PERSISTENT AUDIO MODELLING FOR BACKGROUND DETERMINATION

Simon Moncrieff, Svetha Venkatesh, Geoff West

Department of Computing
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia

ABSTRACT

This paper is concerned with modelling background audio online to detect foreground sounds in complex audio environments for surveillance and smart home applications. We examine and expand upon previous work in the audio and video domains, and propose a new implementation of an audio background modelling algorithm, addressing the complexities of audio data. A number of audio features characterising different aspects of the audio content were analysed to determine the factors relevant to the determination of the background audio. We test the algorithms on three audio data sets of varying complexity. The new approach was successful in modelling the background audio for the test data.

1. INTRODUCTION

A common initial phase in visual tracking applications is the modelling of the background of an image in order to detect the foreground, i.e moving objects. Higher level analysis then focuses on the sections of the image that are of the most interest. Consequently, a logical initial phase in applying audio analysis to surveillance and monitoring applications is the detection of background audio. This would serve to highlight sections of interest in an audio signal. We define background audio to be recurring and persistent audio characteristics that dominate a portion of the signal.

In the visual domain, the intensity values of pixels are modelled over time to determine the background. A basic approach assumes the background is modelled with a single distribution [1]. The adaptive online Gaussian Mixture model (GMM) method [2] expands upon this and is a popular method for visual background detection. The use of multiple distributions and the adaption of the GMM allows multiple background models for a single pixel, and the adaption to changes in the background over time. Cristani *et al.* [3] implemented a version of the GMM method in the audio domain. The authors used eight audio features calculated over $1s$ segments to characterise the audio signal. The background for each feature was then modelled independently using the GMM technique. A classification of foreground for one or more features of an audio segment resulted in the segment being classed as foreground. The algorithm was tested on three short audio clips, of $10s$, $12s$ and $30s$ duration, and successfully detected changes in the audio content.

There are a number of differences between the visual and audio domains with respect to the data. The reduced amount of data in audio results in lower processing overheads, and facilitates a more complex computational approach to analysis. The characteristics of the audio commonly exhibit a higher degree of variability. This is due both to the process by which audio is generated, and the superimposition of multiple audio sources within a single input signal. The result is the formation of complex and dynamic backgrounds. There are potentially two approaches to processing the audio data. Initially, audio features are calculated over segments of the incoming audio signal. A $1D$ GMM can then be used to model the background behaviour for each feature. This method in essence treats each feature as the equivalent of a pixel. Alternatively, the features for the segment can be combined into a single feature vector. A single, multidimensional, GMM is then used to model the background of the audio signal, which is processed as the equivalent of a pixel, or single source of data.

In this paper, we extend the experimental scope of audio background modelling, examining the application of the algorithm [2] to the aural domain. We investigate a $1D$ implementation of the audio background detection algorithm [3], and adapt it for application to longer duration data. We propose an alternative approach based on a multidimensional GMM, which addresses two disadvantages of the $1D$ approach. There is no assumption of independence between the features, and a unified method is used to determine foreground classification. We examine the performance of each implementation, using a number of audio feature sets, over a number of different data sets. The feature sets represent different aspects of the audio signal, and the data sets correspond to audio signals with varying background properties and complexities.

The layout of the paper is as follows. Section 2 outlines the background modelling algorithm, and section 3 de-

0-7803-9332-5/05/$20.00 ©2005 IEEE

scribes the experimental methodology and the experimentation. This is followed by the conclusion.

2. ONLINE ADAPTIVE MODELLING

2.1. Background

Stauffer *et al.* [2] use a GMM to model the background for each pixel, or feature vector. The recent history of a feature vector is modelled by a mixture of K Gaussian distributions. The probability of observing the current feature vector X_t, is

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \quad (1)$$

The weight of each model, $w_{i,t}$, is related to the proportion of recently observed feature vectors accounted for by model i, and η is the Gaussian probability density function

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t-\mu)^T\Sigma^{-1}(X_t-\mu)} \quad (2)$$

Each new observation, X_t, is associated with a model within the GMM using online k-means approximation. The observed X_t is compared to each model in the GMM. A model is considered to represent X_t if it is within n standard deviations of the model. The highest ranking model that represents X_t is selected as the matching model. The models in the mixture are ranked in descending order according to w_i/σ_i. If no match is determined for X_t, the lowest ranked model is replaced by a new model with $\mu_t = X_t$, a high initial variance, and a low initial weight. Upon determining a matching model for X_t, the GMM is updated. The weights for the K distributions at time t, are

$$\omega_{k,t} = (1 - \alpha*)\omega_{k,t-1} + \alpha(M_{k,t}) \quad (3)$$

where $\omega_{k,t}$ is the weight of the k^{th} model at time t, and $M_{k,t}$ is 1 for the matched model, and 0 otherwise. The weights are subsequently normalised. The Gaussian distribution parameters for the matched model are updated;

$$\mu_t = (1 - \rho)\mu_{t-1} + \rho X_t \quad (4)$$

$$\sigma_t^2 = (1 - \rho)\sigma_{t-1}^2 + \rho(X_t - \mu_t)^T(X_t - \mu_t) \quad (5)$$

$$\text{where} \qquad \rho = \alpha * \eta(X_t, \mu_k, \Sigma_k) \quad (6)$$

The first B distributions are chosen as the background, according to equation 7. The weights are summed from the highest ranked model to model b. A higher threshold T_v, means more distributions are regarded as background.

$$B = argmin_b(\sum_{k=1}^{b} \omega_k > T_v) \quad (7)$$

Cristani *et al.* [3] implemented the above approach using a single $1D$ GMM to model the background for each feature. The main difference was the foreground determination method. A model was considered to be foreground if $\sum_{k=1}^{k_{hit}} \omega_k > P$, where 1 was the highest ranked model, k_{hit} was the index of the matched model for observation X_t, and P is a threshold analogous to T_v.

Two problems were encountered when the method was applied to longer data sets. Problem 1: the determination of BG/FG contained the implicit assumption that the background does not dominate the audio. If the portion of background is greater than P, sections of background audio can be misclassified as foreground. Problem 2: This occurred due to the update method in equation 6. An increase in Σ resulted in a decrease in η (eqn. 2), irrespective of the relationship between the observed X_t and the mean of the Gaussian model. This has an adverse effect on the adaption of the model.

2.2. Audio Background Modelling

In this section we describe our implementation the background modelling algorithm. A number of changes were necessary in order to apply the technique to audio data.

In the visual domain, Stauffer *et al.* [2] assume σ decreases as the model is updated over time. As in the visual case, a decreasing σ increases the rank of a model, while an increased σ decreases the rank. This provides a form of constraint on the model. However, σ is not used in the determination of FG/BG as there is no assumption that the initial σ decreases as support for a model increases due to the variability of audio data. Consequently, the models are ranked solely by weight to determine foreground classification. Equation 7 was adapted to determine foreground

$$FG = \sum_{k=K}^{k_{hit}} \omega_k < T \quad (8)$$

where K corresponds to the model of lowest rank. In this case, $T = 1 - T_v$, where T_v is the threshold in equation 7.

The model update was determined as follows. The weights for the K distributions at time t, are

$$\omega_{k,t} = (1 - \alpha_\omega * M_{k,t})\omega_{k,t-1} + \alpha\omega(M_{k,t}) \quad (9)$$

The normalisation process is used to decrease the weights of the unmatched models. This represents a more passive update method, limiting the effect of noisy audio as the weight of the background model is penalised to a lesser extent by the occurrence of a non-background model. Each element of the Gaussian distribution Σ is updated as follows

$$\Sigma_t^{i,j} = (1 - \rho)\Sigma_{t-1}^{i,j} + \rho(X_t^i X_t^j) \quad (10)$$

where $\Sigma_t^{i,j}$ is the (i, j) element of the covariance matrix, and X_t^n is the n^{th} element of the currently observed feature vector X_t, and the value of ρ is

$$\rho = \alpha_g e^{-\frac{1}{2}\frac{1}{d}(X_t-\mu_{t-1})^T\Sigma_{t-1}^{-1}(X_t-\mu_{t-1})} \quad (11)$$

where d is the dimension of X_t. This value of ρ scales to a maximum value of 1, for $X_t = \mu_{t-1}$, with a rate of decay of $e^{-\frac{1}{2}x^2}$ where x is the mean number of standard deviations of X_t from μ_{t-1}. The factor d accounts for the dimensions of X_t in determining the mean of the standard deviations. This method has the advantage of decreasing the influence of the update of the existing distribution parameters the further X_t lies on the distribution from μ_{t-1} (outliers), while

Description	Symbol	Value(s)
Clip resolution	t	0.25s, 1s
Probability	n	2.0, 2.5, 3.0, 3.5
Gaussian update	α_g	0.05, 0.01, 0.005, 0.001
Weight update	α_ω	0.01
No. Distributions	K	10

Table 1. Parameters

not being reliant on the η of the model. This addresses the problem of applying equation 6 to audio (problem 2 outlined in section 2.1).

The parameters α_ω and α_g are determined independently. The value α_ω enables the determination of the rate of adaptation of a model to the background. This is necessary when analysing audio at different resolutions. Determining α_g independently allows a more appropriate model parameter update rate to be set, dependent on the data.

The $1D$ GMM method implemented the above approach, modelling each feature independently, with the exception of the update method. An update of $\rho = \alpha$ was used, which increased the generality of the models, increasing accuracy as the more general models clustered a larger proportion of the data. The multidimensional GMM version of the algorithm was implemented as above, using a single GMM with a multidimensional feature vector as input.

3. EXPERIMENTATION

Our main experimental aim was to determine the performance accuracy of the background detection algorithm on a variety of audio environments.

3.1. Experimental Process

For each data set, the following process was used. The data set was divided into audio clips, or windows, of ts in duration. For each clip, a set of d audio features were calculated to characterise the audio content. For the $1D$ case, d $1D$ GMMs were used. Each GMM modelled the background for a single audio feature. For the multidimensional case, one dD GMM was used to model the background of the audio signal. Two parameters determine the behaviour of the background modelling with respect to the GMMs. The number of standard deviations used to determine the matching between a model and an observation n, and the model update parameter α_g. The algorithm was then used to classify each audio clip in sequence. The parameter T was used to determine background classification. A lower value of $T = 0.5$ was used, which enabled multiple models to be classed as background (equation 8).

Table 1 details the values for each of the parameters used in the experimentation. Where multiple values of each parameter were tested, one parameter value was varied per test implementation for each data set over all feature sets.

3.2. Evaluation of Results

The BG/FG classification result for each clip was then compared with the ground truth. The accuracy of the detection of the background clips was calculated according to

$$BG_{acc} = \frac{TP_{BG}}{N_c - FG_D} \tag{12}$$

where TP_{BG} is the number of background clips classified as background, N_c is the total number of clips, and FG_D is the total number of foreground clips correctly detected. A foreground event was considered to have been detected if one or more clips were classified as foreground within the duration of a ground truth foreground event.

A further factor in determining performance is the failure of a GMM to detect a change in background. When the audio characteristics of a new background are sufficiently similar, given the parameters of the GMM, to the previous background model, the model adapts, and no foreground detection occurs. We term the adaption of a model across backgrounds a *morphing model*.

3.3. Data

Continuous, unedited audio streams ($44.1kHz$, $16bit$, mono, wave format) were used as test data. Foreground events were recorded in the presence of the background audio at the time of capturing.

Three data sets of differing levels of audio complexity were used for analysis. We term background audio that emanates from a single source *simple background*. A *complex background* consists of audio from multiple sources. The lab data, 10.6 minutes in length, consisted of a simple background. The traffic data, 12.1 minutes long, consisted of a complex background of traffic noises from a busy road. In processing, the majority of the audio was considered to be accounted for by the background. The kitchen data, 19.9 minutes long, consisted of multiple backgrounds, both simple and complex, recorded in a kitchen environment. This data consisted of 3 background types, with multiple foreground events ($55s$ in total).

The ground truth for the data sets was defined in terms of the foreground events. Foreground audio was considered to be short duration events that were meaningful in the context of the surrounding audio. The remaining audio was classed as background. The results of this analysis method provide a more accurate indication of real world performance, and usefulness, of the algorithm in detecting the background.

3.4. Audio Feature Set

A number of feature sets were used to encapsulate the characteristics of the audio signal content.

1. *WE* - The mean wavelet energy for 7 frequency subbands.
2. *Xtnd* - The *WE* set with 3 frequency domain features.
3. *RFA* - Predominantly frequency based features determined using an attribute selection method.
4. *RFAL* - *RFA* and the audio amplitude.

To determine appropriate audio features for sets 2 and 3, background audio was extracted from the traffic and lab data sets. A large number of temporal and frequency domain

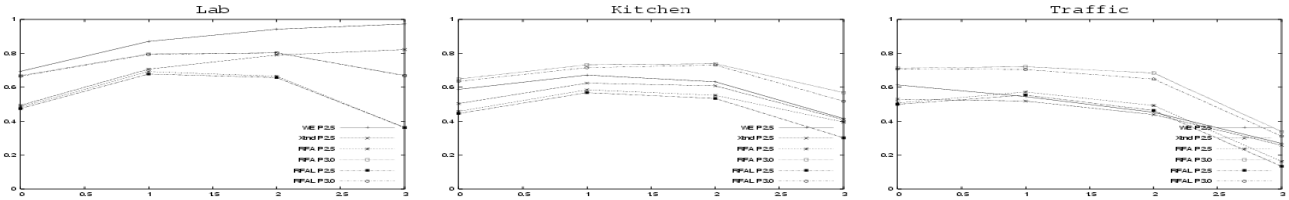

Fig. 1. Background accuracy for $1D$ implementation for the lab, kitchen, and traffic data sets ($1s$ duration).

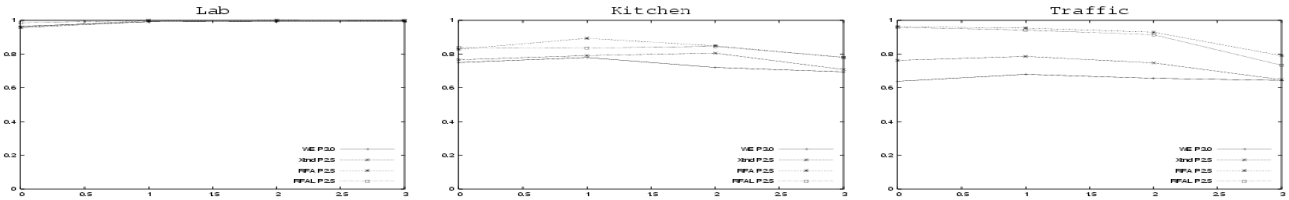

Fig. 2. Background accuracy for nD implementation for the lab, kitchen, and traffic data sets ($1s$ duration).

features were then calculated. The frequency features for set 2 were then selected by choosing the features that exhibited the lowest variance with respect to the scale of the feature. For set 3, an attribute selection method [4] was used.

3.5. Results

The results indicated that the $1D$ GMM method required more generality, both in terms of features and the updating of the models. The higher n and the feature set with the highest variance *WE*, displayed the best overall performance. Figure 1 shows the BG_{Acc} for the best n, for which morphing rarely occurred, across all data sets. The results are graphed according to decreasing α_g from left to right.

The results for BG_{Acc} for the nD model are displayed in figure 2. For the lab data, no foreground events were incorrectly classified, while some short events were not detected in the kitchen data. The *WE* feature sets were more sensitive to the foreground events. The use of a clip duration of $0.25s$ resulted in no foreground false negatives in conjunction with a background detection accuracy that was comparable to the results obtained using a clip size of $1s$. However, due to the increase in the number of clips, the total number of erroneously labelled clips was higher. For a clip size of $1s$, a slower update rate resulted in better performance, accounting for the morphing of models, while the opposite occurred for the $0.25s$ clips. This is in keeping with the theory that higher values of α_g are required for more rapidly changing data. The increase in clip size results in an increased smoothing of the data.

3.6. Comparison of Dimensionality

The results suggest that, in contrast to the $1D$ approach, a more constrained update method and feature variance produce better performance when applied to a multidimensional GMM. This contrast was most noticeable in the performance of the *WE* feature set. The $1D$ GMM was more sensitive to the parameter sets compared with the multidimensional implementation. Overall, the multi-dimensional model achieved a higher background detection accuracy.

For the $1D$ model, a higher value of n produced a higher detection accuracy. This was offset by the more general models resulting in a morphing model for the kitchen data. While a similar problem affects the multi-dimensional method, the background detection accuracy at lower values of n was sufficiently high. The morphing model for the kitchen data is attributed to the superimposition of two background types prior to the change in background.

4. CONCLUSION

We explored a multidimensional GMM implementation of the background detection model, with no assumption of independence. A number of adjustments were made to the original visual background determination method [2] to account for the shift from visual to audio data. We also examined and evaluated a $1D$ GMM implementation of the algorithm. The performance of the algorithms was tested over three data sets, consisting of simple, complex, and multiple background sequences, with the presence of foreground events. A number of different audio feature sets were examined to determine a robust method for encapsulating the properties of the audio for determining the background. The results show that the multidimensional method was more robust in modelling the background for the test data.

5. REFERENCES

[1] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland, "Pfinder: Real-time tracking of the human body," *PAMI*, vol. 19, no. 7, pp. 780–785, 1997.

[2] C. Stauffer and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, Fort Collins, CO USA, 1999, vol. 2, pp. 246–252.

[3] M. Cristani, M. Bicego, and V. Murino, "On-line adaptive background modelling for audio surveillance," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, vol. 2, pp. 399–402.

[4] Ian H. Witten and Eibe Frank, *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, Morgan Kaufmann, 2000.