IRIM at TRECVID 2008: High Level Feature Extraction

Hervé Glotin, Zhongqiu Zhao, Stéphane Ayache and Georges Quénot

Abstract

The IRIM group is a consortium of French teams working on Multimedia Indexing and Retrieval. This paper describes our participation to the TRECVID 2008 High Level Features detection task. We evaluated several fusion strategies and especially rank fusion. Results show that including as many low-level and intermediate features as possible is the best strategy, that SIFT features are very important, that the way in which the fusion from the various low-level and intermediate features does matter, that the type of mean (arithmetic, geometric and harmonic) does matter but which is better depends upon the fused sources. Our best run has a Mean Inferred Average Precision of 0.0885, which is significantly above TRECVID 2008 HLF detection task median performance.

1 Introduction

The IRIM group is a consortium of French teams working on Multimedia Indexing and Retrieval [12] in the context of the ISIS coordinated research group. This paper describes our participation to the TRECVID 2008 [13] High Level Features (HLF) detection task [14]. The IRIM consortium also participated to the rushes summarization task but the work done on this task was already described in [11] and is not reproduced here.

Within the consortium, a subset of 4 teams contributes to the HLF detection task: LIG and LIRIS labs provided the framework for HLF detection based on network of operators [1][2]; LSIS lab provided output scores from a system based on PEF features[6]; INRIA LEAR project provided SIFT bag of features [9][15][5]. Our main interest to this year's TRECVID challenge was to compare various ways to combine low-level features and/or systems output. Our system for video shots indexing is fully based on key frames analysis; our generic classification system follows a classical pipeline architecture which includes low-level features extractor, mid-level semantic classification and fusion modules. We used one or more keyframes per shot according to the shot and subshot segmentation provided by Fraunhofer HHI [10]. The following sections present the considered features, the fusion strategies and the submitted runs.

2 Visual analysis

We performed visual analysis at several level of granularity from global to fine blocks analysis, as well as various semantic level. Our low-level feature extractors first split images on overlapped blocks to form a grid of N×M blocks. For our submissions, we chose N and M such as we obtained a satisfying trade-off between classification performance and time computing. Finally, those values depends on which feature is considered. The analysis first treats each key frame to extract several feature vectors, secondly merge them using standard early or late fusion schemes, or a combination of them and then merge key frames to assign a score to each shot.

2.1 Low-level features

At global level, we consider classical color and texture features. Color is represented by a 3-dimensional histogram on RGB space. We discretize the color space to form a 4×4×4 bins histogram. Texture information is described with Gabor bank of filters; we used 8 orientations and 5 scales. Finally, global features are normalized and concatenated on a 104 dimensions vector. We also extracted color and texture features at block levels, features obtained from each block are then concatenated to form a rich description of key frames:

Color (1): is represented by 3×3×3 3D histogram on a grid of 8×6 blocks. The overall local color feature vector has 1296 dimensions.

Color (2): is represented by the first two moments on a grid of 8×6 blocks. This local color feature vector has 432 dimensions.

Edge Direction Histogram: is computed on a grid of 4×3 blocks. Each bin is defined as the sum of the magnitude gradients from 50 orientations. Thus, overall EDH feature has 600 dimensions. EDH feature is known to be invariant to scale and translation.

Local Binary Pattern: [8] is computed on grid of 2×2 blocks, leading to a 1024 dimensional vector.

The LBP operator labels the pixels of an image by thresholding the 3×3-neighborhood of each pixel with the center value and considering the result as a decimal number. LBP is known to be invariant to any monotonic change in gray level.

2.2 Feature on interest points

One of the more relevant feature for visual indexing is the SIFT descriptor combined with a "bag of words" representation. The SIFT descriptor [7] is describes the local shape of points of interest using edge histograms. To make the descriptor invariant, the interest region is divided into a 4×4 grid and every sector has its own edge direction histogram (8-bin). We used a codebook of 1000 visual word, provided by INRIA LEAR.

2.3 PEF feature

The Profile Entropy Features is based on the entropy of pixel projections. These features of 45 dimensions are derived using the projection in the horizontal orientation. A pixel profile can be a simple arithmetic mean in horizontal (or vertical) direction. The advantage of such feature is to combine raw shape and texture representations in a low CPU cost feature. These feature, associated to mean and color STD, reached the second best rank in the official ImagEval 2006 campaign.

2.4 Semantic feature

This feature aims at modeling co-occurrence between high-level features using a "bag of concepts" approach. First, we consider each block from key frames which are relevant for a concept, as relevant for this concept too. This is a very strong assumption but it could be reasonable depending upon the concepts. Thus, we use existing concepts annotations (from a part of the learning set) at global level, to train SVM classifiers at the blocks level, where blocks are represented with moments color and edge direction histogram features. Then blocks of key frames are classified using models of all the concepts, leading to $nb\_blocks \times nb\_concepts$ classification scores per key frame. The final semantic feature is defined by the sum of scores on nb_blocks for each concepts, leading to a $nb\_concepts$ dimensional feature.

3 HLF extraction framework

3.1 Supervised classification

We use Support Vector Machine [4] as binary classifier for supervised classification of HLF with RBF Kernel and probabilistic output scores. We obtain SVM parameter by testing all combination of parameters (σ, c) with 5-fold cross validation. The models are learned using standard annotation provided by LIG Collaborative Annotation by considering up to 800 randomly positive examples and twice as negative randomly selected.

3.2 Early and Late fusion

We merged our various features with combinations of early and late fusion schemes. While the early fusion proceeds in the feature space, the late fusion combines classification scores. Combination of those two schemes is possible when more than two features is available and yields much flexibility on the way to merge features. For example we can combine features with early fusion once then combine with others feature with late fusion. We know from [3] that such combinations outperform, for some concepts, classical early and late fusion.

3.2.1 Early fusion

Early fusion is basically defined by a simple concatenation of the features from the various modalities. While the number of extracted features depends upon the modalities and the type of the features, an early fusion scheme based on simple vector concatenation is much affected by the vector which has the highest number of inputs. Such fusion should have an impact on the classification, especially with a RBF kernel which is based on Euclidian distance. Thus, we normalize each entry of the concatenated vector so that the average norm of each source vector is about the same. The normalization formula is:

$$x_i\prime = \frac{x_i - min_i}{(max_i - min_i) \times \sqrt{Card(x_i)}}$$

where x_i is an input of the feature vector x, min_i and max_i are respectively the minimum and maximum value of the i^{th} input among the training samples and $Card(x_i)$ is the number of dimensions of the source vector of x_i.

3.2.2 Late fusion

The late fusion scheme first classifies each concept using individual modalities and then merges the scores on top of those classifiers. A second layer of classifier can be considered but it does not always conduct to expected performance as it is needed to split training set to learn this classifier while avoiding over fitting. We chose here to use a usual function to combine modalities with neither parameters nor learning phase:

$$v = \frac{1}{N} \sum_{i=0}^{N} log(v_i)$$

where N is the number of modalities and v_i is the score for the i^{th} modality.

3.3 Rank fusion

Rank fusion is a specific case of late fusion. Its principle is to assign to each shot and for each source a score which is equal to its rank according to a classifier that uses this source and then to build a global score for each shot as a mean of the scores (ranks). This global score is then used to re-rank the shots. Three variants of the rank fusion correspond to the use of the arithmetic (classical) mean, the geometric mean and the harmonic mean. The harmonic mean is often considered as the best choice for rank fusion. All three variants can be implemented with weights for the various sources. We tried the three variants with the expected best results from LIG and LSIS in order to compare them.

4 Experiments

4.1 Runs description and results

As IRIM is a consortium of several groups, we list in table 1 the submitted runs of IRIM, LIG and LIRIS groups, we also show the main run (priority 1) of the LSIS group. For more details for this run, please refer to the corresponding TRECVID paper. We show in bold our three best runs out of our 24 submitted runs. The metrics are the Mean Inferred Average Precision on the 20 target concepts (HLFs).

Here are the details of the two main combination of early and late fusion we tried for the runs LIG_1 and LIG_2:

```
LIG_1 = LATE(
   EARLY(Local-Color , EDH , Global-Features),
   EARLY(Semantic , EDH , Global-Features),
   EARLY(Local-Color , SIFT , Global-Features)
 )

LIG_2 = LATE(
   EARLY(Local-Color , EDH , Global-Features),
   EARLY(Semantic , Global-Features),
   EARLY(Local-Color , SIFT)
 )
```

These were empirically determined as good combinations from tests on the development collection.

4.2 Discussion

LIRIS conducted some runs using only a single feature (LIRIS_4, LIRIS_5 and LIRIS_6). As expected, the performance using single features is low and varies according to the feature. LIRIS also tried arithmetic and harmonic mean based rank fusion. Both do better than each of the single feature but, surprisingly, the arithmetic mean leads to much better results than the harmonic mean (usually considered as better).

LIG compared various combinations of features both in the choice of the features and the way they are combined (see above for example of non-trivial combinations for LIG_1 and LIG_2). Their relative performances were judged on the development set and the run priority was defined according to it. The run performance ordering was quite well predicted (only LIG_5 and LIG_6 were swapped).

The first three IRIM runs are rank fusions of the expected best runs from LIG and LSIS. The expected best run was actually the first run from LIG but this was not the case for LSIS (their best run was LSIS_4 0.0525 versus LSIS_1 0.0334). All three rank fusions performed less well than the LIG_1 run, probably because of the large performance difference between LIG_1 (0.0833) and LSIS_1 (0.0334). The geometric and harmonic mean based rank fusion perform in a similar way and much better than the arithmetic mean based rank fusion. This is conforming to the classical expectation (and the geometric mean has a behavior which is intermediate between arithmetic and harmonic means). There finally seems to be no clear rule about which mean is better.

The last three IRIM runs correspond to still other combinations of features, including the best one at 0.0885. As expected, the more low or intermediate features are included, the better are the results. Additionally, SIFT features really seem to be necessary to get at or close to the best performance.

5 Conclusion

We evaluated several fusion strategies and especially rank fusion. Results show that including as many low-level and intermediate features as possible is the best strategy, that SIFT features are very important, that the way in which the fusion from the various low-level and intermediate features does matter, that the type of mean (arithmetic, geometric and harmonic) does matter but which is better depends upon the fused sources. Our best run has a Mean Inferred Average Precision of 0.0885, which is significantly above TRECVID 2008 HLF detection task median performance.

References

[1] S. Ayache and G. Quénot. Image and video indexing using networks of operators. *EURASIP Journal on Image and Video Processing*, Volume 2007, 2007.

IRIM_1	harmonic rank fusion of LIG_1 and LSIS_1	0.0724
IRIM_2	geometric rank fusion of LIG_1 and LSIS_1	0.0736
IRIM_3	arithmetic rank fusion of LIG_1 and LSIS_1	0.0651
IRIM_4	**late fusion of color, global, edge, lbp, sift, semantic**	**0.0885**
IRIM_5	early fusion of sift, color, global	0.0605
IRIM_6	early fusion of sift, color	0.0505
LIG_1	**combination early/late fusion of color, global, edge, sift, semantic**	**0.0833**
LIG_2	**(an other) combination early/late fusion of color, global, edge, sift, semantic**	**0.0793**
LIG_3	early fusion of color, global, edge, lbp, sift, semantic	0.0785
LIG_4	(an other) combination early/late fusion of color, global, edge, semantic	0.0659
LIG_5	early fusion of global, semantic	0.0423
LIG_6	early fusion of color, edge, global	0.0568
LIRIS_1	early fusion of global, edge, semantic	0.0525
LIRIS_2	harmonic rank fusion of LIRIS_1, 4, 5, 6	0.0267
LIRIS_3	arithmetic rank fusion LIRIS_1, 4, 5, 6	0.0598
LIRIS_4	semantic only	0.0254
LIRIS_5	edge only	0.0221
LIRIS_6	lbp only	0.0189
LSIS_1	PEF features	0.0334

Table 1: Runs description of IRIM, LIG, LIRIS and LSIS submissions

[2] S. Ayache and G. Quénot. LIG and LIRIS at TRECVID 2008: High Level Feature Extraction and Collaborative Annotation. In *Proc. of TRECVID Workshop*, 17-18 Nov. 2008.

[3] S. Ayache, G. Quénot, J. Gensel, and S. Satoh. Using topic concepts for semantic video shots classification. In *International Conference on Image and Video Retrieval (CIVR'06)*, 13-15 Jul. 2006.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: A Library for Support Vector Machines*, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[5] M. Douze, A. Gaidon, H. Jegou, M. Marszalek, and C. Schmid. INRIA-LEAR's Video Copy Detection System. In *Proc. of TRECVID Workshop*, 17-18 Nov. 2008.

[6] H. Glotin and Z. Zhao. LSIS TREC VIDEO 2008 High Level Feature Shot Segmentation using Compact Profil Entropy and Affinity Propagation Clustering. In *Proc. of TRECVID Workshop*, 17-18 Nov. 2008.

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

[8] T. Maenpaa, M. Pietikainen, and T. Ojala. Texture classication by multi-predicate local binary pattern operators. In *Proc. 15th International Conference on Pattern Recognition*, pages 951–954, 2000.

[9] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[10] C. Petersohn. Fraunhofer HHI at TRECVID 2004: Shot Boundary Detection System. In *Proc. of TRECVID Workshop*, 15-16 Nov. 2004.

[11] G. Quénot, J. Benois-Pineau, B. Mansencal, E. Rossi, M. Cord, F. Précioso, D. Gorisse, P. Lambert, B. Augereau, L. Granjon, D. Pellerin, M. Rombaut, and S. Ayache. Rushes Summarization by IRIM Consortium: Redundancy Removal and Multi-Feature Fusion. In *Proc. of TRECVID BBC Rushes Summarization ACM Multimedia Workshop (TVS 2008)*, Oct. 2008.

[12] G. Quénot, P. Joly, J. Benois-Pineau, and M. Cord. Projet IRIM : Indexation et Recherche d'Information Multimdia, http://mrim.imag.fr/irim/, 2008.

[13] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[14] A. F. Smeaton, P. Over, and W. Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In A. Divakaran, editor, *Multimedia Content Analysis, Theory and Applications*. Springer Verlag, Berlin, 2009.

[15] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, jun 2007.