# Eurécom at TRECVid 2005: Extraction of High-level Features

Joakim Jiten, Fabrice Souvannavong, Bernard Merialdo and Benoit Huet
Département Communications Multimédia
Institut Eurécom
2229, route des crêtes
06904 Sophia-Antipolis - France
(Joakim.Jiten, Fabrice.Souvannavong, Bernard.Merialdo, Benoit.Huet)@eurecom.fr

## Abstract

As past years we participated to the high-level feature extraction task and we pursued on the fusion of classifier outputs. New for this year is an experiment with a two-dimensional Hidden Markov Model. Altogether we submitted seven runs. Three runs was based on the SVM model, another three was based on the HMM and one run was done by fusing HMM results with the SVM. To compile the runs, color and texture features were extracted from shot key-frames. Then, SVM and HMM classifiers were build per concept on the training data set. The fusion of classifier outputs is finally provided either by second level SVM or by hierarchical genetic fusion of possibilities (HGFP) on per concept basis. "A_RO1...1" fuses the output of both classifiers trained on color and texture features using HGFP. "A_RO1...2" fuses the output of SVM classifiers build on color and texture features using SVM. "A_RO1...4" fuses the output of SVM classifiers build on color and texture features using HGFP. "A_RO1...6" fuses using HGFP the output of SVM classifiers build on color and texture features and the output of an SVM trained on both features. The comparison of performances of the fusion systems shows that HGFP can efficiently fuse classifier outputs in a simple mannner. We also noticed that including the fusion at an earlier stage could improve retrieval performances.

This year we took the opportunity to experiment with an early version of the implementation of a context-dependant classifier based on a two dimensional Hidden Markov Model. The HMM-model considers an image as a random process of observations. To provide the observations we divide the image into a grid of blocks where each block presents its color and frequency characteristics. We could observe the problem of a well known drawback of the HMMs, that the output probability plays a more important role than the transition probabilities.

**Keywords**: *video content analysis, support vector machine, genetic algorithm, classifier fusion, possibility theory*

## 1 Introduction

With the growth of digital storage facilities, many documents are now archived in huge databases or extensively shared on the Internet. The advantage of such mass storage is undeniable, however the challenging tasks of automatic content indexing, retrieval and analysis remain unsolved, especially for video sequences. TRECVid [13] stimulates the research in this area by providing standard datasets for evaluation and comparison of new techniques and systems.

The paper is organized as follows: section two presents low-level features. The third section presents the classifiers. Section four introduces our fusion technique using genetic algorithm and the new HMM. It is followed by a presentation of results. Finally we conclude with a brief summary and future work.

## 2 Visual feature

To construct low-level features describing a shot for the SVM models, we extract features on its key frame. Two visual features are selected for this purpose: Hue-Saturation-Value color histograms and energies of Gabor's filters [6]. In order to capture the local information in a way that reflects the human perception of the content [1, 3], visual features are extracted on regions of segmented key-frames. Then, region features are quantized and key-frames are represented by a count vector of quantization vectors to have reasonable computation complexity and storage requirements. At this stage, we introduce latent semantic indexing to obtain an efficient region based signature of shots [11]. Finally we combine the signature of the key-frame with the signatures of two extra frames in the shot, as it is described in [12], to get a more robust signature.

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 [9]. 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.

Our HMM uses continuous observation densities represented by a mixture of five Gaussians:

$$b_j(o) = \sum_{m=1}^{5} c_{jm}\pi[o, \mu_{jm}, \Sigma_{jm}] \qquad 1 \leq j \leq N$$

In the light of this fact it was desirable to use features which are Gaussian distributed and are as much uncorrelated as possible. Further as it is well known that histogram output as features has highly skewed probability distributions, we decided to use HSV means and variances for color descriptors and DCT coefficients for their discriminative ability of energies in the frequency domain. As aforementioned the image was split into blocks (figure 5), for which each there are six color features and 16 DCT coefficients $\{H_\mu, S_\mu, V_\mu, H_\sigma, S_\sigma, V_\sigma, D_{ij} : i, j \in (0, 1, 2, 3)\}$. Or in other words, the image is a vector field $O = \{o_{i,j}\}$; where $o_{i,j}$ is the feature vector extracted in block (i,j) and has a dimension of 22.

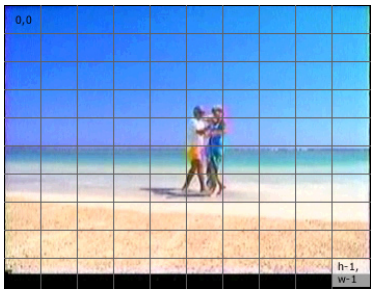


Figure 2: Image decomposed into blocks.

## 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 genetic algorithm 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.

### 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 [4]. The selected kernel, denoted $K(.,.)$ is a radial basis function which normalization parameter $\sigma$ 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.

### 3.2 DTHMM

Conventional block-based classification is based on the labeling of individual blocks of an image, disregarding any adjacency information. When analyzing a small region of an image, it is sometimes difficult even for a person to tell what the image is about. Thus for most images with reasonable resolution; pixels have spatial dependencies which should be enforced during the classification. HMM considers observations (i.e. feature vectors representing blocks of pixels) statistically dependent on neighboring observations through transitions probabilities organized in a Markov mesh, giving a dependency in two dimensions. The state process defined by this mesh is a special case of the Markov Random Field. However, the complexity of the algorithms grows exponentially in higher dimensions, even in dimension two, so that the usage of plain HMM becomes prohibitive in practice [5]. For this reason we use a new type of multi-dimensional Hidden Markov Model: the Dependency-Tree Hidden Markov Model (DTHMM). See Research Report RR-05-128 [7] for a presentation.

The assumption in a 2 D HMM is that the observation sequence was produced by the model, i.e. $P(O|\lambda)$ where O is the observation sequence and $\lambda$ the set of model parameters. The number of states was set as a fixed parameter to sixteen; each one with a Gaussian Mixture Model to represent the continuous observation densities, which had five components. We use a variant of the Baum-Welch algorithms [7] to estimate the model parameters in the training step. To classify an image its low-level features are extracted and then $P(O|\lambda)$ is computed for each model giving a score on how well the model matches the observation, and then search the model with highest a posteriori probability. A general illustration of the classification system is shown in the figure below.

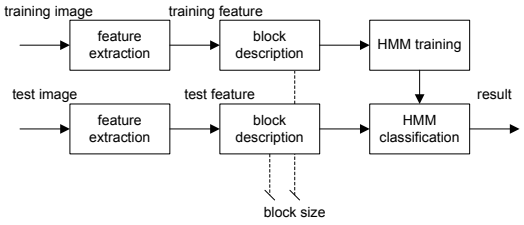


Figure 3: Image classification scheme.

The block description is based on moments of HSV and DCT features as described in the previous section, and the size of the block was set to 8x8 pixels.

## 4 Fusion

In order to combine the output of various classifiers, a fusion algorithm is required. A first approach is to empirically set up a formula to compute the final score using basic operators and functions such as minimum, maximum, sum and product and empiric weights.

Another approach consists in using genetic algorithms to find the best formula using the same operators and a set of weight values. For this purpose we use a hierarchical structure to represent the fusion function. The method is presented in [10] and briefly introduced here.

We assume that the output of an elementary classifier expresses the possibility of its associated class. For example, let $L_{water}(color)$ denote the SVM model that is trained on color ILSA signatures for the class water. Given a shot $s$, the output of $L_{water}(color)$ provides an information about the possibility to have the class water with respect to color features. Possibility logic is then used to achieve the fusion. Different fusion operators exist in this framework: minimum, maximum, t-norm, arithmetic mean, geometric mean, bounded sum, product and probabilistic sum. Each of them have different inference properties. Operators can be conjunctive (highest possibility is preserved), disjunctive (do not favor highest possibility), idempotent

(redundancy is preserved) or reinforcement (redundancy is emphasized).

Unfortunatly, methods to select the right operator do not exist. Moreover, the fusion is conducted by combining two sets of events and when more than two sets are involved the fusion is lead iteratively. In order to find the most appropriate operators and fusion structure, i.e. the order that have to be used to fuse multiple sets, we model the fusion function as a binary tree which is build by a genetic algorithm.

To summarize, the complete fusion chain firstly normalize classifier outputs in $[0, 1]$ thanks to a normalization function; next obtain possibility values are weighted by a priori possibilities; finally, these values are fused with respect to a binary tree and its associated fusion operators. The whole chain is depicted in the figure 4.

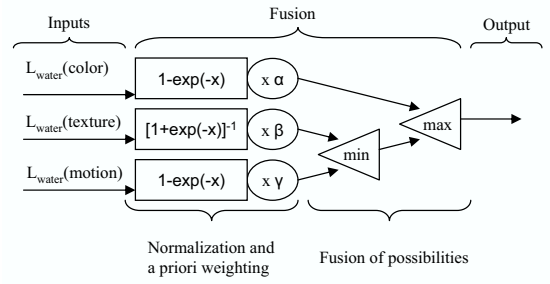


Figure 4: Proposed fusion function.

# 5 Experiments

The classification and the fusion task require annotated data. In June 2005, TRECVid has launched a collaborative effort to annotate a new set of video sequences in order to build a labeled reference database. It is composed of about 80 hours of news videos that are segmented into shots [8]. These shots were annotated with items in a list of 40 labels. The tool described in [14] was used for this time-consuming task. We use this huge annotated database to train classifiers. To train the SVM models the dataset was split into three subsets of equal size. The first one is used to test our systems, the second to train first-level classifiers and the third to compute fusion parameters.

The DTHMM models was trained on the complete training set. We submitted two runs using continuous outputs and one run with discrete output observations. The discrete model uses a vector quantization step to handle multivariate signature vectors, and the continuous model uses a GMM to describe the output probabilities as described in section two. We trained 10 classifiers using the TRECVid training set, one for each semantic class in the High-level feature task. Testing was performed by extracting the images low-level features, compute $P(O|\lambda)$ using the variant of the Baum-Welch algorithms (see [7] section 3.1 ) and return a list of scores for each semantic class.

# 6 Conclusions & Future Work

As in last years, we used visual-, text- and motion features. This year we tried to improve the extraction of salient points, but we believe there is more investigation to be done here. Below is a list of the seven runs in performance order:

1. Genetic Algorithms on all and mixed visual features
2. SVM on all visual features
3. Fusion with Genetic Algorithms on SVM and HMM
4. 16 states continuous DTHMM on HSV and DCT using prior probabilities (block size 44 x 30)
5. 16 states continuous DTHMM on HSV and DCT (block size 44 x 30)
6. 16 states discrete DTHMM on HSV and DCT (block size 44 x 30)

The best model “Genetic Algorithms on all and mixed visual features´´ performed as the median except for “Maps´´, “Flag US´´ and “Sports´´, for which it performed better, at the same time it performed slightly lower than median for “Waterscape´´ and “Mountain´´.

Regarding the DTHMM the effort was to improve semantic classification by examining local context in an image, using a new multidimensional Hidden Markov Model, the Dependency-Tree HMM. We tested the model with the TRECVid collection in order to investigate its performance and to find its point of operation. The results

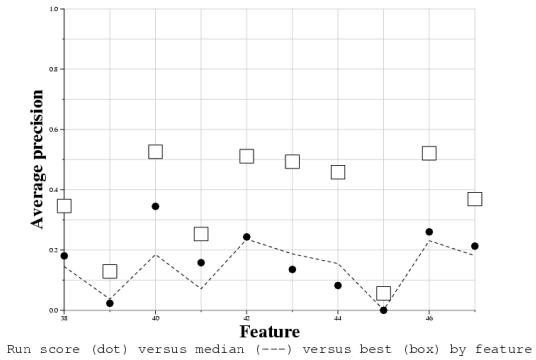


Figure 5: Average precision for best run.

should be considered as preliminary, since there are many parameters involved which have yet not been fully explored. Future works will mainly concern the DTHMM. In particular how the balance between structural information and content description affect the precision in a semantic feature extraction scenario.

# References

[1] Chad Carson, Megan Thomas, and Serge Belongie. Blobworld: A system for region-based image indexing and retrieval. In *Third international conference on visual information systems*, 1999.

[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] Feng Jing, Mingling Li, Hong-Jiang Zhang, and Bo Zhang. An effective region-based image retrieval framework. In *Proceedings of the ACM International Conference on Multimedia*, 2002.

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

[5] R. Levin, E.; Pieraccini. Dynamic planar warping for optical character recognition. volume 3. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992.

[6] Wei-Ying Ma and Hong Jiang Zhang. Benchmarking of image features for content-based image retrieval. In *Thirty-second Asilomar Conference on Signals, System and Computers*, volume 1, pages 253–257, 1998.

[7] B. Merialdo. Dependency tree hidden markov models, research report rr-05-128. Technical report, Institut Eurécom, 2005.

[8] C. Petersohn. Fraunhofer hhi at trecvid 2004: Shot boundary detection system. TREC Video Retrieval Evaluation Online Proceedings, 2004.

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

[10] Fabrice Souvannavong and Benoit Huet. Hierarchical genetic fusion of possibilities. In *Proceedings of the European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, 2005.

[11] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. Video content modeling with latent semantic analysis. In *Third International Workshop on Content-Based Multimedia Indexing*, 2003.

[12] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. Latent semantic analysis for an effective region-based video shot retrieval system. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, 2004.

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

[14] Timo Volkmer, John R. Smith, Apostol (Paul) Natsev, Murray Campbell, and Milind Naphade. A web-based system for collaborative annotation of large image and video collections. In *In Proceedings of the 13th ACM international Conference on Multimedia*, 2005.