Eurécom at Video-TREC 2004: Feature Extraction Task

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

Abstract

Based on the results of last year worldwide submissions to the feature extraction task, we decide to introduce more features to describe the content of shots. In particular, text and motion features are added to existing visual features. The text is by definition a semantic feature, thus it has its importance in the feature extraction task. The motion is necessary to analyze specific features such as *airplane takeoff*. Moreover, to take advantage of the progress of classification systems, support vector machines are used to extract semantic features from low-level features. Finally, genetic algorithms are employed to fuse data from the various classifiers and modalities.

Keywords: *region based indexing, latent semantic indexing, video content analysis, k-nearest neighbor classification, support vector machine classification, genetic algorithm fusion*

1 Introduction

With the growth of numeric 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. Video-TREC [1] stimulates the research in this area by providing standard datasets for evaluation and comparison of new techniques and systems. Based on the analysis of last year submissions to Video-TREC, we introduce more features to describe the content of shots. In particular, text and motion features are added to existing visual features. Moreover, to take advantage of the progress of classification systems, support vector machines are used to extract semantic features from low-level features. Finally, genetic algorithms are employed to fuse data from the various classifiers and modalities.

The paper is organized as follows: the first section presents low-level features. The second section presents k-nearest neighbor and support vector machine classifiers. The third section introduces our fusion technique using genetic algorithm.

2 Shot features

We distinguish three types of features: visual, text and motion features.

2.1 Visual feature

To describe the visual content of a shot, 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 [7]. In order to capture the local information in a way that reflects the human perception of the content [2, 5], visual features are extracted on regions of segmented key-frames [3]. 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 [9], to get a more robust signature.

2.2 Text features

The text or voice are important features. They help to bridge the gap from low-level features to the semantic content by providing a direct information about the semantic content. Text features are based on the automatic speech recognition text provided by LIMSI [4].

First of all, words are stemmed with the widely used Porter's algorithm [8]. Then a dictionary of 2,000 words is created and shots are described by a count vector of the dictionary entries. However, a shot is not a semantic unit, then few words occur in a shot and relevant words might be in surrounding shots. To deal with this synchronization problem, basic text signatures of surrounding shots are included into the current shot signature. This is equivalent to compute a signature over a scene defined as the set of shots that surround the current shot.

2.3 Motion features

For some features like *basket scored, people walking/running, violence* or *airplane takeoff*, it is useful to have an information about the activity present in the shot. Two features are selected for this purpose: the camera motion and the motion histogram of the shot. For sake of fastness, these features are extracted from MPEG motion vectors. The algorithm presented in [10] is used to estimate the camera motion of a frame. The camera motion is approximated by a six parameter affine model. We then compute the average camera motion over the shot. The estimated camera motion is subtracted from macro-block motion vectors to compute the 64 bins motion histogram of moving objects in a frame. Then, the average histogram is computed over frames of the shot.

3 Classifiers

We focus our attention on general models to detect Video-TREC 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 classification is achieved with either the k-nearest neighbor classifier or the support vector machine classifier. In the particular case of text features, we also propose to compute a detection score based on a set of keywords per concept.

3.1 K-nearest neighbors

Since we have no information about the distribution shape of the data, we find natural to use the K-NN classifier as a baseline. Given a shot i, its N nearest neighbors in the training set are identified $(trshot_k), k = 1..N$. Then it inherits from its neighbors a detection score as follows:

$$D_f(shot_i) = \sum_{k=1}^{k=N} cosine(shot_i, trshot_k) * D_f(trshot_k)$$

Where detection scores of training shots, $trshot_k$, are either 1 if the concept f is present or -1 if not.

In order to optimize classifier performances, the algorithm finds the most appropriate number of neighbors for each couple formed by a low-level and a semantic feature. In the particular case of visual features, it also seeks for the best number of factors to be kept by the latent semantic indexing method [9].

K-NN classifiers were trained for all available low-level features: visual, text and motion features.

3.2 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 [6]. 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,

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

SVM classifiers are only trained on visual features.

3.3 Keywords detection

Using full text features as described in section 2, do not provide good classification performances with a k-NN classifier. The idea to efficiently use the text is then to identify important keywords for each concept and then compute a detection score based on the list of important keywords.

First of all, from training data we extract most occurring stemmed words for each concept. Manually we select words that are really related to the concept. Then, we estimate the probability that words related to a concept appear in surrounding shots. This a priori probability is further used to compute the final score. Let $P_f(shot_i+t)$ the probability to detect the concept f in the shot at $(i+t)$. Let $d_f(shot_i)$ the number of times words associated to the concept f occurs in the shot. Then

$$D_f(shot_i) = \sum_{t=-N}^{t=N} P_f(shot_i+t) \times d_f(shot_i+t)$$

4 Fusion and experiments

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.

Figures 1 and 2 show the evaluation result of the presented system. In most cases, the genetic algorithm improves retrieval performances. However combining all features does not always perform the best. The main explanation is that the validation set used for training was not fully representative of the test set.

General performances fluctuate around the mean performance of worldwide submitted systems to Video-TREC. An exception is the *basket scored* feature (numbered 33) where performances are reaching a mean precision of 0.4. Yet, this particular feature was trained using all shots containing the scene feature *basket* in the development set of the year 2003.

References

[1] http://www-nlpir.nist.gov/projects/trecvid/.

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

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

[4] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

[5] Feng Jing, Mingling Li, Hong-Jiang Zhang, and Bo Zhang. An effective region-based image retrieval framework. In *ACM Multimedia*, 2002.

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

[7] 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.

[8] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

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

[10] Roy Wang and Thomas Huang. Fast camera motion analysis from MPEG domain. In *IEEE International Conference on Image Processing*, pages 691–694, 1999.

boat/ship	Albright	Clinton	train	beach	basket	airplane take off	people walk/run	violence	road
28	29	30	31	32	33	34	35	36	37

Table 1: ID and name of Video-TREC features

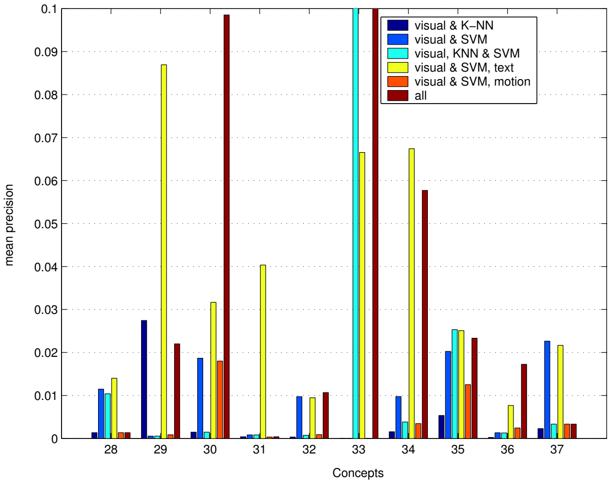

Figure 1: Fusion with a genetic algorithm. The classification outputs of the different modalities are fused using a genetic algorithm. The genetic algorithm estimates the best combination of basic operators: sum, product, minimum and maximum.

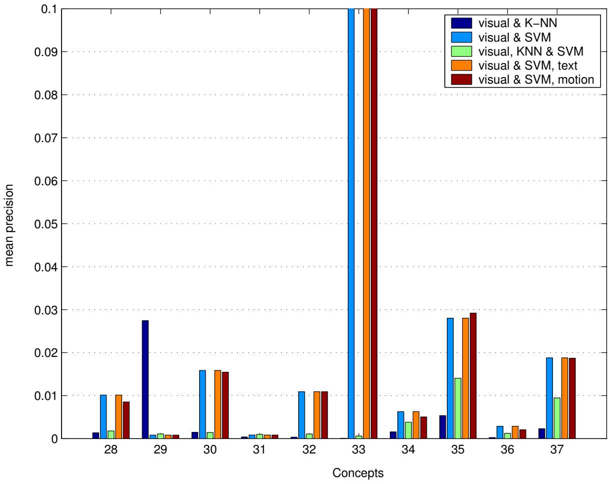

Figure 2: Manual fusion. A fusion formula is empirically selected to fuse classification outputs.