Cepstral Polynomial Regression For Sequential Detection Of Impulsive Waveform In Video Sound-Track

Cyril Hory, William J. Christmas, Anil Kokaram

***Abstract*— A new set of features is introduced for characterization of impulsive events in video clips from the audio signal. The discriminative power of these features to detect and isolate racket hits in tennis video clip is discussed.**

***Index Terms*— cepstral features, sequential classification**

I. Introduction

In digital video analysis it is now apparent that audio cues extracted from a video, along with the visual cues, can provide relevant information for semantic understanding of the content. In [5] for example, audio and visual features are combined within an HMM framework for parsing tennis video. Audio-visual cooperation is ensured through multi-modal conditional density estimation in [1]. Mel-cepstral coefficients are used in [4] to identify specific sounds in a baseball game video-clip in order to detect commercials, speech or music using the maximum entropy method.

In many application domains it is possible to identify some critically informative elementary short-terms event. However, even though visual attributes can be as informative as audio attributes for characterizing short-terms event, audio data are more convenient to handle in terms of computation load. Among short term events impulsive waveforms can be peculiarly informative. For instance accurate percussive sound detection can help with beat analysis. Racket hits are particularly informative events for the understanding a tennis game. From the detection and characterisation of racket hits, it is possible to extract information such as the score, player fitness and skills, or the strategy.

We have proposed in [3] a semi-supervised sequential scheme for detecting events from the audio stream of a video sequence using the Generalized CUSUM procedure. Following this detection step, a system for event identification can be triggered when an event is detected.

II. Cepstral features extraction

We focus here on the identification of impulsive waveforms from the audio content after detection. We propose a new set of features based on cepstral analysis of the recorded events. Denote by $c = [c_1, c_2, \dots, c_N]^T$ the vector of cepstral coefficients of the event e [2]:

$$c = |\mathrm{FT}^{-1}\{\log(|\mathrm{FT}\{e\}|)\}| \ , \tag{1}$$

where FT$\{.\}$ is the discrete Fourier transform. Assume there exists a vector $a_{(p)} = [a_0, a_1, \dots, a_p]^T$ such that

$$c = Q_{(p)} a_{(p)} + \nu_{(p)} \ , \tag{2}$$

where $Q_{(p)}$ is a $N \times (p+1)$ matrix with element $g_{j-1}(q_{i-1})$ on the ith row and jth column, where g_j can be any conveniently chosen polynomial of order j and q_i the ith quefrency index, and $\nu_{(p)}$ is an $N \times 1$ vector of random perturbations. The mean-square error estimate $\hat{a}_{(p)}$ of the vector of regression coefficients $a_{(p)}$ is:

$$\hat{a}_{(p)} = R_{(p)}^{-1} Q_{(p)}^T c \ , \tag{3}$$

with $R_{(p)} = Q_{(p)}^T Q_{(p)}$. The Cepstral Regression Coefficients (CRC) $\hat{a}_{(p)}$ are descriptors of the cepstrum content of the detected events.

In unsupervised sequential classification and learning the amount of available data is often limited and small. Low dimensional feature spaces must be considered in order to cope with the curse of dimensionality. In such a situation the CRC's allow for the encoding of the whole information carried by the cepstrum in a small number of coefficients. Moreover the inversion of the $(p+1) \times (p+1)$ matrix $R_{(p)}$ in (3) can be performed recursively by using a block matrix decomposition and the Schur complement of $R_{(p-1)}$. The dimensionality of the feature space can thus be adaptively updated to match the size of the dataset. The recursive computation of the regression coefficients makes these features particularly appealing for the implementation of a sequential classification system.

III. Experimental validation

A classification experiment was carried on excerpts of tennis video clip to evaluate the capability of the proposed features to discriminate impulsive waveforms. The impulsive waveform (target class of the classification experiment) are the racket hits. A Biased Discriminant Analysis [6] was performed within a supervised learning framework. The experiment has been carried on the second and third game of the Australian Open final tennis game of 2003. The training set contains 203 events that were detected by the CUSUM test including 20 actual racket hits. The test data contains 479 events that were detected

Cyril Hory is with Laboratoire Traitement et Communication de l'Information, CNRS-GET/Télécom-Paris, 37-39 rue Dareau, 75014 Paris.

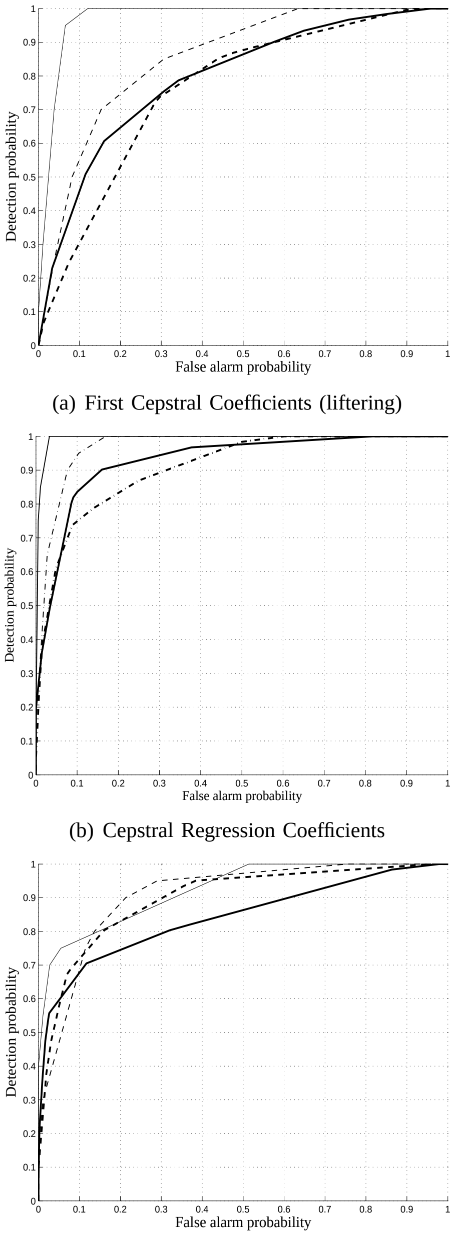
William J. Christmas is with University of Surrey, Centre for Vision Speech and Signal Processing, Guilford GU2 7XH, UK

Anil Kokaram is with University of Dublin, Trinity College, EEE Department, College Green, Dublin 2 Ireland.

This work was carried out during the tenure of a MUSCLE Internal fellowship.

by the CUSUM test including 61 racket hits.

On Fig. 1 are displayed the ROC curves of the classifier based on the First Cepstral Coefficients (FCC), CRC's, and Spectral Regression Coefficients (SRC) of the extracted events. The 3-dimension CRC vector outperforms the FCC's whether the training set or the test set is classified. The 3 FCC's fail to encode enough information about the event.



(a) First Cepstral Coefficients (liftering)

(b) Cepstral Regression Coefficients

(c) Spectral Regression Coefficients

Fig. 1. ROC curve of the Biased Discriminant Analysis for order 2 (dash-dotted line) and order 6 (plain line) of the (a) first cepstral coefficients, (b) cepstral polynomial regression coefficients, and (c) spectral polynomial regression coefficients . The grey lines correspond to the analysis of the training data. The black lines correspond to the analysis of the test data.

The test set classification performances of the CRC's and SRC's are equivalent although the CRC's outperforms the SRC's when applied on the training set.

The 7-dimension FCC's performs better than the 3-dimension FCC's when applied to the training set although the performances on the test set are similar whatever the feature space dimension. When applied to the training set, the 7-dimension FCC vector behaves as the CRC's and outperforms the SRC's. However the performances of the FCC's dramatically deteriorate when applied to the test set. This shows that a classifier based on the FCC's is spoiled by an over-fitting phenomenon. If modelling the waveform as the convolution of a source waveform and a filter impulse response, FCC's encode information about the filter while high quefrency coefficients are characteristic of the source [2]. In the experiment the impulse response of the filter depends on the acoustic characteristics of the hall and on the electronic recording device. It is a common to all the extracted events. Thus the filter characteristics can not provide relevant information for discriminating the various events.

The CRC of order 6 exhibits a better ability to characterise and discriminate the racket hits than the CRC of order 2 except at high false alarm probability. This shows that the CRC of order 6 tends to perform an over-fitting of the training set. As a consequence, outliers racket hits are more often taken into account in the model. In this case, the outliers are two lifted shots. The high probability of detection obtained, even though the characteristic of the class has evolved from the second to the third game, shows that the CRC features are relevant to encode the cepstrum content of a racket hit in a non-stationary context.

IV. Conclusion

The CRC's perform better than the standard FCC's in a low-dimension feature space but the discrepancy between performances of the two feature vectors decreases when the dimension increases. One can conclude that the CRC's seem more appropriate for classification of impulsive waveform when dealing with a small data set. In a higher dimension feature space performances are equivalent but feature vector computed from the polynomial regression (CRC's and SRC's) tends to provide less over-fitting than the standard FCC's.

The high discriminating power in a small dimensional feature space and the recursive computation of the features allows for their integration in a sequential and adaptive learning system. Work is currently being carried to show how the proposed features could improve the retrieval results obtained here in a static supervised learning context.

References

[1] R. Dahyot, A. Kokaram, A. Rea, and H. Denman, "Joint audio visual retrieval for tennis broadcasts," in *Proceedings of IEEE ICASSP'03*, 2003, pp. 561–564.

[2] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. MacMillan, 1993.

[3] C. Hory, A. Kokaram, and W. J. Christmas, "Threshold learning from samples drawn from the null hypothesis for the GLR CUSUM test," in *Proc. IEEE MLSP*, 2005, pp. 111–116.

[4] W. Hua, M. Han, and Y. Gong, "Baseball scene classification using multimedia features," in *Proceedings of IEEE ICME'02*, 2002, pp. 821–824.

[5] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot, "HMM based structuring of tennis videos using visual and audio cues," in *Proc. IEEE ICME'03*, 2003, pp. 309–312.

[6] X. S. Zhou and T. S. Huang, "Small Sample Learning during Multimedia Retrieval using BiasMap," in *Proceedings of IEEE CVPR'01*, December 2001, pp. 11–17.