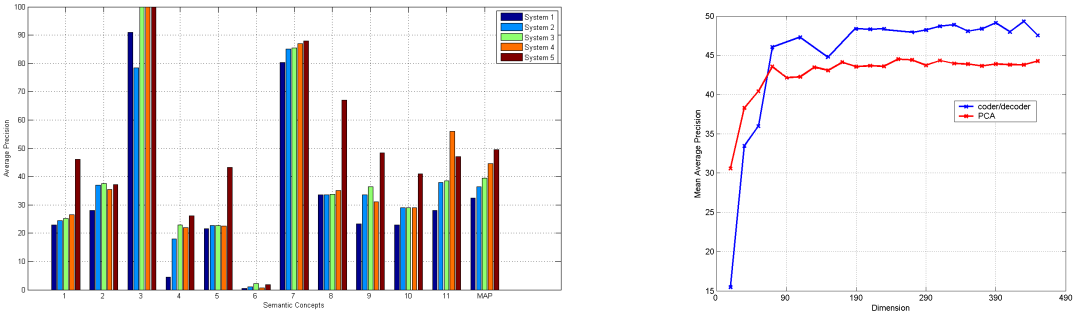
| **Systems** | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **MAP** (%) | 32.30 | 36.40 | 39.70 | 44.50 | 49.70 |

**Table 3**. Mean Average Precision for different systems.

The table 3 summarizes the MAP results obtained with the different systems. We notice that the Neural Network Coder obtains superior scores to those obtained by the other systems (1,2,3,4) for all semantic concepts. This supports further the importance of feature fusion.

## 4. STATISTICAL STUDY

The major quality indicators for description extraction methods are the characteristics of the output descriptor elements. The characteristics can be measured as variance *within* vectors, proximity *between* descriptor elements, distributions



(1)

(2)

**Fig. 4**. (1) Comparison of system configurations results. (2) Mean Average Precision for dynamic feature fusion, from 10 to 450 dimension.

of quantized vector elements, etc. In this work, three methods were used: (1) Mean and standard deviation, (2) factor analysis, (3) hierarchical cluster analysis.

### 4.1. Before feature fusion

The average mean is about $0.6$ and the standard deviation is lower than $0.1$ (Values are normalized to $[0, 1]$). This indicates that descriptors values could be transformed and quantized to a smaller data type.

Interesting findings are obtained thinks to hierarchical cluster structure analysis. It intuitively shows the proximity between elements. It starts by clustering the elements with high color similarity, the distance is less than $50$. Then, it clusters the elements with high edges histograms and the elements with Gabor texture. Finally, it is easy to see that color features are more redundant and dimensionality reduction is suitable comparing to texture dominant elements.

### 4.2. After feature fusion:

Now, standard deviation is $0.31$ for PCA and $0.43$ using coder neural-network with $dim = 270$. In the figure 4.2, we can notice that $70$ factors are able to explain the global variance. The descriptors before fusion are highly redundant. This is not very surprising, because four of the seven investigated descriptors are colors based. Of course, this is a problem if color descriptors should be applied on media objects with monochrome content.

## 5. CONCLUSION

We have presented both static and dynamic low-level feature fusion models based on dimensionality reduction via PCA and information coding neural-network.

We have demonstrated through statistical study and 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 close up action, zoom on player, center view, etc. when the variety of the quality of features used is considered. Results obtained by the dynamic approach on soccer video using precision/recall evaluation protocol report the efficiency of fusion mechanism (before and post classification) and demonstrate the improvement provided by such a combination with an effective signatures coding and lowers computation time by reducing the dimensionality, compared to the system without feature fusion level.

Future works will take several directions. We start a program of work about ontology study between the classes and the exploitation of this semantic information on our classification or fusion system.

## 6. REFERENCES

[1] M. Rautiainen and T. Seppanen, “Comparison of visual features and fusion techniques in automatic detection of concepts from news video based on gabor filters,” *Proceedings of IEEE ICME*, 2005.

[2] F. Souvannavong, B. Merialdo and B. Huet, “Latent semantic analysis for an effective region based video shot retrieval system,” *Proceedings of MIR*, 2004.

[3] R. Benmokhtar, B. Huet,“Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content,” *Proceedings of MMM*, vol. 4351, pp. 196–205, 2007.

[4] L. Duan, M. Xu, T. Chua, Q. Tian and C. Xu, “A mid-level representation framework for semantic sports video analysis,” *Proceedings of ACM MM*, pp. 33–44, 2003.

[5] P. Felzenszwalb and D. Huttenlocher, “Efficiently computing a good segmentation,” *Proceedings of CVPR*, pp. 98–104, 1998.

[6] I. Jolliffe, “Principle component analysis,” *Springer-Verlag*, 1986.

[7] D.D. Lee and H.S. Seung, “Algorithms for Non-negative Matrix Factorization,” *Proceedings of NIPS*, vol. 13, pp. 556–562, 2000.

[8] T. Denoeux, “An evidence-theoretic neural network classifier,” *Proceedings of IEEE SMC*, vol. 3, pp. 712–717, 1995.