Quaero at TRECVID 2011: Semantic Indexing and Multimedi Event Detection (draft)

Bahjat Safadi[1], Nadia Derbas[1], Abdelkader Hamadi[1], Franck Thollard[1], Georges Quénot[1], 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 and KIT participated to the semantic indexing task and LIG participated to the organization of this task. LIG also participated to the multimedia event detection 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 0.1529, which ranked us 3^{rd} out of 19 participants.

We participated to the multimedia event detection task with a system derived from the generic one we have for general purpose concept indexing in videos considering the target events as concepts. Detection scores on videos are produced from the scores on shots.

1 Participation to the organization of the semantic indexing task

For the second 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 50 of which have been officially evaluated at TRECVID.

The 500 concepts are structured according to the LSCOM hierarchy [10]. 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 400 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

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

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 137,327 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 2011 overview paper [13]

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	19856
Duration (total)	∼600 hours
min;max;avg ± sd	11s;1h;132s±93s
# shots	403, 800
# shots (dev)	266, 473
# shots (test)	137, 327

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 [11].

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 [12].

1.3 Annotations on the development set

Shots in the development set have been collaboratively annotated by TRECVID 2010 participants. 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.

A total of about 4.2 M single concept × shots annotations were made, of which about 0.9 M by Quaero, about 2.2 M by the TRECVID 2010 participants and about 1.1M by the TRECVID 2011 participants. Among these, about 87% were done at least once, about 9% were done at least twice and about 3% were done three or more times. The multiple annotations were selected by the active learning tool as those being the more likely to correspond to errors or ambiguities and were made for cleaning as much as possible the annotations made. The resulting 4.2 M annotations were amplified by the use of relations between concepts to about 18 M usable annotations. The relation used included the "implies" and "excludes" relations. These ∼18 M annotations represent about 13% of all the possible annotations on the development set. These have been selected by an active learning procedure that makes them almost as efficient as if the whole annotation was performed [2].

1.4 Assessments

50 (resp. 23) 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 (30 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. A total of 268156 concept × shots assessments were made by NIST and Quaero.

2 Participation to the semantic indexing task

2.1 Introduction

The TRECVID 2011 semantic indexing task is described in the TRECVID 2011 overview paper [1, 13]. 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 2011 semantic indexing task. Annotations on the development part of the collections were provided in the context of the collaborative annotation.

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).

2.2 Descriptors

A total of 47 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.

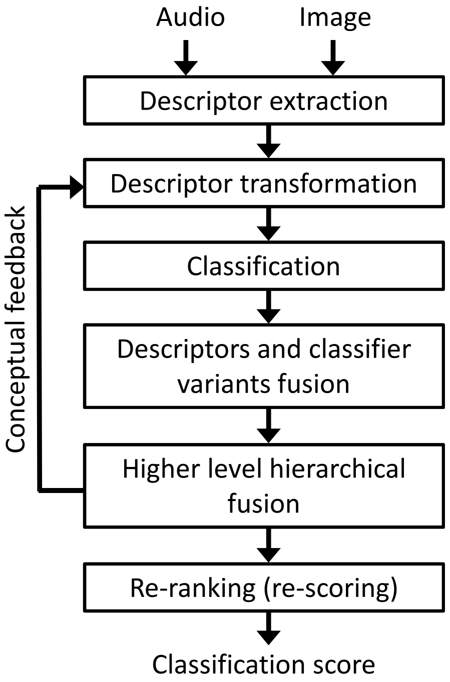

Figure 1: Semantic indexing system

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 × 4 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 and 256; 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 2011 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 48^{th} one in the pool of the 47 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 (1.750 versus 0.150-0.700) for the other ones. This is probably linked with the way the score normalization is performed.

Table 2: Cross-validation performance without and with conceptual feedback, with and without reranking

System	Fusion	Rerank
Original fusion	0.1666	0.1833
Concepts descriptors	0.1144	
Fusion with concepts	0.1697	0.1864

Table 2 shows the effect of including the concepts descriptor in the fusion process. Even though the performance of the descriptor alone is significantly less than the fusion, it can still yield a slight improvement.

2.9 Performances on the semantic indexing task

Four slightly different combinations of hierarchical fusion have been tried. The variations concerned the way the fusion was done: it can be flat or hierarchical, the weighting of components can be uniform, MAP-based or optimized by cross-validation. Not all combinations could be submitted and the following were selected:

F_A_Quaero1_1: Optimized hierarchical combination of all available descriptor/classifier combinations including the concept score feedback descriptor;

F_A_Quaero2_2: Optimized hierarchical combination of all available descriptor/classifier combinations excluding the concept score feedback descriptor;

F_A_Quaero3_3: Flat and uniform combination of available descriptor/classifier combinations excluding the concept score feedback descriptor;

F_A_Quaero4_4: MAP weighted combinations of all available descriptor/classifier combinations including the concept score feedback descriptor.

Table 3 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 the fusion weights by cross-validation performs better than the MAP-based or uniform method.

3 Participation to the multimedia event detection task

The TRECVID multimedia event detection (MED) task is defined as follows: "Given a collection of test videos and a list of test events, indicate whether each of the test events is present anywhere in each of the test videos and give the strength of evidence for each such judgment."

We participated to this task with a system derived from the generic one we have for general purpose concept indexing in videos considering the target events as concepts. As our system is designed for indexing concepts into shots and not within whole videos, we first split all videos into shots. Then, for training, all shots from a positive (resp. negative) video is considered as positive (resp. negative) for the target event. For predicting, a score is computed for all shots within a video and a global score for the video is computed from the shot scores. We explored visual and audio descriptors using different classifiers: kNN, MSVM, Random Forest.

The system is represented in the figure 2. It includes the following steps: descriptor extraction, descriptor transformation, classification, shot-level fusion and video-level fusion. In the next sections, we detail each step of this system. Finally we will show the obtained experimental results.

We did all the system development or adaptation using the MAP metric computed with the trec_eval tool for the evaluation. Though it is significantly different from the MED task one, it is much easier to manage during the development and related enough to the task one for the resulting optimizations to be appropriate.

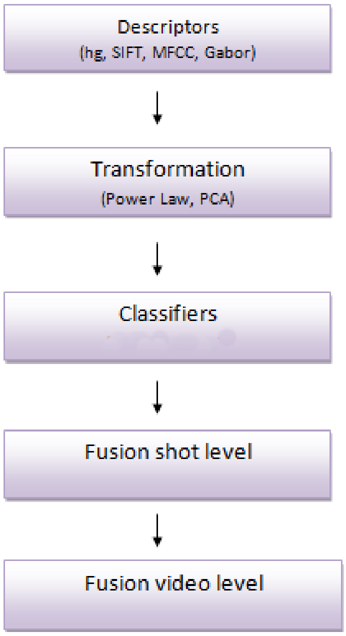

Figure 2: Differents steps of the proposed system

3.1 Descriptors

We used a combination of visual and audio descriptors.

3.1.1 Visual descriptors

For the description of the image track, we used a subset of the visual descriptors shared with the IRIM group [8], including color, texture and interest points:

LIG/h3d64: normalized RGB Histogram 4 × 4 × 4 ⇝ 64 dimensions.

LIG/gab40: normalized Gabor transform, 8 orientations × 5 scales, ⇝ 40 dimensions.

LIG/hg104: early fusion (concatenation) of h3d64 and gab40 ⇝ 104 dimensions.

LIG/opp_sift_<method>[_unc]_1000: bag of word, opponent sift, generated using Koen Van de Sande's software[4] ⇝ 1000 dimensions (384 dimensions per detected point before clustering; clustering on 535117 points coming from 1000 randomly chosen images). **<method>** method is related to the way by which SIFT points are selected: **har** corresponds to a filtering via a

Table 3: InfAP result and rank on the test set for all the 50 TRECVID 2011 evaluated concepts

System/run	MAP	rank
Best submission	0.1731	1
F_A_Quaero1_1	0.1529	8
F_A_Quaero2_2	0.1509	9
F_A_Quaero3_3	0.1497	11
F_A_Quaero4_4	0.1487	12
Median submission	0.1083	34

Harris-Laplace detector and **dense** corresponds to a dense sampling; the versions with **_unc** correspond to the same with fuzziness introduced in the histogram computation.

3.1.2 Audio descriptors

For the audio description, we used a bag of word approach on MFCC vectors.

LIG/mfcc256CB: Bag of word of MFCC (Mel Frequency Cepstral Coefficients), extracted each 10 ms of each video ⇒ 256 dimensions for each extracted vector (12 dimensions for each vector before the clustering).

LIG/mfcc256CB_delta_acc: Bag of word of MFCC coefficients with delta and delta-2 coefficients, extracted each 10 ms of each video ⇒ 256 dimensions for each extracted vector (36 dimensions for each vector before the clustering).

3.2 Descriptor optimization

Two transformations of the original descriptor are considered: power Law and PCA. The first one, power law, is a $x \leftarrow x^{\alpha}$ transformation applied to the vector components (or to its absolute value if negative); it is well suited for histograms where a 0.500 exponent value tend to make the Euclidean distance close to the Chi square one but it appear to improve the performance also for other types of descriptors. The α value can be seen as one of the hyper-parameters of the classification systems and results are generally reported here only for the optimal value found by cross-validation.

The second one, PCA reduction, is performed and only the N components with the highest variance are kept; this reduces the vector size and classification cost with a number of cost speed compromises; generally, the optimal value leads both to a significant vector size reduction and to a (slight) performance improvement.

3.3 Classification

We test three classifiers, the kNN, the Random Forest and the MSVM. kNN is a method of classifying based on closest training examples in the feature space. It is a well known statistical approach applied in many systems and shown good performance. The Random forest (RF) is a method based on decision trees. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Multi-learner SVM (MSVM) is an improved version of SVM classifier; it is a combination of Active Learning with Support Vector Machine with RBF kernel. Multi-learner approaches are designed to handle the problem of the sparse concepts that leads to a strong imbalance between the size of positive and negative sample sets.

We report in tables 4,5 and 6 the classification scores obtained by the different systems with the above described descriptors on the development data collection given for MED2011(DEVT). The scores are obtained with the trec_eval tool provided by TRECVID. They are expressed globally for the 15 events given for MED2011 task with the MAP metrics. The MAP is computed at the video level using the maximum of the shot scores within a video as the video score. The MAP values are computed in the context of "one-fold cross validation": the development set is split into two parts; training is done on one part and prediction and evaluation is done on the second part.

These experiments show that the results obtained by kNN, MSVM and Random forest are very comparable. Although, for the sift descriptors MSVM outperforms KNN and Random forest. Therefore, we decide to keep kNN and MSVM, for the following experiments. As well, we keep only the best descriptors transformations.

Table 4: Results with normalized descriptors

Descriptor	Dims	LIG_KNN	WEKA_RF
LIG/h3d64	64	0.0793	0.1112
LIG/gab40	40	0.1132	0.1084
LIG/hg104	104	0.1132	0.1324
LIG/opp_sift_har_1000	1000	0.1138	0.1144
LIG/opp_sift_dense_1000	1000	0.1026	0.1139
LIG/opp_sift_har__unc_1000	1000	0.1168	0.1150
LIG/opp_sift_dense_unc_1000	1000	0.1029	0.1191
LIG/mfcc256CB	256	0.0875	0.0986
LIG/mfcc256CB_delta_acc	256	0.0819	0.0873

Table 5: Results with normalized descriptors after power transformation

Descriptor	exp.	Dims	LIG_KNN	WEKA_RF	LIG_MSVM
LIG/h3d64	0.300	64	0.1168	0.1017	0.1313
LIG/gab40	0.500	40	0.1149	0.1037	0.1142
LIG/hg104	0.300	104	0.1325	0.1298	0.1750
LIG/opp_sift_har_1000	0.450	1000	0.1136	0.1104	-
LIG/opp_sift_dense_1000	0.450	1000	0.1063	0.1196	-
LIG/opp_sift_har_unc_1000	0.300	1000	0.1188	0.1150	-
LIG/opp_sift_dense_unc_1000	0.450	1000	0.1225	0.1219	-
LIG/mfcc256CB	0.400	256	0.1109	0.0932	0.1044
LIG/mfcc256CB_delta_acc	0.300	256	0.0975	0.0887	0.1119

Table 6: Results with normalized descriptors after PCA reduction

Descriptor	exp.	Dims	LIG_KNN	WEKA_RF	LIG_MSVM
LIG/h3d64	0.300	32	0.1151	0.1039	0.1285
LIG/gab40	0.500	30	0.1144	0.1090	0.1133
LIG/hg104	0.300	52	0.1320	0.1309	0.1720
LIG/opp_sift_har_1000	0.450	150	0.1583	0.1146	0.2248
LIG/opp_sift_dense_100	0.450	200	0.1300	0.1025	0.1996
LIG/opp_sift_har_unc_1000	0.300	200	0.1543	0.0967	0.2346
LIG/opp_sift_dense_unc_1000	0.450	250	0.1287	0.0713	0.1968
LIG/mfcc256CB	0.400	96	0.1114	0.0941	0.1029
LIG/mfcc256CB_delta_acc	0.300	48	0.1088	0.0948	0.1075

3.4 Shot-level fusion

Once the descriptors are computed and classification scores obtained at the shot level, we merge the prediction scores obtained for all shots of a given video by a simple linear combination of scores for various combinations of descriptors and/or scores. This fusion is done hierarchically:

LIG_hg104: late fusion of LIG/h3d64, LIG/gab40 and LIG/hg104;

LIG_sift4: late fusion of LIG/opp_sift_*_1000;

LIG_mfcc: late fusion of LIG/mfcc256CB*;

LIG_hgsift: late fusion of LIG_hg104 and LIG_sift4;

LIG_all: late fusion of LIG_hgsift and LIG_mfcc.

Table 7 shows the performance obtained with these fusions using two-fold cross-validation within the development set. Results are given for the fusion of classification scores obtained using the KNN and MSVM classifiers separately, and by fusing also the scores from both classifiers. Each level of fusion improves over any of its elements, including the first one which is actually a combination of early and late fusions.

Table 7: Results for shot fusion with KNN, MSVM and both classifiers, two-fold cross-validation

Descriptor	KNN	MSVM	both
LIG_hg104	0.1277	0.1595	0.1676
LIG_sift4	0.1442	0.2328	0.2276
LIG_mfcc	0.1056	0.1009	0.1089
LIG_hgsift	0.1685	0.2402	0.2439
LIG_all	0.2130	0.2672	0.2733

3.5 Video-level fusion

Now, we have the prediction scores for each descriptor on video level, we can obtain one prediction score for a video by merging all descriptors scores for the corresponding video, with the following formula:

$$V = \left(\frac{\sum_{i=1}^{N}(V_i^P)}{N}\right)^{\frac{1}{P}}$$

The limit when $P \rightarrow \infty$ correspond to fusion by taking the maximum score:

$$V = \max_{i=1}^{N} V_i$$

which has been used for all the previous evaluations. Table 8 shows the obtained performance for the video fusion using the Max ($P = \infty$) method or the proposed method (Opt) with the optimal value for P and for three global combinations of descriptors at the shot level:

LIG_all: late fusion of LIG_hgsift and LIG_mfcc using only the KNN classifier.

LIG_hgsift: late fusion of LIG_hg104 and LIG_sift4 using both classifiers and only visual descriptors;

LIG_all: late fusion of LIG_hgsift and LIG_mfcc4 using both classifiers and all descriptors.

Table 8: Results for video fusion, two-fold cross-validation

Descriptor	Max	P_{opt}	Opt
LIG_all	0.2145	0.80	0.2534
LIG_hgsifta	0.2436	0.40	0.2720
LIG_alla	0.2733	0.56	0.3040

3.6 Submissions

We made three submissions, one “official” one and two late ones. The official and main one (corresponding to LIG_all in table 8) includes only classification scores from the KNN classifier since MSVM ones were not available early enough. The two late and contrastive submissions correspond to the same one including MSVM classification results (LIG_alla in table 8) and the same with SVM using only the visual information (LIG_hgsifta in table 8).

Table 9 shows the official results for our main submission. The decision threshold was selected as the minimum one in the case of cross-validation within the development set. The prediction was generally quite good or did not have a significant effect on the official NDC evaluation metric since there is not much difference between the actual and minimum NDC.

3.7 Conclusion and future works

In future work, we will explore new features modalities for multimedia event detection particularly descriptors which take into account the temporal information. We will also integrate to the actual system other interesting descriptors like STIP, an extension of Harris and Fostner interest point operator to space-time. Finally, we will test other classifiers for example based on fisher Kernel.

4 Acknowledgments

This work was partly realized as part of the Quaero Programme funded by OSEO, French State agency for innovation. It was also partly supported by the UJF APIMS project. 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

Table 9: Official results on the main submission, A.NDC: actual NDC, M.DNC: minimum NDC

Event	#Targ	#NTarg	#CorDet	#FA	#Miss	A.NDC	Tresh.	M.NDC	Thresh.
E006	172	31865	1	24	171	1.0036	0.3490	1.0001	0.3576
E007	113	31924	3	20	110	0.9813	0.2964	0.8974	0.2599
E008	135	31902	77	668	58	0.6911	0.4006	0.6443	0.4172
E009	83	31954	29	371	54	0.7956	0.2145	0.7840	0.2131
E010	81	31956	8	191	73	0.9759	0.2158	0.9596	0.2459
E011	137	31900	10	27	127	0.9376	0.3022	0.8741	0.2674
E012	187	31850	13	148	174	0.9885	0.2730	0.9721	0.2923
E013	102	31935	21	433	81	0.9634	0.3140	0.9615	0.3142
E014	88	31949	32	69	56	0.6633	0.2681	0.5989	0.2388
E015	82	31955	14	320	68	0.9543	0.2458	0.9298	0.2536
All						0.8955		0.8622	

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 2011: Semantic Indexing and Multimedia Instance Search, In *Proceedings of the TRECVID 2011 workshop*, Gaithersburg, USA, 5-7 Dec. 2011.

[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] 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.

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

[12] 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.

[13] P. Over, G. Awad, J. , B. Antonishek, M.2Michel, A. Smeaton, W. Kraaij, and G. Quénot, TRECVID 2011 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics In *Proceedings of the TRECVID 2011 workshop*, Gaithersburg, USA, 5-7 Dec. 2011.