On-line adaptive background modelling for audio surveillance

Marco Cristani, Manuele Bicego, and Vittorio Murino
Dipartimento di Informatica, Università di Verona - Italy
{bicego,cristanm}@sci.univr.it, vittorio.murino@univr.it

Abstract

In this paper, we investigate the problem of automatic audio surveillance. This aspect of the surveillance, which extends the more investigated area of video surveillance, can be very informative to solve many problems in real situations. Similarly to video surveillance, also in this case it is necessary to build a background (BG) model, so that it is immediate to discover foreground (FG) events. To this end, we first introduce the concepts of audio BG and FG in an automated surveillance scenario. Subsequently, we propose a novel audio BG system able to build in real time an adaptive model of the audio scene BG, and to promptly detect unexpected FG auditory events. The method is based on the probabilistic modelling of the audio data stream using separate sets of adaptive Gaussian mixture models, working on the audio frequency spectrum. This approach is also characterized by the use of only one microphone and on-line functioning, so that it can be directly used in real situations, also to support a video surveillance system. Preliminary results show the effectiveness of the approach to discover different FG audio situations.

1. Introduction

Automated surveillance systems have acquired in the pattern recognition area an increased importance in the last years, due to their effectiveness in discovering and classifying unexpected events in civil areas [6]. In this context, the most important low-level analysis is the so called background modelling, aimed at discriminating the expected (typically visual) information, namely, the background (BG), from the unexpected objects, i.e., the foreground (FG). In general, almost all of the methods work only at the visual level, hence resulting in *video* BG modelling schemes [9, 6]. This could be a severe limitation, since other information modalities are easily available (e.g., audio), which could be effectively used as additional information to discover unusual "activity patterns" in a scene.

In this paper, an on-line probabilistic audio BG modelling approach is proposed, which utilizes the data stream coming from only one microphone, and is adaptive to the several audio situations, so that FG events can be easily detected.

In literature, the audio surveillance problem has never systematically investigated. Only few works have exploited this channel, mainly based on the monitoring of the audio intensity for BG/FG discrimination, or aimed at recognizing specific class of sounds [3]. These methods are not adaptive to the several possible audio situations, and then do not exploit all the potential information conveyed by the audio channel. Other than the automated surveillance context, several approaches to computational audio analysis are present, mainly focused on the computational translation of psychoacoustics results. One class of approaches is the so called *computational auditory scene analysis* (CASA) [1], aimed at the separation and classification of sounds present in a specific environment. Closely related to this field, but not so investigated, there is the *computational auditory scene recognition* (CASR) [7, 2], aimed at overall environment interpretation instead of analyzing the different sound sources. Besides various psycho-acoustically oriented approaches derived from these two classes, a third approach, used both in CASA and CASR contexts, tried to fuse "blind" statistical knowledge with biologically driven representations of the two previous fields, performing audio classification and segmentation tasks [10], and source separation [8, 4] (i.e., blind source separation). In this last approach, many efforts are devoted in the speech processing area, in which the goal is to separate the different voices composing the audio pattern using several microphones [4] or only one monaural sensor [8]. Some of the hypotheses of our approach come out from this third subfield.

In our approach, a multiband frequency analysis was first carried out to characterize the monaural audio signal, by extracting energy features from a parametric estimation of the Power Spectral Density (PSD). For a comparative survey on the efficiency of different acoustic features, see [7, 2]. Subsequently we assume, as in [8], that the energy of the audio signal along time and in different frequency bands can transport independent information. The audio BG model is

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

then obtained by modelling the energy features using a set of adaptive mixtures of Gaussians, one for each frequency subband. Each subband is considered as an independent process, bringing a time-varying acoustic energy intensity, that is classified as *audio BG*, i.e. as expected acoustic information, or as *audio FG*, so that the current audio situation is monitored. More precisely, at each time step, a snapshot of the whole audio spectrum is considered: the current Gaussian components of the mixtures associated to every subband (and classified as audio FG or BG) jointly define a particular *audio event*. The mixtures of Gaussians are continuosly updated to take into account the varying situations and be aware of the different audio FG information that may occur even in presence of a complex audio BG. In summary, the paper introduces novel concepts related to the audio scene analysis, discussing the involved problems, showing potentialities and possible future directions of the research. The key contributions of this work are: 1) the definition of the novel concepts of audio BG and FG 2) the introduction of a multiband audio BG modelling, performing an auditory scene analysis using only *one* microphone; 3) the implementation of these audio principles in a probabilistic framework working on-line and able to deal with complex issues in automated surveillance.
The rest of the paper is organized as follows. In Section 2, the time-adaptive mixture of Gaussians method is presented. The application of this model in the audio context is proposed in Section 3, and preliminary experimental results are reported in Section 4. Finally, in Section 5, conclusions are drawn and future perspectives are envisaged.

2. The Time-Adaptive mixture of Gaussians method

The Time-Adaptive mixture of Gaussians method, well-know in the video surveillance context [9], aims at discovering the deviance of a signal from the expected behavior in an on-line fashion. The general method models a temporal signal with a time-adaptive mixture of Gaussians. The probability to observe the value $z^{(t)}$, at time t, is given by:

$$P(z^{(t)}) = \sum_{r=1}^{R} w_r^{(t)} \mathcal{N}\left(z^{(t)} | \mu_r^{(t)}, \sigma_r^{(t)}\right) \quad (1)$$

where $w_r^{(t)}$, $\mu_r^{(t)}$ and $\sigma_r^{(t)}$ are the mixing coefficients, the mean, and the standard deviation, respectively, of the r-th Gaussian of the mixture associated to the signal at time t. At each time step, the Gaussians in a mixture are ranked in descending order using the w/σ value. The R Gaussians are evaluated as possible match against the occurring new signal value, in which a successful match is defined as a signal value falling within 2.5σ of one of the component. If no match occurs, a new Gaussian with mean equal to the current value, high variance, and low mixing coefficient replaces the least probable component.

If r_{hit} is the matched Gaussian component, the value $z^{(t)}$ is labelled as unexpected (i.e., FG) if $\sum_{r=1}^{r_{hit}} w_r^{(t)} > T$, where T is a threshold representing the minimum portion of the data that supports the "expected behavior". The evolution of the components of the mixtures is driven by the following equations:

$$w_r^{(t)} = (1-\alpha) w_r^{(t-1)} + \alpha M^{(t)}, 1 \leq r \leq R, \quad (2)$$

where $M^{(t)} = 1$ for the matched Gaussian (indexed by r_{hit}), and 0 for the others; α is the adaptive rate that remains fixed along time. It is worthwhile to notice that the higher the adaptive rate, the faster the model is "adapted" to signal (i.e., auditory scene) changes.

The μ and σ parameters for unmatched Gaussians remain unchanged, but, for the matched Gaussian component r_{hit}, we have:

$$\mu_{r_{hit}}^{(t)} = (1-\rho)\mu_{r_{hit}}^{(t-1)} + \rho z^{(t)} \quad (3)$$

$$\sigma_{r_{hit}}^{2\,(t)} = (1-\rho)\sigma_{r_{hit}}^{2\,(t-1)} + \rho (z^{(t)} - \mu_{r_{hit}}^{(t)})^T (z^{(t)} - \mu_{r_{hit}}^{(t)}) \quad (4)$$

where $\rho = \alpha \mathcal{N}\left(z^{(t)} | \mu_{r_{hit}}^{(t)}, \sigma_{r_{hit}}^{(t)}\right)$.

3. The audio background modelling system

The audio BG modelling system aims at extracting information from audio patterns acquired by a *single* microphone. As considered in [8], the energy during time in different frequency bands can transport independent information. Moreover, the energy based features are well suited for recognition and classification tasks [7]. Therefore, we subdivide the audio signal $x(t)$, acquired at sampling frequency F_s, in temporal windows of fixed length W_a. For each time interval $[(t-1)W_a, tW_a]$, a parametric estimation of the Power Spectral Density (PSD) with the Yule-Walker Auto Regressive method [5] is carried out, obtaining the energy samples (in dB) $\{X^{(t)}(f_n)\}$, $n = 1, \ldots, N$, where f_n is the frequency, expressed in Hz, $f_N = F_s/2$, and the desired frequency resolution determines N, directly derived from the temporal window length W_a.

Subsequently, we introduce the *Subband Energy Amount* (SEA):

$$A_i^{(t)} = \sum_{n=D(i)}^{U(i)} X^{(t)}(f_n) \quad (5)$$

The $A_i^{(t)}$ represents an estimation of the sum of the spectral energy relative to the time interval $[(t-1)W_a, tW_a]$ for the i-th subband, $i = 1, 2, \ldots, M$, where each $i-$th subband A_i is bounded by the interval $[D(i), U(i)]$. As suggested in [7], we use M logarithmic spaced subbands in the range $[0, F_s/2]$. Then, we consider each SEA channel as a

Gaussian process $A_i^{(t)} = \mathcal{N}(\mu_i^{(t)}, \sigma_i^{(t)})$. Subsequently, using the adaptive method presented in Sec. 2, we instantiate one time-adaptive mixture of Gaussians for each SEA channel, updated with a fixed learning coefficient $\tilde{\alpha}$, depending on the adaptive rate considered. At this point, we must define the concepts of expected and unexpected acoustic behaviors, i.e., the definition of *audio BG* and *audio FG*. In the real world, we associate to a sound usually present in the scene the meaning of BG, in contrast to an unexpected sound pattern, never heard before. Therefore, we define the probability to observe the value $A_i^{(t)}$ as:

$$P(A_i^{(t)}) = \sum_{k=1}^{K} w_{k,i}^{(t)} \mathcal{N}\left(A_i^{(t)} | \mu_{k,i}^{(t)}, \sigma_{k,i}^{(t)}\right) \qquad (6)$$

where $w_{k,i}^{(t)}$, $\mu_{k,i}^{(t)}$ and $\sigma_{k,i}^{(t)}$ have the same meaning as in Eq. 1. Consequently, considering the SEA value $A_i^{(t)}$ at time step t, and defining k_{hit} the Gaussian component matched for this mixture at this instant, we can identify the acoustic energy of the subband as audio FG, if

$$\sum_{k=1}^{k_{hit}} w_{k,i}^{(t)} \; > \; P, \qquad (7)$$

where the threshold P, together with the audio learning rate $\tilde{\alpha}$, represent the specific parameters in the audio context, for which the considerations previously made in Sec. 2 are still valid. Finally, in order to classify different audio situations, we consider as *audio event* the configuration of the M mixtures determined by the indexes of the matched components. During time, a change in the configuration identifies an occurrence of a further event. In the next session, we prove that this method is able to discriminate different audio events, which can be exploited in perspective also to event recognition purposes.

4. Experimental session

In order to evaluate the proposed method, we have considered several audio sequences, taken from general outdoor and indoor surveillance situations. In all the examples, we have considered the SEA of 8 logarithmically divided subbands, in the range of $[0, F_s/4]$. For each subband we have instantiated a 4-components mixture of Gaussians, with threshold $P = 0.8$ and learning rate $\tilde{\alpha} = 0.01$. These parameters are discriminative for almost all the sequences analyzed. For the sake of clarity and space limitation, we analyze in detail two audio sequences, in an indoor and outdoor environment, respectively, and we report briefly other two tests.

The first audio example, the *Office* sequence, consists in a 10 seconds signal, sampled at 22050 Hz. The sequence is composed by three different audio events that sequentially occur: 1) BG noise situation ; 2) a talking person; 3) a falling chair while the person is talking. The signal is therefore subdivided with windows of length $W_a = 1$s. The proposed audio BG modelling is able to detect and separate the different audio events (Fig. 1). The benefit of using a mixture for each subband, rather than only one Gaussian, is evident: in the latter case only a FG/BG discrimination per time step is possible, while in the former more FG activities can be identifiable, proportionally to the number of components.

The second audio example, the *Engines* sequence, consists in a signal of 12 seconds length, sampled at 18 KHz. The sequence is formed by 4 different audio events sequentially composed as follows: 1) BG noise situation, 2) the first running engine A starts, 3) the second starting engine B overlapping with A, 4) A stops. Also in this situation, our method is able to identify all the audio events, as depicted in Fig. 2, due to the different matched Gaussian components composing the involved mixtures. Two other experiments, here briefly reported, are the *Entering* and the *Two voices* sequences. Both the signals are sampled at 22050 Hz. In the *Entering* sequence three events occur: 1) BG noise situation, 2) knocking on the door, 3) opening the door. In this situation too, the BG model discovers correctly the events. In the last sequence, 30 second long, two female speakers alternatively talk. In this case, the 2 audio events are identified, but not perfectly separated, because the voice spectral energy range is relatively narrow with respect to our original frequency subdivision. However, changing opportunely our spectral interval, and subdividing the interval $[0.1, 0.3]$KHz in two subbands, the BG model perfectly separates the two female voices, identifying the two different speakers. These experiments show the good potentialities of the approach, which is able to detect different FG activities, and, suitably trained to monitor specific situations, can possibly classify non-trivial audio events.

5. Conclusions

In this work, a probabilistic approach for the BG auditory scene modelling is presented. The audio signal, acquired by a single microphone, is managed by considering its frequency spectrum, in particular, by subdividing it in suitable subbands, assumed to convey independent information about the audio events [8]. Each subband is modelled by a mixture of Gaussians, which is adaptive to the possible different BG situations, being on-line updated over time. In this way, at each instant t, FG information is detected by considering the set of subbands which show atypical behaviours. The subbands' energies proved to be good features for audio surveillance purposes as they are able to characterize different audio events, ranging from speech

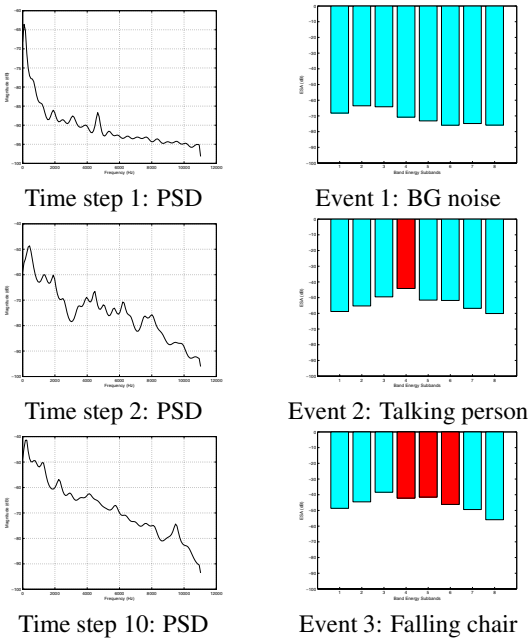

Figure 1. PSD (left), and the BG model (right), that classifies as FG (red) or BG (blue) the SEA of each subband, defining a particular event.

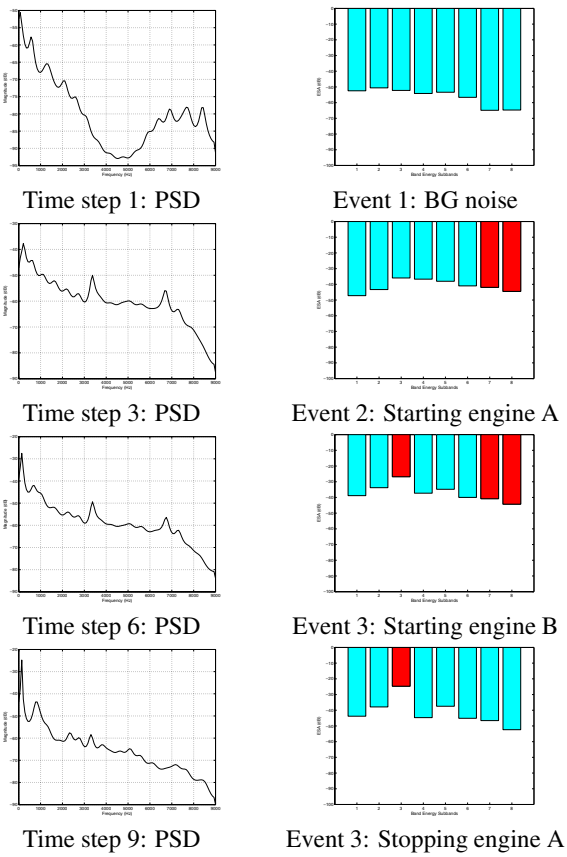

Figure 2. PSD (left), and the BG model (right), that classifies as FG (red) or BG (blue) the SEA of each subband, defining a particular event.

to object sounds, immersed in audio BG of different complexity, considering both indoor and outdoor enviromental sounds.

Moreover, they also show a sufficient discriminative power to be used for the classification of the audio events. Using an adequate classification tool, it is possible to train the proposed system at recognizing BG/FG characteristic situations, so providing an augmented scene understanding. Finally, due to general overall framework, the system can be easily integrated in video surveillance system, so to build a multimodal surveillance system, potentially more robust and efficient at detecting abnormal events. The above cited ideas constitute the natural progress of the work presented in this paper.

References

[1] A. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, London, 1990.

[2] M. Cowling and R.Sitte. Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters*, 24:2895–2907, 2003.

[3] D. Ellis. Detecting alarm sounds. In *Proceedings of Consistent and Reliable Acoustic Cues for sound analysis (CRAC01)*, Aalborg, Denmark, September 2001.

[4] K. Hild II, D. Erdogmus, and J. Principe. On-line minimum mutual information method for time-varying blind source separation. In *Int. Workshop on Independent Component Analysis and Signal Separation*, pages 126–131, 2001.

[5] S. Marple. *Digital Spectral Analysis*. Prentice-Hall, second edition, 1987.

[6] PAMI. Special issue on video surveillance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8), 2000.

[7] V. Peltonen. Computational auditory scene recognition. Master's thesis, Tampere University of Tech., Finland, 2001.

[8] S. Roweis. One microphone source separation. In *NIPS*, pages 793–799, 2000.

[9] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Int. Conf. Computer Vision and Pattern Recognition*, volume 2, 1999.

[10] T. Zhang and C. Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, 9(4):441–457, 2001.