Environmental Sound Classification using Discrete Wavelet Transform

T. Lenin and P. Dhanalakshmi

Annamalai University, Solatharam, Tamil Nadu, India

ABSTRACT

Classifying the sound from environment is an upcoming research area that enables a variety of new applications. Environmental Sound classification is the basic audio signal processing problem. Audio data could be obtained at any moment when the system is functioning, in spite of challenging external conditions such as lack of light or visual obstruction and is relatively cheap to store and compute than visual signals. This work describes an approach to classify the environmental sounds in our daily life such as kitchen, meeting hall, healthcare and factory. These sounds are collected from BBC sound clip database. Features are extracted from the given set of sound to construct feature vector which prominently discriminates one sound class from another. The Discrete Wavelet Transform (DWT) is used for both training and testing phase. DWT is a wave formation that can be used to analyze the temporal and spectral properties of non stationary signal like audio. The multiclass Support Vector Machine (SVM) classifier is used to analyse the performance of Environmental Sound classification.

INTRODUCTION

In living surroundings, there are different sources emitting various types of environmental sound. This is understood here as a chosen subset of sounds that might be found in a given environment. The environment of interest then determines the scope of the recognition problem. It is important to note that although speech may be included as a sound event category, the content of the speech would not be directly interpreted. Instead the detected speech content may be passed to an Automatic Speech Recognition system for recognition.

Most of the research has been developed in the area of health care environmental sound classification. Current work focus on healthcare environment and meeting environment (Boesnach, 2004). The objective in (Barry, 2006) is to recognize and count coughing sound events, to enable the automatic assessment of cough frequency over a long period of time. An audio recording device is placed on the patient, and the detected sound events are classified simply as cough or non-cough. The scope of (Peng, 2009) is expanded to include a wider set of healthcare audio events, including falling-down-stairs, screaming and collapsing.

Detection and classification of sounds in a meeting room environment is another increasing research area in the field of environmental sound classification (Temko, 2005; Temko, 2009). It is also one of the only areas of Sound Event Recognition research that has a standardized database for comparing the performance of competing systems. The collection of database is from the Classification of Events, Activities and Relationships (CLEAR) 2006/7 workshop evaluations (Temko, 2007; Stiefelhagen, 2009), which is part of the Computers in the Human Interaction Loop (CHIL) project (Waibel, 2010). Here, the acoustic events of interest include steps, keyboard typing, applause, coughing, and laughter, among others, while speech during the meetings is ignored. It was found that achieving an accurate detection and segmentation of the sound events from the continuous audio was the most challenging task (Stiefelhagen, 2009). In particular, several systems have difficulties in dealing with overlapping sound events, which occur frequently in the meeting room environment (Temko, 2009).

The development of high speed computer developed a tremendous opportunity to handle large amount of data from different media. The proliferation of world-wide-web has further fueled its growth and this trend is expected to continue at an even faster pace in the coming years. Even in the past decade textual data was the major form of data that was handled by the computer. But above advancement in technology has led to uses of
non textual data such as video, audio and image. These nontraditional data are known to be multimedia data. The database deals with this type of data is called multimedia database. This database has much more capability than traditional database. One of the best features of multimedia database is content base retrieval. In content base retrieval process the main advantage is that it can search data by its content rather textual indexing. This paper deals with the classification of environmental sounds classification with the help of content based retrieval. Environmental sounds are extremely unpredictable so it’s very hard to classify and store in cluster form according to its feature content. Environmental sounds provide many contextual cues that enable us to recognize important aspects of our surroundings. The goal of this project is to consider techniques to allow machines to extract and classify features from predefined classes of sounds in the environment. This work present the different phases of the project: capture of environmental audio, pre-processing of audio data, feature extraction, training and testing.

Fig. 1: Architecture diagram.

Fig.1 shows the architecture diagram of proposed system

Feature extraction:

The Discrete Wavelet Transform:

The wavelet transform (WT) is a technique for analyzing signals. It was developed as an alternative to the Short Time Fourier Transform (STFT) to overcome problems related to its frequency and time resolution properties. More specifically, unlike the STFT that provides uniform time resolution for all frequencies the DWT provides high time resolution and low frequency resolution for high frequencies and high frequency resolution and low time resolution for low frequencies. In that aspect it is similar to the human ear which exhibits similar time-frequency resolution characteristics.

The Discrete Wavelet Transform (DWT) is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently.

The DWT is defined by the following equation:

\[ w(j, k) = \sum_{n} x(n) 2^{-j} \psi(2^{-j} n - k) \]

where \( \psi \) is a time function with finite energy and fast decay called the mother wavelet. The DWT analysis can be performed using a fast, pyramidal algorithm related to multi rate filter banks.

As a multi rate filter bank the DWT can be viewed as a constant Q filter bank with octave spacing between the centers of the filters. Each subband contains half the samples of the neighboring higher frequency sub band. In the pyramidal algorithm the signal is analyzed at different frequency bands with different resolution by decomposing the signal into a coarse approximation and detail information. The coarse approximation is then further decomposed using the same wavelet decomposition step. This is achieved by successive high pass and low pass filtering of the time domain signal and is defined by the following equations.

\[ y_{\text{high}}[k] = \sum_{n} x[n] g[2k - n] \]
\[ y_{\text{low}}[k] = \sum_{n} x[n] h[2k - n] \]

where \( y_{\text{high}}[k] \) is the high pass filter
\( y_{\text{low}}[k] \) is the low pass filter

Because of the down sampling the number of resulting wavelet coefficients is exactly the same as the number of input points. A variety of different wavelet families have been proposed in the literature. In our implementation, the 4 coefficient wavelet family (DAUB4) proposed by Daubechies is used.
Wavelet representation for audio signals:

An adaptive DWT and DWPT signal representation is considered in this work because of its highly flexible family of signal representations that may be matched to a given signal and it is well applicable to the task of audio data compression. In this case the audio signal will be divided into overlapping frames of length 2048 samples. (Peng, 2009) When designing the wavelet decomposition considered some restrictions to have compact support wavelets, to create orthogonal translates and dilates of the wavelet (the same number of coefficients than the scaling functions), and to ensure regularity (fast decay of coefficients controlled by choosing wavelets with large number of vanishing moments). The DWT will act as anormal linear transform. The wavelet transform coefficients are computed recursively using an efficient pyramid algorithm. In particular, the filters given by the decomposition are arranged in a tree structure, where the leaf nodes in this tree correspond to sub bands of the wavelet decomposition. This allows several choices for a basis. This filter bank interpretation of the DWT is useful to take advantage of the large number of vanishing moments. (Peng, 2009)

Wavelets with large number of vanishing moments are useful for this audio compression method, because if a wavelet with a large number of vanishing moments is used, a precise specification of the pass bands of each sub band in the wavelet decomposition is possible. Thus, it can be approximate the critical band division given by the auditory system with this structure and quantization noise power could be integrated over these bands.

Wavelet packet representation:

Given a wavelet packet structure, a complete tree structured filter bank is considered. Once I find the “best basis” for this application, a fast implementation exists for determining the coefficients with respect to the basis. However, in the “best basis” approach, they do not subdivide every sub band until the last level. The decision of whether to subdivide is made based on a reasonable criterion according to the application (further decomposition implies less temporal resolution). The cost function, which determines the basis selection algorithm, will be a constrained minimization problem. The idea is to minimize the cost due to the bit rate given the filter bank structure, using as a variable the estimated computational complexity at a particular step of the algorithm, limited by the maximum computations permitted. At every stage, a decision is made whether to decompose the sub band further based on this cost function. Another factor that influences this decomposition is the tradeoff in resolution. If it is decomposed further down, it will sacrifice temporal resolution for frequency resolution.

The last level of decomposition has minimum temporal resolution and has the best frequency resolution. The decision on whether to decompose is carried out top-down instead of bottom-up. Following that way, it is possible to evaluate the signal at a better temporal resolution before the decision to decompose. It is proved in this paper that the proposed algorithm yields the “best basis” (minimum cost) for the given computational complexity and range of temporal resolution.

Feature Extraction & Classification:

The extracted wavelet coefficients provide a compact representation that shows the energy distribution of the signal in time and frequency. In order to further reduce the dimensionality of the extracted feature vectors, statistics over the set of the wavelet coefficients are used. That way the statistical characteristics of the “texture” or the “Environmental sound” of the piece can be represented. The distribution of energy in time and frequency for music is different for every environment.

The mean of the absolute value of the coefficients in each sub band provides information about the frequency distribution of the audio signal. The standard deviation of the coefficients in each sub band provides information about the amount of change of the frequency distribution. Ratio of the mean values between adjacent sub bands provides information about the frequency distribution.

Techniques:

Support vector machine (SVM) is based on the principle of structural risk minimization (SRM). Like RBFNN, support vector machines can be used for pattern classification and nonlinear regression. SVM constructs a linear model to estimate the decision function using non-linear class boundaries based on support vectors. If the data are linearly separated, SVM trains linear machines for an optimal hyper plane that separates the data without error and into the maximum distance between the hyper plane and the closest training points. The training points that are closest to the optimal separating hyper plane are called support vectors. Fig.2. shows the architecture of the SVM. SVM maps the input patterns into a higher dimensional feature space through some nonlinear mapping chosen a priori. A linear decision surface is then constructed in this high dimensional feature space. Thus, SVM is a linear classifier in the parameter space, but it becomes a nonlinear classifier as a result of the nonlinear mapping of the space of the input patterns into the high dimensional feature space.
**SVM Principle:**

Support vector machine (SVM) can be used for classifying the obtained data (Burges, 1998). SVM are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. Let us denote a feature vector (termed as pattern) by \( x=(x_1, x_2, \cdots , x_n) \) and its class label by \( y \) such that \( y = \{+1, -1\} \). Therefore, consider the problem of separating the set of \( n \)-training patterns belonging to two classes, \( (x_i, y_i) \), \( x_i \in \mathbb{R}^n \), \( y = \{+1, -1\} \), \( i = 1, 2, \cdots , n \).

A decision function \( g(x) \) that can correctly classify an input pattern \( x \) that is not necessarily from the training set.

**SVM for Linearly Separable Data:**

A linear SVM is used to classify data sets which are linearly separable. The SVM linear classifier tries to maximize the margin between the separating hyperplane. The patterns lying on the maximal margins are called support vectors. Such a hyper plane with maximum margin is called maximum margin hyper plane. In case of linear SVM, the discriminate function is of the form:

\[
g(x) = w^T x + b
\]

such that \( g(x_i) \geq 0 \) for \( y_i = +1 \) and \( g(x_i) < 0 \) for \( y_i = -1 \). In other words, training samples from the two different classes are separated by the hyper plane \( g(x) = w^T x + b = 0 \). SVM finds the hyperplane that causes the largest separation between the decision function values from the two classes. Now the total width between two margins \( 2/\|w\| \), which is to be maximized. Mathematically, this hyperplane can be found by minimizing the following cost function:

\[
j(w) = \frac{1}{2} w^T w
\]

Subject to reparability constraints

- \( g(x_i) \geq +1 \) for \( y_i = +1 \)
- \( g(x_i) \leq -1 \) for \( y_i = -1 \)

Equivalently, these constraints can be re-written more compactly as

\[
y_i(w^T x_i + b) \geq 1; i = 1, 2, \cdots , n
\]

For the linearly separable case, the decision rules defined by an optimal hyperplane separating the binary decision classes are given in the following equation in terms of the support vectors.

\[
Y = \text{sign}
\left(
\sum_{i=1}^{N_s} y_i \alpha_i (x_i, x) + b \right)
\]

where \( Y \) is the outcome, \( y_i \) is the class value of the training example \( x_i \), and represents the inner product. The vector corresponds to an input and the vectors \( x_i, i = 1, \ldots , N_s \), are the support vectors. \( b \) and \( \alpha_i \) are parameters that determine the hyperplane.

**SVM for Linearly Non-separable Data:**

For non-linearly separable data, it maps the data in the input space into a high dimension space \( x \in \mathbb{R}^l \rightarrow \Phi(x) \in \mathbb{R}^H \) with kernel function \( \Phi(x) \), to find the separating hyperplane. A high-dimensional version of Eq.

\[
Y = \text{sign}
\left(
\sum_{i=1}^{N} y_i \alpha_i K(x_i, x) + b \right)
\]

**Fig. 2: Support vector machine**
Determining Support Vectors:

The support vectors are the (transformed) training patterns. The support vectors are (equally) close to hyperplane. The support vectors are training samples that define the optimal separating hyperplane and are the most difficult patterns to classify. In-formally speaking, they are the patterns most informative for the classification task. SVM example to classify a person into two classes: overweighted, not overweighted; two features are pre-defined: weight and height. Each point represents a person. Dark circle and star points denote overweighted and not overweighted respectively. Circles over the points denote support vectors.

Experimental result:

Dataset:

The database for the experiments contains 400 samples which are taken from BBC database. The recordings are categorized into general classes according to common characteristics of the scenes 110 from kitchen noises, 90 from hospital noises, 80 from factory, 120 from meeting sounds and events hearing, pan boiling, steel plate, music player, paper scrap, washing machine, flush, overlapped speech, footsteps, typewriter, dust bin, etc. The recordings are manually labelled and are separated into 1-second fragments. Each sound signal was stored with some properties that are also the initial conditions and criteria for the well-functioning of the algorithm. The sample database is split into training sets and test sets. Then select 80% sounds of each class for the training set. The remaining 20% sounds form the test set. Thus we have taken different proportion of samples based on class dependency in each category as shown in table1.

Table 1: Acoustic database descriptor.

<table>
<thead>
<tr>
<th>Category of environment sound</th>
<th>No. Of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kitchen</td>
<td>27.5%</td>
</tr>
<tr>
<td>Hospital</td>
<td>22.5%</td>
</tr>
<tr>
<td>Factory</td>
<td>20%</td>
</tr>
<tr>
<td>Meeting</td>
<td>30%</td>
</tr>
</tbody>
</table>

Preprocessing:

The signal is sampled at 16kHz and training set has extracted from 1 sec to 5 seconds.

The training data are segmented into fixed length overlapped frame (in this experiment 20 ms frames with 10 ms overlapping is used). Since a 16KHz sampling rate is deployed, 20 ms frames consists of 320 values which are converted into 6 dimension for one frame.

Here 400 clips were used for training data, 40 clips for testing data and each clips must be mono channel.

Conclusion:

This work is a pioneering work in the area of sound classification by focusing on the acoustic events which are naturally produced in a meeting-room environment, hospital environment etc., SVM classifier is chosen as a basic classification technique in this work. The main contribution of this work is an attempt to deal with the problem of classifying acoustic events. The proposed work focused on acoustic events that may take place in meeting-rooms or kitchen and on the preliminary task of classifying isolated sounds. Various environment sounds were used for training and testing. The system showed a performance of 88.5% and the results are significant and satisfactory.

REFERENCES


