Multimed Tools Appl (2014) 73:663–689
DOI 10.1007/s11042-011-0936-5

An ontology-based evidential framework for video indexing using high-level multimodal fusion

Rachid Benmokhtar · Benoit Huet

Published online: 24 December 2011

Abstract This paper deals with information retrieval and semantic indexing of multimedia documents. We propose a generic scheme combining an ontology-based evidential framework and high-level multimodal fusion, aimed at recognising semantic concepts in videos. This work is represented on two stages: First, the adaptation of evidence theory to neural network, thus giving *Neural Network based on Evidence Theory (NNET)*. This theory presents two important information for decision-making compared to the probabilistic methods: belief degree and system ignorance. The NNET is then improved further by incorporating the relationship between descriptors and concepts, modeled by a weight vector based on entropy and perplexity. The combination of this vector with the classifiers outputs, gives us a new model called *Perplexity-based Evidential Neural Network (PENN)*. Secondly, an ontology-based concept is introduced via the influence representation of the relations between concepts and the ontological readjustment of the confidence values. To represent this relationship, three types of information are computed: low-level visual descriptors, concept co-occurrence and semantic similarities. The final system is called *Ontological-PENN*. A comparison between the main similarity construction methodologies are proposed. Experimental results using the TRECVid dataset are presented to support the effectiveness of our scheme.

Keywords Video shots indexing · Semantic gap · Classification · Classifier fusion · Inter-concepts similarity · Ontology · LSCOM-lite · TRECVid

R. Benmokhtar (✉) · B. Huet
Département Communications Multimédia, Eurécom, 2229, route des crêtes,
06904 Sophia-Antipolis, France
e-mail: rachid.benmokhtar@eurecom.fr

B. Huet
e-mail: benoit.huet@eurecom.fr

1 Introduction

The growing amount of image and video available either online or in one's personal collection has attracted the multimedia research community's attention. There are currently substantial efforts investigating methods to automatically organize, analyze, index and retrieve video information. This is further stressed by the availability of the MPEG-7 standard that provides a rich and common description tool for multimedia contents. Moreover, it is encouraged by TRECVid evaluation campaigns which aim at benchmarking progress in video content analysis and retrieval tools developments.

Retrieving complex semantic concepts such as CAR, ROAD, FACE or NATURAL DISASTER from images and videos requires to extract and finely analyze a set of low-level features describing the content. In order to generate a global result from the various potentially multimodal data, a fusion mechanism may take place at different levels of the classification process. Generally, it is either applied directly on extracted features (*feature fusion*), or on classifier outputs (*classifier fusion*).

In most systems concept models are constructed independently [34, 46, 55]. However, the binary classification ignores the fact that semantic concepts do not exist in isolation and are interrelated by their semantic interpretations and co-occurrence. For example, the concept CAR co-occurs with ROAD while MEETING is not likely to appear with ROAD. Therefore, multi-concept relationship can be useful to improve the individual detection accuracy taking into account the possible relationships between concepts. Several approaches have been proposed. Wu et al. [55] have reported an ontological multi-classification learning for video concept detection. Naphade et al. [34] have modeled the linkages between various semantic concepts via a Bayesian network offering a semantics ontology. Snoek et al. [46] have proposed a semantic value chain architecture for concept detection including a multi-concept learning layer called *context link*. In this paper, we propose a generic and robust scheme for video shots indexing based on ontological reasoning construction. First, each individual concept is constructed independently. Second, the confidence value of each individual concept is re-computed taking into account the influence of other related concepts.

This paper is organized as follows. Section 2 reviews existing video indexing techniques. Section 3 presents our system architecture. Section 4 gives the proposed concept ontology construction, including three types of similarities. Section 5 reports and discusses the experimentation results conducted on the TRECVid collection. Finally, Section 6 provides the conclusion of the paper.

2 Review of existing video indexing techniques

This section presents some related works from the literature in the context of semantic indexing. The field of indexing and retrieval has been particularly active, especially for content such as text, image and video. In [2, 11, 45, 50, 52], different types of visual content representation, and their application in indexing, retrieval, abstracting, are reviewed.

Early systems work on the basis of query by example, where features are extracted from the query and compared to features in the database. The candidate images are ranked according to their distance from the query. Several distance functions can be used to measure the similarity between the query and all images in the database. In Photobook [39], the user selects three modules to analyze the query: face, shape or texture. The QBIC system [13] offers the possibility to query on many features: color, texture and shape. VisualSeek [44] goes further by introducing spatial constraints on regions. The Informedia system [53] includes camera motion estimation and speech recognition. Netra-V [58] uses motion information for region segmentation. Regions are then indexed with respect to their color, position and motion in key-frames. VideoQ [9] goes further by indexing the trajectory of regions. Several papers touch upon the semantic problem. Nephade et al. [33] built a probabilistic framework for semantic video indexing to map low-level media features with high-level semantic labels. Dimitrova [11] presents the main research topics in automatic methods for high-level description and annotation. Snoek et al. [45] summarize several methods aiming at automating this time and resource consuming process as state-of art. Vembu et al. [52] describe a systematic approach to the design of multimedia ontologies based on the MPEG-7 standard and sport events ontology. Chang et al. [20] exploit the audio and visual information in generic videos by extracting atomic representations over short-term video slices.

However, models are constructed to classify video shots in semantic classes. Neither of these approaches satisfy holistic indexing, where a user wants to find high level semantic concepts such as an OFFICE or a MEETING for example. The reason is, that there is a semantic gap [52] between low-level features and high-level semantics. While it is difficult to bridge this gap for every high level concept, multimedia processing under a probabilistic framework and ontological reasoning facilitate, bridging this gap for a number of useful concepts.

3 System architecture

The general architecture of our system can be summarized in five steps as depicted in Fig. 1: (1) features extraction, (2) classification, (3) perplexity-based weighted descriptors, (4) classifier fusion and (5) ontological readjustment of the confidence values. Let us detail each of those steps:

3.1 Features extraction

Temporal video segmentation is the first step toward automatic annotation of digital video for browsing and retrieval. Its goal is to divide the video stream into a set of meaningful segments called shots. A shot is defined as an unbroken sequence of frames taken by a single camera. 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 [31]. Our system employs five types of MPEG-7 visual descriptors: Color, texture, shape, motion and face descriptors. These descriptors are briefly defined as follows:

3.1.1 Scalable Color Descriptor (SCD)

is defined as the hue-saturation-value (HSV) color space with fixed color space quantization. The Haar transform encoding is used to reduce the number of bins of the original histogram with 256 bins to 16, 32, 64, or 128 bins [17].

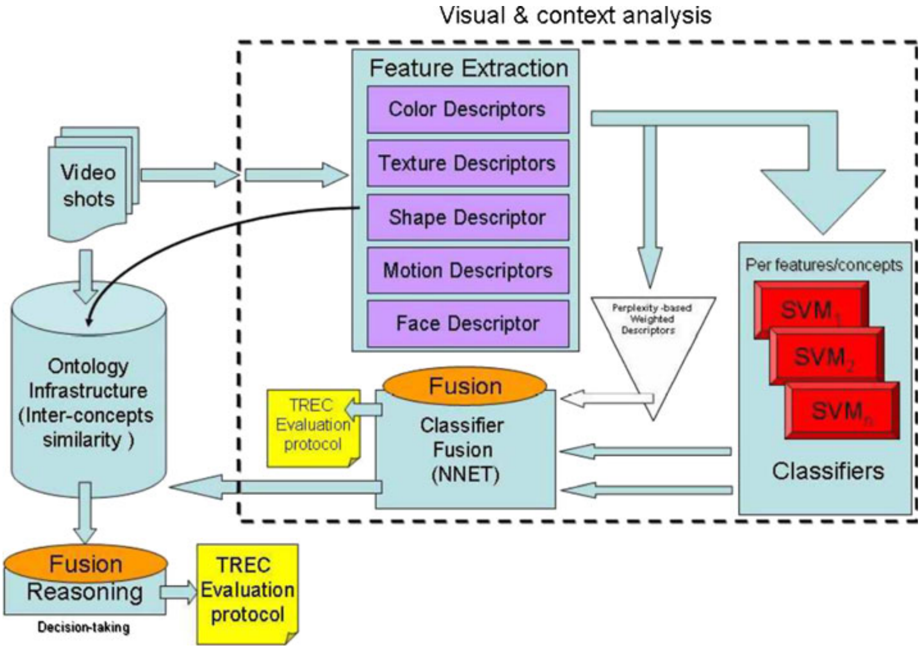


Fig. 1 General indexing system architecture

3.1.2 Color Layout Descriptor (CLD)

is a compact representation of the spatial distribution of colors [21]. The color information of an image is divided into (8×8) block. The blocks are transformed into a series of coefficient values using dominant color descriptor or average color, to obtain $CLD = \{Y, Cr, Cb\}$ components. Then, the three components are transformed by 8×8 DCT (Discrete Cosine Transform) to three sets of DCT coefficients. Finally, a few low frequency coefficients are extracted using zigzag scanning and quantized to form the CLD for a still image.

3.1.3 Color Structure Descriptor (CSD)

encodes local color structure in an image using a structuring element of (8×8) dimension. CSD is computed by visiting all locations in the image, and then summarizing the frequency of color occurrences in each structuring element location on four HMMD color space quantization possibilities: 256, 128, 64 and 32 bins histogram [32].

3.1.4 Color Moment Descriptor (CMD)

provides some information about color in a way which is not explicitly available in other color descriptors. It is obtained by the mean and the variance on each layer of the LUV color space of an image or region.

3.1.5 Edge Histogram Descriptor (EHD)

expresses only local edge distribution in the image. An edge histogram in the image space represents the frequency and the directionality of the brightness changes in the image. The EHD basically represents the distribution of 5 types of edges in each local area called a sub-image. Specifically, dividing the image into (4×4) non-overlapping sub-images. Then, for each sub-image, we generate an edge histogram. Four directional edges (0°, 45°, 90°, 135°) are detected in addition to non-directional ones. Finally, it generates a 80 dimensional vector (16 sub-images, 5 types of edges). We make use of the improvement proposed by [38] for this descriptor, which consist in adding global and semi-global levels of localization of an image.

3.1.6 Homogeneous Texture Descriptor (HTD)

characterizes a region's texture using local spatial frequency statistics. HTD is extracted by Gabor filter banks (6 frequency times, 5 orientation channels), resulting in 30 channels in total. Then, computing the energy and energy deviation for each channel to obtain 62 dimensional vector [31, 56].

3.1.7 Statistical Texture Descriptor (STD)

is based on statistical methods of co-occurrence matrix such as: energy, maximum probability, contrast, entropy, etc [1], to model the relationships between pixels within a region of some grey-level configuration in the texture; this configuration varies rapidly with distance in fine textures, slowly in coarse textures.

3.1.8 Contour-based Shape Descriptor (C-SD)

presents a closed 2D object or region contour in an image. To create *Curvature Scale Space (CSS)* description of contour shape, N equidistant points are selected on the contour, starting from an arbitrary point and following the contour clockwise. The contour is then gradually smoothed by repetitive low-pass filtering of the x and y coordinates of the selected points, until the contour becomes convex (no curvature zero-crossing points are found). The concave part of the contour is gradually flattered out as a result of smoothing. Points separating concave and convex parts of the contour and peaks (maxima of the CSS contour map) in between are then identified. Finally, eccentricity, circularity and number of CSS peaks of original and filtered contour are should be combined to form more practical descriptor [31].

3.1.9 Camera Motion Descriptor (CM)

details what kind of global motion parameters are present at what instance in time in a scene provided directly by the camera, supporting 7 camera operations: fixed, panning (horizontal rotation), tracking (horizontal transverse movement), tilting (vertical rotation), booming (vertical transverse movement), zooming (change of the focal length), dollying (translation along the optical axis), and rolling (rotation around the optical axis) [31].

3.1.10 Motion Activity Descriptor (MAD)

shows whether a scene is likely to be perceived by a viewer as being slow, fast paced, or action paced [48]. Our MAD is based on intensity of motion. The standard deviations are quantized into five activity values. A high value indicates high activity and the low value of intensity indicates low activity.

3.1.11 Face Descriptor (FD)

detects and localizes frontal faces within the keyframes of a shot and provides some face statistics (e.g, number of faces, biggest face size), using the face detection method implemented in OpenCV. It uses a type of face detector called a Haar Cascade classifier, that performs a simple operation. Given an image, the face detector examines each image location and classifies it as "face" or "not face" [37].

3.2 Classification

The classification consists in assigning classes to videos given some description of its content. The literature is vast and ever growing [24]. This section summarizes the classifier method used in the work presented here: "Support Vector Machines".

SVMs have become widely employed in classification tasks due to their generalization ability within high-dimensional pattern [51]. 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 its kernel function. In this paper, a single SVM is used for each low-level feature and is trained per concept under the *"one against all"* approach. At the evaluation stage, it returns for every shots a normalized value in the range [0, 1] using (1). This value denotes the degree of confidence, to which the corresponding shot is assigned to the concept.

$$y_i^j = 1/\left(1 + \exp\left(-\alpha d_i\right)\right) \tag{1}$$

Where (i, j) represents the ith concept and jth low-level feature, d_i is the distance between the input vector and the hyperplane and α is the slope parameter which is obtained experimentally.

3.3 Perplexity-based weighted descriptors

Each concept is best represented or described by its own set of descriptors. Intuitively, the color descriptors should be quite appropriate to detect certain concepts such as: SKY, SNOW, WATERSCAPE, and VEGETATION, while inappropriate for STUDIO, MEETING, MEETING, CAR, etc.

For this aim, we propose to weight each low-level feature according to the concept at hand, without any feature selection (Fig. 2). The variance as a simple second order vector can be used to give the knowledge of the dispersion around the mean between descriptors and concepts. Conversely, the entropy depends on more parameters and measures the quantity of informations and uncertainty in a probabilistic distribution. We propose to maps the visual features onto a term weight vector via entropy and perplexity measures. This vector is then combined with the original classifier outputs[1] to produce the final classifier outputs. As presented in Fig. 2, we shall now define the four steps of the proposed approach [6].

[1] We can also use the weight in the feature extraction step.

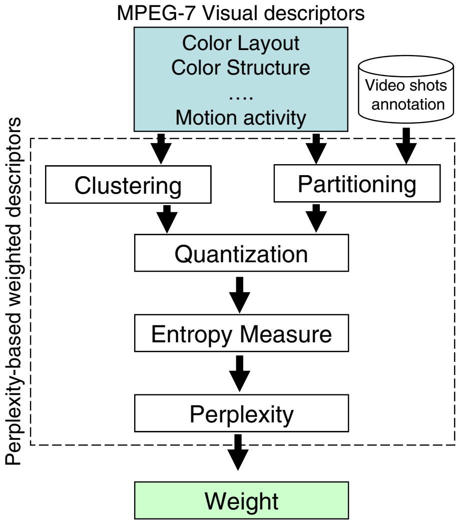


Fig. 2 Perplexity-based weighted descriptors structure

3.3.1 K-means clustering

It computes the k centers of the clusters for each descriptor, in order to create a "visual dictionary" of the shots in the training set. The selection of k is an unresolved problem, and only tests and observation of the average performances can help us to make a decision. In Souvannavong et al. [47], a comparative study of the classification results *vs* the number of clusters used for the quantization of the region descriptors of TRECVid 2005 data, shows that the performances are not deteriorated by quantization of more than 1,000 clusters. Based on this result, our system will employ $k_r = 2{,}000$ for the clustering the MPEG-7 descriptors computed from image regions, and $k_g = 100$ for the global ones. This presents a good compromise between efficiency and a low computation times.

3.3.2 Partitioning

Separating data into positive and negative sets is the first step of the model creation process. Typically, based on the annotation data provided by TRECVid, we select the positive samples for each concept.

3.3.3 Quantization

To obtain a compact video representation, we vector-quantize features. Based on the vocabulary size $k_r = 2{,}000$ (number of visual words) which has empirically shown good results for a wide range of datasets. All features are assigned to their closest vocabulary word using Euclidean distance.

3.3.4 Entropy measure

The entropy H (2) of a certain feature vector distribution $P = (P_0, P_1, ..., P_{k-1})$ gives a measure of concepts distribution uniformity over the clusters k [27]. In [22], a good model is such that the distribution is heavily concentrated on only few clusters, resulting in low entropy value.

$$H = -\sum_{i=0}^{k-1} P_i \log(P_i) \tag{2}$$

where P_i is the probability of cluster i on the quantized vector.

3.3.5 Perplexity measure

In [15], perplexity (PPL) or normalized perplexity value ($\overline{PPL}$) (3) can be interpreted as the average number of clusters needed for an optimal coding of the data.

$$\overline{PPL} = \frac{PPL}{PPL_{\max}} = \frac{2^H}{2^{H_{\max}}} \tag{3}$$

If we assume that k clusters are equally probable, we obtain $H(P) = \log(k)$, and then $1 \leq \overline{PPL} \leq k$.

3.3.6 Weight

In speech recognition, handwriting recognition, and spelling correction [15], it is generally assumed that lower perplexity/entropy correlates with better performance, or in our case, to a very concentrated distribution. So, the relative weight of the corresponding feature should be increased. Many formula can be used to represent the weight such as Sigmoid, Softmax, Gaussian, etc. In our paper, we choose Verhulst's evolution model (4). This function is non exponential, it allows a brake rate α_i to be defined, as well as reception capacity (upper asymptote) K, and β_i defines the decreasing speed of weight function.

$$w_i = K \frac{1}{1 + \beta_i \exp\left(-\alpha_i (1/\overline{\mathrm{PPL}_i})\right)} \tag{4}$$

$$\beta_i = \begin{cases} K \exp(-\alpha_i^2) & \text{if } Nb_i^+ < 2 * k \\ 1 \quad \text{Otherwise} \end{cases} \tag{5}$$

β_i is introduced to decrease the negative effect of the training set limitation, due to the low number of positive samples ($Nb_i^+ << k$) of certain concepts such as WEATHER, DESERT, MOUNTAIN,... (see Table 2). We observe a lower perplexity value, which could not be interpreted as a relevant relation between descriptor and concept. So, we increase β_i (5) to obtain a rapid weight decrease for each concept presenting less than $2 * k$ positive samples.

The relevance of the various descriptors at identifying high level concepts can be obtained through the perplexity distribution (see Fig. 3). The Boxplot provides a good visual summary of many important aspects of a distribution. The lower and upper lines express the data range, the lower and upper edges of the box indicate the 25th and 75th percentile. The line inside the box indicates the median value of the data. Figure 3 shows the normalized perplexity for each descriptor and its best concept presented by the minimum observation, such as: SCD is more effective to detect the concept SKY "13", EDH for ROAD "12", etc. The first observation concerns the same value of median perplexity obtained for SCD, CLD, CMD, CSD, where

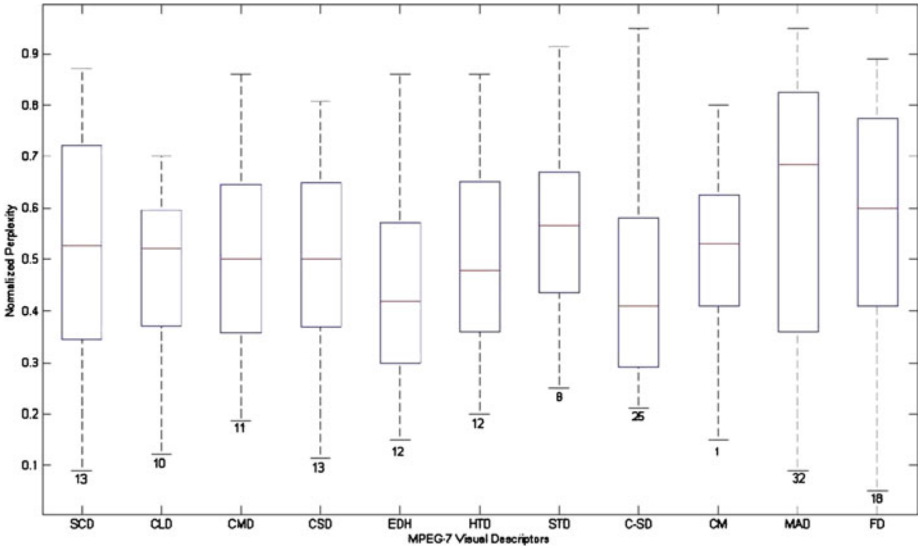


Fig. 3 Normalized perplexity Boxplot

color is more discriminant. Secondly, C-SD gives the smallest 25th percentile of normalized perplexity for all data, followed by EDH and SCD. Thirdly, it seems that EHD is very useful in the detection of the contour as in the SPORT and ROAD concepts. Identical observation is given for C-SD. Conversely, MAD presents a large interval of perplexity but gives small value for the concepts WALKING-RUNNING, PEOPLE-MARCHING where the motion activity can be detected. Finally, FD is a relevant descriptor to detect FACE and PERSON concepts which was to be expected from the very nature of this descriptor.

This approach is proposed to weight each low-level feature per concept, within an adaptive classifier fusion step (Section 3.4). The combination provides a new classification system that we call PENN “Perplexity-based Evidential Neural Network”. We will now present the classifier fusion step.

3.4 Classifier fusion

Classifier fusion is an important step of the classification task. It improves recognition reliability by taking into account the complementarities between classifiers, in particular for multimedia indexing and retrieval. Several schemes have been proposed in the literature according to the type of information provided by each classifier as well as their training and adaptation capacity. The state of the art and the comparison study about the effectiveness of the classifier fusion methods are given in [4].

In [12], Duin et al. have distinguished the combination methods of different classifiers and the combination methods of weak classifiers. Another kind of grouping using only the type of classifiers outputs (class, measure) is proposed in [57]. Jain [18] built a dichotomy according to two criteria of equal importance: the type of classifiers outputs and their capacity of learning. This last criteria is used by [25, 26] for grouping

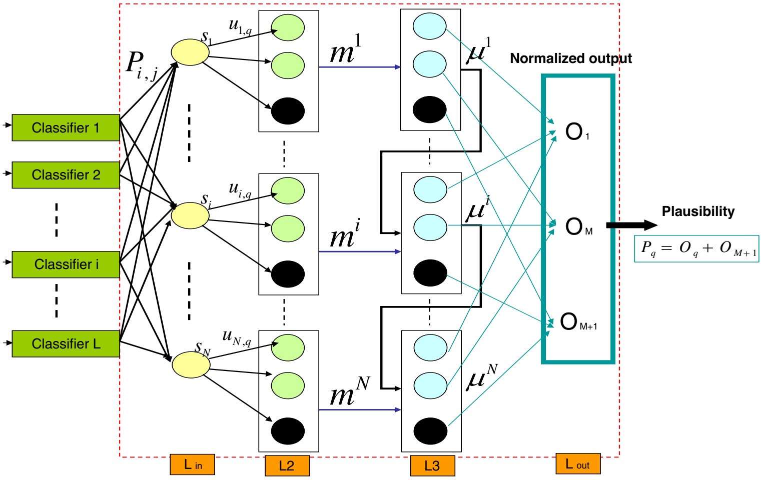


Fig. 4 NNET classifier fusion structure

the combination methods. The trainable combiners search and adapt the parameters in the combination. The non trainable combiners use the classifiers outputs without integrating another *a priori* information of each classifiers performances.

In this part, we describe our proposed neural network based on evidence theory (NNET) [5] to address classifier fusion (Fig. 4).

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

$$\begin{cases} s^i = \alpha^i \phi(d^i) \\ \phi(d^i) = \exp\left(-\gamma^i (d^i)^2\right) \end{cases} \tag{6}$$

2. **Layer** L_2: Computes the belief masses m^i (7) associated to each unit. The units of module i are connected to neuron i of the previous layer.

$$\begin{cases} m^i(\{w_q\}) = \alpha^i u_q^i \phi(d^i) \\ m^i(\Omega) = 1 - \alpha^i \phi(d^i) \end{cases} \tag{7}$$

   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 (8):

$$m(A) = (m^1 \oplus ... \oplus m^N) = \sum_{B_1 \bigcap ... \bigcap B_N = A} \prod_{i=1}^{N} m^i(B_i) \tag{8}$$

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

$$\begin{cases} \mu^1 = m^1 \\ \mu^i_j = \mu^{i-1}_j m^i_j + \mu^{i-1}_j m^i_{M+1} + \mu^{i-1}_{M+1} m^i_j \\ \mu^i_{M+1} = \mu^{i-1}_{M+1} m^i_{M+1} \end{cases} \tag{9}$$

4. **Layer** L_{output}: In [10], the output is directly obtained by $O_j = \mu^N_j$. 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 (10). 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.

$$O_j = \frac{\sum_{i=1}^{N} \mu^i_j}{\sum_{i=1}^{N} \sum_{j=1}^{M+1} \mu^i_j} \tag{10}$$

$$P_q = O_q + O_{M+1} \tag{11}$$

The different parameters (Δu, $\Delta\gamma$, $\Delta\alpha$, ΔP, Δs) 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.

Therefore, the combination between perplexity-based weighted low-level feature per concept, within the adaptive NNET classifier fusion provides a novel system that we call PENN “Perplexity-based Evidential Neural Network”.

4 Concept ontology construction

The ontology has been historically used to achieve better performance in the multimedia retrieval system [8]. It defines a set of representative concepts and the inter-relationships among them. It is therefore important to introduce some constraints to the development of the similarity measures before proceeding to the presentation of our method. Psychology demonstrates that similarity depends on the context, and may be asymmetric [30]. However, when ontologies have been defined for multimedia they have not been extensively used at the decision making stage of high level concept detection.

Most indexing models are based on binary classification, ignoring possible relationships between concepts. However, concepts do not exist in isolation and are interrelated by both their semantic interpretations and co-occurrence. Wu et al. [55] have reported an ontological multi-classification learning for video concept detection in the NIST TREC-2003 Video Retrieval Benchmark.[2] Ontology-based multi-classification learning consists of two steps. At the first step, each single concept model is constructed independently based on SVM (Support Vector Machine). At

[2] NIST TREC-2003 Video Retrieval Benchmark defines 133 video concepts, organized hierarchically and each video data belong to one or more concepts [35].

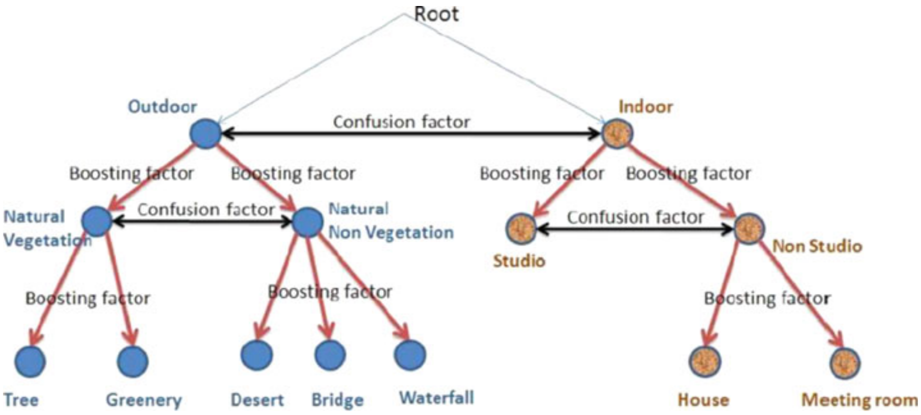


Fig. 5 An example of ontology used by Wu et al. [55]

the second step, ontology-based concept learning improves the accuracy of individual classifiers based on updating confidence scores from single concept models. Two kinds of influences have been defined: *confusion factor* β and *boosting factor* λ. The *confusion factor* β is the influence between concepts that can not be co-existent. The *boosting factor* λ is the top-to-down influence from big to small concepts in the ontology hierarchy. A small example of such influence is presented in Fig. 5. The factors are obtained using a correlation study of the training data. Then, an update of the novel confidence is applied, as shown in (12) and (13).

$$\begin{cases} \underline{p}(x/C_i) = p(x/C_i) + \sum_{j\in\psi} \lambda_j^i p(x/C_j) \\ \lambda_j^i = \dfrac{A}{B + \exp\left(C|p(s/C_i) - p(s/C_j)|\right)} \end{cases} \tag{12}$$

$$\begin{cases} \underline{p}(x/C_i) = \dfrac{p(x/C_i)}{\beta} \\ \beta = \dfrac{1}{f\left(p(x/C_i) - \max_{j\in\theta}(p(x/C_j))\right)} \end{cases} \tag{13}$$

The parameters A, B and C of (12) are empirically obtained as described in the works of Li et al. [28]. $f(.)$ is a positive and increasing function for the (13).

Naphade et al. [34] have modeled the linkages between various semantic concepts via a Bayesian network offering a semantics ontology. The central theme to this approach is the concept of Multijects or Multimedia Objects. A Multiject has a semantic label and summarizes a time sequence of low level features of multiple modalities in the form of a probability. It has 3 main aspects: The first aspect is the semantic label. The second aspect of a Multiject is, that it summarizes a time sequence of low level features. The detection of a certain Multiject can increase or decrease the probability of occurrence of other Multiject. For example, if the Multiject BEACH is detected with a very high probability, then the probability of occurrence of the Multiject YACHT or the Multiject SUNSET increases. This is the third aspect of Multijects, i.e. their interaction in a network.

Authors assume that all concepts have the same semantic level, related by the conditional dependence relation with the associated low-level descriptors.

Fan et al. [14] have proposed a hierarchical classification for image annotation.[3] This approach introduce the contextual dependences of the WordNet ontology and the co-occurrence relationship, as presented by the following equation:

$$\begin{cases} \lambda(C_m, C_n) = \rho(C_m, C_n)\pi(C_m, C_n) \\ \text{ou} \quad \rho(C_m, C_n) = log\left(\dfrac{P(C_m, C_n)}{P(C_m)P(C_m)}\right) \\ \text{et} \quad \pi(C_m, C_n) = -log\left(\dfrac{dist(C_m, C_n)}{2D}\right) \end{cases} \tag{14}$$

with $\rho(C_m, C_n)$ is the joint probability between two concepts. It is obtained by the computation of the frequency for the co-occurrence of the relevant C_m and C_n. $\pi(C_m, C_n)$ is the contextual dependency, extracted in the ontology structure (*dist* is the length of the shortest path between two concepts, and D is the maximum depth of the WordNet).

Hauptmann et al. [16] have presented a comparison between the unimodal and the multimodal indexing. The multimodal system learn the dependence between concepts using the following graphical models: *Conditional Random Field "CRF"* and *Bayesian network*. The two models provide closer results in term of precision but are better than the unimodal approach. Koskela et al. [22] have exploited the correlations between the concepts to build a clustering method.

In another development, Li et al. [29] have proposed a study of various linear and non-linear functions $S = f(f_1, f_2, f_3)$ depending on the shortest path length l, depth of subsumer concept in the hierarchy h, and the local semantic density d, as shown in (15).

$$\begin{cases} f_1 = \exp(-\alpha l) \\ f_2 = \dfrac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \\ f_3 = \dfrac{e^{\delta d} - e^{-\delta d}}{e^{\delta d} + e^{-\delta d}} \end{cases} \tag{15}$$

where α is a constant, $\beta > 0$ is a smoothing factor. $\delta = max_{c \in CS(c_m, c_n)}(-\log p(c))$ represents the semantic similarity measured by the information content.

Several combinations have been applied and evaluated such as: $S_1 = f_1$, $S_2 = f_1 f_2$, $S_3 = S_2 f_3$, $S_4 = S_2 + f_3$, etc. The obtained results with different parameters (α and β) indicate that different functions have satisfactory performances, particularly those that use the three influences.

Discussion The work of Wu et al. [55] uses a confidence update using the correlation of data, and a fixed ontology structure. Naphade et al. [34] have trained the low-level features, and the co-occurrence between concepts. Koskela et al. [22] have included the co-occurrence and visual information in the construction of

[3] Three image datasets are used: *Corel Images, Google Images, and LabelMe*.

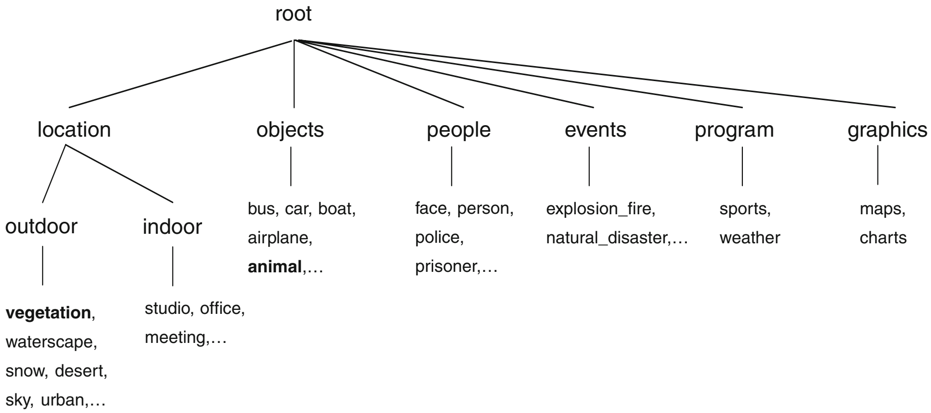

root
location
objects
people
events
program
graphics
outdoor
indoor
bus, car, boat, airplane, **animal**,...
face, person, police, prisoner,...
explosion_fire, natural_disaster,...
sports, weather
maps, charts
vegetation, waterscape, snow, desert, sky, urban,...
studio, office, meeting,...

Fig. 6 Fragment of the hierarchical LSCOM-Lite

this relationship. Fan et al. [14] as in Li et al. [29] have incorporated contextual dependencies of the WordNet ontology, and co-occurrences. This paper extends preceding works in term of the inter-concepts similarity construction. We use the co-occurrence in the corpus, the visual information outcome from low-level description, and finally the hybrid semantic similarity obtained from the ontology architecture.

In LSCOM-lite ontology[4] [35], we notice positive relationships such as (ROAD, CAR), (VEGETATION, MOUNTAIN), and negative relationships like (BUILDING, SPORTS), (SKY, MEETING).

Here, we will investigate how the relationship between different semantic concepts can be extracted and used. One direct method for similarity calculation is to find the minimum path length of connecting two concepts [40]. For example, Fig. 6 illustrates a fragment of the semantic hierarchy of LSCOM-Lite. The shortest path between VEGETATION and ANIMAL is VEGETATION-OUTDOOR-LOCATION-ROOT-OBJECTS-ANIMAL. The minimum length of a path is 5. Or, the minimum path length between VEGETATION and OUTDOOR is 1. Thus, we could say in LSCOM-lite ontology, OUTDOOR is more similar semantically to VEGETATION than ANIMAL. But, we should not say ANIMAL is more similar to CAR. In an other way, OUTDOOR contains many different concepts such as "DESERT, URBAN, ROAD,etc" each with different colors and textures scene descriptions. Therefore, the linking of concepts can infer new and more complex concepts, or improve the recognition of concepts previously detected. Thus, the presence or absence of certain concepts suggests a high or low capability to find other concepts (e.g. detection of SKY and SEA increases the probability of the concept BEACH and reduces the likelihood of DESERT). For this, more information between the concepts are introduced, so that it becomes a function of attributes

[4] The LSCOM-lite (Large-Scale Concept Ontology for Multimedia) [36] annotations include 39 concepts, which are interim results from the effort in developing a LSCOM. The dimensions consist of program category, setting/scene/site, people, object, activity, event, and graphics. A collaborative effort among participants in the TRECVid benchmark was completed to produce the annotations. Human subjects judge the presence or absence of each concept in the video shots.

“co-occurrence, low-level visual descriptors, path length, depth and local density” to boost the performance of specific indexing system, as:

$$\lambda(C_m, C_n) = \sum_{i=\{\text{Cos, Vis, Sem}\}} (Sim_i(C_m, C_n)) \tag{16}$$

Below, we explain with more details the similarity forms used in our architecture.

4.1 Co-occurrence

The first similarity is obtained by considering the co-occurrence statistics between concepts, where the presence or absence of certain concepts may predict the presence of other concepts. Intuitively, documents (video shots) that are “close together” in the vector space relate to similar things. Many methods are proposed in literature to represent this proximity such as: Euclidean, Hamming, Dice, etc. Here, we use Cosine similarity because it reflects similarity in terms of relative distributions of component. Cosine is not influenced by one document being small compared to others like the Euclidean distance tends to be [23]:

$$Sim_{\cos}(P^m, P^n) = \frac{\sum_{i=0}^{k-1} P_i^m P_i^n}{\sqrt{\sum_{i=0}^{k-1} (P_i^m)^2 \sum_{i=0}^{k-1} (P_i^n)^2}} \tag{17}$$

4.2 Visual similarity

The second similarity is based upon low level visual features. In Section 3.3, we have used perplexity to build a weighted descriptor per concept. Now, in order to compute the visual similarity d_{vis}, we are interested in mutual information presented as a measure of divergence. To this end, several measures are proposed in the literature: *Jensen–Shannon (JS), Kullback–Leibler (KL)*, etc. We decided to use d_{JD} *Jeffrey divergence* [23] which is like d_{KL}, but is numerically more stable.

$$d_{JD}(P^m, P^n) = \sum_{i=0}^{k-1} \left(P_i^m \log \frac{P_i^m}{\hat{P}_i} + P_i^n \log \frac{P_i^n}{\hat{P}_i} \right) \tag{18}$$

where $\hat{P}_i = \frac{P^m + P^n}{2}$ is the mean distribution. The visual distance between two concepts is:

$$Sim_{\text{vis}}(C_m, C_n) = \frac{1}{\sum_{i=1}^{Nb\ \ features} \frac{1}{2}(w_i^m + w_i^n) d_{JD}^i (P^m, P^n)} \tag{19}$$

where w_i^m is the ith perplexity-based weighted descriptors for the concept m.

4.3 Semantic similarity

The semantic similarity between the concepts has been widely studied in the literature and can be classified in three major approaches [43]:

4.3.1 Distance-based approach

It estimates the distance (edge length) between nodes which correspond to the concepts being compared. Two concepts C_m and C_n are similar if their path is short,

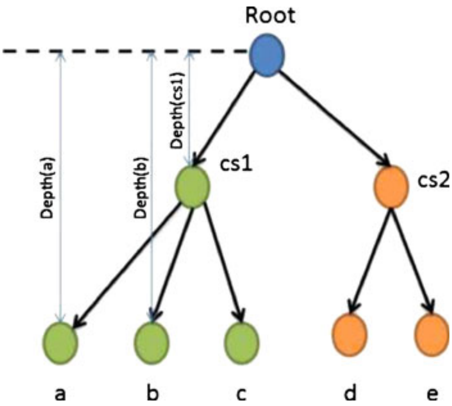


Fig. 7 The concept similarity measure

presented by the minimum number of edges that separates the two concepts. Rada et al. [40] propose the following equation:

$$Sim_{sem}(C_m, C_n) = 1/\left(1 + dist_{Rada}(C_m, C_n)\right) \tag{20}$$

Wu and Palmer [54] propose a similarity-based (see Fig. 7) on the depth of the concept subsumes CS[5] and the two concepts (21).

$$Sim_{sem}(C_m, C_n) = \frac{2 * depth(CS)}{depth(C_m) + depth(C_n)} \tag{21}$$

The drawbacks of this approach are its dependence on the concepts position in the hierarchy, and that all edges have the same weight, which imposes difficulties in defining and controlling the distance edges.

4.3.2 Information content-based approach

It takes into account the information shared by the concepts in terms of entropy measure. Two methods exist. The first uses a learning corpus and compute the probability $p(C_i)$ to find the concept C_i or one of its descendants. For Resnik [41], the semantic similarity can be obtained per the frequency of appearance in the corpus, and defined by:

$$Sim_{sem}(C_m, C_n) = \max\left(IC(CS(C_m, C_n))\right) \tag{22}$$

with $IC(C_i) = -\log(p(C_i))$ is the information content of the concept C_i (i.e, the entropy of a class C_i). The probability $p(C_i)$ is computed by dividing the number of instances of C_i by the total number in the corpus. This measure does not seem complete and precise because it depends on the specific subsumed concept only.

The second method computes the information content of nodes from WordNet instead of a corpus. Seco et al. [42] use descendant hyponyms of the concepts to obtain the information content. This approach can produce a similarity between two neighbor concepts of an ontology, exceeding the value of two concepts contained in the same hierarchy. This is inadequate in the context of information retrieval.

[5] The concept subsumes is the most common specific concept.

4.3.3 Hybrid approach

The hybrid approach combines the two previous approaches. Often, it reuses the information content of nodes and the smallest common ancestor, as with the equation of Lin et al. [30], or with the distance of Jiang and Conrath $dist_{J\&C}$ [19].

$$Sim_{sem_{Lin}}(C_m, C_n) = \frac{2 * \log P(CS)}{\log P(C_m) + \log P(C_n)} \tag{23}$$

$$\begin{cases} dist_{J\&C}(C_m, C_n) = IC(C_m) + IC(C_n) - 2 * IC(CS(C_m, C_n)) \\ Sim_{sem_{J\&C}}(C_m, C_n) = 1/\left(dist_{J\&C}(C_m, C_n)\right) \end{cases} \tag{24}$$

For the ontology presented in the Fig. 8, we compare the last two hybrid approaches with the novel one as presented in the (26), that it is the combination of Rada [40] and J&C [19].

$$\begin{cases} Sim_{sem_{J\&C}}(C_m, C_n) = 1/d_{J\&C}(C_m, C_n) \\ d_{J\&C}(C_m, C_n) = IC(C_m) + IC(C_n) - 2 * IC(CS(C_m, C_n)) \end{cases} \tag{25}$$

$$Sim_{sem}(C_m, C_n) = 1/\left(d_{Rada}(C_m, C_n) + d_{J\&C}(C_m, C_n)\right) \tag{26}$$

where $d_{Rada}(C_m, C_n)$ is the length of the shortest path between C_m and C_n.

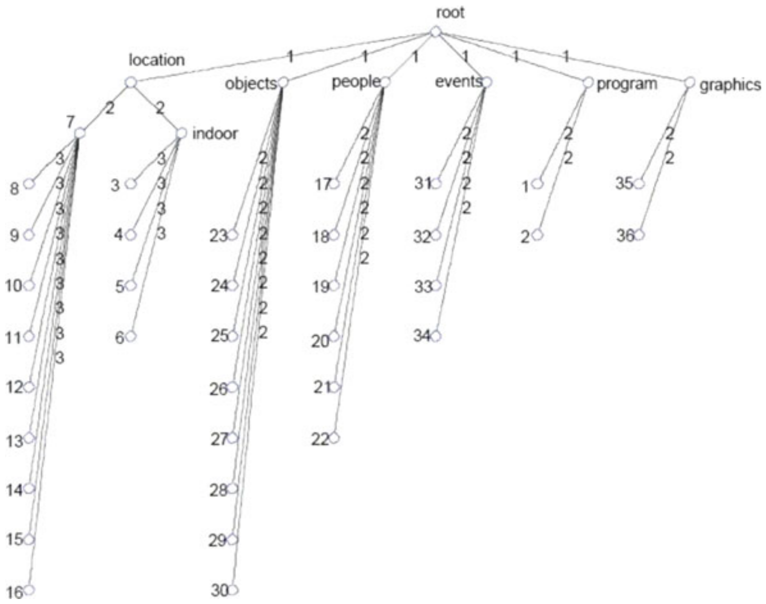


Fig. 8 Hierarchical ontology model

4.4 Concept-based confidence value readjustment (CCVR)

The proposed framework (Fig. 1) introduces a *reranking* or confidence value readjustment to refine the PENN results for concept detection [7], and is computed using:

$$\underline{P(x/Ci)} = P(x/Ci) + \frac{1}{Z}\sum_{j=1}^{Nb\ arc} \lambda_{i,j}(1-\zeta_j)P(x/Cj) \tag{27}$$

where $\underline{P(x/Ci)}$ corresponds to the multi-modal PENN result, $\lambda_{i,j}$ is the causal relationship between concepts C_i and C_j, ζ_j is the classifier error in the validation set and Z is a normalization term.

5 Experimentations

The experiments provided here are conducted on the TRECVid 2007 dataset [49] containing science news, news reports, documentaries, etc. Of the 100 hours of video segmented into shots and annotated [3] with semantic concepts from the 36 defined labels. Half is used to train the feature extraction system and the other half is used for evaluation purposes. The evaluation is realized in the context of TRECVid using mean average precision MAP in order to provide a direct comparison of the effectiveness of the proposed approach with other published work using the same dataset. Precision provides a measure of the ability of a system to present only relevant sequence.

$$\text{AP} = \frac{\left(\frac{\text{number of relevant video sequences retrieved}}{\text{total number of video sequences retrieved}}\right)}{\text{total number of relevant video sequences}} \tag{28}$$

Other metrics are introduced in our evaluation to have a global comparison: F-measure, classification rate CR, and balanced error rate BER.[6] The classifier results can be represented in a confusion matrix (Table 1), where a, b, c and d represent the number of examples falling into each possible outcome:

$$\text{F-measure} = 2\frac{P.R}{P+R} \tag{29}$$

$$\text{BER} = \frac{1}{2}(\frac{b}{a+b} + \frac{c}{c+d}) \tag{30}$$

Figure 9 shows the variation of average precision results *vs* semantic concepts, for three systems: NNET,[7] PENN,[8] and Onto-PENN.[9] First, we observe that PENN

[6] The balanced error rate is the average of the errors on each class. BER is used in “Performance Prediction Challenge Workshop”.

[7] NNET: Neural Network based on Evidence Theory.

[8] PENN: Perplexity-based Evidential Neural Network.

[9] Onto-PENN: Ontological readjustment of the PENN. The results presented in the rest of paper for the Onto-PENN, are given by (26) for the semantic similarity computation.

Table 1 Confusion matrix representation

		Prediction	
		Class 0	Class 1
Real	Class 0	a	b
Class	Class 1	c	d

and Onto-PENN systems have the same performance on average for several concepts, and present a significant improvement compared to NNET for the concepts 4,6,17,18,19,23,31 and 32. This is not surprising considering the manner the MAP (Mean Average Precision) is computed (using only the first 2,000 returned shots as in TRECVid) (see Table 2). Furthermore, low performances on several concepts can be observed due to both numerous conflicting classification and limited training data regardless of the fusion system employed. This also explains the rather low retrieval accuracy obtained for concepts 3, 22, 25, 26, 33 and 34.

To evaluate the inter-concepts similarity contribution in the video shots indexing system, we need to study the results in all test set. For this, the comparisons of the detection performances are carried out by thresholding the soft-decisions at the shot-level before and after using the inter-concepts similarity via F-meas, CR$^+$ and BER. Note that the MAP is not sensitive to *Threshold* values τ. Figure 10 compares the three experimental systems along with the variation of $\tau \in [0.1, 0.9]$, by step of 0.1. We can clearly see that for any τ value the Onto-PENN dominates and obtains higher performances for F-meas, CR$^+$ as well as lower BER comparing to PENN and NNET. The BER$_{min}$ = 40.38% is given by $\tau = 0.2$, for F-meas= 16.98% and CR$^+$ = 34.48%. The best results are obtained for $\tau \in [0.2, 0.5]$. With $\tau = 0.40$, the

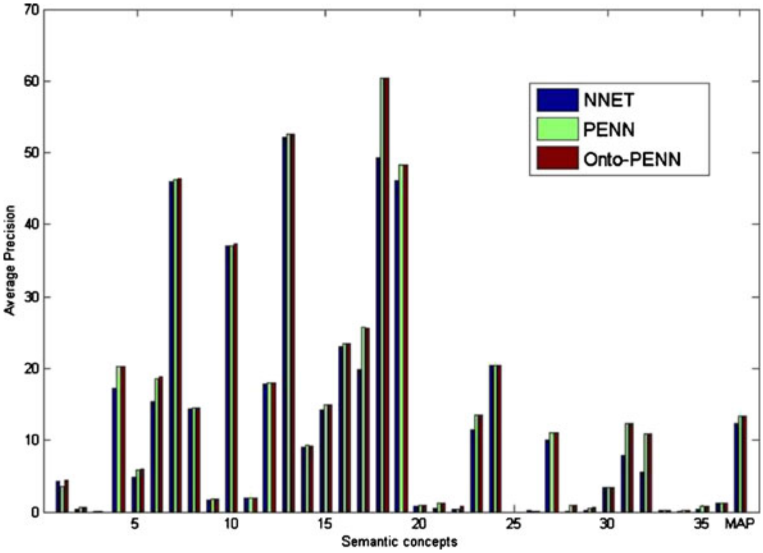


Fig. 9 Average precision evaluation

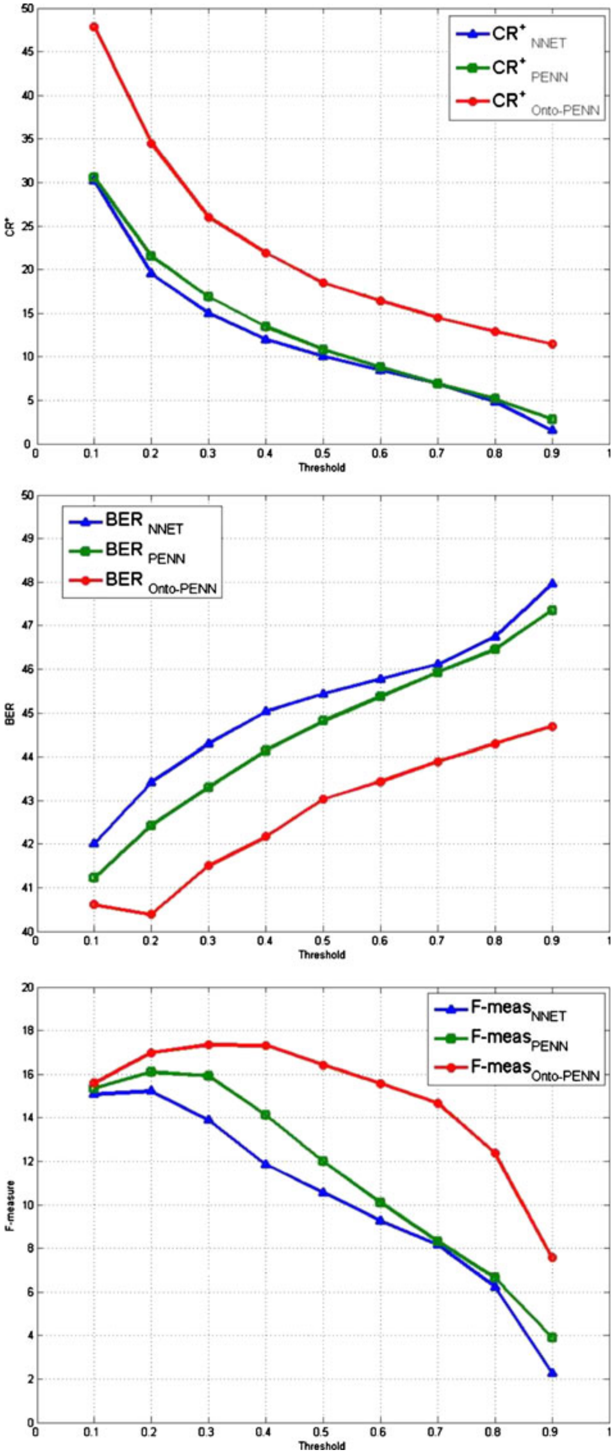
Table 2 Id of the TRECVid 2007 concepts

Id	Concepts	Neg.train	Pos.train	Pos.test
1	Sports	11,974	106	42
2	Weather	12,029	51	34
3	Court	11,967	113	5
4	Office	11,159	921	453
5	Meeting	11,532	548	270
6	Studio	11,722	358	468
7	Outdoor	8,643	3,437	1,812
8	Building	10,964	1,116	477
9	Desert	12,019	61	15
10	Vegetation	10,615	1,465	499
11	Mountain	12,004	76	17
12	Road	11,420	660	297
13	Sky	10,777	1,303	853
14	Snow	12,044	36	91
15	Urban	10,746	1,334	537
16	Waterscape	11,725	355	414
17	Crowd	11,159	921	552
18	Face	6,596	5,484	2,325
19	Person	4,981	7,099	2,972
20	Pol. security	11,824	256	63
21	Military	11,848	232	74
22	Prisoner	12,067	13	7
23	Animal	11,675	405	271
24	Computer Tv	11,617	463	202
25	US Flag	12,070	10	0
26	Airplane	12,052	28	7
27	Car	11,663	417	187
28	Bus	12,033	47	40
29	Truck	11,985	95	19
30	Boat/ship	11,979	101	151
31	Walk. running	11,221	859	385
32	Peop. marching	11,960	120	82
33	Exp. fire	11,068	12	19
34	Nat. disaster	12,061	19	21
35	Maps	12,030	50	31
36	Charts	11,954	126	80

CR^+ is improved by 10.14% to achieve 22.07%, and decreasing the BER of 2.91% compared to NNET.

Figure 11 presents the performance evolution per concepts using $\tau = 0.4$. Some points can be noticed: The three systems produce a certain non-detection (F-meas $=$ 0, $CR^+ = 0$) for the concepts 2, 3, 9, 11, 25, 26, 28, 29, 33, 34, and 36. Then, NNET can not detect any of the following concepts 1, 5, 6, 20, 21, 22, 31, 32, and 35. Identically, for PENN in 5,20,22, and 35. Finally, Onto-PENN resolves the limitation previously mentioned and achieves a high improvement for the concepts 1, 4, 7, 8, 10, 12, 13, 15, 16, 17, 18, 19, 22, 23, 24, and 31, due to the strong relationship between the connected concepts, allowing for better, more accurate decision-making.

Fig. 10 Evaluation of the metrics (CR$^+$, BER and F-measure) vs *Threshold* $\tau \in [0.1, 0.9]$



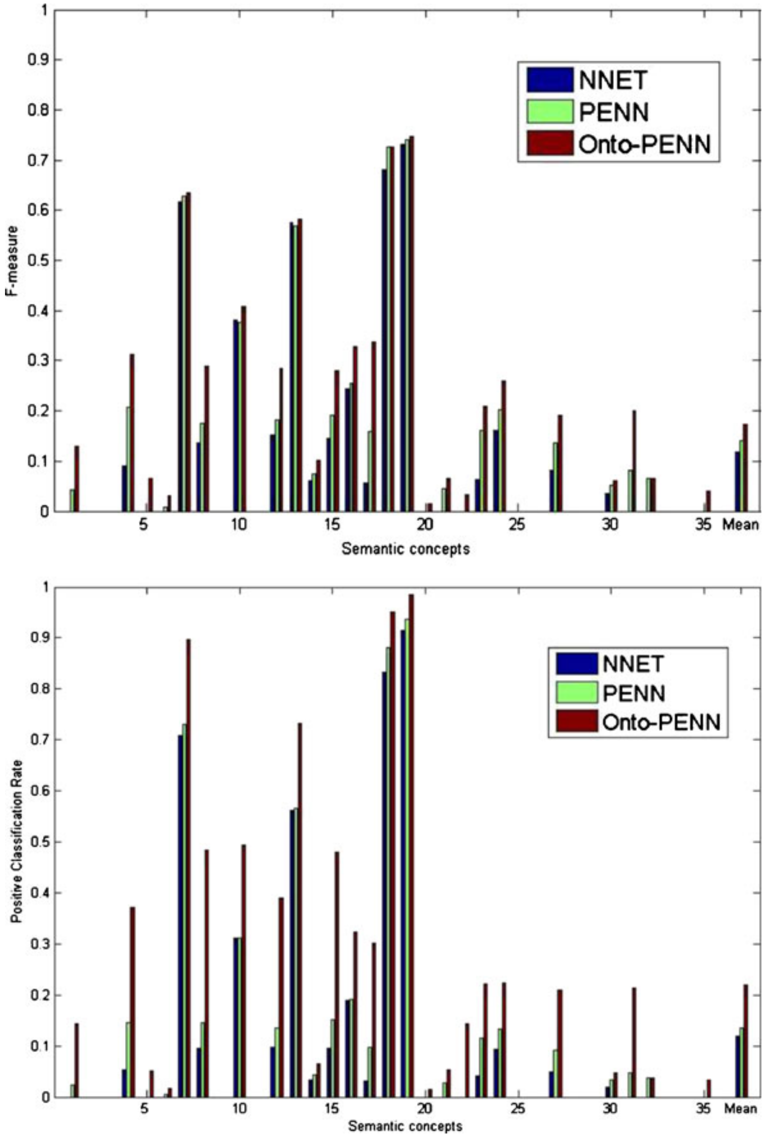


Fig. 11 F-measure and CR$^+$ evaluation

As an example, to detect FACE, PERSON, MEETING, or STUDIO concepts, PENN gives more importance to *FaceDetector, ContourShape, ColorLayout, ScalableColor, EdgeHistogram* than others descriptors. For the "PERSON" concept, the improvement was as high as 11%, making it the best performing run. The Onto-PENN system introduces the relationship between the connected concepts (i.e. concepts that are likely to co-occur in video shots), increasing the performance in term of accuracy (see

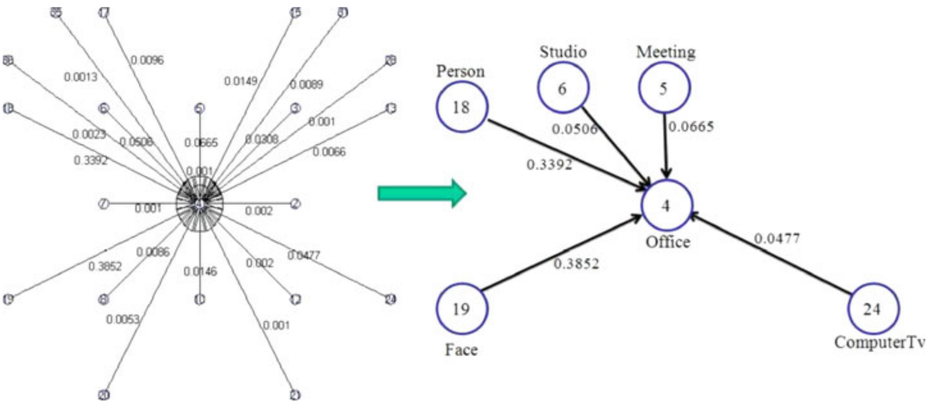


Fig. 12 Inter-concept connections graphical model for the concept OFFICE. We observe that 20 concepts are connected with OFFICE, but only 5 are strong and significant (MEETING: 6.65%, STUDIO: 5.06%, FACE: 33.92%, PERSON: 38.52%, and COMPUTERTV: 4.77%) presenting 88.92% of the global information

Fig. 12). The co-occurring concept constitute some type of contextual information about the content of the shot under consideration.

Table 3 summarizes the overall performances for the content-based video shots classification systems using a fixed *Threshold*($\tau = 0.4$). We compute the above mentioned statistics for all concepts, and for a subset composed of the 10 most frequent concepts in the dataset. All hybrid semantic similarities-based Onto-PENN allow an overall improvement of the system and a significant increase of F-meas and CR^{+}. They achieve a respectable result for MAP, and significantly decrease the balanced error rate "BER" compared to NNET and PENN. Finally, the results given by the two equations (25 and 26) are very close, with a slight advantage for the (26). However, it can be observed that the MAP declines using the equations of Rada,

Table 3 Performance comparisons between the three experimental systems: NNET, PENN and Onto-PENN

Methods/	NNET	PENN	Onto-PENN			
eval. (%)			Rada	Lin	J&C	B&H
MAP	12.70	13.29	12.94	13.01	13.31	13.37
MAP@10	33.70	35.30	34.12	34.91	35.30	35.36
F-meas	11.84	14.10	15.97	16.17	17.07	17.30
F-meas@10	38.75	40.79	41.83	43.41	44.67	44.74
CR^{+}	11.93	13.43	18.12	20.58	21.76	22.07
CR^{+}@10	40.69	41.74	53.76	57.80	59.45	59.71
BER	45.02	44.13	43.93	43.62	42.32	42.11
BER@10	38	36.52	36.02	35.45	34.03	33.96

We present in term of accuracy the effect of each similarity method (Rada (20), Lin (23), J&C (25)), and our proposed method B&H (26) in the Onto-PENN system, for $\tau = 0.4$

and Lin compared to the two equations used, which underlines the importance of the semantic similarity choice.

6 Conclusions

In this paper, we have presented a generic and robust ontology-based video shots indexing scheme. One of the particular aspect of the proposed framework is to employ contextual information during the classification phase. To learn the influence of the relation between concepts, three types of influence are computed: co-occurrence, visual descriptors and hybrid semantic similarity. A comparison of some approaches to automatically construct the semantic similarity has been presented. Based on the newly defined simulated user principle, we evaluate the results of four alternative methodologies. We demonstrate through statistical study and empirical testing the potential of multimodal fusion, to be exploited in video shots retrieval. In TRECVid 2007 benchmark, a significant improvement is obtained with our system, about 18.75% in terms of correct positive recognition rate (CR^+), 5.99% for the F-measure, 1.66% for the mean average precision (MAP), and decreases the balanced error rate of 2.91% on average. Our proposed "Onto-PENN" method outperforms clearly both the NNET and PENN methods which are not using any contextual information. In addition, we have shown that perplexity-based weighted vector integration in the indexing papeline increases the performances of our system.

In the future works, we plan to extend application to WordNet instead of a corpus, integration of richer semantics and broader knowledge.

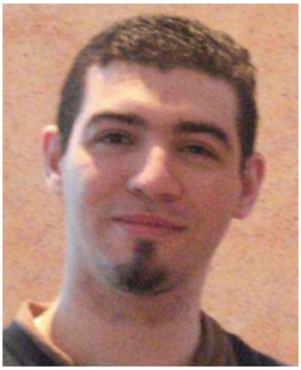
Acknowledgements This research was supported by Eurecom's industrial members: Ascom, Bouygues Telecom, Cegetel, France Telecom, Hitachi, ST Microelectronics, Motorola, Swisscom, Texas Instruments and Thales.

References

1. Adamek T (2007) Extension of MPEG-7 low-level visual descriptors for TRECVid07. Kspace Technical Report, FP6-027026
2. Aigrain P, Joly P (1994) The automatic real-time analysis of film editing and transition effects and its applications. Comput Graph 18(1):93–103
3. Ayache S, Quènot G (2007) TRECVid 2007 collaborative annotation using active learning. In: TRECVid, 11th international workshop on video retrieval evaluation, Gaithersburg, USA
4. Benmokhtar R, Huet B (2006) Classifier fusion: combination methods for semantic indexing in video content. In: International conference on artificial neural networks, pp 65–74
5. Benmokhtar R, Huet B (2007) Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content. In: International multimedia modeling conference, pp 196–205
6. Benmokhtar R, Huet B (2008) Perplexity-based evidential neural network classifier fusion using MPEG-7 low-level visual features. In: ACM international conference on multimedia information retrieval, pp 336–341
7. Benmokhtar R, Huet B (2009) Hierarchical ontology-based robust video shots indexing using global MPEG-7 visual descriptors. In: Proceedings of the international workshop on content-based multimedia indexing, pp 195–200
8. Berners T, Hendler J, Lassila O (2001) The semantic web. Scientific American, pp 29–37

9. Chang S-F, Chen W, Meng H, Sundaram H, Zhong D (1998) A fully automated content-based video search engine supporting spatiotemporal queries. In: IEEE transactions circuits and systems for video technology, pp 602–615
10. Denoeux T (1995) An evidence-theoretic neural network classifer. In: International conference on systems, man and cybernetics, vol 31, pp 712–717
11. Dimitrova N (2003) Multimedia content analysis: the next wave. In: International conference on image and video retrieval. Lecture notes in computer science, vol 25, pp 8–17
12. Duin R, Tax D (2000) Experiements with classifier combining rules. In: Proc. first int. workshop MCS 2000, vol 1857, pp 16–29
13. Faloutsos C, Barber R, Flickner M, Hafner J, Niblack W, Petkovic D, Equitz W (1994) Efficient and effective querying by image content. JIIS 3(3), 231–262
14. Fan J, Gao Y, Luo H (2007) Hierarchical classification for automatic image annotation. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, pp 111–118
15. Gao J, Goodman J, Li M, Lee K (2001) Toward a unified approach to statistical language modeling for chinese. In: ACM transactions on Asian language information processing
16. Hauptmann A, Christel M, Concescu R, Gao J, Jin Q, Lin W, Pan J, Stevens S, Yan R, Yang J, Zhang Y (2005) CMU informedia's TRECVid 2005 skirmishes. In: TREC video retrieval evaluation online proceedings
17. ISO/IEC 14496-2. Information Technology (2001) Coding of moving pictures and associated audio information.
18. Jain A, Duin R, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 20(1), 4–37
19. Jiang J, Conrath D-W (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: International conference research on computational linguistics
20. Jiang W, Cotton C, Chang S-F, Ellis D, Loui A (2009) Short-term audio-visual atoms for generic video concept classification. In: MM '09: proceedings of the seventeen ACM international conference on multimedia, pp 5–14
21. Kasutani E, Yamada A (2001) The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/ video retrieval. In: Proceedings of the IEEE international conference on image processing, vol 1, pp 674–677
22. Koskela M, Smeaton A (2006) Clustering-based analysis of semantic concept models for video shots. In: Proceedings of the international conference on multimedia and expo, pp 45–48
23. Koskela M, Smeaton A, Laaksonen J (2007) Measuring concept similarities in multimedia ontologies: analysis and evaluations. IEEE Trans Multimedia 9:912–922
24. Kotsiantis S-B (2007) Supervised machine learning: a review of classification techniques. Informatica 31:249–268
25. Kuncheva L (2003) Fuzzy versus nonfuzzy in combining classifiers designed by bossting. IEEE Trans Fuzzy Syst 11(6),729–741
26. Kuncheva L, Bezdek JC, Duin R (2001) Decision templates for multiple classifier fusion: an experiemental comparaison. Pattern Recogn 34:299–314
27. Laaksonen J, Moskela M, Oja E (2004) Class distributions on SOM surfaces for feature extraction and object retrieval. Neural Netw 17:1121–1133
28. Li B, Goh K (2003) Confidence-based dynamic ensemble for image annotation and semantics discovery. In: Proceedings of the eleventh ACM international conference on multimedia, pp 195–206
29. Li Y, Bandar ZA, Mclean D (2003) An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 15(4):871–882
30. Lin D (1998) An information-theoretic definition of similarity. In: Proceedings of the 15th international conference on machine learning. Morgan Kaufmann, pp 296–304
31. Manjunath B, Salembier P, Sikora T (2002) Introduction to MPEG-7: multimedia content description interface. Wiley, New York
32. Messing D-S, Beek PV, Errico J-H (2001) The MPEG-7 color structure descriptor: image description using color and local spatial information. In: Proceedings of the IEEE international conference on image processing, vol 1, pp 670–673
33. Naphade MR, Kozintsev I, Huang T (2000) Probabilistic semantic video indexing. In: Proceedings of neural information processing systems, pp 967–973

34. Naphade M, Kristjansson T, Frey B, Huang T (1998) Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval in multimedia systems. In: Proceedings of the IEEE international conference on image processing, pp 536–540
35. Naphade M, Kennedy L, Kender J, Chang S, Smith J, Over P, Hauptmann A (2005) A light scale concept ontology for multimedia understanding for TRECVid 2005 (LSCOM-Lite). IBM Research Technical Report
36. Naphade M, Kennedy L, Kender J, Chang S, Smith J, Over P, Hauptmann A (2005) A light scale concept ontology for multimedia understanding for trecvid 2005. IBM Research Technical Report
37. OpenCV (2010) Intelcorporation: open source computer vision library: reference manual. http://opencvlibrary.sourceforge.net
38. Park D, Jeon YS, Won CS (2000) Efficient use of local edge histogram descriptor. In: Proceedings of ACM workshop on multimedia, pp 51–54
39. Pentland A, Picard R, Sclaroff S (1994) Photobook: content-based manipulation of image databases. In: Proceedings of SPIE conference on storage and retrieval for image and video databases
40. Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 19(1):17–30
41. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on artificial intelligence, pp 448–453
42. Seco N, Veale T, Hayes J (2004) An intrinsic information content metric for semantic similarity in WordNet. In: Proceedings of European conference on artificial intelligence
43. Slimani T, BenYaghlane B, Mellouli K (2007) Une extension de mesure de similarité entre les concepts d'une ontologie. In: International conference on sciences of electronic, technologies of information and telecommunications, pp 1–10
44. Smith J-R, Chang S-F (1996) VisualSEEk: a fully automated content-based image query system. In: Proceedings of ACM international conference on multimedia, pp 87–98
45. Snoek C-M, Worring M (2005) Multimodal video indexing: a review of the state-of-the-art. Multimedia Tools and Applications 25:5–35
46. Snoek C, Worring M, Geusebroek J-M, Koelma D-C, Seinstra F-J (2004) The mediamill TRECVid 2004 semantic viedo search engine. In: TREC video retrieval evaluation online proceedings
47. Souvannavong F (2005) Indexation et recherche de plans vidéo par le contenu sémantique. PhD thesis, Eurécom, France
48. Sun X, Manjunath B, Divakaran A (2002) Representation of motion activity in hierarchical levels for video indexing and filtering. In: Proceedings of the IEEE international conference on image processing, pp 149–152
49. TRECVID (2010) Digital video retrieval at NIST. http://www-nlpir.nist.gov/projects/trecvid/
50. Tsinaraki C, Polydoros P, Christodoulakis S (2004) Interoperability support for ontology-based video retrieval applications. In: Proceedings of the third international conference on image and video retrieval
51. Vapnik V (2000) The nature of statistical learning theory. Springer, New York
52. Vembu S, Kiesel M, Sintek M, Baumann S (2006) Towards bridging the semantic gap in multimedia annotation and retrieval. In: Proceedings of the 1st international workshop on semantic web annotations for multimedia
53. Wactlar H, Kanade T, Smith MA, Stevens SM (1996) Intelligent access to digital video: the informedia project. In: IEEE computer, vol 29
54. Wu Z, Palmer M (1994) Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on association for computational linguistics, pp 133–138
55. Wu Y, Tseng B, Smith J (2004) Ontology-based multi-classification learning for video concept detection. In: Proceedings of the international conference on multimedia and expo, vol 2, pp 1003–1006
56. Xu F, Zhang Y (2006) Evaluation and comparison of texture descriptors proposed in MPEG-7. J Vis Commun Image Represent 17:701–716
57. Xu L, Krzyzak A, Suen C (1992) Methods of combining multiple classifiers and their application to hardwriting recognition. IEEE Trans Syst Man Cybern 22:418–435
58. Yining D, Manjunath B (1998) Netra-V: toward an object-based video representation. In: Proceedings of IEEE conference of multimedia and expo, vol 8, no 5, pp 616–627



Rachid Benmokhtar received the engineering degree in 2004 from USTHB University of Algiers-Algeria, and the Master diploma in June 2005 from University of technology of Compiègne, France. In October 2005, he joined the Multimedia Communications Department at Eurecom to study toward the Ph.D. degree under the supervision of *Benoit Huet* and *Bernard Mérialdo*. He defended his Ph.D thesis in June 2009, with a Very Honorable mention. During 2009–2011, he worked with *Ivan Laptev* as a research engineer under the Quaero project at Willow-TexMex INRIA research teams. In June 2011, he moved to TexMex-INRIA of Rennes in the *Hervé Jégou* group. His research interests include multimedia indexing, content-based video retrieval, object and scene recognition from video and still images, event detection, tracking, multi-level fusion, classification, and machine learning.

Benoit Huet received his BSc degree in computer science and engineering from the Ecole Superieure de Technologie Electrique (Groupe ESIEE, France) in 1992. In 1993, he was awarded the MSc degree in Artificial Intelligence from the University of Westminster (UK) with distinction, where he then spent two years working as a research and teaching assistant. He received his DPhil degree in Computer Science from the University of York (UK) for his research on the topic of object recognition from large databases. He is currently working as a research and teaching assistant in the multimedia information processing group of the Eurecom (France). His research interests include computer vision, content-based retrieval, multimedia data mining and indexing (still and/or moving images) and pattern recognition. He has published over 80 papers in journals, edited books and refereed conferences. He is a member of IEEE, ACM and ISIF. He has served in many international conference organization and technical program committee. He is regularly invited to serves as reviewer for prestigious scientific journals as well as expert for project proposal at national, European and International level.