International Journal of scientific research and management (IJSRM)
\||Volume||3||Issue||5||Pages|| 2910-2919||2015||
Website: www.ijsrm.in ISSN (e): 2321-3418

Development of an Intelligent Eye for Automatic Surveillance Based on Sound Recognition

Lutfun Nahar Nipa[a] Md. Rokunuzzaman[b] Tamanna Tasnim Moon[c]

Department of Mechanical Engineering, Email: (Corresponding author) nahar_nipa@yahoo.com

Department of Mechanical Engineering, Email: rzaman_me.ruet@yahoo.com

Department of Mechanical Engineering, Email: tamanna.t.moon@gmail.com

[a, b & c]Rajshahi University of Engineering & Technology, Rajshahi-6204, Bangladesh.

Abstract

In surveillance or homeland security most of the systems aiming to automatically detect abnormal situations are only based on visual clues while, in some situations, it may be easier to detect a given event using the audio information. A new platform for sustainable development of automatic surveillance is introduced based on intelligent eye system which gathers information of human behavior, activities and environmental changes. The present research deals with audio events detection in noisy environments for surveillance application. The increasing availability of forensic audio surveillance recordings covering days or weeks of time makes human audition impractical and error prone. The ability of a normal human listener to recognize objects in the environment from only the sounds they produce is extraordinarily robust even in adverse acoustic conditions. In this research, we have developed a surveillance system which can recognize sound sources and detect events. This system can cover large area which is cost efficient. Sound sources can be recognized by comparing the frequency of sounds. This proposed intelligent eye system can recognize different sound sources accurately in real time and pretty much quick.

Keywords: Intelligent eye; automatic surveillance; sound recognition

1. Introduction

When the vision system is unable to detect events occurring at a high speed, sound is an important cue for perception. To become intelligent, systems or robots should have to understand situation, make decisions and interact accordingly. Vision system is susceptible to adverse conditions like fog, mist, rain, dark etc. In these conditions sound system can be very effective. If a camera can be made by an intelligent system to respond accordingly for specific sound recognition, an event can be detected instantly. If an intelligent sound system is introduced, it will be easy to detect an event of any unnatural sound for the operator and take actions accordingly. The perspective of using such a recognition system in surveillance and security applications is therefore possible, on the condition that sound class models could be learned and built at the place to control.

Long-term audio surveillance recordings may contain speech information and also non-speech sounds such as environmental noise, audible warning and alert signals, footsteps, mechanical sounds, gunshots, and other acoustic information of potential forensic interest. Security system should focus on the robustness of the detection against variable and adverse conditions which is particularly important in surveillance applications. Research in the area of automatic surveillance systems is mainly focused on detecting abnormal events based on the acquired video information [1, 2]. In addition to the traditional video cameras, the use of audio sensors in surveillance and monitoring applications is becoming increasingly important. Audio based surveillance has been studied earlier for detecting various types of acoustic events such as

human's coughing in the office environment [3], impulsive sounds like gunshot detection [4], glass breaks, explosions or door alarm [5]. In order to determine whether surveillance technology is actually improving surveillance, the effectiveness of surveillance must be expressed in terms of these higher purposes. This paper investigates techniques to recognize environmental sounds and their direction, with the purpose of using intelligent techniques in an autonomous mobile surveillance robot. It also presents advanced methods to improve the accuracy and efficiency of these techniques. We specifically focus on the robustness of the detection against variable and adverse conditions and the reduction of the false rejection rate which is particularly important in surveillance applications.

2. Mathematical Foundation of Sound Recognition

Sound effected camera control is the technology to detect sound sources with the help of the installed sensors in a definite platform and orient the camera accordingly to the direction from where the sound is created or occurred [6, 7]. According to sound events camera movement is controlled and sound source is localized automatically [8, 9]. Our approach to sound classification is inspired by the human auditory system in that we extract auditory features as known from auditory scene analysis from the input signal [10]. At the highest level, all sound recognition systems contain two main modules feature extraction and feature matching [11] Feature extraction plays a very important in the sound recognition process. This is basically a process of dimension reduction or feature reduction as this process eliminates the irrelevant data present in the given input while maintaining important information [12]. The whole process is divided into two stages: training phase and testing phase. Detection is the first step of sound analysis system and is necessary to extract the significant sounds before initiating the classification step. The classification stage uses a Gaussian Mixture Model classifier with classical acoustical parameters like MFCC [13]. Detection and classification stages are evaluated in experimental recorded noise condition which is non-stationary and more aggressive than simulated white noise. The basic steps of sound recognition are as follows:

2.1 MFCC (Mel-frequency cepstrum coefficients)

In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstrum coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip. The extraction of the best parametric representation of sound signals is an important task to produce a better recognition performance. In this paper MFCC have been used for feature extraction which is mainly used for sound recognition system [13]. A block diagram of the structure of an MFCC processor is shown here

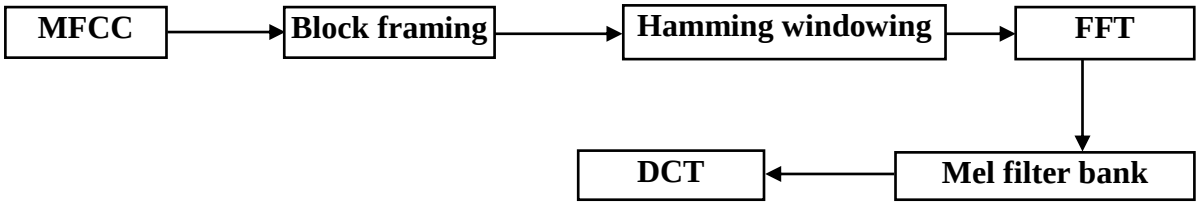


Figure 1: Block diagram of the structure of an MFCC processor

2.2 Blocking and framing

In this step the continuous 1D signal are blocked into small frames of N samples, with next frames separated by M samples (M<N) with this the adjacent frames are overlapped by N-M samples. As per many researches the standard value taken for N= 256 and M= 100 with a reason of dividing the given 1D signal into small frames having sufficient samples to get enough information. Because, if the frame size smaller than this size is taken then the number of samples in the frames will not be enough to get the reliable information and with large size frames it can cause frequent change in the information inside the frame. So, while working with

MFCC these parameters are very common in practice. This process of breaking up the signals into frames will continue until the whole 1D signal is broken down into small frames [12].

2.3 Hamming windowing

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window is represented as shown in Equation (1). If the window is defined as W (n), $0 \leq n \leq N-1$ where N = number of samples in each frame; Y[n] = Output signal; X (n) = input signal; W (n) = Hamming window, then the result of windowing signal is shown below [13]:

$$Y[n] = X(n) * W(n) \quad \text{.....(1)}$$

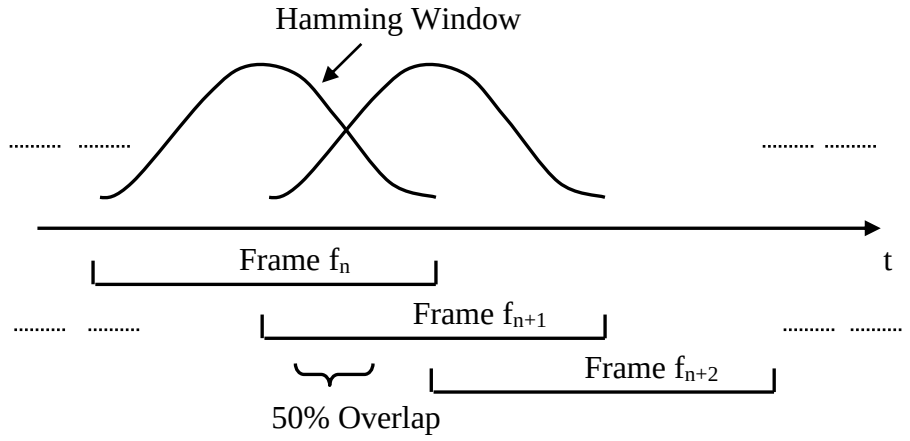


Figure 2: Frame blocking in the conventional MFCC extraction algorithm

2.4 FFT (Fast Fourier transform)

To convert each frame of N samples from time domain into frequency domain FFT is being used. The Fourier Transform is used to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement supports as shown in Equation (2) below [13]:

$$Y[w] = FFT[h(t) * X(t)] = H(w) * X(w) \quad \text{.....(2)}$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

2.5 Mel filter bank

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the Centre frequency and decrease linearly to zero at center frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components. After that the following equation as shown in Equation (3) is used to compute the Mel for given frequency f in Hz [13]:

$$F[Mel] = [2595 * \log 10[1 + f / 100] \quad \text{.....(3)}$$

2.6 Discrete cosine transform

This process of carrying out DCT is done in order to convert the log Mel spectrum back into the spatial domain. For this transformation either DFT or DCT both can be used for calculating Coefficients from the given log Mel spectrum as they divide a given sequence of finite length data into discrete vector. However, DFT is generally used for spectral analysis whereas DCT used for data compression as DCT signals have

more information concentrated in a small number of coefficients and hence, it is easy and requires less storage to represent Mel spectrum in a relative small number of coefficients. The output after applying DCT is known as MFCC (Mel Frequency Cepstrum Coefficient) which can be written by the Equation (4) as

$$C_n = \sum_{k-1}^{k} (\log D_k) \cos[m(k - \frac{1}{2}) \frac{\pi}{k}] \quad \text{.....(4)}$$

where $m = 0, 1 \ldots$ k- 1 & where C_n represents the MFCC and m is the number of the coefficients.[12]

2.7 Vector Quantization

A vector quantizer maps k-dimensional vectors in the vector space R^k into a finite set of vectors $Y = \{y_i: i = 1, 2, ..., N\}$. Each vector y_i *is* called a code vector or a code word and the set of all the code words is called a code book. Associated with each code word, y_i, is a nearest neighbor region called Voronoi region, and it is defined by Equation (5) as:

$$V_i = \{x \in R^k : \|x - y_i\| \leq \|x - y_j\|, \text{ for all } j \neq i\} \quad \text{.....(5)}$$

The set of Voronoi regions partition the entire space R^k can be written by the Equations (6) and (7) such that:

$$\bigcup_{i=1}^{N} V_i = R^k \quad \text{.....(6)}$$

$$\bigcap_{i=1}^{N} V_i = \phi \quad \text{for all } i \neq j \quad \text{.....(7)}$$

As an example we take vectors in the two dimensional case without loss of generality. Figure 1 shows some vectors in space. Associated with each cluster of vectors is a representative code word. Each code word resides in its own Voronoi region. These regions are separated with imaginary lines in figure 3 for illustration. Given an input vector, the code word that is chosen to represent it is the one in the same Voronoi region.

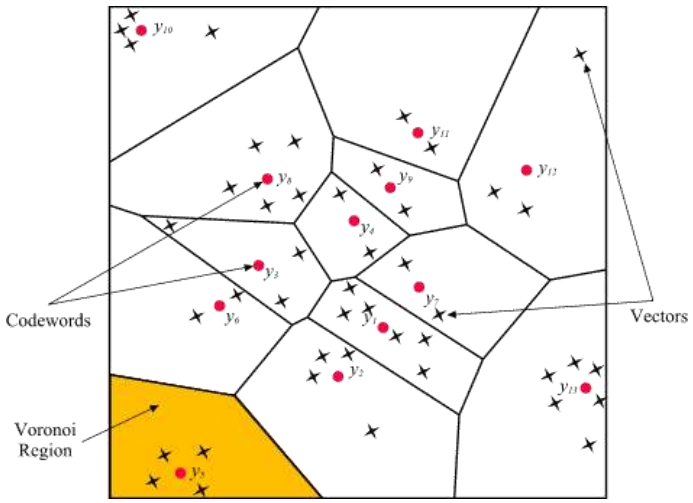


Figure 3: Vectors in space and quantization in a Voronoi region

Here, Code words in 2-dimensional space. Input vectors are marked with an x, code words are marked with red circles, and the Voronoi regions are separated with boundary lines.

The representative code word is determined to be the closest in Euclidean distance from the input vector. The Euclidean distance is defined by the Equation (8) as:

$$d(x, y_i) = \sqrt{\sum_{j=1}^{k} (x_j - y_{ij})^2} \quad \text{.....(8)}$$

where x_j is the jth component of the input vector, and y_{ij} is the jth is component of the code word y_i.

2.8 LBG Design Algorithm

The LBG VQ design algorithm is an iterative algorithm which alternatively solves the above two optimality criteria. The algorithm requires an initial code book $C^{(0)}$ .This initial codebook is obtained by the splitting method. In this method, an initial code vector is set as the average of the entire training sequence. This code vector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The final two code vectors are splitted into four and the process is repeated until the desired number of code vectors is obtained. The algorithm is summarized as below:

1. Given τ . Fixed $\varepsilon > 0$ to be a "small" number.
2. Let $N = 1$ and

   $$c_1^* = \frac{1}{M} \sum_{m=1}^{M} x_m$$

   Calculate

   $$D_{ave}^* = \frac{1}{Mk} \sum_{m=1}^{M} \left\| x_m - c_1^* \right\|^2$$

3. **Splitting:** For $i = 1, 2, ..., N$, set

   $c_1^{(0)} = (1+\varepsilon) c_i^*$,

   $c_{N+i}^{(0)} = (1-\varepsilon) c_i^*$,

   Set $N = 2N$

4. **Iteration:** Let $D_{ave}^{(0)} = D_{ave}^*$. Set the iteration index $i = 0$.

   i. For $m = 1, 2, ..., M$ find the minimum value of

   $\left\| x_m - c_n^{(i)} \right\|^2$, over all $n = 1, 2, ..., N$.

   Let n^* be the index which achieves the minimum. Set

   $Q(x_m) = c_{n^*}^{(i)}$

   ii. For $n = 1, 2, ..., N$, update the code vector

   $$c_n^{(i+1)} = \frac{\sum_{Q(x_m) = c_n^{(i)}} x_m}{\sum_{Q(x_m) = c_n^{(i)}} 1}$$

   iii. Set $i = i + 1$

   iv. Calculate

$$D_{ave}^{(i)} = \frac{1}{Mk}\sum_{m=1}^{M}\left\|x_m - Q(x_m)\right\|^2$$

v. If $(D_{ave}^{(i-1)} - D_{ave}^{(i)}) / D_{ave}^{(i-1)} > \varepsilon$, go back to Step (i).

vi. Set $D_{ave}^{*} = D_{ave}^{(i)}$. For $n = 1, 2, ..., N$, set $c_n^{*} = c_n^{(i)}$ as the final code vectors.

5. Repeat Steps 3 and 4 until the desired number of code vectors is obtained.

3. System architecture

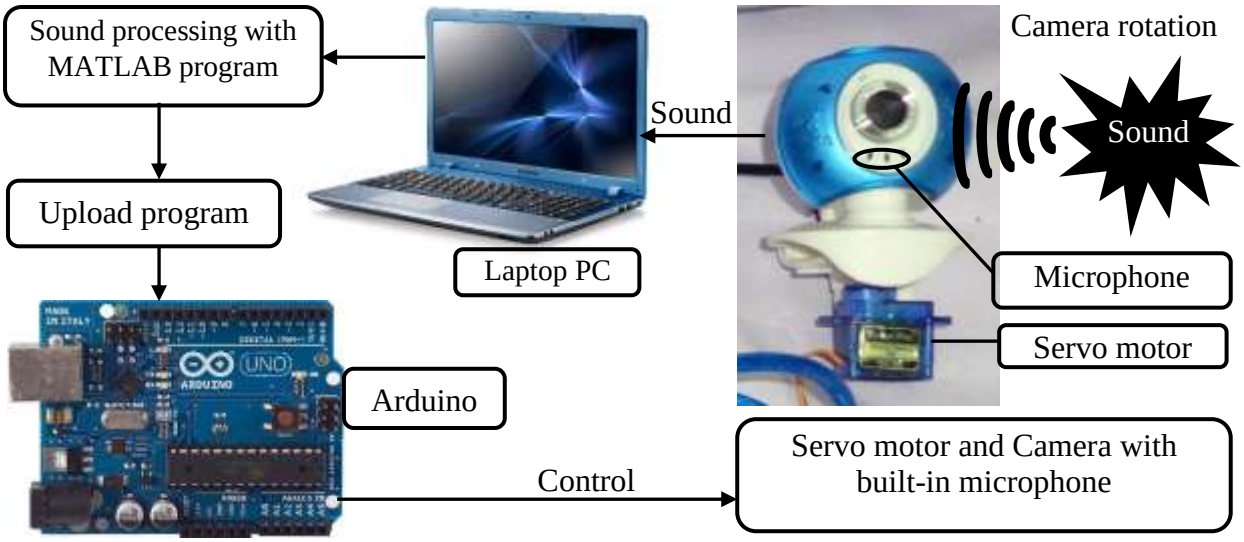
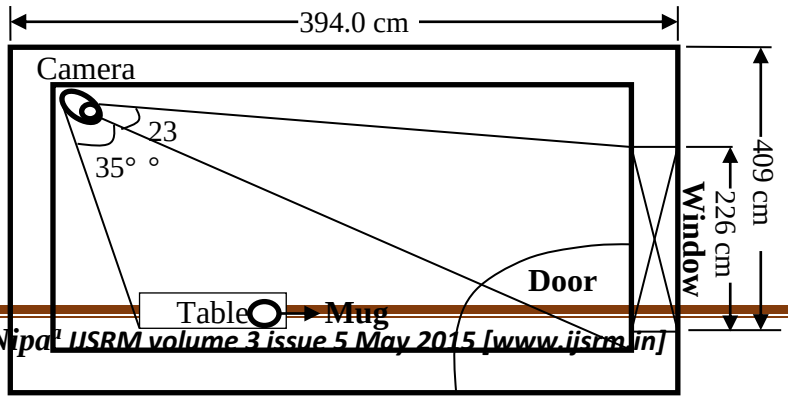


Figure 4: System architecture for development of intelligent eye

Figure 4 shows the system architecture for development of intelligent eye. When a sound event happens, then it is captured by built-in micro-phone of the camera. The sound signal is passed to the laptop for processing. The sound is processed by the algorithms discussed in section 2 and implemented with MATLAB. The program is then uploaded to an Arduino board. The output signal of the arduino is then fed to the servo motor input to control the movement of the camera toward the sound source.

3.1 Configuration & Position of Sound Source

Three sound input are taken for experiments and their position including camera position with specific angle and room layout are shown in Figure 5.



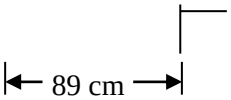


Figure 5: Project room layout

3.2 Intelligent Eye algorithm

The intelligent eye algorithm is based on sound recognition and pointing of the camera towards the recognized sound to track objects. The sound recognition is based on feature matching of sound signals between trained signals and input sound. The feature extraction, training and matching is done through a series of operations namely MFCC computation, Vector Quantization and LBG Design algorithms. Figure 6 shows the complete flow chart of the intelligent eye algorithm.

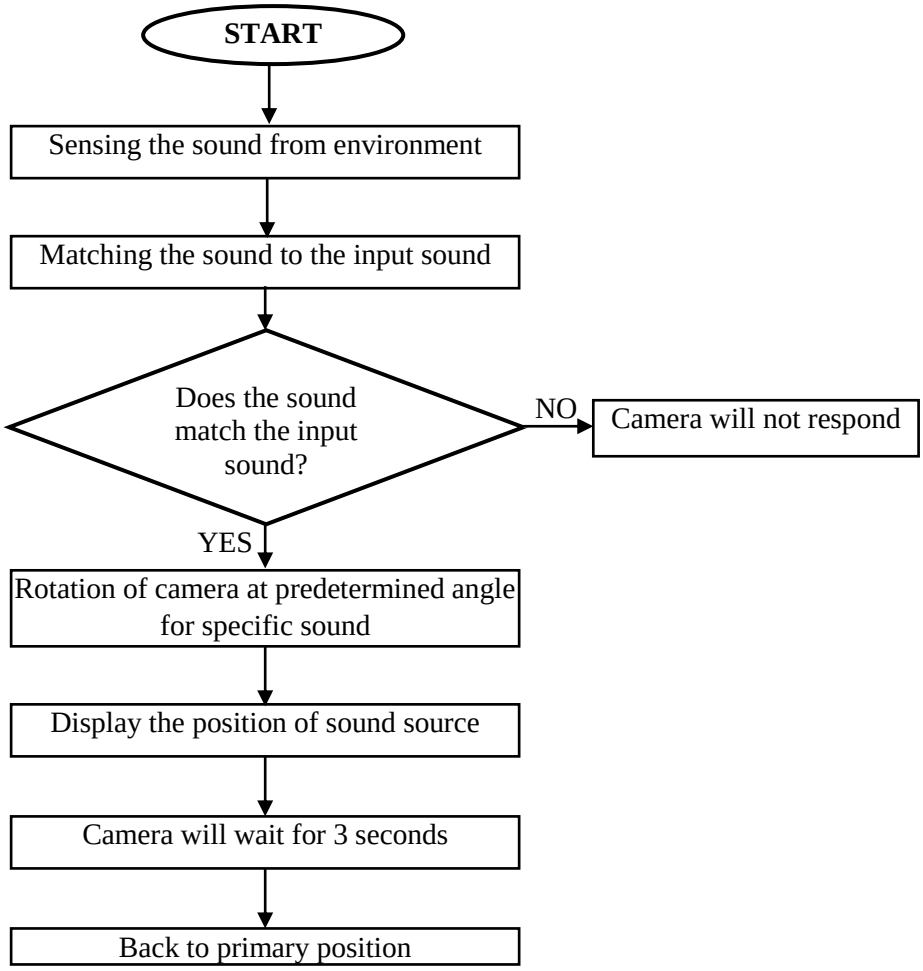


Figure 6: Flow chart of the intelligent eye algorithm

3.3 Camera movement algorithm

After recognition of sound, the recognized signal is transmitted from serial port of PC to Arduino. According to the configuration and position of sound source as shown in figure 5, the Arduino computes the required angle and rotates the camera with the help of a servomotor. The flowchart of camera movement algorithm is depicted with figure 7.

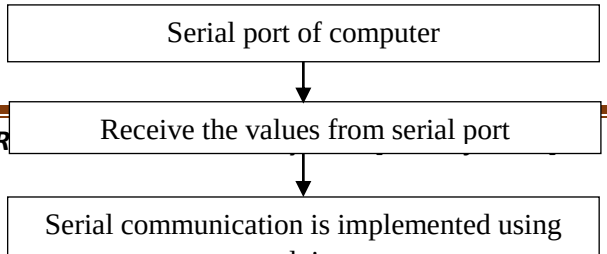


Figure 7: Flow chart of the Camera movement algorithm

4. Results

Different types of sounds from objects have been tested for recognition. The results have been shown in Table1. For each type of sound from object 5 trials have been taken and success rates have been calculated. The success rate is defined by Equation (9) as

$$Success\ rate\ (\%) = \frac{No.\ of\ trials\ of\ sucessful\ detection}{Total\ no.\ of\ trials} \times 100\% \quad \text{.....(9)}$$

Table 1: Success Rates of 3 different sounds

Sound from Object	Successful detection (Y/N)	Success rate (%)
Door	Yes	100
	Yes	
	Yes	
	Yes	
	Yes	
Window	Yes	80
	Yes	
	Yes	
	Yes	
	No	
Plastic mug	Yes	100
	Yes	
	Yes	
	Yes	
	Yes	

Table 2: Recognition time and Success Rates of different sounds

Objects	Average Response Time (s)	Success Rate (100 %)
Door	5	100
Window	5	80
Plastic Mug	5	100

The response times of camera in recognizing specific sound source and the success rates of sound source detection of the system are shown in Table 2. Sample sounds of door, window and mug were tested for several times and response time was recorded using stop watch. The obtained average response time is 5 seconds which means it will take 5 seconds to response after the actual sound detection from environment. Success rate of plastic mug is 100% as its sound is different from door and window. As the sound of door and window is similar so success rate of window was not 100%.

Moreover, it is tested that if two sounds occurred at the same time then sound of maximum intensity was detected. Such as when sound of door and mug or window and mug were occurred simultaneously then sound of mug was detected always due to its high intensity. Similarly when sound of door and window were recorded at the same time then door was detected as the sound source as the sound of door has higher intensity than window.

Table 3: Identification rates of different sounds with code size book

Code size book	**Hamming**
1	57.14
2	85.7
4	90.47
8	95.24
16	100
32	100
64	100

The identification rates are shown when hamming window is used for framing in a linear frequency scale. Table 3 clearly shows that as codebook size increases, the identification rate for each of the three cases increases when code book size is 16, 32 and 64.

5. Conclusion

Hearing is an important part of normal human interaction, yet we understand surprisingly little about how our brains make sense of sound. This project is driven by the desire to understand how human auditory perception works. This is also to identify the nature of sound and focus the camera directly on the sound source. The time wasted to search the source of the sound will be saved in this project. So this project will be more developed and efficient from conventional security system. But the system can be far developed by reducing the response time of camera and more sound can be stored to the system so that it can recognize variety of sounds from the environment. Though the system should have been more robust, the performance of our developed system is quite perfect. This project can play an important role in intelligent security system.

References:

[1] Harma, A, McKinney, M. F, Skowronek, J, Automatic surveillance of the acoustic activity in our living environment, *IEEE International Conference on Multimedia and Expo*, July 2005:1-4.

[2] Michael Cowling. Non- speech environmental sound classification system for autonomous surveillance, *PhD Thesis, Griffith University, Gold Coast Campus*, March 2004

[3] Clavel, C, Ehrette, T, Richard, G, Events detection for an audio-based surveillance system, *IEEE International Conference on Multimedia and Expo*, July 2005: 1306-1309

[4]Valenzise, G, Gerosa, L, Tagliasacchi, M, Antonacci, F, Sarti, A, Scream and Gunshot Detection and Localization for Audio-Surveillance Systems, *IEEE Conference on Advanced Video and Signal Based Surveillance*, September 2007: 21-26

[5] Dufaux, A, Besacier, L, Ansorge, M, Pellandini, F, Automatic sound detection and recognition for noisy environment, *Proc. of the X European Signal Processing Conference*, September, 2000

[4] Brandstein, M. S, Adcock, J. E, Silverman, H. F, A localization-error-based method for microphone-array design, *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1996 ; 2: 901- 904

[6] Wang, H, Chu, P, Voice source localization for automatic camera pointing system in videoconferencing, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1997; 1: 187-190

[7] Cornaz, C, Hunkeler, U, Velisavljevic, V, An automatic speaker recognition system, *Digital Signal Processing Laboratory*, Federal Institute of Technology, Lausanne, Switzerland, 2003

[8] Allegro, S, Buchler, M, Launer, S, Automatic sound classification inspired by auditory scene analysis, *Eurospeech*, Aalborg, Denmark, September 2001

[9] Vacher, M, Istrate, D, Serignat, J. F, Sound detection and classification through transient models using wavelet coefficient trees, *12th European Signal Processing Conference*, September 2004

[10] Zhong-Xuan Yuan, Bo-Ling Xu, Chong-Zhi Yu, Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification, *IEEE Transactions on Speech and Audio Processing*, January 1999; 7(1):70-78

[11] Gupta, S, Jaafar, J, Fatimah, W, Bansal, A, Feature extraction using MFCC, *Signal & Image Processing: An International Journal (SIPIJ)*, August 2013; 4(4): 101-108

[12] Soong, F, Rosenberg, E, Juang, B, Rabiner, L, A Vector Quantization Approach to Speaker Recognition, *AT&T Technical Journal*, March/April 1987; 66: 14-26

[13] Bala A, Kumar A, Birla N, Voice Command Recognition System based on MFCC and DTW, *International Journal of Engineering Science and Technology*, 2010; 2 (12): 7335-734