MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Systematic Acquisition of Audio Classes for Elevator Surveillance

Regunathan Radhakrishnan, Ajay Divakaran, Paris Smaragdis

TR2005-076 March 2005

Abstract

We present a systematic framework for arriving at audio classes for detection of crimes in elevators. We use our time series analysis framework proposed in5 to low-level features extracted from the audio of an elevator surveillance content to perform an inlier/outlier based temporal segmentation. Since suspicious events in elevators are outliers in a background of usual events, such a segmentation help bring out such events without any a priori knowledge. Then, by performing an automatic clustering on the detected outliers, we identify consistent patterns for which we can train supervised detectors. We apply the proposed framework to a colleciton of elevator surveillance audio data to systematically acquire audio classes such as banging, footsteps, non-neutral speech and normal speech etc. Based on the observation that the banging audio class and non-neutral speech class are indicative of suspicous events in the elevator data set, we are able to detect all of the suspicious activities without any misses.

SPIE Image and Video Communications and Processing

Systematic Acquisition of Audio Classes for Elevator Surveillance

Regunathan Radhakrishnan, Ajay Divakaran and Paris Smaragdis
Mitsubishi Electric Research Laboratory,
Cambridge, MA 02139
E-mail: {regu,ajayd,paris}@merl.com

We present a systematic framework for arriving at audio classes for detection of crimes in elevators. We use our time series analysis framework proposed in[5] to low-level features extracted from the audio of an elevator surveillance content to perform an inlier/outlier based temporal segmentation. Since suspicious events in elevators are outliers in a background of usual events, such a segmentation help bring out such events without any a priori knowledge. Then, by performing an automatic clustering on the detected outliers, we identify consistent patterns for which we can train supervised detectors. We apply the proposed framework to a collection of elevator surveillance audio data to systematically acquire audio classes such as banging, footsteps, non-neutral speech and normal speech etc. Based on the observation that the banging audio class and non-neutral speech class are indicative of suspicious events in the elevator data set, we are able to detect all of the suspicious activities without any misses.

1 Introduction

Past work on surveillance has mainly focussed on video analysis. Based on an adaptive background model, foreground video objects are segmented and tracked in the input video sequence. Then, based on trajectory features a model is learnt to characterize the "usual" behavior of objects in the scene. A suspicious event is flagged in case an object's behavior deviates from the learnt "usual" model,[2] ,[3] .[1]

In this paper, we focus on a framework based on audio classification for surveillance. Audio analysis would take us closer to semantics than video analysis would and also is computationally more efficient. This motivates us to detect events in surveillance based on a audio classification framework that classifies every time segment of audio into one of a set of trained audio classes. Since certain sound classes are indicative of suspicious events, such an audio classification framework can be used to detect suspicious events. Therefore, in this framework, the choice of audio classes is crucial. However, the typical large volume of audio data for surveillance and lack of domain knowledge rule out the option of selecting these audio classes based on intuition. There is a definite need for a framework to systematically select these audio classes to characterize the domain to be able to detect events successfully.

We propose one such framework to systematically acquire domain knowledge to arrive at the audio classes that would characterize the different sounds in a given domain. We treat the low-level audio features as a time series and use the time series analysis framework proposed in[5] to perform an inlier/outlier based temporal segmentation of the audio content. The analysis framework in[5] models suspicious events as "unusual" events in a background of "usual" events. Then, by performing an automatic clustering on the detected outliers, we identify consistent patterns for which we can train supervised detectors. We apply the proposed framework to a collection of elevator surveillance audio data to systematically acquire audio classes such as banging, footsteps, non-neutral speech and normal speech etc. Based on the observation that the banging audio class and non-neutral speech class are indicative of suspicious events in the elevator data set, we are able to detect all of the suspicious activities without any misses.

The rest of the paper is organized as follows. In section 2, we provide a brief introduction to the time series analysis framework for inlier/outlier based temporal segmentation. In section 3, we present our experimental results on elevator surveillance data. In section 4, we finally conclude with some discussions.

2 [illegible]

2.1 [illegible]

In this section, we present a brief description of the time series clustering framework proposed in.[5] The proposed framework is motivated by the observation that "interesting" events in multimedia happen sparsely in a background of usual or "uninteresting" events. Some examples of such events are:

- [illegible]: A burst of overwhelming audience reaction following a highlight event in a background of commentator's speech.

  [illegible]: A burst of screaming noise following a suspicious event in a silent or static background.

This motivates us to formulate the problem of discovering "interesting" events in multimedia as that of detecting outlier subsequences or "unusual" events by statistical modelling of a stationary background process in terms of low/mid-level audio-visual features. Note that the background process may be stationary only for small period of time and can change over time. This implies that background modelling has to be performed adaptively throughout the content. It also implies that it may be sufficient to deal with one background process at a time and detect outliers. In the following subsection, we elaborate on this more formally.

2.2 [illegible]

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

$$...p_1p_1p_1p_1p_1p_2p_2p_1p_1p_1...$$

then the problem of outlier subsequence detection is that of finding the times of occurrences of realizations of $\mathbf{P_2}$ without any a priori knowledge about $\mathbf{P_1}$ or $\mathbf{P_2}$.

To begin with, the statistics of the class $\mathbf{P_1}$ are assumed to be stationary. However, there is no assumption about the class $\mathbf{P_2}$. The class $\mathbf{P_2}$ can even be a collection of a diverse set of random processes. The only requirement is that the number of occurrences of $\mathbf{P_2}$ is relatively rare compared to the number of occurrences of the dominant class. Note that this formulation is a special case of a more general problem namely clustering of a time series in which a single highly dominant process does not necessarily exist. We treat the sequence of low/mid level audio-visual features extracted from the video as a time series and perform a temporal segmentation to detect transition points and outliers from a sequence of observations.

2.3 [illegible]

Given the problem of detecting times of occurrences of $\mathbf{P_1}$ & $\mathbf{P_2}$ from a time series of observations from $\mathbf{P_1}$ and $\mathbf{P_2}$, we propose the following time series clustering framework:

1. Sample the input time series on a uniform grid. Let each time series sample at index 'i' (consisting of a sequence of observations) be referred to as a context, C_i.
2. Compute a statistical model M_i from the time series observations within each C_i.
3. Compute the affinity matrix for the whole time series using the context models and a commutative distance metric ($d(i,j)$) defined between two context models (M_i & M_j). Each element, A(i,j), in the affinity matrix is $e^{\frac{-d(i,j)}{2\sigma^2}}$ where σ is a parameter that controls how quickly affinity falls off as distance increases .

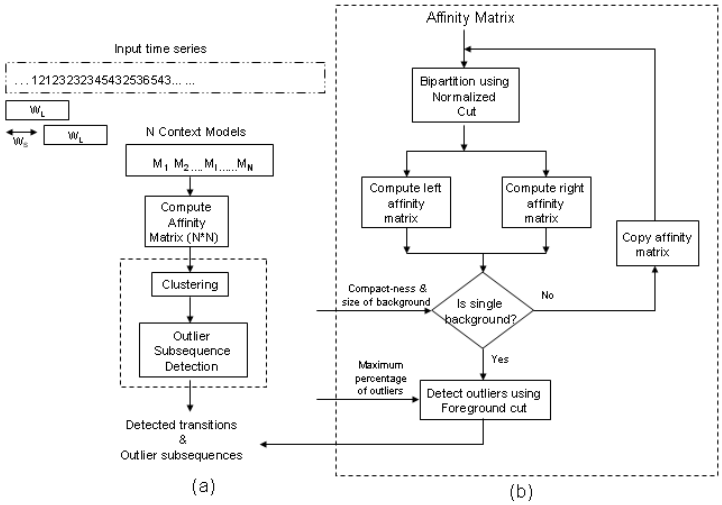

Figure 1. Proposed outlier subsequence detection framework

4. The computed affinity matrix represents an undirected graph where each node is a context model and each edge is weighted by the similarity between the nodes connected by it. Then, we can use a combination of the normalized cut solution and the modified normalized cut solution to identify distinct clusters of context models & "outliers context models" that do not belong to any of the clusters. Note that the second generalized eigenvector of the computed affinity matrix is an approximation to the cluster indicator vector for bi-partitioning the input graph, as discussed in section **?** .

Figure 1 illustrates the proposed framework. The portion of the figure (b) is a detailed illustration of the two blocks: (clustering & outlier detection) in figure (a).

For more details, please see.[5] In the next subsection, we describe our framework for systematic acquisition of key audio classes.

[redacted]

The proposed time series analysis framework can be used for systematic acquisition of key audio classes as shown in Figure Figure 2.

The proposed framework not only gives an inlier/outlier based temporal segmentation of the content but also distinguishable sound classes for the chosen low-level features in terms of distinct backgrounds and outlier sound classes. Then, by examining individual clusters from the detected outliers one can identify consistent patterns in the data that correspond to the events of interest and build supervised statistical learning models.

In the following section, we use this framework for selecting the sound classes to characterize the elevator surveillance audio data and achieve accurate detection of notable events.

3 [redacted]

In this section, we apply the proposed analysis to a collection of elevator surveillance audio data for systematic acquisition of key audio classes for this domain. The data set contains recordings of suspicious activities in elevators as well as some event free clips. Since most of the suspicious events are outliers in a background of usual events, we are motivated to apply the proposed outlier subsequence detection framework for the task of inlier/outlier based segmentation of the surveillance content. Then, by examining the detected outliers from the suspicious clips we can systematically select key audio classes that are indicative of suspicious events. By

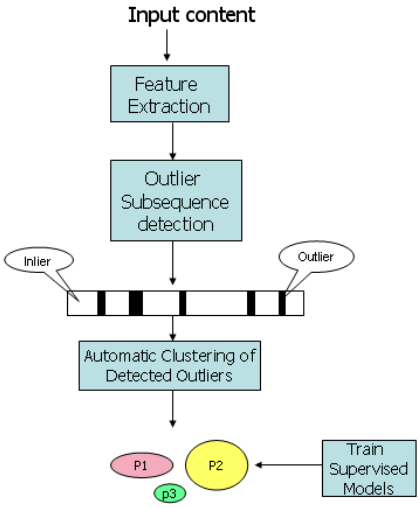

Figure 2. Systematic acquisition of domain knowledge using the inlier/outlier based representation framework

examining inliers and outliers from the segmentation of event free clips, we can identify sound classes that characterize usual events in the elevator.

The elevator surveillance audio 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). Then, for each of the clips we perform inlier/outlier based segmentation with the proposed framework to detect outlier subsequences. A subsequent clustering of the detected outliers will help us discover key audio classes that correlate with the event of interest.

The parameters of the proposed framework were set as given below:

- Context window size(W_L) = 4 sec
- Step size(W_S) = 2 sec
- Frame rate at which MFCC features are extracted = 125 frames per second
- Maximum percentage of outliers = 30%
- Compactness constraint on the background = 0.5
- Relative time span constraint on the background = 0.35 and the context model is a 2 component GMM

A 2 component GMM was used to model the PDF of the low-level audio features in the 4s context window. Figure 3(A)-(D) shows an example hierarchical segmentation result of an event-free surveillance audio clip using the proposed time series analysis framework. The first normalized cut solution of the affinity matrix 3(B) is shown in Figure 3(A). It was verified that the one of the clusters corresponds to time segments during which the elevator was not used. The other cluster corresponds to segments during which the elevator was in use. Further splitting this cluster as shown in Figures 3(C) through 3(E) finds the following three types of outliers:

- when there are sounds of footsteps.
- when there are screeching sounds due to elevator door opening or closing.
- when there are conversations within the elevator.

Label	Sound Class
1	Alarm
2	Banging
3	Elevator Background
4	Door Opening & Closing
5	Elevator Bell
6	Footsteps
7	Non-neutral Speech
8	Normal speech

Table 1. Systematically acquired sound classes from detected outliers and inliers after analysis of surveillance content from elevators

These outliers are consistent patterns detected by applying the analysis framework to event-free clips. This implies that we can characterize the usual elevator activity by training a supervised detector for the following four sound classes: elevator background (i.e inliers from event free clips), footsteps, normal speech within the elevator and elevator door opening and closing sounds.

In order to characterize the sounds during suspicious events in elevators, we carry out a similar temporal segmentation for each of the 61 clips. The outliers in this case turned out to be from the following two categories: banging sounds against elevator walls and non-neutral (excited) speech. In all the clips with suspicious activity, the outliers consistently turned out to be clips of banging sounds against elevator walls and excited speech.

Finally, by performing a simple clustering operation on the detected outliers from the analysis of event-free clips and suspicious clips, we collected training data for the following sound classes to characterize the complete surveillance audio data.

Since these sound classes were obtained as distinguishable sounds from the data, we already know what features and supervised learning method are to be used for modelling them. Hence, we use a Minimum Description Length GMM to model the distribution of low-level features. The learned models were used to classify every second of audio from the remaining clips with suspicious activity in elevators. The classification system correctly classified banging sounds and non-neutral speech in all of the suspicious clips thereby enabling detection of these events. Figure 4 shows classification results for every second of audio in each of the four types of suspicious activity in elevators. Class labels that are indicative of suspicious activity (banging sounds and non-neutral speech) have labels 2 and 7.

4.

In this paper, we have presented a systematic way to choose audio classes for event detection in surveillance. We model the suspicious events as "unusual events" in a "usual" background. We treat the low-level audio features as a time series and perform an inlier/outliet based temporal segmentation of the content. Then, by performing automatic clustering on the detected outliers we collected training data for consistent patterns. We used a MDL-GMM to model the distribution of features of each of the chosen audio class. Of the chosen audio classes, we found that the sound class for banging sound and the sound class of non-neutral speech are indicative of suspicious activity in elevators. Our audio classification framework detects all of the suspicious activity without any false alarms.

One drawback of the proposed audio classification based surveillance is that it would only be able to detect known kinds of suspicious activity (e.g the ones that are accompanied by banging sounds and screaming). However, a real surveillance system should also be capable of detecting unusual events that the system has not seen before. The proposed time series analysis framework can be adapted to perform online event detection by computing a adaptive model of "usual" background sounds and flagging deviations from that. In order to keep the false alarms under control, currently we are working on a hybrid approach that combines the results of audio classification with results of low-level time series analysis.[4]

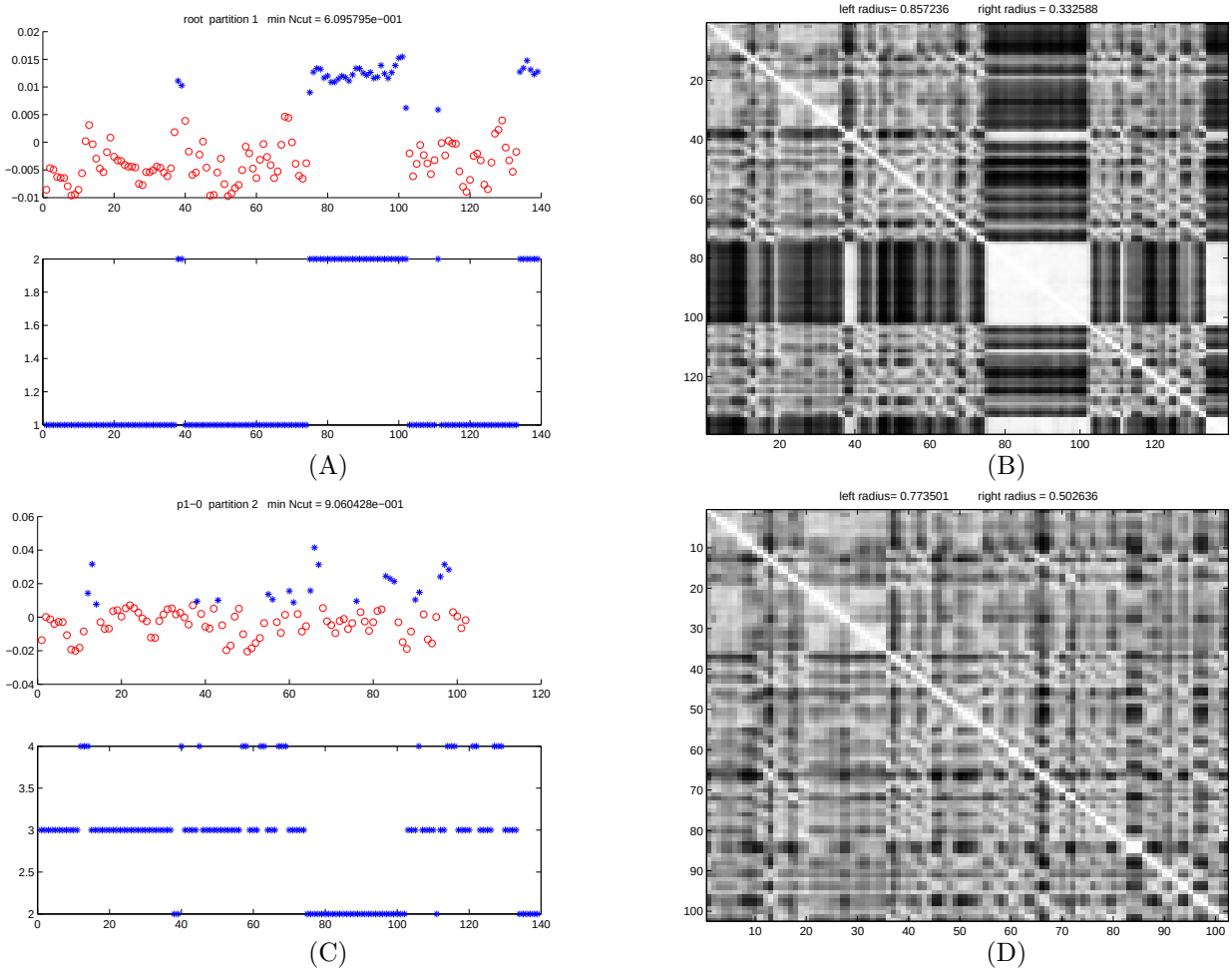

Figure 3. Example Hierarchical Segmentation for a event-free clip in elevator surveillance content

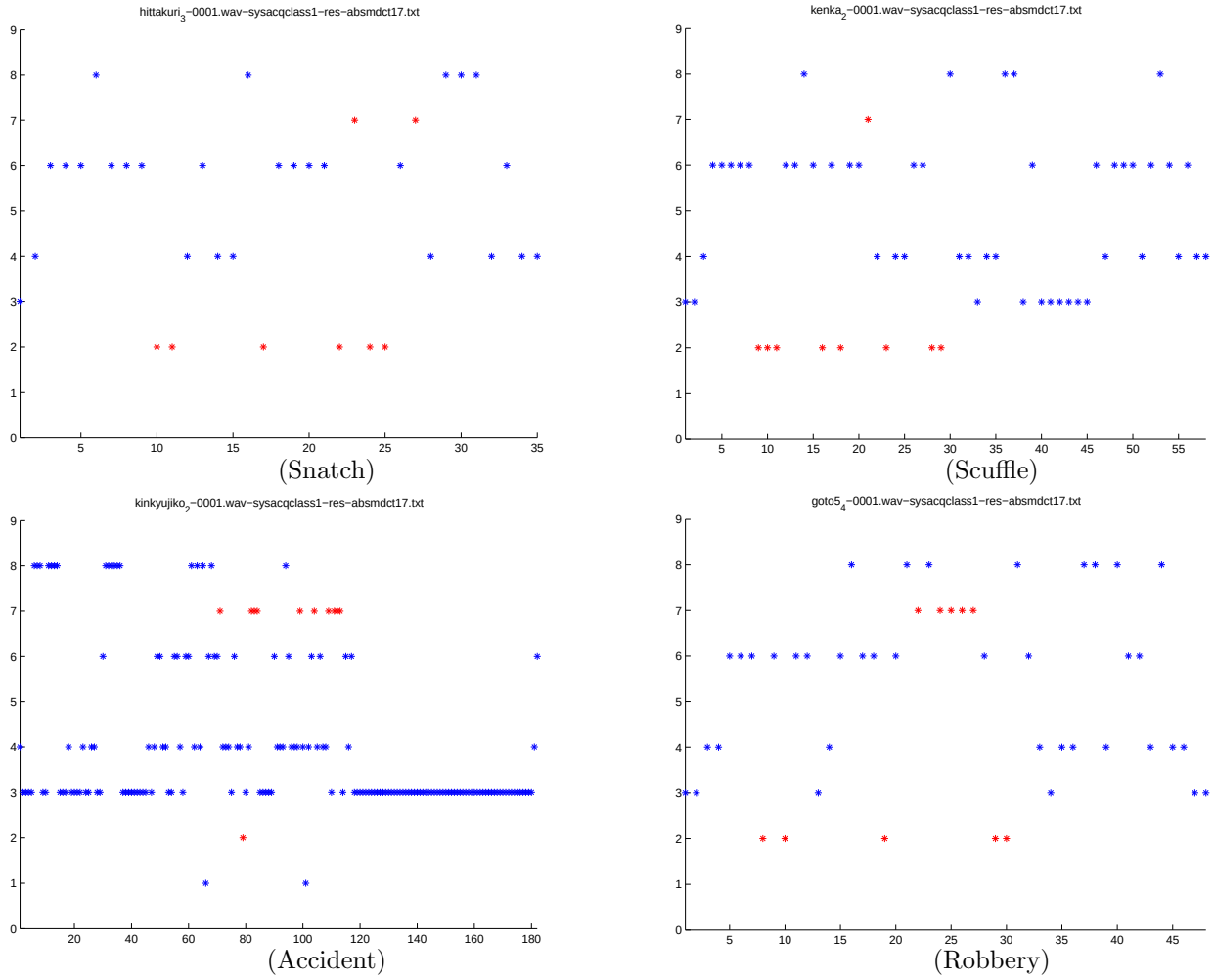

Figure 4. Classification results on test clips with suspicious activity

1. C.Stauffer, and W.E.Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2000), 747–757.
2. G. Wu, Y. Wu, L. Jiao, Y.-F. Wang, and E. Chang. Multi-camera spatio-temporal fusion and biased sequence-data learning for security surveillance. *ACM Multimedia* (2003).
3. Porikli, F.M. and Haga, T. Event detection by eigenvector decomposition using object and frame features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (June 2004).
4. R.Radhakrishnan. A content-adaptive analysis & representation framework for video summarization using audio cues. *http://isis.poly.edu/ regu/ReguThesis.pdf* (Dec. 2004).
5. R.Radhakrishnan, A.Divakaran, Z.Xiong and I.Otsuka. A content-adaptive analysis & representation framework for audio event discovery from "unscripted" multimedia. *submitted to Eurasip Journal on Applied Signal Processing* (2005).