IRIM at TRECVID 2015: Semantic Indexing

Hervé Le Borgne[1], Philippe Gosselin[2], David Picard[2], Miriam Redi[3], Bernard Mérialdo[3], Boris Mansencal[4], Jenny Benois-Pineau[4], Stéphane Ayache[5], Abdelkader Hamadi[6,7], Bahjat Safadi[3,6,7], Nadia Derbas[6,7], Mateusz Budnik[6,7], Georges Quénot[6,7], Boyang Gao[8], Chao Zhu[8], Yuxing Tang[8], Emmanuel Dellandrea[8], Charles-Edmond Bichot[8], Liming Chen[8], Alexandre Benoît[9], Patrick Lambert[9], and Tiberius Strat[9]

[1]CEA, LIST, Laboratory of Vision and Content Engineering, Gif-sur-Yvettes, France.
[2]ETIS UMR 8051, ENSEA / Université Cergy-Pontoise / CNRS, Cergy-Pontoise Cedex, F-95014 France
[3]EURECOM, Campus SophiaTech, 450 Route des Chappes, CS 50193, 06904 Biot Sophia Antipolis cedex, France
[4]LABRI UMR 5800, Université Bordeaux 1 / Université Bordeaux 2 / CNRS / ENSEIRB, Talence Cedex, France
[5]LIF UMR 6166, CNRS / Université de la Méditerranée / Université de Provence, F-13288 Marseille Cedex 9, France
[6]Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
[7]CNRS, LIG, F-38000 Grenoble, France
[8]LIRIS, UMR 5205 CNRS / INSA de Lyon / Université Lyon 1 / Université Lyon 2 / École Centrale de Lyon, France
[9]LISTIC, Domaine Universitaire, BP 80439, 74944 Annecy le vieux Cedex, France

Abstract

The IRIM group is a consortium of French teams supported by the GDR ISIS and working on Multimedia Indexing and Retrieval. This paper describes its participation to the TRECVID 2015 semantic indexing (SIN). 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 evaluated a number of different descriptors and tried different fusion strategies. The best IRIM run has a Mean Inferred Average Precision of 0.2947, which ranked it 4th out of 15 participants.

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 collection were provided for 346 concepts including the 60 target ones in the context of a collaborative annotation effort [13].

Eight French groups (CEA-LIST, ETIS, EURECOM, LABRI, LIF, LIG, LIRIS, LISTIC) collaborated to participate to the TRECVID 2015 semantic indexing task. Xerox (XRCE), though not being member of IRIM, also contributed descriptors.

The IRIM approach uses a six-stages processing pipeline that computes scores reflecting 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 produced by the participants (section 2).

2. Descriptor optimization. A post-processing of the descriptors allows to simultaneously improve their performance and to reduce their size (section 3).

3. Classification. Two types of classifiers are used as well as their fusion (section 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 6).

5. Higher-level fusion. We fuse here descriptors of different types, e.g. color, texture, interest points, motion (section 7).

6. Re-ranking. We post-process here the scores using the fact that videos statistically have an homogeneous content, at least locally (section 8).

This approach is quite similar to the one used by the IRIM group last years [14, 17]. The main novelties are the inclusion of more deep learning based descriptors and the use of multiple key frames (section 5).

2 Descriptors

Eight IRIM participants (CEA-LIST, ETIS, EURECOM, LABRI, LIF, LIG, LIRIS and LISTIC) provided a total of 74 descriptors, including variants of a same descriptors and early fusions of them. Xerox (XRCE) also provided two descriptors with us. 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 can be split into "hand crafted" or "engineered" (designed by human experts) descriptors and learnt (obtained using deep neural networks) descriptors. The relative performance of all the IRIM and XEROX descriptors has been separately evaluated using a combination of LIG classifiers (see LIG paper [18]).

2.1 Engineered descriptors

CEALIST/tlep: texture local edge pattern [3] + color histogram ⇝ 576 dimensions.

CEALIST/bov_dsiftSC_8192: : bag of visterm[28]. Dense SIFT are extracted every 6 pixels. The codebook of size 1024 is built with K-means. Bags are generated with soft coding and max pooling. The final signature result from a three levels spatial pyramid ⇝ $1024 \times (1 + 2 \times 2 + 3 \times 1) = 8192$ dimensions: see [15] for details.

CEALIST/bov_dsiftSC_21504: : bag of visterm[28]. Same as CEALIST/bov_dsiftSC_8192 with a different spatial pyramid ⇝ $1024 \times (1 + 2 \times 2 + 4 \times 4) = 21504$ dimensions.

ETIS/global_<feature>[<type>]x<size>: (concatenated) histogram features[4], where:

<**feature**> is chosen among lab and qw:

lab: CIE L*a*b* colors

qw: quaternionic wavelets (3 scales, 3 orientations)

<**type**> can be:

m1x1: histogram computed on the whole image

m1x3: histogram for 3 vertical parts

m2x2: histogram on 4 image parts

<**size**> is the dictionary size, sometimes different from the final feature vector dimension.

For instance, with <type>=m1x3 and <size>=32, the final feature vector has $3 \times 32 = 96$ dimensions.

ETIS/vlat_<desc type>_dict<dict size>_<size>: compact Vectors of Locally Aggregated Tensors (VLAT [6]). <desc type> = low-level descriptors, for instance hog6s8 = dense histograms of gradient every 6 pixels, 88 pixels cells. <dict size> = size of the low-level descriptors dictionary. <size> = size of feature for one frame. Note: these features can be truncated. These features must be normalized to be efficient (e.g. L_2 unit length).

EUR/sm462: The Saliency Moments (SM) feature [5] is a holistic descriptor that embeds some locally-parsed information, namely the shape of the salient region, in a holistic representation of the scene, structurally similar to [8].

LABRI/faceTracks: OpenCV+median temporal filtering, assembled in tracks, projected on key frame with temporal and spatial weighting and quantized on image divided in 16×16 blocks ⇝ 256 dimensions.

LIF/percepts_<x>_<y>_1_15: 15 mid-level concepts detection scores computed on x × y grid blocks in each key frames with (x,y) = (20,13), (16,6), (5,3), (2,2) and (1,1), ⇝ 15 × x × y dimensions.

LIG/raw32x24: (baseline) RGB image resized to 32×24 ⇝ 2304 dimensions.

LIG/h3d64: normalized RGB Histogram $4 \times 4 \times 4$ ⇝ 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[9] ⇝ 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 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.

LIG/concepts: detection scores on the 346 TRECVID 2011 SIN concepts using the best available fusion with the other descriptors, ⇝ 346 dimensions.

LIRIS/OCLPB_DS_4096 : Dense sampling OCLBP [29] bag-of-words descriptor with 4096 k-means clusters. We extract orthogonal combination of local binary pattern (OCLBP) to reduce original LBP histogram size and at the same time preserve information on all neighboring pixels. Instead of encoding local patterns on 8 neighbors, we perform encoding on two sets of 4 orthogonal neighbors, resulting two independent codes. Concatenating and accumulating two codes leads to a final 32 dimensional LBP histogram, compared with original 256 dimensions. The 4096 bag-of-words descriptors are finally generated by the pre-trained dictionary.

LIRIS/MFCC_4096: MFCC bag-of-words descriptor with 4096 k-means clusters. To reserves video's sequential information, we keep 2 seconds audio wave around the key frame, 1 second before and after. 39 dimensional MFCC descriptors with delta and delta delta are extracted with 20ms window length and 10ms window shift. The 4096 bag-of-words descriptors are finally generated by the pre-trained dictionary.

LISTIC/SIFT_*: Bio-inspired retinal preprocessing strategies applied before extracting Bag of Words of Opponent SIFT features (details in [23]) using the retinal model from [10]). Features extracted on dense grids on 8 scales (initial sampling=6 pixels, initial patch=16x16pixels, using a linear scale factor 1.2). K-means clusters of 1024 or 2048 visual words. The proposed descriptors are similar to those from [23] except the fact that multiscale dense grids are used. Despite showing equivalent mean average performance, the various prefiltering strategies present different complementary behaviors that boost performances at the fusion stage [36].

LISTIC/trajectories_*: Bag of Words of trajectories of tracked points. Various ways of describing a trajectory are used, such as the spatial appearance along a trajectory, the motion along a trajectory or a combination of both. Each type of trajectory description generates its own Bag of Words representation. K-means clustering of 256-1024 visual words, depending on the type of description [37].

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 [11].

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

2.2 Learnt descriptors

The learnt descriptors have been produced by LIG and Eurecom using the caffe tool [30] developed by the Vision group of the University of Berkeley, for which both the source code and the trained parameter values (models) have been made available. All of the following descriptors are obtained using models that were trained only on data different from TRECVid data and no retraining or tuning was performed on TRECVid data.

EUR/caffe1000: The AlexNet model trained on the ImageNet data only [31] has been applied unchanged on the TRECVID key frames, both on training and test data, providing detection scores for 1000 concepts. These are accumulated into a 1000 dimension semantic feature vector for the shot.

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 [31], VGG16 and VGG19 [32][33]). 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/caffeb1000: This descriptor is equivalent to the EUR/caffe1000 one and was also computed using a variant of the AlexNet model [31] but with a different (later) version ⇝ 1000 dimensions.

LIG/caffe_fc[6|7]_4096 : This descriptor correspond to the LIG/caffeb1000 one and was also computed using a variant of the AlexNet model [31] but is made of the 4096 values of the last two hidden layers, see [18] for more details ⇝ 4096 dimensions.

LIG/googlenet_pool5b_1024 : This descriptor is obtained by extracting the output of the last but one layer (pool5) of the GoogLeNet model [34] ⇝ 1024 dimensions.

LIG/vgg_all_fc8 : This descriptor is obtained by extracting the output of the last layer of the VGG19 model [32][33] before the last normalization stage ⇝ 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 3 ⇝ 1931 dimensions.

IRIM/all_dcnn_early : Early fusion of EUR/b4096 and LIG/alex_goog_vgg_early after descriptor optimization as described in section 3 ⇝ 604 dimensions.

3 Descriptor optimization

The descriptor optimization consists of a principal component analysis (PCA) based dimensionality reduction with pre and post power transformations [22]. A L_1 or L_2 unit length normalization can optionally by applied after the first power transformation.

4 Classification

The LIG participant ran two types of classifiers on the contributed descriptors as well as their combination, see [16] for details.

LIG/KNNB: The first classifier is kNN-based.

LIG/MSVM: The second one is based on a multiple learner approach with SVMs.

LIG/FUSEB: Fusion between classifiers. The fusion is simply done by a MAP weighted average of the scores produced by the two classifiers.

All the descriptors contributed by the IRIM participants have been evaluated for the indexing of the 346 TRECVID 2012 concepts. This has been done by the LIG participant and is reported in the TRECVid 2014 LIG paper [16].

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 [27]. 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).

6 Performance improvement by fusion of descriptor variants and classifier variants

As in previous years, we started by fusing classification scores from different variants of a same descriptor and from different classifiers of a same variant of a same descriptor. This is done as first levels of hierarchical late fusion, the last ones being done using dedicated methods as described in section 7. Three levels are considered when applicable: fusions of different classifiers of a same variant of a same descriptor, fusion of different variants of a same descriptor according to a dictionary size, and fusion of different variants of a same descriptor according to a pyramidal decomposition. While the last levels of fusion attempt to improve the overall performance by fusing information of different types (e.g. color, texture, percepts or SIFT), the first fusion levels attempt to improve the robustness of the classification from a given type. More details on this approach can be found in the previous TRECVid IRIM papers [20, 14].

7 Final fusion

The IRIM participant LISTIC worked on the automatic fusion of the classification results (experts) while LIG used manual (human expert guided) fusion. The fusion started with the original classification scores and/or with the results of previous fusions of descriptor variants and/or classifier variants as described in the previous section. A comparison of the LIG, LISTIC and LIMSI automatic fusion methods, along with an arithmetic mean and the best attribute per concept, is given in [25]. With the idea that a number of weak experts can be combined for producing a strong expert [26], all of the FUSEB experts (sections 2, 4 and 5) as well as some KNNB experts (corresponding to retina-enhanced SIFT/SURF/FREAK Bags of Words [36]) were included in the fusion process whatever their cross-validation estimated performance was, though some of them may be eventually filtered out by the automatic fusion process.

7.1 Automatic fusion (LISTIC)

LISTIC experimented with various automatic late fusion strategies, covered in detail in [25], ranging from simple arithmetic weighted fusions to more elaborate fusions such as AdaBoost or correlation-based methods. All these fusions work in a concept-per-concept manner, and for each concept, they fuse a subset of sufficiently-good experts. Two strategies appear to be the most effective, which constitute runs M_D_IRIM.15_2 and M_D_IRIM.15_3:

M_D_IRIM.15_2: combines a large set of 139 experts of varying input data and semantic level. More specifically, some experts are extracted from *keyframes* only and others come from *I-frame* of the video shot as described in section 5. The semantic level varies from low-level engineered features such as SIFT Bag of Words, Fisher vectors or high semantic level deep features etc. However, we do not include high level experts generated from the manual fusion detailed in next section. In this configuration, the AdaBoost fusion [26] is the most effective and manages well the high diversity of inputs.

M_D_IRIM.15_3: involves the same initial set of experts, but they are combined in two stages. The first stage fuses the set of *keyframe* experts and the set of *I-frame* experts independently. For each subset, the AdaBoost fusion performs the best for most concepts, but we keep flexibility by selecting the best fusion algorithm on a concept-per-concept basis. The second stage combines the two results of the first stage with a concept-per-concept weighted average.

7.2 Manual fusion (LIG)

LIG implemented the first fusion level grouping descriptor and classifier variants as described in section 6. These can be used in conjunction with the previous levels in automatic fusion runs or as the basis for further steps of supervised (manual) hierarchical fusion. The same hierarchical fusion as last year was done for engineered descriptors as no new ones were available this year. A new hierarchical fusion of DCNN-based descriptors was performed including early and late fusion schemes. Additionally, as the use of I-frames was introduced this year, fusions were performed separately using only key-frame based descriptors and I-frame based descriptors and a final level fusing both was added. More information on how this was performed and on the corresponding results can be found in the LIG paper [18]. The following runs were submitted by LIG for the IRIM group:

M_D_IRIM.15_1: late fusion of the two best LIG manual fusion results, key frames and I-frames;

M_D_IRIM.15_4: manual fusion of only DCNN (LIG and Eurecom) features, key frames only.

The last was only submitted for a contrast experiment.

8 Temporal re-scoring and conceptual feedback

At the end, temporal re-scoring (re-ranking) and conceptual feedback are performed. Temporal re-scoring consists in modifying the detection score of a given video shot for a given concept according to the detection scores of adjacent video shots for the same concept [21]. Conceptual feedback consists in modifying the detection score of a given video shot for a given concept according to the detection scores of other concepts for the same video shot [24].

9 Evaluation of descriptors

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 [19]. The performance is measured by the inferred Mean Average Precision (MAP) computed on the 346 concepts. Results are presented for the two classifiers used, as well as for their fusion. Results are presented only for the best combinations of the descriptor optimization hyper-parameters.

Tables 1 and 2 show respectively for the engineered and learnt descriptors the two-fold cross-validation performance (trec_eval MAP) within the development set and the performance (sample_eval MAP) on the 2015 test set with the LIG_FUSEB classifier combination; dim is the original number of dimensions of the descriptor vector, Pdim is the number of dimensions of the descriptor vector kept after PCA reduction, and α_1 and α_2 are the optimal values of the pre- and post-PCA power transformation coefficients. Additionally, the last column (iacc.2) displays the average of the MAP score on the 2013, 2014 and 2015 test sets. All these values are for descriptors computed only using one key frame (the reference one) per shot.

Table 3 shows for both the engineered and learnt descriptors the same values for the descriptors that were computed using the I-frames (as well as the reference key frame). For time and cost reasons, only a small fraction of the descriptors have been re-computed on the I-frames. These include the best DCNN-based descriptors computed by LIG, a number of SIFT-based descriptors computed by LISTIC and a SIFT-based pyramidal bag of word descriptor from CEA-LIST. The MAP values can be directly compared with those from

Table 1: Performance of engineered descriptors (key frames only)

Descriptor	dim	α_1	Unit length	Pdim	α_2	MAP dev	MAP 2015	MAP iacc.2
CEALIST/tlep_576	576	0.424	-	120	0.719	0.1237	0.0797	0.0832
CEALIST/bov_dsiftSC_8192	8192	0.700	-	292	0.575	0.1486	0.0826	0.0979
CEALIST/bov_dsiftSC_21504	21504	0.600	-	364	0.714	0.1557	0.1102	0.1297
ETIS/labm1x1x256	256	0.334	-	132	0.641	0.1096	0.0678	0.0663
ETIS/labm1x1x512	512	0.340	-	178	0.712	0.1115	0.0685	0.0670
ETIS/labm1x1x1024	1024	0.345	-	208	0.742	0.1122	0.0679	0.0669
ETIS/labm1x3x256	768	0.338	-	208	0.633	0.1213	0.0816	0.0849
ETIS/labm1x3x512	1536	0.351	-	310	0.651	0.1215	0.0813	0.0844
ETIS/labm1x3x1024	3072	0.380	-	333	0.720	0.1211	0.0789	0.0835
ETIS/labm2x2x256	1024	0.324	-	240	0.577	0.1173	0.0752	0.0793
ETIS/labm2x2x512	2048	0.353	-	308	0.621	0.1175	0.0742	0.0782
ETIS/labm2x2x1024	4096	0.378	-	324	0.739	0.1184	0.0738	0.0785
ETIS/qwm1x1x256	256	0.450	-	144	0.742	0.0982	0.0555	0.0580
ETIS/qwm1x1x512	512	0.437	-	166	0.718	0.1044	0.0612	0.0654
ETIS/qwm1x1x1024	1024	0.449	-	182	0.724	0.1088	0.0668	0.0713
ETIS/qwm1x3x256	768	0.421	-	205	0.696	0.1134	0.0717	0.0783
ETIS/qwm1x3x512	1536	0.413	-	230	0.725	0.1193	0.0757	0.0856
ETIS/qwm1x3x1024	3072	0.410	-	253	0.666	0.1225	0.0795	0.0902
ETIS/qwm2x2x256	1024	0.431	-	203	0.720	0.1098	0.0711	0.0736
ETIS/qwm2x2x512	2048	0.427	-	229	0.771	0.1150	0.0743	0.0799
ETIS/qwm2x2x1024	4096	0.423	-	277	0.788	0.1184	0.0772	0.0843
ETIS/vlat_hog3s4-6-8-10_dict64_4096	4096	0.875	L_1	4096	1.000	0.1624	0.1201	0.1457
EUR/sm462	462	0.167	-	215	0.380	0.1269	0.0791	0.0842
LIF/percepts_1_1_1_15	15	0.495	-	15	0.735	0.0860	0.0238	0.0307
LIF/percepts_2_2_1_15	60	0.470	-	60	0.669	0.1056	0.0436	0.0538
LIF/percepts_5_3_1_15	225	0.623	-	148	0.575	0.1092	0.0464	0.0561
LIF/percepts_10_6_1_15	900	0.619	-	169	0.381	0.1092	0.0466	0.0544
LIF/percepts_20_13_1_15	3900	0.550	-	193	0.420	0.1093	0.0520	0.0594
LABRI/faceTracks16x16	256	0.240	-	210	0.480	0.0180	0.0068	0.0069
LIG/raw32x24	2304	1.100	-	91	0.700	0.0991	0.0482	0.0487
LIG/gab40	40	0.629	-	40	0.629	0.0809	0.0206	0.0249
LIG/h3d64	64	0.286	-	52	0.813	0.0916	0.0425	0.0442
LIG/hg104	104	0.348	-	89	0.700	0.1148	0.0653	0.0666
LIG/opp_sift_har_1000	1000	0.513	-	103	0.782	0.1194	0.0635	0.0773
LIG/opp_sift_dense_1000	1000	0.489	-	206	0.466	0.1276	0.0764	0.0890
LIG/opp_sift_har_unc_1000	1000	0.331	-	116	0.592	0.1262	0.0746	0.0876
LIG/opp_sift_dense_unc_1000	1000	0.415	-	303	0.384	0.1354	0.0847	0.0975
LIG/opp_sift_har_1024_fu8	1024	0.409	-	170	0.324	0.1264	0.0697	0.0826
LIRIS/MFCC_4096	4096	0.426	L_2	200	1.000	0.0584	0.0157	0.0173
LIRIS/OCLBP_4096	4096	0.374	L_2	167	0.681	0.1122	0.0773	0.0915
LISTIC/SIFT_1024	1024	0.444	-	272	0.436	0.1274	0.0841	0.0983
LISTIC/SIFT_2048	2048	0.912	-	175	0.420	0.1115	0.0679	0.0774
LISTIC/SIFT_retina_1024	1024	0.504	-	204	0.515	0.1288	0.0781	0.0906
LISTIC/SIFT_retina_2048	2048	0.768	-	134	0.455	0.1208	0.0665	0.0798
LISTIC/SIFT_retinaMasking_1024	1024	0.400	-	136	0.399	0.1274	0.0797	0.0911
LISTIC/SIFT_retinaMasking_2048	2048	0.434	-	171	0.187	0.1013	0.0525	0.0572
LISTIC/SIFT_multiChannels..._1024	1024	0.398	-	123	0.369	0.1287	0.0802	0.0949
LISTIC/SIFT_multiCh...Dual1024_2048	2048	0.438	-	160	0.258	0.1291	0.0861	0.1015
LISTIC/SIFTbased_1024	1024	0.450	-	277	0.432	0.1291	0.0839	0.0989
LISTIC/SIFTbased_retina_1024	1024	0.479	-	218	0.285	0.1333	0.0800	0.0925
LISTIC/SIFTbased_Retina..._parvo_1024	1024	0.450	-	181	0.492	0.1308	0.0826	0.0939
LISTIC/SIFTbased_multiCh..._L2norm_3072	3072	0.500	-	135	0.460	0.1390	0.0922	0.1041
LISTIC/expe6_trajectories_13_1024	1024	0.542	-	64	0.849	0.0726	0.0601	0.0705
LISTIC/expe6_trajectories_14_1024	1024	0.547	-	64	0.849	0.0724	0.0595	0.0703
XEROX/ilsvrc2010	1000	0.575	-	592	0.650	0.1710	0.1539	0.1824
XEROX/imagenet10174	10174	0.200	-	1024	0.650	0.1721	0.1570	0.1886

Table 2: Performance of learnt descriptors (key frames only)

Descriptor	dim	α_1	Unit	Pdim	α_2	MAP	MAP	MAP
			length			dev	2015	iacc.2
EUR/caffe1000	1000	0.297	-	670	0.547	0.2025	0.1627	0.1847
EUR/b4096	4096	1.900	-	800	0.700	0.2521	0.2398	0.2641
LIG/caffe_fc6_4096	4096	0.449	-	662	0.558	0.2157	0.1776	0.2043
LIG/caffe_fc7_4096	4096	0.766	-	738	0.558	0.2133	0.1717	0.1985
LIG/caffeb1000	1000	0.210	-	754	0.558	0.1982	0.1572	0.1796
LIG/googlenet_pool5b_1024	1024	0.650	-	660	0.670	0.2291	0.2087	0.2323
LIG/vgg_all_fc8	1000	0.650	-	609	0.400	0.2301	0.2042	0.2291
LIG/alex_goog_vgg_early	1931	1.000	-	294	0.610	0.2536	0.2322	0.2621
IRIM/all_dcnn_early	604	0.970	-	275	0.710	0.2615	0.2453	0.2728

tables 1 and 2 for measuring the benefit of using I-frames in addition to the reference key frames.

We can observe that the performance actually decreases in cross-validation within the development set. This is due to the fact that within the development set the annotation for a shot is generally done on the basis of only the reference key frame while within the test sets the assessment is done on the basis of the whole shot (including all the I-frames). On the test sets, a significant improvement is observed as reported in [27]. On the 2015 test set, the relative improvement is of 25% (respectively 30%) for the engineered (respectively learnt) features. On the whole iacc.2 (2013, 2014 and 2015) test set, the relative improvement is of 25% (respectively 22%). The average gain is thus very important. Unfortunately, this gain does not cumulate well with the gain brought by the re-scoring post processing. This is likely because both use similar or redundant information from adjacent frames or shot.

10 Evaluation of the submitted runs

We submitted 4 runs, each using the same input experts:

M_D_IRIM.15_1: late fusion of the two best LIG manual fusion results, key frames and I-frames;

M_D_IRIM.15_2: automatic fusion based on Adaboost, key frames and I-frames;

M_D_IRIM.15_3: automatic two-stage fusion of key frames and I-frames;

M_D_IRIM.15_4: manual fusion of only DCNN (LIG and Eurecom) features, key frames only.

All include re-ranking and manual fusion runs include conceptual feedback.

Table 4 presents the result obtained by the four runs submitted as well as the best and median runs for comparison. The best IRIM run corresponds to a rank of 4 within the 15 participants to the TRECVID 2014 main SIN task.

Table 4: InfMAP result and rank on the test set for all the 30 TRECVID 2015 evaluated concepts (main task, over 56 submissions (2015 only, excluding the progress ones).

System/run	MAP	rank
Best run (*)	0.3624	1
M_D_IRIM.15_1	0.2953	12
M_D_IRIM.15_2	0.2907	15
M_D_IRIM.15_3	0.2893	16
M_D_IRIM.15_4	0.2648	23
Median run	0.2398	28
Random run	0.0009	-

(*) This run uses extra annotations.

These results show that, regarding automatic fusion, directly fusing all the experts works slightly better. The use of *I-frame* descriptors improves results compared to relying on a *keyframe*-only description, however there is no need to treat the two sets separately. Our resulting fused experts outperform the best deep network based expert, nevertheless the margin remains about 15%. This confirms the superiority of deep architectures versus classical shallow architectures.

11 Additional contrast experiments

We conducted several contrast experiments for isolating and measuring the effect of various methods and/or parameters. We have not been able to consider all the possible combinations because, for instance, not all descriptors were computed for all descriptors.

11.1 Learnt versus engineered features

We first compare the best fusion of engineered features and the best fusion of learnt features, without the use

Table 3: Performance of descriptors using I-frames

Descriptor	dim	α_1	Unit length	Pdim	α_2	MAP dev	MAP 2015	MAP iacc.2
CEALIST/bov_dsiftSC_21504	21504	0.600	-	364	0.714	0.1519	0.1392	0.1577
LIG/caffe_fc6_4096	4096	0.449	-	662	0.558	0.2039	0.2233	0.2472
LIG/googlenet_pool5b_1024	1024	0.650	-	660	0.670	0.2172	0.2770	0.2906
LIG/vgg_all_fc8	1000	0.650	-	609	0.400	0.2177	0.2657	0.2814
LIG/alex_goog_vgg_early	1931	1.000	-	294	0.610	0.2380	0.3001	0.3152
LISTIC/SIFTbased_1024	1024	0.450	-	277	0.432	0.1256	0.1024	0.1223
LISTIC/SIFTbased_retina_1024	1024	0.479	-	218	0.285	0.1292	0.0960	0.1157
LISTIC/SIFTbased_Retina... _parvo_1024	1024	0.450	-	181	0.492	0.1313	0.1061	0.1201
LISTIC/SIFTbased_multiCh... _L2norm_3072	3072	0.500	-	135	0.460	0.1385	0.1155	0.1326

of I-frames, without re-scoring and without conceptual feedback.

Table 5: Learnt versus engineered features

Fusion	MAP 2013	MAP 2014	MAP 2015	MAP iacc.2
IRIM/engineered	0.2546	0.2056	0.1774	0.2125
IRIM/learnt	0.3091	0.2752	0.2477	0.2773
IRIM/all	0.3190	0.2849	0.2553	0.2863

As can be seen in table 5 and as this was observed last year, learnt features significantly outperform engineered features but the combination of both still perform even better.

11.2 Use of adjacent frames and shots and use of conceptual feedback

We consider the two “basic” manual hierarchical fusions: “LIG/dcnn” that contains all DCNN-based (learnt) descriptors computed by LIG and “IRIM/all” that contains all (engineered and learnt) descriptors computed by the IRIM participants. All the descriptors included in the LIG/dcnn descriptors have also been computed on the I-frames but this has been the case only for part of those included in the IRIM/all fusion (only those mentioned in table 3). Therefore, “LIG/dcnn+if” (including I-frames) is directly comparable with “LIG/dcnn” while “IRIM/all+if” is not directly comparable with “IRIM/all” (the “IRIM/all+if” is under-evaluated).

Table 6 (respectively table 7) show the effect of using I-frames (+if), Temporal Re-Scoring (+trs) and conceptual feedback (+cf) on the 2015 test set (respectively on the iacc.2 test set). Results in table 7 average the performance on the 2013, 2014 and 2015 test sets and is therefore expected to be more stable than results in table 6 that correspond to the 2015 test set only. Results between both tables are not always consistent, indicating that the small differences observed are not always statistically significant.

Table 6: Effect if I-frames (+if), Temporal Re-Scoring (+trs) and conceptual feedback (+cf), MAP on the 2015 test set

processing	base	+trs	+trs+cf
LIG/dcnn	0.2350	0.2491	0.2533
LIG/dcnn+if	0.3004	0.2974	0.2935
IRIM/all	0.2553	0.2625	0.2670
IRIM/all+if	0.2917	0.2938	0.2933

Table 7: Effect of I-frames (+if), Temporal Re-Scoring (+trs) and conceptual feedback (+cf), average of MAP on the iacc.2 (2013, 2014 and 2015) test set

processing	base	+trs	+trs+cf
LIG/dcnn	0.2660	0.2869	0.2947
LIG/dcnn+if	0.3170	0.3204	0.3257
IRIM/all	0.2863	0.3002	0.3075
IRIM/all+if	0.3196	0.3258	0.3310

Considering only the results in table 6 (2015 test set), the best performance is obtained with only the DCNN features computed using the I-frames. Neither engineered features, temporal re-scoring or conceptual feed back can help. Without I-frames, all of engineered features, temporal re-scoring and conceptual feed back do help.

Considering only the results in table 7 (2013, 2014 and 2015 test sets), with and without I-frames, all of engineered features, temporal re-scoring and conceptual feed back do help and the best performance is obtained by combining them all with I-frames.

12 Data sharing

As in previous years, we propose to reuse and extend the organization that has been developed over six years within the members of the IRIM project of the French ISIS national Research Group (see [12] and section 1 of this paper). It is based on a limited number of simple data formats and on a (quite) simple directory organization. It also comes with a few scripts and procedures as well as with some sections for reporting intermediate results. The supporting structure is composed of a wiki (http://mrim.imag.fr/trecvid/wiki) and a data repository (http://mrim.imag.fr/trecvid/sin15). The wiki can be accessed using the TRECVid 2013 active participant username and password and the data repository can be accessed using the TRECVid 2013 IACC collection username and password.

A general rule about the sharing of elements is that:

- any group can share any element he think could be useful to others with possibly an associated citation of a paper describing how it was produced;
- any group can use any element shared by any other group provided that this other group is properly cited in any paper presenting results obtained using the considered element,

exactly as this was the case in the previous years for the shared elements like shot segmentation, ASR transcript or collaborative annotation. Groups sharing elements get "rewarded" via citations when their elements are used.

Shared elements can be for instance: shot or key frame descriptors, classification results, fusion results. For initiating the process, most IRIM participants agreed to share their descriptors. Most classification and fusion results obtained are also shared. These are available on the whole 2010-2015 TRECVID SIN collection. Descriptors, classification scores or fusion results from other TRECVid participants are most welcome. See the wiki for how to proceed.

Acknowledgments

This work has been carried out in the context of the IRIM (Indexation et Recherche d'Information Multimédia) of the GDR-ISIS research network from CNRS.

Experiments presented in this paper were carried out using the Grid'5000 experimental test bed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr). LISTIC carried out experiments using the MUST computing center of the University of Savoie Mont Blanc.

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] Y.-C. Cheng and S.-Y. Chen. Image classification using color, texture and regions. In *Image and Vision Computing,* 21:759-776, 2003.

[4] P.H. Gosselin, M. Cord, Sylvie Philipp-Foliguet. Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval. In *Computer Vision and Image Understanding,* Special Issue on Similarity Matching in Computer Vision and Multimedia. Volume 110, Issue 3, Pages 403-441, 2008.

[5] M. Redi and B. Merialdo. Saliency moments for image categorization, In *ICMR 2011, 1st ACM International Conference on Multimedia Retrieval,* April 17-20, 2011, Trento, Italy.

[6] D. Picard and P.H. Gosselin. Efficient image signatures and similarities using tensor products of local descriptors, In *Computer Vision and Image Understanding,* Volume 117, Issue 6, Pages 680-687, 2013.

[7] P.H. Gosselin, N. Murray, H. Jegou and F. Perronnin. Revisiting the Fisher vector for fine-grained classification. In *Pattern Recognition Letters,* Volume 49, Pages 92-98, November 2014.

[8] A. Oliva and A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, In *International Journal of Computer Vision,* vol 42, number 3, pages 145-175, 2001.

[9] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of Color Descriptors for Object and Scene Recognition. In *IEEE Trans. Pattern Anal. Mach. Intell.,* 32(9):1582-1596, September 2010.

[10] A. Benoit, A. Caplier, B. Durette, and J. Herault. Using Human Visual System Modeling for Bio-inspired Low Level Image Processing, In *Computer Vision and Image Understanding,* vol. 114, no. 7, pp. 758-773, 2010.

[11] 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.

[12] Ballas et al. IRIM at TRECVID 2012: Semantic Indexing and Multimedia Instance Search, In *Proceedings of the TRECVID 2011 workshop*, Gaithersburg, USA, 26-28 Nov. 2012.

[13] Safadi et al. Quaero at TRECVID 2013: Semantic Indexing and Collaborative Annotation, In *Proceedings of the TRECVID 2013 workshop*, Gaithersburg, MD, USA, 20-22 Nov. 2013.

[14] Ballas et al. IRIM at TRECVID 2013: Semantic Indexing and Multimedia Instance Search, In *Proceedings of the TRECVID 2013 workshop*, Gaithersburg, MD, USA, 20-22 Nov. 2013.

[15] Nicolas Ballas, Benjamin Labbé, Hervé Le Borgne, Aymen Shabou CEA LIST at TRECVID 2013: Instance Search, In *Proceedings of the TRECVID 2013 workshop*, Gaithersburg, MD,USA, 20-22 Nov. 2013.

[16] Safadi et al. LIG at TRECVID 2014: Semantic Indexing, In *Proceedings of the TRECVID 2014 workshop*, Orlando, FL, USA, 10-12 Nov. 2014.

[17] Ballas et al. IRIM at TRECVID 2014: Semantic Indexing and Instance search, In *Proceedings of the TRECVID 2014 workshop*, Orlando, FL, USA, 10-12 Nov. 2014.

[18] Safadi et al. LIG at TRECVID 2015: Semantic Indexing, In *Proceedings of the TRECVID 2015 workshop*, Gaithersburg, MD, USA, 16-18 Nov. 2015.

[19] 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.

[20] 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.

[21] 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.

[22] Bahjat Safadi, Georges Quénot. Descriptor Optimization for Multimedia Indexing and Retrieval. *,Multimedia Tools and Applications* 74(4) pp 1267-1290, Feb. 2015.

[23] Strat, S.T. and Benoit, A. and Lambert, P., Retina enhanced SIFT descriptors for video indexing, *CBMI 2013, 11th International Workshop on Content-Based Multimedia Indexing,* Veszprem, HUNGARY, jun 2013.

[24] Abdelkader Hamadi, Georges Quénot, Philippe Mulhem. Conceptual Feedback for Semantic Multimedia Indexing, *,Multimedia Tools and Applications* 74(4) pp 1225-1248, Feb. 2015.

[25] S. T. Strat, A. Benoit, P. Lambert, H. Bredin and G. Quénot. Hierarchical Late Fusion for Concept Detection in Videos, In *Fusion in Computer Vision - Understanding Complex Visual Content, Springer Advances in Computer Vision and Pattern Recognition*, 2014.

[26] Y. Freund and R.E. Schapire, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, In Journal of Computer and System Sciences, 55(1):119-139,1997,

[27] 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.

[28] A. Shabou and H. Le Borgne. Locality-constrained and spatially regularized coding for scene categorization, In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3618–3625, 2012.

[29] C. Zhu, C.-E. Bichot, L. Chen. Color orthogonal local binary patterns combination for image region description. In *Technical Report, LIRIS UMR5205 CNRS*, Ecole Centrale de Lyon.

[30] Y. Jia, E. helhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell, Caffe: Convolutional Architecture for Fast Feature Embedding, In *Proceedings of the ACM International Conference on Multimedia, MM'14*, Orlando, Florida, USA, 2014.

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

[32] 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).

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

[34] 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.

[35] A. Mikulik, M. Perdoch, O. Chum and J. Matas. Learning vocabularies over a fine quantization. *International Journal of Computer Vision*, vol 103; no 1, pp 163–175, 2013.

[36] T. Strat, A. Benoit, P. Lambert Retina enhanced bag of words descriptors for video classification Eusipco 2014, Lisbon, Portugal, 2014.

[37] Strat, S.T., Benoit, A. and Lambert, P., Bags of Trajectory Words for video indexing, In *Content-Based Multimedia Indexing (CBMI)*, Klagenfurt, Austria, June 2014