Quaero at TRECVid 2012: Semantic Indexing

Bahjat Safadi[1], Nadia Derbas[1], Abdelkader Hamadi[1], Franck Thollard[1], Georges Quénot[1], Jonathan Delhumeau[2], Hervé Jégou[2], Tobias Gehrig[3], Hazim Kemal Ekenel[3], and Rainer Stifelhagen[3]

[1]UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France
[2]INRIA Rennes / IRISA UMR 6074 / TEXMEX project-team / 35042 Rennes Cedex
[3]Karlsruhe Institute of Technology, P.O. Box 3640, 76021 Karlsruhe, Germany

Abstract

The Quaero group is a consortium of French and German organizations working on Multimedia Indexing and Retrieval[1]. LIG, INRIA and KIT participated to the semantic indexing task and LIG participated to the organization of this task. This paper describes these participations.

For the semantic indexing task, our approach uses a six-stages processing pipelines for computing scores for the likelihood of a video shot to contain a target concept. These scores are then used for producing a ranked list of images or shots that are the most likely to contain the target concept. The pipeline is composed of the following steps: descriptor extraction, descriptor optimization, classification, fusion of descriptor variants, higher-level fusion, and re-ranking. We used a number of different descriptors and a hierarchical fusion strategy. We also used conceptual feedback by adding a vector of classification score to the pool of descriptors. The best Quaero run has a Mean Inferred Average Precision of 0.2692, which ranked us 3^{rd} out of 16 participants. We also organized the TRECVid SIN 2012 collaborative annotation.

1 Participation to the organization of the semantic indexing task

For the third year, UJF-LIG has co-organized the semantic indexing task at TRECVid with the support of Quaero. A list of 500 target concepts has been produced, 346 of which have been collaboratively annotated by the participants and 46 of which have been officially evaluated at TRECVid.

The 500 concepts are structured according to the LSCOM hierarchy [11]. They include all the TRECVid "high level features" from 2005 to 2009, the CU-VIREO374 set plus a selection of LSCOM concepts so that we end up with a number of generic-specific relations among them. We enriched the structure with two relations, namely *implies* and *excludes*. The goal was to promote research on methods for indexing many concepts and using ontology relations between them.

TRECVid provides participants with the following material:

- a development set that contains roughly 600 hours of videos;
- a test set that contains roughly 200 hours of videos;
- shot boundaries (for both sets);
- a set of 500 concepts with a set of associated relations;
- elements of ground truth: some shots were collaboratively annotated. For each shot and each concept, four possibilities are available: the shot has been annotated as positive (it contains the concept), the shot has been annotated as negative (it does not contain the concept), the shot has been skipped (the annotator cannot decide), or the shot has not been annotated (no annotator has seen the shot).

The goal of the semantic indexing task is then to provide, for each of the 346 annotated concepts, a ranked list of 2000 shots that are the most likely to contain the concept. The test collection contains 145,634 shots. A light version of the task has also been proposed in order to facilitate the access to small and/or new groups. More information about the organization of this task can be found in the TRECVid 2012 overview paper [14]. In a *light* version of the task, results have to be given

[1] http://www.quaero.org

only for a subset of 50 out of the 346 annotated concepts. A *pair* version of the task in which 10 pairs of concepts (e.g. Car+Bicycle) has also been proposed this year.

1.1 Development and test sets

Data used in TRECVid are free of right for research purposes as it comes from the Internet Archive (http://www.archive.org/index.php). Table 1 provides the main characteristics of the collection set.

Table 1: Collection feature

Characteristics	TRECVid 2010
#videos	27,964
Duration (total)	∼800 hours
min;max;avg ± sd	11s;270s;132s±93s
# shots	545, 923
# shots (dev)	403, 800
# shots (test)	145, 634

The whole set of videos has been split into two parts, the development set and the test set. Both sets were automatically split into shots using the LIG shot segmentation tool [12].

1.2 The evaluation measure

The evaluation measure used by TRECVid is the MAP (Mean Average Precision). Given the size of the corpus, the inferred MAP is used instead as it saves human efforts and has shown to provide a good estimate of the MAP [13].

1.3 Annotations on the development set

Shots in the development set have been collaboratively annotated by TRECVid 2010-2012 participants and by Quaero annotators. As concepts density is low, an active learning strategy has been set up in order to enhance the probability of providing relevant shots to annotators [2]: the active learning algorithm takes advantage of previously done annotations in order to provide shots that will more likely be relevant. Although this strategy introduces a bias, it raises the number of examples available to systems. Moreover, it exhibits some trend in the concept difficulty. As an example, the number of positive examples for the concept *Person* is larger than the number of negative examples. This means that the active learning algorithm was able to provide more positive examples than negative ones to annotators, meaning that *Person* is probably a "too easy" concept.

346 concepts were annotated on IACC.1.B (tv11 test). 818000 raw annotations were made on these with about 1000 to 2000 annotations for most concepts and up to over 35,000 for a few very infrequent ones so that a minimum number of positive samples can be found for all. After fusion of multiple annotations and propagation using *implies* and *excludes* relations, 1,789,687 new annotations are available on IACC.1.B, which have been added to the 18,936,471 already available from the last two years. In total volume, this amounts to only about 9.5% new ones but these were optimally selected using an active learning approach bootstrapped with a fusion of all the TRECVid SIN 2011 submissions. This also ensures that the active learning based annotation is not biased in favour of the system used for the active learning process. The fused system had a MAP on the tv11 test collection of 0.191 while the best individual system had a MAP of only 0.173 (using the 2011 version of sample_eval). Additionally, an improved algorithm for annotation cleaning has been introduced in the annotation tool this year [10]

1.4 Assessments

46 (resp. 15) concepts were selected for evaluation out of the 346 (resp. 50) ones for which participants were asked to provide results for the full (resp. light) SIN task. Assessments were done partly by NIST (20 concepts) and by Quaero (26 concepts). Assessments were done by visualizing the whole shot for judging whether the target concept was visible or not at any time within the shot. Additionally, all the 10 concept pairs were selected for evaluation. A total of 282,949 concept × shots assessments were made by NIST and Quaero.

2 Participation to the semantic indexing task

2.1 Introduction

The TRECVid 2012 semantic indexing task is described in the TRECVid 2012 overview paper [1, 14]. Automatic assignment of semantic tags representing high-level features or concepts to video segments can be fundamental technology for filtering, categorization, browsing, search, and other video exploitation. New technical issues to be addressed include methods needed/possible as collection size and diversity increase, when the number of features increases, and when features are related by an ontology. The task is defined as follows: "Given the test collection, master shot reference, and concept/feature definitions, return for each feature a list of at most 2000 shot IDs from the test collection ranked according to the possibility of detecting the feature." 346 concepts have been selected for the TRECVid 2012 semantic indexing task. Annotations on the development part of the collections were provided in the context of the collaborative annotation.

As last year, our system uses a six-stages processing pipelines for computing scores for the likelihood of a video shot to contain a target concept. These scores are then used for producing a ranked list of images or shots that are the most likely to contain the target concept. The pipeline is composed of the following steps:

1. Descriptor extraction. A variety of audio, image and motion descriptors have been considered (section 2.2).

2. Descriptor optimization. A post-processing of the descriptors allows to simultaneaously improve their performance and to reduce their size (section 2.3).

3. Classification. Two types of classifiers are used as well as their fusion (section 2.4).

4. Fusion of descriptor variants. We fuse here variations of the same descriptor, e.g. bag of word histograms with different sizes or associated to different image decompositions (section 2.5).

5. Higher-level fusion. We fuse here descriptors of different types, e.g. color, texture, interest points, motion (section 2.6).

6. Re-ranking. We post-process here the scores using the fact that videos statistically have an homogeneous content, at least locally (section 2.7).

Additionally, our system includes a conceptual feedback in which a new descriptors is built using the prediction scores on the 346 target concepts is added to the already available set of 47 audio and visual descriptors (section 2.8). Compared to last year, our system has been improved by the inclusion of new descriptors, an improved desctiptor-classifier joint optimization and an improved scheme for hierarchical late fusion.

2.2 Descriptors

A total of 127 audio and visual descriptors have been used. Many of them have been produced by and shared with the IRIM consortium. These include variants of a same descriptors (e.g. same methods with different histogram size or image decomposition). These descriptors do not cover all types and variants but they include a significant number of different approaches including state of the art ones and more exploratory ones. They are described and evaluated in the IRIM consortium paper [8]. They include color histogram, Gabor transform, quaternionic wavelets, a variety of interest points descriptors (SIFT, color SIFT, SURF, STIP), local edge patterns, saliency moments, percepts, and spectral profiles for audio description. Many of them rely on a bag of words approach.

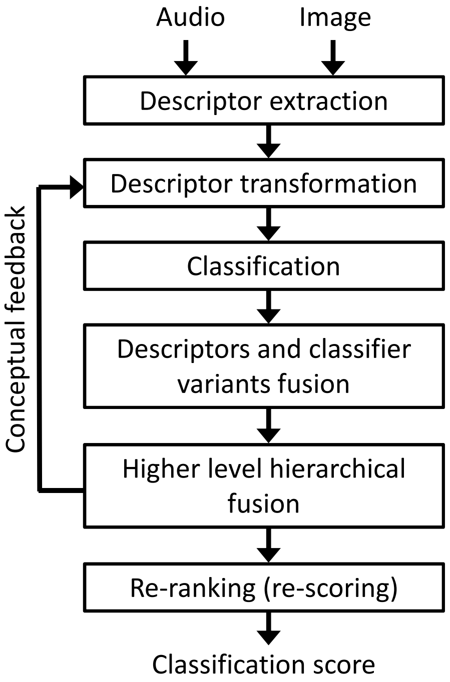

Figure 1: Semantic indexing system

2.3 Descriptor optimization

The descriptor optimization consists of two steps: power transformation and principal component analysis (PCA).

2.3.1 Power transformation

The goal of the power transformation is to normalize the distributions of the values, especially in the case of histogram components. It simply consists in applying an $x \leftarrow x^\alpha$ ($x \leftarrow -(-x)^\alpha$ if $x < 0$) transformation on all components individually. The optimal value of α can be optimized by cross-validation and is often close to 0.5 for histogram-based descriptors.

The optimization of the value of the α coefficient is optimized by two-fold cross-validation within the development set. It is done in practice only using the LIG_KNNB classifier (see section 2.4) since it is much faster when a large number of concepts (346 here) has to be considered and since it involves a large number of combinations to be evaluated. Trials with a restricted number of varied descriptors indicated that the optimal values for the kNN based classifier are close to the ones for the multi-SVM based one. Also, the overall performance is not very sensitive to the precise values for this hyper-parameter.

2.3.2 Principal component analysis

The goal of PCA reduction is both to reduce the size (number of dimensions) of the descriptors and to improve performance by removing noisy components.

The number of components kept in the PCA reduction is also optimized by two-fold cross-validation within the development set using the LIG_KNNB classifier. Also, the overall performance is not very sensitive to the precise values for this number.

2.4 Classification

The LIG participant ran two types of classifiers on the contributed descriptors as well as their combination.

LIG_KNNB: The first classifier is kNN-based. It is directly designed for simultaneously classifying multiple concepts with a single nearest neighbor search. A score is computed for each concept and each test sample as a linear combinations of 1's for positive training samples and of 0's for negative training samples with weights chosen as a decreasing function of the distance between the test sample and the reference sample. As the nearest neighbor search is done only once for all concepts, this classifier is quite fast for the classification of a large number of concepts. It is generally less good than the SVM-based one but it is much faster.

LIG_MSVM: The second one is based on a multiple learner approach with SVMs. The multiple learner approach is well suited for the imbalanced data set problem [5], which is the typical case in the TRECVid SIN task in which the ration between the numbers of negative and positive training sample is generally higher than 100:1.

LIG_ALLC: Fusion between the two available classifiers. The fusion is simply done by averaging the classification scores produced by the two classifiers. Their output is naturally or by designed normalized in the the [0:1] range. kNN computation is done using the KNNLSB package [6]. Even though the LIG_MSVM classifier is often significantly better than the LIG_KNNB one, the fusion is most often even better, probably because they are very different and capture different things.

2.5 Performance improvement by fusion of descriptor variants and classifier variants

In a previous work, LIG introduced and evaluated the fusion of descriptor variants for improving the performance of concept classification. We previously tested it in the case of color histograms in which we could change the number of bins, the color space used, and the fuzziness of bin boundaries. We found that each of these parameters had an optimal value when the others are fixed and that there is also an optimal combination of them which correspond to the best classification that can be reached by a given classifier (kNN was used here) using a single descriptor of this type. We also tried late fusion of several variants of non-optimal such descriptors and found that most combinations of non-optimal descriptors have a performance which is consistently better than the individual performance of the best descriptor alone. This was the case even with a very simple fusion strategy like taking the average of the probability scores. This was also the case for hierarchical late fusion. In the considered case, this was true when fusing consecutively according to the number of bins, to the color space and to the bin fuzziness. Moreover, this was true even if some variant performed less well than others. This is particularly interesting because descriptor fusion is known to work well when descriptors capture different aspects of multimedia content (e.g. color and texture) but, here, an improvement is obtained using many variants of a single descriptor. That may be partly due to the fact that the combination of many variant reduces the noise. The gain is less than when different descriptor types are used but it is still significant.

We have then generalized the use of the fusion of descriptor variants and we evaluated it on other descriptors and on TRECVid 2010. We made the evaluation on descriptors produced by the ETIS partner of the IRIM group. ETIS has provided 3 × 6 variants of two different descriptors (see the previous section). Both these descriptors are histogram-based. They are computed with four different number of bins: 64, 128, 192, 256, 512 and 1024; and with three image decomposition: 1x1 (full image), 1x3 (three vertical stripes) and 2x2 (2 by 2 blocks). Hierarchical fusion is done according to three levels: number of bins, “pyramidal” image decomposition and descriptor type.

We have evaluated the results obtained for fusion within a same descriptor type (fusion levels 1 and 2) and between descriptor types (fusion level 3) [7]. The fusion of the descriptor variants varies from about 5 to 10% for the first level and is of about 4% for the second level. The gain for the second level is relative to the best result for the first level so both gains are cumulated. For the third level, the gain is much higher as this could be expected because, in this case, we fuse results from different information sources. The gain at level 3 is also cumulated with the gain at the lower levels.

2.6 Final fusion

Hierarchical fusion with multiple descriptor variants and multiple classifier variants was used and optimized for the semantic indexing task. We made several experiment in order to evaluate the effect of a number of factors. We optimize directly the first levels of the hierarchical fusion using uniform or average-precision weighting. The fusion was made successively on variant of the same descriptors, on variant of classifiers on results from the same descriptors, on different types of descriptors and finally on the selection of groups of descriptors.

2.7 Re-ranking

Video retrieval can be done by ranking the samples according to their probability scores that were predicted by classifiers. It is often possible to improve the retrieval performance by re-ranking the samples. *Safadi and Quénot* in [9] propose a re-ranking method that improves the performance of semantic video indexing and retrieval, by re-evaluating the scores of the shots by the homogeneity and the nature of the video they belong to. Compared to previous works, the proposed method provides a framework for the re-ranking via the homogeneous distribution of video shots content in a temporal sequence. The experimental results showed that the proposed re-ranking method was able to improve the system performance by about 18% in average on the TRECVid 2010 semantic indexing task, videos collection with homogeneous contents. For TRECVid 2008, in the case of collections of videos with non-homogeneous contents, the system performance was improved by about 11-13%.

2.8 Conceptual feedback

Since the TRECVid SIN 2012 task considers a quite large number (346) of descriptors and since these are also organized according to a hierarchy, one may expect that the detection scores of some concept help to imrpove the detection score of related concepts. We have made a number of attempts to use the explicit *implies* or *excludes* provided relations but these were not successful so far, maybe due to a normalization problem between the scores of the different concepts. We tried then an alternative approach using the implicit relations between concepts by creating a vector with the classification scores of all the available concepts. We used for that the best hierarchical fusion result available. This vector of scores was then included as a 128^{th} one in the pool of the 127 already available descriptors and processed in the same way as the others, including the power and PCA optimization steps and the fusion of classifier outputs. The found optimal power value was quite different of the ones for the other descriptors (1.800 versus 0.150-0.700) for the other ones. This is probably linked with the way the score normalization is performed.

2.9 Performances on the semantic indexing task

Four slightly different combinations of hierarchical fusion have been tried. The variations concerned the way the re-ranking was done: it can be locally temporal, globally temporal and or conceptual. Not all combinations could be submitted and the following were selected:

F_A_Quaero1_1: post-processed version of Quaero2 with a temporal re-ranking method that attempts to exploit the assumed homogeneity of video documents at a global level;

F_A_Quaero2_2: baseline yet already complex run with a manually built hierarchical fusion of a large number (over 100) of jointly optimized descriptor-classifier combinations including also a conceptual feedback;

F_A_Quaero3_3: post-processed version of Quaero2 with a temporal re-ranking method that attempts to exploit the assumed homogeneity of video documents at a local level;

F_A_Quaero4_4: post-processed version of Quaero2 with a temporal and conceptual re-ranking method that attempts to exploit the assumed temporal homogeneity of video documents and the observed co-occurrence of concepts.

Quaero2 is a baseline and Quaero1, 3 and 4 are contrastive runs aiming at studying the effect of different post-processing (re-ranking) methods exploiting the assumed temporal coherency of video documents and/or the co-occurrences between detected concepts.

Table 2 shows the performance of the four submitted variants. Our submissions ranked between 8 and 12 in a total of 68 for the full SIN task. Our best submission ranked us as the third group out of 19 for the full SIN task. The improvement brought by the conceptual feedback is quite small and less than what was expected from cross-validation within the development set but it is significant. The hierarchical fusion performs better than the flat one and the optimization of

Table 2: InfAP result and rank on the test set for all the 46 TRECVid 2012 evaluated concepts

System/run	MAP	rank
Best submission	0.3210	1
F_A_Quaero1_1	0.2692	8
F_A_Quaero4_4	0.2536	9
F_A_Quaero3_3	0.2534	10
F_A_Quaero2_2	0.2486	11
Median submission	0.1944	26

the fusion weights by cross-validation performs better than the MAP-based or uniform method.

3 Acknowledgments

This work was partly realized as part of the Quaero Programme funded by OSEO, French State agency for innovation. Some results from the IRIM network were also used in these experiments.

References

[1] A. Smeaton, P. Over and W. Kraaij, Evaluation campaigns and TRECVid, In *MIR'06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp321-330, 2006.

[2] Stéphane Ayache and Georges Quénot. Video Corpus Annotation using Active Learning, In *30th European Conference on Information Retrieval (ECIR'08)*, Glasgow, Scotland, 30th March - 3rd April, 2008.

[3] S. Ayache, G. Quénot, J. Gensel, and S. Satoh. Using topic concepts for semantic video shots classification. In Springer, editor, *CIVR – International Conference on Image and Video Retrieval*, 2006.

[4] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. A comparison of color features for visual concept classification. In *ACM International Conference on Image and Video Retrieval*, pages 141–150, 2008.

[5] B. Safadi, G. Quénot. Evaluations of multi-learners approaches for concepts indexing in video documents. In *RIAO,* Paris, France, April 2010.

[6] Georges Quénot. *KNNLSB: K Nearest Neighbors Linear Scan Baseline*, 2008. Software available at `http://mrim.imag.fr/georges.quenot/freesoft/knnlsb/index.html`.

[7] D. Gorisse et al., IRIM at TRECVid 2010: High Level Feature Extraction and Instance Search. In *TREC Video Retrieval Evaluation workshop*, Gaithersburg, MD USA, November 2010.

[8] Delezoide et al. IRIM at TRECVid 2012: Semantic Indexing and Multimedia Instance Search, In *Proceedings of the TRECVid 2012 workshop*, Gaithersburg, USA, 26-28 Nov. 2012.

[9] B. Safadi, G. Qunot. Re-ranking by Local Rescoring for Video Indexing and Retrieval, *CIKM 2011: 20th ACM Conference on Information and Knowledge Management,* Glasgow, Scotland, oct 2011.

[10] B. Safadi, S. Ayache, G. Qunot. Active Cleaning for Video Corpus Annotation. *International MultiMedia Modeling Conference,* 7131:518-528, Klagenfurt, Austria, jan 2012. Glasgow, Scotland, oct 2011.

[11] Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13:86–91, 2006.

[12] Georges Quénot, Daniel Moraru, and Laurent Besacier. CLIPS at TRECvid: Shot boundary detection and feature detection. In *TRECVid'2003 Workshop*, Gaithersburg, MD, USA, 2003.

[13] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR*. ACM 978-1-60558-164-4/08/07, July 2008.

[14] P. Over, G. Awad, J. , B. Antonishek, M.2Michel, A. Smeaton, W. Kraaij, and G. Quénot, TRECVid 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics In *Proceedings of the TRECVid 2012 workshop*, Gaithersburg, USA, 26-28 Nov. 2012.