THRESHOLD LEARNING FROM SAMPLES DRAWN FROM THE NULL HYPOTHESIS FOR THE GENERALIZED LIKELIHOOD RATIO CUSUM TEST

C. Hory, A. Kokaram[*]

University of Dublin, Trinity College
EEE Department
College Green, Dublin 2 Ireland

W. J. Christmas

University of Surrey
Centre for Vision Speech and Signal Processing
Guilford GU2 7XH, UK

ABSTRACT

Although optimality of sequential tests for the detection of a change in the parameter of a model has been widely discussed, the test parameter tuning is still an issue. In this communication, we propose a learning strategy to set the threshold of the GLR CUSUM statistics to take a decision with a desired false alarm probability. Only data before the change point are required to perform the learning process. Extensive simulations are performed to assess the validity of the proposed method. The paper is concluded by opening the path to a new approach to multi-modal feature based event detection for video parsing.

1. INTRODUCTION

Change detection in the temporal evolution of a signal is of key importance in a wide range of applications from system failures control and diagnosis to event detection in video streams. Event detection in video streams is important for parsing the media to create keyframes, or edit points or summaries [1]. The essential idea is that changes in the nature of measured features reflect a change in the semantics of the video event. For instance, in sports like snooker and tennis, a change in geometry of the scene generally reflects a change in camera view, which itself is an indicator for further analysis [2].
Obviously, using both audio and video features, would yield much better performance in video parsing and recently several efforts have been made to exploit that power [3]. The Hidden Markov Model has been the principal framework for inference with temporally varying multimodal data thus far. Mainly offline algorithms exploiting the HMM are considered for video parsing with mixed audio and visual features. Gish [4] proposed a distance measure for speech detection which can be linked to the offline approach described in [5] of the CUSUM (for CUmulative SUM) algorithm. Here again, Boreczky and Wilcox proposed to use this distance in a HMM framework for video parsing.
However, the HMM treats audio and visual features as components of the state vector with no discrimination regarding the nature of the data. In this paper, the parsing problem is decomposed into two parts: i) citing that *some* event has occurred by detecting a change in the statistical model of the audio data and then ii) identifying that event by processing of *some* relevant visual features. In so doing it is more easy to attribute varying weights to the power of inference for each feature. It transpires that for sport, quite often the audio stream is more beneficially used in this manner than as a component of an HMM state space. A good example is tennis, in which the sound of a ball hitting the racket is an essential feature of the game.

Sequential analysis provides a straightforward framework for online event detection from audio signal in document exploration. Sequential analysis was introduced by Wald in the framework of statistical decision in the late 40's [6]. He proved that for a given power, the Average Sample Number (ASN) required to perform a Sequential Probability Ratio Test (SPRT) is smaller than the number of sample required to perform the corresponding fixed sample size Likelihood Ratio Test (LRT). This appealing result opened the path to research in the field of sequential detection. In particular the problem of determining whether a change in a model occurs can be formulated within Wald's framework. Page [7] introduced the CUSUM test as a solution to this problem. An extensive study of the existing work in change detection is provided in [5]. More recently, Nikiforov [8] addressed the problem of change detection and isolation, namely the problem of detecting and identifying a change in a statistical model. This can be seen as the equivalent of multiple hypothesis testing where the probability density function (PDF) parameter under the alternative hypothesis is likely to take one among several values. Note that isolation [8] requires a set of known possible values of the parameter after change. If such knowledge is not available, a composite hypothesis scheme is invoked. The parameter after change is considered to belong to a domain of the pa-

[*]This work was partly supported by ERCIM within the MUSCLE European Network of Excellence.

rameter space. Some reinforcement is then to be provided to perform the isolation task.
Solutions to the simple and multiple hypothesis testing problems exist which are shown to be optimal with respect to certain criterion [5, 9]. For the composite hypothesis testing, a sequential optimal equivalent to the Generalized Likelihood Ratio test has not been found so far. However, some algorithms are commonly accepted as almost optimal. One can cite for example the Generalized Likelihood Ratio (GLR) CUSUM test proposed by Lorden [10]. Another drawback of the current state-of-the-art in change detection is the test parameter tuning issue. A sequential learning scheme has been proposed by Bershad and Sklansky where the threshold was considered as solution of a diffusion equation controled by positive and negative reinforcement [11]. Besides the need for a training sample drawn from both hypotheses, this approach provides analytic solutions in only few simple cases. Usually, thresholds are manually set, assuming a priori knowledge on the analysed data [5].
In this communication we propose an approach to automatically set the threshold to apply to the GLR CUSUM statistics. Optimality criteria of sequential tests such as the CUSUM test are defined as a lower bound on a detection performance index for a given false alarm index. Thus, it is possible to tune the threshold in order to perform a test with upper-bounded false alarm by using a training sample only drawn from the PDF under null hypothesis (i.e. PDF before the change point). We make use of the formulation of the CUSUM test as a set of parallel open-ended SPRT [10] to perform a jackknife-like estimation of the threshold.

After reviewing necessary results on sequential analysis in section 2, we propose three criteria for the tuning of the threshold in section 3. In section 4 we perform Monte-Carlo simulations for comparing and validating the efficiency of the proposed criteria. We conclude in section 5 by proposing a potential application of the presented results to event detection in tennis broadcasting.

2. BACKGROUND ON SEQUENTIAL ANALYSIS

In this section, we recall the results of theory of change detection that are used to derive the threshold criteria proposed in this paper.

2.1. Sequential Probability Ratio Test

Suppose a set of independent and identically distributed random variables $x_1, x_2, ...$ are sequentially sampled from the parent random variable X having PDF $f_\theta(x)$ parameterized by the scalar θ. the problem is to decide as soon as possible between the two hypotheses:

$$\begin{cases} H_0: & x_i \sim f_{\theta_0}(x), \quad \forall i \\ H_1: & x_i \sim f_{\theta_1}(x), \quad \forall i \end{cases} \tag{1}$$

Let

$$S_1^n(\theta_0, \theta_1) = \sum_{i=1}^{n} \ln \left\{ \frac{f_{\theta_1}(x_i)}{f_{\theta_0}(x_i)} \right\}, \tag{2}$$

be the log-likelihood ratio of the n first sampled data. The dependence of the test statistic on the parameters of the problem θ_0 and θ_1 is omitted when no confusion is possible. Set two constants A and B and the sequence of instants $T_1, T_2, ...$ such that $S_1^{T_k} \geq A$ or $S_1^{T_k} \leq B$. Then the stopping time T of the SPRT is defined as $T = \min_k\{T_k\}$. Hypothesis H_0 is accepted if $S_1^T \leq B$ and hypothesis H_1 is accepted if $S_1^T \geq A$. Wald has shown that the false alarm probability α and miss detection probability β and the bound A and B satisfy the following inequalities [6]:

$$\begin{cases} A \leq \ln \left\{ \frac{1-\beta}{\alpha} \right\}, \\ B \geq \ln \left\{ \frac{\beta}{1-\alpha} \right\}, \end{cases} \tag{3}$$

Under specific conditions widely accepted in practice, inequalities (3) can be replaced by approximations and are called the Wald's approximations. A special case of SPRT is the open-ended test for which the test statistic presents no lower bound B. In other words, the stopping time of an open-ended test depends only on the upper threshold A of the log likelihood ratio statistic.
The SPRT is said to be optimal in the sense that for given error probabilities, it provides a minimum Average Sample Number $E_{\theta_1}\{T\}$.

2.2. CUSUM test

Suppose now that we are to decide whether a change in the distribution of the sampled data has occurred at instant t_0. The null and alternative hypotheses are now:

$$\begin{cases} H_0: & x_i \sim f_{\theta_0}(x), \quad \forall i \\ H_1: & x_i \sim f_{\theta_0}(x), \quad i < t_0, \\ & x_i \sim f_{\theta_1}(x), \quad i \geq t_0, \end{cases} \tag{4}$$

The statistic S_c^n of the CUSUM test is built from the partial log-likelihood ratio $S_l^n = S_1^n - S_1^{l-1}$ of the first n data sampled:

$$S_c^n = \max_{1 \leq l \leq n} \{S_l^n\}. \tag{5}$$

Given threshold A, the stopping time T is defined by $T = \min\{T_k, k = 1, 2...\}$ where the T_k are such that $S_c^{T_k} \geq A$ [5].
Lorden has proved in [10] that for a given rate of false alarm, the mean time delay for detection of the CUSUM test reaches the minimum achievable by a sequential test. This states the

optimality of the CUSUM test under a minimax type of criterion.

Consider now that there is less a priori information on θ_1. More specifically, parameter θ after the change point is only known to belong to a domain Θ_1 such that $\theta_0 \notin \Theta_1$. The problem is to decide as fast as possible between the two hypotheses:

$$\begin{cases} H_0: & x_i \sim f_{\theta_0}(x), & \forall i \\ H_1: & x_i \sim f_{\theta_0}(x), & i < t_0, \\ & x_i \sim f_{\theta_1 \in \Theta_i}(x), & i \geq t_0, \end{cases} \quad (6)$$

The GLR CUSUM test resolves this problem by applying a threshold A to the test statistics

$$S_g^n = \max_{1 \leq l \leq n} \{ \max_{\theta \in \Theta_1} \{ S_l^n(\theta, \theta_0) \} \}. \quad (7)$$

The CUSUM test is a special case of the GLR CUSUM test in the case where Θ_1 reduces to the singleton $\{\theta_1\}$.

3. THRESHOLD LEARNING IN PRESENCE OF SAMPLE DRAWN FROM THE NULL HYPOTHESIS

Suppose independent identically distributed random variables $x_1, x_2, ...$ are sequentially sampled from a PDF $f_\theta(x)$ of the exponential family with varying but unknown θ. Suppose also it is known that for the N first samples $\theta = \theta_0$. Some unknown changes in θ occur at unknown instants $t_i > N$, $i = 0, 1, \ldots$ The problem is to detect each change of parameter θ.
A CUSUM-like algorithm is obviously the best solution to this problem since each decision has to be taken as soon as possible, i.e. before a new change occurs. Moreover, the parameter after the change point can take any value so we have to apply a GLR CUSUM test where $\Theta_1 \subset I\!R \setminus \theta_0$.
Since no information is available concerning domain Θ_1, it is impossible to assess a threshold value A to the test which will satisfy a desired mean time delay to take a decision for a given rate of false alarm. We propose to use the N first samples drawn from the known PDF $f_{\theta_0}(x)$ to set a threshold that ensures at most a required false alarm probability.

3.1. Principle of the learning procedure

The Kulback-Leibler (KL) information $I(\theta_0, \theta_1)$ shared by two PDF's $f_{\theta_0}(x)$ and $f_{\theta_1}(x)$ is defined by:

$$I(\theta_0, \theta_1) = E_1 \left\{ \ln \left\{ \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \right\} \right\} \quad (8)$$

The KL information is the mean increment of the likelihood ratio statistic under hypothesis for data drawn from $f_{\theta_1}(x)$. It can be seen as a degree of detectability between hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ [5]. When $\theta_0 = \theta_1$, the Kulback-Leibler information is null; performing a test of detection is actually meaningless in such a case.
It is easy to see that up to instant N, the GLR CUSUM stopping rule defined previously is equal to the extended stopping time of N parallel open-ended SPRT triggered at instants $k = 1, 2, ...N$ [10]. When performing a GLR CUSUM test as a set of parallel open-ended tests, $N(N-1)/2$ statistics are computed. So, considering the GLR CUSUM test as a set of open-ended tests allows straightforwardly to place ourselves in a jackknife [12] approach: we consider that the statistics and the maximum likelihood estimates of θ_1 under null hypothesis are samples of the corresponding random variables. We propose to tune the GLR CUSUM parameters by evaluating a minimum degree of detectability using the jackknife approach.

3.2. Threshold selection criteria

Consider that the parameter after the change point is known to belong to domain Θ_1 and define:

$$\theta_m = \arg \min_{\theta \in \Theta_1} \{ I(\theta_0, \theta) \}, \quad (9)$$

i.e. the minimum measure of detectability of the change. There exists the following relation between false alarm probability α of the open-ended SPRT and threshold A [5]:

$$A = \ln \left\{ \frac{-3 \ln\{\alpha\} \left[1 + \frac{1}{I(\theta_0, \theta_m)} \right]^2}{\alpha} \right\}. \quad (10)$$

However, as no information is available about the parameter after the change point, θ_m is unknown in our problem. So, during the learning session, we propose to look for θ_m such that $I(\theta_0, \theta_m)$ is a measure of affordable detectability.

Denote $\hat{\theta}_l^n = \arg\max_{\theta \in \Theta_1} \{ S_l^n(\theta_0, \theta) \}$, the maximum likelihood estimator of θ computed from samples between instant l and n. We propose three criteria for evaluating θ_m:

- first criterion: *mean parameter criterion*.
  Parameter θ_m in (9) is the mean of the $N(N-1)/2$ estimates $\hat{\theta}_l^n$ required to compute the test statistics:

$$\theta_m^1 = \frac{2}{N(N-1)} \sum_{l,n} \hat{\theta}_l^n. \quad (11)$$

- second criterion: *mean information criterion*.
  Parameter θ_m in (9) is the mean of the KL information between $f_{\theta_0}(x)$ and $f_\theta(x)$ for each $\theta = \hat{\theta}_l^n$:

$$I(\theta_0, \theta_m^2) = \frac{2}{N(N-1)} \sum_{l,n} I(\theta_0, \hat{\theta}_l^n) \quad (12)$$

- third criterion: *bound criterion*.
  For an open ended test with threshold A, and PDF of the exponential family, it can be shown using the Wald's approximation that for $N = \frac{A}{I(\theta_0,\theta_m)}$ the false alarm probability of an open ended test can be approximated by:

$$P_{\theta_0}(T \leq N) \approx N \exp\{-A\}. \qquad (13)$$

  So we choose the θ_m which minimizes the difference between the left and right hand sides of approximation (13):

$$\theta_m^3 = \arg\min_{\hat{\theta}_l^n} \{N \exp\{-NI(\theta_0, \hat{\theta}_l^n)\} - \frac{\text{card}\{S_g^n \geq nI(\theta_0, \hat{\theta}^n)\}}{N}\}, \qquad (14)$$

  where $\hat{\theta}^n = \frac{1}{n}\sum_{l=1}^{n} \hat{\theta}_l^n$. The choice of the second term in the right-hand side of (14) to estimate the probability of false alarm is motivated by considering S_g^n as the test statistic of a reverse-time sequential test performed up to time n.

4. EXPERIMENTAL VALIDATION

An experiment has been conducted to compare the learning capability of the method when applied with the three proposed criteria. We focussed on Gaussian data of constant mean presenting a change in the variance. As discussed in section 5, this model is relevant for event detection from the audio recording of a video clip.

6000 trials with data sets consisting of 800 samples having a Gaussian distribution with zero-mean have been performed. For each trial, change in the variance σ^2 of the distributions occurred at instant $t_0 = 301$. Before the change point $\sigma^2 = \sigma_0^2 = 1$. The 6000 values σ_1^2 of the variance after the change point were drawn from a uniform distribution $\mathcal{U}([1, 15])$. During the experiment, a change was considered as not being detected if no decision was taken after 500 samples. So, the experiment is performed as a truncated sequential test. It is most likely that both false alarm probability and detection probability should be higher if the stopping time could tend to infinity. However, in this case, the false alarm probability would increase less than the detection probability [13].

Fig. 1 shows the rate of false alarm computed from the 6000 GLR CUSUM tests performed versus the probability of false alarm α of the open-ended tests. Three different length $N = 100, 200, 300$ points of the learning set have been tested. The three criteria presented in the previous section are considered. The threshold A computed with the mean parameter criteria offers the smaller rate of false alarm while the threshold A bound criterion provides the higher rate of false alarm.

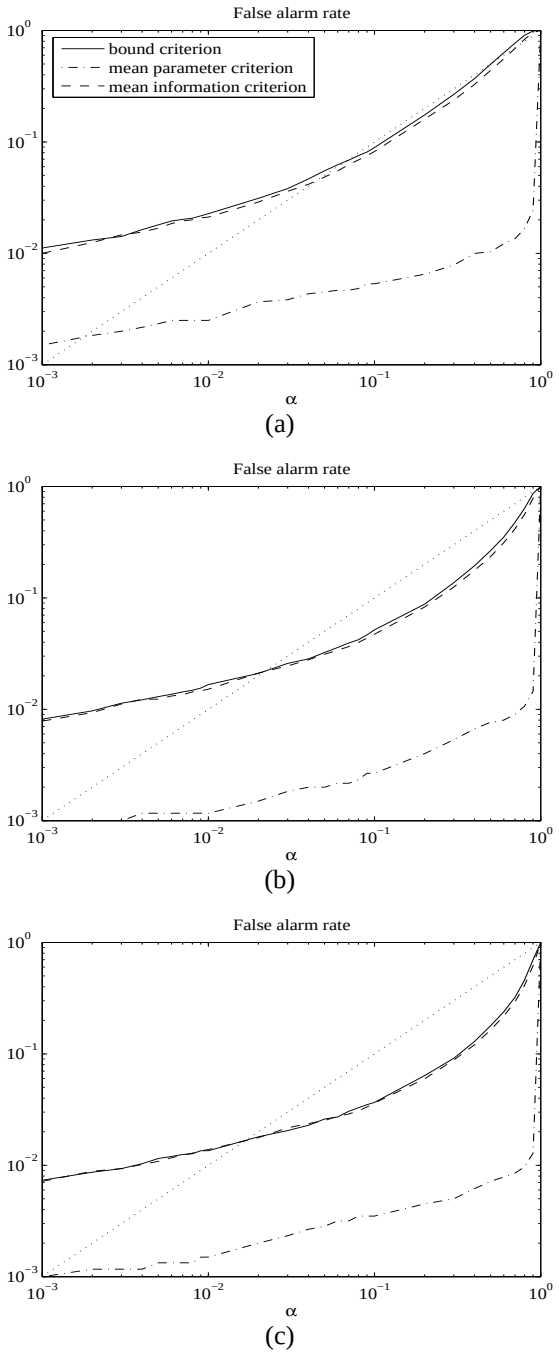

Fig. 1. False alarm probability of the GLR CUSUM test with respect to the false alarm probability α of the corresponding open-ended tests. The data set is of 800 samples. The learning length is $N = 100$ samples (a), $N = 200$ samples (b) and $N = 300$ samples (c).

However, for bound criterion and mean information criterion the probability of false alarm is closed to α. For small α the probability of false alarm get higher than α. We are not able to provide a definite explanation for such a behavior event though we suspect that the small number of trials performed (6000) can be considered as partly responsible.

To apply the bound criterion or the mean information criterion provides a probability of false alarm which is closed to α. However, it seems that when the learning length N increases, the difference between α and the actual probability of false alarm increases too. The mean parameter criterion provides a much smaller probability of false alarm. Furthermore, to apply the mean parameter ensures a loose upper bound to the actual probability of false alarm. This is to the expense of the probability of detection which in turn, is smaller. Indeed, on Fig. 2 is presented the detection rate for 250 trials performed from Gaussian data with a parameter after the change point $\sigma_1^2 = 1.5$. The probability of detection of the GLR CUSUM performed using the mean parameter criterion is smaller than when applying any of the two other criteria.

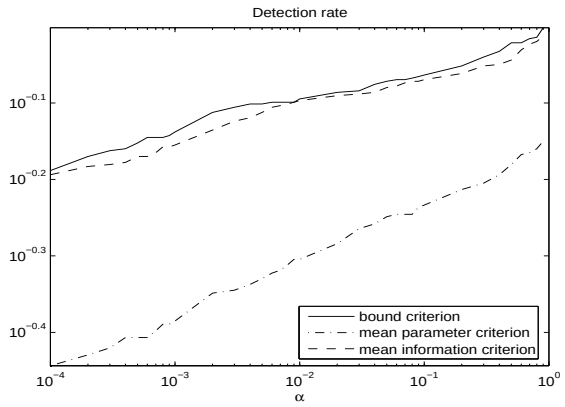

Fig. 2. Detection rate of the GLR CUSUM test versus the false alarm probability α of the corresponding open-ended tests. 250 experiments have been run with $\sigma_0^2 = 1$ and $\sigma_1^2 = 1.5$.

On Fig. 3 is plotted the variance of threshold A computed using the bound criterion and the mean information. It appears that when applying the bound criterion, the variance of A is smaller than when applying the mean information criterion. That makes the bound criterion a more reliable criterion.

5. DISCUSSION ON APPLICATION TO VIDEO PARSING

We have proposed a scheme for automatically tuning the threshold of a GLR CUSUM test to meet a desired probability of false alarm. It is only required that the learning set is drawn from the distribution before change. Three criteria have been proposed to evaluate the threshold, namely, the bound criterion, the mean parameter criterion and the mean information criterion. These criteria evaluate a minimum degree of detectability of a change in the parameter of interest from the learning set.

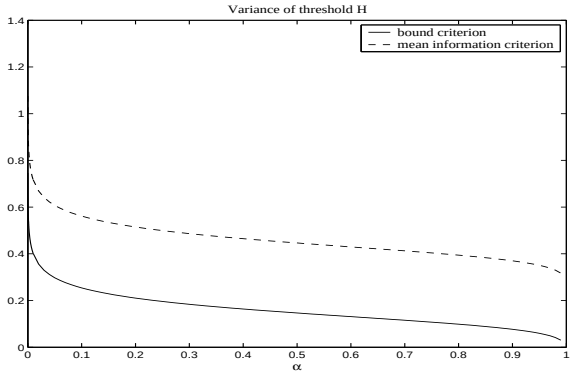

Fig. 3. Variance of the thresholds evaluated using the bound criterion and the mean information criterion. The learning length is $N = 200$ samples.

Experimental comparison of the three criteria has been conducted on simulated data. It turns out that the mean parameter criterion provides a probability of false alarm which is much smaller than the desired probability of false alarm. Since this implies a lowest probability of detection, we conclude that this criterion is not relevant. The two other criteria perform similarly in terms of probability of false alarm. However the bound criterion provides a smaller variance of the derived threshold than the mean information criterion. That makes this criterion eligible as the standard criterion.

Racket hit detection in tennis broadcast. We present now a methodology for video parsing which involves a sequential detection of event using the tuned GLR CUSUM. We consider the problem of detecting racket hits in a tennis rally from the sequential analysis of the audio recording. A racket hit is characterized by a very impulsive waveform with high energy. Thus we aim at detecting a change in the variance of the data when a racket hit occurs.

Fig.4 presents a typical audio recording of a rally. Five racket hits are to be detected. Moreover, one of the player shouted before the third racket hit. Before processing, data were subsampled so that independency could be assumed. In order to take account of echoes due to the acoustical characteristics of the tennis court, we assumed the duration t_1 of a racket hit was $t_1 = 450$ ms. The GLR CUSUM test was performed recursively as follows: denote T_k the instant of the kth alarm. The kth point of change $t_0^{[k]}$ is estimated us-

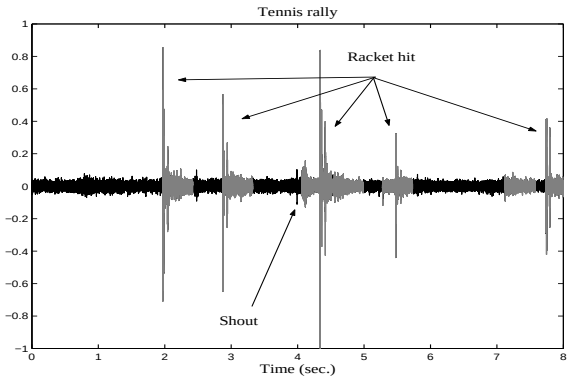

Fig. 4. Parsing of the audio recording of a tennis rally.

ing the maximum likelihood estimator:

$$\hat{t}_0^{[k]} = \arg\max_l \{S_l^{T_k}\}. \quad (15)$$

A new test is performed from starting point $t_{in} = \hat{t}_0^{[k]} + t_1$. Threshold H was tuned so that the probability of false alarm $\alpha = 10^{-3}$ is met. The threshold was trained from the 200 first samples. This corresponds to a learning time of 453 ms.
The algorithm has detected 7 change points $\hat{t}_0^{[k]}$ in the variance of the data. On Fig.4, the 7 segments $[\hat{t}_0^{[k]}, \hat{t}_0^{[k]} + t_1]$ are plotted in grey color. The five racket hits were successfully detected. The shout was also detected as well as a small variation in the variance of the noise, due to a variation in the recording system tuning. The times of occurrence of the first, second and fifth racket hits were accurately estimated. The two errors are due to the detection of a slight variation of the magnitude of the signal. The first error is due to the detected shout around instant $t = 4$ seconds. The other error is due to an unidentified event happening at instant 5.5 ms.

In spite of the diversity in the magnitude of the waveform characterizing the detected events, the threshold was correctly tuned for insuring a good rate of detection. However, additional information is required to isolate the different events in the sense defined by Nikiforov [8]. For example, specific to a tennis game, an a priori probability for a shout to happen right before a racket hit can allow to discriminate the voice shout happening at instant 4 ms and the subsequent racket hit. In a multi-modal cooperation framework [14], we are currently investigating reinforcement provided by visual features such as player or ball tracking or motion field estimate to isolate the detected events.

6. REFERENCES

[1] C. Colombo, A. Del Bimbo, and P. Pala, "Semantics in Visual Information Retrieval," *IEEE Multimedia*, vol. 3, no. 6, pp. 38–53, 1999.

[2] R. Dahyot, N. Rea, and A. C. Kokaram, "Sport Video Shot Segmentation and Classification," in *Proceedings of Visual Communications and Images Processing*, Lugano, Ch., July 2003.

[3] Yao Wang, Zhu Liu, and Jin-Cheng Huang;, "Multimedia Content Analysis Using Both Audio and Visual Clues," *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12 – 36, November 2000.

[4] H. Gish, M. Siu, and R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification," in *Proceedings of IEEE ICASSP'91*, Toronto, Ca., April 1991, vol. 2, pp. 873–876.

[5] M. Basseville and I. Nikiforov, *Detection of abrupt changes. Theory and applications*, Information and system sciences series. Englewood Cliffs, NJ: Prentice Hall, 1993.

[6] A. Wald, *Sequential Analysis*, Wiley and Sons, New-York, 1947.

[7] E. S. Page, "Continuous Inspection Schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, June 1954.

[8] I. V. Nikiforov, "A Generalized Change Detection Problem," *IEEE Trans. on Info. Theory*, vol. 41, no. 1, pp. 171–187, 1995.

[9] T. L. Lai, "Sequential analysis: some classical problems and new challenges," *Statistica Sinica*, vol. 11, pp. 303–408, 2001.

[10] G. Lorden, "Procedures for Reacting to a Change in Distribution," *Ann. Math. Stat.*, vol. 42, pp. 1897–1908, December 1971.

[11] N. J. Bershad and J. Sklansky, "Threshold Learning and Brownian Motion," *IEEE Trans. on Information Theory*, vol. 17, no. 3, pp. 350–352, May 1971.

[12] B. Efron and R. Tibshirani, *An Introduction to Bootstrap*, Chapman & Hall, New York, 1993.

[13] D. Siegmund, *Sequential Analysis*, Springer-Verlag, 1985.

[14] J. Kittler, W.J. Christmas, A. Kostin, F. Yan, I. Kolonias, and D. Windridge, "A memory architecture and contextual reasoning framework for cognitive vision," in *Invited paper for SCIA 2005*.