Towards Event Detection in an Audio-Based Sensor Network

Alan F. Smeaton*
Centre for Digital Video Processing and
Adaptive Information Cluster
Dublin City University Glasnevin, Dublin 9,
Ireland
Alan.Smeaton@DCU.ie

Mike McHugh
Centre for Digital Video Processing and
Adaptive Information Cluster
Dublin City University Glasnevin, Dublin 9,
Ireland
mmchugh@computing.DCU.ie

ABSTRACT

In this paper, we describe an experiment where we gathered audio information from a series of conventional wired microphones installed in a typical university setting. We also obtained visual information from cameras located in the same area. We set out to see if audio analysis could be used to assist our existing visual event detection system, and to note any improvements. We were not concerned with identifying or classifying what was detected in the audio. Our aim was to keep audio processing to a minimum, as this would enable wireless sensor networks to be used in the future. We present the results of analysis of audio information based on the mean of the volume, the zero-crossing rate, and the frequency. We found that detecting events based on their volume returned satisfactory results.

Categories and Subject Descriptors

H.5.5 [**Sound and Music Computing**]: Signal analysis, synthesis, and processing; I.6 [**Simulation and modelling**]: Applications

Keywords

Sensor networks, security monitoring, audio surveillance

General Terms

Measurement, Experimentation

1. INTRODUCTION

Conventional approaches to surveillance and monitoring usually involve video coverage and the industry to support this is well developed. CCTV deployment is widespread and commonplace. However, conventional CCTV has well-documented problems of lighting (will not work at night unless using infra-red or thermal spectrum), and is susceptible to poor weather conditions if outdoor, like mist, fog, rain, etc.

Maximising the potential of CCTV for surveillance involves confronting an even greater problem. Monitoring conventional video surveillance is a mind-numbing task which requires concentration and effort which we humans do not have the capacity to do repeatedly and effectively. If we consider surveillance on a large area, like a university campus, which may have several hundred CCTV cameras, then monitoring activity to any real degree is simply impossible and thus CCTV footage is used almost exclusively to gather evidence *after* some incident has already occurred. The automatic detection of "events" does not yet occur with video-based CCTV and although there is much work being done on this, it is still at the research stage and far from deployment on a large scale.

If we consider the nature of an "event" we may wish to detect, the amount of information generated is more than just visual in nature. Many events which are significant from a monitoring point of view are accompanied by audio information that would be useful to examine. Examples of these kinds of events include shouting in a public area, the sound of an engine in a pedestrian zone, or even a door opening or a desk banging in an office out of business hours. The significance of these events is not provided solely by their semantic information, but rather by their temporal context. A monitoring system that does not need to distinguish between a door opening and glass breaking, but rather that it should expect to hear one and not the other at a given time and location, should be feasible to create.

In our work we are setting out to build a sensor network for audio monitoring of an environment using inexpensive sensor nodes. One of the reasons these should be inexpensive is because they perform relatively simple signal processing on the audio signals received, in a power-efficient way. Thus they are simple devices and could be mass-produced at a reasonable cost, and easily deployed. They should each be capable of creating an environmental audio signature for a specific time and place, and recognising deviations from that signature. In conjunction with CCTV information, this should make recognition of significant events after the fact easier, and could potentially highlight them as they occur. In this paper, we present an experiment to gather and analyse audio and visual data, and we show that simple analysis of audio can be used to augment event detection in CCTV.

*Author for correspondence.

VSSN'05, November 11, 2005, Singapore.

2. RELATED WORK

The sophistication of the technology available to the end-user increases year on year, and some of the most striking advances have been seen in information capture. What was once considered transient information can now be captured, stored and recalled at the touch of a button. As the scope and amount of information we can store increases, so too does the amount of applications that become useful, or even viable. The feedback available in a thoroughly monitored location such as that described in [1] allows for changes in human behaviour to be monitored and theoretically responded to automatically. In such an environment, audio, seismic, and visual sensors all work together, with the goal of making life easier for the residents. A more focussed application in the same vein can be seen in [11], which describes a monitoring system comprised entirely of audio sensors. This approach was suggested by privacy concerns that were raised by subjects concerning video surveillance, and shows that the potential for the use of audio monitoring is recognised. On a more personal scale, the image suggested by [7] is that of the replacement of diaries with dictaphones, on a much greater scale than that allowed by the humble tape cassette.

As more and more data is generated, though, it becomes harder to identify key points of interest. One approach that has been investigated is the classification of audio events [8, 9]. This kind of classification relies on extracting kinds of events from training data, with this extraction performed automatically or with the aid of user interaction. A variation on this can be seen in [6], which uses a predefined database of events of interest as its training base. These, and similar papers [4, 10], extract various features from the audio that correspond to events of interest, and then use these features to aid event detection. This method, while allowing for high precision and recall of defined events, requires a certain amount of data processing to operate. When we examine the properties of wireless sensor motes (see Section 3.1)[3], we will see the attraction of alternate methods of event detection.

Another approach is to "scavenge" information from other sources, such as GPS, timestamps or application logs [7], or to take advantage of a detailed knowledge of the capture system's setup [5]. This type of information, external to the audio captured, can allow differentiation between significant and non-significant events, as well as allowing features like spatial location to be calculated. To do so demands a certain level of expertise, either in the installation of the system, or in the amalgamation and interpretation of the various pieces of information.

Considering all the above, we can see that the overlap between visual data and audio data for the same physical location is already seen to be advantageous when attempting event detection. We therefore set out to see if audio analysis could be brought to bear on our existing visual event detection system, and note any improvements.

3. THE CASE FOR AUDIO EVENT DETECTION AS A SENSOR NETWORK

Audio surveillance is traditionally performed using one or more microphones that are wired up to a central unit. This central unit may be a dedicated microphone amp, or a direct link to a computer. The microphone capture equipment may be a component of an integrated CCTV installation – in which the microphones are physically located within the camera housing – or it may be a separate system that monitors the same area as a specific CCTV device. The audio information for that location is captured, and analysis of the captured information is then carried out by a separate processor, either continuously as the audio is streamed in [6], or after the fact on the stored audio [9].

An alternative approach is to create an audio sensor network, where each sensor is a stand-alone unit with wireless communication capabilities. The sensors would consist of a microphone, a processor, and a power-unit. Once installed in the desired location, the sensors should generate their own network, which would also contain a cental base unit. The sensors would work in combination with each other to detect significant/abnormal events for their specific location. Thus, the analysis to detect events is performed in-situ, and the result of that analysis is communicated back to base.

When comparing the traditional wired approach to the wireless one, we can see that each has distinct features. Wired installs allow detailed knowledge of the location of the microphones – which can be used to spatially locate events [5] – pre-determined power consumption for the microphones, and the components are widely available. However, the installation process is long and involved, and physical constraints are imposed upon the system by the requirements of cabling: the microphones can only be located so far from their central unit, and in many cases the cabling must be unobtrusive, accessible for maintenance and yet protected from damage or interference. Another factor that needs to be considered is the length of focus of the CCTV camera. If the positioning of the microphones is tied to the physical location of the camera, a situation may arise where audio events that are significant are not recognised. As sound propagation obeys the inverse square rule [12], significant sounds may not have sufficient power to reach the microphones. An example of this may be street scene monitoring, where for safety and security reasons the camera is usually mounted at a height.

A wireless approach, on the other hand, has a different set of issues. Care needs to be taken in the synchronisation of signals from each sensor, and the processing performed would need to make efficient use of the limited power that would be available. Advantages include the fact that installation is greatly simplified, as there are no cabling constraints. This would allow the most efficient placement of sensors to be used, aligning more directly with the focus of the CCTV camera. As the sensors would be small, discreet placement would be possible, and because a desired property of all sensor networks is that the individual components are cheap, maintenance costs would be reduced.

A design decision that needs to be made in both kinds of network is the amount and type of data that is to be kept. In a wireless sensor network, the immediately obvious choice is to store only notification of events, which can then be correlated with CCTV imagery to determine what the event was. In a wired environment, if the processing is to take place post-event, adequate storage needs to be provided. If real-time processing is desired, then CPU and memory need to be capable of the task: as the number of microphones in the network goes up, so do the processing requirements.

We believe that these are compelling reasons for examining the use of a wireless audio sensor network based system to complement conventional video-based CCTV. While it

may not be advisable in all situations, expanding the possibilities available to professionals in the field would seem to be desirable.

3.1 Sensor Mote Properties

Our proposed wireless audio sensor network is a specific implementation of a wireless sensor network. Such a network is formed by many small, self-contained computers – known as *motes* – communicating with each other. While we have not implemented such a network yet, the results of the experiment described in this paper show that implementation is feasible given the capabilities of typical motes.

Examples of these motes include the mica2 and mica2dot motes developed at the University of California at Berkeley. An analysis of their capabilities was carried out in [2], which illustrates both the advantages and constraints that they come with. They consist of a 4Mz, 8-bit Atmel microprocessor, 512KB of logger memory, and a sensor, in our proposed system, an audio sensor. They communicate in a half-duplex fashion, and have a theoretical maximum throughput of 19.2Kbps. The throughput rate is highly dependent on the message size, however, and was shown to be 4.6Kbps with a message size of 36 bytes.

One of the factors that was shown to influence the sensor network behaviour adversely was environmental conditions. The communication range between sensors drops significantly in the presence of rain or fog, and their installation in outdoor locations would need to take this into account. Communication is also impacted by their physical height above the ground.

The above description shows that for an audio sensor network, attention needs to be paid both to the on-board operations and to the system layout. The processor is usually not powerful enough for complicated operations, and so the signal processing performed on each mote needs to be simple, and the system as a whole should take advantage of the aggregation of information from each mote to perform more complicated tasks than a single mote can. Less complicated on-board processing also reduces power usage, which extends the useful life of the mote. The layout needs to factor in the communication properties of the motes in the prevalent environment.

4. EXPERIMENTAL SETUP

For our initial experiment to determine the audio characteristics which need to be analysed in order to recognise events, we implemented a traditional, wired approach. At this point, our aim was to gather the maximum amount of information. In contrast to the usual approach of capturing the basic signals and using processing to enrich the information, we decided to capture and store far more information than we expected to need in order to determine what audio features work, and what is necessary, and based on the results we will determine the the most appropriate level of detail to capture in the future.

Our goal for the final deployed system is to not rely on stereo information, nor on directional (and expensive) microphones. Instead, we want to create as cheap a device with maximum benefit – specifically for features for events – as is possible. In its current form, the setup approaches this goal, with the exception of the storage requirements.

We used the most economical components available to us that captured the most amount of information. We laid out a system consisting of 3 CCTV cameras, and 7 microphones hooked up to a MOTU Firewire interface, which provided pre-amps for each microphone. The specifications of the components can be seen in Table 1.

Table 1: Equipment specification

Microphone specifications	
Model	Behringer ECM8000 Omnidirectional Measuement Microphones
Impedance	600 Ohms
Sensitivity	60 dB
Bandwidth	15 Hz - 20 kHz
Amp specifications	
Model	Mark Of The Unicorn 896
Input/Output	8 XLR/2 IEEE 1394
Sampling Rates	44.1, 48, 88.2, 96 kHz
Camera specifications	
Model	Axis 2100 Network Camera
Settings	Constant capture @ 640*480 with low compression
Avg Img Size	34k

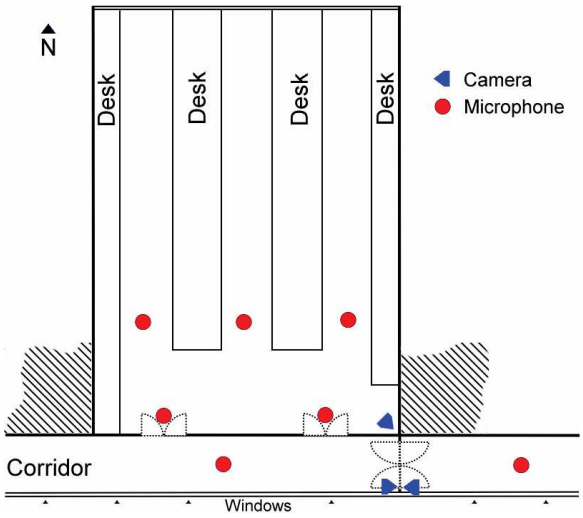

Figure 1: Capture equipment layout

Each camera contains a web server, which was used to display the image stream. The capture program captured 1 image per second from each camera, accessing each web server to do so. To control the possibilities of network collisions, the 3 cameras were connected to a 10mbit hub which ran directly to the computer running the capture program.

Each microphone was calibrated using normal environmental noise, so that there was constant activity on each microphone at the same level when the only noise was the air conditioning in the specific locations.

The equipment was set up in an undergraduate computer lab and along the corridor which provides access to that lab (Figure 1). There are two doors leading to the lab from the corridor, and the corridor itself contains three sets of internal connecting doors. The southern wall of the corridor contains windows for its entire length. The corridor also allows access to other labs, and so is a major source of pedestrian activity.

This environment is publicly accessible between 07:00 and 22:00. The computers in the undergraduate lab are always on, and after a certain amount of idle time their screensavers activate.

Each corridor was monitored by one camera and one microphone. Each camera was mounted above the central connecting door between the corridors. The microphone in the east corridor was 480cm from the camera, and the one in the west corridor was 600cm from the camera. In the lab, the camera was mounted in the south-east corner, and the microphones were mounted above the entrances and exits to both the lab and the individual rows of desks (Figure 1). Each microphone was connected to the amplifier using the same length of cable.

The audio data was captured at a sampling frequency of 48kHz using 16 bits. This gave rise to storage requirements of 5.49Mb/minute per microphone, or 329.4 Mb/h. For all 7 microphones, 2.25Gb of data was stored for each hour of sound recorded. For this experiment, 92 hours 32 minutes of activity was recorded. While this is indeed a lot of data, we remind the reader that the purpose here was to capture far more data than is needed in order to allow us to determine minimum requirements for audio event detection.

The visual data was captured at a sampling frequency of 1/second. As each image was 34Kb in size, the storage requirements were just under 2Mb/minute per camera, or 120Mb/hour. For all 3 cameras, 360Mb of data was stored for each hour of images stored. For the purposes of this experiment, the amount of images that were considered matches the length of audio recorded. The system for capturing the images was Java-based, and at times would experience clock drift using the Real Time API. In the time frame of this experiment, it was decided that this was allowable, but it remains an issue that should be addressed in the future.

4.1 Software tools

Once the visual data had been gathered, analysis began using software developed by us. This pre-processing calculates frame differences between successive images, and determines an event as having occurred when the difference exceeds a certain threshold.

The process of the difference calculation involves converting each image to greyscale, and then breaking it into blocks of 8x8 pixels. The average luminance of each of these blocks is then obtained. Comparison between images is done by comparing each block with its equivalent in the comparison image. If the weighted sum of the absolute differences between the each block's average luminance exceeds a specified threshold, the block is deemed to be part of an event.

Filtering is then applied to all blocks, to reduce the separate event blocks down into one (or more) connected event block(s). Each block is analysed to determine how representative it is of its surroundings, and so how likely it is to form part of an actual event. If movement in any of its neighbours is detected, that neighbour's value is set to 1. The median value of its neighbours is then determined. If the result is 1, then the block is considered to be part of an event. If it is 0, then the block is treated as containing no movement. This acts to filter out isolated blocks that, on the first pass, are flagged as movement.

Events are deemed to have occurred when the number of blocks which report events reaches a certain threshold. Once the minimum threshold is breached, supplementary thresholds are used. These thresholds are used to divide the events into different classes: major, significant and minor. The need for these supplementary thresholds was discovered during implementation as without them, the technique described would classify screensavers activating on computer monitors as being an event. With them, they are still reported as an event, but classified as a minor one.

Problems that arise with this method of event detection include the detection of insignificant results, which are so prevalent as to diminish the effectiveness of the approach, and the age old problem of changes in light levels. This is particularly evident in the pictures obtained by the corridor cameras, as one wall consists of south facing windows (see Figure 1).

5. ANALYSIS OF AUDIO AND CORRELATION WITH VIDEO

Because of the large number of visual "events" detected automatically, we decided to examine the audio recordings and see if these could be used as a key to detecting events. We correlated detected audio events with observed events from the video CCTV footage, based on timestamping of the data. As the audio data and the CCTV images were collected by the same machine, time synchronisation was not an issue. Gathering took place in a variety of area types comprising undergraduate laboratories and public corridors, detailed in Section 4.

In a surveillance system, the aim is to detect outlier events. In our proposed system, events would be classified as outliers based on their frequency of occurrence: an unexpected drop or an unexpected rise in event frequency would be something to which attention would need to be drawn. This approach is also proposed in [11].

This paper presents analysis of events detected in the west corridor of Figure 1. For the purposes of this paper, an audio event is classified purely by its noise level: no attempt is made at categorising given events as major or minor. This is an obvious avenue to pursue in future work.

Three different forms of analysis were performed on the same data – event detection based on the difference between the event and the average root mean squared volume, event detection based on the difference between the event and the average zero crossing frequency rate, and event detection based on energy levels in the frequency spectrum. As the eventual aim is to use sensor motes to identify audio events, and given the restrictions described in the section above, a critical restriction was the amount of computational power required to generate the results.

5.1 Volume-based event detection

The first step of the audio analysis was performed on the volume of the recorded data. The data was analysed using 5 second windows with 1 second overlap, and the root mean squared (RMS) value of the volume across the window was determined. The choice of window size reflected the events we were interested in identifying, as it is large enough that minor (measurable in milliseconds) noises will not significantly affect the mean, yet small enough that human generated events – which typically last for longer than a second – are detected. The analysis of 1 hour of data took 1 minutes, 46 seconds. (Analysis of all the data gathered,

Table 2: Air-conditioning Settings

Air-conditioning Settings	
Full	08:00 - 17:00
Half	07:00 - 08:00, 17:00 - 23:00
Off	23:00 - 07:00

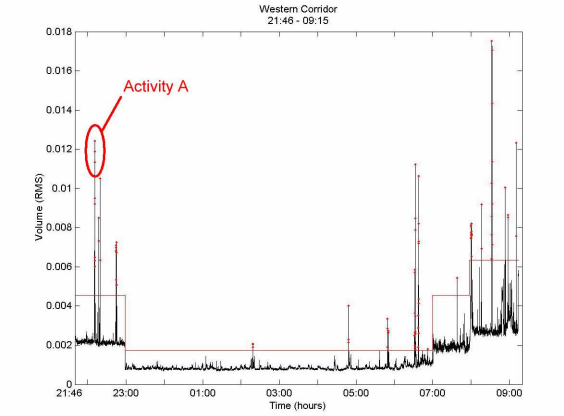

Figure 2: Events by volume, 21:40 - 09:15

both audio and visual, was performed on a P4 3Ghz with 512MB of RAM.)

When the data collected on a weekday was plotted (Figures 2 and 3), 3 different noise levels were visible. These correspond to the building's air-conditioning settings, described in Table 2. Figures 2 and 3 were plotted separately because of the differences of scale in the volume of events observed during the night as opposed to during the day.

For each period, the mean RMS value was computed to give a consistent value for a large time period. Thus, from 23:00 to 07:00, the mean values for each 5 second window were summed, and the mean of this figure was determined. This figure was taken to represent the average volume for that time period, which represented the air-conditioning in an off state. This process was repeated for the other two states of the air-conditioning.

While this knowledge was available to us from external sources, and from visual examination of the data, it is obvious that the next step is to determine the presence of these arbitrary periods automatically. This will be addressed in future work.

Once the mean volume was identified, we could move on to identifying events. A threshold for an audio event was set as twice the mean volume for the event time, which meant that the time at which a given sound was recorded would determine whether or not it was classified as an event. The stepped line running across Figures 2 and 3 indicates the threshold level. This threshold was chosen arbitrarily: future work should address optimising this threshold.

The RMS value of each 5 second sound window was then compared to the threshold, and if it exceed the threshold it was classified as an event component. Temporally adjacent event components were then joined together to form an event. On doing so, it was found that 29 distinct events occurred between 21:40 and 09:15.

When the procedure was repeated on the time period between 09:26 and 21:01, the number of events that were detected was 218. This conforms to expectations, as it is a period of more intensive activity than in the evening.

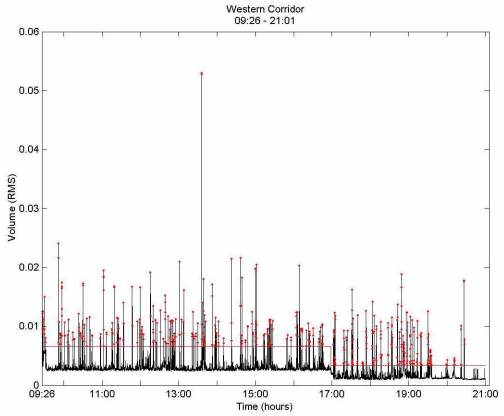

Figure 3: Events by volume, 09:26 - 21:01

On Figure 2, there is an activity A circled at 22:10. This corresponds to security guard activity in the western corridor detected by the camera, and will be referred to in later sections.

5.2 Zero Crossing Frequency Rate-based detection

The next form of audio analysis applied was based around the Zero Crossing Frequency Rate (ZCR) of the recorded data. Again, the data was analysed using 5 second windows with 1 second overlap, and the ZCR of each window was determined. The ZCR measures the number of times in the given time interval that the signal amplitude passes through a value of zero, moving from negative to positive. This means that the value we expect during sustained noise is lower than that associated with fluctuating background noise, and so we can associate events with values that are less than a certain threshold. The analysis of 1 hour of data took 2 minutes, 1 second.

Using the same time periods from the previous section, the mean ZCR value was determined. A arbitrary threshold was set to be, in this case, half the mean ZCR for the event time. This was felt to be analogous to the threshold used in the volume-based event detection. The stepped line running across Figures 4 and 5 represents the threshold level. Again, this raises an obvious goal for future work in optimising the choice of the threshold level.

The ZCR value of each 5 second sound window was then compared to the threshold, and if it was less than the threshold, it was classified as an event component. Again, temporally adjacent event components were merged to form an event, and using this measure, 22 distinct events were detected.

Repeating the procedure on the data collected between 09:26 and 21:01, the number of events detected was 108.

On Figure 4, activity A is circled again.

5.3 Frequency based-event detection

The final form of audio analysis undertaken was an examination of frequency-based information. It was quickly decided that the computational demands that this required

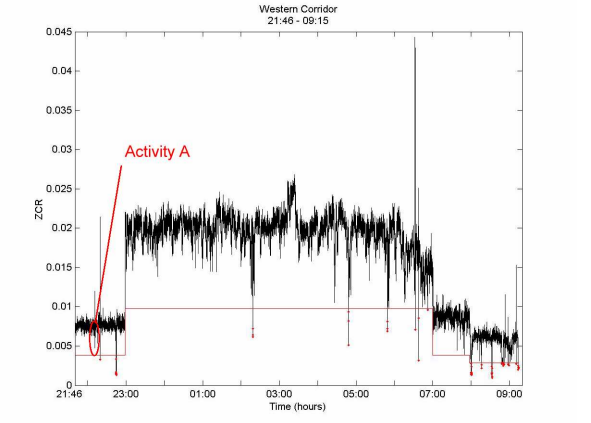

Figure 4: Events by Zero Crossing Frequency Rate, 21:40 - 09:15

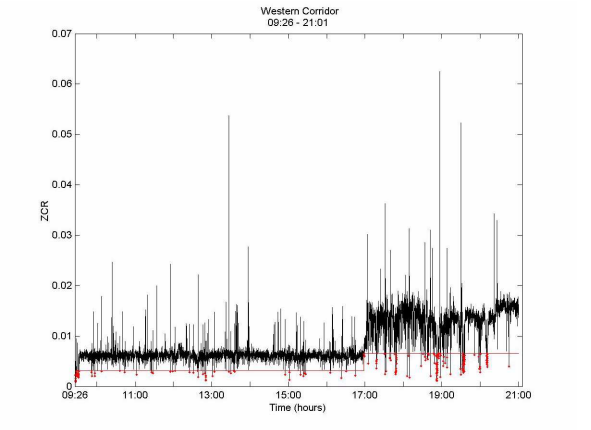

Figure 5: Events by Zero Crossing Frequency Rate, 09:26 - 20:01

meant that its use would be improbable on sensor motes. However, it was felt to be worthwhile to use frequency values as a baseline to ensure that the information that was available from it was duplicated by other methods.

To this end, a 2.5 minute period was analysed using the FFT, corresponding to the activity A referenced in Figures 2 and 4. An more detailed view of Activity A for Volume (RMS) can be seen in Figure 7, which shows the threshold breaches clearly. A corresponding view of Activity A for ZCR is shown in Figure 8, with the threshold unbroken. Examining the frequency spectrum generated (6), we can see that there are readily identifiable peaks that correspond to threshold breaches in Figure 7. This is reassuring, as it corroborates the information obtained through volume analysis.

The time taken to analyse this 2.5 minute sample was in excess of 15 seconds. This is an order of magnitude slower than the volume of ZCR analysis, and as it was calculated on a desktop computer with far more processing power than a sensor mote, it is clear that this is not a useful approach to take.

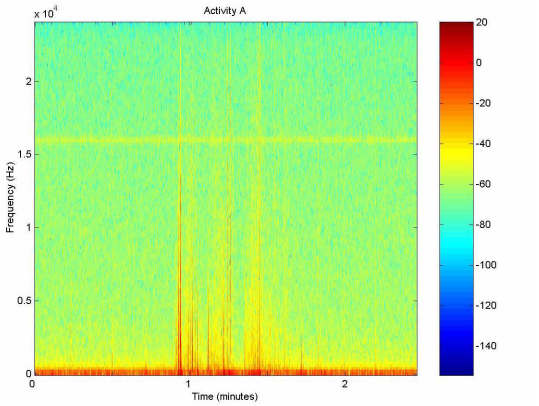

Figure 6: Activity A - FFT

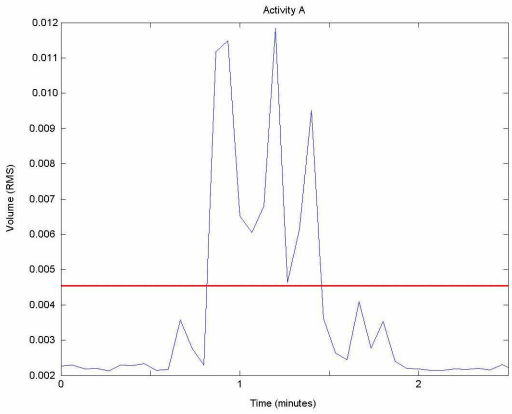

Figure 7: Activity A - Volume (RMS)

5.4 CCTV event detection

Figure 9 is an example of correct event detection in the visual domain. We can see that the source image contains a security guard, and the highlighted image correctly identifies the guard. These two images are the same frame, before and after processing by our event detection software.

Where the change between frames is relatively large, as is the case where a person is moving towards or away from the camera, the accurate identification of the event is much more likely. As a result, it is possible to categorise visual events into major or minor.

An example of incorrect event detection can be seen in Figure 10. This was caused by the movement of reflections in the glass, which meant that, when combined with the changing light levels on the wall, the system categorised the frame as containing a major event. As human observers, we can tell that this is incorrect, but the system has no semantic knowledge to do this.

5.5 Analysis of results

The aim of the experiment reported here was to investigate the use of audio information in confirming the presence of events in a specific area when combined with a CCTV system, and determine the minimum amount of audio signal needed to achieve this.

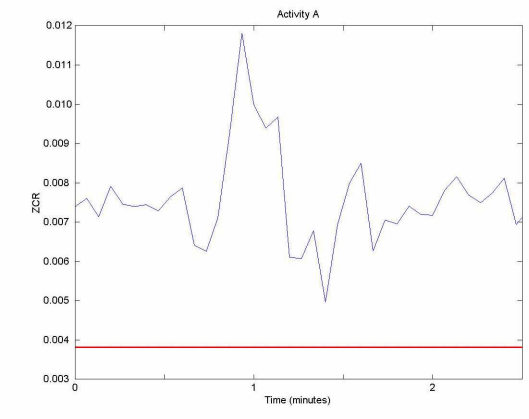

Figure 8: Activity A - ZCR

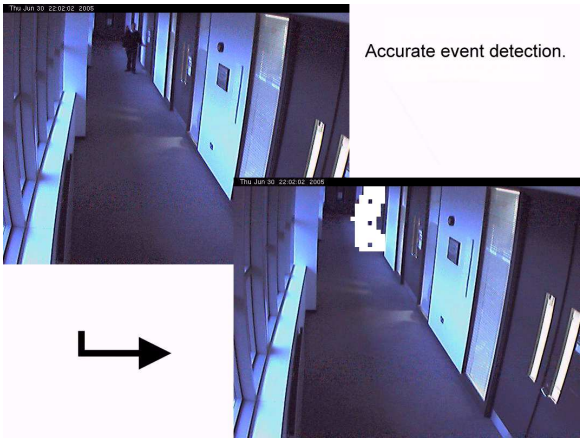

Figure 9: Correct event detection

After we discarded the use of frequency based event detection for more than isolated corroborative checks, we were left with events detected by Zero Crossing Frequency Rate, and events detected by volume(RMS). Comparing the results of these methods, we find that 130 events were detected across both time periods (21:40 - 21:15, 21:26 - 21:01) by the ZCR method, as opposed to 247 by Volume (RMS). A breakdown of these figures by time period can be seen in Table 3.

On further analysis of the 21:40 - 09:15 period, we find that there is substantial overlap between the events detected: there are 15 events in common. However, of the 7 events that were detected by ZCR but not by Volume (RMS), 5 were false positives where no event could be identified when the source audio was listened to. The remaining 2 events were speech events that could not be associated with any source within range of the system. This means that the conversations were picked up from outside the surveillance area,

Table 3: Events detected

Method	21:40 - 09:15	09:26 - 21:01
Volume(RMS)	29	218
ZCR	22	108

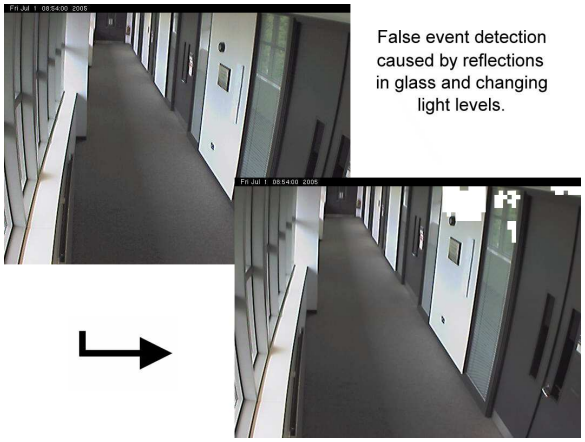

Figure 10: Incorrect event detection

Table 4: False Positives in Event Detection

Method	21:40 - 09:15
Volume(RMS)	2/29 (6.89%)
ZCR	5/22 (22.73%)

either out of doors or in another lab. Of the 14 events that were detected by Volume(RMS) but not by ZCR, the false positive rate was much lower: only 2 out of the 14 had no event that could be identified when the source was examined (See Table 4). Looking at the data returned for Activity A for Volume (RMS) and ZCR (Figures 7 and 8 respectively) illustrates the type of event detected by one approach but not the other.

This result implies that ZCR is more sensitive to speech events than Volume (RMS) is, and less sensitive to other events such as doors opening, or car engines outside. Volume (RMS), on the other hand, responds to changes in the local environment more readily. In the case of a surveillance system, this could be considered an advantage, as it is not affected quite as much by events outside its frame of reference/interest.

Examination of the ZCR results also suggests that modification of the threshold may deliver more accuracy. As can be seen from Figure 5, there are a number of dips that come close to breaking the threshold. The same time period examined in Figure 3 shows that the Volume (RMS) peak at the same time has breached the threshold, and so formed an event.

When taking the events detected by audio and examining CCTV data for the appropriate time, we find that there is a high degree of accuracy in using audio as an event signifier. In addition to positively weighting the presence of an event in the visual domain, the audio data can also be used to negatively weight visual results. The false positive example shown above has no associated audio event, and so it can be discarded from consideration. This property of the system allows us to ignore with greater ease changes in light levels, and silent non-events such as screensavers on monitors.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have shown that audio data can be a significant aid in the analysis of video data for surveillance and security applications. Our working environment for signal processing on the audio data is deliberately constrained by the processing capabilities of a sensor node such as a Berkeley mote. Where does our work on audio-based event detection go from here? Our aim is not to try to replace CCTV but on the contrary, to try to enhance its navigability. Given the large amount of data generated by CCTV system, narrowing down the information presented to the end user of a security system in such a way as to preserve notable events is of great use.

As we have shown in our work, we can have "events" in CCTV video which can be automatically detected and can be categorised on a graded scale from minor (screensavers moving) to major events. We can use these ranges of event sizes to provide a searchable index on CCTV but we can not do more than say "something small moved over a period of X seconds" or "there was a lot of movement here over Y seconds", so it is not really that helpful: the amount of information provided is not significantly reduced. In the absence of semantic or situational knowledge, outlier visual movement is difficult to detect/define. Unless we detect movement in a part of the frame not usual (e.g. on the actual desks in the labs), differentiation between a significant and a non-significant event requires human interaction and actually viewing the source image. Significant events that should be brought to the attention of security may be lost among the events that are unimportant, when the only information provided is the size of the event.

With audio we can also have event detection on a graded scale, from minor event to abnormal sounds, and we have shown that it is possible to detect outlier audio events using simple forms of analysis. It is potentially cheap to deploy, so it is a good complement to CCTV.

When an event is detected in both of the systems – CCTV and audio – the likelihood of it being a significant event, of interest to the monitor, increases. By expanding the range of information available to the system, we can improve the accuracy of its operation. The aim of an audio sensor network would be to assist the end user to sift through data and return points of interest, and do this not through overwhelming them with even more data, but by drawing attention to information that they already have access to, but might not find.

Future work will include further analysis of the data collected to determine what are the most significant pieces of data: is there a specific frequency range that registers events in a more pronounced manner? Is there a constant weighting that can be used when combining the two forms of information, audio and video, or does it vary with time; that is, is audio more reliable in darkness/low light situations, or vice-versa? Is there any situation where it would be useful to transmit the audio information captured in its raw state, rather than just flagging the fact that an event has occurred? More important of all, the constraints that are imposed by the wireless sensor network architecture need to be examined to determine the impact that they will make on the system.

Acknowledgments

This work is partly supported by Science Foundation Ireland under grant 03/IN.3/I361.

7. REFERENCES

[1] The ubiquitous home project, available at http://www2.nict.go.jp/jt/a135/eng/research/ubiquitous_home.html.

[2] G. Anastasi, A. Falchi, A. Passarella, M. Conti, and E. Gregori. Performance measurements of motes sensor networks. In *MSWiM '04: Proceedings of the 7th ACM international symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 174–181, New York, NY, USA, October 4-6 2004. ACM Press.

[3] M. Beigl, A. Krohn, T. Zimmer, and C. Decker. Typical sensors needed in ubiquitous and pervasive computing. In *First International Workshop on Networked Sensing Systems (INSS) 2004, Tokyo, Japan*, pages 153–158, June 22-23 2004.

[4] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai. Highlight sounds effects detection in audio stream. In *Proceedings of the IEEE Int. Conf. on Multimedia and Expo (ICME 2003), Baltimore, 6-9 July 2003.*

[5] J. Chen and K. Phua. An adaptive microphone array with local acoustic sensitivity. In *Proceedings of the IEEE Int. Conf. on Multimedia and Expo (ICME 2005), Amsterdam, 6-8 July 2005.*

[6] C. Clavel, T. Ehrette, and G. Richard. Events detection for an audio-based surveillance system. In *Proceedings of the IEEE Int. Conf. on Multimedia and Expo (ICME 2005), Amsterdam, 6-8 July 2005.*

[7] D. P. Ellis and K. Lee. Minimal-impact audio-based personal archives. In *CARPE'04: Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 39–47, New York, NY, USA, 2004. ACM Press.

[8] R. S. Goldhor. Recognition of environmental sounds. In *Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 149–152.

[9] M. Hrm, M. F. McKinney, and J. Skowronek. Automatic surveillance of the acoustic activity in our living environment. In *Proceedings of the IEEE Int. Conf. on Multimedia and Expo (ICME 2005), Amsterdam, 6-8 July 2005.*

[10] M. F. McKinney and J. Breebaart. Features for audio and music classification. In *Proceedings of the 4th International Conference on Music Information Retrieval, ISMIR 2003*, 26-30 October 2003.

[11] M. Vacher, D. Istrate, L. Besacier, E. Castelli, and J.-F. Serignat. Smart audio sensor for telemedecine. In *Smart Object Conference, Grenoble, France*, pages 222–225, 15-17 May 2003.

[12] H. E. White and D. H. White. *Physics and Music.* Saunders College, Philadelphia, 1980.