Eurecom at TRECVid 2007: Extraction of High-level Features

Rachid Benmokhtar Eric Galmar Benoit Huet
Institut Eurécom - Département Multimédia
2229, route des crêtes
06904 Sophia Antipolis - France
(rachid.benmokhtar, eric.galmar, benoit.huet)@eurecom.fr

ABSTRACT

In this paper we describe our experiments for the high level features extraction task of TRECVid 2007. Our approach is different than previous submissions in that we have implemented a multi-descriptors system. Five (5) experimentations are submitted based on:

- **Run 1:** MPEG-7 global descriptors,
- **Run 2:** MPEG-7 global and GET audio descriptors,
- **Run 3:** MPEG-7 region descriptors using region based automatic segmentation method RBAS (A region merging approach incorporating geometric properties),
- **Run 4:** Color and texture descriptors are extracted using three segmentation methods (A fixed image grid, watersheds and a technique based on minimum spanning trees MST),
- **Run 5:** Combination of global and regions descriptors.

The experimental results show that the performance can be improved with suitable concept models. Secondly, using audio features did not lead to performance improvement in our experiments.

Keywords

Video semantic analysis, multi-level fusion, feature fusion, classifier fusion, neural network, evidence theory.

1. INTRODUCTION

This year the Eurécom Institute submitted a total of five fully automatic runs to the high level features extraction task. Furthermore, the retrieval system was also utilized in the collaborative experiments for the manual search task performed within the K-Space project [1, 2], combining the work done in 9 partner organizations and coordinated by DCU. The rest of the paper is organized as follows. The retrieval system itself and the features used in these experiments are described in Section 2. The experiments submitted for the fully automatic features extraction task are described in Section 3. Conclusions are then presented in Section 4.

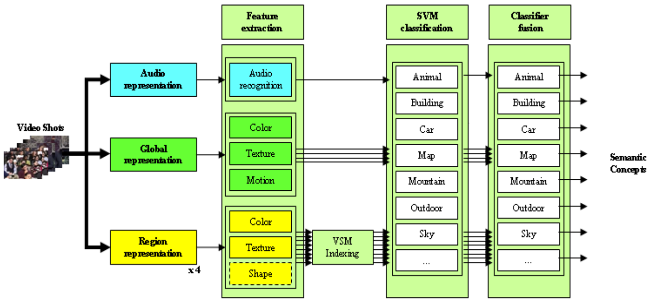


Figure 1: Overall system for high level feature extraction

2. SYSTEM ARCHITECTURE

This section describes the approach used for detecting the presence of concepts within the feature extraction task. Since no particular feature is relevant to all concepts and each concept is *a priori* best described by different features, we build concept detectors by combining different low-level representations including visual representation (image and regions) and semantic audio descriptions.

From these descriptions, we propose to build shot descriptors that are introduced in a SVM-based classification system, which outputs a detection score per low-level feature and per concept. Final concept detectors are obtained by combining the different detection scores within a classifier fusion framework based on the evidence theory [3].

The overall system is illustrated figure 1.

2.1 Video shot representation

For the TRECVid 2007 database, we consider visual and audio description of video shots. To handle the temporal evolution of the shots and improve the robustness of the description, each visual shot is further subdivided into a set of keyframes using a fixed temporal interval (2s). Visual

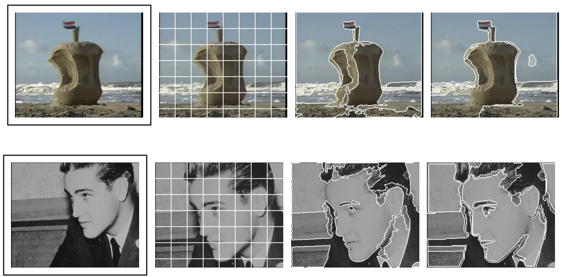
TRECVid'07, November 5–6, 2007, National Institute of Standards and Technology, Gaithersburg, Maryland, USA.

modalities can then be extracted at different granularity levels which are respectively the global shot, the keyframes and the keyframe regions.

At the global level, we consider MPEG-7 color and texture descriptors. Color descriptors are represented by ColorLayout and ColorStructure, and EdgeHistogram is used for texture. These features are extracted for every keyframe and are then aggregated into a single descriptor using median histograms.

Region representations are obtained using four different image segmentation methods: a fixed image grid, watersheds [4], a technique based on minimum spanning trees [5], and a region merging approach incorporating geometric properties [6] which aims to give accurate segmented regions. An illustration of these segmentation approaches is provided in figure 2. For the first three mentioned representations, only color and texture descriptors are extracted, in the same way as [7]. For the last algorithm, we use a set of MPEG-7 descriptors including colour (DominantColor, ColorStructure, ColorLayout), texture (HomogeneousTexture, EdgeHistogram), and shape (ContourShape). These MPEG-7 descriptors are completed by Haralick texture descriptors derived from the pixel coocurrence matrix (StatisticalTexture) [8] and color moments in CIE-LUV space (ColorMoment).

For each region feature we build a dictionary of visual terms by k-means clustering of all training feature vectors. Shot signatures are then obtained using the Vector Space Model (VSM). In this approach, region features are quantized to their nearest visual terms. The signature of each image is then built by counting the number of occurrences of each term in the shot.



(a) Original image (b) Image grid (c) Watershed segmentation (d) MST segmentation

Figure 2: Example of segmentation outputs.

2.2 SVM Classification

SVM is a recent alternative for classification [9]. The hypothesis is the existence of a high-dimensional hyperplane (or, in the general case, a non-linear function) that separates two classes. The original vector space is transformed into a Hilbert space by means of a mapping performed with a function called kernel. The hyperplane is calculated using training vectors. The optimal hyperplane is orthogonal to the shortest line connecting the convex hull of the two classes, and intersects it half-way. To this end, the selected kernel denoted $\mathcal{K}(.)$ is a radial basis function which normalization parameter σ is chosen depending on the performance obtained on a validation set. The radial basis kernel is chosen for his good classification results comparing to polynomial and linear kernels [7].

2.3 Classifier fusion

Classifier fusion is a necessary step to efficiently classify the video semantic content from multiple cues [10, 11, 12]. For this aim, an improved version of RBF neural network based on evidence theory witch we call NN-ET is used, with one input layer L_{input}, two hidden layers L_2 and L_3 and one output layer L_{output} (figure 3). Each layer corresponds to one step of the procedure described in [3]:

1. **Layer L_{in}:** Contains N units (prototypes). It is identical to the RBF network input layer with an exponential activation function ϕ and d a distance computed using training data.

2. **Layer L_2:** Computes the belief masses m^i associated to each prototype. It is composed of N modules of $M+1$ units each. The units of module i are connected to neuron i of the previous layer.

3. **Layer L_3:** The Dempster-Shafer combination rule combines N different mass functions in one single mass.

4. **Layer L_{out}:** We build the normalized output. To take final decision, we compute the maximum of plausibility of each class.

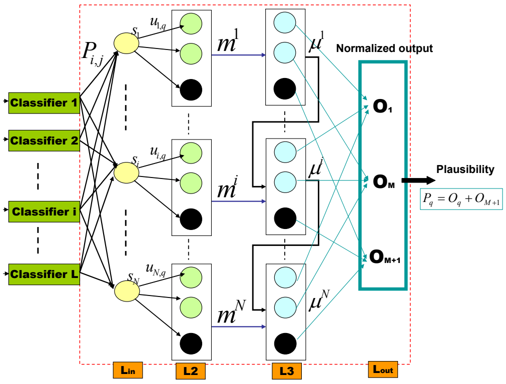


Figure 3: Neural network based on evidence theory (NN-ET) classifier fusion structure

3. EXPERIMENTATIONS

Experiments are conducted on the TRECVid 2007 database [13] of news magazine, science news, news reports, documentaries, educational programs, and archival video. This database is substantially different from the previous year, which will lead us to see how will our methods apply to the new types of content, particularly when we introduce monochrome videos. About 50 hours are used to train the feature extraction system, that are segmented into shots. These shots were annotated with items in a list of 36 labels and 50 hours are used for the evaluation purpose. The training set is divided into two subsets in order to train classifiers and subsequently the fusion parameters. For evaluation, we use the common measure from the information retrieval community: the Average Precision.

This year, the measures are based on 50% random sample of the submission pool. The table 1 gives the name of evaluated concepts.

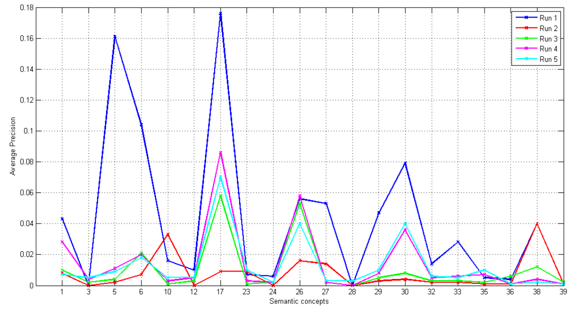


Figure 4: Performance results for the TRECVid'07 concepts chosen.

Our first observation is about the very low number of positive samples in the table 1. We have preferred to use only TRECVid'07 data for training without any introduction of video shots of TRECVid'05-06 in order not to bias the training set. We notice that *Office* (9.1%) , *Meeting* (5.53%) and *Waterscape* (5.75%) concepts have respectable results with an average precision of $(16,1\%, 10,4\%, 17,6\%)$ respectively, given the limits of training data. The second observation shown in figure 4 is about the performance of runs. The run (1) with MPEG-7 global descriptors obtain the best score comparing to regions descriptors runs (3,4) and combination descriptors via the run (5). This is probably due to the choice of the small dictionary size (50 clusters).

The introduction of audio descriptors provided by GET [14] in the run (2) with a global descriptors have a negative behavior due to the difference between audio classes and visual classes. Where the system is more adaptive to detect concepts like *Person* and *Face* as shown in the table 2 (See the numbers of *Speech* audios).

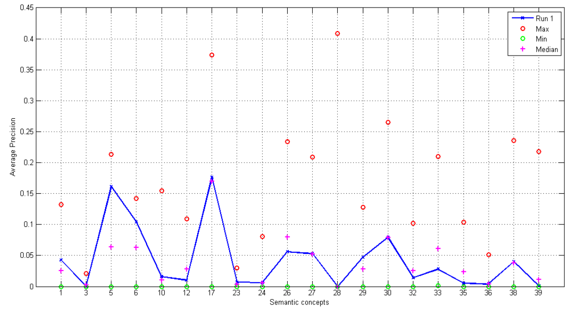


Figure 5: Statistical results of our best submitted run for each concept.

The figure 5 shows some statistical results and our best run (1). *Max*, *Min* and *Median* are computed for all TRECVid'07 participants. It's easy to see the low scores of *Max* for concepts (1,3,6,10,12,23,24,29,32,35,36). It confirm that our results for these concepts are affected by the low number of training sample, as shown in table 1 which does not exceed an average precision of 20%.

For the concept (5), we are closest to the *Max* result, where the number of training samples is acceptable. The comparison of *Median* detection results with our results shown that we have a system able to detect semantic concepts (Except for concepts with low training samples).

For certain semantic concepts such as (29: US flag, 38: Maps, 39: Charts) our system doesn't have a good detection comparing to the *Max* of TRECVid'07 participants. It can be improved with the introduction of new positive samples in the training set or via development of a specific detector.

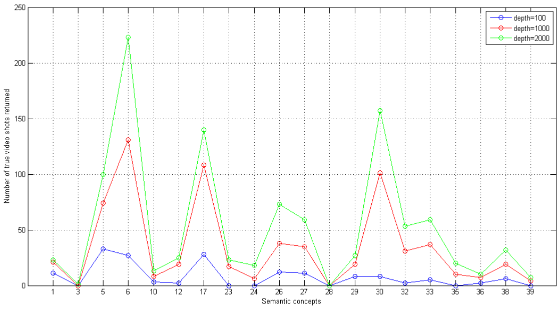


Figure 6: Evolution of video shots returned number at $depth = \{100, 1000, 2000\}$.

Figure 6 shows the evolution of returned video shots using three (3) evaluations, at $depth = \{100, 1000, 2000\}$. For instance, we obtain for concept (5) 33 video shots on the first 100 video shots returned by the system, which explain that 33% are good answers. The same results are noticed for the concepts (6,17). The numbers of video shots can not ensure a good or bad results missing the idea of the numbers of positive samples in the test set.

4. CONCLUSION

In our submission, we adapted an automatic retrieval system described in this paper. The reported system first employs the MPEG-7 global and local features, in order to obtain a compact and effective representation, followed by SVM based classification to solve the challenging task of video shot content detection. Finally, neural network based on evidence theory appears therefore to be particularly well suited for the task of classifier fusion. This approach is based on a feeling of uncertainty to the classification model, considering complete or partial knowledge of the class.

Rather surprisingly, the global MPEG-7 features give best results for many topics comparing to the region MPEG-7 features. This seems to indicate that global features should be considered a valuable information source also for general topics, at least with such difficult topics that were included in this year's high level feature extraction task. Secondly, using audio features did not lead to performance improvement in our experiments.

We start a program of work about ontology study between the classes. Indeed, the concepts are not remotely expressed and a strong correlation exists between certain semantic concepts as *Mountain, vegetation, sky, waterscape, outdoor,....* A first difficulty lives in the elaboration of an ontology describing the existing relations between the concepts. A second difficulty which is of particular interest to

Id	Concepts	Negative	Skipped	Positive	% Posit.
1	Sports	15488	164	200	1.26
3	Weather	15537	254	96	0.60
5	Office	13639	810	1446	9.10
6	Meeting	14421	602	879	5.53
10	Desert	15590	203	88	0.55
12	Mountain	15529	234	109	0.69
17	Waterscape	14719	210	910	5.75
23	Police	15334	167	383	2.41
24	Military	15213	329	362	2.28
26	Animal	14895	190	768	4.84
27	Computer Tv	14877	278	758	4.76
28	US flag	15837	5	12	0.08
29	Airplane	15793	65	50	0.31
30	Car	14948	188	691	0.58
32	Truck	15500	239	128	0.81
33	Boat	15476	113	288	1.81
35	People marching	15443	202	246	1.55
36	Explosion/Fire	15834	13	46	0.29
38	Maps	15684	70	98	0.62
39	Charts	15303	304	254	1.60

Table 1: Id of the TRECVid Concepts

us, is in the exploitation of this semantic information on our classification or fusion system.

Acknowledgments

This work was partially supported by the European Commission under contract FP6-027026 (K-Space).

5. REFERENCES

[1] A. Smeaton, P. Over and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR*, pp. 321–330, California, USA, 2006.

[2] P. Wilkins, T. Adamek, P. Ferguson, M. Hughes, G. Jones,G. Keenan, K. McGuinness, J. Malobabic, N. O'Connor, D. Sadlier, A. Smeaton, R. Benmokhtar, E. Dumont, B. Huet, B. Mrialdo, E. Spyrou, G. Koumoulos, Y. Avrithis,R. Moerzinger, R. Schallauer, W. Bailer, Q. Zhang, T. Piatrik, K. Chandramouli, E. Izquierdo, L. Goldmann, M. Haller, T. Sikora, P. Praks, J. Urban, X. Hilaire, J. Jose. K-Space at TRECVid 2006 In *International Workshop on Video Retrieval Evaluation*, Gaithersburg, USA , November 2006.

[3] R. Benmokhtar and B. Huet,"Neural network combining classifier based on Dempster-Shafer theory for semantic indexing in video content,"*International MultiMedia Modeling Conference*, vol. 4351, pp. 196–205, Singapore, 2007.

[4] L. Vincent and P. Soille, Watersheds in digital space: an efficient algorithm based on immersion simulations. *IEEE Trans. PAMI*, 13(6):583–598, 1991.

[5] P. F. Felzenszwalb and D. P. Huttenlocher, Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[6] T. Adamek, N. O'Connor, and N. Murphy, Region-based segmentation of images using syntactic visual features. In *WIAMIS*, Montreux, Switzerland, 2005.

[7] F. Souvannavong, B. Mérialdo, and B. Huet, Region-based video content indexing and retrieval. In *CBMI*, 2005.

[8] Z. Zheng, Z. Jixian, H. Guoman, and B. Rong, The textural analysis and interpretation of high resolution air sar images. In *ISPRS*, Istanbul, Turkey, 2004.

[9] V. Vapnik, "The nature of statistical learning theory," *Springer*, 1995.

[10] L. Xu, A. Krzyzak, and C. Suen, "Methods of combining multiple classifiers and their application to hardwriting recognition," *IEEE Trans.Sys.Man.Cyber*, vol. 22, pp. 418–435, 1992.

[11] R. Duin and D. Tax, "Experiments with classifier combining rules," *Proc. First Int. Workshop MCS 2000*, vol. 1857, pp. 16–29, 2000.

[12] R. Benmokhtar and B. Huet, "Classifier fusion: Combination methods for semantic indexing in video content," *International Conference on Artificial Neural Networks*, vol. 4132, pp. 65–74, Greece, 2006.

[13] TrecVid, "Digital video retrieval at NIST," *http://www-nlpir.nist.gov/projects/trecvid/.*

[14] G. Richard, M. Ramona and S. Essid, Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hawaii, USA, 2007.

Id	Concepts	Nb Posit.
1	Clean Speech	25409
2	Noisy Speech	63983
3	Music & Speech	30177
4	Pure Music	15569
5	Environmental Sound	7823
6	Pause	13628
7	Airplane	713
8	Applause	541
9	Crowd	3320
10	Dogs	162
11	Explosion	21
12	Gun shot	80
13	Helicopter	9
14	Race car	109
15	Siren	664
16	Truck	77
17	Motorcycle	565
18	Bus	28
19	Car	73

Table 2: Id of the GET audio samples