IRIT @ TRECVid 2010: Hidden Markov Models for Context-aware Late Fusion of Multiple Audio Classifiers

Hervé Bredin*, Lionel Koenig† & Jérôme Farinas†
* LIMSI-CNRS, BP 133, F-91403 Orsay Cedex, France
† University of Toulouse, IRIT, 118 Route de Narbonne, F-31062 Toulouse, France

Abstract

This notebook paper describes the four runs submitted by IRIT at TRECVid 2010 Semantic Indexing task. The four submitted runs can be described and compared as follows:

- Run 4 – late fusion (weighted sum) of multiple audio-only classifiers output
- Run 3 – context-aware re-rank of run 4 using hidden Markov model
- Run 2 – context-aware late fusion of multiple audio classifiers output with hidden Markov model
- Run 1 – late fusion (weighted sum) of multiple audio & video classifiers output

1 Introduction

In this notebook paper, we describe the systems submitted by IRIT to the semantic indexing task as defined by NIST for the TRECVid evaluation campaign: detecting the presence of visual concepts in video shots [5]. Most systems rely on the late fusion of multiple binary classifiers based on numerous visual descriptors extracted from video keyframes. Given a representative keyframe of a video shot, visual features are extracted and provided as input of various classifiers which, in turn, return a score analogous to a probability that the shot contains the considered visual concept. Those scores are then combined into one single score, meant to be more robust than each combined scores taken individually [1].

Those approaches simply consider a video as a set of unrelated shots. Thus, knowing that a concept was detected in one shot of the video tells us nothing about the presence of the very same concept in another shot of the same video. *Yang & Hauptmann* showed that this assumption is wrong for most videos [6]: they found that the probability for shot k to contain a concept (i.e. $q_k = 1$ in Figure 1) is higher if the concept is present in the previous shot (i.e. if $q_{k-1} = 1$). Denoting o_k the baseline score for shot k, they managed to slightly improve the performance of a baseline system by temporally smoothing o_k, based on the scores o_{k-1} and o_{k+1} of its neighboring shots.

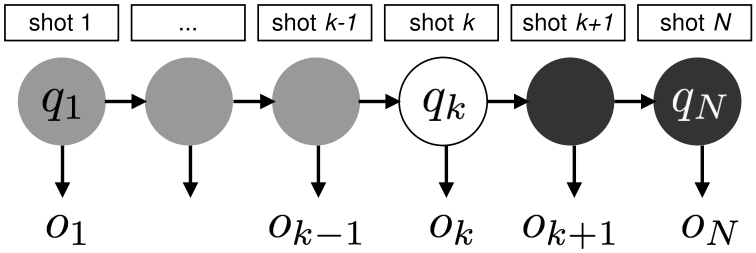

Figure 1: For each shot k, o_k is the score output by the classifier and q_k is a binary variable indicating whether shot k contains a given concept ($q_k = 1$) or not ($q_k = 0$). N is the number of shots in the video.

We propose to achieve both late fusion of multiple classifiers **and** contextual smoothing in one single step, using hidden Markov models (HMMs).

Section 2 describes our baseline run (run 4) that makes use of audio clues only, and gives a short description of its audio/visual extension (run 1). Then, in Section 3, we describe runs 3 and 2 that make use of hidden Markov models to introduce temporal context awareness to the baseline run.

2 Visual concept detection using audio

In this section, we describe our baseline approach to video semantic indexing based on the late fusion of multiple audio classifiers. To our knowledge, it is one of the very few attempts to design a visual concept detectors using audio clues only.

2.1 Audio features

A collection of audio features is extracted every 10ms with a 20ms window using YAAFE [3]. It includes Mel Frequency Cepstral Coefficients (MFCC, 13 coefficients), MFCC first derivatives (13 coef.), MFCC second derivatives (13 coef.), loudness (24 coef.), spectral crest factor per log-spaced band of 1/4 octave (spcrestb, 23 coef.) and spectral flatness per log-spaced band of 1/4 octave (spflatb, 23 coef.).

In the rest of the paper, variables marked with exponent d indicate they were obtained in the d^{th} feature (**d**escriptor) space (D being the total number of feature spaces).

2.2 GMM/UBM classifiers

For a given feature space d, a large set of feature vectors is used to train a gaussian mixture model (GMM) which models their global distribution. This model is called universal background model (UBM, denoted Ω^d) as it is representative of any vector of the selected feature space. Given a visual concept to be detected, feature vectors extracted from positive shots (i.e. shots actually containing the visual concept) are used to perform a MAP adaptation of the UBM model: we denote ω_+^d the resulting GMM.

The score resulting from the application of the GMM/UBM classifier on the k^{th} shot of a test video is computed as follows:

$$o_k^d = \frac{1}{\#\mathcal{S}_k^d} \sum_{\mathbf{x} \in \mathcal{S}_k^d} \frac{\Pr\left(\mathbf{x}|\omega_+^d\right)}{\Pr\left(\mathbf{x}|\Omega^d\right)} \tag{1}$$

where $\mathcal{S}_k^d$ is the set of features extracted from shot k, $\#\mathcal{S}_k^d$ its cardinal. In practice, we used the GMM/UBM implementation of Mistral/Alizé toolkit [2] to obtain Ω^d, ω_+^d and o_k^d.

2.3 Run 4: Weighted sum late fusion

Our baseline system (run 4) is based on the weighted sum fusion of scores provided by GMM/UBM classifiers in the D feature spaces – as such, it will be denoted $\mathbf{\Sigma}$ in the rest of this article.

$$\mathbf{\Sigma}(o_k) = \sum_{d=1}^{D} w_d o_k^d \tag{2}$$

where optimal weights $\{w_d\}$ are tuned on the development set.

Run 1 is similar to run 4, except that 3 additional systems (based on visual descriptors) are added to the pool of scores:

- Support Vector Machine (SVM) applied on HSV color histograms;
- SVM on 250-dimensional aggregated bag-of-SIFT representation;
- SVM on Discrete Cosine Transform (DCT) of gray scale frames.

3 Context-aware classification using hidden Markov models

As highlighted in the introduction, context should not be ignored when trying to decide whether a shot contains a given concept. Chances that it does are much higher if the concept is also appearing in other shots of the same video.

However, up to here, the baseline system is completely unaware of the context: the score it provides is based on the sole observation of the considered shot. We propose to introduce context awareness using hidden Markov models.

3.1 Modelling context with hidden Markov models

The evolution of the presence of a concept in the shots of a video is modelled by a two-states (numbered 0 and 1 in Figure 2) hidden Markov model (HMM) [4]. As described in Figure 1, the state of shot k is denoted q_k. $q_k = 1$ indicates that the concept is present in shot k, while $q_k = 0$ indicates the contrary.

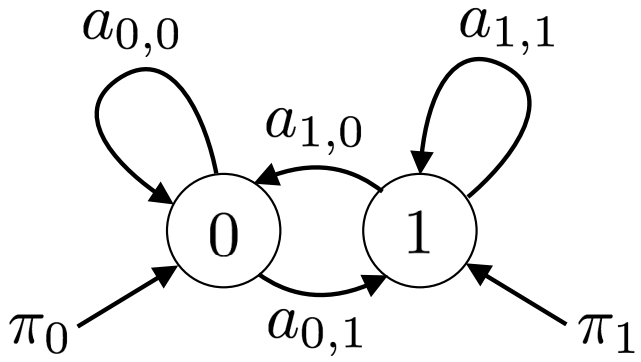

Figure 2: Topology of hidden Markov models

An observation probability density function (pdf) b_i is associated to each state i. It does not depend on the shot:

$$\forall k, \qquad \Pr\left(\mathcal{O}|q_k = i\right) = b_i(\mathcal{O})$$

where $\mathcal{O}$ lies in the observation space. Transitions between states follow the Markov property. The conditional probability of the future state only depends on the present state:

$$\begin{aligned}\Pr\left(q_{k+1} = j|q_k = i, q_{k-1}, \cdots\right) &= \Pr\left(q_{k+1} = j|q_k = i\right) \\ &= a_{i,j}\end{aligned} \tag{3}$$

Based on the knowledge of $a_{i,j}$ and b_i, it is possible to derive the following probability:

$$\gamma\left(\mathcal{O}_k\right) = \Pr\left(q_k = 1|\mathcal{O}_1, \ldots, \mathcal{O}_k, \ldots, \mathcal{O}_N\right) \tag{4}$$

The whole video context is available for a classifier based on γ.

Note that, up to this point, no restriction was defined on the choice of the nature of the observation $\mathcal{O}$. Sections 3.2 and 3.3 introduces two possibilities.

3.2 Run 3: HMM for contextual smoothing / $\mathcal{O}_k \sim \mathbf{\Sigma}\left(o_k\right)$

Using the output of the baseline system $\mathbf{\Sigma}$ (run 4) as the observation, we define a novel classifiers, denoted $\gamma \circ \mathbf{\Sigma}$:

$$\gamma \circ \mathbf{\Sigma}\left(o_k\right) = \Pr\left(q_k = 1|\mathbf{\Sigma}\left(o_1\right), \ldots, \mathbf{\Sigma}\left(o_k\right), \ldots, \mathbf{\Sigma}\left(o_N\right)\right) \tag{5}$$

This system can be seen as a way of smoothing the output of a reference system using the video context and is the one used for run 3.

3.3 Run 2: HMM for late fusion / $\mathcal{O}_k \sim o_k = \left[o_k^1, \ldots, o_k^D\right]$

Nowhere is it said that $\mathcal{O}$ has to be mono-dimensional. Therefore, we propose to use the HMM framework as a context-aware late fusion tool by using the D-dimensional vector $o_k = \left[o_k^1, \ldots, o_k^D\right]$ as the observation – o_k^d being the score for shot k provided by the classifier in the d^{th} feature space. Consequently, we can derive a novel classifier, denoted $\mathbf{\Gamma}$:

$$\mathbf{\Gamma}\left(o_k\right) = \Pr\left(q_k = 1 | o_1, \ldots, o_k, \ldots, o_N\right) \tag{6}$$

This system performs – in one single step – late fusion of multiple classifiers **and** contextual smoothing, and is the one used for run 2.

4 Training

Videos from the development set provided by NIST are divided into three subsets $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ – carefully chosen to make sure positive samples for each concept are equally distributed among them. By doing so, we aim at avoiding overfitting in the two-steps training process:

- GMM/UBM classifiers (with 512 gaussians) are trained and tuned using $\mathcal{A}$ as the training set and $\mathcal{B}$ as the development set.
- Late fusion approaches are trained and tuned using $\mathcal{B}$ as the training set and $\mathcal{C}$ as the development set.

1. Weights $\{w_d\}$ are chosen to maximize the infered average precision.

2. Observation pdfs are assumed to be gaussian

$$b_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right) \tag{7}$$

3. Transition probabilities are estimated as

$$a_{i,j} = \frac{\#\left\{q_{k-1} = i, q_k = j\right\}}{\#\left\{q_{k-1} = i\right\}} \tag{8}$$

where $\#\left\{q_{k-1} = i\right\}$ is the number of shots in state i, and $\#\left\{q_{k-1} = i, q_k = j\right\}$ the number of times a shot in state j follows a shot in state i.

5 Remark

Our runs 2, 3 and 4 rely entirely and exclusively (partially, for run 1) on the audio stream extracted from the videos. Yet, for approximately one eighth of the test videos, the length of the audio stream did not match the one of the visual stream (or for some videos, there was no audio at all). This is probably due to a bad MPEG-4 encoding. Consequently, we did not provide any score for those videos.

6 Conclusion

In this paper, we introduced novel approaches to perform late fusion of multiple classifiers in the framework of content-based video semantic indexing. They are based on two-states hidden Markov models allowing to bring video context awareness to any existing reference systems.

In the future, we plan on applying this approach to more efficient concept detectors – indeed, audio-only classifiers do not achieve baseline performance as good as current video-based state-of-the-art systems (based on SIFT descriptors, for instance).

References

[1] Pradeep Atrey, M. Answar Hossain, Abdulmotaleb El Saddik, and Mohan Kankanhalli. Multimodal Fusion for Multimedia Analysis: a Survey. *Multimedia Systems*, pages 1–35, 2010. 10.1007/s00530-010-0182-0.

[2] Jean-François Bonastre, Frédéric Wils, and Sylvain Meignier. ALIZE, a Free Toolkit for Speaker Recognition. In *ICASSP'05, IEEE*, Philadelphia, PA (USA), March, 22 2005.

[3] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard. YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, 2010.

[4] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[5] Alan F. Smeaton, Paul Over, and Wessel Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In Ajay Divakaran, editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin, 2009.

[6] Jun Yang and Alexander G. Hauptmann. Exploring Temporal Consistency for Video Analysis and Retrieval. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 33–42, New York, NY, USA, 2006. ACM.