Sound Classification using Image Feature Extraction Technique

Khine Zar Thwe
University of Computer Studies, Mandalay(UCSM)
Mandalay, Myanmar
khinezarthwe@ucsm.edu.mm
Mie Mie Thaw
University of Computer Studies, Mandalay(UCSM)
Mandalay, Myanmar
miemiethaw@ucsm.edu.mm

Abstract

This paper presents texture feature extraction methods for sound classification. Nowadays, many researchers interest in combination of digital signal processing and digital image processing fields to get higher efficiency. In this paper, feature extraction methods used in image processing are applied in classification of signal processing. Signals are converted into image format and then features are extracted using bi-directional local binary pattern. Feature vector is constructed using these features and then label the input signal by checking similarly value from known dataset using multi support vector machine classifier. Evaluation is tested on benchmark dataset namely ESC10 Dataset, ESC50 Dataset and UrbanSound8K Dataset.
This paper presents texture feature extraction methods for sound classification. Nowadays, many researchers interest in combination of digital signal processing and digital image processing fields to get higher efficiency. In this paper, feature extraction methods used in image processing are applied in classification of signal processing. Signals are converted into image format and then features are extracted using bi-directional local binary pattern. Feature vector is constructed using these features and then label the input signal by checking similarly value from known dataset using multi support vector machine classifier. Evaluation is tested on benchmark dataset namely ESC10 Dataset, ESC50 Dataset and UrbanSound8K Dataset.

1. Introduction

Audio classification is the process that takes an audio sample as input and gives the corresponding class mark as output. There are many types of machine learning paradigm classifiers that can perform this task. Machine learning can be used to efficiently classify the sound after the source. Feature extraction methods have been the most widely used concept for classifying sound [2]. Audio classification has many applications, such as remote monitoring, home automation, hands free communication, etc.

There are two general types of sound based on the source: generated by humans and not by humans. Sounds made by the man consist of speech and other non-vocal sounds. Non-human sounds include everything else that sounds made of animals and things [3].

Environmental sounds [5, 7] consist of several non-human sounds in everyday life. In recent years, many attempts have been made in recognition of environmental issues [12]. Improvements in classifying visual scenes in recent years are leading researchers to begin to perfect the task of environmental sound classification.

The purpose of classification of audio event is to develop a system capable of achieving human appearance with various listening tasks. Firstly, sound event characteristics differ only in the nature of the speech with different content, duration, and profile. Second, as with the speech that the word can divide into its constituent phonemes, there are no word of dictionary for environmental sound. Finally, there are unstructured environments in which the sound occurs, often the noise and distortion. It is therefore important that Sound Event Classification (SEC) that can solve these challenges are better classified.

The most common approach is to use a system based on one-dimensional frame level functions like Mel Frequency Cepstral Coefficients (MFCC), modeled using Gaussian Model Mixes (GMMs), with time sequencing captured by Hidden Mark's Model (HMM). The disadvantages of using this approach to SEC are, however, two-fold.

Firstly, there is a large variation in the time-frequency structure of different audio events, and this information may not be the best to capture HMM, which requires independence between adjacent observations over time. Secondly, frame-based functions only record audio information in a narrow time frame, and usually represent a full spectrum of frequencies. This causes problems under certain conditions, where features can contain elements of noise and challenge them to separate them.

Therefore, the signal becomes formally two-dimensional using Fourier Transform, and the texture function extraction method is used in the audio classification file. This paper presents the use of the texture extraction method for classifying environmental sound.

There are four portions in this paper. Firstly, the state of the art related with this paper is described. Secondly, the detail step of process is clearly explained. The next portion is evaluation. Previous paper [10] evaluated only some portion of ESC-10 dataset. No all labels are included. In this paper, three datasets are calculated and all labels are considered. Conclusion and further work are finally followed.

2. Related Works

The approaches in [6] generate two projection-based local binary pattern features to capture the texture information of the spectrogram. That paper use gammatone-like spectrogram instead of short-time Fourier Transform (STFT) spectrogram to extract the features because gammatone-like spectrogram approximates the human auditory system's response. Support vector machine (SVM) is used for classification and the accuracy rate is 84%.

In [8] Piczak used a deep convolutional neural network to classify audio events in the UrbanSound8K datasets. Like the input features, the Log-scaled Mel-spectrograms were used. The DCNN architecture consisted of 2 convolutions, followed by 3 fully linked layers. A MaxPooling layer followed the first convolutional layer and a Dropout layer followed the first dense layer. With this architecture, the reported categorical accuracy was 73%.

Dennis et al. [1] have used block-based features to classify the audio signals. They partition the spectrogram image into 9x9 blocks and then for each block they have computed second order and third order central moments. These collections of moments are used as the feature vector and 84% of accuracy is achieved. This approach used image processing techniques, such as grey-scale normalization, dynamic range quantization and mapping, and pixel distribution statics, to generate a representation of the visual information.

In [9], the authors proposed log-Gabor-filters features in addition to MFCCs coefficients to analyze environmental sounds. Support Vector Machine is used for classification. Classification type is one-against-one configuration with the Gaussian kernel. 94.5% accuracy is achieved based on ten class labels.

G. Yu and J. Slotine [11] proposed the method to classify multiple musical instruments by treating the spectrogram as texture image with the feature extraction scheme based on time-frequency matching. The average accuracy is 85.5%. This study analyzed various features dimension and the final feature dimension is regarded as 420 features. Each of 100 audio files is used for testing.

Data are collected using both manual and internet resources because there is not enough dataset in public. Eight type of labels are used in that study. They are violin, cello, piano, harpsichord, trumpet, tuba, flute and drum. In classification accuracy, harpsichord is mostly misclassified with piano and trumpet is misclassified with violin, flute and so on.

3. Processing Steps

3.1. Pre-processing Step

Firstly, input signal processing is partitioned into 0.75second frame length and each frame is checked whether it is silent or non-silent portion. To check it is silent portion, firstly 0.05s (50ms) of input audio signal is fetched and calculate energy. This calculated energy is represented as silent threshold. Energy of each frame is compared with this silent threshold. If energy of frame greater than silent threshold value, this frame is represented as non-silent frame. Silent frame is not needed to do next processes.

Non-silent frame is then partitioned into 0.011s (11ms) window segment. This segment is converted into spectrogram image using Short-time Fourier Transform. Short-time Fourier Transform is described in equation (1).

$$= \sum_{n=0}^{N-1} x_t[n]\omega[n]e^{-2\pi\frac{f}{f_s}n}$$

Where f=kfs/N for k=0, 1, 2. . . N-1, $\omega(.)$ is the Hamming window, and t is the frame index, as multiples of N/fs.

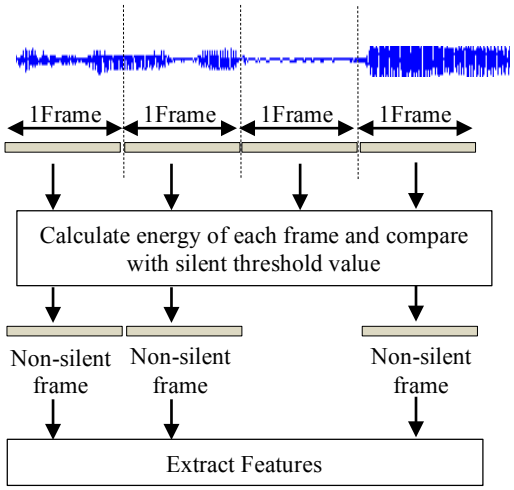


Figure 1. Simple Process of Pre-processing

3.2. Feature Extraction Step

In this step, spectrogram that achieved from pre-processing step is used to extract features. In feature extraction method, bi-directional local binary pattern is used. Bi-directional local binary pattern is modified from traditional local binary pattern.

The original LBP operator identifies pixels of a decimal-digit image, called Local Binary Patterns, which encode the local structure around each pixel. It thus continues as shown in Figure 1: each pixel is compared to its eight neighbors in a quarter of 3x3 by subtracting the value of the central pixel; The resulting strictly negative values are coded with 0 and the others with 1; A binary number is obtained by linking all these binary codes clockwise from the top left and the corresponding decimal value is used for labeling. The derived binary numbers are known as local binary patterns or LBP codes. Local binary pattern equation is described in equation (2).

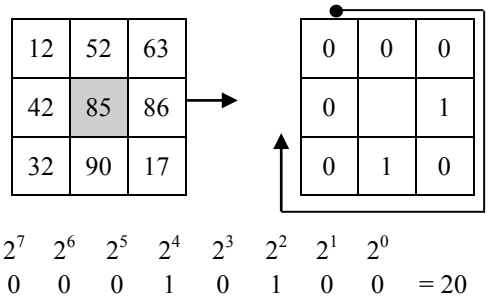
$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p$$

Where

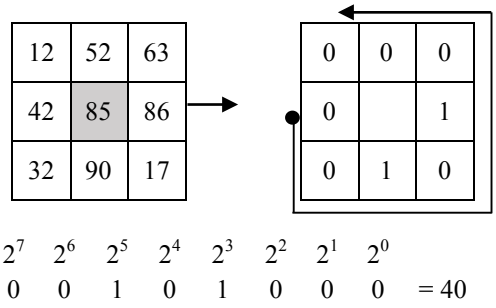
$$s(x) = \begin{cases} 1 & if(x \geq 0) \\ 0 & if(x < 0) \end{cases}$$

g_p are gray values of pixels regularly spaced on circle and g_c is the gray value of the center pixel. Gray values at non-integer positions are obtained by interpolation. P is the neighborhood size and R is the radius.

Conventional local binary pattern uses one direction only such as clockwise direction or anti-clockwise directional. Method used in this paper use both of clockwise and anti-clockwise direction namely bi-directional local binary pattern. Feature length of local binary pattern is 256 (28). In bi-directional local binary pattern has 512 feature length because of (2 x 256). These 512 features are used as feature vector.



This feature vector is passed through classification step to label the input signal.



Process of Local Binary Pattern of Clockwise Direcion

Figure 2. Process of Local Binary Pattern of Counter Clockwise Direction

3.2. Classification Step

Feature vector that resulted from feature extraction step are used as input in classification step. There are mainly three classification type namely supervised classification, unsupervised classification and semi-supervised classification. In this paper, one type of supervised classification step called multi support vector machine method is used.

3.2.1.Support Vector Machine

The Support Vector Machine (SVM) (Mayoraz, E. and Alpaydin, E., 1999; Wang, 2005) is a state-of-the-art classification method introduced in 1992 by Boser, Guyon, and Vapnik. The SVM classier is widely used in bioinformatics due to its high accuracy, ability to deal with high-dimensional data such as gene expression, and flexibility in modeling diverse sources of data.

Support vector machine (SVM) is a powerful tool used in solving either two class or multi class classification problem. In a two-class problem, the input data has to be separated into two different categories wherein each category is assigned a unique class label. A multi class classification problem can be solved by dividing it into multiple two class classification problems and later aggregating the individual results to get the final result of the multi class problem. In this paper one-vs-one support vector machine is used.

SVM can be categorized into non-linear and linear SVM. Linear SVM can be geometrically represented by a line which divides the data space into two different regions thus resulting in classifying the said data which can be assigned two class labels corresponding to the two regions. In the case of non-linear SVM, the input data space can be generalized onto a higher dimensional feature space so called the kernel function. Kernel function can be computed using an inner dot product in the feature space. There is various kernel function such as liner kernel function, polynomial kernel function, Gaussian radial based kernel function and sigmoid kernel function and so on.

3.4. Dataset

ESC-10 dataset [4], ESC-50 [4] dataset, and UrbanSound8K [13] datasets are used in this paper. There are 10 labels in each dataset.

In ESC-50 dataset, there are five categories. Each category has 10 class labels and the label's name are different. The categories names are Animals category, Natural Soundscapes, Human, Interior Sounds, and Urban Sounds. In UrbanSound8K, there are 10 categories and each category have 10 class labels. Label names in each category are the same labels. The label name of each dataset is described in the following tables.

Table 1. LABEL NAMES OF EACH DATASET

Dataset Name	Category Name	Label Names
ESC-10	C1	Dog bark, Rain, Sea waves, Baby Cry, Clock tick, Person sneeze, Helicopter, Chainsaw, Rooster, Fire crackling.
ESC-50	C1	Dog, Rooster, Pig, Cow, Frog, Cat, Hen, Insects, Sheep, Crow.
	C2	Rain, Sea waves, Crackling fire, Crickets, Chirping birds, Water drops, Wind, Pouring water, Toilet flush, Thunderstorm.
	C3	Crying baby, Sneezing, Clapping, Breathing, Coughing, Footsteps, Laughing, Brushing teeth, Snoring, Drinking sipping.
	C4	Door knock, Mouse click, Keyboard typing, Door-wood creaks, Can opening, Washing machine, Vacuum cleaner, Clock alarm, Clock tick, Glass breaking.
	C5	Helicopter, Chainsaw, Siren, Car horn, Engine, Train, Church bells, Airplane, Fireworks, Handsaw.
UrbanSound-8K	C1, C2, C3, C4, C5, C6, C7, C8, C9 C10	Air conditioner, Car Horn, Children Playing, Dog bark, Drilling, Engine idling, Gun shot, Jackhammer, Siren, Street Music.

4. Experimental Results

Classification result, confusion matrix and for each dataset are described as the following.

4.1. Classification Results

In table1, classification results of baseline method and bidirectional local binary pattern is described. Three datasets are used in this paper namely ESC-10, ESC-50 and UrbanSound8K dataset. 10-fold cross validation is used in all three datasets. Therefore, one-tenth of data is used for testing and remaining portion is used for training. Generally, mostly number of maximum classification results are achieved in bidirectional local binary pattern although these are not sharply increased.

Table 2. Classification Result of Bidirectional Local Binary Pattern

Dataset Name	Category Name	LBP	Bidirectional LBP
ESC-10	C 1	74.64%	**74.78 %**
ESC-50	C 1	63.02%	**62.50 %**
	C 2	78.46%	**78.60 %**
	C 3	**66.82%**	65.50 %
	C 4	**67.62%**	67.40 %
	C 5	67.61%	**67.70 %**
UrbanSound8K	C1	**83.37%**	83.10 %
	C 2	78.52%	**78.90 %**
	C 3	81.83%	**81.90 %**
	C 4	**81.89%**	81.80 %
	C 5	79.65%	**79.70 %**
	C 6	77.99%	**78.20 %**
	C 7	**77.89%**	77.80 %
	C 8	82.70%	82.70 %
	C 9	81.50%	81.50 %
	C 10	**78.78%**	78.70 %

4.2 Confusion Matrix

The accuracy of a classification can be estimated by calculating the number of class conflicts that are correctly recognized (true positives), the number of correctly recognized examples that do not belong to the class (true negatives) and the examples incorrectly assigned to the class (false positives)) or which were not recognized as class examples (false negatives). True positives, true negatives, false positives and false negatives can be known from Confusion Matrix.

The Confusion matrix is one of the easiest matrices used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes.

Table3 shows the Confusion Matrix (CM) of esc-10 dataset. The values in the diagonal represents the correctly classified numbers. ESC-10 dataset includes 10 labels. There are 10 rows and 10 columns in this table for each label; dog bark for first row, first column; rain for second row, second column and so on.

In rain labels (second rows), 0.10% (= $\frac{27}{276}$) of rain are misclassified with sea waves. The nature of rain sound and sea waves sound are similar and some rain sound are heavy sounds. This fact can cause the similarity problems with sea waves. In rooster (ninth row) label, 25% (= $\frac{38}{148}$) of rooster are misclassified with baby cry.

In table4, 17% (= $\frac{31}{180}$) of cow labels (fourth rows) is misclassified with sheep label because histogram pattern of cow and sheep are similar. These histogram patterns are described in figure 4. Likewise, in table5, 13% (= $\frac{32}{249}$) of wind label (seventh row) are misclassified with thunderstorm label because one type of sound is included in another portion of sound; the thunderstorm sound includes wind sound. When the sound of wind is heavy windy sound, this sound type is similar with thunderstorm.

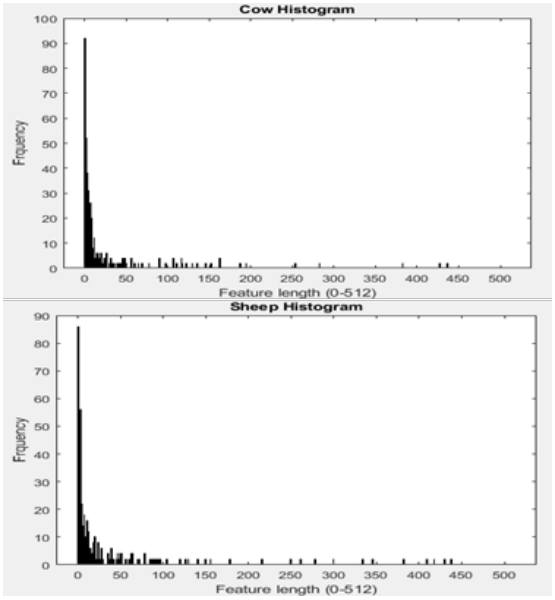


Figure 3. Histogram Features of Cow Audio and Sheep Audio

Table 3. CM of ESC10 Dataset

label	1	2	3	4	5	6	7	8	9	10
1	416		26	1		10				7
2	1	25	5	1	4			2	5	13
3	14		361	22	3	12		4	11	49
4	2		48	204	8	2		1	32	18
5		2	16	8	293	14		7	12	20
6			20	3	3	495		5	1	20
7			1	3	1	3	19	7		
8			17	3	10	8		471		15
9	4		35	12	6	1			191	36
10	11	3	87	5	7	25		4	13	325

Table 4. CM of ESC50 Dataset (C1)

label	1	2	3	4	5	6	7	8	9	10
1	190	9		17	3		26	9	5	1
2	12	25	2	19	14		5	9	2	14
3		2	205	2	3	9	1	6	1	11
4	10	9	1	130	1		16	2	13	7
5		15	6	10	37		8	3	5	8
6			9			207		6	1	19
7	28	8	1	24	8		84	5	1	6
8	5	3	1	8		8	3	184	4	15
9	4	4	1	29	2	2	4	8	86	58
10	3	5	3	10	8	30	5	14	6	94

Table 5. CM of ESC50 Dataset (C2)

label	1	2	3	4	5	6	7	8	9	10
1	75	9	9	3	9		1	1	15	
2	2	81	20	2	16	9	1		41	11
3		29	173	3	15	5			6	
4	11	4	22	94	5	7	8	6	13	
5	1	24	11	1	62	4	4	1	16	7
6		4	7	3		179	26		10	
7		4			2	22	186	3	10	2
8		4	2	2	6	1	4	201	6	6
9	2	36	8	3	5	21	5	3	165	1
10		13	10	1	7		1	3	14	41

Table 6. CM of ESC50 Dataset (C3)

label	1	2	3	4	5	6	7	8	9	10
1	144	6		7	7	20		36		4
2	5	173	1	9	33	4	12	9	3	7
3	4	1	179	16		1	23	3		
4	11	6	18	57	2	31	14	21	4	1
5	10	15		4	160	10		15	3	25
6	14	5		19	2	142	2	32	5	
7	3	14	13	5		10	180	8		
8	25	14		16	4	44	3	111	1	
9	4	9		9	3	12		14	165	
10	1	12			18				1	206

Table 7. CM of ESC50 Dataset (C4)

label	1	2	3	4	5	6	7	8	9	10
1	467		9	2		2		3		19
2	1	47	13	5	1				3	6
3	10		341	15	7	3		7	1	58
4	12	1	37	263	6	4		8	1	40
5	12		16	12	289	7	1	4		10
6	14		3	12		521		5	2	4
7			1	5	4	5	17	8		
8			5	1	8	4		943		5
9	2		32	7	2	1		7	175	31
10	34		92	25	4	13		5	5	288

Table 8. CM of ESC50 Dataset (C5)

label	1	2	3	4	5	6	7	8	9	10
1	467		9	2		2		3		19
2	1	47	13	5	1				3	6
3	10		341	15	7	3		7	1	58
4	12	1	37	263	6	4		8	1	40
5	12		16	12	289	7	1	4		10
6	14		3	12		521		5	2	4
7			1	5	4	5	17	8		
8			5	1	8	4		943		5
9	2		32	7	2	1		7	175	31
10	34		92	25	4	13		5	5	288

Table 9. CM of UrbanSound8K Dataset (C1)

label	1	2	3	4	5	6	7	8	9	10
1	150		2	11	18	9	12	7	6	4
2		227	27		3		5	3		11
3		27	209		4		1	10		1
4	14		1	206	9	9	4	2	15	1
5	15	7	7	9	189	8	10	13	1	4
6	18	4		10	23	32	2	5	2	5
7	10	1	2	2	9	1	205	13	1	2
8	6	6	6	7	10	4	7	207		12
9	19	2		38	6	3	2	1	76	1
10	3	13	1	3	14	6	1	10		222

Table 10. CM of UrbanSound8K Dataset (C2)

label	1	2	3	4	5	6	7	8	9	10
1	367		8	3	1	19		2	14	19
2		93	5	7					3	4
3	5		301	32	3	29	1	13	16	40
4	1		36	252	13	6		3	31	23
5	1		9	9	360	1			17	10
6			22	1		440	2	10	4	11
7	1		6	2		6	39	3	5	3
8	2		5	2		14		473	1	17
9	1		11	13	1	9		2	483	29
10	3	1	51	16	3	13		7	22	285

Table 11. CM of UrbanSound8K Dataset (C3)

label	1	2	3	4	5	6	7	8	9	10
1	204	21	13			1			4	7
2	26	177			2		8		12	6
3	5	2	205		2	15			19	
4	3	1		167	35	7	2	1	11	5
5	2	9		16	168	4	6		5	4
6	10	1	10	2	4	121		5	24	1
7	2	23			10		176			38
8	2	2	8	1		6		180	8	
9	9	5	1		10	4	2	1	229	1
10	2	7	3		1	3	33		4	194

Table 12. CM of UrbanSound8K Dataset (C4)

label	1	2	3	4	5	6	7	8	9	10
1	467	1	11	3	4	1		8		10
2	2	145	27	2	20	4		16	5	6
3	33	12	316	20	2			20	12	39
4	20		73	184	8	1		1	22	18
5	12	16	16	3	516			13	5	10
6	4		4	1		514		10	2	10
7			2		3	2	24	15		5
8	6	5	16		25	7		376	2	13
9		5	6	9	1				738	38
10	18	4	44	8	7	8		23	27	353

Table 13. CM of UrbanSound8K Dataset (C5)

label	1	2	3	4	5	6	7	8	9	10
1	457		8	6	10	2			4	29
2	2	290	29	11	3			1	2	25
3	19	27	266	28	23	4		1	6	42
4	18	9	42	211	21	8		1	14	35
5	2	1	23	26	348	6		1	3	17
6	5		20	6	18	424		1		12
7	1		3	7	8	2	24			1
8			2	3	5	4		547		3
9	7	12	37	11	3				240	6
10	2	26	51	18	5	2			12	341

Table 14. CM of UrbanSound8K Dataset (C6)

label	1	2	3	4	5	6	7	8	9	10
1	398		14		1	11				41
2	2	41	9	1					4	15
3	3	3	305	30	1	18	1	3	7	63
4	2	1	52	237	14	19	1	8	2	22
5	2		5	24	341	3		2	3	11
6	7		8	16	2	403		5		22
7		2	2	3	1	5	27	1	3	4
8	5			10	1	17		277		2
9	4	2	23	7		1	2		157	23
10	9		44	24	2	34		2	8	328

Table 15. CM of UrbanSound8K Dataset (C7)

label	1	2	3	4	5	6	7	8	9	10
1	426		20	11	5	4		3	6	2
2		4	20	13	1	1		1		22
3	38	2	286	20	15	13		12	6	39
4	20	1	32	251	11	11	3	2	17	20
5	22		5	7	475	6		1	9	6
6	9		32	18	9	435		3		6
7	1		4	4	5	2	79	3		2
8	8		10	1	8	6		316	1	6
9	11		35	23	7	4			198	25
10	4	1	69	4	16	4		11	21	327

Table 16. CM of UrbanSound8K Dataset (C8)

label	1	2	3	4	5	6	7	8	9	10
1	454		11	4	3	7				8
2		43	4	6					5	6
3	8		286	38	2	10	5		2	54
4	8	2	66	224	17	11	1	1	1	15
5	11		12	16	551	5		5	2	12
6	8		20	4	4	610		2		5
7			6	7	8		6			2
8	1		1	4	5	7		290		1
9	1	2	20	2					203	19
10	20	2	63	16	9	13		1	7	294

Table 17. CM of UrbanSound8K Dataset (C9)

label	1	2	3	4	5	6	7	8	9	10
1	402		30	4	7	1				17
2	1	22	5	8	4				1	17
3	28		290	28	2	1				49
4	21		46	243	14	2			6	21
5	12		13	11	410	4		1		7
6			4	4	9	273	4		7	8
7	2		1	12	8	9	20			
8	2		25		2			411		1
9	3		14	14		2			235	20
10	7	1	63	3	3	7			9	359

Table 18. CM of UrbanSound8K Dataset (C10)

label	1	2	3	4	5	6	7	8	9	10
1	420		17	2	5	5			6	6
2		40	19	19	1			1	3	10
3	18	2	215	48	10	12			19	40
4	6	2	52	222	9	12		2	20	33
5	6		7	12	304	11	1	1	26	16
6	31	1	7	6	5	331			2	6
7	1		3		4	2	21	3		
8			4	2	1	7		744		2
9	3		22	20	18	6			188	12
10	25		41	20	9	5		1	22	283

4.3 Classification Performance

Multi-class classification measurement is described as the following equations respectively. True positives, true negatives, false positives and false negatives can be known from Confusion Matrix.

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It is described in equation (3).

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

Recall or sensitivity (true positive rate) is the proportion of positive cases that are properly identified and can be calculated using equation: (4).

$$Recall\ (Sensitivity) = \frac{TP}{TP+FN} \qquad (4)$$

Table 19. Precision and Recall

Dataset Name	Category Name	Precision	Recall
ESC-10	C 1	72.60 %	71.03 %
ESC-50	C 1	63.29 %	61.98 %
	C 2	79.30 %	78.30 %
	C 3	61.80 %	60.40 %
	C 4	67.90 %	64.10 %
	C 5	67.70 %	66.70 %
UrbanSound8K	C 1	84.90 %	74.80 %
	C 2	81.80 %	72.50 %
	C 3	84.30 %	79.50 %
	C 4	82.40 %	75.90 %
	C 5	81.20 %	76.00 %
	C 6	80.90 %	74.30 %
	C 7	76.60 %	70.90 %
	C 8	79.90 %	74.50 %
	C 9	84.00 %	73.70 %
	C 10	78.70 %	71.90 %

Specificity or true negative rate was defined as the proportion of negatives cases which are classified correctly, and is calculated using the equation (5).

$$Specificity = \frac{TN}{TN+FP} \qquad (5)$$

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

$$F-score = \frac{2*TP}{(2*TP+FP+FN)} \qquad (6)$$

Table 20. Specificity and F-score

Dataset Name	Category Name	Specificity	F-score
ESC-10	C 1	97.18 %	71.44 %
ESC-50	C 1	95.82 %	62.34 %
	C 2	97.60 %	78.50 %
	C 3	96.20 %	60.50 %
	C 4	96.30 %	65.40 %
	C 5	96.40 %	67.00 %
UrbanSound 8K	C 1	98.10 %	78.10 %
	C 2	97.60 %	75.70 %
	C 3	97.90 %	81.40 %
	C 4	98.00 %	77.90 %
	C 5	97.70 %	77.70 %
	C 6	97.50 %	76.80 %

	C 7	97.50 %	71.90 %
	C 8	98.00 %	76.30 %
	C 9	97.90 %	76.60 %
	C 10	97.60 %	74.10 %

Error rate (ERR) is calculated as the number of all incorrect predictions divided by the total number of the dataset. The best error rate is 0.0, whereas the worst is 1.0. Error rate is described in equation (7).

$$Error\ Rate = \ (1 - Specificity) \qquad (7)$$

Coefficient, symbolized by the lower-case Greek letter, k (7) is a robust statistic useful for either inter-rater or intra-rater reliability testing. Similar to correlation coefficients, it can range from −1 to +1, where 0 (0%) represents the amount of agreement that can be expected from random chance, and 1 (100%) represents perfect agreement between the raters. This equation is described in following equation.

$$Kappa = \frac{(Observed\ Agreement - Expected\ Agreement)}{(1 - Expected\ Agreement)} \qquad (8)$$

$$Oberved\ Agreement = \sum_{i=1}^{k} p_{ii} \qquad (9)$$

$$Expected\ Agreement = \sum_{i=1}^{k} p_{i+} * \ p_{+i} \qquad (10)$$

Table 21. Error Rate and Kappa Coefficient

Dataset Name	Category Name	Error Rate	Kappa Coefficient
ESC-10	C 1	2.82 %	71.74 %
ESC-50	C 1	4.18 %	58.22 %
	C 2	2.40 %	76.14 %
	C 3	3.80 %	61.24 %
	C 4	3.70 %	63.29 %
	C 5	3.60 %	64.09 %
UrbanSound 8K	C 1	1.90 %	80.31 %
	C 2	2.40 %	75.82 %
	C 3	2.10 %	79.48 %
	C 4	2.00 %	79.34 %
	C 5	2.30 %	77.12 %
	C 6	2.50 %	75.16 %
	C 7	2.50 %	74.69 %
	C 8	2.00 %	80.12 %
	C 9	2.10 %	78.90 %
	C 10	2.40 %	75.49 %

5. Conclusion and Future Work

This paper presents audio classification using texture feature extraction called bidirectional local binary pattern. Multi SVMs is used to classify the label of input audio. Evaluation is tested on three bench mark datasets called ESC-10 dataset, ESC-50 dataset and UrbanSound8K datasets. The further extension of this paper is to assistant noise present and to detect and classify real time condition.

References

[1] Dennis, J., Tran, H.D. and Li, H., 2011. Spectrogram image feature for sound event classification in mismatched conditions. IEEE signal processing letters, 18(2), pp.130-133.

[2] Gencoglu, O., Virtanen, T. and Huttunen, H., 2014, September. Recognition of acoustic events using deep neural networks. In Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European (pp. 506-510). IEEE.

[3] McLoughlin, I., Zhang, H., Xie, Z., Song, Y. and Xiao, W., 2015. Robust sound event classification using deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(3), pp.540-552.

[4] Piczak, Karol J. "ESC: Dataset for environmental sound classification." In Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015-1018. ACM, 2015.

[5] Kalantarian, Haik, Nabil Alshurafa, Mohammad Pourhomayoun, Shruti Sarin, Tuan Le, and Majid Sarrafzadeh. "Spectrogram-based audio classification of nutrition intake." In Healthcare Innovation Conference (HIC), 2014 IEEE, pp. 161-164. IEEE, 2014.

[6] Ren, J., Jiang, X., Yuan, J. and Magnenat-Thalmann, N., 2017. Sound-Event Classification Using Robust Texture Features for Robot Hearing. IEEE Trans. Multimedia, 19(3), pp.447-458.

[7] Lu, Fuxiang, and Jun Huang. "An improved local binary pattern operator for texture classification." In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pp. 1308-1311. IEEE, 2016.

[8] Piczak, K.J., 2015, September. Environmental sound classification with convolutional neural networks. In Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on (pp. 1-6). IEEE.

[9] Sameh, S. and Lachiri, Z., Using Spectro-Temporal Features for Environmental Sounds Recognition.

[10] Thwe, K.Z., 2017, July. Sound event classification using bidirectional local binary pattern. In Signal

Processing and Communication (ICSPC), 2017 International Conference on (pp. 501-504). IEEE.

[11] Yu, G. and Slotine, J.J., 2009, April. Audio classification from time-frequency texture. In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1677-1680). IEEE.

[12] Zhang, Z. and Schuller, B., 2012, March. Semi-supervised learning helps in sound event classification. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on (pp. 333-336). IEEE.

[13] https://serv.cusp.nyu.edu/projects/urbansounddataset/urbansound8k.html