VIDEO CLASSIFICATION BASED ON LOW-LEVEL FEATURE FUSION MODEL

Mickael Guironnet, Denis Pellerin, and Michele Rombaut

Laboratoire des Images et des Signaux
46 Avenue Felix Viallet, 38031, Grenoble, France
email: firstname.lastname@lis.inpg.fr
web: www.lis.inpg.fr

ABSTRACT

This article presents a new system for automatically extracting high-level video concepts. The novelty of the approach lies in the feature fusion method. The system architecture is divided into three steps. The first step consists in creating sensors from a low-level (color or texture) descriptor, and a Support Vector Machine (SVM) learning to recognize a given concept (for example, “beach” or “road”). The sensor fusion step is the combination of several sensors for each concept. Finally, as the concepts depend on context, the concept fusion step models interaction between concepts in order to modify their prediction. The fusion method is based on the Transferable Belief Model (TBM). It offers an appropriate framework for modeling source uncertainty and interaction between concepts. Results obtained on TREC video protocol demonstrate the improvement provided by such a combination, compared to mono-source information.

1. INTRODUCTION

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

Recent advances in content analysis have allowed video annotation systems to be developed. However, the difficulty consists in bridging the gap between low-level features and semantic concepts. Semantic video classification techniques are divided into two categories: *(i)* Rules-based approaches exploit a priori knowledge on a particular domain (like for example sport video) for extracting concepts [1, 2], *(ii)* Statistical approaches try to achieve annotations with an independent analysis of videos. The statistical methods generally require a training to categorize video scenes from low-level features automatically. Different classifiers can be proposed, for example, Bayesian networks [1] or Support vector machine (SVM) classifier [3]. In this last case, different combinations of features (only color descriptor, concatenation of color and texture descriptors,. . .) are considered to create the SVM model and the combination that gives the best results is chosen as the optimal combination. In [4], several features are extracted and each feature is used to train an Artificial Neural Networks (ANN) classifier. From these classifiers, a combination method is achieved based on Dempster-Shafer theory. However this combination requires a training again to minimize the mean square error between the combined output and the target output of a training set.

In this paper, a system framework for automatically semantic video annotation is described. The system architecture is depicted in figure 1. We distinguish three steps: *sensor*, *sensor fusion* and *concept fusion*. Firstly, the *sensor* step consists in creating a set of sensors from a low-level descriptor (based on color or texture) and a Support Vector Machine (SVM) learning to allow a given concept (for example, boat, beach or basketball. . .) to be predicted. Secondly, the *sensor fusion* step corresponds to the combination of sensors for each concept. Finally, the *concept fusion* step models interaction between concepts. The fusion method is achieved using the Transferable Belief Model (TBM). This is appropriate to model source uncertainty and combine information. We have applied our system using TREC video protocol [5]. The interest of working with the TREC video base is the great quantity of data (reference video segmentation, ground truth for high-level concept extraction. . .)

The rest of this paper is organized as follows: In section 2 the image descriptors are presented. Section 3 describes the method of high-level concept extraction. In section 4, the results of the method are shown. Finally, we present our conclusions.

2. LOW-LEVEL DESCRIPTORS

It is well-known that color and texture are visual cues used for image classification or similarity searches. Color and texture are important features in image perception.

2.1 Color descriptor

Among the color descriptors, we retain color histogram which captures global color distribution in an image. Selected color space is YCbCr space, which is used in compression MPEG. However, we do not use a uniform quantification of the color space which gives the same weight to the pixels near the centre of a bin as those that are located at the edges. The use of fuzzy sets allows each pixel to associate a membership degree to each bin. Each component of YCbCr is quantified into 8 bins as shown in figure 2. For each pixel p, the bins b_i, b_j and b_k are computed, respectively for each component of YCbCr as:

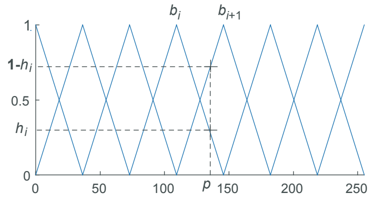

Figure 2: A membership degree of a pixel to each bin.

$$\left\{ \begin{array}{l} h(b_i) = h_i \\ h(b_{i+1}) = 1 - h_i \end{array} \right. \tag{1}$$

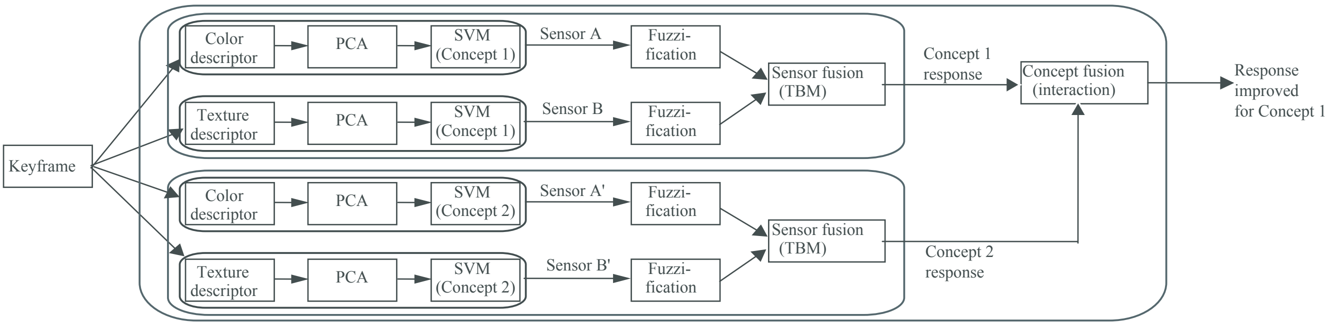

Figure 1: System architecture for sensor fusion and concept fusion.

The histogram is updated as follows:

$$For\, i,j,k = 0\, to\, 1,\; h_{3d}(b_i,b_j,b_k) = h(b_i)h(b_j)h(b_k) + h_{3d}(b_i,b_j,b_k) \quad (2)$$

Finally, after a normalization by the image size, a fuzzy 3D histogram with 8x8x8 components is obtained.

2.2 Texture descriptor

Texture has been widely studied in recognition tasks. While many computational approaches have been proposed, we chose to design a descriptor inspired by human perception and adapted to describe video content. This descriptor is divided into two steps: retinal filter followed by a Gabor decomposition.

At the first level of image processing, the retinal filter [6] performs an adaptive compression of brightness intensity followed by high-pass filtering. It provides a relative insensitivity to local illumination variations and then carries out a spectral whitening compensating for the 1/f image amplitude spectrum of natural images.

In the primary visual cortex, cells are sensitive to stimuli having a certain orientation and a certain frequency with a specific position in the visual field. Here, we chose to model this using two-dimensional Gabor function. Figure 3 shows the bank of Gabor filters used in our experimentation. We carried out this filtering by directly multiplying the retina output image weighted by a Hanning window with the Gabor filter in the Fourier domain. Finally, we obtained the energy $E(f_k,\theta_i)$ according to 7 spatial frequencies and 7 orientations. A normalization [7] is then applied to be invariable to image quality. Indeed, the blur is an isotropic function of the $G(f)$ frequency and the normalization carried out by frequency band removes this term.

$$E(f_k,\theta_i) = \frac{E(f_k,\theta_i)G(f_k)}{\sum_j E(f_k,\theta_j)G(f_k)} = \frac{E(f_k,\theta_i)}{\sum_j E(f_k,\theta_j)} \quad (3)$$

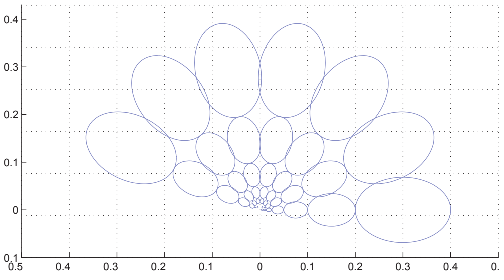

Figure 3: Bank of 49 Gabor filters.

Each keyframe is characterized by a matrix 7x7 where each component corresponds to energy for an orientation and a frequency.

3. HIGH-LEVEL FEATURE EXTRACTION

The fusion method aims to combine the response of the different sensors for a given concept and then model interaction between concepts.

3.1 Sensors

The sensors are created from low-level descriptors (color or texture). After a dimensionality reduction of descriptors, a classifier is used to recognize each concept. A principal component analysis (PCA) was performed to reduce feature dimensionality. The training is only applied on the keyframes of TREC 2004 development data. For each descriptor, the selected dimension is chosen to retain at least 99% of the variance. In our experiments, we went from 512 to 64 components for color descriptor and from 49 to 32 for texture descriptor.

Then, the Support Vector Machine (SVM) classifier was applied to learn each concept of TREC video. SVM is successfully used in a variety of pattern recognition tasks. We will quickly describe the principle of the SVM, a more detailed description can be found in [8]. Let $\{x_1 \cdots x_n\}$ be a set of training data which are feature vectors of labeled images. We are also given their labels $\{y_1 \cdots y_n\}$ where $y_i \in \{-1,1\}$. The problem consists in approximating an unknown function g such as:

$$g(x) = \sum_{i=1}^{L} y_i \cdot w_i \cdot K(x_i,x) + b \quad (4)$$

where $K(\,,\,)$ is the kernel function, x_i are called support vectors determined from training data, L is the number of support vectors, y_i is the label associated with each x_i, and w_i, b are constants determined from training. In this study, the commonly used radial basis function (RBF) kernel is considered. SVM classifier consists in finding the hyperplane that separates the training data with a maximal margin. The classification of a new vector x is given by the sign of decision function g. Nevertheless we prefer to consider a confidence measure for classification. The perpendicular distance from the hyperplane to vector x is used as a confidence measure. We apply the SVM_light developed by Thorsten Joachims [9] with the default parameters.

A ground truth of each concept is carried out on TREC 2004 development data and allowed SVM model to be created. The classifiers are then applied on the keyframes of

TREC 2004 test data. The important thing in this article is not the classifier used but the way in which we combine the classifiers. Finally, for each concept, two sensors are created from color descriptor and texture descriptor.

3.2 Sensor fusion step

The fusion method is based on Transferable Belief Model [10] coming from the Dempster-Shafer theory. This tool is adapted to deal with imprecise information.

3.2.1 Transferable Belief Model

The set of hypotheses is defined: $\Omega = \{H_1, \cdots, H_n\}$. The different sources of information will give a belief to subsets $A_i \in \Omega$. For each source, a Basic Belief Assignment (BBA) is defined as:

$$\begin{array}{rccc} m: & 2^\Omega & \rightarrow & [0,1] \\ & A_i & \rightarrow & m(A_i) \end{array} \tag{5}$$

where 2^Ω is the set of all subsets of Ω and $m(A_i)$ is called basic belief masses and represents a confidence measure that is assigned to the subset A_i. The attribution of BBA for each information source is constrained by the following rules:

$$\begin{array}{l} m(\emptyset) = 0 \\ \sum_{A_i \in 2^\Omega} m(A_i) = 1 \end{array} \tag{6}$$

where $\emptyset$ is the empty set. Let m_1 and m_2 be the BBA respectively attributed by source 1 and source 2, their conjunctive combination is defined as:

$$m_{12}(A_i) = \sum_{A_j \cap A_k = A_i} m_1(A_j) \cdot m_2(A_k) \tag{7}$$

3.2.2 BBA definition

Without statistical knowledge, the fuzzy sets can be used to model the BBA from the output x of the SVM learned on a given concept. Figure 4 shows how SVM output is used to attribute a confidence measure. The distribution of SVM output (fig 4.a) corresponds to the perpendicular distance from the hyperplane of SVM model to the image vector. The closer the distance is to zero, the more likely the classifier is to make an error and reciprocally. We then defined the BBA as shown in figure 4.b. If the output of the "Beach" concept is considered, the set of hypotheses is $\Omega = \{H_1 = Beach, \overline{H_1} = \overline{Beach}\}$. $m_x(H_1)$ is the confidence that is assigned to the "Beach" concept , $m_x(H_1 \cup \overline{H_1})$ represents the doubt about the "Beach" concept and $m_x(\overline{H_1})$ is the confidence that is not the "Beach" concept.

3.2.3 Sensor fusion

The purpose of sensor fusion is to model the fusion of several sensors on the same concept. Each sensor carries out an observation and assigns its confidence over the set Ω. In our experimentation, two sensors are defined from color descriptor and texture descriptor. Table 1 illustrates the combination of two sensors which relates the same concepts H_i. A mass is often assigned to the empty set. This is interpreted such as a conflict between the sensors. The conflict means that one of the sensors makes a mistake but that one does not know which. The mass in conflict $H_i \cap \overline{H_i} = \emptyset$ is then transfered on $H_i \cup \overline{H_i}$. This avoids making decision. Thus the combination with other sensors will be able to allow a correct classification.

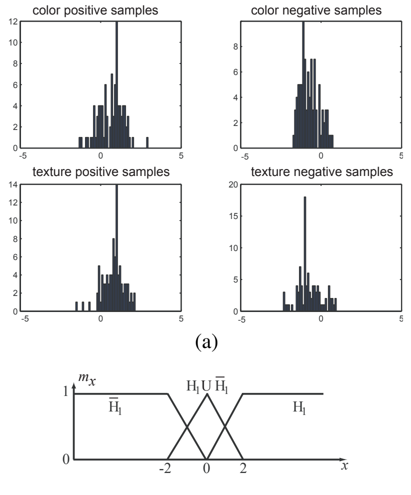

Figure 4: (a) Example of distribution of the SVM output on the training set of "Beach" concept. (b) BBA definition from SVM output.

Table 1: Interaction of two sensors m_1 and m_2 for the same concept.

		m_2		
		H_i	$H_i \cup \overline{H_i}$	$\overline{H_i}$
m_1	H_i	H_i	H_i	$H_i \cup \overline{H_i}$
	$H_i \cup \overline{H_i}$	H_i	$H_i \cup \overline{H_i}$	$\overline{H_i}$
	$\overline{H_i}$	$H_i \cup \overline{H_i}$	$\overline{H_i}$	$\overline{H_i}$

3.2.4 Decision

We use the TREC protocol to evaluate classification and this requires a ranked shot list per concept to be submitted. The rules of decision are then applied to singleton hypotheses H_i. The mass $m(H_i)$ is selected for the final decision and represents the degree of confidence placed exactly on the concept H_i. A shot can contain one or more keyframes and the passage of keyframes to shot is carried out in the following way:

$$v_i = \max_{I_k \in S_i}(m(H)\{I_k\}) \tag{8}$$

where I_k keyframes belong to a S_i shot and v_i is the confidence degree of the concept H placed on the S_i shot.

3.3 Concept fusion step

The concept fusion step models interaction between concepts. The principle consists in combining a concept with another concept having a good reliability to improve the classification. The BBA on a set $\Omega_1 = \{H_1, \overline{H_1}\}$ can also be combined with the BBA of another set $\Omega_2 = \{H_2, \overline{H_2}\}$ if a relation exists between H_1 and H_2. For instance, if they are exclusive, Table 2 shows how the combination is carried out. Several strategies can be adopted to deal with the empty set. As previously, the mass is transfered on the union $H_1 \cup \overline{H_1}$.

4. EXPERIMENTS AND DATA ANALYSIS

In this section, our purpose is to judge the effectiveness of the fusion method previously described.

Table 2: Combination of the m_1 concept with a m_2 concept having good reliability.

		m_2		
		$H_2 = \overline{H_1}$	$H_2 \cup \overline{H_2} = H_1 \cup \overline{H_1}$	$\overline{H_2} = H_1 \cup \overline{H_1}$
m_1	H_1	$H_1 \cup \overline{H_1}$	H_1	H_1
	$H_1 \cup \overline{H_1}$	$\overline{H_1}$	$H_1 \cup \overline{H_1}$	$H_1 \cup \overline{H_1}$
	$\overline{H_1}$	$\overline{H_i}$	$\overline{H_i}$	$\overline{H_i}$

4.1 Data

We considered a high-level feature extraction task over a keyframe dataset of TREC Video data. The elementary unit in the context of TREC is a shot and a shot can contain one or more keyframes. This data is provided by the National Institute of Standards and Technology (NIST). The TREC 2004 development data consists of 254 videos which is represented by 138823 keyframes and the TREC 2004 test data has 128 videos with 48818 keyframes. The video collection contains CNN or ABC news and advertisements.

4.2 Results of fusion

The evaluation of the ranked shot list is performed by average precision (AP) and total number of relevant documents (NRD) returned for a given concept. Average precision is the mean of the precision value obtained after each relevant shot is retrieved. Table 3 illustrates the method of fusion. The results show that the number of documents found by combining the color and texture sensors (Sensor fusion) is higher or equal to the number of documents found independently by the sensors. Obviously, the results depend on the kind of descriptor and the training carried out with the classifier. If the sensors taken separately are not very effective, their combination will not be able to find all the documents. The important point is that the combination of sensors improves the number of documents found. A studied concept can be put in competition with another concept of better quality in order to improve the precision. It allows false alarms of ranked shot list to be removed. A new concept is then defined containing mono-color and few textured images because often black images are inserted in videos. The concepts interact with the "Monocolor" concept and the results are shown in Table 3 (Concept fusion). The results show that the number of documents found and the precision increases.

Table 3: Results of sensor and concept fusion. The first line of each concept corresponds to total number of relevant documents and the second one is average precision.

Concept	Tot. Number Relevant Doc.	Texture sensor	Color sensor	Sensor fusion	Concept fusion
Boat	441	62	120	120	120
Ship		0.0054	0.0185	0.0182	0.0191
Beach	374	84	139	139	145
		0.0143	0.0358	0.036	0.0381
Basket	103	11	22	34	34
scored		0.0006	0.0049	0.0065	0.0071
Airplane	62	18	14	20	28
takeoff		0.0045	0.002	0.0049	0.0327
People	1695	153	167	171	191
walking		0.009	0.0084	0.0081	0.0102
Physical	292	28	41	50	54
violence		0.0016	0.0024	0.0036	0.0048
Road	938	205	162	243	279
		0.0418	0.0128	0.0322	0.0429

This process is iterative and the concept fusion output can be combined with other concepts. For example, if a "Natural landscapes" concept is created, we will be able to combine it with the "Basket scored" concept but not with the others because they are not exclusive of landscapes. This combination still improves the results with 37 relevant shots found against 34 and a mean average precision equals 0.0117 against 0.0065. The difficulty consists in finding new concepts which will be able to remove the false alarms as well as possible.

5. CONCLUSION

We have presented a method of high-level concept extraction. This approach is divided into three steps. First, sensors are created for each concept from color or texture descriptors and SVM learning. Then, the sensor fusion is performed for each concept to improve the classification. Finally, the concept fusion models interaction between concepts. The fusion method is based on the Transferable Belief Model. This tool is adapted to model imprecise information of sensors. Results obtained on TREC video protocol demonstrate the improvement provided by such a combination, compared to mono-source information. The fusion method can be applied with other sensors (audio information) and can also be used to model other interactions between concepts.

REFERENCES

[1] H. Shih, and C.Huang, "Semantic network modeling for understanding baseball video," in *Proc. ICASSP 2003*, Hong-Kong, Apr. 6-10. 2003.

[2] W. Zhou, A. Vellaikal, and C. C. Jay Kuo, "Rule-based video classification system for basketball video indexing," in *Proc. ACM Mult*, San Franscico, California, USA, June 3-5. 2000, pp. 213–216.

[3] M. R. Naphade, "On supervision and statistical learning for semantic multimedia analysis," *Journal of Visual Communication and Image Representation*, 15(3):348–369, 2004.

[4] A. Al-Ani and M. Deriche, "A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence," *Artificial Intelligence*, 17:333–361, 2002.

[5] A Smeaton, W. Kraaij and P. Over, "TRECVID 2004 An introduction", *13th Text Retrieval Conference*, USA, 2004.

[6] W. H. A. Beaudot, "Sensory coding in the vertebrate retina: towards an adaptive control of visual sensitivity," *Journal Network: Computation in Neural Systems*, vol. 7, no. 2, pp. 317–323, May 1996.

[7] N. Guyader, and J. Herault, "Representation espace-frequence pour la categorisation d'images," in *Proc. GRETSI 2001*, Toulouse, France, Sept. 10-13. 2001.

[8] B. Schlkopf, "SVMs - a practical consequence of learning theory," *Journal IEEE Intelligent Systems*, vol. 13, no. 4, 18-21, 1998.

[9] T. Joachims, "Advances in Kernel Methods - Support Vector Learning," Chapter Making large-scale svm learning practical. *MIT-Press*, 1999.

[10] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, 66(2):191–234, 1994.