Conceptual feedback for semantic multimedia indexing

Abdelkader Hamadi, Philippe Mulhem and Georges Quénot
{firstname.lastname}@imag.fr
UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France

***Abstract*—In this paper, we consider the problem of automatically detecting a large number of visual concepts in images or video shots. State of the art systems involve feature (descriptor) extraction, classification (supervised learning) and fusion when several descriptors and/or classifiers are used. Though direct multi-label approaches are considered in some works, detection scores are often computed independently for each target concept. We propose here a method that we call "conceptual feedback" for improving the overall detection performance that implicitly takes into account the relations between concepts. The vector of normalized detection scores is added to the pool of available descriptors. It is then processed just as the other descriptors for the normalization, optimization and classification steps. The resulting detection scores are finally fused with the already available detection scores obtained with the original descriptors. The feedback of the global detection scores in the pool of descriptors can be iterated several times. It is also compatible with the use of the temporal context that also improves the overall performance by taking into account the local homogeneity of video contents. The method has been evaluated in the context of the TRECVID 2012 semantic indexing task involving the detection of 346 visual or multimodal concepts. Combined with temporal re-scoring, the proposed method increased the global system performance (MAP) from 0.2613 to 0.3014 (+15.3% of relative improvement) while the temporal re-scoring alone increased it only from 0.2613 to 0.2691 (+3.0%).**

Keywords*—Semantic Indexing, Multimedia, Fusion, Conceptual Feedback.

I. Introduction

Semantic multimedia indexing consists in automatically assigning tags (or labels or concepts) to multimedia samples. We shall consider here video segments (shots) but the proposed approach could be used too for image or audio and even text samples. Semantic multimedia indexing is a key element for video retrieval when textual metadata is not available, incomplete or inaccurate, which is frequently the case in ever growing video collections. Concept-based queries have been proven to be efficient for video passage retrieval [1]. Semantic multimedia indexing is difficult because of the so-called "semantic gap" [2] or the discrepancy between the level at which the multimedia samples are represented (raw signal) and the level of the concepts and relations making sense to human beings (semantic).

Semantic multimedia indexing is generally carried out by supervised learning. Multimedia samples are represented by descriptors (vectors of fixed dimensionality) extracted from the raw contents. A classifier is trained from annotated samples and then applied for prediction to unannotated (test) samples. When several descriptors types are available (e.g. from color, texture, feature points) fusion can be applied to improve the overall performance. Early (descriptors) and late (classification scores) fusion schemes have been proposed by Snoek et al. [3]. In this case, fusion is performed between different classifiers (using different descriptors and/or different learning methods) but for a single concept or separately for each concept. In this work, we consider fusion between classifiers for different concepts.

The classical extraction/classification/fusion pipeline has been successfully applied for semantic multimedia indexing but because of the semantic gap the overall performance remains low: the Mean Average Precision (MAP) of state of the art systems at TRECVid is in the 0.1-0.3 range [4]. At the last TRECVid evaluation, the best system reached a MAP of 0.32. Over the last years, the performance has been progressively increased using better descriptors, more descriptors, better classification methods, better fusion schemes and better annotation. This classical approach can further be improved by taking into account the temporal coherency of video contents and/or the relationships between the target concepts when a number of them have to be detected simultaneously. These aspects have been taken into account in several recent works and the approach presented in this paper also aims at exploiting them for improving the performance of a semantic indexing system.

II. Related work

In the case of video indexing where the indexed units are shots, two types of contexts can be considered: the temporal context between shots from a same video documents and the conceptual context between the different target concepts. These two types of contexts constitute distinct and orthogonal dimensions (all target concepts have to be indexed for all shots). Both dimensions can be considered and used separately or jointly for improving system performance.

978-1-4799-0956-8/13/$31.00 ©2013 IEEE

Approaches considering the conceptual dimension include Naphade et al. [5] Multinet approach in which concepts are organized in an graph and relationship between them are modeled in a Bayesian framework. While several approaches mainly consist in post-processing classification scores obtained from independent classifiers, Qi et al. [6] directly proposed to build and train a global classifier for all target concepts at once.

Considering the temporal dimension, Safadi et al. proposed a re-scoring approach applied to classification scores obtained independently for each shot. The principle is to re-score the video shots as a combination of the original scores with an average of these scores at the level of the whole video (for short and homogeneous videos) [7] or within a fixed-size neighborhood (for long and heterogeneous videos) [8].

Both dimensions can be considered simultaneously. For instance, Qi et al. extended their original approach [6] in order to also take into account the temporal dimension through the use of temporal kernels [9]. Weng et al. [10] also proposed a direct method for taking into account both dimensions while also focusing on the cross-domain adaptation problem. Methods considering both dimensions at once are theoretically able to capture all the available "conceptuo-temporal" relations between concepts (e.g. "concept A appears n shots after concept B with a probability p") into a single statistical model. However, they might also be prone to over-fit or require large amounts of training data in order to accurately train their models. They might also not be easily compatible with sparse or incomplete multi-concept annotations.

Another possibility for taking into account both dimensions is to apply temporal and conceptual approaches sequentially (or alternatively in case of iterative variants), this can be done easily for post-processing based methods. Though possibly less optimal than methods considering both dimensions at once, they may be less prone to over-fit and easier to implement.

The method proposed in this paper is primarily considering the conceptual dimension. Its originality is that it involves a feedback of the detection scores via an additional descriptor, similar to the "model vectors" proposed by Smith et al. [11], that also goes through the classification and fusions stages. The temporal dimension can also be taken into account through a sequential or alternative application.

III. Proposed method

We consider a "baseline" system which is a classical three-stage "extraction/classification/fusion" pipeline as shown in figure 1. n_e *extractors* are used to extract descriptors from the image or audio tracks of the video samples (typically shots). $n_e \times n_c$ *classifiers* are then trained or applied separately for each (descriptor, concept) pair. n_c *(late) fusion modules* are finally used for producing n_c scores each representing the likeliness of a video sample to contain the target concepts. The same set of descriptors and the same fusion scheme are used for all the target concepts. It is up to the classifiers and/or to the fusion tuning procedure to select and/or weight the relevant elements. The training and tuning is done using annotated training data. Once trained/tuned, the system can be applied to unseen video samples for producing detection scores for them. The exact type of descriptors, classifiers or fusion methods used does not need to be specified at this point so the proposed method is as general as possible.

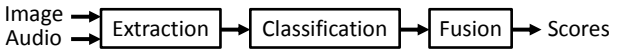

Fig. 1. Baseline system pipeline architecture.

In this baseline system, the processing is done completely independently for all video samples and for all target concepts, usion is done only between descriptors or learning methos but not between concepts. As previously mentioned, this is not optimal because it does not take into account the temporal relationships between video samples (shots) and semantic or statistic relationships between the target concepts. As an alternative to existing approaches, we propose to build a conceptual descriptor which is a normalized version of the vector of scores produced by the baseline system. This descriptor is similar to Smith et al.'s model vector [11]. After normalization, it is added as a feedback to the pool of already available descriptors coming directly from the samples' image or audio signal as shown in figure 2.

For being usable as a classical descriptor, the conceptual descriptor must be available both on the training (development) data and on the prediction (test) data. While it is naturally available on the test data as a result of the training on the development data, we also need to build it also on the development data. In this case, cross-validation is used. This approach is also used when learning has to be done in the fusion process. Though it is less homogeneous since the conceptual descriptor is not built exactly in the same way on all the parts of the data, the processus is similar enough for an efficient operation.

The new descriptor is finally processed exactly as the already available ones: it is used for classification and included in the fusion process. Indeed, in practice, there is no loop in the processing flow and the feedback is actually unfolded as shown in figure 3. The conceptual feedback can be applied several times until some kind of convergence is reached.

The conceptual descriptor normalization can be done in different ways. The simplest one is to apply a linear gain and offset normalization so that all components have a zero-centered and unit-variance distribution on the training set. Additionally, each component can be assigned a weight related for instance to an estimation of the performance (e.g. AP) of the system for the corresponding concept.

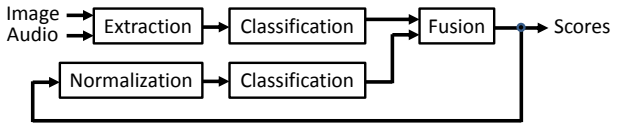

Fig. 2. Semantic indexing system with conceptual feedback.

Fig. 3. Semantic indexing system with conceptual feedback, unfolded (single iteration).

As just described, the conceptual feedback does not take into account the temporal coherency of video documents. For this, we rely on a separate approach, "temporal re-ranking", proposed by Safadi et al. [7], [8]. This is done separately and in the same way for each target concept. The temporal re-ranking is actually performed via a re-scoring which is done in two steps. First, a global score is computed for each video as its likeliness of containing the target concept; this score is computed from the scores of all the shots within the video. Then, the score of each shot is re-evaluated according to the global score of the video it belongs to.

We consider a collection of videos $V = (v_1, v_2, \ldots, v_m)$, m being the number of videos in the collection. Each video v_i composed of a sequence of shots $v_i = (s_{i1}, s_{i2}, \ldots, s_{in_i})$, n_i being the number of shots of v_i.

For each shot s_{ij}, an initial classification score x_{ij} is computed from supervised learning on the development set. Many options are possible for the computation of a global score x_i for a video v_i from the shots that it contains. Safadi et al. [7], [8] tried several formulas and found that the following one which is a generalization of the mean of the shot scores was the most efficient:

$$z_i = \left(\frac{\sum_{j=1}^{n_i} (x_{ij})^\alpha}{n_i} \right)^{1/\alpha}, \quad (1)$$

where α is a parameter that has to be tuned by cross-validation within the development collection. Then, the score of each shot is updated according to its previous score and the global score of the video it belongs to. Again, many options were possible and a weighted multiplicative fusion was chosen:

$$x'_{ij} = x_{ij}^{1-\gamma} \times z_i^\gamma, \quad (2)$$

where γ is a parameter that controls the "strength" of the re-ranking. It also has to be tuned by cross-validation within the development collection.

From the conceptual feedback framework perspective, temporal re-scoring can be seen as integrated as a post-processing step inside the fusion module (as descriptor normalization or optimization can be seen as a preprocessing step integrated inside the classification module). With this approach, the temporal and conceptual aspects are both taken into account but sequentially (or alternatively in the case of iterative feedback) instead of jointly as this is done in other approaches. This might appear less optimal but it is probably so only slightly because previous experiments showed that the way in which the optimal temporal window size depends upon the concept is small and unstable between development and test sets. Also, this sub-optimality should be reduced in the case of iterative feedback.

The novelty of the proposed approach comes from the use of a conceptual descriptor in addition to a pool of already available ones as a kind of feedback and from the combination of the conceptual feedback and the temporal re-ranking (or re-scoring) in an iterative way so that both aspects can be taken into account jointly.

IV. Evaluation

Evaluations are carried out in the context of the TRECVid 2012 [12] semantic indexing (SIN) task. This task is defined as follows: "Given the test collection, master shot reference, and concept definitions, return for each concept a list of at most 2000 shot IDs from the test collection ranked according to their likeliness of containing the concept." Table I displays some statistics about the development and test collections. Annotations of 346 concepts were provided on the development set in the context of a collaborative work [13] and relevance judgments were provided for 46 of them on the test set. A set of "implies" or "excludes" relations between the 346 concepts was also provided. The evaluation metric is the inferred Average Precision (infAP) on the 46 evaluated concepts which is an estimation of the classical Mean Average Precision (MAP) obtained using the Yilmaz et al. [14] method.

TABLE I. TRECVid SIN 2012 collection.

	Development	Test
Hours of video	∼600	∼200
Number of files	19,701	8,263
Number of shots	400,289	145,634

The baseline system used for the evaluation includes a large number of descriptors as well as associated detection scores made available to TRECVID participants by the IRIM project [15]. A first set of 14 descriptor types has been used in a first set of experiments. These include (see [15] for a detailed description): CEALIST/tlep_576 (color/texture), CEALIST/bov_dsiftSC (pyramidal dense SIFT bag of words), CEALIST/2012_motion1000_tshot (motion), ETIS/vlat_hog3s4-6-8-10_dict64 (VLAT descriptors), EUR/sm462 (saliency moments), INRIA/dense_sift (dense

SIFT bag of words), INRIA/vlad (VLAT descriptors), LABRI/faceTracks (fac tracks statistics),LIF/percepts (percepts, local categories), LIRIS/MFCC_4096 (audio MFCC), LIRIS/OCLBP_DS_4096 (OC-LBP descriptors), LISTIC/SIFT_L2 (filtered SIFT bag of words), LSIS/mlhmslbp_spyr (pyramidal multi-scale features) and MTPT/superpixel_color_sift_k1064 (SIFT on superpixel regions).

As 5 new descriptor types were made available later, they have been inserted in a second set of experiments. These include: ETIS/lab (pyramidal color histograms), ETIS/qw.bin (pyramidal quaternionic wavelets), hg (color histogram and gabor transform), LIG/opp_sift (opponent SIFT bag of words) and LIG/stip (HOG and HOF STIP bag of words). The inclusion of these additional descriptors did not significantly changed the performance of the baseline system, probably because the type of information they convey is redundant with the one conveyed by the already available decriptors.

A kNN/SVM combination was used for the classification stage. A common descriptor pre-processing step is integrated in the classification process; it includes a normalization combined with a PCA-based dimensionality reduction [16]. A hierarchical late fusion approach was used for the last stage [17]. The normalization step for the concept descriptor is simply a gain and offset normalization done separately on each component.

A classifier has been trained for each (descriptor, concept) pair. Their output was normalized using Platt's method [18]. The conceptual descriptor was processed in the same way as descriptors directly extracted from the video signal. The late fusion was done hierarchically by linear combination of the normalized scores. The most similar descriptors were fused first, and more and more dissimilar descriptors or previous combinations were then successively fused [17]. Fusion of similar descriptors was done using uniform weights for getting robustness. Fusion of dissimilar descriptors was done using AP based weights or weights directly optimized by cross-validation within the training set. Fusion with the conceptual descriptor was done at the last stage.

In the baseline system considered here, the similarity between descriptors as well as the sequence of the hierarchical fusions is manually designed according to the type of feature represented in the descriptor. The order is as follows: merge different classifiers kNN and MSVM, merge pyramid levels if applicable, merge different variants of a same descriptor (e.g. same BOW descriptor with different dictionary size), merge different feature types (color, texture, SIFT, percepts, faces...), merge different modalities (visual, audio, text) and finally merge classical descriptors and conceptual feedback). Methods based on measured similarities and automatic clustering could also have been considered but they currently perform less well that the manual one [17].

Results are presented here on the test set only but the tuning of all parameters was done only within the development set by cross-validation. Results on the test set are consistent with results within the development set cross-validation.

A. *Conceptual feedback versus temporal re-scoring.*

In this first experiment, we compare different strategies for combining temporal re-scoring (TRS) with conceptual feedback. Table II shows the infAP performance for various system variants.

TABLE II. COMBINATION OF CONCEPTUAL FEEDBACK AND TEMPORAL RE-SCORING (TRS).

Method	infAP
Signal descriptor fusion (baseline)	0.2600
Concept descriptor	0.2431
Signal and concept fusion	0.2749
Signal and concept fusion with TRS	0.2850
Signal descriptor fusion with TRS	0.2694
Concept with TRS descriptor	0.2679
Signal and concept with TRS fusion	0.2915
Signal and concept with TRS fusion with TRS	0.2959

In the first part (first four lines) of the table, we study the conceptual feedback when applied directly from the baseline system output (without TRS). TRS is then applied on the output of the fusion with feedback as shown in figure 4. The performance of the concept descriptor (0.2431) is lower than the performance of the original baseline fusion from which it is derived (0.2600) indicating that the classifier is unable to do better than replicating the corresponding component in the vector and is unable to find that this would be a better strategy for maximizing the global infAP metric. However, it does capture different and complementary information as this can be seen when it is included in the fusion along with the signal-based descriptor where it leads to a significant improvement (0.2749/0.2600: +5.7% of relative improvement). Posterior TRS produces a further significant improvement (0.2850/0.2749: +3.7%). The gain from posterior TRS is similar to the gain obtained by TRS on the original fusion output (0.2694/0.2600: +3.5%).

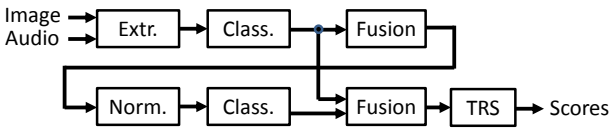

Fig. 4. Conceptual feedback before temporal re-scoring.

In the second part (last four lines) of the table, we study the conceptual feedback when applied after TRS as shown in figure 5. The performance of the concept with TRS descriptor (0.2679) is still lower than the performance of the original

fusion from which it is derived (0.2694) but it is much closer in this case indicating that the conceptual feedback is more efficient after TRS. The concept descriptor also leads to a more significant improvement (0.2915/0.2694: +8.2%). Posterior (second) TRS produces a further but smaller improvement (0.2959/0.2915: +1.5%). This could have been expected since a first pass of TRS is already included in the conceptual feedback.

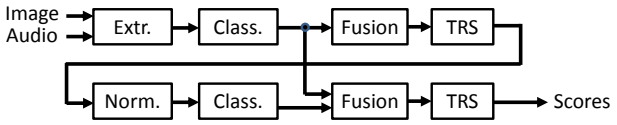

Fig. 5. Conceptual feedback after temporal re-scoring.

B. Iterative conceptual feedback

Once established (by cross-validation) that the conceptual feedback performs better when applied after TRS, we investigate the effect of iterating it several times. Table III shows the system performance after 0 (baseline), 1, 2 and 3 iteration, before and after TRS. The vector of scores used for the concept descriptor is always taken after TRS in the previous iteration.

TABLE III. ITERATIVE CONCEPTUAL FEEDBACK WITH TEMPORAL RE-SCORING (TRS).

Method	by shot	+TRS
Baseline fusion	0.2613	0.2691
Feedback iteration 1	0.2925	0.2981
Feedback iteration 2	0.2984	0.3014
Feedback iteration 3	0.2980	0.3011

The values for the baseline system are slightly different from those in the previous experiments because a few more descriptors were added in the system as they become available later. This slightly increased the performance before TRS and slightly decreased it (probably not significantly) after. This means that the new descriptor did not actually bring signifincant new information and the differences probably comes from added noise. We can observe that there is a performance gain only in the first two iterations. The gain obtained with the second iteration is much smaller than the gain obtained with the first one and no gain (actually a small loss) is obtained from the third one. This probably because most of the information about relation between concepts is extracted in the first iteration, the remainder is extracted in the second, and only noise is amplified in the subsequent ones. The gain provided by TRS also decreases with the iteration. These results are consistent with those obtained by cross-validation within the development set.

The iterative conceptual feedback combined with a TRS provides a total relative gain in MAP of 15.3%. Previous results on TRS showed that the provided gain is greater on low performance systems than on higher performance ones: the better the original system is, the harder it becomes to improve its performance. Though it has not been checked directly, a similar effect is expected for the conceptual feedback and we start here with an already quite good baseline system.

The overall infAP performance of the resulting system on the TRECVid 2012 semantic indexing task is of 0.3014. For comparison, the best official submission conducted in the same conditions had an infAP of 0.2978 (other submissions from the same group reached an infAP of 0.3210 but these used additional and non official annotations). Though the experiments reported here were not officially submitted to TRECVid, they were conducted exactly in the same conditions.

V. CONCLUSION

We have proposed a conceptual feedback approach for improving the performance of a semantic indexing system. This approach can be applied to a system that follows an "extraction/classification/fusion" pipeline for computing detections scores on video shots for a number of concepts and using a number of shot content descriptors. The principle is to build an additional descriptor from the set of scores obtained for the target concepts, to include it in the pool of signal-based descriptors, and to then process it in the same way as them. This approach can be combined with a temporal re-scoring approach by simply applying the latter as a last step in the fusion module before the feedback. The combination of both takes into account both the conceptual and temporal contexts for improving the final detection scores. Finally, the approach can be used iteratively yielding an additional gain. Without the temporal re-scoring, which is specific to video samples, the approach should be applicable to any system based on an "extraction/classification/fusion" pipeline (which is very general) for the simultaneous detection of a set of concepts.

The approach has been evaluated in the context of the TRECVid 2012 semantic indexing task for the simultaneous detection of 346 target concepts. Combined with temporal re-scoring, the proposed method increased the global system performance (MAP) from 0.2613 to 0.3014 (+15.3% or relative improvement) while the temporal re-scoring alone increased it only from 0.2613 to 0.2691 (+3.0%). The overall performance of the resulting system is comparable (even slightly higher) to the performance of the best system officially evaluated at TRECVid 2012 in the same conditions (0.2978). The approach is very easy to implement and, without the temporal re-scoring which is specific to video samples, it could in principle be applicable to any system based on an "extraction/classification/fusion" pipeline (which is very general) for the simultaneous detection of a set of concepts.

The approach could be extended for taking into account simultaneously instead of alternatively the conceptual and

temporal contexts. This could be done by building a concept vector from the scores not only of the current video shot but also from those of the few previous and next shots or from those of the complete video for short ones. A dimensionality reduction technique (e.g. PCA-based) could be applied to keep the resulting vector to a manageable size. Independently, an early fusion scheme could be used for directly mixing the concept vector with the signal vectors for better taking into account the correlation between detected scores and extracted features.

ACKNOWLEDGMENTS

This work was partly realized as part of the Quaero Program funded by OSEO, French State agency for innovation. This work was supported in part by the French project VideoSense ANR-09-CORD-026 of the ANR. Experiments presented in this paper were carried out using the Grid'5000 experimental test bed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr). The authors wish to thanks the participants of the IRIM (Indexation et Recherche d'Information Multimédia) group of the GDR-ISIS research network from CNRS for providing the descriptors used in these experiments.

REFERENCES

[1] C. G. M. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, vol. 4, no. 2, pp. 215–322, 2009.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000. [Online]. Available: http://dx.doi.org/10.1109/34.895972

[3] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of ACM Multimedia*, Nov. 2005.

[4] A. F. Smeaton, P. Over, and W. Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements," in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran, Ed. Berlin: Springer Verlag, 2009, pp. 151–174.

[5] M. Ramesh Naphade, I. V. Kozintsev, and T. S. Huang, "Factor graph framework for semantic video indexing," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 12, no. 1, pp. 40–52, Jan. 2002. [Online]. Available: http://dx.doi.org/10.1109/76.981844

[6] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proceedings of the 15th International Conference on Multimedia 2007, Augsburg, Germany, September 24-29, 2007*, R. Lienhart, A. R. Prasad, A. Hanjalic, S. Choi, B. P. Bailey, and N. Sebe, Eds. ACM, 2007, pp. 17–26.

[7] B. Safadi and G. Quénot, "Re-ranking for Multimedia Indexing and Retrieval," in *ECIR 2011: 33rd European Conference on Information Retrieval*. Dublin, Ireland: Springer, Apr. 2011, pp. 708–711.

[8] ——, "Re-ranking by Local Re-scoring for Video Indexing and Retrieval," in *CIKM 2011: 20th ACM Conference on Information and Knowledge Management*, ser. CIKM '11. Glasgow, Scotland: ACM, Oct. 2011, pp. 2081–2084.

[9] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, M. Wang, and H.-J. Zhang, "Correlative multilabel video annotation with temporal kernels," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 5, no. 1, pp. 3:1–3:27, Oct. 2008.

[10] M.-F. Weng and Y.-Y. Chuang, "Cross-domain multicue fusion for concept-based video indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1927–1941, Oct. 2012. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2011.273

[11] J. R. Smith, M. R. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors." in *ICME*. IEEE, 2003, pp. 445–448.

[12] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quénot, "TRECVID 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[13] S. Ayache and G. Quénot, "Video Corpus Annotation using Active Learning," in *European Conference on Information Retrieval (ECIR)*, Glasgow, Scotland, Mar. 2008, pp. 187–198.

[14] E. Yilmaz and J. A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *Proceedings of the 15th ACM international conference on Information and knowledge management*, ser. CIKM '06. New York, NY, USA: ACM, 2006, pp. 102–111. [Online]. Available: http://doi.acm.org/10.1145/1183614.1183633

[15] N. Ballas, B. Labbé, A. Shabou, H. Le Borgne, P. Gosselin, M. Redi, B. Merialdo, H. Jégou, J. Delhumeau, R. Vieux, B. Mansencal, J. Benois-Pineau, S. Ayache, A. Hamadi, B. Safadi, F. Thollard, N. Derbas, G. Quénot, H. Bredin, M. Cord, B. Gao, C. Zhu, Y. Tang, E. Dellandrea, C.-E. Bichot, L. Chen, A. Benoît, P. Lambert, T. Strat, J. Razik, S. Paris, H. Glotin, T. Ngoc Trung, D. Petrovska Delacrétaz, G. Chollet, A. Stoian, and M. Crucianu, "IRIM at TRECVID 2012: Semantic Indexing and Instance Search," in *Proc. TRECVID Workshop*, Gaithersburg, MD, USA, Nov. 2012.

[16] B. Safadi and G. Quénot, "Descriptor optimization for multimedia indexing and retrieval," in *Proc. of Content Based Multimedia Ingexing (CBMI) Workshop*, Veszprém, Hungary, June 2013.

[17] S. Tiberius Strat, A. Benot, H. Bredin, G. Qunot, and P. Lambert, "Hierarchical late fusion for concept detection in videos," in *ECCV 2012, Workshop on Information Fusion in Computer Vision for Concept Recognition*, Firenze, Italy, Oct. 2012.

[18] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press, 1999, pp. 61–74.