Semantic Feature Extraction using Mpeg Macro-block Classification

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

In this paper, we present some first results in the extraction of semantic features from video sequences. Our approach is based on the classification of Mpeg DCT macro-blocks. Although it is clear that using macro-blocks imposes severe restrictions on the analysis accuracy of the image, it has the advantage of avoiding the complete decoding of the Mpeg stream. Our objective is to evaluate the quality of the Semantic Feature Extraction that can be obtained with this direct approach, to serve as a comparative baseline with more elaborate approaches.

Keywords: *Semantic classification, Discrete Cosine Transform, Gaussian Mixture Models, Compressed Domain.*

1 Introduction

The large amount of visual information, carried by video documents as well as still images, requires efficient and effective indexing and search tools [2, 6]. The U.S. Institute of Standards and Technology sponsors the serie of TREC [1] 2002 conferences to promote progress in content-based retrieval from digital video. Our work takes place in this context where we focus on the feature extraction task; video shots should be classified into the high level semantic concepts *indoor, outdoor, cityscape, landscape, text overlay, face* and *people*.

To extract relevant features, the content should in principle be decoded first. Since this operation is time consuming, especially when a large video database should be processed, feature extraction directly from the compressed domain would be particularly interesting by providing fast and reliable information analysis and selection tools. Lots of work have been conducted to achieve image or video analysis [3], however only few researchers have given solutions to this challenging task with limited decoding of the mpeg stream [10, 4].

In this paper, we propose to extract semantic features from 16 by 16 pixels DCT macro-block classification. We have distinguished two types of features in the TREC set, the **region-level** features like *face* and *text overlay* and the **frame-level** features like *indoor, outdoor, cityscape, landscape* and *people* that require elementary concepts like *building, greenery, sky* and *water* to be detected.

The next section details the supervised classification process via Gaussian Mixture Models [9, 7] of macro-blocks. Then we explain how the final decision is taken by introducing new elementary concepts to describe frame-level semantics. Finally, we will outline future improvements.

[1] TREC is a series of conferences which high-level goal is the investigation of content-based retrieval from digital video.
See http://www-nlpir.nist.gov/projects/t2002v/t2002v.html

2 Macro-block Classification

In the context of supervised classification, three steps are involved: feature extraction and representation, class modelisation and parameter estimation, finally classification with respect to decision rules.

In our approach, features are directly provided by the video stream after parsing since we work only on I-frames, which are encoded somehow like jpeg pictures. These frames are composed of macro-blocks that contain 6 DCT blocks, 4 for Y color component, 1 for U and 1 for V i.e. 4:2:0 video format. We can represent a DCT macro-block by a vector of size 64 corresponding to the zigzag scan of the DCT block coefficients and then make the concatenation of the 6 vectors to obtain the feature vector of the whole region. Since the first DCT coefficients are the most important i.e: to eye sensitivity and noise, the feature space dimension is simply reduced to 60 by truncation. Moreover coefficients are scaled with respect to their importance in order to increase the sensitivity of the classifier to important components and at the same time to slightly improve the initialisation of the training algorithm, which is usually obtained via k-means algorithm as explained in the next subsection.

We assume a mixture model to describe the distribution of macro-blocks for each class, and specifically a multi-dimensional Gaussian distribution. Gaussian models can capture the characteristics of a macro-block, while modeling the variation due to motion or lighting conditions. Moreover in [5], E.Y. Lam and J.W. Goodman have proven that the distribution of macro-block DCT coefficients can be well approximated by a Gaussian *when the variance is constant*; in the classification situation, the latter hypothesis is more or less true and mixtures should compensate it. So the probability density function can be written as follows:

$$\text{For } X \in C_i, P(X \mid \Phi_i) = \sum_j \alpha_j p_j(X)$$

$$\text{where } \alpha_i \in \Re, \Phi_i = (\mu_j, \sigma_j) \text{ and } p_j(X) \sim \mathcal{N}(\mu_j, \sigma_j)$$

The GMM parameters α_j, μ_j and σ_j are estimated using the traditional Expectation-Maximization algorithm [1] which is initialized with a classical k-means algorithm. In our current experiments, we also make the hypothesis that feature vector components are independent, thus σ_i is a diagonal matrix, or that only color components of the same frequency are correlated, thus σ_i is a matrix diagonal by block. Finally the choice of the number of mixtures is simply achieved by looking at the test set loglikelihood evolution of the EM algorithm for various mixture numbers. It should not increase to much in order to avoid data overfitting.

Given an unlabeled macro-block X, the maximum a posteriori rule:

$$\hat{C} = arg \max_i P(\Phi_i \mid X)$$

gives an estimation of the class it belongs to. The posterior probabilities can be expanded by Baye's rule:

$$P(\Phi_i \mid X) = \frac{P(X \mid \Phi_i)P(\Phi_i)}{P(X)}$$

finally,

$$P(\Phi_i \mid X) \propto P(X \mid \Phi_i)$$

since we assume the *equiprobability* of classes and vectors.

However, it is possible that a macro-block does not belong to any predefined class. Thus we introduce for each model i a minimum bound $-mb_i$ for the loglikelihood which is selected to eliminate 10% of the training data set. Of course there is a trade-off to find between precision and recall, see figure 1. Finally the decision rule can be written:

$$\hat{C} = arg \max_i \{P(X \mid \Phi_i) \mid -\log(P(X \mid \Phi_i)) \leq mb_i\}$$

3 Feature detection

The presented classification method allows to detect **region-level** features only. In our previous work [8] we have underlined that macro-blocks could not carry **frame-level** semantic information but succeed well in providing a lower level semantic. Thus a heuristic two-step hierarchy, depicted in figure 2, was introduced to detect **frame-level** concepts via additional elementary semantics. The hierarchy contains three kinds of elements:

- Elementary concepts at the leaves of the hierarchy that are perceivable from macro-blocks,

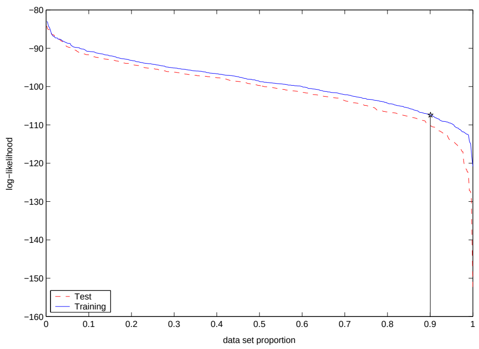


Figure 1: Threshold selection

- Higher-level semantics on the upper part of the graph that are difficult to extract directly, but can be induced by a combination of lower-level features.

- Trec concepts, enclosed in boxes, that are spread in either of the previously stated categories.

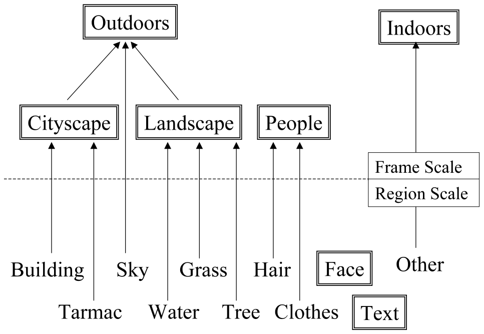


Figure 2: Concepts hierarchy.

The detection of features present in one shot is finally achieved with respect to the following procedure:

1. Classify all macro-blocks of the shot into elementary concepts with respect to preliminary trained Gaussian Mixtures,

2. Compute a detection score for each feature.

The detection score of the feature i whose elementary childrens are J is simply defined by:

$$Ds_i = \sum_j P(j) \text{ where } j \in J$$

$$P(j) = \frac{\text{Number of macro-blocks with label j}}{\text{Total number of macro-blocks in the shot}}$$

It represents the posterior probability of a feature to be in the given shot. Finally, for each feature, shots are ordered by decreasing detection score.

4 Experiments

Nine video sequences were randomly selected in the development set in order to create training and test samples. Some Macro-blocs of these sequences were labeled with **region-level** concepts; half to perform the training of semantic classes and half to evaluate models. The fastidious annotation task was achieved over 232 frames and table 1 gives a summary of the accomplished task.

We have finally modeled classes by fifteen gaussian mixtures and truncated the space dimension to six by fifteen features. This values reveal to be a good compromise between performance and complexity. For the same reasons, we have approximated the co-variance matrix to a diagonal and not diagonal by block matrix, see table 2 that emphasizes the small improvement acquired by using a diagonal by block co-variance matrix.

Finally figures 3 and 4 show the performance of our method thanks to the assesor's judgement provided by TREC. To evaluate the feature extraction task, we have represented the classical precision and recall curves for the four Trec features *outdoors*, *cityscape*, *text and face*. Encouraging results were obtained for *outdoors* and *cityscape* features, however we were expecting better results for *text* and *face* features since they are relevant at the macro-block level and the development analysis was forecasting good classification capacities, see table 2. Several explanations can be envisaged: heterogenous sizes of training sets leading to overtrained and undertrained models and too few training variety conducting to restricted models (for example, no cartoon sequences were used to train models). The results we obtained using a *single framework* for all visual features, are closely comparable to submitted runs of other labs. In particular we get

Selected sequences	00616	01859	06085a	08131b	08261	08325	16683	19567b	35435b
Nb of selected frames	46	18	14	38	42	26	21	17	10
Nb of selected blocks	4647	1464	2133	2745	6200	3549	1522	1879	1401

Features	text	skin	clothes	sky	tree	building	grass	tarmac	hair	water	ground	total
Number	1472	4097	7158	4636	1839	2550	558	639	857	297	1437	25540

Table 1: Summary of the manual annotation.

Features	text	skin	clothes	sky	tree	building	grass	tarmac	hair	water	ground
Diagonal by block											
precision	38	26	23	49	22	12	34	17	10	58	40
recall	75	56	45	88	69	45	66	84	44	68	81
Diagonal											
precision	33	30	22	41	18	12	26	12	6	19	30
recall	75	53	42	83	64	36	67	77	43	80	78

Table 2: Precision and recall during the development of the low level features.

surprisingly good ranking in text ovelay detection.

5 Conclusion

We have presented a method based on DCT information of macro-blocks to detect Trec visual features from video shots in a *single framework*. Since macro-blocks carry only local information, a heuristic hierarchy was introduced to build the final decision rule at the **frame-level** and **region-level**. In gereral this evaluation is encouraging knowing the small extract of the development set used. In future works we plan to investigate methods to automatically elaborate the hierarchy. This will set up a complete probabilistic framework to detect features from low level observations and a more realistic manual annotation at the shot level will be required to train models.

References

[1] J. A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. ICSI-TR, 1997.

[2] S.-F. Chang, W. Chen, H. Meng, H. Sundaram, and D. Zhong. A fully automated content-based video search engine supporting spatiotemporal queries. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 8, pages 602– 615, 1998.

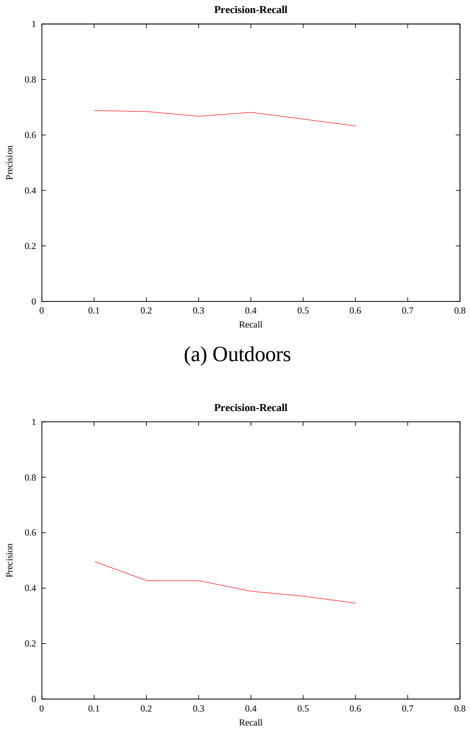
[3] S.-F. Chang and H. Sundaram. Structural and semantic analysis of video. In *ICME*, 2000.

[4] A. Girgensohn and J. Foote. Video classification using transform coefficients. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3045–3048, 1999.

[5] E. Y. Lam and J. W. Goodman. A mathematical analysis of the dct coefficient distribution for images. In *IEEE Transactions on Image Processing*, volume 9, pages 1661–1666, October 2000.

[6] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. In *SPIE Storage and Retrieval for Image and Video Databases*, february 1994.

[7] M. Saeed, W. Karl, T. Nguyen, and H. Rabiee. A new multiresolution algorithm for image segmentation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 2753–2756, 1998.



(a) Outdoors

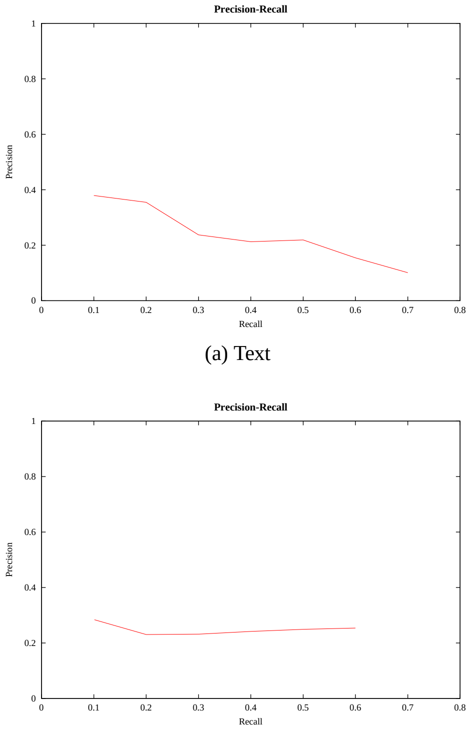
(b) Cityscape

Figure 3: Classification Evaluation

[8] F. Souvannavong, B. Merialdo, and B. Huet. Classification semantique des macro-blocs mpeg dans le domaine compresse. In *Compression et Representation des Signaux Audiovisuels*, pages 235–238, 2003.

[9] J. Verbeek, N. Vlassis, and B. Kr. Greedy gaussian mixture learning for texture segmentation. In *Workshop on Kernel and Subspace Methods for Computer Vision*, 2001.

[10] H. Wang and S.-F. Chang. A highly efficient system for automatic face region detection in mpeg video. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 7, pages 615–628, August 1997.



(a) Text

(b) Face

Figure 4: Classification Evaluation