Eurécom at TRECVid 2004: Feature Extraction Task

Fabrice Souvannavong, Bernard Merialdo and Benoit 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 decided 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. To efficiently use this feature, we propose a solution to the problem of synchronization of visual and text events. The motion is also necessary to analyze specific features such as *airplane takeoff* and it will be used through two 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. This final step is very important to efficiently take into account the necessary information from all features 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. TRECVid [12] 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 TRECVid, 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. It is followed by a presentation of results. Finally we conclude with a brief summary and future work.

2 Shot features

We distinguish three types of features: visual, text and motion features that are presented in next sections.

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 [1, 4], visual features are extracted on regions of segmented key-frames [2]. 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 [9]. Finally we combine the signature of the key-frame with the signatures of two extra frames in the shot, as it is described in [10], 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 [3].

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 [13] 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 bin 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 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 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 [10].

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 [5]. 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, does 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

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. An extension of this structure using dynamic binary trees is presented in [11]. The selected structure for TRECVid is fixed as it is depicted in the figure 1.

The fusion is realized as follows:

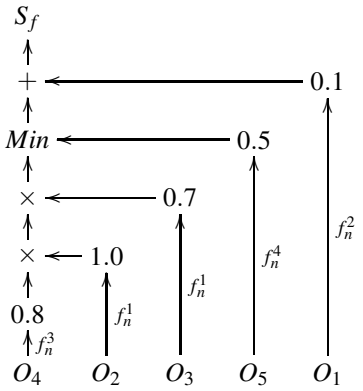

Figure 1: Structure of the fusion function.

- $\{O_i\}$ is the set of classifier outputs,
- Each of them is normalized thanks to a function from the set $\{f_i\}$ that includes min-max and two Gaussian-like normalization functions,
- Normalized scores are then weighted by an a priori probability,
- Finally, scores are merged using simple operators on two operands.

Genetic algorithms are used to find the best parameters for the fusion that are: the order of inputs in the tree structure, normalization functions, a priori weights and operators. The criterion to select best parameters is the one used for the evaluation, i.e. mean precision value at 2,000. It is computed on a subset of the initial training set.

5 Experiments

The classification and the fusion task require annotated data. In June 2003, TRECVid has launched a collaborative effort to annotate video sequences in order to build a labeled reference database. It is composed of about 63 hours of news videos that are segmented into shots. These shots were annotated with items in a list of 133 labels which root concepts are the event taking place, the context of the scene and objects involved. The tool described

in [6] was used for this time-consuming task. We use this huge annotated database to train classifiers. The training dataset was split into two subsets. The first one is used to train classifiers while the second one is used to validate classifiers and train the fusion system.

Figures 2 and 3 show the evaluation results 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 all features might not be relevant and the current structure and algorithm do not allow to discard them. Presented experiments reveal the importance of the text as it was already underlined in last year experiments from other groups.

General performances fluctuate around the median performances of worldwide submitted systems to TRECVid (figure 4). An exception is the *basket scored* feature (numbered 33) where performances are reaching a mean precision of 0.4 (plots were truncated to allow a better reading of other concepts performances) while the best mean precision is 0.55. Yet, this particular feature was trained using all shots containing the scene feature *basket* in the development set of the year 2003.

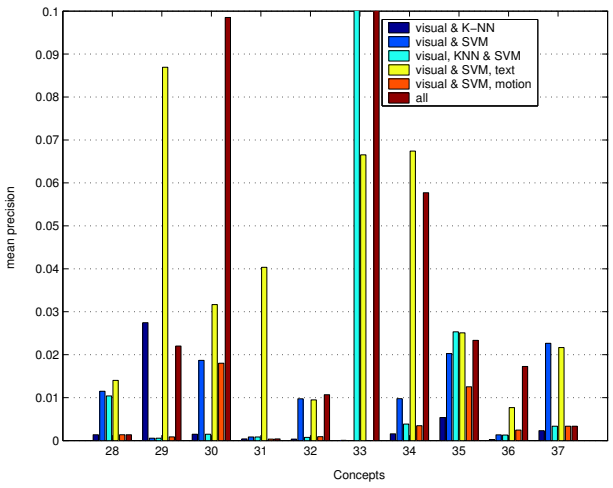

Figure 2: Fusion with a genetic algorithm. The classification outputs of the different modalities are fused using a genetic algorithm. The genetic algorithm estimates the best combination of basic operators: sum, product, minimum and maximum.

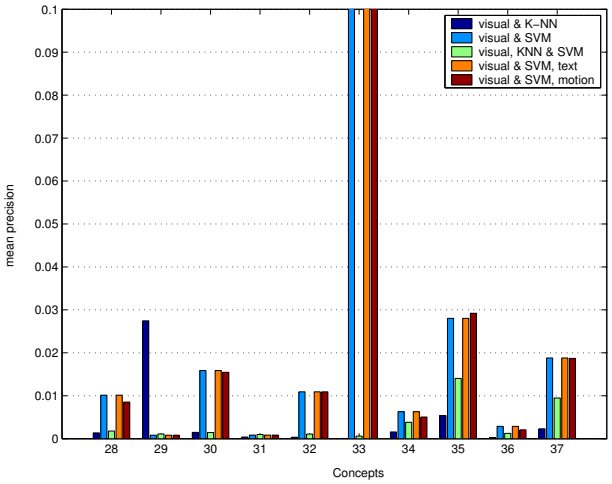

Figure 3: Manual fusion. A fusion formula is empirically selected to fuse classification outputs.

6 Future Work

In our previous participations to TRECVid, only visual cues were used 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. This year we introduced a text feature that is a major self semantic containing feature of the shot. We further addressed the problem of synchronization between text and visual events. Experiments confirmed the importance of the text feature for content-based retrieval. Two motion feature were also used: camera motion and shot activity. However they did not really improved system performances. Then SVM were added to the bench of classifiers and they are achieving good performances on TRECVid datasets compared to K-NN classifiers.

Future works will mainly concern the fusion mechanism. In particular a dynamic hierarchical structure should better model fusion possibilities.

References

[1] Chad Carson, Megan Thomas, and Serge Belongie. Blobworld: A system for region-based image index-

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 TRECVid features

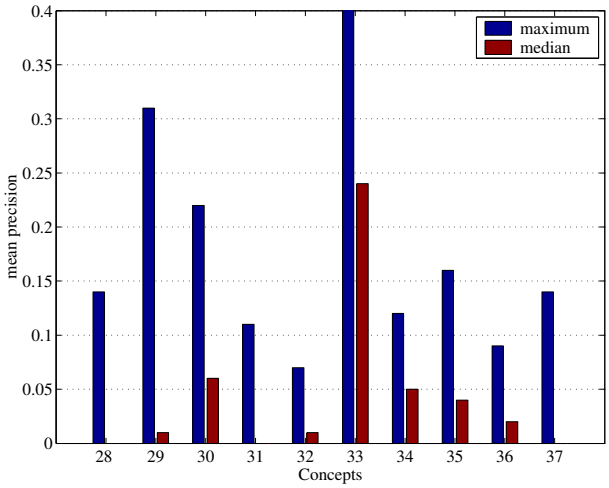

Figure 4: Median versus best performances of all runs submitted to TRECVid

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

[4] 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.

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

[6] Ching-Yung Lin, Belle L. Tseng, and John R. Smith. Video collaborative annotation forum: Establishing ground-truth labels on large multimedia datasets. In *Proceedings of the TRECVID 2003 Workshop*, 2003.

[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 Benoit Huet. Video content modeling with latent semantic analysis. In *Third International Workshop on Content-Based Multimedia Indexing*, 2003.

[10] 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.

[11] Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet. Multi-modal classifier fusion for video shot content retrieval. In *Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services*, 2005.

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

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