# CLIPS and NII at TRECvid: Shot segmentation and feature extraction

Stéphane Ayache, Georges Quénot, Jérôme Gensel, Shin'Ichi Satoh

# CLIPS-LSR-NII Experiments at TRECVID 2005

*Stéphane Ayache*[1], *Georges M. Quénot*[1], *Jérôme Gensel*[2] *and Shin'Ichi Satoh*[3]

[1] CLIPS-IMAG, BP53, 38041 Grenoble Cedex 9, France
[2] LSR-IMAG, BP53, 38041 Grenoble Cedex 9, France
[3] NII, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Stephane.Ayache, Georges.Quenot@imag.fr

## Abstract

This paper presents the systems used by CLIPS-IMAG laboratory. We participated to shot segmentation and high-level extraction tasks. We focus this year on High-Level Features Extraction task, based on key frames classification. We propose an original and promising framework for incorporating contextual information (from image content) into the concept detection process. The proposed method combines local and global classifiers with stacking, using SVM. We handle topologic and semantic contexts in concept detection performance and proposed solutions to handle the large amount of dimensions involved in classified data.

## 1 Introduction

The CLIPS-IMAG team have participated to the Shot Segmentation task with few modifications from previous participations. The emphasis have been laid on the High-Level Feature Extraction Task, with the use on image cue only.

We previously proposed and evaluated a framework for incorporating contextual information into image indexing process [16]. We assume that both semantic and local contexts can increase the accuracy of a classifier. We present a general framework for contextual image indexing using vector-based classifier. The main difficulty is to fuse a lot of information while managing the curse of dimensionality problem. We want to incorporate both local and inter-concept information in the decision process, which can need a lot of dimensions. In addition, supervised classifiers need a number of samples, which is correlated to the dimensionality problem.

## 2 Context

New issues in Content-Based Image Indexing (CBII) field are arising for the reduction of the well-known semantic gap. In order to improve concept extraction, many approaches take into account the context. To do so, some approaches fuse local and global descriptors [1, 2] combined using boosted classifiers [3].

Other define context as spatial relationships between objects within an image [4, 5, 6] using probabilistic frameworks. However, in order to deal with a large amount of local descriptors and simplify computation, such approaches first detect points of interest and then assume them independent.

In their work, [7, 8, 9] handle semantic relationships [10] between concepts using Stacked Classifiers [11, 12]. They first classify intermediate concepts, and learn their relationships in the context of a higher level concept by a second-level classifier. In TRECVID'04 experiments, [9] used a basis of 22 intermediate concepts. By adding the 10 TRECVID's concepts, they learned semantic context of 32 concepts with Stacking.

## 3 A Three-Level framework

We use several classifiers and arrange them into networks, using stacking. The idea is that the correlation between the input (low-level features) and the output (concepts) is too weak to be efficiently recovered by a single "flat" classifier even if the low-level features have been carefully chosen. Combining classifiers in a multi-level framework allows extraction of intermediate-level data from low-level features and other classifiers. Detection of concepts from these intermediate-level data is expected to improve the overall performance (both correctness and computation time) of the system.

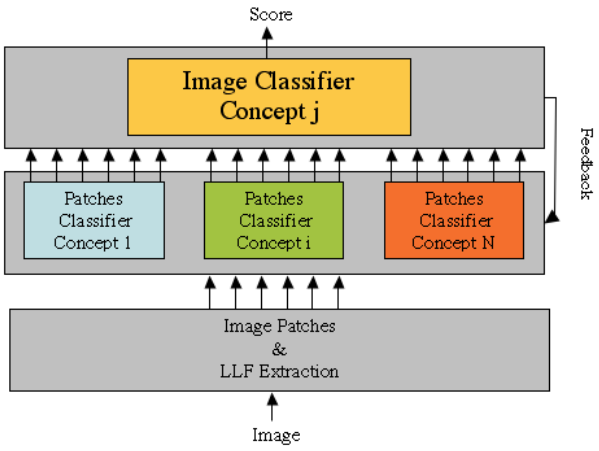


Figure 1: Overall classifier architecture

Classifiers from each level bridge a small part of the semantic gap and are expected to do it well because the correlation between their inputs and their outputs is expected to be better than the correlation between the inputs and the outputs of a single classifier that would bridge alone the whole semantic gap. Also, not only the levels are cascaded but many intermediate data are computed in parallel and the outputs of all of them are combined as inputs to the next level. Also, such architecture allows either inference if "Concept j" is not classified as patches in the second layer, either whole image labeling if "Concept j" is already classify at patch-level.

The objective of the present work is to validate these assumptions and to quantify the benefits that can be obtained from this approach. We propose a multi-layer framework in order to integrate the different contextual information. The first layer extracts Low-Level Features for each patch, the second layer provides a score for patches according to a concept, and the higher level layer assigns a value for a whole image. The first layer is basically the only one that handles low-level features.

We focus on three kinds of concepts:

- Patches-Concepts can be quite easily extracted from a piece of an image (a patch). We consider them as middle-level concepts as they are useful to help extraction of higher-level concepts. Examples of such concepts are Sky, Greenery, Water, Wall, Sand, Building,
- Scenes-Concepts are higher-level concepts as they refer to the whole image. Thus, their extraction necessitates a deeper analysis of image's context. Examples are Beach, City, Studio settings, Mountain ... Such concepts cannot be extracted as patches.
- Objects-Concepts are specific concepts. They are usually detected using specific object-based detectors, which are much highly time-consuming. In spite of it, we expect to extract them using our contextual framework.

## 4 Low-level features extraction

As we want to handle topologic context, we need to compute low-level features for parts of the image. In order to compute local features, many approaches have been proposed. While automatic and a priori segmented regions are too far from semantic meaning of image, we decided to split images into patches. By doing so, we should be very far from semantic, but with such granularity one patch is more likely to contain a single concept.

The low-level features extractor's process first splits image into overlapped patches. Basically, we used 260 overlapped patches of 32 $\times$ 32 pixels (in 352 $\times$ 240 images). And then, extracts 9

moments color (3 means + 6 co-variances), Gabor wavelets for texture (3 scales × 8 orientations), and coordinates of patches.

These choices have been made for a baseline system. The main goal here is to explore the use of context for concept indexing. We want to study and evaluate various ways of doing it by combining classifiers into networks. In further work, we plan to enrich and optimize the set and characteristics of low level features, especially for video content indexing. Currently, we expect to obtain representative results from the current set of low-level features.

# 5 Use of the topologic and semantic contexts

The idea behind the use of topologic context is that the confidence (or score) for the whole image could be computed more accurately by taking into account the confidences obtained for each patch in the image for the same concept. By doing so, the classifier learn topological distribution of a concept into an image.

The idea behind the use of semantic context is that the confidence (or score) for a single concept could be computed more accurately by taking into account the confidences obtained for other concepts in the same image. We are considering concepts as related one to each other.

Using both contexts, we expect to improve the performance of concept detection at the image level by combining the output of patches-level detectors (classifiers).

We have $N$ level-1 classifiers, each with $F$ inputs and 1 output and $N$ level-2 classifiers, each with $N \times P$ inputs and 1 output. All level-1 classifiers are callled $P$ times on a given image and their $N \times P$ output values are passed to the corresponding level-2 classifier which is called only once.

First we classify patches for each concept, and then classify the whole image by merging every output. With many concepts, the number of inputs for the second stage can be too large. We can apply different approach to reduce this number, using well known feature selection techniques such as PCA. In TRECVID05 experiments, we manually choose a set of 5 semantic context concepts per concept.

## 5.1 Baseline, no context, one level

In order to evaluate the patch level alone, and to compare the use of contexts, we define an image score based on the patch confidence values. To do so, we simply compute the average of all of the patch confidence scores. This baseline is very basic it does not take into account any spatial or semantic context. We have here $N$ classifiers, each with $F$ inputs and 1 output. Each of them is callled $P$ times on a given image and the $P$ output values are averaged.

# 6 Experiments

We had many hardware problems for the official submission, so we recompute them in the same conditions. We now present these results.

According to our typology of concepts, 8 out of 10 TRECVID'05 concepts are patches-concepts: `People-Marching`, `Explosion_Fire`, `Maps`, `Flag-US`, `Building`, `Prisoner`, `Sports` and `Car`. We manually modify official annotations in order to have annotated regions. For this task, we had considered `Prisoner` as `Prisoner clothes` and `Sports` as `Sports grounds` (tennis, basketball, soccer, ...). We also enriched these intermediate concepts with `Sky`, `Greenery`, `Skin_Face`, `Water` and `Road`. The two others concepts, which are not patches-concepts (Waterscape and Mountain) were "inferred" using others concepts, such as `Water` and `Greenery`.

In order to handle the curse of dimensionality problem, we need to limit the amount of intermediate concepts to merge. For these experiments, we manually did that. Each concept's context is defined by at most 5 intermediate concepts (patches-concept).

We used SVM classifier [15] with RBF Kernel, because it has shown good classification results in many fields, especially in CBIR [14]. It is important to use cross validation for parameter selec-

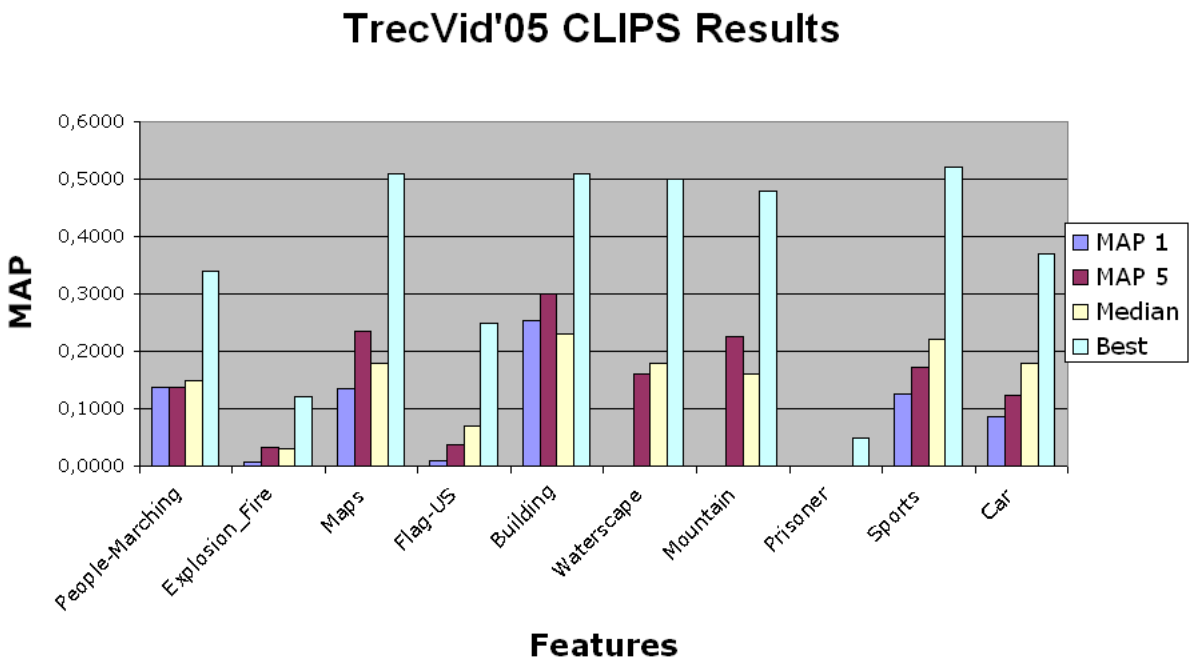


Figure 2: TrecVideo 2005 Results

tion. We use grid search tool to select the best combination of parameters C and gamma (out of 110).

In order to obtain the training set, we extract patches from annotated regions, it is easy to get many patches by performing overlapped patches. Annotating whole images is harder as annotators must observe each one.

We collected many positive samples for patches classification, and defined experimentally a threshold for maximum numbers of positive samples. We found that 2048 positive samples is a good compromise to obtain good accuracy with smaller training time. Also, we found that twice negative samples is a good compromise. Finally, we randomly choose negative samples. The Table 2 shows the number of positive image examples, for each concept.

Figure 2 shows Mean Average Precision results for our baseline and the proposed approach. We compare them with Median and Best scores of TrecVid'05 evaluation. It is interesting to notice that Median and Best results include approaches, which also used ASR data.

The use of contexts (MAP5) improves the performance over the baseline (MAP1).

## 7 Conclusion

We present a framework for incorporate semantic and topologic contexts into CBII. We compare our approach with a baseline system which don't handle context information. We show that both contextual information improves concepts detection. We now plan to combine such information with other video cues such as ASR and motion.

## 8 Acknowledgments

This work has been spported by the ISERE CNRS ASIA-STIC project and the Video Indexing INPG BQR project.

## References


[1] K. Murphy, A. Torralba, D. Eaton and W. Freeman Object detection and localization using local and global features. Sicily Workshop on Object Recognition, 2005. Lecture Notes in Computer Science (submitted)

[2] A Garg, S Agarwal, T.S. Huang Fusion of Global and Local Information for Object Detection. In 16th International Conference on Pattern Recognition (ICPR'02) - Volume 3, 2002.

[3] J. Amores, N. Sebe, P. Radeva, T. Gevers and A. Smeulders Boosting Contextual Information in Content-Based Image Retrieval. In MIR, 2004.

[4] A. Singhal, J. Luo, and W. Zhu Probabilistic spatial context models for scene content understanding. In CVPR, 2003.

[5] A. Torralba, K. Murphy, and W. Freeman Contextual models for object detection using boosted random felds. In Advances in Neural Info. Proc. Systems, 2004.

[6] P. Carbonetto, N. d. Freitas, and K. Barnard. A statistical model for general contextual object recognition. In ECCV, 2004.

[7] G. Iyengar and H.J. Nock Discriminative model fusion for semantic concept detection and annotation in video, MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia, 2003.

[8] M. Naphade. On supervision and statistical learning for semantic multimedia analysis. Journal of Visual Communication and Image Representation, 15(3):348369, 2004.

[9] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma and F.J. Seinstra The MediaMill TRECVID 2004 Semantic Video Search Engine. InTRECVID Workshop, 2004.

[10] M. Fink P. Perona Mutual boosting for contextual influence. In Advances in Neural Info. Proc. Systems, 2003.

[11] D.H. Wolpert Stacked Generalization. Neural Networks, Vol. 5, pp. 241-259, Pergamon Press.

[12] D. A. Lisin, M. A. Mattar, M B. BlMark C. Benfield and E. G. Learned-Miller Combining Local and Global Image Features for Object Class Recognition In CVPR, 2005.

[13] P. Howarth and S. Rueger. Evaluation of Texture Features for Content-Based Image Retrieval. In P. Enser et al. (Eds.): CIVR 2004, LNCS 3115, pp.326-334, 2004.

[14] P.H. Gosselin and M. Cord. A Comparison of Active Classification Methods for Content-Based Image Retrieval. Int. Workshop on Computer Vision Meets Database, CVDB, 2004.

[15] C. Chang and C Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

[16] S. Ayache, G. Quénot and S. Satoh. Context-based Cconceptual Image Indexing, In ICASSP, 2006.