GENERATIVE PROCESS TRACKING FOR AUDIO ANALYSIS

Regunathan Radhakrishnan and Ajay Divakaran

Mitsubishi Electric Research Laboratory,
Cambridge, MA 02139
E-mail: {regu, ajayd}@merl.com

ABSTRACT

The problem of generative process tracking involves detecting and adapting to changes in the underlying generative process that creates a time series of observations. It has been widely used for visual background modelling to adaptively track the generative process that generates the pixel intensities. In this paper, we extend this idea to audio background modelling and show its applications in surveillance domain. We adaptively learn the parameters of the generative audio background process and detect foreground events. We have tested the effectiveness of the proposed algorithms using synthetic time series data and show its performance on elevator audio surveillance.

1. INTRODUCTION

The problem of generative process tracking involves detecting and adapting to changes in the underlying generative process that creates a time series of observations. This problem has been widely studied for visual background modelling. The intensity observation at each pixel is considered to be generated by a generative process that can be modelled by a Gaussian Mixture Model(GMM). Then, by detecting and adapting to changes in the intensity at that pixel one is able to perform background-foreground segmentation from a sequence of images.

Visual background modelling is a fundamental problem in computer vision and has been an active area of research [1],[2],[3]. The approach in [2] adaptively estimates the probability distribution of pixel intensities using a parametric model such as a mixture of Gaussians. It has been extended to audio analysis in [4]. In [4], Cristani et al propose a method based on the probabilistic modelling of the audio data stream using separate sets of adaptive Gaussian mixture models, working for each sub-band of the spectrum. The main drawback with this approach is that one has to maintain a GMM for each sub-band to detect outlier events in that sub-band and then decide whether it is a foreground event or not. However, it would be easier to detect and interpret foreground events if we can detect them from a multivariate time series of cepstral/spectral features instead of treating each sub-band energy independently and then having another layer for interpretation.

Towards that end, in our previous work, we had detected audio "backgrounds" and "foregrounds" from a time series of cepstral features extracted from stored content [5]. Recall our definition of "background" and "foreground" in time series. The generative process that generates most of the observations in the time series is referred to as a "background". A generative process that generates a burst of observations amidst the observations generated by the "background" is referred to as the "foreground". We have shown that one can detect highlight segments in sports audio and suspicious events in surveillance audio by detecting such "backgrounds" and "foregrounds". However, the computational complexity and the latency of the proposed approach in [5] is high.

In this paper, we present an application where we arrive at the "backgrounds" and the "foregrounds" using only the causal time series and using fewer computational resources thereby making the approach in [5] more practical and realizable. The proposed approach adaptively tracks the background process to detect outliers that cannot be explained by the background model estimate at that time. Such an extension of the visual background modelling to cepstral features extracted from audio, can help us perform background-foreground segmentation for event detection in surveillance audio content.

The rest of the paper is organized as follows. In section 2, we motivate and describe the proposed framework for surveillance audio analysis. In section 3, we present the experimental results on elevator surveillance content. In section 4, we present our conclusions.

2. PROPOSED FRAMEWORK

The proposed framework is motivated by the observation that suspicious events in surveillance audio data happen sparsely as outliers in a background of usual events. For instance, there is a burst of screaming sounds in the vicinity of a suspicious event. Therefore, with the help of an adaptive model of the generative background process, one can detect these

1-4244-0469-X/06/$20.00 ©2006 IEEE

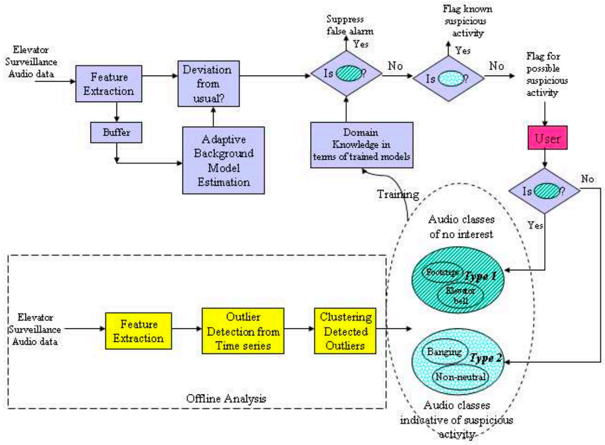
Elevator Surveillance Audio data
Feature Extraction
Deviation from usual?
Buffer
Adaptive Background Model Estimation
Suppress false alarm
Yes
Is ?
No
Flag known suspicious activity
Yes
Is ?
No
Flag for possible suspicious activity
Domain Knowledge in terms of trained models
User
Is ?
No
Yes
Training
Audio classes of no interest
Footsteps
Type 1
Elevator bell
Elevator Surveillance Audio data
Feature Extraction
Outlier Detection from Time series
Clustering Detected Outliers
Banging
Type 2
Non-neutral
Offline Analysis
Audio classes indicative of suspicious activity

Figure 1: Proposed Framework for suspicious event detection using audio analysis

events as foreground processes. In the following, we elaborate on this formally.

2.1. Problem Formulation

Let p_1 represent a realization of the "usual" process ($\mathbf{P_1}$) which can be thought of as the background process. Let p_2 represent a realization of the "unusual" process $\mathbf{P_2}$ which can be thought of as the foreground process. Given any causal time sequence of observations or low-level audio features from the two the classes of events ($\mathbf{P_1}$ and $\mathbf{P_2}$), such as

$$...p_1p_1p_1p_1p_1p_1p_1p_1p_2p_2p_2p_2...$$

then the problem is to find the onsets and times of occurrences of realizations of $\mathbf{P_2}$ without any a priori knowledge about $\mathbf{P_1}$ or $\mathbf{P_2}$.

2.2. Proposed Approach for Adaptive Background Modelling

Given a causal time series of audio features, we estimate the background process P_1 by training a GMM from a window of W_L initial observations,$\{O_1, O_2, ...O_{W_L}\}$. The number of mixtures for this model is obtained by using minimum description length principle. Let us represent this GMM as G_b. Let us denote of the number of mixtures in G_b by K. We use the notation π, μ, and R to denote the parameter sets $\{\pi_k\}_{k=1}^K$, $\{\mu_k\}_{k=1}^K$, and $\{R_k\}_{k=1}^K$, respectively, for mixture coefficients, means, and variances.

Then for every subsequent new observation, O_n, the current estimate of the background model G_b is updated according to the following steps:

- **Step 1:** Initialize a dummy mixture component, $K+1$, with a random mean vector and high variance diagonal covariance and a low mixture probability. Normalize the mixture probability matrix(π) accordingly.
- **Step 2:** Compute the log likelihood of O_n under G_b.
- **Step 3:** If the computed log likelihood in step 2 is greater than a specified threshold, then find the most likely mixture component that generated O_n given by $j = argmax_m(\frac{P(O_n/\{\mu_m,R_m\})\pi_m}{P(O_n/G_b)})$. And, update the parameters of component j according to the following equations for online EM learning [2].

$$\pi_{j,t} = (1-\alpha)\pi_{j,t-1} + \alpha \quad (1)$$

$$\mu_{j,t} = (1-\rho)\mu_{j,t-1} + \rho O_n \quad (2)$$

$$R_{j,t} = (1-\rho)R_{j,t-1} + \rho(O_n-\mu_{j,t})^T(O_n-\mu_{j,t}) \quad (3)$$

where α and ρ are related to the learning rate of the background model.

For other mixture components ($h \neq j$), update only the mixture probabilities as given below

$$\pi_{h,t} = (1-\alpha)\pi_{h,t-1} \quad (4)$$

Finally, normalize the matrix π.

- **Step 4:** If the computed log likelihood in step 2 is lesser than the specified threshold, then it implies that the current background model with its K mixture components cannot explain O_n. Recall that $(K+1)_{th}$ mixture component is a dummy one. Replace the mean of this component with that of O_n and create a dummy component. As a result, we have added a new mixture component to the background model to account for O_n that cannot be explained by the current background model. Also, create a new dummy component for prospective observations from the foreground process in future.
- **Step 5:** For each O_n, record the most likely mixture component from the background that explains it. Then, by examining the pattern of memberships to mixture components of the background model, one can detect the observations from foreground.

There are two differences in the proposed extension for audio when compared to the visual background modelling proposed by Stauffer and Grimson in [2]. First, we do not assume a diagonal covariance for the multivariate cepstral time series for they are not like R, G and B components.

Second, we use the likelihood value of the incoming data under the current model to determine whether it is a foreground point unlike in [2] where every new point is checked against the mixture components in the current model until a match is found. A match is declared when the new point is within 2.5 standard deviations of the component's mean.

There are four parameters to choose in this proposed approach namely W_L (the initial window size for background model estimate), α and ρ (related to learning rate of the background model) and the likelihood threshold to decide whether an observation could have been generated by the current background model.

3. EXPERIMENTAL RESULTS

We test the proposed generative process tracking in the context of an audio surveillance application for elevators. In [6], we proposed a hybrid audio analysis framework for surveillance that combines the results of unsupervised audio analysis with the results of a audio classification framework. The proposed framework for suspicious event detection is shown in Fig. 1.The audio classes for the classification framework are obtained from off-line time series analysis of cepstral features and training. The chosen audio classes can be further grouped into one of the following two types: ***Type 1:***"classes that are **NOT** indicative of suspicious events" and ***Type 2:***"classes that are indicative of suspicious events".

It also adaptively learns a Gaussian Mixture Model(GMM) to model the background sounds and updates the model incrementally as new audio data arrives and has been shown to detect suspicious events effectively. However, the adaptive background modelling algorithm used in [6] first estimates a GMM from a small window (W_S) of incoming observations and then updates the parameters of statistically equivalent components in the background GMM. This has the following two disadvantages: first, the window (W_S) has to be large enough to include enough samples to estimate a GMM from the new data; second, this limits the resolution at which events can be detected. In this paper, we avoid both of these issues by using background modelling approach introduced in the previous section which is updated for every new incoming data vector.

Now, let us look at the how the background model can be used to detect new types of suspicious events. The likelihood of new audio data under the current background model is compared against a threshold to flag an outlier. In the case of an inlier, the background model is updated according to equations 1-3. In the case of an outlier, it is checked to see if it belongs to ***Type 1:*** or ***Type 2:***. ***Type 1:*** implies it is a known false alarm to be suppressed and ***Type 2:*** implies it is a known suspicious event. When encountered with a new type of outlier that has a small likelihood under the models for ***Type 1:*** and ***Type 2:*** classes, end user's feedback is sought to create a model for the new type of outlier.

In the following, we will illustrate the ability of the proposed framework to detect foreground audio events from elevator surveillance content. The elevator audio surveillance data consists of 126 clips (2 hours of content) with suspicious events and 4 clips (40 minutes of content) that are without events. We extract low-level features from 61 clips (1 hour of content) of all the suspicious event clips and 4 clips of normal activity in the elevators (40 minutes of content). The low-level features were 12 MFCC features extracted for 8ms frame of audio. The audio classification framework consists of the following four audio classes, namely, Banging, Footsteps, non-neutral speech, normal speech. Banging and non-neutral speech belong to ***Type 2:*** outlier category whereas the other two belong to ***Type 1:***. We got a 91% classification accuracy on 90/10 cross-validation split of this data.

Now, let us look at the results of online adaptive background modelling on the elevator surveillance data. Figure 2 shows the log-likelihood of every audio frame under the background model estimate at that time for a 8min sequence. The initial background model was learnt from low-level features corresponding to 200s of audio. There are two thresholds involved in the online adaptive background modelling described earlier: one on the value of log-likelihood to determine foreground audio objects (consequently to decide whether or not to update the background model) and the second threshold to determine which component from the current background GMM could have generated the observed feature vector. The value of α and ρ were empirically determined and set to 0.008 and 0.05 respectively for all the clips. Notice from the Figure 2, that there are peaks during onsets of bursts of outlier observations. Also, we listened to the detected outliers and labelled them as one of $\{$F:footsteps, E: elevator door opening/closing sound, C: man coughing, S:normal speech $\}$ as shown in the Figure 2.

Figure 3 shows the most likely mixture component of the current background model that could have generated the incoming audio frame for each of these clips. Notice that adaptation tracks the process change over time.

For every detected onset of outlier burst from the Log-likelihood values (as in Figure 2), one would like to find out whether the detected outlier is of ***Type 1:*** or ***Type 2:***. We use the audio classification framework obtained from off-line analysis, for this purpose. The input to the classification framework is cepstral features corresponding to 1s of audio from the onset of each burst. Table 1 shows the likelihood of the detected outliers under models for the four classes in the classification framework (Banging, Footsteps, normal speech, non-neutral/agitated speech).

Also, from the training data for each of these classes one can learn the mean and variance of the likelihood val-

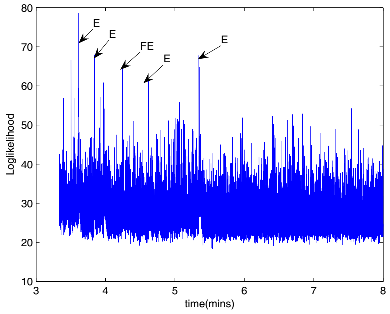

Figure 2: *Log-likelihood from adaptive background model for clip 4*

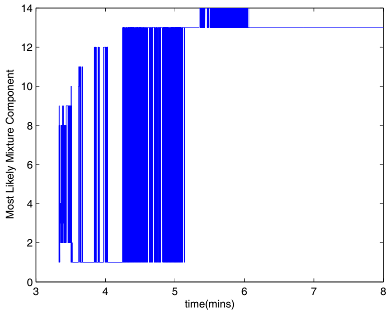

Figure 3: Most likely mixture component of adaptive background model for every frame of clip 4

ues under the trained model for that class. For instance, let us denote the computed mean and variance of the likelihood values of training data for class 1 by $\mu_{\lambda_1}^{lik}$ and $\sigma_{\lambda_1}^{lik}$ respectively. Now, for each of the detected outlier that is being tested with this model, the likelihood value is checked to see if it is within $\mu_{\lambda_1}^{lik} \pm 2.7\sigma_{\lambda_1}^{lik}$. If all the trained models flag the likelihood value to be out of the usual range, then the detected outlier is probably a new audio class. These likelihood values are marked as bold and italicized values in the likelihood table 1. Outliers that correspond to elevator door opening and closing sounds and a man coughing sound fell into the category of outliers for which the likelihood values were out of range. For other outliers, the classification results help us find out if they are of ***Type 1****:* or ***Type 2****:*.

4. CONCLUSION

In this paper, we proposed algorithms for the problem of generative process tracking in audio. It involves detecting and adapting to changes in the underlying generative process that creates a time series of observations. It has been widely used for visual background modelling to adaptively track the generative process that generates the pixel intensities. We have extended this idea to audio background modelling and applied it to surveillance audio analysis. The experimental results show that the proposed extension for generative process tracking is able to adapt to changes in audio background and also detect outliers in those backgrounds. Then, by using an audio classification framework on the detected outliers, we can detect whether it is a false alarm, a known suspicious event or a new audio class.

t(secs)	class	λ_1	λ_2	λ_3	λ_4
211	2	-2084.1	-1982.8	-2098.1	-2112.1
218	2	***-2506.2***	***-2244.5***	***-2795.6***	-2427.7
231	2	-1980.2	-1912.1	-2014.0	***-2026.3***
255	3	-2120.9	-2135.7	-2091.6	-2206.7
321	2	-2117.1	-2062.3	-2180.2	-2190.4
322	2	***-2585.6***	***-2398.5***	***-2894.1***	-2487.7

Table 1: Audio Classification Results on Detected Outliers for clip 4; Likelihood under λ_1: Banging; λ_2: Footsteps; λ_3: normal speech; λ_4: non-neutral;

5. REFERENCES

[1] K.P.Karman and A. von Brandt, “Moving object recognition using an adaptive background memory,” *in Capellini, editor, Time-varying Image Processing and Moving Object Recognition*, pp. 297–307, 1990.

[2] C.Stauffer, and W.E.Grimson, “Learning patterns of activity using real-time tracking,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 747–757, 2000.

[3] D.Harwood A.Elgammal and L.Davis, “Non-parametric model for background subtraction,” *in Proc. ECCV*, 2000.

[4] M.Cristani, M.Bicego and V.Murino, “Online adaptive background modelling for audio surveillance,” *Proc. of ICPR*, 2004.

[5] R.Radhakrishnan, A.Divakaran, Z.Xiong and I.Otsuka, “A content-adaptive analysis & representation framework for audio event discovery from “unscripted” multimedia,” *accepted for publication in Eurasip Journal on Applied Signal Processing, Special Issue on Information Mining from Multimedia*, 2005.

[6] R.Radhakrishnan, A. Divakaran, P. Smaragdis, “Audio analysis for surveillance applications,” *IEEE WASPAA*, Oct 2005.