# LIG at TRECVID 2009: Hierarchical Fusion for High Level Feature Extraction

Bahjat Safadi and Georges Quénot
Laboratoire d'Informatique de Grenoble (LIG)
385 rue de la Bibliothèque - BP 53 - 38041 Grenoble - Cedex 9 - France

## Abstract

We investigated in this work a hierarchical fusion strategy for fusing the outputs of hundreds of descriptors × classifier combinations. Over one hundred descriptors gathered in the context of the IRIM consortium were used for HLF detection with up to four different classifiers. The produced classification scores are then fused in order to produce a unique classification score for each video shot and HLF. In order to cope with the redundancy of the information obtained from similar descriptors and from different classifiers using them, we propose a hierarchical fusion approach so that 1) each different source type gets an appropriate global weight, 2) all the descriptors × classifier combinations from similar source type are first combined in the optimal way before being merged at the next level. The best LIG run has a Mean Inferred Average Precision of 0.1276, which is significantly above TRECVID 2009 HLF detection task median performance. We found that fusion of the classification scores from different classifier types improves the performance and that even with a quite low individual performance, audio descriptors can help.

## 1 Introduction

The classical approach for concept classification in images or video shots is based on a three-stage pipeline: descriptors extraction, classification and fusion. In the first stage, descrifptors are extracted from the raw data (video, image or audio signal). Descriptors can be extracted in different ways and from different modalities. In the second stage, a classification score is generated from each descriptor and, for each image or shot, and for each concept. In the third stage, a fusion of the classification scores obtained from the different descriptors is performed in order to produce a global score for each image or shot and for each concept. This score is generally used for producing a ranked list of images or shots that are the most likely to contain a target concept.

We investigated in this work a hierarchical fusion strategy for fusing the outputs of hundreds of descriptors × classifier combinations. Over one hundred descriptors gathered in the context of the IRIM consortium were used for HLF detection with up to four different classifiers. The produced classification scores are then fused in order to produce a unique classification score for each video shot and HLF. In order to cope with the redundancy of the information obtained from similar descriptors and from different classifiers using them, we propose a hierarchical fusion approach so that 1) each different source type gets an appropriate global weight, 2) all the descriptors × classifier combinations from similar source type are first combined in the optimal way before being merged at the next level.

## 2 Descriptors and classifiers

The IRIM consortium of the ISIS "Groupe De Recherche" (GDR) from CNRS leaded by LIG, IRIT, LABRI and LIP6 has produced and evaluated a large number of image, motion and audio descriptors for video shot classification [3]. These descriptors were evaluated in the context of TRECVID 2008 and 2009 High Level Features (HLF) detection task [1]. These HLFs are actually concepts, objects or events to be detected in video shots.

Twelve IRIM participants (CEA-LIST, ETIS, Eurecom, GIPSA, IRIT, LABRI, LEAR, LIF, LIG, LIP6, LSIS and XLIM-SIC) provided descriptors and three participants (LIF, LIG ans ETIS) provided classification results using them allowing for comparing the relative performances of these descriptors. 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. Three IRIM participants evaluated these descriptors using a total of four different classifiers. The evaluations were conducted on TRECVID 2008 concepts annotated on the TRECVID 2007 collection (which is the trec2008 development collection). The training and evaluation were done respectively on the development and test parts of the TRECVID 2007 collection. More infor-

mation about these descriptors and evaluations can be found in [3, 2]

Previous experiments have shown that combining many weak classifiers can produce a strong classifier, that using classifiers based on very different principles can be very efficient and that even classifiers with a poor individual performance can positively contribute to a global classifier, especially if they can capture something which is not captured by others. Therefore, any of the above evaluated classifier with a performance significantly higher than the random one can be useful and should be considered in the fusion process.

# 3 Hierarchical fusion

We have made a lot of experiments for evaluating various fusion strategies and try to obtain the best classification performance using the available set of descriptors. Many of these experiments only involved image descriptors but the organization of the experiments and evaluation procedure was the same when motion and audio descriptors were used as well and this is why we present the experiments and results in this section. One can also consider different ways of looking at images (like color, texture or SIFTs) as different modalities. From the fusion point of view, this does not make a significant difference.

We again used the two parts (dev and test) of the TRECVID 2007 video collection for training and validation but we used TRECVID 2009 HLFs (concepts) in this case. We conducted most experiments on late fusion. The fusion parameters were tuned using the classification scores of the individual descriptors using three classifiers LIF_SVM, LIG_KNNC and LIG_KNNG.

We do not display here all the results of these experiments. We only explain the type of experiments we conducted and the general conclusion that we obtained from them. Finally, we display the results that we obtained from our official submissions at TRECVID 2009. These submissions were based on the best strategies that we found in our fusion experiments. We also explored some variants that were not expected to lead to the best performance in order to evaluate the effect of various parameters.

We performed a few experiments on early and late fusion. It turned out that sometimes early fusion was better and sometimes late fusion was better. Considering this and the fact that late fusion is much easier to implement, we conducted the next experiments by the means of the late fusion.

We compared various late fusion methods, including: weighted sums and products, max, min and harmonic-, geometric- or arithmetic-mean based rank fusions. Again, it turned out that none of these strategies has a clear advantage once the prediction scores from the individual classifiers are properly normalized. The relative weighting of the different classifiers in the global combination is much more important.

Several methods can be used for the weighting of the classifiers. A uniform weighting is quite often a good choice because everything else tends to overfit the data and to generalize poorly. Another good choice is a weighting based on the individual performance of the classifier, evaluated by cross-validation. A third possibility is to globally optimize the weight for maximizing the global performance evaluated again by cross-validation. All these methods can lead to an overfit, especially if they are applied separately for each concept.

Another important aspect is the selection of the classifier that will be used for the global fusion. One selection criterion could be the individual performance of the classifier but this is already somehow handled by the weighting schemes. The main problem is the presence of a large number of descriptors that capture something similar with a consistent quality while a small number of descriptors capture something different. The descriptor types that are the most represented tend to dominate in the global system and mask the contribution of the least represented. In order to solve this problem, we proposed a hierarchical fusion based on some heuristics. All descriptors of the same type are first fused together. The results of their fusion are the merged with similar weights. This is done by variant, by type, by classification engine and by modality. Several corresponding strategies were tried and validated within the TRECVID 2007 collection. We obtained the following results:

- The hierarchical fusion can do better than all the flat strategies if properly organized.
- Fusion of classifier outputs using different variants (e.g. dictionary size) usually do slightly better than any single variant.
- Fusion of classifier outputs using different classification engines usually do slightly better than that of any single variant.
- The better strategy seems to fuse elements in the following order: descriptor variants, descriptor types, classification engine types and finally modalities though the order in the last levels is less important.

Table 1: Official TRECVID 2009 submissions and results

| Run | MAP | Description |
|---|---|---|
| A_LIG_RUN1_1 | 0.1269 | Late fusion of runs A_LIG_RUN3_3 and A_LIG_RUN6_6 |
| A_LIG_RUN2_2 | **0.1276** | Late fusion of runs A_LIG_RUN4_4 and A_LIG_RUN6_6 |
| A_LIG_RUN3_3 | 0.1047 | Late fusion of run A_LIG_RUN4_4 plus face detection |
| A_LIG_RUN4_4 | 0.1042 | Late fusion of run A_LIG_RUN5_5 plus audio features |
| A_LIG_RUN5_5 | 0.1002 | Late fusion of KNN on various visual features |
| A_LIG_RUN6_6 | 0.1165 | Late fusion of SVM on visual and audio features plus face detection |

# 4 Official TRECVID 2009 submissions and results.

We submitted six runs. These runs are described in Table 1. The run names include a priority number corresponding to our prediction of performance from the best to the worse:

- RUN5 is a baseline run. It is a hirarchical combination of KNN scores from almost all the available visual features including motion ones but excluding those based on face detection.
- RUN4 is a combination of RUN5 and of KNN scores from audio features.
- RUN3 is a combination of RUN4 and of KNN scores from features based on face detection.
- RUN6 is a flat combination of SVM scores from heuristically selected features from all categories, including audio features and features based on face detection.
- RUN2 is a combination of runs RUN4 and RUN6.
- RUN1 is a combination of runs RUN3 and RUN6.

RUN5, RUN4 and RUN3 combines scores obtained with a KNN classifier. RUN6 combines scores obtained with a SVM classifier. RUN1 and RUN2 combines scores obtained with both a KNN classifier and a SVM classifier. Il all cases, the weights of the various components were determined by cross-validation on the development collection.

The best LIG submission (RUN2) has a performance of 0.1276 while the best performance was of 0.2285 and the median performance was of 0.0516. The results confirmed that:

- Hierarchical fusion is a good way of choosing weights.
- Fusion of the classification scores from different classifier types improves the performance: from 0.1042 (RUN4) and 0.1165 (RUN6) to 0.1276 (RUN2).
- Even with a quite low individual performance, audio descriptors can help: from 0.1002 (RUN5) to 0.1042 (RUN4) with a performance of only 0.0213.
- Face detection does not help much globally if at all. The slight difference between RUN3 and RUN4 and between RUN1 and RUN2 is not stable and probably not statistically significant. This is at the global level however, for some individual concepts (HLFs), this might be different.

# 5 Acknowledgments

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. This work was partly realized in the context of the IRIM (Indexation et Recherche d'Information Multimédia) of the GDR-ISIS research group from CNRS.

# References

[1] Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330.

[2] Bredin H., Koenig L., Lachambre H. and El Khoury E., IRIT @ TRECVid HLF 2009 – Audio to the Rescue, In TREC2009 notebook, Gaithersburg, USA, 16-17 Nov. 2009.

[3] Quénot G. et al., IRIM at TRECVID 2009: High Level Feature Extraction, In TREC2009 notebook, Gaithersburg, USA, 16-17 Nov. 2009.