

Enhancing Video Surveillance with Audio Events

Ruben Gonzalez
NICTA, 300 Adelaide St, Brisbane, Australia
and
Institute for Intelligent Integrated Systems
Griffith University, PMB 50, Gold Coast Mail Centre, QLD, 4217
R.Gonzalez@griffith.edu.au
and

Abstract

Video surveillance systems are only as good as their ability to capture events of interest. The automatic detection of acoustic events of interest coupled with steerable cameras greatly increases the ability of surveillance cameras to capture relevant data. This paper describes an approach for retrofitting this capability to existing surveillance camera networks.

1. Introduction

Video surveillance is receiving renewed attention in the currently political climate. While new technology is being developed in this area there are already many existing systems that have been deployed for a number of years. Many installations consist of dozens if not more fixed or steerable cameras. The operation of many multi-camera sites involves views from cameras being displayed in succession on one or more monitors. Generally only a subset of all available cameras are visible on the display at any one time, and if an event occurs in the vicinity of a camera not being monitored it may go undetected at the time. Hence the main benefit of such surveillance systems is to collect forensic data for analysis after an event rather than detection of events and their intervention as they occur [1].

In the situation where proactive monitoring is required to be able to intervene some kind of event detection mechanism is required in the system to alert operators. A number of efforts into detecting physical events though image and analysis have been developed [2]. These approaches however are only as good as their ability to capture events of

interest which in these cases is limited to the cameras' field of view. In the case where multiple cameras are used and the event occurs within the field of view of one of the cameras that camera can be immediately monitored or some other action taken. If the event does not occur within the field of view of any camera then it will not be detected and cannot be monitored. This is the same whether the cameras are fixed or steerable.

In this paper we discuss the development of an audio-based event detection system that can be retrofitted to existing systems. Being audio based it is not limited to detecting events that fall within the cameras' field of view. An added benefit is that if the cameras' are steerable triangulation can be used to direct the cameras field of view to include the location of the source of the acoustic event.

2. System Design

The system consists of set of remote acoustic sensors connected to a centralised processor which communicates with an existing camera network to select the camera closest to and pointing in the location of the event in the case of fixed cameras and in the case of the steerable cameras to select the closest camera and pan and or tilt it to point in the direction of the location of the event. The centralized processing of the audio is due to the need interface to centralized camera control units.

In order to easily retrofit existing surveillance systems where wiring may have been preinstalled a wireless sensor network solution is proposed. The ZigBee specification is based around the IEEE 802.15.4 standard for wireless personal area networks (WPANs). It is designed for secure, low power secure operation at data rates up to 240 kbps and with self-forming and self-healing mesh

topologies with up to 65,000 devices. The typical range of a single radio is about 100m outdoors the network size can be much greater as ZigBee nodes can act as transponders. The sensor network consists of a ZigBee coordinator (ZC) device to form the root of the network tree that will interface with the surveillance system. ZigBee coordinator (ZC) devices are used as the sensor nodes and can also act as intermediate routers to pass information from other devices. Alternatively for very small networks ZCs could be replaced with lower cost ZigBee End Devices (ZED) that have the limitation that they cannot relay data from other ZigBee devices.

We chose the XBeePro modules from MaxStream (Figure 1) for our initial implementation. These modules provide out-door line of sight range of 1500 metres and a maximum serial interface of 115200 bps. The XBeePro provides an integrated 10 bit ADC with 6 direct analogue inputs, however the maximum sampling rate on any input is only 1kHz. This sample rate is insufficient for sampling the environmental sounds and so an embedded processor is required with the XBeePro module to capture the audio, perform some local processing and pass it over to the XbeePro module via its UART interface.



Figure 1 XBee Module from MaxStream

The microprocessor requirement was for a low power and simple to interface device with integrated high-speed UART and 8 bit ADC supporting 16kHz sample rates. The PIC 16F688 is a 20MHz, 8 bit microprocessor, with an 8 channel, 10 bit ADC. It has a 4096 word program memory, with 512 bytes of data memory shared equally between FLASH memory and EEPROM. This processor requires very few support components. In order to provide the best quality audio for processing the spare processing capacity of the microprocessor is used to implement G.726 ADPCM based compression. This permits 16 bit audio sampling at 16kHz and transmission at 64kbps. As we have spare capacity on the link, while G.726 is designed to operate with 8kHz sampled inputs to produce 24 kbps, 32 kbps or 40 kbps output data rates, we can improve audio

quality by doubling the input rate which has the effect of doubling the output rate. It is possible to trade signal to noise ratio against frequency response by using different input rates and compressed bits per sample configurations as shown in Table 1. We are using the 48 kbps configuration for a 2 sensor node configuration and the 32kbps configuration for 3 sensor node networks. This is due to the limitation for a total maximum interface rate of 115200 bps at the root node.

Table 1 Alternatives

Input Rate	bps	Data Rate
8 kHz	2	24 kbps
	3	32 kbps
	4	40 kbps
16 kHz	2	48 kbps
	3	64 kbps
	4	80 kbps
24 kHz	2	72 kbps
	3	96 kbps
32 kHz	2	96 kbps

The audio is captured using an electret condenser microphone. Dynamic microphones cannot provide sufficient sensitivity to capture audio beyond a 0.5 m range. A second order Butterworth filter is used to provide antialiasing. A block diagram of the components in a sensor node is shown in Figure 2

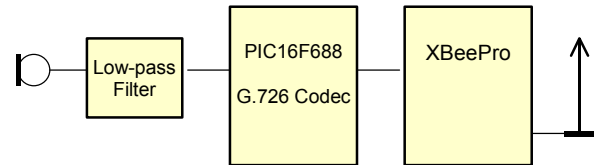


Figure 2. Sensor Node

3. Event Localisation

Deploying a number of acoustic sensor nodes permits localization of acoustic events by creating a sensor array. A number of methods for localisation have been proposed in the literature. The most common method is to use the time differences of arrivals (TDOAs) estimates calculated as the peak of the cross-correlation vectors between 4 microphone pairs. Sensor array based approaches generally require the distance between microphones

to be at least a half of the wavelength of the emitted sound. For narrowband sounds this aperture problem makes trying to localize sources emitting at a higher frequencies difficult. This is less of an issue for broadband sound due to the existence of higher frequency components within the signal that can be used for correlation. This produces the following sensor spacing as in Table 2:

Table 2 Sensor Spacing

Frequency	Sensor Spacing
125 Hz	135 cm
250 Hz	70 cm
500 Hz	35 cm
1 kHz	17 cm
4 kHz	5 cm

With a distributed sensor network, far field and uniform sensor spacing assumptions are not valid for localisation. Ajdler [3] looked at using time differences of arrivals (TDOAs) estimates for microphone pairs, followed by maximum likelihood (ML) estimation for the source position is performed. This was found to be robust to noise and provide high precision in large aperture, sparse arrays. Birchfield [4] also looked at arbitrary microphone configurations but using compact microphone arrays. His approach also used TODAs, but instead of taking the peak of the correlation vectors as estimates for the time delay between the microphones, the correlation vectors are accumulated in a common unit hemisphere coordinate system, centred on the microphone array. Omologo [5] explored three delay estimation methods using microphone arrays; Normalised Cross Correlation (NCC), LMS Adaptive Filters (LMS), and Cross-Power Spectrum Phase (CSP) and found that CSP provided almost twice the accuracy of NCC. Svaizer also [6] used CSP followed by a maximum likelihood approach to derive the sound source coordinates from eight microphones arranged as squares on two orthogonal planes.

In a ZigBee based implementation the use of TODA is straightforward when the network configuration consists of a single root node with a single hop to sensor nodes. In this case the network latency to all sensor nodes is equal. In the case where data must be routed through intermediary nodes, the network latency on routed data will vary and synchronization will be required. Periodically measuring round trip delay to each sensor node and using the result to correctly align the start of audio

frames relative to each other prior to the calculation of localization compensates for this problem. Additionally the microphone array needs to be calibrated when first installed to establish relative sensor locations in the network.

4. Audio Classification

Apart from sound localization, determining the significance of acoustic events is also important. Most sounds are irrelevant and diverting attention to these is counter-productive as focusing on spurious acoustic events may overload the system and cause it to not have the capacity to process a significant event. Filtering events is hence an important requirement for system robustness and is addressed through acoustic event recognition (AER).

AER and the related problem of computational auditory scene recognition (CASR) attempt to separate ambient or background sounds from data that may be of potential interest. This has proved to be a non-trivial process in the general case. Hence much of the research in this area has approached this problem by applying environmental constraints such as limiting the range of all input sounds to only specific well behaved classes or eliminating background sounds altogether by controlling the environment and capture process. For example audio analysis may be limited to only sounds that occur in an elevator [7] or only to gunshot sounds [8], or discriminating a small number of classes such as music genre, speech and noise [9, 10, 11].

The analysis process, which involves class recognition or classification generally, follows the feature extraction. There are a variety of different classification methods that are commonly used including Gaussian Mixture Models (GMM), K-Nearest Neighbour (kNN), Neural Networks (NN), support vector machines (SVM), and Hidden Markov Models (HMM) [10,12,13]. Various experiments have found that in the case of general audio classification the choice of classifier only makes a small difference to the overall performance. Arias [12] compared GMM and SVM to classify four audio classes (speech, music, applause, laughter) using features consisting of 8 Mel-Frequency Cepstral features plus energy and their derivatives. He found that performance was relatively similar with only a 2% difference in classification error. Selina [13] investigated the problem of CASR comparing the effectiveness of SVM, k-NN, and GMM, with a range of spectral features. The task was to correctly determine between one of five different classes; Hallway,

Café, Lobby, Elevator and Sidewalk. Features used included MFCC values plus spectral features (centroid, bandwidth, asymmetry, and spectral), zero-crossing rate, energy range, and frequency roll-off. An accuracy of around 90% was achieved with all classification methods. The SVM was found to be superior by only 2% - 6%, with a large amount of training, while GMM performed the worst.

A more significant factor in audio analysis appears to be the selection of analysis features. A wide variety of features have been presented in the literature, being extracted from the audio signal more or less directly in either the temporal or frequency domains. The Mel-Frequency Cepstral features (MFCC), which are frequency transformed and logarithmically scaled, appear to be universally recognised as the most effective. Peltonen [14] evaluated a wide range of time domain and frequency domain features in the context of the CASR problem for a total of 26 different acoustic scenes. Stereo recordings were made to provide information about directionality of sources. A range of time and frequency domain and MFCC features were evaluated using both kNN and GMM classifiers. Peltonen was able to classify 17 out of 26 scenes with an accuracy of 68.4%. He showed that increasing the length of test sequence improved overall recognition rates with about 30 seconds being required to reach about 80% of the final recognition rates. He also showed that for sequences over 60 seconds in length the band energy ratio performed better than MFCC and that the k-NN performed better with simpler features while GMM performed better for multidimensional features such as MFCC and band energy ratios. Overall however the multidimensional features which included the MFCC, Band-energy, LP-cepstra and LPC performed on average twice as well as simpler features such as the Centroid, ZCR, short time average energy, flux, and Roll-off.

In this paper we take a different approach inspired by image processing. Recently in the area of face recognition the Trace Transform [15] has been proposed as a robust feature space from which to perform analysis. We apply a modified form of the Trace Transform to the audio classification problem and provide improved performance against MFCC features. The Trace transform is a generalization of the Radon transform of which the Hough transform is a special case. It consists of tracing an image with straight lines along which are calculated certain functionals of the image function. In the Radon transform the functional is just a line

integral while in the trace transform and example functional may be the sum($x-c$) along the line where c is the weighted median of x . This process results in a 2D function of the parameters P and A for each tracing line. These parameters can be considered to define a vector V of distance P from the centre of the parameter space at angle A . Hence the value of the function at instance (P, A) is the line integral for all lines t orthogonal to vector V and passing through point (P,A) . This is depicted in Figure 3.

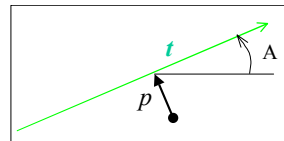


Figure 3 Tracing Line Parameters

The trace transform is normally applied to affine invariant image analysis; in this paper we use it for audio analysis beginning with a spectrogram of the sound to be analyzed. The normal Trace Transform considers the line t to go from one extreme of the spectrogram, pass through the point (P,A) to the other extreme as t goes from $0 \rightarrow 1$. In this case Trace Transform data appears as shown in Figure 4.

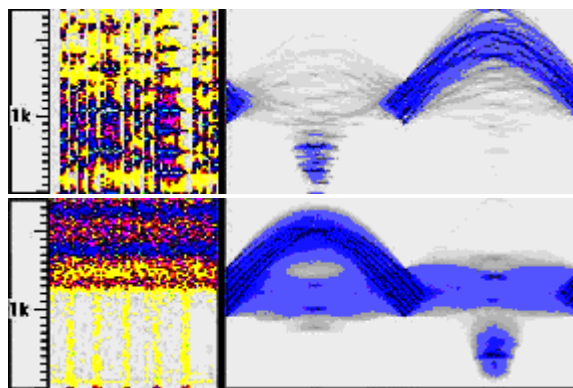


Figure 4 Normal Trace Transform for Bugle reveille and Cricket recordings

Alternatively the line t can be considered to travel in both negative and positive directions away from point (P, A) towards the extremes of the spectrogram. In this case the energy in the Trace diagrams becomes more focused. This is shown in figures 5 – 12. Features are extracted from the trace transform diagram by calculating the diagonal line integrals for each x position.

5. Experimental results

We used a set of 200 different monophonic sounds of consisting of the following nine classes: Sirens, music, screams, environment, insects, animals, humans, City/road noise, Gunshots and explosions. The sounds were recorded at 11kHz using 16 bit samples. A spectrogram was generated for each using the FFT with 512 sample windows. The trace transform was calculated for each spectrogram as described previously and a 360 element vector of line integrals along the diagonals of the trace transforms produced. These vectors were averaged to end up with a total of 60 bins. A k-NN classifier was used to recognize acoustic events. In our experiments the use of the trace transform resulted in improvements of 8% in recall and 14% in overall classification accuracy over the MFCC features.

6. Conclusions

Wireless sensor networks can be retrofitted to existing video surveillance systems. This permits automatic selection of cameras within the vicinity of relevant acoustic event to focus their attention in the direction of the event to capture to bring the source of the event into the field of view.

Each node in the sensor network consists of a ZigBee module with a PIC processor, an electret microphone and related support components. Audio is captured and compressed using G.726 by the PIC processor and relayed to the central node for localization and classification.

The use of the Trace transform derived features provided noticeable improvements in classification accuracy over the standard MFCC features.

7. Acknowledgements

The author wishes to express his appreciation to NICTA for providing the hardware resources for this project. NICTA is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

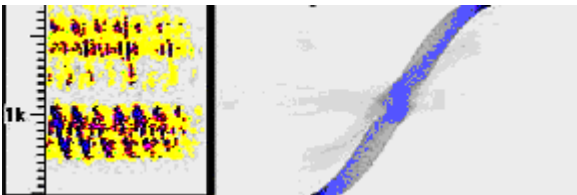


Figure 5 Trace Transform for ambulance siren

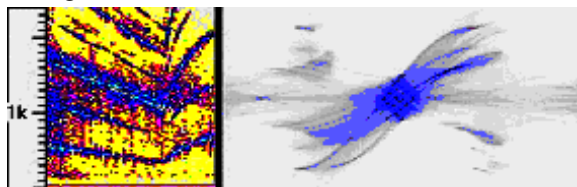


Figure 7 Trace Transform for police siren

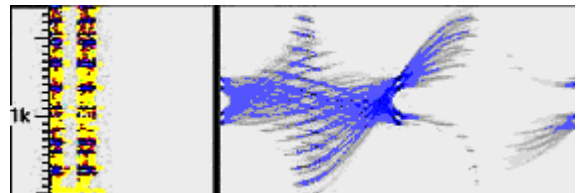


Figure 9 Trace Transform for car horn beep

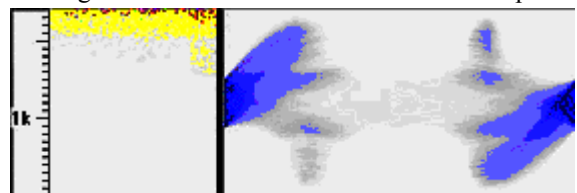


Figure 11 Trace Transform for cicada recording

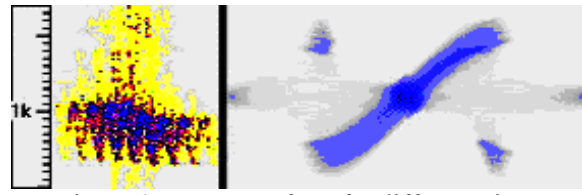


Figure 6 Trace Transform for different siren

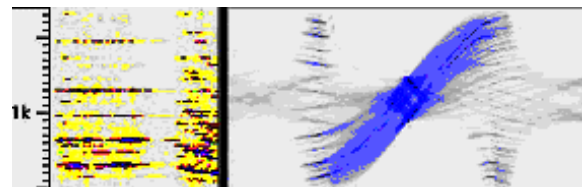


Figure 8 Trace Transform for bach organ music

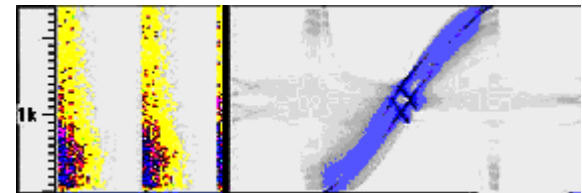


Figure 10 Trace Transform for gun shots

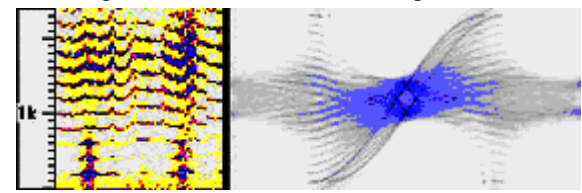


Figure 12 Trace Transform for mosquito buzz

8. References

- [1] Helene A Wells, Troy Allard and Paul Wilson, "Crime and CCTV in Australia : understanding the relationship". Online Resource: http://epublications.bond.edu.au/hss_pubs/70, Gold Coast, Qld : Bond University, 2006.
- [2] G. Medioni, I. Cohen, S. Hongeng, F. Bremond and R. Nevatia. Event Detection and Analysis from Video Streams, IEEE Transaction on Pattern Analysis and Machine Intelligence, 8(23) pages 873-889, August 2001
- [3] T. Ajdler, I. Kozintsev, R. Lienhart and M. Vetterli, "Acoustic Source Localization in Distributed Sensor Networks", Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, Vol. 2, pp. 1328-1332, 2004
- [4] Birchfield, S. T., and Gillmor, D. K. "Acoustic source direction by hemisphere sampling". IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 5 (2001), 3053--3056.
- [5] Maurizio Omologo, Piergiorgio Svaizer, "Acoustic Event Localization Using A Crosspower-Spectrum Phase Based Technique" Proc. ICASSP '94 Adelaide, Australia, vol.2, pp. 273-276,
- [6] Svaizer, P., Matassoni, M., Omologo, M., "Acoustic Source Location in a Three-dimensional Space using Cross-power Spectrum Phase." Proc. of ICASSP, Vol.1, p.231, Munich, Germany, April 1997.
- [7] Regunathan Radhakrishnan, Ajay Divakaran, Paris Smaragdīs, "Audio analysis for surveillance applications," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, p.158 – 161, October 2005.
- [8] L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi, A. Sarti, *Scream and gunshot detection in noisy environments*, EURASIP European Signal Processing Conference, Poznan, Poland, September 2007
- [9] M.F. McKinney and J. Breebaart. "Features for audio and music classification." In Proc. of the International Conference on Music Information Retrieval (ISMIR 2004), pages 151--158, Plymouth MA, 2004.
- [10] Silvia Allegro, Stefan Launer, Michael Buehler, "Automatic Sound Classification Inspired by Auditory Scene Analysis," Eurospeech, Aalborg, Denmark, 2001
- [11] Ravindran, S. Anderson, D. Slaney, M. "Low-power audio classification for ubiquitous sensor networks" Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. (ICASSP '04), 17-21 May 2004, Vol.4, pp:337-340.
- [12] José Anibal Arias, Julien Pinquier and Régine André-Obrecht, "Evaluation Of Classification Techniques For Audio Indexing," Proceedings of 13th European Signal Processing Conference, September 4-8, 2005. EUSIPCO'2005, Antalya, Turkey.
- [13] Selina Chu, Shrikanth Narayanan, C.-C. Jay Kuo, and Maja J. Mataric. "Where am i? scene recognition for mobile robots using audio features". In Proceedings of ICME, Toronto, Canada, July 2006.
- [14] Peltonen, V. Tuomi, J. Klapuri, A. Huopaniemi, J. Sorsa, T., "Computational auditory scene recognition", Proceeding of. International Conference on Acoustics, Speech, and Signal Processing, 2002. (ICASSP '02). IEEE, May 13-17, 2002, Orlando, FL, USA, vol.2, pp:1941-1944.
- [15] A Kadyrov and M Petrou. "The Trace transform and its applications". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 23, pp 811-828. , 2001