IRIM at TRECVID2009: High Level Feature Extraction

Bertrand Delezoide, Hervé Le Borgne, Pierre-Alain Moëllic, David Gorisse, Frédéric Precioso, Feng Wang, Bernard Merialdo, Philippe-Henri Gosselin, Lionel Granjon, Denis Pellerin, et al.

IRIM at TRECVID 2009: High Level Feature Extraction

Bertrand Delezoide (CEA-LIST) Hervé le Borgne (CEA-LIST),
Pierre-Alain Moëllic (CEA-LIST) David Gorisse (ETIS) Frédéric Precioso (ETIS)
Feng Wang (Eurecom) Bernard Merialdo (Eurecom) Philippe Gosselin (ETIS)
Lionel Granjon (GIPSA) Denis Pellerin (GIPSA) Michèle Rombaut (GIPSA)
Hervé Bredin (IRIT) Lionel Koenig (IRIT) Hélène Lachambre (IRIT)
Elie El Khoury (IRIT) Boris Mansencal (LABRI) Yifan Zhou (LABRI)
Jenny Benois-Pineau (LABRI) Hervé Jégou (LEAR) Stéphane Ayache (LIF)
Bahjat Safadi (LIG) Georges Quénot (LIG) Jonathan Fabrizio (LIP6)
Matthieu Cord (LIP6) Hervé Glotin (LSIS) Zhongqiu Zhao (LSIS)
Emilie Dumont (LSIS) Bertrand Augereau(XLIM-SIC)

Abstract

The IRIM group is a consortium of French teams working on Multimedia Indexing and Retrieval. This paper describes our participation to the TRECVID 2009 High Level Features detection task. We evaluated a large number of different descriptors (on TRECVID 2008 data) and tried different fusion strategies, in particular hierarchical fusion and genetic fusion. The best IRIM run has a Mean Inferred Average Precision of 0.1220, which is significantly above TRECVID 2009 HLF detection task median performance. We found that fusion of the classification scores from different classifier types improves the performance and that even with a quite low individual performance, audio descriptors can help.

1 Introduction

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

2 Evaluation of image descriptors

We have evaluated a large number of image descriptors for image classification in the context of TRECVID 2008 and 2009 [6]. We used for this the data, annotations, ground truth, protocol and metrics of the TRECVID High Level Features (HLF) detection task. These HLFs are actually concepts, objects or events to be detected in video shots. Video shots are often represented by key frames and feature extraction, classification and fusion are often performed only on key frames. In this case, the task can be considered as a still image classification task. In some other cases, motion information is taken into account or image sequences are considered and the task is truly a video shot classification task. Finally, in some cases, features from the audio track are also taken into consideration and the task is truly a multimodal shot classification task.

We have considered here a number of image descriptors. These descriptors were produced in the context of the IRIM action of the ISIS "Groupe De Recherche" (GDR) from CNRS leaded by LIG, IRIT, LABRI and LIP6.

Twelve IRIM participants (CEA-LIST, ETIS, Eurecom, GIPSA, IRIT, LABRI, LEAR, LIF, LIG, LIP6, LSIS and XLIM-SIC) provided descriptors and three participants (LIF, LIG ans ETIS) provided classification results using them allowing for comparing the relative performances of these descriptors. These descriptors do not cover all types and variants but they include a significant number of different approaches including state of the art ones and more exploratory ones. Three IRIM participants evaluated these descriptors using a total of four different classifiers. The evaluation was conducted on TRECVID 2008 concepts annotated on the TRECVID 2007 collection (which is the trec2008

development collection). The training and evaluation were done respectively on the development and test parts of the TRECVID 2007 collection.

We evaluated the following image descriptors:

- **CEALIST_global_tlep**: texture (local edge pattern [1], 512-dim) and color (histo RGB, 64-dim).
- **CEALIST_global_cime**: 4-connexity histogram for 64 RGB colors (64-dim) derived from [2].
- **CEALIST_global_projection**: horizontal and vertical projections (sums) of gray level of the image, rescaled at 100×100 pixels, according to a 2×2 grid (400-dim), as originally proposed in [3].
- **CEALIST_global_probe**: co-occurrence in color space (164-dim)
- **CEALIST_global_cciv**: color histogram on pixels of same color in regions larger that 5% of the image area (345-dim) (initialy described in [4]).
- **CEALIST_global_pigment** $5 \times 5 \times 5$ RGB color histogram (125-dim).
- **CEALIST_local**: bag of SIFT, using a Harris-Laplace detector and a K-mean algorithm to create the dictionnary (5000-dim).
- **CEALIST_scribe**: 4 semantic descriptors developped in [5]:
  Dimension 1: 1=photo, 2=B&W photo, 3=color photo, 4=painting, 5=clipart, 6=map, 0=unknown
  Dimension 2: 1=indoor, 2=outdoor, 0=unknown
  Dimension 3: 1=day, 2=night, 0=unknown
  Dimension 4: 1=urban, 2=nature, 0=unknown
- **ETIS_global_<attr>[<type hist>]<dict size>**: histograms computed for different visual attributes and dictionary size.
  <attr> = lab: Lab colors, qw: norm of quaternionic wavelets coefficients, 3 scales.
  <type hist> = (nothing): histogram computed on the whole image, bic: 2 histograms on interior and border pixels, m1 × 3: 3 histograms on 3 vertical stripes, m2 × 2: 4 histograms on four image quarters.
- **Eurecom_surf**: bag of SURF descriptor (SIFT-like, 500-dim) [14].
- **GIPSA_faces_<n>**: number of faces detected in the key frame. <n> is a tuning parameter of the recall versus precision tradeoff. The face detector is described in [10].
- **IRIT_BoW-SIFT**: SIFT Bag-of-Words (TF-IDF / 250-dim).
- **IRIT_BoW-Color**: Color Bag-of-Words (TF-IDF / 250-dim).
- **LABRI_BlackAndWhite_Mean**: Prediction of black and white shots.
- **LABRI_facesOCV_Mean**: mean number per shot. Faces are detected by OpenCV combined with particle filter tracking [11].
- **LABRI_faces**: mean and variance of number of faces, ratio of faces bounding boxes area and frame area, x coordinate and y coordinate of faces bounding boxes.
- **LEAR_bow_sift_<dict_size>**: Bag Of SIFT Words vectors with dict_size = 500, 1000, 2000, 4000, 8000 and 16000.
- **LIF_Global**: Global color and texture descriptor. Similar to LIG_hg104 but with different normalisation.
- **LIF_MMCOL_HxW**: Color RGB moments (3 means, 3 variances and 3 covariances) on a H × W grid.
- **LIF_HISTCOL_HxW**: Color $3 \times 3 \times 3$ RGB histogram on a H × W grid.
- **LIF_GABOR_HxW**: Texture descriptor, Gabor filter with 8 orientations and 5 scales on a H × W grid.
- **LIF_HISTEDGE_HxW**: Histogram of edge orientation (50 bins) on a H × W grid.
- **LIF_LBP_2x2**: Local Binary Pattern descriptor on a 2×2 grid (1024-dim).
- **LIF_SIFT_1000**: bag of SIFT features [17] (1000-dim).
- **LIF_SEMANTIC**: Detection of other concepts on image blocks and summation of scores on these blocks (20-dim).
- **LIF_Percepts**: Intermediate level descriptor contains the prediction scores of 15 visual concepts a 29×13 grid (3900-dim).
- **LIG_h3d64** : normalized RGB Histogram $4 \times 4 \times 4$ (64-dim).
- **LIG_gab40** : normalized Gabor transform, 8 orientations × 5 scales (40-dim).
- **LIG_hg104** : early fusion (concatenation) of LIG_h3d64 and LIG_gab40 (104-dim).

- **LIG_opp_sift_har** : bag of word, opponent sift with Harris-Laplace detector [18], generated using Koen Van de Sande's software (4000-dim).
- **LIP6_text** : Text detection: number of detected text zone, number of detected characters and area of text zones for the whole image, the top two-third and the bottom third (9-dim).
- **LSIS_PEF_150** : Profile Entropy Features (150-dim).
- **LSIS_PEF_45** : Profile Entropy Features (45-dim).
- **LSIS_HSV** : Color HSV Histogram ($7 \times 3 \times 3$)
- **LSIS_EDGE** : Global edge histogram (72-dim)
- **LSIS_GABOR** : Histogram of Gabor filters, $4 \times 3 \times 5$ dimensions.

Four classifiers were used for the evaluation of the descriptors (the same classifiers were also used for producing the predictions of the TRECVID submission using the same descriptors):

- **LIF_SVM: SVM** classifier trained with all positive examples and twice as many randomly chosen negative samples [15]. RBF kernel with C and gamma values optimized by cross-validation.
- **LIG_KNNC** and **LIG_KNNG**: KNN-based classifier with hyper-parameters obtained by cros validation with an optimization respectively by concept or globally [16].
- **ETIS_SVM**: SVM with triangular kernel and progressive inclusion of negative samples.

Tables 1 and 2 shows the individual performance of each descriptor with up to four different classifiers. The perfomance of the ten best descriptors tested for each classifier are displayed in bold. For comparison, tests with randomly generated output indication a performance of 0.0022 ± 0.0005 for a random submission while a perfect submission would have a performance of 1.0000.

Many interesting observations can be made from these results:

- The individual performance of the descriptors is low. The typical MAP performance of a good "monomodal" descriptor is in the 0.0300-0.0500 range. This is about 20 times more than a random prediction but still 20 to 30 times less than a perfect prediction. This is indeed due to the fact that the task is difficult and this indicates that the fusion of several descriptors is necessary.
- The best individual performance is obtained by a color variant of a SIFT descriptor (0.0644 for LIG_opp_sift_har with the LIG_KNNC classifier) but the second best performance (0.0639 for LIF_Percepts with the LIG_KNNG classifier) is very close with a very different method.
- The bag of work approach is frequently used and when various dictionaries sizes are used, the larger sizes often leads to the best performances, sometimes with sizes up to several thousands.
- Similarly, decomposing the images into spatial regions and computing descriptors on all of them generally improves the performance and large number of regions, when possible, often leads to the best performances.
- The performance of different classifiers on a same descriptor are different, sometimes with a factor of up to two. As this has been previously observed, the KNN classifiers, when properly optimized, can perform as well as or better than SVM classifiers. Here, they perform quite often better, possibly indicating a non optimal use for the SVM classifiers. Other tests (not in the table) have shown a better performance of SVM classifiers when better managing the imbalanced class problem (there are always much more negative examples than positive examples and this causes problems to SVM-based approaches in this context).
- Some descriptors do no better or even worse than random. Those having a score less than 0.0025 are probably not at all or insufficiently correlated to the targeted concepts. A concept by concept analysis might indicate however that some of them might be useful for some concepts.
- The ranking of the descriptor depends upon the used classifier and vice-versa suggesting that descriptor $\times$ classifier combinations are better than others.
- The LIG_KNNC and LIG_KNNG classifier are identical except that the former optimizes its parameters by cross-validation separately for each concept, while the latter does it globally. The latter is quite often better than the former. This suggests that the former does over-fitting, possibly because the training set is too small.

Previous experiments have shown that combining many weak classifiers can produce a strong classifier, that using classifiers based on very different principles can be very efficient and that even classifiers with a poor individual performance can positively contribute to a

Table 1: Performance of still image descriptors, part 1

	Dims	LIF_SVM	LIG_KNNC	LIG_KNNG	ETIS_SVM
CEALIST_global_cciv	345	0.0171	0.0238	0.0232	0.0107
CEALIST_global_cime	64	0.0093	0.0164	0.0157	0.0072
CEALIST_global_pigment	125	0.0161	0.0257	0.0259	0.0102
CEALIST_global_probe	164	0.0116	0.0214	0.0227	0.0108
CEALIST_global_projection	400	0.0158	0.0260	0.0262	0.0131
CEALIST_global_tlep	576	**0.0410**	**0.0536**	**0.0492**	**0.0457**
CEALIST_local	5000	**0.0410**	**0.0480**	**0.0452**	N/A
CEALIST_scribe	4	0.0060	0.0066	0.0057	0.0011
ETIS_global_lab16	16	0.0130	0.0207	0.0223	0.0089
ETIS_global_lab32	32	0.0167	0.0282	0.0261	0.0146
ETIS_global_lab64	64	0.0196	0.0307	0.0297	0.0150
ETIS_global_lab128	128	0.0213	0.0282	0.0297	0.0205
ETIS_global_lab256	256	0.0222	0.0295	0.0302	0.0159
ETIS_global_labbic16	32	0.0178	0.0248	0.0264	0.0128
ETIS_global_labbic32	64	0.0181	0.0300	0.0283	0.0151
ETIS_global_labbic64	128	0.0225	0.0293	0.0314	0.0188
ETIS_global_labbic128	256	0.0235	0.0295	0.0306	0.0178
ETIS_global_labbic256	512	0.0257	0.0291	0.0317	0.0180
ETIS_global_labm1x3x16	48	0.0306	0.0344	0.0360	0.0202
ETIS_global_labm1x3x32	96	0.0312	0.0369	0.0379	0.0195
ETIS_global_labm1x3x64	192	**0.0346**	0.0386	0.0407	0.0220
ETIS_global_labm1x3x128	384	**0.0384**	0.0359	0.0329	0.0250
ETIS_global_labm1x3x256	768	**0.0358**	0.0330	0.0314	0.0243
ETIS_global_labm2x2x16	64	0.0253	0.0309	0.0324	0.0150
ETIS_global_labm2x2x32	128	0.0272	0.0358	0.0329	0.0152
ETIS_global_labm2x2x64	256	0.0291	0.0418	0.0335	0.0199
ETIS_global_labm2x2x128	512	0.0318	0.0305	0.0315	0.0214
ETIS_global_labm2x2x256	1024	0.0290	0.0339	0.0322	0.0221
ETIS_global_qw16	16	0.0117	0.0204	0.0227	0.0066
ETIS_global_qw32	32	0.0149	0.0250	0.0243	0.0123
ETIS_global_qw64	64	0.0145	0.0264	0.0266	0.0186
ETIS_global_qw128	128	0.0203	0.0349	0.0351	0.0236
ETIS_global_qw256	256	0.0229	0.0354	0.0353	0.0281
ETIS_global_qwbic16	32	0.0121	0.0204	0.0237	0.0104
ETIS_global_qwbic32	64	0.0141	0.0247	0.0254	0.0200
ETIS_global_qwbic64	128	0.0148	0.0268	0.0274	0.0220
ETIS_global_qwbic128	256	0.0213	0.0353	0.0357	0.0294
ETIS_global_qwbic256	512	0.0265	0.0351	0.0359	**0.0316**
ETIS_global_qwm1x3x16	48	0.0183	0.0328	0.0313	0.0162
ETIS_global_qwm1x3x32	96	0.0196	0.0390	0.0360	0.0216
ETIS_global_qwm1x3x64	192	0.0202	0.0411	0.0405	0.0254
ETIS_global_qwm1x3x128	384	0.0273	**0.0450**	**0.0450**	**0.0315**
ETIS_global_qwm1x3x256	768	**0.0349**	**0.0503**	**0.0498**	**0.0329**
ETIS_global_qwm2x2x16	64	0.0127	0.0284	0.0266	0.0163
ETIS_global_qwm2x2x32	128	0.0154	0.0317	0.0307	0.0169
ETIS_global_qwm2x2x64	256	0.0186	0.0359	0.0352	0.0196
ETIS_global_qwm2x2x128	512	0.0236	0.0391	0.0406	**0.0299**
ETIS_global_qwm2x2x256	1024	0.0312	**0.0423**	**0.0451**	**0.0316**
Eurecom_surf	500	0.0273	**0.0438**	0.0362	0.0155
GIPSA_faces_1	1	0.0021	0.0043	0.0047	0.0028
GIPSA_faces_5	1	0.0022	0.0066	0.0063	0.0019
GIPSA_faces_9	1	0.0028	0.0054	0.0049	0.0018
IRIT_BoW-Color	250	0.0066	0.0129	0.0136	0.0092
IRIT_BoW-SIFT	250	0.0233	0.0274	0.0280	0.0192

Table 2: Performance of image descriptors, part 2

	Dims	LIF_SVM	LIG_KNNC	LIG_KNNG	ETIS_SVM
LABRI_BlackAndWhite_Mean	1	0.0039	0.0030	0.0030	0.0010
LABRI_facesOCV_Mean	1	0.0058	0.0078	0.0072	0.0027
LABRI_faces	8		0.0061	0.0046	
LEAR_bow_sift_500	500	**0.0371**	0.0341	0.0313	0.0199
LEAR_bow_sift_1000	1000	**0.0407**	0.0353	0.0364	0.0266
LEAR_bow_sift_2000	2000	**0.0336**	0.0333	0.0388	0.0263
LEAR_bow_sift_4000	4000	0.0310	0.0350	0.0381	**0.0304**
LEAR_bow_sift_8000	8000	0.0285	**0.0407**	0.0458	N/A
LEAR_bow_sift_16000	16000	0.0226	**0.0383**	0.0442	N/A
LIF_Global	104	0.0252	0.0414	0.0407	0.0247
LIF_MMCOL_2x2	36	0.0115	0.0208	0.0230	0.0100
LIF_MMCOL_4x3	108	0.0144	0.0243	0.0289	0.0146
LIF_MMCOL_8x6	432	0.0162	0.0318	0.0335	0.0167
LIF_HISTCOL_2x2	108	0.0165	0.0240	0.0267	0.0111
LIF_HISTCOL_4x3	324	0.0200	0.0335	0.0328	0.0184
LIF_HISTCOL_8x6	1296	0.0291	0.0380	0.0373	**0.0325**
LIF_GABOR_2x2	160	0.0171	0.0294	0.0311	0.0130
LIF_GABOR_4x3	480	0.0225	0.0361	0.0387	0.0237
LIF_GABOR_8x6	1920	0.0302	0.0367	0.0420	0.0278
LIF_HISTEDGE_2x2	200	0.0162	0.0302	0.0295	0.0155
LIF_HISTEDGE_4x3	600	0.0290	0.0379	0.0410	0.0228
LIF_HISTEDGE_8x6	2400	0.0239	**0.0441**	**0.0442**	**0.0325**
LIF_LBP_2x2	1024	0.0227	0.0234	0.0258	0.0163
LIF_SEMANTIC	20	0.0281	0.0410	0.0440	0.0168
LIF_Percepts	3900	**0.0495**	**0.0575**	**0.0639**	**0.0376**
LIF_SIFT_1000	1000		0.0353	0.0364	
LIG_h3d64	64	0.0120	0.0304	0.0295	0.0123
LIG_gab40	40	0.0207	0.0303	0.0294	0.0139
LIG_hg104	104		**0.0433**	0.0429	
LIG_opp_sift_har	4000		**0.0644**	**0.0575**	
LIP6_text	9		0.0031	0.0025	
LSIS_PEF45	45		0.0344	0.0382	0.0200
LSIS_PEF150	150		0.0332	0.0330	0.0285
LSIS_EDGE	72		0.0205	0.0213	
LSIS_GABOR	60		0.0307	0.0328	
LSIS_HSV	63		0.0239	0.0249	

global classifier, especially if they can capture something which is not detected by others. Therefore, any of the above evaluated classifier with a performance significantly higher than the random one can be useful and should be considered in the fusion process.

3 Evaluation of motion descriptors

We evaluated the following motion descriptors:

- **GIPSA_motion**: estimation of the residual motion based on the Motion2D algorithm [12].
- **LABRI_varianceResMov_Mean**: variance of the residual motion, horizontal and vertical components.
- **XLIM-SIC_qmvt_L<n>** : primary (n=1) and secondary (n=2) motions quantity).

Table 3 shows the performance of these descriptors with the four classifiers.

The performance of these descriptors is close to the one on the random prediction but a bit higher for some descriptor × classifier combinations. This is only a global result however and a concept by concept analysis remains to be done. It is likely that motion information is irrelevant for many concepts but can significantly help for a few of them. Also, the evaluated concepts are quite simple. More sophisticated motion descriptors could obtain higher performances.

Table 3: Performance of motion descriptors

	Dims	LIF_SVM	LIG_KNNC	LIG_KNNG	ETIS_SVM
GIPSA_motion.bin	1	**0.0030**	**0.0030**	0.0018	
LABRI_varianceResMov_Mean	2	**0.0039**	**0.0035**	**0.0033**	0.0016
XLIM-SIC_qmvt_L1	1	**0.0031**	0.0023	0.0017	0.0012
XLIM-SIC_qmvt_L2	1	0.0021	0.0021	0.0021	0.0016

Table 4: Performance of audio descriptors

	Dims	LIF_SVM	LIG_KNNC	LIG_KNNG	ETIS_SVM
GIPSA_AudioCS	1	0.0043	0.0023	0.0058	0.0020
GIPSA_audio_intensity	1	0.0025	0.0024	0.0019	0.0015
GIPSA_AudioSpectro	30	0.0048	0.0134	**0.0204**	0.0068
GIPSA_AudioSpectroN	30	0.0051	**0.0188**	**0.0213**	0.0068
GIPSA_spectral_centroid	1	0.0065	0.0028	0.0023	0.0013
GIPSA_spectral_centroid_var	1	0.0043	0.0022	0.0029	0.0013
IRIT_MFCC-Avg-Avg	16		0.0092	0.0122	
IRIT_MFCC-Var-Avg	16		**0.0093**	**0.0109**	
IRIT_MFCC-Var-Min	16		0.0087	0.0082	
IRIT_MFCC-Var-Max	16		0.0101	0.0094	
IRIT_YIN	4		0.0056	0.0055	

4 Evaluation of audio descriptors

We have evaluated audio descriptors from GIPSA and IRIT [13]. These descriptors are not related to the spoken contents but only to the spectral composition of the audio signal. We evaluated the following audio descriptors:

- **GIPSA_AudioCS**: spectral centroid, different method.
- **GIPSA_audio_intensity**: audio intensity in the neighborhood of the key frame, in percentile on the whole video.
- **GIPSA_AudioSpectro**: spectral profile in 30 bands on a Mel scale.
- **GIPSA_AudioSpectroN**: spectral profile in 30 bands on a Mel scale, normalized.
- **GIPSA_spectral_centroid**: spectral centroid.
- **GIPSA_spectral_centroid_var**: standard deviation of the spectral centroid.
- **IRIT_MFCC-Avg-Avg**: Mean of MFCC on homogeneous segments (16-dimensions) mean on TRECVID shot,
- **IRIT_MFCC-Var-Avg**: Standard deviation of MFCC on homogeneous segments (16-dimensions) mean on TRECVID shot,
- **IRIT_MFCC-Var-Min**: Standard deviation of MFCC on homogeneous segments (16-dimensions) minimum on TRECVID shot,
- **IRIT_MFCC-Var-Max**: Standard deviation of MFCC on homogeneous segments (16-dimensions) maximum on TRECVID shot,
- **IRIT_YIN**: Concatenation of four YIN-based descriptors.

Table 4 shows the classification performance obtained using these descriptors.

The performance of these descriptors is significantly above the performance of a random prediction even if these descriptors do not capture any linguistic information about the signal. The performance of these descriptors is significantly lower than the performance of image descriptors (typically three times lower). We can expect a significant contribution from them to the global fusion however since they obviously capture something very different, at least for a few concepts.

5 Relation between descriptor dimension and performance

Figure 1 displays the MAP performance of the considered descriptors according to their number of dimensions (logarithmic scale). A small number of dimension is better for the speed of classification and search tasks that use them.

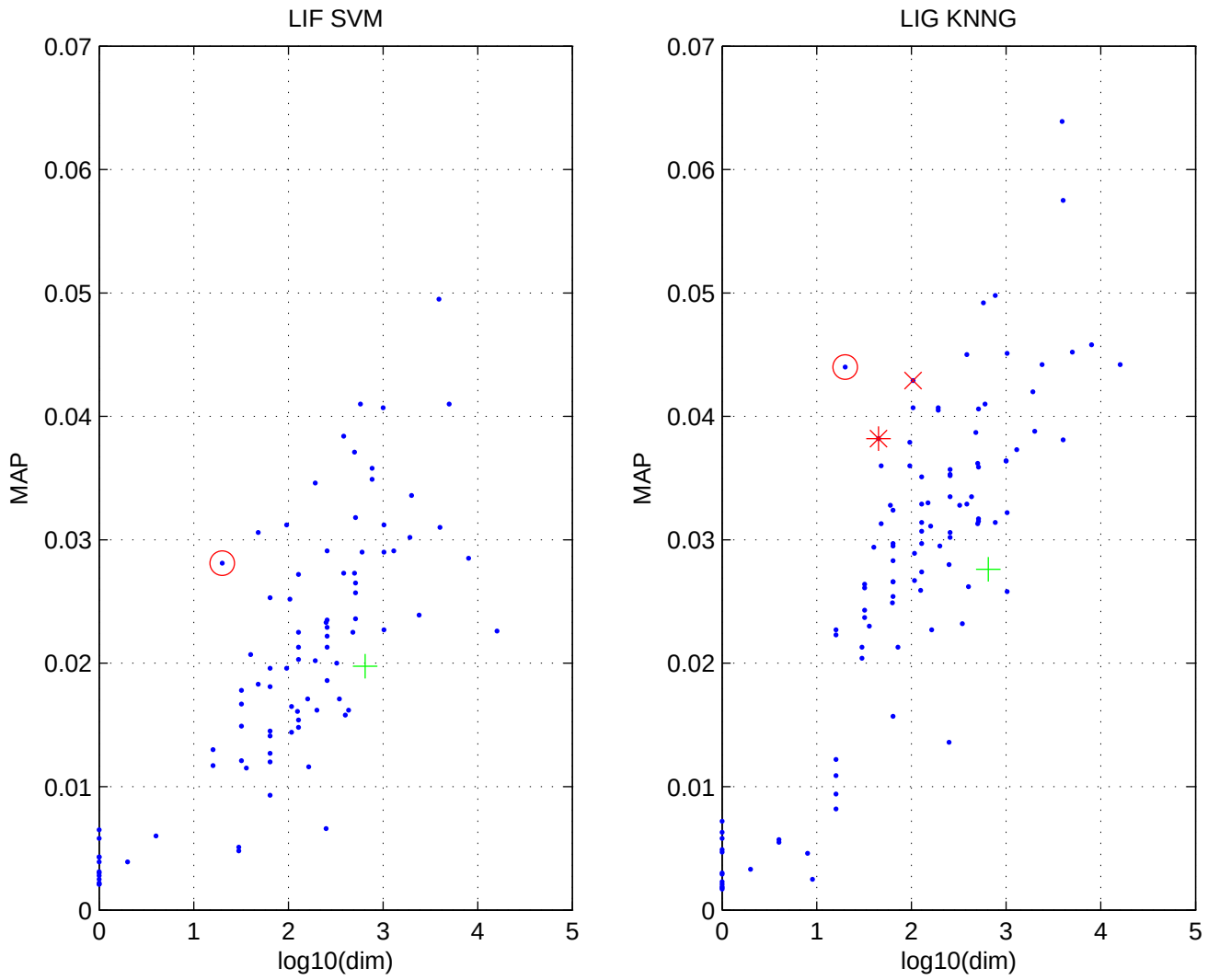

Figure 1: Feature dimensions versus their MAP according to the LIF SVM and LIG KNNG evaluations, for all the IRIM features (when evaluated). The red circle is the LIF Semantic feature. The red '*' is the LSIS PEF45. The red 'x' is the LIG HG 104. The green '+' is the centroide of all the results.

According to the ratio MAP to Dimension, the LIF Semantic features are the most efficient (they are actually outputs of a classifier). The LIG HG104 and the PEF45 are the second best efficient features, yielding to a good compactness and performances. The PEF are new features integrating entropy texture and basic color features that are very fast to compute [8, 9]. More details on TRECVID09 with PEF can be found in [7].

6 Specific work on face detection

For some concepts, such as "Female-human-face-closeup" or "Classroom", the presence of mid-level features, such as faces, is important (see Fig 2).

LaBRI used OpenCV face detector on key frames and filtered the detection with tracking of detected objects. The tracking was performed by multi-object particle filter tracking with consistency check.

In order to evaluate the performance of this detector, we have manually produced the groundtruth for 10 movies randomly selected from the development set. Our experiments show a real improvement when tracking is added. For OpenCV only, we have Recall=0.4386 Precision=0.7547 F-measure=0.5548. For OpenCV and tracking, we get Recall=0.5270 Precision=0.7993 F-measure=0.6352.

Figure 2: Classroom concept: some frames with faces.

7 Multi-modal fusion

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

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

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

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

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

Several methods can be used for the weighting of the classifiers. A uniform weighting is quite often a good choice because everything else tends to overfit the data and to generalize poorly. Another good choice is a weighting based on the individual performance of the classifier, evaluated by cross-validation. A third possibility is to globally optimize the weight for maximizing the global performance evaluated again by cross-validation. Finally, we also tried a weight optimization using a genetic algorithm [15]. All these methods can lead to an overfit, especially if they are applied separately for each concept.

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

Table 5: Official TRECVID 2009 submissions and results

Run	MAP	Description
A_IRIM_RUN1_1	0.1194	Genetic fusion of runs IRIM3, IRIM4, IRIM5 and IRIM6 with Context
A_IRIM_RUN2_2	0.1189	Genetic fusion of runs IRIM3, IRIM4, IRIM5 and IRIM6
A_IRIM_RUN3_3	0.0992	Genetic fusion of KNN classifiers on numerous visual and audio features
A_IRIM_RUN4_4	**0.1220**	Genetic fusion of SVM and KNN classifiers on selected visual and audio features
A_IRIM_RUN5_5	0.1116	Genetic fusion of SVM classifiers on selected visual and audio features
A_IRIM_RUN6_6	0.1014	Genetic fusion of KNN classifiers on selected visual and audio features
A_LIF_RUN1_1	0.0998	Genetic fusion of SVM classifiers on selected visual features with Context
A_LIF_RUN2_2	0.0972	Genetic fusion of SVM classifiers on selected visual features
A_LIF_RUN3_3	**0.1317**	Late fusion of runs IRIM3, IRIM4, IRIM5, and IRIM6
A_LIF_RUN4_4	0.0929	Late fusion of SVM and KNN classifiers on selected visual and audio features with Context
A_LIF_RUN5_5	0.0924	Late fusion of SVM and KNN classifiers on selected visual and audio features
A_LIF_RUN6_6	0.0943	Late fusion of SVM classifiers on selected visual features
A_LIG_RUN1_1	0.1269	Late fusion of runs LIG3 and LIG6
A_LIG_RUN2_2	**0.1276**	Late fusion of runs LIG4 and LIG6
A_LIG_RUN3_3	0.1047	Late fusion of run LIG4 plus face detection
A_LIG_RUN4_4	0.1042	Late fusion of run LIG5 plus audio features
A_LIG_RUN5_5	0.1002	Late fusion of KNN on various visual features
A_LIG_RUN6_6	0.1165	Late fusion of SVM on visual and audio features plus face detection

eral corresponding strategies were tried and validated within the TRECVID 2007 collection. We obtained the following results:

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

8 Official TRECVID 2009 submissions and results.

The IRIM consortium has submitted six runs. Other related runs have been submitted by LIG and LIF as individual participants [16, 15]. All these runs are described in Table 5. The runs are named IRIM, LIG or LIF, depending upon they were actually submitted by the IRIM consortium, by LIG or by LIF. The run names also include a priority number corresponding to our prediction of performance from the best to the worse.

The best IRIM submission has a performance of 0.1220 and a LIF submission including IRIM runs has a performance of 0.1317 while the best performance was of 0.2285 and the median performance was of 0.0516. The results confirmed that:

- Genetic fusion is a good way of choosing weights.
- Hierarchical fusion is also a good way of choosing weights.
- Fusion of the classification scores from different classifier types improves the performance.
- Even with a quite low individual performance, audio descriptors can help.

A few bugs were discovered after the submission. After they have been corrected, the run performance are slightly better. Table 6 shows the original and corrected results for LIG runs.

Table 6: Official TRECVID 2009 submissions and corrected results

Run	Submitted	Corrected
A_LIG_RUN1_1	0.1269	0.1352
A_LIG_RUN2_2	0.1276	0.1358
A_LIG_RUN3_3	0.1047	0.1221
A_LIG_RUN4_4	0.1042	0.1221
A_LIG_RUN5_5	0.1002	0.1189
A_LIG_RUN6_6	0.1165	0.1171

9 Acknowledgments

This work has been carried out in the context of the IRIM (Indexation et Recherche d'Information Mul-

timédia) of the GDR-ISIS research group from CNRS. This work was also partly realized as part of the Quaero Programme funded by OSEO, French State agency for innovation.

References

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

[2] R. O. Stehling , M. A. Nascimento , A. X. Falco. A compact and efficient image retrieval approach based on border/interior pixel classification. Proceedings of the eleventh international conference on Information and knowledge management, pp 102-109, McLean, Virginia, USA 2002

[3] Special issue on Optical Character Recognition, La Recherche, 186, october 1981.

[4] Pass, G., Zabih, R. Ramin, Histogram Refinement for Content-Based Image Retrieval, Proceedings Third IEEE Applications Of Workshop On Computer Vision pp 96-102, Dept. of Comput. Sci., Cornell Univ., Ithaca, NY ; 1996.

[5] Christophe Millet. Annotation automatique d'images : annotation cohrente et cration automatique d'une base d'apprentissage, PhD thesis, 2008, ENST, Paris.

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

[7] Glotin H., Zhao Q., Dumont E., LSIS TREC VIDEO 2009 High Level Features Retrieval using Compact Profile Entropy, In TREC2009 notebook, Gaithersburg, USA, 16-17 Nov. 2009.

[8] Tollari S., Glotin H., Learning optimal visual features from web sampling in online image retrieval, in Proc. IEEE ICASSP, vol. 4p, Las Vegas, 2008.

[9] Glotin H., Zhao Q., Ayache S., Efficient Image Concept Indexing by Harmonic and Arithmetic Profiles Entropy, In IEEE International Conference on Image Processing, Egypt, 2009

[10] M. Nilsson, J. Nordberg, I. Claesson, Face Detection using Local SMQT Features and Split Up SNoWClassifier, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), (2007), pp. 589-592, Honolulu.

[11] Zhou, Y., Nicolas, H., and Benois-Pineau, J., A multi-resolution particle filter tracking in a multi-camera environment, In IEEE International Conference on Image Processing, Egypt, 2009.

[12] J.-M. Odobez, P. Bouthemy, Robust multiresolution estimation of parametric motion models, Journal of visual communication and image representation, Vol. 6, (1995), pp. 348-365.

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

[14] Wang F. and Merialdo B., Eurecom at TRECVID 2009: High Level Feature Extraction, In TREC2009 notebook, Gaithersburg, USA, 16-17 Nov. 2009.

[15] Ayache S., LIF TREC VIDEO 2009 High Level Feature Extraction Using Genetic Fusion, In TREC2009 notebook, Gaithersburg, USA, 16-17 Nov. 2009.

[16] Safadi B. and Quénot G., LIG at TRECVID 2009: Hierarchical Fusion for High Level Feature Extraction, In TREC2009 notebook, Gaithersburg, USA, 16-17 Nov. 2009.

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[18] 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.