Audio feature extraction for sequential learning in video content analysis

Cyril Hory

Centre for Vision, Speech and Signal Processing
University of Surrey

NETWORK OF EXCELLENCE

UniS

On the recent performance of French football

How did Zidane manage to come back to fitness?????????

On the recent performance of French football

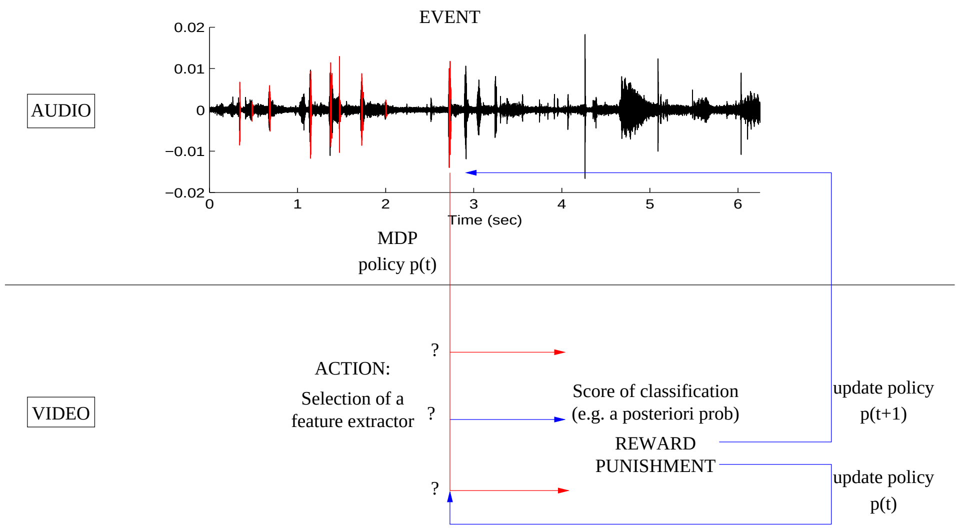
Sequential video analysis

Sequential Learning

- Sequential framework allows for on-line learning (intuitive)
- Context provides rules for sequences of events (coarse to fine)

Audio/Video cooperation

- Audio data easier to handle
- Audio-based event detection / Video-based event isolation
- grounded theory available for 1D signal detection (Optimality of CUSUM test)



Fine to coarse approach

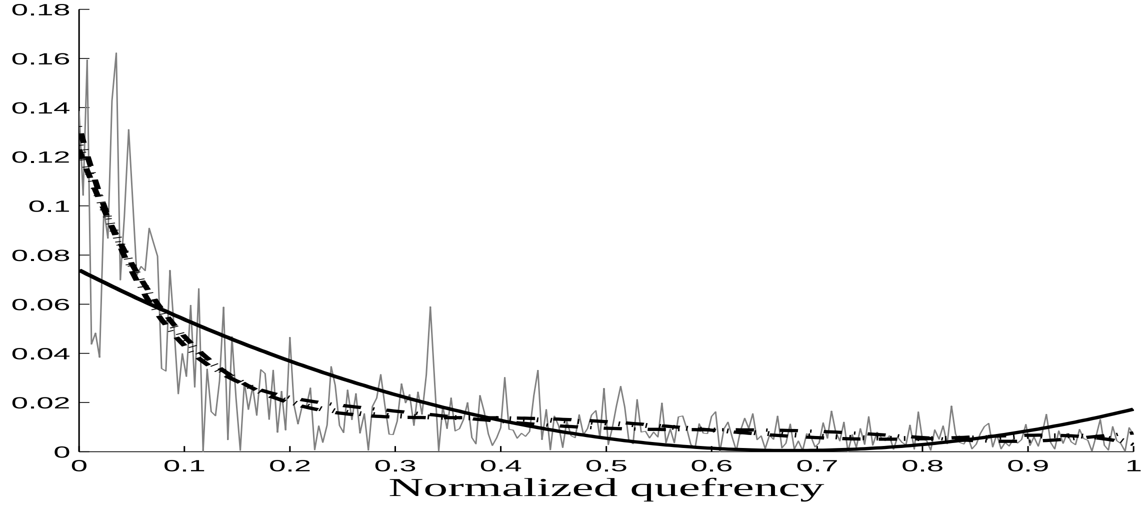
Outline of the talk

- Feature extraction
  - Proposed features
  - Experimental setup: Racket hit detection in tennis video clip
- Assessment of sequential learning
  - Two issues:
    * Is the dimension of the feature space critical (curse of dimensionality)?
    * Is their a need for adaptivity (non-stationarity)?
  - Relevance Feedback framework
  - Average Normalized Modified Retrieval Rank (ANMRR)

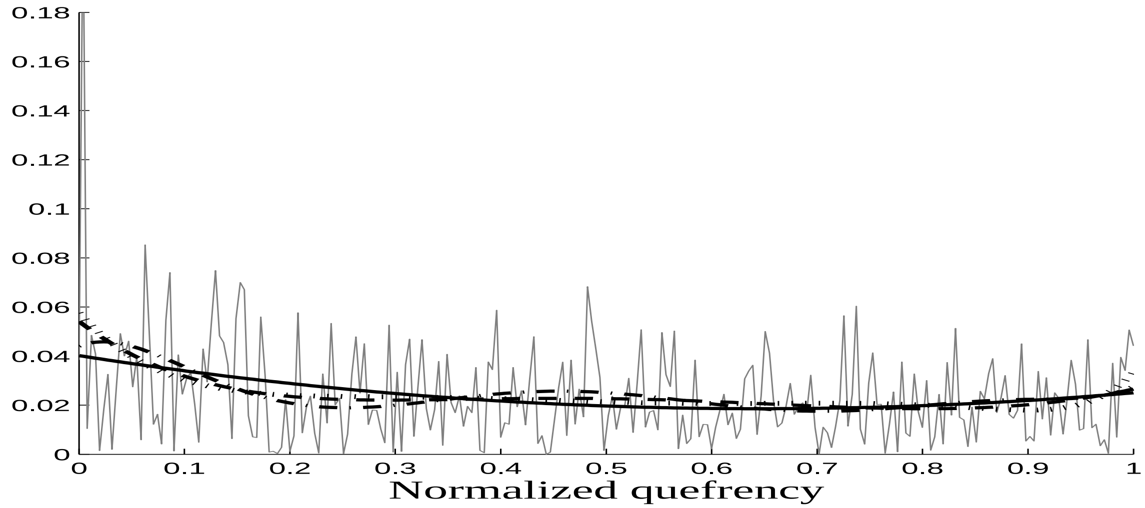
Proposed features

- Spectral Domain
  - Spectral Subband Coefficients
  - First Cepstral Coefficients (Spectral smoothing)
- Spectral smoothing?
  - Relevant for characterizing the resonant system
  - Possible overlapping of the high-quefrency and low-quefrency bands
  - number of coefficients is constrained by the size of the dataset
- Cepstral Domain
  - Cepstral Subband Coefficients
  - Cepstral Regression Coefficients

Cepstral Regression



racket hits



Crowd applauses

Recursive computation of the RC's

1D signal model:

$$c = Q_{(p)} a_{(p)} + \epsilon$$

Estimation of the regression coefficients:

$$\hat{a}_{(p)} = R_{(p)}^{-1} Q_{(p)}^T c$$

where $R_{(p)} = Q_{(p)}^T Q_{(p)}$ and $u_{(p)} = Q_{(p)}^T c$

Using Schur decomposition:

$$\hat{a}_{(p)} = \begin{bmatrix} I + R_{(p-1)}^{-1} r_p s_{(p)}^{-1} r_p^T & -R_{(p-1)}^{-1} r_p s_{(p)}^{-1} \\ -s_{(p)}^{-1} r_p^T & s_{(p)}^{-1} \end{bmatrix} \begin{bmatrix} \hat{a}_{(p-1)} \\ u_p \end{bmatrix}$$

where $s_{(p)}^{-1}$ is the (scalar) Schur complement of $R_{(p-1)}$.

Recursive computation is appealing for a sequential scheme.

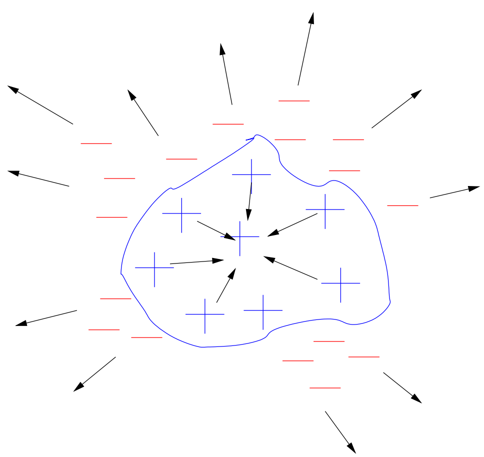
Experimental setup

- Dataset
  - 2nd and 3rd game of the 2003 Australia open final game
  - Synthetic surface; Indoor
  - type of events: racket hits; voice (player grunts, umpire, broadcaster); applause
  - events: 20ms time signal sequentially detected with GCUSUM [MLSP'06]
- Aim: identifying the racket hits (isolation)
- $1 - x$ class problem
  - Training set (2nd game): 203 detected events including 20 racket hits
  - Tested set (3rd game): 556 detected events including 73 racket hits

Bias Discriminant Analysis (BDA)

Modification of Linear Discriminant Analysis designed to address $1 - x$ class problem [Zhou et al/ 01].
Between scatter matrix S_b and within scatter matrix S_w:

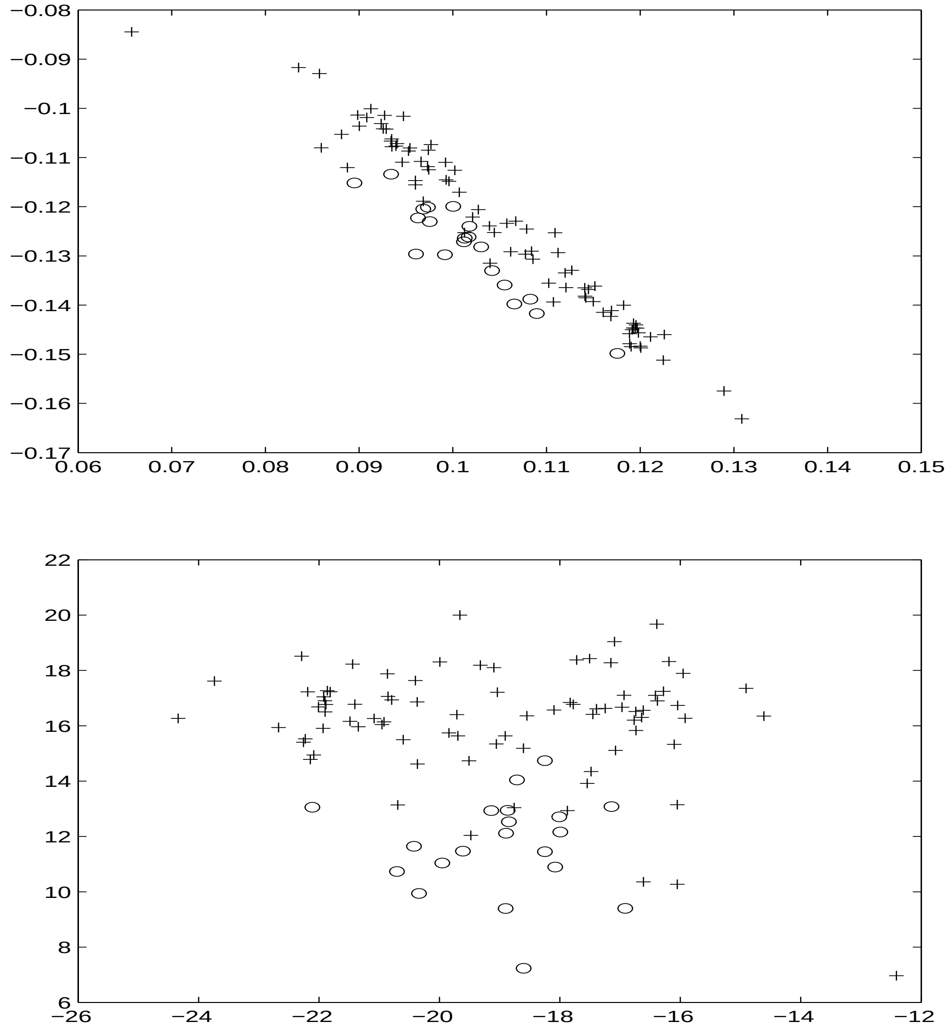
$$
\begin{aligned}
S_b &= \sum_{\mathcal{R}} (a^k - m_r)(a^k - m_r)^T, \\
S_w &= \sum_{\overline{\mathcal{R}}} (a^k - m_r)(a^k - m_r)^T,
\end{aligned}
$$



- a^k: kth feature
- $\mathcal{R}$: set of positive features
- $\overline{\mathcal{R}}$: set of negative features
- m_r: mean of positive features

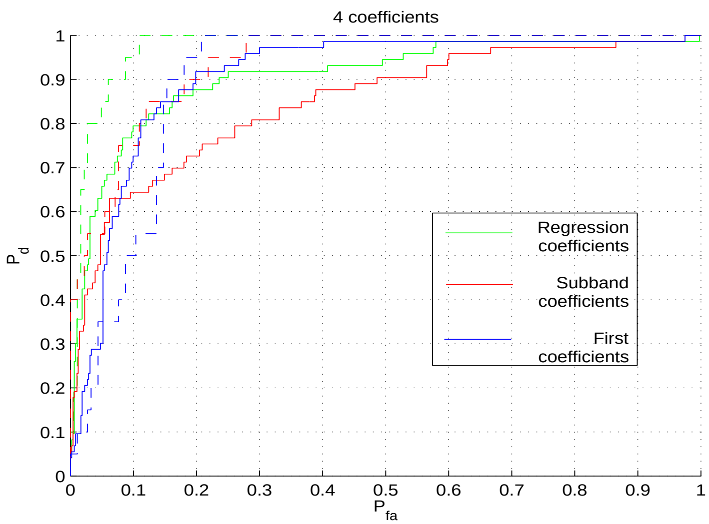
BDA tends to cluster $\mathcal{R}$ while spreading $\overline{\mathcal{R}}$

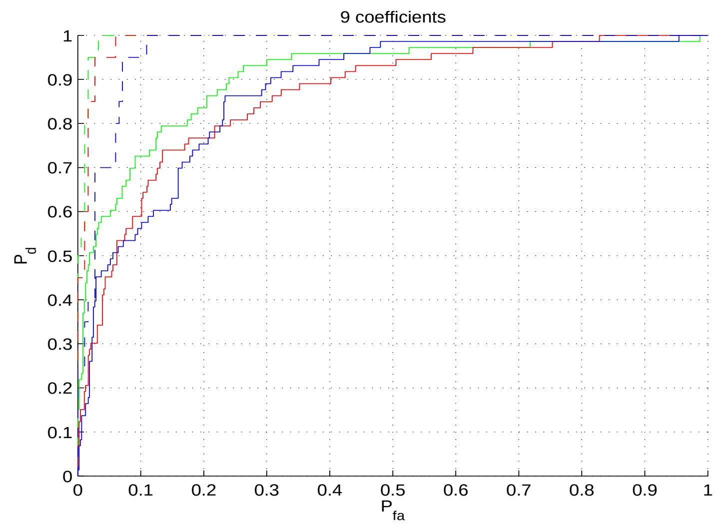
Bias Discriminant Transform

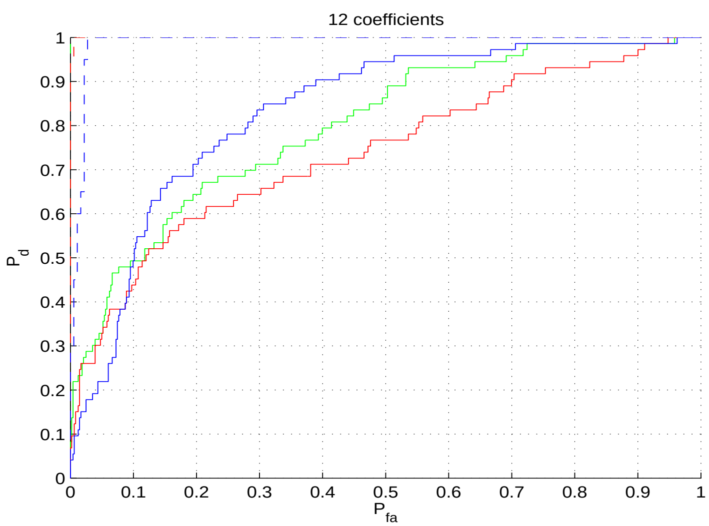


Fixed sample size: Results

Detection by thresholding the Likelihood of the positive cluster (Gaussian assump.)







Detection of training set (dashed line):

- **upper bound on detection**
- **index of characterization of training set**

perf. on training set increases with # coeff. while perf. on tested set decreases

Evolution of the positive cluster with time

Relevance Feedback Experiment

- aim: to evaluate the need for adaptive scheme for identification from audio features
- user play the role of video agent
- recursive data retrieval scheme (MPEG-7 style)
  1. current events: retrieval of positive events between the m_cth and $(m_c + T)$th events
  2. the M events including M_r positive events (racket hits) previous to the m_cth event are used to compute the BD transform
  3. BDA and ranking is applied to the T current events
  4. go to the $m_c + T$ event and repeat from 1.

Assessment of the Ranking performances

Dataset of size N with N_r positive instances in the first T ranks (top T).

- Average Normalized Modified Retrieval Rank (ANMRR) [Müller et al.01], [Manjunath et al.01]:

$$\tilde{r} = \frac{1}{S_{max} - S_{min}} \left(\sum_{k=1}^{N_r} r_k - S_{min} \right),$$

$$\tilde{r} = \frac{1}{(T - N_r)N_r} \left(\sum_{k=1}^{N_r} r_k - \frac{N_r(N_r + 1)}{2} \right).$$

- Properties
  - Minimum ANMRR (best case)

$$\tilde{r}_{\mathrm{min}} = 0.$$

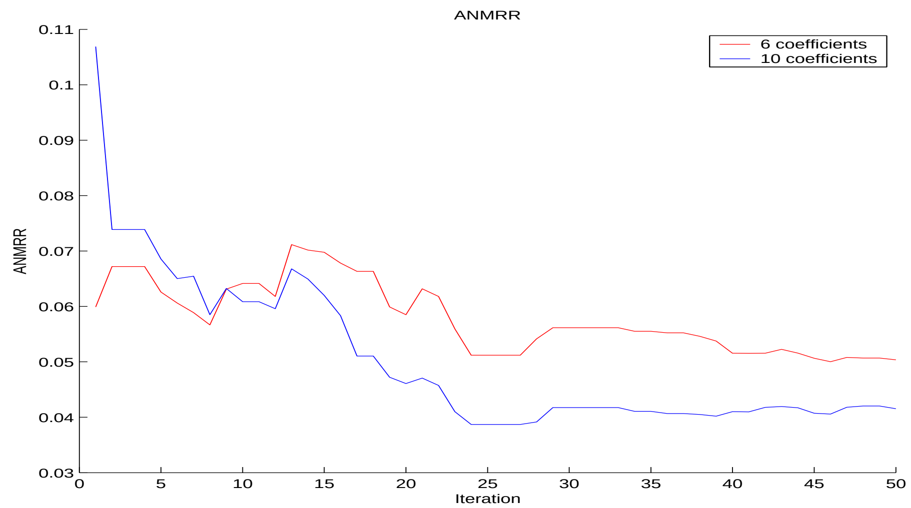
  - Maximum ranking (worst case)

$$\tilde{r}_{\mathrm{max}} = 1.$$

  - Expected value for a uniform random ranking

$$\mathrm{E}\{\tilde{r}\} = 0.5.$$

Results



- Good overall performance of both features (i.e. $\tilde{r} < 0.5$)
- increasing trend of the performance of the sequential retrieval
- From iteration 9, the dimension of feature space is worth being incremented to 10 dimensions.

How reliable is the ANMRR?