LIG at TRECVid 2015: Semantic Indexing

Bahjat Safadi[1,2], Nadia Derbas[1,2], Abdelkader Hamadi[1,2], Mateusz Budnik[1,2], Philippe Mulhem[1,2], and Georges Quénot[1,2]

[1]Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
[2]CNRS, LIG, F-38000 Grenoble, France

Abstract

LIG participated to the semantic indexing main task. LIG also participated to the organization of this task. This paper describes these participations which are quite similar to our previous year's participations (within the Quaero consortium).

For the semantic indexing main 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 main innovation this year consisted in the inclusion of semantic descriptors computed using a deep learning method. We also used multiple frames for some features and this did lead to a significant improvement. The best LIG run has a Mean Inferred Average Precision of 0.2935, which ranked it 5^{th} out of 15 participants.

1 Participation to the organization of the semantic indexing task

For the Sixth year, LIG has co-organized the semantic indexing task at TRECVid [1]. From 2010 to 2013 included, this was done with the support of Quaero[1] but this project has been completed by the end of 2013. The task is the same as in 2013 with the same set of 60 target concepts of which 30 were evaluated by NIST on the 2015 section of the test data.

A list of 500 target concepts has been produced, 346 of which have been collaboratively annotated by the participants and by Quaero annotators. A subset of 60 of them was selected for participants' submissions, 30 of which have been officially evaluated in 2015.

The 500 concepts are structured according to the LSCOM hierarchy [17]. 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 SIN provides participants with the following material:

- a development set that contains roughly 800 hours of videos;
- a test set that contains roughly 600 hours of videos, decomposed in three parts or roughly equal sizes, respectively for the 2013, 2014 and 2015 evaluations;
- 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).

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

The goal of the semantic indexing task is then to provide, for each of the 60 selected concepts, a ranked list of 2000 shots that are the most likely to contain the concept. The 2015 test collection contains 113,467 shots. More information about the organization of this task can be found in the TRECVid 2015 overview paper [2]. The *pair* version of the task that was proposed in 2012 and 2013 has been discontinued. The localization subtask, introduced in 2013 is also proposed and organized by NIST.

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	**IACC 2010-2015**
#videos	27,964
Duration (total)	∼1,400 hours
# shots	879,873
# shots (dev)	545,923
# shots (test 2013)	112,677
# shots (test 2014)	107,806
# shots (test 2015)	113,467

The whole set of videos has been split into two parts, the development set and the test set. The test set has been split in three part dedicated to the TRECVid SIN evaluations of 2013, 2014 and 2015. This has been done in order to be able to measure the performance progress over the three years. All sets were automatically split into shots using the LIG shot segmentation tool [18].

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 [19].

1.3 Annotations on the development set

Shots in the development set have been collaboratively annotated by TRECVid 2010-2013 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 [3]: 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. An improved algorithm for annotation cleaning has also been used in the annotation tool [14]. 8,158,517 were made directly by annotators and a total of 28,864,844 was obtained by propagating them using "implies" or "excludes" relations.

No new annotations were produced for 2014 and 2015; the development set is frozen so that difference of system performance is due only to algorithmic innovation and not to additional training data. 346 concepts were annotated on the development collection.

1.4 Assessments

30 concepts were selected for evaluation out of the 60 ones for which participants were asked to provide results for the main SIN task. Assessments were done part by NIST. Assessments were done by visualizing the whole shot for judging whether the target concept was visible or not at any time within the shot.

2 Participation to the semantic indexing main task

2.1 Introduction

The TRECVid 2015 semantic indexing task is described in the TRECVid 2015 overview paper [1, 2]. 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." 60 concepts have been selected for the TRECVid 2015 semantic indexing task. Annotations on the development part of the collections were provided in the context of the collaborative annotation and by Quaero.

As last years, our system uses a six-stage processing pipeline 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 simultaneously 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 2.7).

5. Higher-level fusion. We fuse here descriptors of different types, e.g. color, texture, interest points, motion (section 2.8).

6. Re-ranking. We post-process here the scores using the fact that videos statistically have an homogeneous content, at least locally (section 2.9).

Our system also 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 2.10). Compared to last year, our system include more semantic descriptors computed using a deep learning method (section 2.2.2) and the use of multiple key frames (section 2.5).

2.2 Descriptors

A total of 57 audio and visual descriptors have been used. Many of them have been produced by and shared with the IRIM consortium and two of them were provided by Xerox (XRCE). 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 in the IRIM consortium paper [11] and they are separately evaluated in section 2.6. They are decomposed into "classical" and "semantic" descriptors.

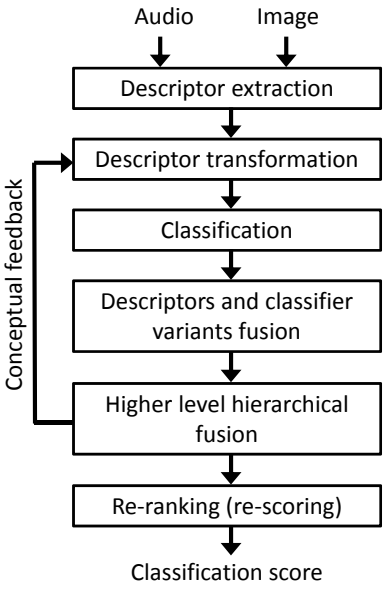

Figure 1: Semantic indexing system

2.2.1 Classicel descriptors

Classical descriptors include color histogram, Gabor transform, quaternionic wavelets, a variety of interest points descriptors (SIFT, color SIFT, SURF), local edge patterns, saliency moments, and spectral profiles for audio description. Many of them rely on a bag of words approach.

2.2.2 Semantic descriptors

Semantic of "high-level" descriptors are vectors of classification scores computed on the current data (here IACC) using classifier trained on other data and also with (generally) different target concepts (e.g. TRECVid HLF 2003 or ImageNet). They are opposed to classical or "low-level" ones in the sense in which the latter are computed using explicit algorithmic procedures (e.g. histograms or Gabor transforms) while the former comes from learning using annotated data.

We introduced in 2013 two semantic descriptors computed using Fisher vectors on ImageNet images and annotations:

XEROX/ilsvrc2010: Attribute type descriptor constituted as vector of classification score obtained with classifiers trains on external data with one vector component per trained concept classifier. For XEROX/ilsvrc2010, 1000 classifiers were trained using annotated data from the Pascal VOC / ImageNet ILSVRC 2010 challenge. Classification was done using Fisher Vectors [21].

XEROX/imagenet10174: Attribute type descriptor similar to XEROX/ilsvrc2010 but with 10174 concepts trained using ImageNet annotated data.

These were completed in 2014 by similar descriptors computed using deep convolutional networks on ImageNet images and annotations:

EUR/caffe1000: This descriptor was computed by Eurecom using the CAFFE Deep Neural Net [22] developed by the Vision group of the University of Berkeley, for which both the source code and the trained parameter values have been made available. The network has been trained on the ImageNet data only, and provides scores for 1000 concepts. The network is applied unchanged on the TRECVID key frames, both on training and test data. The resulting scores are accumulated in a 1000 dimension semantic feature vector for the shot.

LIG/caffeb1000: This descriptor is equivalent to the EUR/caffe1000 one and was also computed using the CAFFE Deep Neural Net [22] but with a different (later) version.

We also used descriptors based on the hidden layers of the deep convolutional network used for the computation of the LIG/caffeb1000 descriptor. We considered only the last two hidden layers (fc6 and fc7) since they were expected to also extract high-level information close to the semantics though not yet being tuned for other final target concepts:

LIG/caffe_fc[6|7]b_4096 : This descriptor correspond to the LIG/caffeb1000 one and was also computed using the CAFFE Deep Neural Net [22] but is made of the 4096 values of the last two hidden layers.

We introduced this year new DCNN-based descriptors and some early fusion of them:

EUR/b4096: descriptor of dimension 4096 obtained by early fusion of several other descriptors, including various local and global features, and the output of several pre-trained Deep Networks (Caffe [23], VGG16 and VGG19 [24][25]). The fusion is done by selecting the components for which the average conditional entropy of concepts given the component is the lowest. The selection is done independently for each component.

LIG/googlenet_pool5b_1024 : This descriptor is obtained by extracting the output of the last but one layer (pool5) of the GoogLeNet model [26] $\rightsquigarrow$ 1024 dimensions.

LIG/vgg_all_fc8 : This descriptor is obtained by extracting the output of the last layer of the VGG19 model [24][25] before the last normalization stage $\rightsquigarrow$ 1000 dimensions.

LIG/alex_goog_vgg_early : Early fusion of LIG/caffe_fc6_4096, LIG/googlenet_pool5b_1024 and LIG/vgg_all_fc8 after descriptor optimization as described in section 2.3 $\rightsquigarrow$ 1931 dimensions.

IRIM/all_dcnn_early : Early fusion of EUR/b4096 and LIG/alex_goog_vgg_early after descriptor optimization as described in section 2.3 $\rightsquigarrow$ 604 dimensions.

2.3 Descriptor optimization

The descriptor optimization consists into a PCA-based dimensionality reduction with pre and post power transformation [15]. Optionally, a L_1 or L_2 unit length normalization can also be performed before the PCA-based dimensionality reduction.

2.3.1 First 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.

2.3.3 Second power transformation

A second power transformation can be applied after PCA dimensionality reduction/ It has an affect which is similar to a post-PCA whitening but is has been proven to be more efficient and easy to tune. The optimal value of α_2 can be optimized by cross-validation and is often close to 0.7.

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 combination 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 [8], 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_BUSEB: Fusion between the two available classifiers. The fusion is simply done by a MAP weighted average of the scores produced by the two classifiers. Their output is naturally (or by construction) normalized in the the [0:1] range. kNN computation is done using the KNNLSB package [9]. 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 in term of information type capture. The MAP values used for the weighting are obtained by a two-fold cross-validation within the development set.

2.5 Use of multiple key frames

All descriptors (except audio and motion ones) have been computed on the reference key frame provided in the master shot segmentation. Additionally, some of them have been computed on all the I-frames extracted from the video shots (typically one every 12 video frames and about 13 per shot in average). Classification scores are computed in the same way both for the regular key frames and all the additional I-frames and a max pooling operation is performed over all the scored frames within a shot [6]. This max pooling operation is performed right after the classification step and before any fusion operation (though it would probably have been better to postpone it after).

2.6 Evaluation of classifier-descriptors combinations

We evaluated a number of image descriptors for the indexing of the 346 TRECVID 2012 concepts. This has been done with two-fold cross-validation within the development set. We used the annotations provided by the TRECVID 2013 collaborative annotation organized by LIG and LIF [3]. The performance is measured by the inferred Mean Average Precision (MAP) computed on the 346 concepts. Results are given in the IRIM paper [11].

2.7 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) [10]. 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.

2.8 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 variants of the same descriptors, on variants of classifiers on results from the same descriptors, on different types of descriptors and finally on the selection of groups of descriptors.

2.9 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 [13] 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 less homogeneous contents, the system performance was improved by about 11-13%.

2.10 Conceptual feedback

Since the TRECVid SIN 2013 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 improve 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 [16]. We used for that the best hierarchical fusion result available. This vector of scores was then included as a $(n+1)^{th}$ one in the pool of the N 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 (about 1.800 versus 0.150-0.700) for the other ones. This is probably linked with the way the score normalization is performed. Even though the 2013-2015 evaluation is done on 60 concepts only, as the annotations are available for 346 concepts, we used the full set for the conceptual feedback. The conceptual feedback vectors of concepts scores were built and used several times for different fusion processes corresponding to different sets of selected descriptors or to different ways of fusing them.

2.11 Performances on the semantic indexing task

In order to evaluate the systems' progress between 2013 and 2015 as suggested in the main SIN task, we shortly describe here the system variants that we used for our 2013, 2014 and 2015 submissions (four runs for each). The 2013 submissions were labeled as "Quaero" but, as this project is now finished, they are now labeled "LIG".

Four slightly different combinations of hierarchical fusion have been tried in 2013. The variations concerned

the way the re-ranking was done: it can be locally temporal, globally temporal and or conceptual. The variations also concerned the use or not of the uploader field available in the metadata [12]. Not all combinations could be submitted and the following were selected:

M_A_LIG,13_1 (was M_A_Quaero-2013-1_1): combination of M_A_LIG,13_3 with uploader information with 3:1 weights;

M_A_LIG,13_2 (was M_A_Quaero-2013-2_2): combination of M_A_LIG,13_3 with uploader information with 7:1 weights;

M_A_LIG,13_3 (was M_A_Quaero-2013-3_3): manually built hierarchical fusion of a large number (over 100) of jointly optimized descriptor-classifier combinations including two iterations of conceptual feedback combined with temporal re-ranking;

M_A_LIG,13_4 (was M_A_Quaero-2013-4_4): manually built hierarchical fusion of a large number (over 100) of jointly optimized descriptor-classifier combinations including a single iterations of conceptual feedback combined with temporal re-ranking.

Four slightly different combinations of hierarchical fusion have been tried in 2014. The variations concerned the use or not of the uploader field and the use of extended conceptual feedback versus basic conceptual feedback. Not all combinations could be submitted and the following were selected:

M_D_LIG,14_1: combination of M_D_LIG,14_2 with uploader information with 9:1 weights;

M_D_LIG,14_2: manually built hierarchical fusion of a large number (over 100) of jointly optimized descriptor-classifier combinations with extended conceptual feedback and temporal re-ranking.

M_D_LIG,14_3: combination of M_D_LIG,14_4 with uploader information with 9:1 weights;

M_D_LIG,14_4: manually built hierarchical fusion of a large number (over 100) of jointly optimized descriptor-classifier combinations with conceptual feedback and temporal re-ranking. Extended conceptual feedback is a version of conceptual feedback in which the components are weighted according to the correlation between the source and target concepts.

Four different combinations of hierarchical fusion have been tried in 2015. The variations concerned the use or not of the uploader field and the use of extended conceptual feedback versus basic conceptual feedback. Not all combinations could be submitted and the following were selected:

M_D_LIG,15_1: is similar to the LIG-2015-2C-3 submission but it additionally includes the I-frames in the prediction. However, the I-frames descriptors were available only for some of the available descriptors. We therefore made a I-frame pooling for those for which the I-frames version was available followed by an ad hoc late fusion between the predictions with I-frames and predictions without I-frames. The overall gain would probably have been higher if all the descriptors have been available on I-frames;

M_D_LIG,15_2: is similar to the LIG-2015-2C-4 baseline but it additionally includes the I-frames in the prediction. This cannot be seen in the submissions but the gain is significantly higher before temporal re-ranking: the gain brought by the multiple key frames and by the temporal re-ranking cumulates only partially, probably because the information obtained from the adjacent shots is partially redundant with the information obtained from the adjacent frames;

M_D_LIG,15_3: is similar to the LIG-2015-2C-4 submission but it additionally includes all the engineered descriptors, including the semantic ones from Xerox and the DCNN-based descriptors from Eurecom. The engineered descriptors alone lead to a performance much lower that the one obtained using only the DCNN features so we made no submission using only them; this also correspond to the performance of our 2013 system. Also, even though some new descriptors were made available from LISTIC and ETIS, we were not able to do better (at least using the ad hoc LIG strategy) than the 2014 fusion of them;

M_D_LIG,15_4: is the LIG baseline, it involves only DCNN features extracted by the LIG using the caffe software and three publicly available trained networks (AlexNet, GoogLeNet and VGG). It also uses only the main key frames (one per shot). This baseline is very good and above the best 2013 LIG submission.

Note: 2014 and 2015 rune were submitted as "type D" while 2013 ones were submitted as "type A". There is actually no real difference in training type but the rules regarding run types have been clarified in a more conservative way. Under the 2014 and 2015 understanding, 2013 runs would also have been labelled as "type D",

mostly because of the use of ImageNet data and annotations for the computation of the semantic descriptors.

Table 2: Mean InfAP result on the test set for all the 38 TRECVid 2013 evaluated concepts and/or for all the 30 TRECVid 2014 evaluated concepts

System/run	MAP 2013	MAP 2014	MAP 2015
Best run (*)	0.3211	0.3320	0.3624
M_A_LIG,13_3	0.2848	0.2416	0.2011
M_A_LIG,13_2	0.2846	-	-
M_A_LIG,13_4	0.2835	-	-
M_A_LIG,13_1	0.2827	0.2408	0.2012
M_D_LIG,14_3	0.3058	0.2659	-
M_D_LIG,14_4	0.3049	0.2643	-
M_D_LIG,14_2	0.3087	0.2586	0.2199
M_D_LIG,14_1	0.3094	0.2582	0.2254
M_D_LIG,15_1	0.3539	0.3460	0.2933
M_D_LIG,15_2	0.3421	0.3416	0.2935
M_D_LIG,15_3	0.3407	0.3151	0.2670
M_D_LIG,15_4	0.3288	0.3021	0.2533
Median submission	0.1275	0.2063	0.2398

(*) This run uses extra annotations.

Table 2 shows the performance of the three times four submitted variants in 2013, 2014 and 2015 for the 2013, 2014 and 2015 test collections, including the "progress" runs.

The addition of engineered descriptors (and of the Eurecom DCNN-based descriptors) brings a significant improvement (+5.4% in relative value) when using only the key frames (M_D_LIG,15_3 versus M_D_LIG,15_4) but none when using also the I-frames (M_D_LIG,15_1 versus M_D_LIG,15_3). However, this is the case only for the 2015 test data (otherwise an improvement is also observed) and it has to be taken into account that only some engineered descriptors were computed over the I-frames.

The use of I-frames brings a very important improvement (M_D_LIG,15_1 and M_D_LIG,15_2 versus M_D_LIG,15_3 and M_D_LIG,15_4): +15.4% and +9.9% respectively for the DCNN-based descriptors only and for all the descriptors.

Concerning the progress over years aspect, the values for the 2013, 2014 and 2015 test collections for a given run are not directly comparable because the test data are different and (probably mostly) because the evaluated concepts are different subsets of the 60 submitted ones, the 2014 subset looking harder than the 2013 one, and the 2015 subset looking harder than the 2014 one.

Considering our 2013 and 2014 runs, they are not directly comparable either because the variants have different tunings or because they were bugged. Only the M_A_LIG,13_3 and M_D_LIG,14_3 are built exactly with the same principles, the difference being in the use of additional semantic concepts coming from deep convolutional networks. These new descriptors yielded an improvement from 0.2848 to 0.3058 (+7.4% relative) on 2013 test data and from 0.2416 to 0.2659 (+10.0% relative) on 2014 test data.

Considering our 2014 and 2015 runs, they are not directly comparable because the variants have different tunings. The gain between our 2014 runs and our 2015 runs comes mostly from the use of multiple frames per shot (I-frames) and partly from the use of more DCNN-base descriptors.

3 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 as part of the CHIST-ERA Camomile Project funded by ANR, French national research agency.

Most of the computations presented in this paper were performed using the Froggy platform of the CIMENT infrastructure (https://ciment.ujf-grenoble.fr), which is supported by the Rhne-Alpes region (GRANT CPER07_13 CIRA) and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the programme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche.

Results from the IRIM network were also used in these experiments [11].

The authors also wish to thank Florent Perronnin from XRCE for providing descriptors based on classification scores from classifiers trained on ILSVRC/ImageNet data.

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] Paul Over, Georges Awad, Martial Michel, Johnatan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, Georges Quénot and Roeland Ordelman, TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics In *Proceedings of TRECVID 2015*, Gaitherburg, MD, USA, 16-18 Nov. 2015.

[3] 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.

[4] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News In *Transcription System. Speech Communication*, 37(1-2):89-108, 2002.

[5] 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.

[6] C.G.M. Snoek, M. Worring, J.-M. Geusebroek, D. Koelma and F.J. Seinstra, On the surplus value of semantic video analysis beyond the key frame, In IEEE International Conference onMultimedia and Expo (ICME), 6-8 July 2005.

[7] 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.

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

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

[10] D. Gorisse et al., IRIM at TRECVid 2010: High Level Feature Extraction and Instance Search. In *Proceedings of TRECVid 2010*, Gaithersburg, MD USA, November 2010.

[11] H. Le Borgne et al. IRIM at TRECVID 2015: Semantic Indexing, In *Proceedings of the TRECVID 2015 workshop*, Gaithersburg, MD, USA, 16-18 Nov. 2015.

[12] U. Niaz, M. Redi, C. Tanase, B. Merialdo, EURECOM at TrecVid 2012: The Light Semantic Indexing Task, In *Proceedings of TRECVid 2012*, Gaithersburg, USA, 25-28 Nov. 2012.

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

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

[15] Bahjat Safadi, Nadia Derbas and Georges Quénot. Descriptor Optimization for Multimedia Indexing and Retrieval. *,Multimedia Tools and Applications* Published online, May 2014.

[16] Abdelkader Hamadi, Georges Quénot, Philippe Mulhem. Conceptual Feedback for Semantic Multimedia Indexing, *,Multimedia Tools and Applications* Published online, May 2014.

[17] 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.

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

[19] 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.

[20] Stéphane Ayache, Georges Quénot. Image and Video Indexing using Networks of Operators. In *EURASIP Journal on Image and Video Processing*, 2007.

[21] Jorge Sánchez, Florent Perronnin, Thomas Mensink, Jakob Verbeek Image Classification with the Fisher Vector: Theory and Practice In *International Journal of Computer Vision,* Volume 105, Issue 3, pp 222-245, December 2013.

[22] Jia, Yangqing, Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding, 2013

[23] A. Krizhevsky, I. Sutskever and G.E. Hinton, *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–105,2012.

[24] K. Chatfield, K. Simonyan, A. Vedaldi and A. Zisserman, Return of the Devil in the Details: Delving Deep into Convolutional Nets, In *British Machine Vision Conference, 2014* (arXiv ref. cs1405.3531).

[25] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, Going Deeper with Convolutions arXiv:1409.4842.