Eurécom at TRECVid 2006: Extraction of High-level Features and BBC Rushes Exploitation

Rachid Benmokhtar, Emilie Dumont, Benoit Huet and Bernard Mérialdo
Institut Eurécom
Département Communications Multimédia
2229, route des Crêtes - B.P. 193
06904 Sophia-Antipolis cedex - France
{benmokhtar, dumont, huet, merialdo}@eurecom.fr

Abstract

For the four year we have participated to the high-level feature extraction task and we pursued our effort on the fusion of classifier outputs. Unfortunatly a single run was submitted for evaluation this year, due to lack of computationnal ressources during the limited time available for training and tuning the entire system. This year's run is based on a SVM classification scheme. Localised color and texture features were extracted from shot key-frames. Then, SVM classifiers were build per concept on the training data set. The fusion of classifier outputs is finally provided by a multilayer neural network.

In BBC rushes exploitation, we explore the description of rushes through a visual dictionary. A set of non-redundant images are segmented into blocks. These blocks are clustered in a small number of classes to create a visual dictionary. Then, we can describe each image by the number of blocks of each class. After, we evaluate the power of this visual dictionary for retrieving images from rushes: if we use one or more blocks from an image as a query, are we able to retrieve the original image, and in which position in the result list. And finally, we organize and present video using this visual dictionary.

Keywords: *video content analysis, support vector machine, neural network, classifier fusion, shot detection, visual dictionary*

1 First Task : Extraction of High-level Features

1.1 Introduction

The retrieval system we use for the feature detection tasks is functionally similar to that used for TRECVID 2005 [1], but with a different classifier fusion method. This year, we participate to the high-level feature extraction task and we pursued on the fusion of classifier outputs. We use color and texture features extracted around the location of the salient points, then these features are introduced in the SVM classification system built on a per concepts basis according to the dataset. The fusion of classifiers outputs is finally provided by a multilayer neural network.

The paper is organized as follows: the first section presents the low-level features. The second section presents the classifier system. The third section introduces the fusion technique using neural network. It is followed by a presentation of results. Finally we conclude with a brief summary and future work.

1.2 System Architecture

This section describes the workflow of the semantic feature extraction process that aims to detect the presence of semantic classes in video shots, such as building, car, U.S. flag, water, map, etc ...

The segmentation of key-frames is provided either by the algorithm presented in [2] or by the detection of salient points. The latter method first extracts salient points as described in [3]. The idea is to track and keep salient pixels at different scales. We then propose to build two rectangular regions around each salient point, one region on the left and the other on the right for vertical edges and one on the top and the other on the bottom for horizontal edges. The depth of rectangles is proportional to the scale level at which corresponding points were detected. We propose to have smaller rectangles for high frequencies. An illustration of both segmentation approaches is provided on the figure 1.

(a) Region segmentation (b) Salient segmentation

Figure 1: Example of segmentation outputs.

Representative elements are then used as visual keywords to describe video shot content. To do so, features computed on a single video shot are matched to their closest visual keyword with respect to the Euclidean distance (or another distance measures).

Then, the occurrence vector of the visual keywords in the shot is build and this vector is called the Image Vector Space Model (IVSM) signature of the shot. Image

Latent Semantic Analysis (ILSA) is applied on these features to obtain an efficient and compact representation of video shot content. Finally, support vector machines (SVM) are used to obtain the first level classification which output will then be used by the fusion mechanism [4]. The overall chain is presented in figure 4.

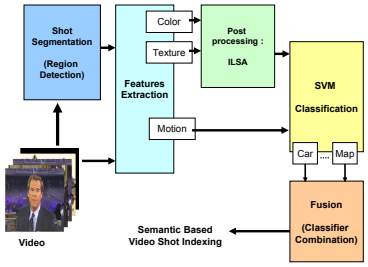

Figure 2: General framework of the application.

1.2.1 Visual features extraction

For the study presented in this paper we distinguish two types of visual modalities: Hue-Saturation-Value color histograms and energies of Gabor's filters [5]. In order to capture the local information in a way that reflects the human perception of the content, visual features are extracted on regions of segmented key-frames [6]. Then, to have reasonable computation complexity and storage requirements, region features are quantized and key-frames are represented by a count vector of quantization vectors. At this stage, we introduce latent semantic indexing to obtain an efficient region based signature of shots. Finally, we combine the signature of the key-frame with the signatures of two extra frames in the shot, as it is described in [4], to get a more robust signature.

1.3 Classifiers

We focus our attention on general models to detect TRECVid features. We have decided to compute a detection score per low-level feature at a first level. The different classifiers fusion presented in the next section will then take care of the fusion of all detection scores at a second level.

The first level of the classification is achieved with support vector machines.

1.3.1 Support Vector Machine

Support vector machine classifiers compute an optimized hyperplane to separate two classes in a high dimensional space. We use the implementation SVMLight detailed in [7]. The selected kernel, denoted $K(.,.)$ is a radial basis function which normalization parameter σ is chosen depending on the performances obtained on a validation set. Let $\{sv_i\}, i = 1, ..., l$ be the support vectors and $\{\alpha_i\}, i = 1, ..., l$ corresponding weights.

Then,

$$D_s(shot_i) = \sum_{k=1}^{k=l} \alpha_k K(shot_i, sv_k)$$

We used the second third of the training set in order to train our SVM models. The last third is used to compute fusion parameters and the first one to test our systems.

1.4 Fusion

Contrary to vote based methods, many fusion methods use a learning step to combine results. The training set can be used to adapt the combining classifiers to the classification problem. Here the optimal combination of classifiers output is obtained by training a multilayer perceptron

1.4.1 Neural Network (NN)

Multilayer perceptron (MLP) networks trained by back propagation are among the most popular and versatile forms of neural network classifiers. In the work presented here, a multilayer perceptron networks with a single hidden layer and sigmoid activation function [8] is employed. The number of neurons contained in the hidden layer is calculated by heuristic. A description of the feature vectors given to the input layer is given in section 2.1.

1.5 Experimentations

Experiments are conducted on the TrecVid'05 (About 85 hours are used to train the feature extraction system, that are segmented into shots. These shots were annotated with items in a list of 39 labels) and TrecVid'06 (250 hours are used for the evaluation purpose) databases [1] of broadcast news videos from US, Chinese, and Arabic sources. The training set is divided into two subsets in order to train classifiers and subsequently the fusion parameters. The evaluation is realized in the context of TrecVid'06 and we use the common evaluation measure from the information retrieval community: the Average Precision.

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

The results obtained in our system are not as good as expected, comparing to the average results, this is can be explained by the small number of features used (due to the technical problems of memory, and computational complexity given the size of the development data). The couple of feature we used was not adapted to concepts detection chosen (figure 3,4). The above mentioned issues have also prevented from testing other fusion approaches that we had implemented and were hoping to evaluate on the TRECvid06 dataset. We will be working hard to be ready for next year event.

1	3	5	6	10
Airplane	Boat/Ship	Bus	Car	Court
12	17	22	23	24
Desert	Gouvernement	Natural Disaster	Office	Outdoor
26	27	28	29	30
Person	Police Security	Prisoner	Road	Sky
32	35	36	38	39
Sports	Urban	Vegetation	Waterscape	Weather

Table 1: Id of the TrecVid Concepts evaluation

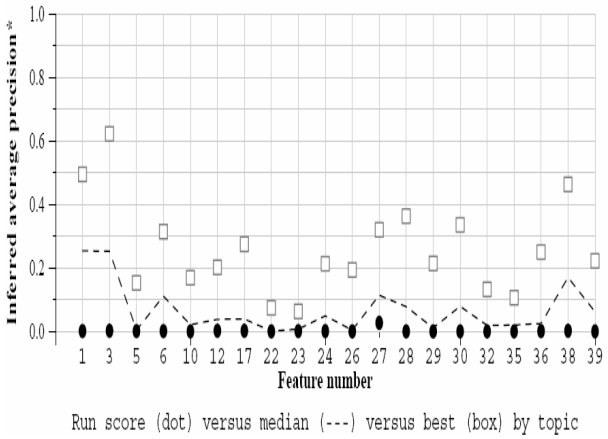

Figure 3: Performance results for the TrecVid'06 concepts chosen.

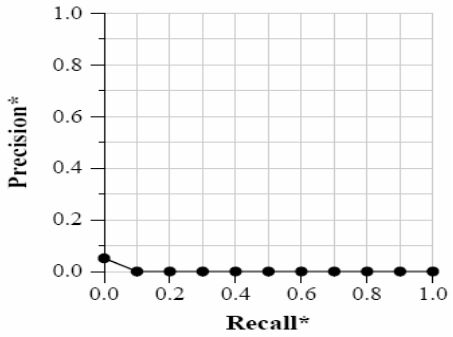

Figure 4: Precision/Recall results.

2 Second Task : BBC Rushes Exploitation

BBC Rushes are a set of video files containing unedited footage recorded for the preparation of video programs. There is no edition, and the data is highly redundant, as typically only 5% of it would be kept in the final program. The task in TrecVid is defined with three objectives:

- remove redundancies so that non-redundant information can be presented,
- organize non-redundant information in some way, so that it can be browsed or searched,
- provide some kind of evaluation for the quality of the results.

Our approach to the BBC Rushes Exploitation Task is to define a fixed number of visual elements to optimally describe the contents of the video files. These visual elements may then be used to organize the video files for presentation, or as keys for searching inside their content. In this section, we describe the various steps of our approach:

- remove redundancies through shot segmentation and hierarchical clustering,
- define a visual dictionary of visual elements,
- optimize the visual dictionary for an artificial search task with automatic evaluation,
- organize and display the visual content based on the visual dictionary

2.1 Extraction of a non-redundant image set

To extract a set of non-redundant images, we first perform a shot boundary detection. We remove the intra-shot redundancy by only selecting the middle keyframe as the representative frame for the shot. Then, we remove the inter-shot redundancy by clustering hierarchically those keyframes. The non-redundant set of images is defined as the set of medioids of the clusters obtained.

2.1.1 Shot boundary detection

We perform shot boundary detection using a method similar to the one proposed in [9]. We consider a sliding window over video frames, with the current frame located in the center. To compute the distance between frames, we build a 16-region HSV histogram for each frame, we drop the 4 central regions, and use the Euclidean distance of the remaining vectors. For hard cuts, we compare the ranking of pre- and post-frames, and we detect a cut when the number of top ranked pre-frames is maximum. For gradual transitions, we compute the average similarity of pre- and post-frames, and we detect the end of a transition when the ratio is minimal. For each shot, we select the central frame as the representative keyframe for this shot.

2.1.2 Hierarchical agglomerative clustering algorithm

Keyframes extracted by the shot boundary detection are classified by a hierarchical agglomerative clustering algorithm. Each image is represented by a HSV histogram and 12 Gabor filters. The distance between two images is computed as the Euclidean distance, and the distance between two clusters is the average distance across all possible pairs of images of each cluster. Figure 5 shows an illustration of the hierarchical agglomerative clustering. When the clustering is finished, we select for each cluster the image which is closest to the centroid of the cluster. Those selected images will compose the set of non-redundant images for the video files.

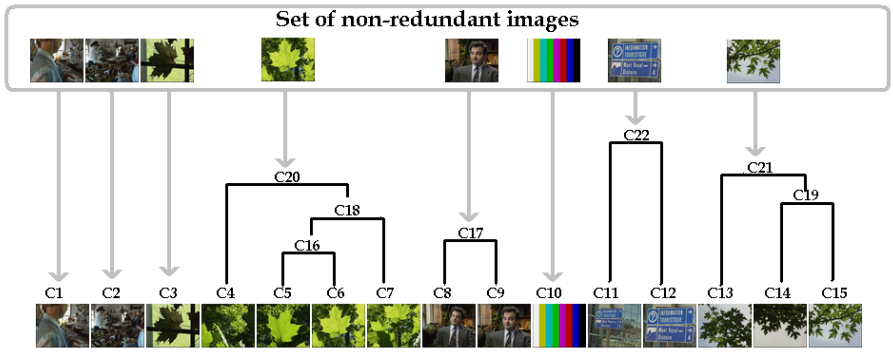

Figure 5: Illustration of the hierarchical clustering

2.2 Visual Elements

Our approach is based on the idea of using a small and fixed number of visual elements. Those visual elements should be automatically computable, so that a video can automatically be described in terms of those visual elements. They should also have some interpretable representation for the user, so that the user can understand the relation between the representation in visual elements and the content of the video, and also that he can select some of those elements to compose a query during a search activity. While a large number of visual elements may be considered, for example the ”indoor/outdoor” attribute could be such a visual element, we focus in the present work on the construction of a visual dictionary of image blocks, either through a color representation or a texture representation. In the following sections, we recall some related work on visual dictionaries, we describe our approach to the construction and the optimization of a visual dictionary, and we propose a method for the evaluation of the visual dictionary.

2.2.1 Construction of the visual dictionary

We start from the set of non-redundant images. Each image is divided into blocks, and for each block we construct the color feature vector (HSV histogram) and the texture feature vector (Gabor filters). We cluster independently the color and the texture vectors using the K-Means algorithm, with a predefined number of clusters (Nc and Nt). We build two dictionaries Dc and Dt by selecting the feature vector which is closest to the centroid for each cluster. Each image block is then associated to one color and one texture visual element.
Finally, we construct our visual dictionary D by selecting the most discriminative vectors: let T be the total number of image blocks, df_v the number of image blocks associated with feature vector v, we define the discriminative power of vector v as $\log(\frac{1+T}{1+df_v})$. The visual dictionary D is composed of the N feature vectors with the highest discriminative power (note that color and texture vectors are mixed in this process).

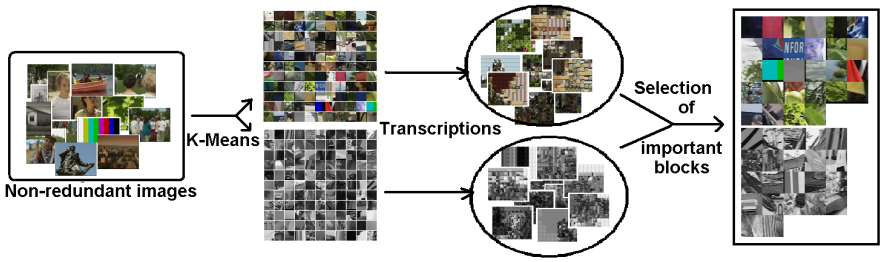

Figure 6: Creation of the visual dictionary

2.2.2 Automatic Evaluation

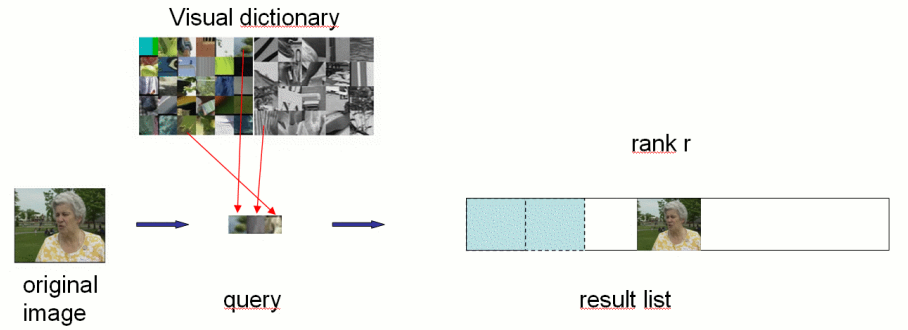

Figure 7: Simulated Search Experiment

We propose to evaluate the quality of the visual dictionary through a Simulated Search Experiment. The experiment works as follows: from the set of non-redundant

images, we select a random one and show it to a user, then ask him(her) to try to retrieve this image through a query based on visual elements only. That is, based on the image, the user should select a number of visual elements to compose a query, which will return a ranked list of images where the original image should be as highly ranked as possible. Figure 7 shows an illustration of this process.

We can simulate this experiment by providing a reasonable mechanism to automatically select visual elements based on the original image. Then, we can define the performance of the vocabulary as the average rank of the original image in the result list. This evaluation can be conducted completely automatically.
To implement this evaluation, we use the set of 1759 non-redundant images. We experiment with different block sizes and different dictionary sizes. From the original image, we construct the query by selecting the two most important visual elements which appear in the image. For example, Figure 8 shows the average rank of the original image for various sizes of dictionaries, in the case where images are split into 15x10 blocks. The X axis indicates the size of the visual dictionaries, while the color bars indicate the size of the intermediate color and texture dictionaries.

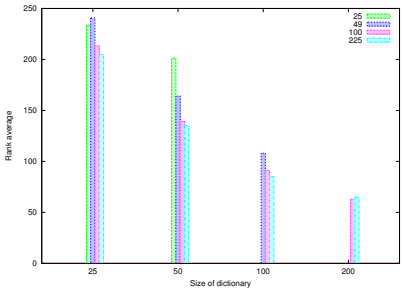

Figure 8: Dictionary evaluation

2.3 BBC Rushes Organization

2.3.1 Presentation

The use of a visual dictionary allows us to describe the visual content of video files in the same way as textual documents are described with words. For example, we can build a tf-idf vector representation in where each component corresponds to a visual element. For a global representation of the set of video files, we propose to rank the visual elements by decreasing discriminative power, then to rank the video files by decreasing importance, where the importance of a video file is computed as the norm of its vector representation. Figure 9 is the illustration of such a presentation. The color is an indication of the importance of each block for each video file.

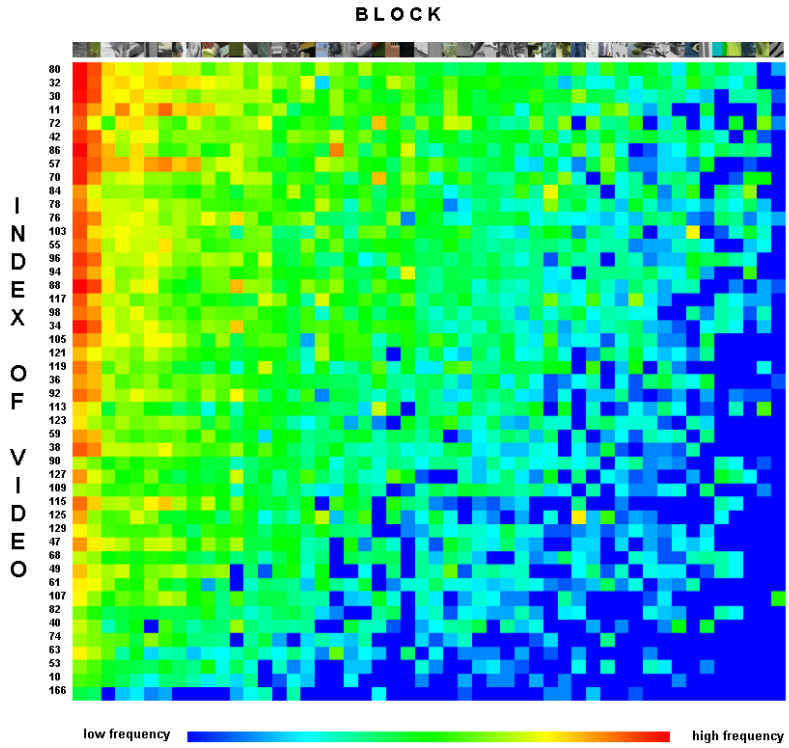

Figure 9: Representation of BBC Rushes

2.3.2 Interactive Search

The visual dictionary may also be used to search for information inside the content of video files. One possible approach is the following (we show simulations of the interface, although we have not built the corresponding system yet). Initially, all video files are available, and represented as a line of micro-icons. Then, the user may select one of the visual elements of the visual dictionary. This identifies a set of video files which contain this visual element, and the icons of those video files are presented in a larger format. In the dictionary, the visual elements which do not appear anymore in the list of selected video files are greyed, so that the user may select another relevant visual element, to filter the selected list further. Figure 10 shows a simulation of this progressive refinement process.

Figure 10: Simulated interface to search among BBC Rushes

3 Conclusion and Future Works

A neural network based approach for classifier output fusion has been implemented and evaluated on this year's TRECVid dataset. A single run was submitted for evaluation due to lack of computing power and implementation issues during the training phase. Unfortunatly, this year's results are not representative of the approach since the fusion is only performed on two visual features. In the previous participations to TRECVid, only visual cues were used by EURECOM to describe shot contents. However it reveals that it was not sufficient to address the difficult problem of semantic content retrieval through the feature extraction task. Future works will mainly concern the fusion mechanism. In particular a neural network based on Dempster Shafer Theory.

For the first year, we participated in BBC Rushes Exploitation, we developed and demonstrated a basic toolkit for support of exploratory search on highly redundant rushes data through a visual dictionary. We proposed a method to evaluate the power of our dictionary. But, our presentation of rushes is not completed and polished, and our organization is just simulated. Then, the future work is to continue to develop the ideas proposed in this paper.

References

[1] TRECVID. Digital video retrieval at NIST. http://www-nlpir.nist.gov/projects/trecvid/.

[2] P Felzenszwalb and D Huttenlocher. Efficiently computing a good segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–104, 1998.

[3] Nicu Sebe and Michael S. Lew. Salient points for content-based retrieval. In *BMVC*, 2001.

[4] Fabrice Souvannavong. Semantic video indexing and retrieval. *PhD Thesis*, 2005.

[5] W. Ma and H. Zhang. Benchmarking of image features for content-based image retrieval. In *Thirtysecond Asilomar Conference on Signals, System and Computers*, pages 253–257, 1998.

[6] C. Carson, M. Thomas, and S. Belongie. Blobworld: A system for region-based image indexing and retrieval. In *Third international conference on visual information systems*, 1999.

[7] T. Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter 11 (Making large-Scale SVM Learning Practical). MIT Press, 1999.

[8] G. Cybenko. Approximations by superposition of a sigmoidal function. In *Mathematics of Control, Signal and Systems*, volume 2, pages 303–314, 1989.

[9] Timo Volkmer, S.M.M. Tahaghoghi, and Hugh E. Williams. RMIT University at TREC 2004. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the TRECVID 2004 Workshop*, Gaithersburg, Maryland, USA, 2004.

[10] R. Picard. Toward a visual thesaurus, 1995.

[11] Rosalind W. Picard. A society of models for video and image libraries. *IBM Systems Journal*, 35(3/4):292–312, 1996.

[12] Ruofei Zhang and Zhongfei (Mark) Zhang. Hidden semantic concept discovery in region based image retrieval. *cvpr*, 02:996–1001, 2004.

[13] Joo-Hwee Lim. Categorizing visual contents by matching visual "keywords". In *VISUAL '99: Proceedings of the Third International Conference on Visual Information and Information Systems*, pages 367–374, London, UK, 1999. Springer-Verlag.

[14] J. Fauqueur and N. Boujemaa. New image retrieval paradigm: logical composition of region categories, 2003.

[15] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. *iccv*, 02:1470, 2003.

[16] Lei Zhu, Aidong Zhang, Aibing Rao, and Rohini K. Srihari. Keyblock: an approach for content-based image retrieval. In *ACM Multimedia*, pages 157–166, 2000.