Hierarchical Ontology-based Robust Video Shots Indexing using Global MPEG-7 Visual Descriptors

Rachid Benmokhtar and Benoit Huet
EURECOM - Département Multimédia
2229, route des crêtes
06904 Sophia-Antipolis - France
(rachid.benmokhtar, benoit.huet)@eurecom.fr

Abstract

This paper proposes to improve our previous work on the concept-based video shot indexing, by considering an ontological concept construction in the TRECVid 2007 video retrieval, based on two steps. First, each single concept is modeled independently. Second, an ontology-based concept is introduced via the representation of the influence relations between concepts and the ontological readjustment of the confidence values. The main contribution of this paper is in the exploitation manner of the inter-concepts similarity in our indexing system, where three measures are represented: co-occurrence, visual similarity and LSCOM-lite ontology path length contribution. The experimental results report the efficiency and the significant improvement provided by the proposed scheme.

1. Introduction

The expansion of image and video collections on the Web has attracted the research community's attention for effective retrieval system, visual information management and video content analysis. Retrieving complex semantic concepts requires to extract and finely analyze a set of low-level features describing the content. A fusion mechanism can take place at different levels of the classification process. Generally, it is either applied directly on extracted features (*feature fusion*), classifier outputs (*classifier fusion*), or at the decision-making level (*decision fusion*).

Most systems concept models are constructed independently. 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. [1] have reported an ontological multi-classification learning for video concept detection. Naphade et al. [2] have modeled the linkages between various semantic concepts via a Bayesian network offering a semantics ontology. Snoek et al. [3] 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 robust schema for video shots indexing based on two levels ontological reasoning /decision 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 presents our system architecture. Section 3 gives the proposed concept ontology construction, including three types of similarities. Section 4 reports and discusses the experimentation results conducted on the TRECVid 2007 collection. Finally, section 5 provides the conclusion of the paper.

2. System architecture

The general architecture of our system can be summarized in five steps: 1. visual descriptors extraction, 2. classification, 3. perplexity-based weighted descriptors, 4. classifier fusion and 5. ontological readjustment of the confidence values.

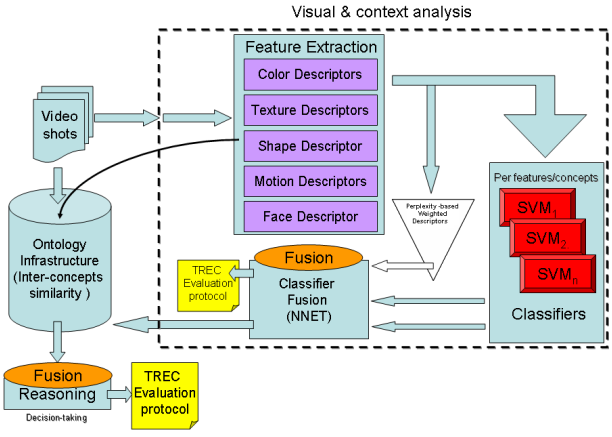

Figure 1. General indexing system architecture.

2.1. Visual Descriptors

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. Five types of MPEG-7 global visual descriptors are extracted on the selected keyframes : Color (ScalableColor, ColorLayout, ColorStructure, ColorMoment), texture (EdgeHistogram, HomogeneousTexture, StatisticalTexture), shape (ContourShape), motion (CameraMotion, MotionActivity), and FaceDescriptor (For more details, see [4]).

2.2. SVM-based Classification

The main idea is similar to the concept of a neuron: Separate classes with a hyperplane [5]. However, samples are indirectly mapped into a high dimensional space due 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. A sigmoid function is employed to compute the degree of confidence $y_i^j = 1/\left(1+\exp\left(-\alpha d_i\right)\right)$. Where (i,j) represents the i^{th} concept and j^{th} low-level feature. d_i: distance between the input vector and the hyperplane. α: slope parameter obtained experimentally.

2.3. Perplexity-based Weighted Descriptors

In [4], we have proposed a novel approach to weight each low-level feature per concept within an adaptive classifier fusion step (section 2.4) that we call PENN "Perplexity-based Evidential Neural Network". The proposed approach, as presented in Fig. 2 will now be briefly defined.

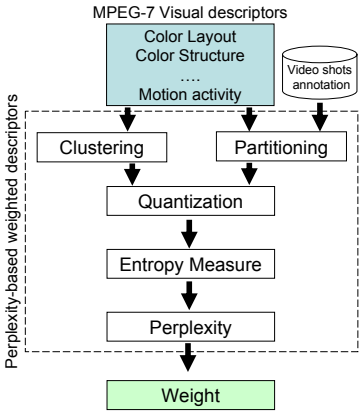

Figure 2. Perplexity-based weighted descriptors steps.

1. K-means Clustering: computes the k center of the clusters for each descriptor, in order to create a "visual dictionary" of the shots. 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 [6], 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 diminished by quantization of more than 1000 clusters. Therefore, $k_r = 2000$ are used for the quantization of MPEG-7 region descriptors, and $k_g = 100$ for the global, that present a compromise between efficiency and a low time-consuming computation.

2. Partitioning: selects the positive samples per concept.

3. Quantization: computes Euclidean distance between each partitioning data set and dictionary.

4. Entropy measure: The entropy $H = -\sum_{i=0}^{k-1} P_i \log(P_i)$ 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. In [7], a good model is such that the distribution is heavily concentrated on a few clusters only, resulting in a low entropy value .

5. Perplexity measure: In [8], perplexity PPL or normalized perplexity value $\overline{PPL}$ (Eq. 1) can be interpreted as the average number of clusters needed for an optimal coding of the data.

$$\overline{PPL} = \frac{PPL}{PPL_{max}} = \frac{2^H}{2^{H_{max}}} \tag{1}$$

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

6. Weight: It is generally assumed that lower perplexity (or entropy) correlates with better performance [8], or in our case, to a very concentrated distribution. So the relative weight of the corresponding feature should be increased. Many formulae can be used to represent the weight such as Sigmoid, Softmax, etc. Here, we choose Verhulst's evolution model (Eq. 2). This function is non exponential and allows for brake rate α_i, reception capacity K, and decreasing speed of weight function β_i (For more details, see [4]).

$$w_i = K/\left(1+\beta_i \exp\left(-\alpha_i(1/\overline{\mathrm{PPL}_i})\right)\right) \tag{2}$$

2.4. 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. In this work, we have used our recently proposed neural network based on evidence theory (NNET) to address classifier fusion [9].

3. Concept Ontology Construction

The ontology has been historically used to achieve better performance in the multimedia retrieval system. 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 [10]. In LSCOM-lite ontology [11], we notice positive relationships such as (ROAD, CAR), (VEGETATION, MOUNTAIN), and negative relationships like (BUILDING, SPORTS), (SKY, MEETING).

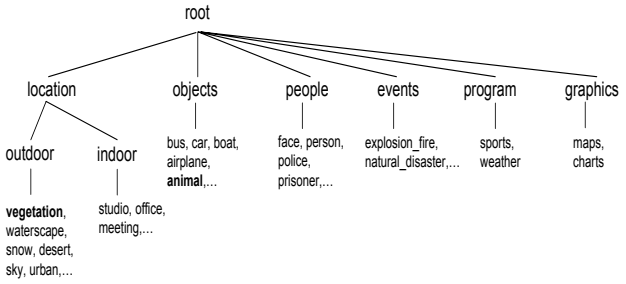

Figure 3. Fragment of the hierarchical LSCOM-Lite.

In this section, we 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 [12]. For example, Fig. 3 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. To address this weakness, more information between the concepts are introduced, so that it becomes a function of attributes co-occurrence, low-level visual descriptors, path length, depth and local density to boost the performance of specific indexing system, as: $\lambda(C_m, C_n) = (Sim_{cos} + Sim_{vis} + Sim_{sem})(C_m, C_n)$.

3.1 Co-occurrence: 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 components [7].

$$Sim_{cos}(P^m, P^n) = \frac{\sum_{i=0}^{k-1} P_i^m P_i^n}{\sqrt{\sum_{i=0}^{k-1}(P_i^m)^2 \sum_{i=0}^{k-1}(P_i^n)^2}} \quad (3)$$

3.2 Visual similarity: is based upon low level visual features. In section 2.3, 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* [7] which is like d_{KL}, but is numerically more stable.

$$d_{JD}(P^m, P^n) = \sum_{i=0}^{k-1} \left(P_i^m \log \frac{P_i^m}{\hat{P}_i} + P_i^n \log \frac{P_i^n}{\hat{P}_i} \right) \quad (4)$$

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

$$Sim_{vis}(C_m, C_n) = \frac{1}{\sum_{i=1}^{Nb\ features} \frac{1}{2}(w_i^m + w_i^n) d_{JD}^i(P^m, P^n)} \quad (5)$$

where w_i^m is the i^{th} perplexity-based weighted descriptors for the concept m.

3.3 Semantic similarity: between the concepts has been widely studied in the literature and can be classified in three major approaches [13]: (1) distance-based approaches (i.e, based on the ontology structure), (2) information content-based approaches (IC) and finally (3) the hybrid approaches (i.e, combine the two previous approaches).

For the hierarchical LSCOM-lite ontology presented in Fig. 4, we have decided to use hybrid approach proposed by Jiang & Conrath measure [14] (Equ. 6), associating probabilities to each concept in the ontology based on occurrences in a given corpus. The IC is then obtained by considering the negative log likelihood: $IC(C_i) = -log(p(C_i))$. We also propose a novel hybrid combination form of semantic similarity as presented in Equ. 7 which will be compared with the standard $J\ \&\ C$ approach.

$$\begin{cases} Sim_{sem_{J\&C}}(C_m, C_n) = 1/d_{J\&C}(C_m, C_n) \\ d_{J\&C}(C_m, C_n) = IC(C_m) + IC(C_n) - 2 * IC(CS(C_m, C_n)) \end{cases} \quad (6)$$

$$Sim_{sem}(C_m, C_n) = 1/\left(d_{Rada}(C_m, C_n) + d_{J\&C}(C_m, C_n)\right) \quad (7)$$

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

3.1. 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, and is computed using:

$$\underline{P\left(x/Ci\right)} = P\left(x/Ci\right) + \frac{1}{Z}\sum_{j=1}^{Nb\ arc}\lambda_{i,j}(1-\zeta_j)P\left(x/Cj\right) \tag{8}$$

where $\underline{P\left(x/Ci\right)}$ 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.

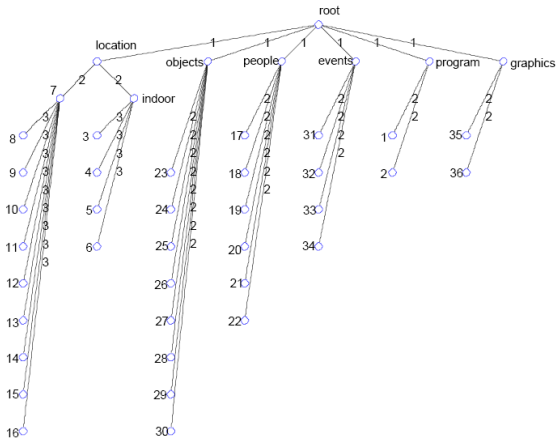

Figure 4. Hierarchical ontology model.

4. Experimentations

The experiments provided here are conducted on the TRECVid 2007 dataset [15] containing science news, news reports, documentaries, etc. Of the 100 hours of video segmented into shots and annotated [16] with semantic concepts from the 36 defined labels [1]. 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. Other metrics are introduced in our evaluation: F-measure, positive classification rate CR^+, and balanced error rate BER.

Fig. 5 shows the variation of average precision results *vs* semantic concepts, for three systems: NNET [2], PENN [3],

1. The feature extraction task consists in retrieving shots expressing one of the following 36 semantic concepts: (1)SPORTS, (2)WEATHER, (3)COURT, (4)OFFICE,..., (35)MAPS, (36)CHARTS [11].

2. PENN: Perplexity-based Evidential Neural Network.

3. NNET: Neural Network based on Evidence Theory.

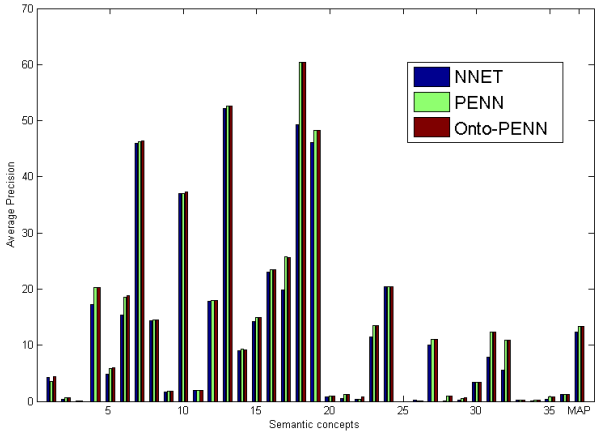

Figure 5. Average precision evaluation.

and Onto-PENN [4]. First, we observe that PENN 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 2000 returned shots as in TRECVid). 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. Fig. 6 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 CR^+ is improved by 10.14% to achieve 22.07%, and decreasing the BER of 2.91% compared to NNET.

Fig. 7 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,

4. Onto-PENN: Ontological readjustment of the PENN. The results presented in the rest of paper for the Onto-PENN, are given by Equ. 7 for the semantic similarity computation.

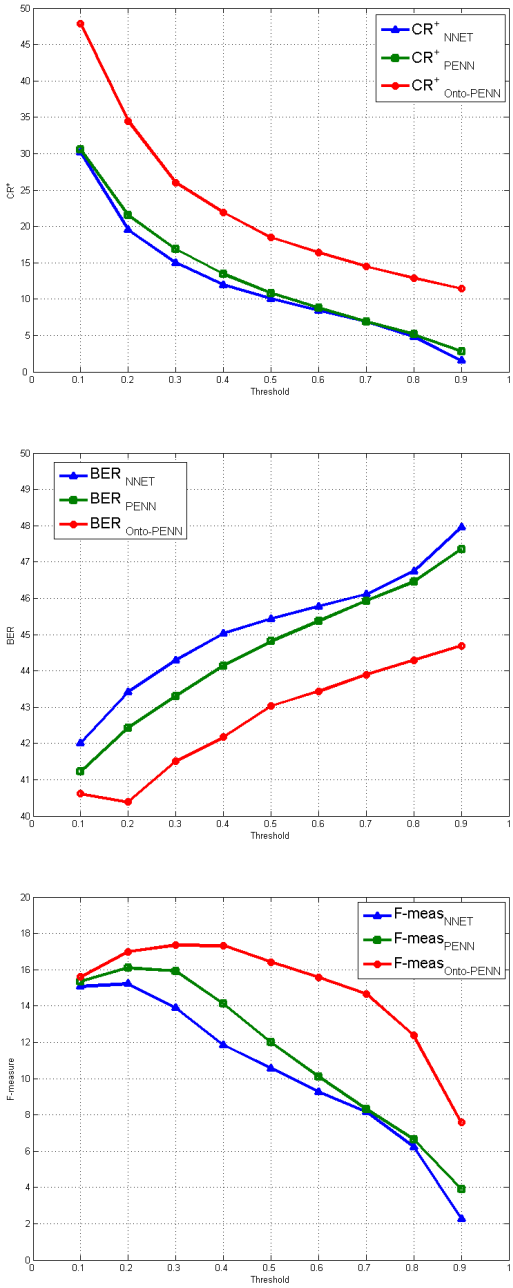

Figure 6. Evaluation of the metrics (CR^+, BER and F-measure) vs *Threshold* $\tau \in [0.1, 0.9]$.

$CR^+ = 0$) 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.

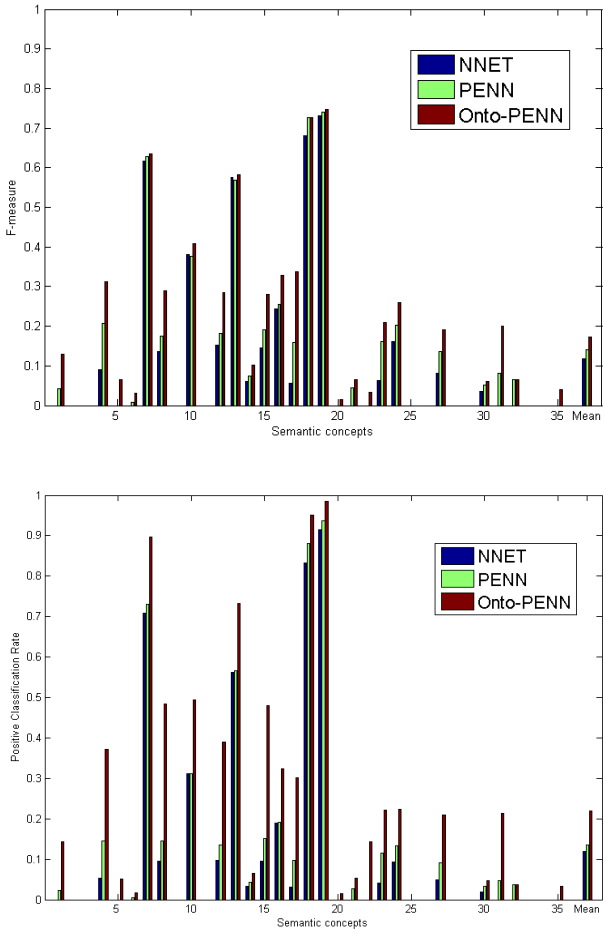

Figure 7. F-measure and CR^+ evaluation.

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. 8).

Table 1 summarizes the overall performances for the content-based video shots classification systems using a fixed $\tau = 0.4$. We compute the above mentioned statistics for all concepts, and for a subset composed of the 10 most frequent concepts in the dataset. Both hybrid semantic similarity-based Onto-PENN allow an overall improvement of the system and a significant increase of F-meas and CR^+. They achieve a respectable result of MAP, significantly decrease the "BER" compared to NNET and PENN.

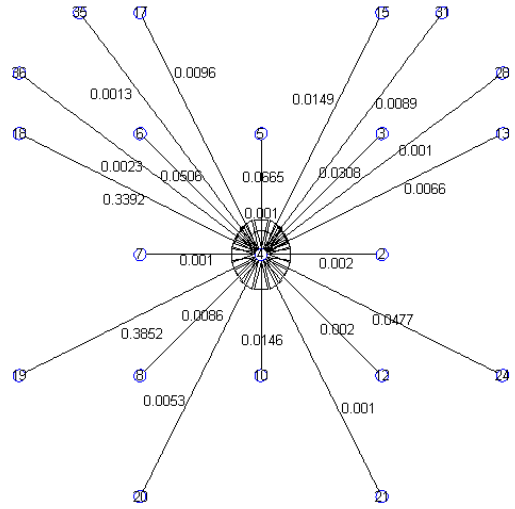

Figure 8. Inter-concept connections graphical model for the concept OFFICE. We observe that 20 concepts are connected with OFFICE, but only 5 are strong and significant (MEETING:6.65%, STUDIO:5.06%, FACE:33.92%, PERSON:38.52%, and COMPUTERTV:4.77%) presenting 88.92% of the global information.

Table 1. Performances comparisons (*Threshold*$= 0.4$).

Methods / Evaluation (%)	NNET	PENN	Onto-PENN (Equ. 6)	Onto-PENN (Equ. 7)
MAP	12.70	13.29	13.31	13.37
MAP@10	33.70	35.30	35.30	35.36
F-meas	11.84	14.10	17.07	17.30
F-meas@10	38.75	40.79	44.67	44.74
CR^+	11.93	13.43	21.76	22.07
CR^+@10	40.69	41.74	59.45	59.71
BER	45.02	44.13	42.32	42.11
BER@10	38	36.52	34.03	33.96

Finally, the results given by the two equations (Equ. 6 and Equ.7) used in the semantic similarity construction are very close, with a slight advantage for the Equ. 7. Other experiments have been made for choosing the semantic similarity, but due to space constraints are not reported in this paper. However, it can be observed that performance declines using the equation of Lin et al. [10] compared to the two used equations, which underlines the importance of the semantic similarity.

5. Conclusions

In this paper, we have described an ontological-based robust video shots indexing. Ontology is defined for learning the influence of the relation between concepts. Three types of influence are used: co-occurrence, visual similarity and semantic similarity to improve the accuracy of the independent concept classifiers, on the TRECVid 2007 benchmark. Our proposed approach obtains a significant improvement, about 18.75% of CR^+, 5.99% of F-meas, 1.66% of MAP, and deceases the BER with 2.91%. Future works will concern the similarities from WordNet instead of a corpus.

References

[1] Y. Wu, B-L. Tseng, and J-R. Smith. Ontology-based multi-classification learning for video concept detection. In *the proceedings of IEEE ICME*, 2:1003–1006, 2004.

[2] M-R. Naphade et al. Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. In *the Proceedings of IEEE ICIP*, pp. 536–540, 1998.

[3] C-G. Snoek et al. The Mediamill TRECVid 2004 semantic video search engine. In *the Proceedings of TRECVid*, 2004.

[4] R. Benmokhtar and B. Huet. Perplexity-based evidential neural network classifier fusion using MPEG-7 low-level visual features. In *the proceedings of ACM MIR*, pp. 160-169, 2008.

[5] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[6] F. Souvannavong. *Indexation et Recherche de Plans Vidéo par le Contenu Sémantique*. PhD Thesis, Eurécom, France, 2007.

[7] M. Koskela, A-F. Smeaton, and J. Laaksonen. Measuring concept similarities in multimedia ontologies: Analysis and evaluations. In *the Proceedings of IEEE Trans. on Multimedia*, 9(5):912–922, 2007.

[8] J. Gao et al. Toward a unified approach to statistical language modeling for chinese. In *the proceedings of ACM TALIP*, 2001.

[9] R. Benmokhtar and B. Huet. Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content. In *the proceedings of MMM*, vol. 1, pp. 196-205, 2007.

[10] D. Lin. An information-theoretic definition of similarity. In *the Proceedings of ICML*, pp. 296–304, 1998.

[11] M-R. Naphade et al. A Light Scale Concept Ontology for Multimedia Understanding for trecvid 2005 (LSCOM-Lite). *IBM Research Technical Report*, 2005.

[12] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30, 1989.

[13] T. Slimani, B. BenYaghlane, and K. Mellouli. Une extension de mesure de similaritè entre les concepts d'une ontologie. In *the Proceedings of SETIT*, pp. 1–10, 2007.

[14] J. Jiang and D-W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *the Proceedings of ROCLING*, 1997.

[15] TRECVID. Digital video retrieval at NIST. *http://www-nlpir.nist.gov/projects/trecvid/*.

[16] S. Ayache and G. Quénot. TRECVid 2007 collaborative annotation using active learning. In *the proceedings of TRECVid*, 2007.