IRIM at TRECVID 2014: Semantic Indexing and Instance Search

Nicolas Ballas[1], Benjamin Labbé[1], 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], Bahjat Safadi[3,6], Nadia Derbas[6], Mateusz Budnik[6], Georges Quénot[6], Boyang Gao[7], Chao Zhu[7], Yuxing Tang[7], Emmanuel Dellandrea[7], Charles-Edmond Bichot[7], Liming Chen[7], Alexandre Benoît[8], Patrick Lambert[8], and Tiberius Strat[8]

[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]UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France
[7]LIRIS, UMR 5205 CNRS / INSA de Lyon / Université Lyon 1 / Université Lyon 2 / École Centrale de Lyon, France
[8]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 2014 semantic indexing (SIN) and instance search (INS) tasks. For the semantic indexing task, our approach uses a six-stages processing pipelines for computing scores for the likelihood of a video shot to contain a target concept. These scores are then used for producing a ranked list of images or shots that are the most likely to contain the target concept. The pipeline is composed of the following steps: descriptor extraction, descriptor optimization, classification, fusion of descriptor variants, higher-level fusion, and re-ranking. We evaluated a number of different descriptors and tried different fusion strategies. The best IRIM run has a Mean Inferred Average Precision of 0.2796, which ranked us 5th out of 15 participants.

For INS 2014 task IRIM participation, the classical BoW approach was followed, trained only with *eastenders* dataset. Shot signatures were computed on one key frame, or several key frames (at 1fps) and average pooling. A dissimilarity, computing a distance only for words present in query, was tested. A saliency map, build from object ROI to incorporate background context, was tried. Late fusion of two individual BoW results, with different detectors/descriptors (Hessian-Affine/SIFT and Harris-Laplace/Opponent SIFT), was used. The four submitted runs were the following:

- Run F_D_IRIM_1 was the late fusion of BOW with SIFT, dissimilarity L_2p, on several key frames per shot, with context for queries, and BOW with Opponent SIFT, dissimilarity L_1p, on one key frame per shot.
- Run F_D_IRIM_2 was similar to F_D_IRIM_1 but context for queries used also for second BoW.
- Run F_D_IRIM_3 was similar to F_D_IRIM_1 but no context for queries used.
- Run F_D_IRIM_4 was similar to F_D_IRIM_2 but using δ_1 dissimilarity[46] (from INS 2013 best run).

We found that extracting several key frames per shot coupled with average pooling improved results. We confirmed than including context in queries was also beneficial. Surprisingly, our dissimilarity performed better than δ_1.

1 Semantic Indexing

1.1 Introduction

The TRECVID 2014 semantic indexing task is described in the TRECVID 2014 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 2014 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 [17].

Eight French groups (CEA-LIST, ETIS, EURECOM, LABRI, LIF, LIG, LIRIS, LISTIC) collaborated to participate to the TRECVID 2014 semantic indexing task. Xerox (XRCE), though not being member of IRIM, also shared descriptors with us.

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 1.2).
2. Descriptor optimization. A post-processing of the descriptors allows to simultaneously improve their performance and to reduce their size (section 1.3).
3. Classification. Two types of classifiers are used as well as their fusion (section 1.4).
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 1.5).
5. Higher-level fusion. We fuse here descriptors of different types, e.g. color, texture, interest points, motion (section 1.6).
6. Re-ranking. We post-process here the scores using the fact that videos statistically have an homogeneous content, at least locally (section 1.7).

This approach is quite similar to the one used by the IRIM group last year [16]. The main novelties are the inclusion of new deep learning based descriptors and improvements in the automatic fusion methods.

1.2 Descriptors

Eight IRIM participants (CEA-LIST, ETIS, EURECOM, LABRI, LIF, LIG, LIRIS and LISTIC) provided a total of 71 descriptors, including variants of a same descriptors. 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. The relative performance of these descriptors has been separately evaluated using a combination of LIG classifiers (see LIG paper [19]). Here is a description of these descriptors:

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

CEALIST/bov_dsiftSC_8192: : bag of visterm[38]. 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 $\rightsquigarrow 1024 \times (1 + 2 \times 2 + 3 \times 1) = 8192$ dimensions: see [18] for details.

CEALIST/bov_dsiftSC_21504: : bag of visterm[38]. Same as CEALIST/bov_dsiftSC_8192 with a different spatial pyramid $\rightsquigarrow 1024 \times (1 + 2 \times 2 + 4 \times 4) = 21504$ 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 [7].

EUR/caffe1000: We used the CAFFE Deep Neural Net [44] developped 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.

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

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

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

LIG/opp_sift_<method>[_unc]_1000: bag of word, opponent sift, generated using Koen Van de Sande's software[8] ⇝ 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/caffeb1000: This descriptor is equivalent to the EUR/caffe1000 one and was also computed using the CAFFE Deep Neural Net [44] but with a different (later) version.

LIG/caffe_fc[6|7]_4096 : This descriptor correspond to the LIG/caffeb1000 one and was also computed using the CAFFE Deep Neural Net [44] but is made of the 4096 values of the last two hidden layers, see [19] for more details.

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 [39] 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 [26]) using the retinal model from [9]). 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 [26] 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 [58].

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 [61].

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 [12].

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

1.3 Descriptor optimization

The descriptor optimization consists of a principal component analysis (PCA) based dimensionality reduction with pre and post power transformations [25]. A L_1 or L_2 unit length normalization can optionally by applied after the first power transformation.

1.4 Classification

The LIG participant ran two types of classifiers on the contributed descriptors as well as their combination, see /citetrec14a 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 /citetrec14a.

1.5 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 1.6. 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 [21, 16].

1.6 Final fusion

The IRIM participant LISTIC worked on the automatic fusion of the classification results (experts). 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 LISTIC and LIMSI automatic fusion methods, along with another fusion method tried in the context of the Quaero group using some of the same classification results, and an arithmetic mean and the best attribute per concept, is given in [37].

We combine all of the available FUSEB experts (71 experts in total) as well as 13 KNNB experts (corresponding to retina-enhanced SIFT/SURF/FREAK Bags of Words [58]), in a concept-per-concept manner, by performing five late fusions in parallel. The first fusion is the agglomerative clustering approach which we have previously seen in [37] and in [15]. The second fusion is based on optimising classification scores by using AdaBoost. The third fusion also uses AdaBoost, but this time attempting to optimize the rankings of video shots instead of their scores. The fourth fusion is a weighted arithmetic mean of the input experts, with weights given by the average precisions of the expert for the semantic concept in question. The fifth fusion consists in taking just the best expert for the concept in question. All of these five fusions are combined, by choosing for the concept in question, the late fusion approach that worked best on the training set. The fusion approach is described in more detail in [62].

1.7 Temporal re-scoring (re-ranking), conceptual feedback and uploader model

At the end, conceptual feedback [27] and temporal re-scoring [24] are performed. For reasons of time constraints, conceptual feedback is performed using information from a manual hierarchical late fusion [17] instead of our own fusions. At the end, information about the uploader of each shot is also included, although with a low weight [63].

1.8 Evaluation of the submitted runs

We submitted 4 runs, each using the same 84 input experts:

- M_D_IRIM.14_1 - the best of the 5 fusion approaches for each concept, followed by conceptual feedback and temporal re-scoring;
- M_D_IRIM.14_2 - similar to the above system, with the addition of the uploader model
- M_D_IRIM.14_3 - the best of the 5 fusion approaches for each concept, followed only by temporal re-scoring;
- M_D_IRIM.14_4 - similar to the above system, with the addition of the uploader model

IRIM officially submitted the four M_D_IRIM.14_1 to M_D_IRIM.14_4 runs that are described in section 1.6. Table 1 presents the result obtained by the four runs submitted as well as the best and media runs for comparison. The best IRIM run corresponds to a rank of 5 within the 15 participants to the TRECVID 2014 main SIN task.

Table 1: InfMAP result and rank on the test set for all the 30 TRECVID 2014 evaluated concepts (main task).

System/run	MAP	rank
Best run	0.3320	1
M_D_IRIM.14_1	0.2590	14
M_D_IRIM.14_2	0.2587	15
M_D_IRIM.14_3	0.2449	20
M_D_IRIM.14_4	0.2460	19
Median run	0.2075	28
Random run	0.0009	-

Table 1 shows the results of our submitted runs. The best run is IRIM.14_1, which is the concept-per-concept selection of the best of the 5 late fusion approaches, with added conceptual feedback and temporal re-scoring. The result is almost identical to IRIM.14_2, which adds the uploader model, The uploader was only given a small weight, hence the close result, because preliminary testing showed that the uploader can sometimes decrease results.

Thanks to the conceptual feedback, IRIM.14_1 and IRIM.14_2 perform 5.7% better than IRIM.14_3 and 5.2% better than IRIM.14_4. Between IRIM.14_3 and IRIM.14_4, the results differ by less than 0.5%, again due to the small weight of the uploader model. This time however, the run using the uploader model performs slightly better.

2 Instance Search

Given visual examples of entities of limited number of types: person, object or location, Instance Search (INS) task [2] consists in finding shots which contain instances of these entities. Each instance is represented by a few (4) example images. Hence if we can consider the set of video clips as a visual database, the problem consists in retrieval of each instance in this database.

2.1 Related work

The bag-of-words (BoW) model [29] is one of the most effective content-based approaches for large scale image and visual object retrieval. First features are detected on regions of each image and described by a feature descriptor. Feature descriptors are then quantized into visual words, creating a visual vocabulary. A similarity is then computed between quantized vector of query image and database images. At last, a spatial verification step may be performed on the top results ranked by similarity. It may be optionally followed by an automatic query expansion step that uses the verified results to build a new query. BoW is the approach the most often followed by INS participants. To be applied to videos, it requires first to extract key frames from shots. The various aspects of the BoW method have been intensively studied.

For features detection, various detectors are available. The most frequently used and effective detectors are the Harris-affine, Hessian-affine and MSER [45]. For INS 2013, NII [47], participant with best run, used several detectors: Hessian-affine, Harris-Laplace, MSER. PKU [48] also used these three descriptors (and also a Laplace of Gaussian detector). Detected interest regions are then most often described by the SIFT descriptor [40]. Arandjelović and Zisserman [49] propose a small modification, RootSIFT, that improves performances at virtually no cost. NII [47] used RootSIFT for INS 2013. SIFT descriptors also have been extended to color. Among these color descriptors, OpponentSIFT have been found to perform well on some benchmarks [8]. For INS 2013, NII [47] and NTT [51] used a compact color SIFT of dimension 192 (128 for luminance SIFT and 64 for chrominance), PKU [48] used CSIFT.

For vocabulary construction, the method of choice is the approximate k-means, introduced by Philbin *et al.* [34]. Approximate K-means optimizes the step of retrieving nearest neighbors between feature points and cluster centers by using an approximate nearest neighbor technique, such as FLANN [35]. A forest of multiple randomized kd-trees is built over the cluster centers at the beginning of each iteration. This Approximate K-means gives similar results to exact K-means, at a fraction of the computational cost. Most often a vocabulary size of 1 million words is chosen [49, 52]. NII [47, 46] used a vocabulary of 1 million words for INS 2013. Indeed, Philbin *et al.* predicted a drop in performance with vocabularies larger than 1 million. Mikulik *et al.* [53] attribute this result to a too small dataset (16.7M descriptors). They propose an hybrid approach: approximate hierarchical k-means. They test dictionaries up to 64 million words (on 11 billion descriptors), and show that object retrieval performance increase with the size of the vocabulary.

To apply BoW approach to videos, it is first necessary to extract key frames from video shots [29]. For INS 2013, NII [47, 46] extracted key frames uniformly at the rate of 5 frames per second. Other participants with good runs also extracted several frames per shot but a lower rates (2.6fps for NTT [51], 0.3fps for VIREO). Zhu *et al.* [55] also showed performance increases on past INS datasets with increasing sampling

rates (tested up to 3fps). Using multiple key frames requires a fusion operation. NII [47, 46] did an early fusion and used a joint-average scheme or average pooling: each shot is represented by a single BoW vector that is the average of the BoW vectors of its multiple key frames. In [55], they compare different aggregation methods, and find that these early fusion by average pooling gives the best results.

Various methods are also used to compute similarity between BoW vectors of query and database images or shot. Zhu *et al.* [46] introduces a query-adaptive asymmetrical dissimilarity. NII [47] used this dissimilarity for the best 2013 run.

The spatial verification step, followed by automatic query expansion, is a well-studied way to improve results [29, 34, 49, 52]. For INS 2011 and 2012, spatial checking did improve retrieval results on some but not all queries [15, 59]. For INS 2013, few of the best runs used this step favorably. Maybe due to the versatility of queried objects, that may be non-rigid, taken from fairly different viewpoints; it seems that spatial re-ranking did not perform so well. The NII [47] run with spatial re-ranking is their weakest one. CEA-List and IRIM spatial verification tests on INS 2013 also showed no improvement. Zhang *et al.* [59, 57] propose instead a topology checking technique. It improved VIREO results on INS2013. Due to this spatial verification step not working so well and interactive time nature of INS task, few of the best runs in INS 2013 used automatic query expansion. Besides, for INS task, several (four) example images are already available for query. Arandjelović and Zisserman [56] investigated various fusion methods when multiple queries are available. They found that having multiple queries is always beneficial, and late fusion of individual scores obtained for each example image gave the best results. They did only a qualitative evaluation on TRECVID Known-item search 2011 dataset, and found that late fusion with maximum of individual scores (CombMax operator) performed best. On the other hand, Zhu *et al.* [46, 55] found, on various INS datasets, that early fusion of BoW vectors of query images gave better results than late fusion. NII [47] used late fusion with average pooling for queries vectors (similar to what they did on shots key frames).

Another problem when doing visual object retrieval, where we have a Region-Of-Interest (ROI), is to model the background context of the object. Indeed, for certain objects, static objects or locations in particular, the whole image may show the context of the object and thus it may be better to not use only the ROI for the query. For other objects, thus appearing in various contexts, like cars or people, limiting the query to the ROI may be more precise and using the whole image would only bring more noise. Based on human perception, Zhang *et al.* [59, 57] propose a simple weighting model. Features inside ROI are considered in focus and affected a weight of 1. Features outside the ROI are considered out of focus when they are distant from the ROI and thus are down-weighted according to their distance to the center of the ROI. Mikulik [54] also tries to model the context before query expansion, from spatially verified first results.

2.2 IRIM approach

All IRIM members participating in the INS task provide individual results for their methods. They produce an intermediary result where with a similarity $S_{e,i,s}$ computed between each example image (e) of each instance (i) and each video shot (s). We then do a late fusion of these intermediary results to obtain a final result that is similarity $S_{i,s}$ between each instance (i) and each shot (s)

2.3 Members methods

This year only two members of IRIM Consortium, LaBRI and LISTIC, participated in INS task. Their methods are all based on the BoW approach.

2.3.1 LISTIC

Experts detailed in [58] were experimented in the INS task as well as the SIN task. All those experts are Bag of Visual Words based on the Opponent color SIFT signature but applied on a preprocessed video key frame or sequence around the key frame. Preprocessing consists in applying a retina model able to enhance both details (foveal vision) and transient signals (peripheral vision) with various strategies. Key frame preprocessing (SIFT_1024* and SIFT_retina_1024*) and video preprocessing (SIFT_retinaMasking*and SIFT_multiChannels*) appear to provide similar individual average performance (around 0.10 infAp on SIN task, REFER TO PAPER TABLE VALUES) and, when fused, show efficient complementary (0.16 infAp, i.e. +60% improvement, refer to paper table values). Experiments show that such descriptors using a low number of visual words (1024) perform well on the SIN task while keeping a reasonable computational cost. However, this low dimensionality limits efficiency on the INS task thus not being a competing approach. Further visual word clustering strategies must be experimented to adapt to this task.

2.3.2 LaBRI

LaBRI BoW system uses two types of descriptors: SIFT descriptors (of dimension 128) extracted with Hessian-Affine detector [60] and Opponent SIFT descriptors (of dimension 384) extracted with Harris-

Laplace detector [50]. The RootSIFT [49] post-processing step is applied.

Approximate k-means algorithm [34] is then used to compute a vocabulary of k=1 million visual words, for each type of descriptor. The size of the random forest was set to 8 kd-trees. Vocabularies on SIFT and Opponent SIFT descriptors were computed respectively on 117K and 24K randomly selected images from the shots, with one image extracted per shot (that is respectively 25% and 5% of the 470K shots). Hard assignment is used to compute the histogram of visual words occurrences. This vector is then weighted by the tf-idf scheme.

To represent a shot, two approaches were tested. First, only one key frame is extracted per shot. This key frame is chosen arbitrary at the middle of the shot. A BoW vector is computed only for this image. For INS 2013, it represents ∼830M of SIFT for the 470K shots, that is a mean of ∼1.7K SIFT per image, and ∼694M Opponent SIFT for all the shots, that is a mean of ∼1.5K Opponent SIFT per image. Otherwise, several key frames are uniformly extracted per shot, at a given frame rate. A global histogram is computed for all the key frames of the shot and averaged. This is the joint average scheme or average pooling used in [46, 47]. We tested only a frame rate of 1 frame per second. It corresponds to ∼1.57M images. It represents ∼2.96G of SIFT for all the shots, that is a mean of ∼1.8K SIFT per image, and ∼2.59G Opponent SIFT, that is ∼1.6K Opponent SIFT per image. It is noteworthy that NII [47] used a similar approach in INS 2013, but with a much higher frame rate of 5 fps. It corresponds to ∼7.7G images, that would represent between 12.7G and 14.5G features.

In the standard BoW method, each signature vector is first L_m-normalized (with $m = 1$ or $m = 2$). Then a similarity or a distance is computed between the query BoW vector and the database BoW vector to obtain the final ranking. Various similarities have been employed: cosine similarity, histogram intersection, or a similarity computed from a distance. We have tested various combination of L_m-normalization and a L_n distance metric (with $1 <= m, n <= 2$). We used a dissimilarity, noted L_np with $n = 1$ or $n = 2$, that correspond to the L_n distance computed on the non-zero space of the query. The L_n distance is only computed for the words present in the query, that is between the non-zero bins of the query BoW vector and the corresponding bins of the shot BoW vector. Then a similarity $s = \frac{1}{d+\epsilon}$ is computed from this dissimilarity. We also tested the query-adaptive asymmetrical dissimilarity $\delta 1$ proposed by Zhu *et al.* [46], and used in best INS 2013 NII run [47]. See equations 1 and 2 where T_j and Q_i represent respectively test image and query image vectors, weighted by *idf* term. The rationale of this dissimilarity is to penalize features that are detected in the query object region and have no corresponding features in the database image. $\overline{\omega}$ balances the impact of clutter and positive matches in the scoring. It is computed on-the-fly, to adapt to the database and to the query. We used $\alpha_1 = 0.5$ in our tests. Both $\delta 1$ and our L_np dissimilarity can be computed efficiently with the help of an inverted file.

Figure 1: Example of instance: original image (programme material copyrighted by BBC), mask, saliency map generated from mask to weight the features.

$$\delta_1(Q_i, T_j, \overline{\omega}) = \|T_j\|_1 - \overline{\omega}\|min(Q_i, T_j)\|_1 \quad (1)$$

$$\overline{\omega} = \alpha_1 \frac{\sum_{j=1}^{N} \|T_j\|_1}{\sum_{j=1}^{N} \|min(Q_i, T_j)\|_1} \quad (2)$$

For various instances, locations or static objects in particular, it may be interesting to take into account the background context in which they appear. It is expected that using the background context will bring more information than using only the ROI but also more noise. We build a saliency map, or a stare model as called in [59, 57], that will weigh the contribution of points detected on the query image. Similar to [59], we define a function to down-weight the features distant from the object mask or ROI. A point outside the ROI has its weight computed according to its distance to the ROI contour. The further away from the ROI contour, the less its weight. The weighting function is described in 3, where $\delta^2 = -\frac{diag^2}{4ln0.08}$ and $diag$ is the diagonal axis of the query image. For a point x, p is its projection on the ROI contour. That is, p is the closest point of the ROI contour to point x. A point detected inside the ROI has a weight of 1. The figure 1 shows the saliency map obtained with this weighting function on a 2014 topic example image. As we do not use the center of object ROI to compute our weighting function, it will give a more even saliency map that the one described in [59] for elongated objects and/or ROI composed of several connected components.

$$k(x) = \begin{cases} 1 & \text{if } x \in ROI \\ exp(-\frac{\|x-p\|^2}{2\delta^2}) & \text{otherwise} \end{cases} \quad (3)$$

2.4 Late fusion

Each described members method produces intermediary results. Thus for each method (m), we have a similarity $S_{m,e,i,s}$ between each example image (e) of each

Name	Formula
CombMAX	MAX(individual similarities)
CombSUM	SUM(individual similarities)
CombANZ	CombSUM / Number of non zero similarities
CombMNZ	CombSUM * Number of non zero similarities
CombProd	Geometric Mean of individual similarities

Table 2: Definitions of different combination operators

instance (i) and each shot (s). We have to fuse these similarities to obtain a similarity $S_{i,s}$ for an instance (i) and a shot (s). We considered only queries where all the four example images were used. Thus, with four example images (e) for each instance (i), it means that we have to fuse $4m$ similarity to get the similarity $S_{i,s}$ for an instance (i) and a shot (s).

We have tested a limited number combination operators: CombMAX, CombSUM, CombANZ, CombMNZ, CombProd [43], defined in table 2.

These fusion operators can be applied to similarity scores or ranks. A Combination operator will be noted $Comb[S]$ if applied to score, and $Comb[R]$ if applied to rank.

2.5 Results

Here we present various results for different parameters of our methods. The results noted INS 2013 concern 2013 topics on *eastenders* dataset. They were computed (unless otherwise noted) before the submission of our 2014 runs, and helped to define those runs. The results noted INS 2014 concern 2014 topics evaluated, on the same dataset, after 2014 runs submission. They were computed once the ground truth for 2014 topics has been provided by NIST.

Table 3 shows results of LaBRI BoW approach, individually for SIFT and Opponent SIFT descriptors (with RootSIFT applied, 1M vocabulary, tf-idf applied and L1 normalization), computed only on 1 key frame per shot (noted *1kf*). As we use the four available query images per topic, we have to fuse their individual results. We tested different fusion operators, both on similarity scores and ranks. CombProd[S] and CombSUM[R] gave the best results. Here we will only reference CombProd[S] results.

Table 3 allows comparing results using different dissimilarities for different types of query. Results referenced $a\{1..9\}$ correspond to our BoW approach applied on SIFT descriptors (noted *SIFT*). For $a1, a2, a3$, we use our dissimilarity L_1p with distance L1 respectively with the whole image, only the ROI, or the ROI with context obtained with our saliency map used for the queries (respectively noted *image*, *ROI* and *ctxt*). For $a4, a5, a6$, we use our dissimilarity L_2p with the distance L2 squared. Results referenced $b\{1..9\}$ correspond to the same methods but applied to Opponent SIFT descriptors (noted *OppSIFT*).

ref.	description	INS 2013	INS 2014
a1	SIFT 1kf L_1p image	0.0920	0.0811
a2	SIFT 1kf L_1p ROI	0.0456	0.0004
a3	SIFT 1kf L_1p ctxt	0.0884	0.0879
a4	SIFT 1kf L_2p image	0.0897	0.0860
a5	SIFT 1kf L_2p ROI	0.0584	0.0008
a6	SIFT 1kf L_2p ctxt	0.1045	0.1023
a7	SIFT 1kf δ_1 image	0.0765	0.0708
a8	SIFT 1kf δ_1 ROI	0.0392	0.0006
a9	SIFT 1kf δ_1 ctxt	0.0812	0.0811
b1	OppSIFT 1kf L_1p image	0.1171	0.1041
b2	OppSIFT 1kf L_1p ROI	0.0683	0.0868
b3	OppSIFT 1kf L_1p ctxt	0.1110	0.1148
b4	OppSIFT 1kf L_2p image	0.1037	0.0949
b5	OppSIFT 1kf L_2p ROI	0.0777	0.0884
b6	OppSIFT 1kf L_2p ctxt	0.1156	0.1195
b7	OppSIFT 1kf δ_1 image	0.1084	0.0981
b8	OppSIFT 1kf δ_1 ROI	0.0829	0.0891
b9	OppSIFT 1kf δ_1 ctxt	0.1266	0.1105

Table 3: Results on individual methods for descriptors computed on 1 key frame per shot. We compare SIFT and Opponent SIFT descriptors ($a\{1..9\}$ vs $b\{1..9\}$), L_1p, L_2p and δ_1 dissimilarities ($\{a,b\}\{1,2,3\}$ vs $\{a,b\}\{4,5,6\}$ vs $\{a,b\}\{7,8,9\}$), and queries with whole image, only ROI or ROI and context ($\{a,b\}\{3i\}$ vs $\{a,b\}\{3i+1\}$ vs $\{a,b\}\{3i+2\}$ $\forall i \in \{1,2,3\}$).

We see in table 3 that results with Opponent SIFT ($b\{1..9\}$) are systematically better than those for SIFT ($a\{1..9\}$). Results for query with whole image or with context ($\{a,b\}\{1,4\}$ & $\{a,b\}\{3,6\}$) are better than ROI only ($\{a,b\}\{2,5\}$). Context does not always improve the results. L_1p and L_2p give similar results. However L_2p with context (a6 & b6) is always better than L_1p with context (a3 & b3), and often better overall.

Query adaptive asymmetrical dissimilarity δ_1 did not perform well. Results are always inferior to our results using L_1p or L_2p dissimilarities ($\{a,b\}\{7,8,9\}$ vs $\{a,b\}\{6..7\}$). It is in contradiction with Zhu *et al.* [46] findings. We have to investigate further.

Table 4 shows results with the same approaches but with descriptors extracted on several key frames per shot, at the rate of 1fps and average pooled to build the BoW vector of the shot (noted *1fps*).

Results for Opponent SIFT ($d\{1..9\}$) on INS 2013 were not computed on time for runs submission.

Comparing results in tables 3 and 4, we see that results with several key frames per shot are always better than with only one key frame per shot. It is consistent with observations by Zhu *et al.* [55] that increasing number of key frames per shot (to at least to 3 fps) increases performances. Otherwise, same observations made for table 3 may be done for table

ref.	description	INS 2013	INS 2014
c1	SIFT 1fps L_1p image	0.1370	0.1243
c2	SIFT 1fps L_1p ROI	0.0617	0.0010
c3	SIFT 1fps L_1p ctxt	0.1435	0.1450
c4	SIFT 1fps L_2p image	0.1357	0.1411
c5	SIFT 1fps L_2p ROI	0.0828	0.0029
c6	SIFT 1fps L_2p ctxt	0.1682	0.1640
c7	SIFT 1fps δ_1 image	0.0991	0.1021
c8	SIFT 1fps δ_1 ROI	0.0659	0.0019
c9	SIFT 1fps δ_1 ctxt	0.1020	0.1067
d1	OppSIFT 1fps L_1p image	0.1610	0.1648
d2	OppSIFT 1fps L_1p ROI	0.0915	0.1096
d3	OppSIFT 1fps L_1p ctxt	0.1610	0.1829
d4	OppSIFT 1fps L_2p image	0.1473	0.1545
d5	OppSIFT 1fps L_2p ROI	0.1096	0.1176
d6	OppSIFT 1fps L_2p ctxt	0.1675	0.1829
d7	OppSIFT 1fps δ_1 image	0.1347	0.1300
d8	OppSIFT 1fps δ_1 ROI	0.1110	0.1193
d9	OppSIFT 1fps δ_1 ctxt	0.1522	0.1409

Table 4: Results on individual methods for descriptors computed on several key frames per shot, extracted at 1fps. We compare SIFT and Opponent SIFT descriptors ($c\{1..9\}$ vs $d\{1..9\}$), L_1p and L_2p and δ_1 dissimilarities ($\{c,d\}\{1,2,3\}$ vs $\{c,d\}\{4,5,6\}$), and queries with whole image, only ROI or ROI and context ($\{c,d\}\{3i\}$ vs $\{c,d\}\{3i+1\}$ vs $\{c,d\}\{3i+2\}$ $\forall i \in \{1,2,3\}$).

4: Opponent SIFT is better than SIFT, and queries with whole image or context are better than with ROI only. However, here, using context for query always improves the results ($\{c,d\}3$ vs $\{c,d\}\{1,2\}$, and $\{c,d\}6$ vs $\{c,d\}\{4,5\}$). Once again L_2p with context (c6 & d6) gives the best overall results.

We then tried to fuse these individuals results two by two, using one result on SIFT and one on Opponent SIFT. As we used all four example images for each query, it means that for each shot we had (at most) eight results to fuse. In table 5 we present some of these results. First we present results of fusions on individual methods obtained with one key frame per shot ($f1$ to $f4$). Then, results of fusion of individual methods obtained for key frames extracted at 1fps for SIFT and 1 key frame per shot for Opponent SIFT ($f5$ to $f8$). At last, we display results of fusion of methods obtained for key frames extracted at 1fps for both SIFT and Opponent SIFT ($f9$ to $f12$). As $d\{1..9\}$ results were obtained after submission deadline, $f9$ to $f12$ results could not be submitted as runs. We used $f5$ to $f8$ as runs : respectively as run3, run2, run1 and run4.

For all the fusion results presented in table 5, except the first one, the late fusion of two results is better than the individual results. That is, $f_k = CombProd[s](m_i, m_j) > MAX(CombProd[s](m_i), CombProd[s](m_j))$.

The use of context is systematically better than using the whole image (f_{2i+1} vs f_{2i}, but not for f7 & f8).

ref.	description	INS 2013	INS 2014
f1	a1+b1	0.1156	0.1067
f2	a3+b3	0.1162	0.1154
f3	a4+b4	0.1056	0.1047
f4	a6+b6	0.1321	0.1303
f5	c4+b1 = run3	0.1443	0.1556
f6	c6+b3 = run2	0.1766	0.1741
f7	c6+b1 = run1	0.1884	0.1763
f8	c9+b9 = run4	0.1190	0.1381
f9	c1+d1	0.1614	0.1681
f10	c3+d3	0.1751	0.1840
f11	c4+d4	0.1510	0.1704
f12	c6+d6	0.1964	0.2054

Table 5: Results of fusion of two individual methods.

2.6 Conclusions

This year IRIM participation to INS task brought useful information

- As highlighted by NII 2013 participation, extracting several key frames per shot and doing an early fusion on their BoW vectors, by average pooling for example, at the shot level, allows to greatly improves retrieval results on videos. As we extracted only one frame per second this year, we want to investigate further the use of more key frames per shot. It will however depend on the availability of larger computing resources. Indeed, extracting key frames at a rate of 5 fps, like NII [47] did in 2013, would require to handle around five more features that we have. It would represent around 15 billion features. Therefore, it would certainly also be beneficial to use a larger vocabulary. Mikulik *et al.* [53] tested vocabularies up to 64 million words for 11 billion features, and found that large vocabularies always improved performances.
- Using the background context of the query object also proved to bring better retrieval performance. The proposed simple weighting function, allowing to incorporate some of the background features in the query BoW, showed better results that ROI only or whole image queries. We have to look more precisely for which topics it improved results, and for which it did not help.
- Surprisingly, our dissimilarity, constructed from a distance applied only on non-zero query space, gave better results than δ_1 query adaptive dissimilarity by Zhu *et al.* [46]. We have to investigate further and in particular check on other datasets.

3 Data sharing

As in previous years, we propose to reuse and extend the organization that has been developed over five years within the members of the IRIM project of the French ISIS national Research Group (see [15] 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/sin12). 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.

4 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).

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 and Georges Quénot, TRECVID 2014 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics In *Proceedings of TRECVID 2014*, Orlando, USA, 10-12 Nov. 2014.

[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] 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.

[8] 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.

[9] 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.

[10] S. T. Strat, A. Benoit, P. Lambert and A. Caplier, Retina Enhanced SURF Descriptors for Spatio-Temporal Concept Detection, In *Multimedia Tools ans Applications,* to appear, 2012.

[11] S. Paris, H .Glotin, Pyramidal Multi-level Features for the Robot Vision@ICPR 2010 Challenge, In *20th International Conference on Pattern Recognition,* pp.2949-2952, 2010

[12] 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.

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

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

[15] 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.

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

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

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

[19] Safadi et al. LIG at TRECVID 2014: Semantic Indexing, In *Proceedings of the TRECVID 2014 workshop*, Orlando, USA, 10-12 Nov. 2014.
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] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News In *Transcription System. Speech Communication*, 37(1-2):89-108, 2002.

[21] 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.

[22] Alice Porebski, Color texture feature selection for image classification. Application to flaw identification on decorated glasses printing by a silk-screen process. *Phd thesis,* Université Lille 1, Sciences et Technologies, Nov. 2009

[23] V. D. Blondel and J. Guillaume and R. Lambiotte and E. Lefebvre, Fast Unfolding of Community Hierarchies in Large Networks, In *Computing Research Repository,* abs/0803.0, 2008.

[24] 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.

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

[26] 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.

[27] Abdelkader Hamadi, Georges Quénot, Philippe Mulhem. Conceptual Feedback for Semantic Multimedia Indexing, *CBMI 2013, 11th International Workshop on Content-Based Multimedia Indexing,* Veszprem, HUNGARY, jun 2013.

[28] H. Bay, Herbert, T.Tuytelaars,and L. Van Gool. SURF: Speeded Up Robust Features, In *ECCV 2006*, pp 404-417, 2006.

[29] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV'03*, volume 2, pages 1470–1477, 2003.

[30] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–22, 2004.

[31] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, 2006.

[32] L. Liu, L. Wang, and X. Liu, "In Defense of Soft-assignment Coding," in *IEEE International Conference on Computer Vision*, 2011.

[33] R. Arandjelović, A. Zisserman. Three things everyone should know to improve object retrieval, In *CVPR 2012*, 2012.

[34] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman Object retrieval with large vocabularies and fast spatial matching In *CVPR 2007*, 2007.

[35] M. Muja, D. G. Lowe. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration, In *VISAPP'09*, 2009.

[36] Hamming Embedding and Weak geometry consistency for large scale image search, In *ECCV 2008*, 2008.

[37] S. T. Strat, A. Benoit, H. Bredin, G. Quénot and P. Lambert. Hierarchical Late Fusion for Concept Detection in Videos. In *ECCV workshop on Information Fusion in Computer Vision for Concept Recognition,* Firenze, Italy, 13 Oct. 2012.

[38] 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.

[39] 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.

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

[41] R. O. Stehling, M. A. Nascimento, and A.X. Falcão. A compact and efficient image retrieval approach based on border/interior pixel classification In *11th International Conference on Information and Knowledge Management* 2002

[42] E. Fox and J. Shaw Combination of Multiple searches In *Proceedings of the 2nd Text Retrieval Conference* Gaithersburg, USA, 1994

[43] G. Csurka and S. Clinchant An empirical study of fusion operators for multimodal image retrieval In *10th Workshop on Content-Based Multimedia Indexing* Annecy, France, 2012

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

[45] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool. A comparison of affine region detectors. In *IJCV'05*, vol. 65, no 1–2, pp 43–72, 2005.

[46] C.-Z. Zhu, H. Jégou, S. Satoh Query-adaptive asymmetrical dissimilarities for visual object retrieval In *ICCV - International Conference on Computer Vision*, Dec. 2013.

[47] C.-Z. Zhu, H. Jégou, and S. Satoh. NII team: Query-adaptive asymmetrical dissimilarities for instance search. In *Proceedings of the TRECVID 2013 workshop*, Gaithersburg, USA, Nov. 2013.

[48] Y. Peng, X. Zhai, J. Zhang, L. Huang, N. Li, P. Tang, X. Huang, and Y. Zhao. PKU ICST at TRECVID2013 : Instance Search Task. In *Proceedings of the TRECVID 2013 workshop*, Gaithersburg, USA, Nov. 2013.

[49] R. Arandjelović, A. Zisserman. Three things everyone should know to improve object retrieval, In *CVPR 2012*, 2012.

[50] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Empowering Visual Categorization with the GPU. In *IEEE Transactions on Multimedia*, vol. 13, no. 1, pp 60–70, 2011.

[51] M. Murata, H. Nagano, K. Kunio and S. Satoh NTT Communication Science Laboratories and National Institute of Informatics at TRECVID 2013 Instance Search Task In *Proceedings of the TRECVID 2013 workshop*, Gaithersburg, USA, Nov. 2013.

[52] O. Chum, A. Mikulík, M. Perdoch, and J. Matas Total Recall II: Query Expansion Revisited In *Computer Vision and Pattern Recognition (CVPR)*, 2011

[53] 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.

[54] A. Mikulik, Large-Scale Content-Based Sub-Image Search. PhD Thesis, 2014.

[55] C.-Z. Zhu, Y.-H. Huang, S. Satoh Multi-image aggregation for better visual object retrieval In *Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[56] R. Arandjelović, A. Zisserman. Multiple queries for large scale specific object retrieval. In *British Machine Vision Conference*, 2012.

[57] C. Ngo, F. Wang, W. Zhang, C. Tan, Z. Sun, S. Zhu, T. Yao VIREO/ECNU @ TRECVID 2013: A Video Dance of Detection, Recounting and Search with Motion Relativity and Concept Learning from Wild In *Proceedings of the TRECVID 2013 workshop*, Gaithersburg, USA, Nov. 2013.

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

[59] W. Zhang, C.W Ngo Searching Visual Instances with Topology Checking and Context Modeling In *ICMR 2013*, 2013.

[60] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. In *IJCV'04*, 1, pages 63–86, 2004.

[61] 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

[62] Strat, Sabin Tiberius, Benoit, Alexandre, Lambert, Patrick, Bredin, Herv and Quénot, Georges, Hierarchical Late Fusion for Concept Detection in Videos, In *Fusion in Computer Vision, Springer*, 2014.

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