ONTOLOGICAL RERANKING APPROACH FOR HYBRID CONCEPT SIMILARITY-BASED VIDEO SHOTS INDEXING

Rachid Benmokhtar and Benoit Huet

EURECOM - Département Multimédia
2229, route des crêtes - Sophia-Antipolis, France
(rachid.benmokhtar, benoit.huet)@eurecom.fr

ABSTRACT

This paper proposes to compare three hybrid concept similarity measures for video shots indexing and retrieval [1], based on two steps. First, individuals concepts are modeled independently. Second, an ontology is introduced via the representation of the relationship between concepts and the ontological readjustment of the confidence values. Our contribution lies in the manner in which inter-concepts similarities are exploited in the indexing system using co-occurrence, visual descriptors, and hybrid semantic similarities. Experimental results report the efficiency and the significant improvement provided by the proposed scheme.

1. INTRODUCTION

Most indexation 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. [2] have reported an ontological multi-classification learning for video concept detection. Naphade et al. [3] have modeled the linkages between various semantic concepts via a Bayesian network offering a semantics ontology, etc. In this paper, we propose a robust ontological indexing system (Fig. 1), that can be summarized in five steps: (1) visual descriptors extraction, (2) SVM classification, (3) perplexity-based weighted descriptors [1], (4) NNET classifier fusion [4] and (5) ontological readjustment of the confidence values. Here, we focus on the last step for ontological reasoning and decision construction, taking into account the relationships between concepts.

This paper is organized as follows. Section 2 presents the proposed concept ontology construction, including co-occurrence, visual descriptors and hybrid concept similarities. Section 3 reports and discusses the experimentation results conducted on the TRECVid 2007 collection. Finally, section 4 provides the conclusion of the paper.

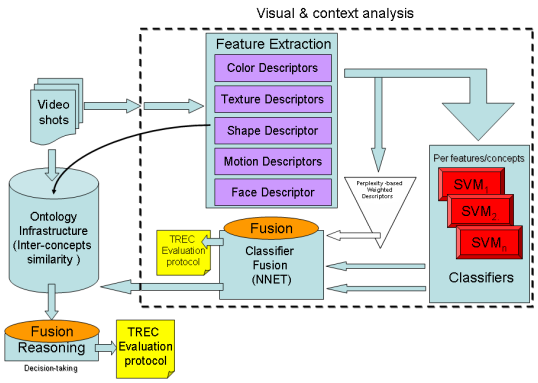

Fig. 1. General indexing system architecture.

2. CONCEPT ONTOLOGY CONSTRUCTION

In multimedia retrieval systems, ontology has been historically used to achieve better performance by defining representative concepts and their relationships. Psychology demonstrates that similarity depends on the context, and may be asymmetric [5]. The LSCOM-lite ontology [6], contains two types of relationships: Positive such as (BUILDING, OUTDOOR), (ROAD, CAR), and negative like (SKY, MEETING), (ROAD, OFFICE). 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 connecting two concepts [7], taking into account several information between the concepts such as co-occurrence, low-level visual descriptors, path length, depth and local density to boost the performance of specific indexing system.

2.1. 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 others. Many methods are proposed in literature to represent this proximity. Here, we use Cosine similarity because it reflects similarity in terms of relative distributions of components. Cosine is not

influenced by one document being small compared to others like the Euclidean distance tends to be [8]:

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

2.2. Visual similarity

The second similarity is based upon low level visual feature. In [1], we used entropy/perplexity to build a weighted descriptor per concept. Here, the visual similarity is computed with *Jeffrey divergence* d_{JD} which is like $d_{Kullback-Leibler}$, but is numerically more stable [8].

$$d_{JD}(P^m, P^n) = \sum_{i=0}^{k-1}\left(P_i^m \log\frac{P_i^m}{\hat{P}_i} + P_i^n \log\frac{P_i^n}{\hat{P}_i}\right) \quad (2)$$

with $\hat{P}_i = \frac{P^m + P^n}{2}$ is the mean distribution.

The visual distance between concepts (C_m, C_n) is :

$$Sim_{vis}(C_m, C_n) = \sum_{i=1}^{Nb\ Feature} \frac{1}{2}(w_i^m + w_i^n) d_{JD}(P^{m,i}, P^{n,i}) \quad (3)$$

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

2.3. Semantic similarity - Contribution of Path Length

The semantic similarity between the concepts has been widely studied in the literature and can be classified in three major categories approaches.

2.3.1. 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, presented by the minimum number of edges that separates the two concepts. Rada et al. [7] propose the following equation:

$$Sim_{sem}(C_m, C_n) = 1/\left(1 + dist_{Rada}(C_m, C_n)\right) \quad (4)$$

Wu and Palmer [9] propose a similarity-based on the depth of the concept subsumes CS [1] and the two concepts (Equ. 5).

$$Sim_{sem}(C_m, C_n) = \frac{2 * depth(CS)}{depth(C_m) + depth(C_n)} \quad (5)$$

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.

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

2.3.2. 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 [10], the semantic similarity can be obtained per the frequency of appearance in the corpus, and defined by :

$$Sim_{sem}(C_m, C_n) = \max\left(IC(CS(C_m, C_n))\right) \quad (6)$$

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. [11] 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.

2.3.3. Hybrid approach

It 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. [5], or with the distance of Jiang & Conrath $dist_{J\&C}$ [12].

$$Sim_{sem_{Lin}}(C_m, C_n) = \frac{2 * \log P(CS)}{\log P(C_m) + \log P(C_n)} \quad (7)$$

$$\begin{cases} dist_{J\&C}(C_m, C_n) = IC(C_m) + IC(C_n) - 2 * IC(CS(C_m, C_n)) \\ Sim_{sem_{J\&C}}(C_m, C_n) = 1/\left(dist_{J\&C}(C_m, C_n)\right) \end{cases} \quad (8)$$

For the ontology presented in the Fig. 2, we compare the last two hybrid approaches with the novel one as presented in the Equ. 9, that it is the combination of Rada and J&C.

$$Sim_{sem}(C_m, C_n) = 1/\left((dist_{Rada} + dist_{J\&C})(C_m, C_n)\right) \quad (9)$$

2.4. Concept-based Confidence Value Readjustment (CCVR)

The proposed framework (Fig. 1) introduces a *reranking* or confidence value readjustment to refine the results. It is computed using the following equation:

$$\underline{P\left(x/Ci\right)} = P\left(x/Ci\right) + \frac{1}{Z}\sum_{j=1}^{Nb\ arc} \lambda_{i,j}(1 - \zeta_j) P\left(x/Cj\right) \quad (10)$$

where $\underline{P\left(x/Ci\right)}$ corresponds to the multi-modal result, $\lambda_{i,j} = \left(Sim_{cos_{i,j}} + Sim_{vis_{i,j}} + Sim_{sem_{i,j}}\right)$ 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.

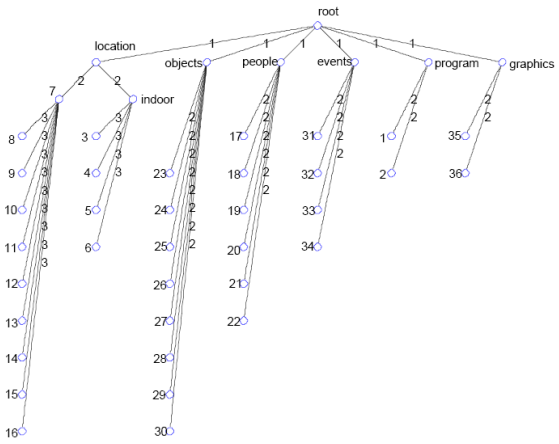

Fig. 2. Hierarchical LSCOM-lite ontology model.

3. EXPERIMENTATIONS

The experiments provided here are conducted on the TRECVid 2007 dataset [13]. Of the 100 hours of video segmented into shots and annotated with semantic concepts from the 36 defined labels [2], half is used to train the feature extraction system and the rest for the evaluation purposes. The evaluation is realized in the context of TRECVid using Mean Average Precision MAP. Other metrics are introduced in our evaluation: F-measure, positive classification rate CR^+, and balanced error rate BER. Five types of MPEG-7 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 [1]).

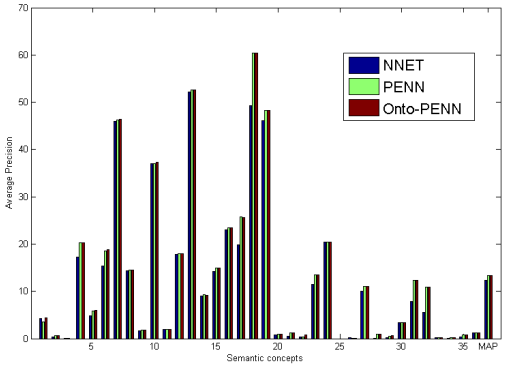

Fig. 3. Average precision evaluation.

[2]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 [6].

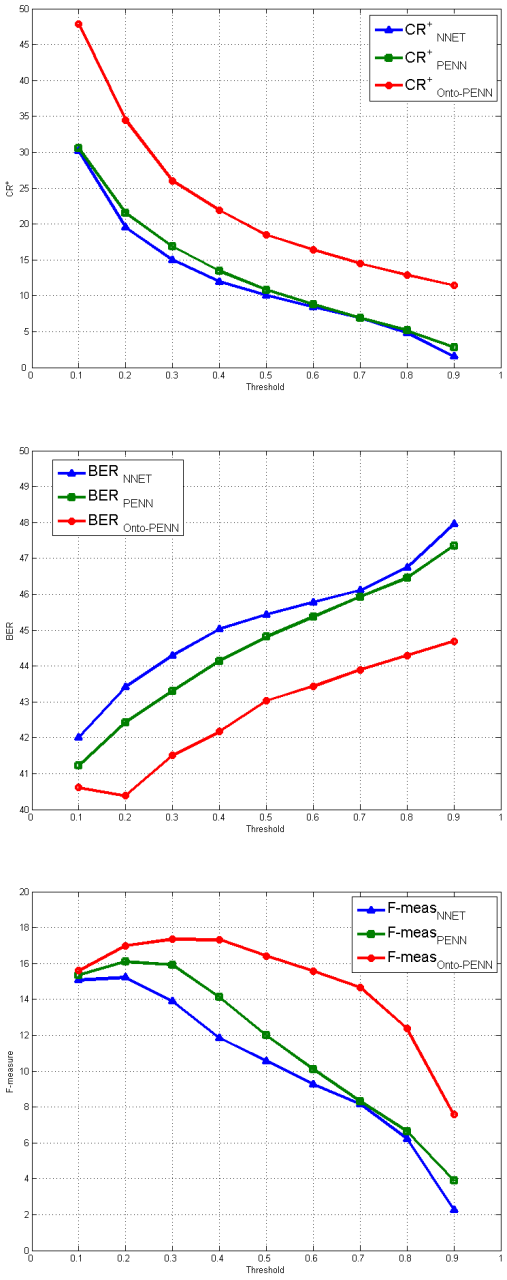

Fig. 4. Evaluation of the metrics (CR^+, BER and F-measure) vs *Threshold* $\in [0.1, 0.9]$.

Fig. 3 shows the variation of average precision results *vs* semantic concepts, for three systems: NNET [3], PENN [4], and Onto-PENN [5]. First, we observe that PENN and Onto-PENN

[3]NNET: Neural Network based on Evidence Theory [4].

[4]PENN: Perplexity-based Evidential Neural Network [1].

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

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 is computed (using only the first 2000 returned shots as in TRECVid [13]). 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-measure, CR^+ and BER. Note that the MAP is not sensitive to *Threshold* values. Fig. 4 compares the three experimental systems along with the variation of *Threshold* $\in [0.1, 0.9]$ by step of 0.1. We can clearly see that for any *Threshold* 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 *Threshold*$= 0.2$, for F-meas$= 16.98\%$ and $CR^+ = 34.48\%$. The best results are obtained for *Threshold* $\in [0.2, 0.5]$. With the *Threshold* fixed at 0.40, the CR^+ is improved by 10.14% to achieve 22.07%, and decreasing the BER of 2.91% compared to NNET.

Table 1 summarizes the overall performances for the content-based video shots classification systems using a fixed *Threshold*$= 0.4$. 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 (Equ. 8 and Equ. 9) are very close, with a slight advantage for the Equ. 9. However, it can be observed that the MAP declines using the equation of Lin et al. [5] compared to the two used equations, which underlines the importance of the semantic similarity.

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

Methods / Eval.(%)	NNET	PENN	Onto-PENN		
			Lin	J&C	(Equ. 9)
MAP	12.70	13.29	13.01	13.31	13.37
MAP@10	33.70	35.30	34.91	35.30	35.36
F-meas	11.84	14.10	16.17	17.07	17.30
F-meas@10	38.75	40.79	43.41	44.67	44.74
CR^+	11.93	13.43	20.58	21.76	22.07
CR^+@10	40.69	41.74	57.80	59.45	59.71
BER	45.02	44.13	43.62	42.32	42.11
BER@10	38	36.52	35.45	34.03	33.96

4. CONCLUSIONS

In this paper, we have presented an ontological-based robust video shots indexing to learn the influence of the relation between concepts. Three types of influence are used: co-occurrence, visual descriptors and semantic similarity based on hybrid approaches to improve the accuracy of the independent concept classifiers. Thought TRECVid 2007 benchmark, it obtains a significant improvement of our system, about 18.75% of CR^+, 5.99% of F-measure, 1.66% of MAP, and decreases the balanced error rate with 2.91%. The future works will concern the similarities from WordNet instead of a corpus.

5. REFERENCES

[1] R. Benmokhtar and B. Huet, "Perplexity-based evidential neural network classifier fusion using MPEG-7 low-level visual features," in *Proceedings of ACM MIR*, 2008, pp. 336–341.

[2] Y. Wu, B-L. Tseng, and J-R. Smith, "Ontology-based multi-classification learning for video concept detection," *Proceedings of IEEE ICME*, vol. 2, pp. 1003–1006, 2004.

[3] M-R. Naphade, T. Kristjansson, B. Frey, and T-S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems," in *Proceedings of IEEE ICIP*, 1998, pp. 536–540.

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

[5] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of ICML*. 1998, pp. 296–304, Morgan Kaufmann.

[6] M-R. Naphade, L. Kennedy, J-R. Kender, S-F. Chang, J-R. Smith, P. Over, and A. Hauptmann, "A Light Scale Concept Ontology for Multimedia Understanding for TRECVid 2005 (LSCOM-Lite)," *IBM Research Technical Report*, 2005.

[7] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *Proceedings of IEEE SMC*, vol. 19, no. 1, pp. 17–30, 1989.

[8] M. Koskela, A.F. Smeaton, and J. Laaksonen, "Measuring concept similarities in multimedia ontologies: Analysis and evaluations," *IEEE Trans. on Multimedia*, pp. 912–922, 2007.

[9] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of Annual Meeting of ACL*, 1994, pp. 133–138.

[10] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of IJCAI*, 1995, pp. 448–453.

[11] N. Seco, T. Veale, and J. Hayes, "An intrinsic information content metric for semantic similarity in WordNet," in *Proceedings of ECAI*, 2004.

[12] J. Jiang and D-W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of ICCL*, 1997.

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