International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Acoustic Scene Classification by using Combination of MODWPT and Spectral Features

Mie Mie Oo, Lwin Lwin Oo

University of Computer Studies (UCSM), Mandalay, Myanmar

How to cite this paper: Mie Mie Oo | Lwin Lwin Oo "Acoustic Scene Classification by using Combination of MODWPT and Spectral Features" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.2518-2522, **https://doi.org/10.31142/ijtsrd27992**

IJTSRD27992

ABSTRACT

Acoustic Scene Classification (ASC) is classified audio signals to imply about the context of the recorded environment. Audio scene includes a mixture of background sound and a variety of sound events. In this paper, we present the combination of maximal overlap wavelet packet transform (MODWPT) level 5 and six sets of time domain and frequency domain features are energy entropy, short time energy, spectral roll off, spectral centroid, spectral flux and zero crossing rate over statistic values average and standard deviation. We used DCASE Challenge 2016 dataset to show the properties of machine learning classifiers. There are several classifiers to address the ASC task. We compare the properties of different classifiers: K-nearest neighbors (KNN), Support Vector Machine (SVM), and Ensembles Bagged Trees by using combining wavelet and spectral features. The best of classification methodology and feature extraction are essential for ASC task. In this system, we extract at level 5, MODWPT energy 32, relative energy 32 and statistic values 6 from the audio signal and then extracted feature is applied in different classifiers.

KEYWORDS: *Acoustic Scene Classification; DCASE 2016; K-nearest neighbors (KNN); Support Vector Machine (SVM); and Ensembles Bagged Trees; MODWPT; Statistic values*

I. INTRODUCTION

Acoustic scene classification aims to recognize the environmental sounds that occur for a period of time.

Many approaches have been proposed for acoustic scene classification including feature representation, classification models, and post-processing [1]. These environments could be indoor (home, office, library etc.), outdoor (busy-street, forest, beach etc.), and a moving vehicle (car, bus, train etc.). Audio scene classification is a complex problem due to the wide variety of individual sound events occurring in an audio scene while only few of them give some information about the scene [2].

The proposed system uses the combination of maximal overlap wavelet packet transform (MODWPT) level 5 and spectral features over statistic values. The preprocessing of the input audio file is pre-processed by full 44,100 Hz sampling frequency. The original and the divided segments all were decomposed to the MODWPT bases in five decomposition levels. At level 5 (MODWPT), the feature 32 energy (E), the relative energy (RE) of feature 32, and spectral features are calculated average and standard deviation as statistics six-values. In total 70 features are obtained from the pre-processed audio signal. Then these 70 features are classified by using machine learning technique. Detection and Classification of Acoustic scenes and Events 2016 dataset is used to extant the properties of the proposed feature.

This system employs the combination of maximal overlap wavelet packet transform and six sets of time domain and frequency domain features over compute the statistic values. We evaluated the classification accuracy in the k-fold cross-validation structure. For k-fold cross validation, we evaluate classification accuracy for 4-fold cross validation. Comparison of (MODWPT) level 3feature set and statistic value, level 4 feature set and statistic value, level 5 feature set and statistic value, level 6 feature set and statistic value, level 7 feature set and statistic value.

The classification accuracy of the features we propose meets acceptable conditions. However, under certain conditions, all acoustic scenes are correctly marked; some scene audio is incorrectly flagged in some cases. This paper consists of five sections. The introduction and literature reviews are presented in section I and section II. Proposed Methodology is presented in section III. Experimental results and conclusion are presented in section IV and section V respectively.

II. RELATED WORK

Acoustic scene classification (ASC) is an important problem of the computational auditory scene analysis. Solving this problem will allow the device to recognize the surrounding environment through the sound it captures, enabling a wide range of applications such as surveillance, robot navigation and context aware services [3]. Multi-width frequency-delta data augmentation which uses static mel-spectrogram as well as frequency-delta features as individual examples with same labels for the network input, and the experimental result shows that this method significantly improves the performance [4].

A novel technique for diagnosis of epileptic seizures based on non-linear entropy features extracted from maximal overlap discrete wavelet packet transform (MODWPT) of EEG signals. Discriminative features are selected by a t-test criterion and used for the classification with two different classifiers [5]. The real-time estimation of the root mean square (RMS) values, primary power quantities (active, total apparent, non-active power, and power factor), and distortion power using the maximal overlap wavelet packet transform (MODWPT) [6].

Non-linear wavelet features such as entropy based features and energy based features have been employed in recent studies [7].

Six electroencephalographic (EEG) and two electro-oculographic (EOG) channels were used in this study. The maximum overlap discrete wavelet transform (MODWT) with the multi-resolution Analysis is applied to extract relevant features from EEG and EOG signals [8]. The characteristics of each audio data must be selected and extracted to find a valid feature set. Most audio analysis studies use values that represent the characteristics of the audio [9].

An acoustic scene classification system based on block based MFCC features and few traditional audio features. The number of extracted features is determined by the number of filter banks used in MFCC feature extraction [10]. Deep learning method adopted from the field of computer vision research. Convolutional neural networks are employed to solve the problem of audio-based scene classification. Specifically, the classifier is built using the intra-network network architecture. In the feature extractor, mel frequency spectral coefficient (MFCC) is used as the input vector for the classifier [11].

Discriminating spatio-temporal models are trained to classify the environment through deep convolutional neural networks. This system experiment with four distinct audio information increases (deformations), leading in five augmentation sets. Each transformation is applied directly to the audio signal before being converted to the network input representation. The extended set of deformations and results are time stretch, pitch shift, pitch shift, background noise, and dynamic range compression. [12].

Three feature selection algorithms for the aggregation of acoustic and visual functions and acoustic scene classification. In addition, comparison was made using six classifiers to obtain the optimal classification system. The classification accuracy is 87.44%, which is the fourth best of all non-neural network based systems. [13].

Frame level statistics supplied to the spectrum function of the support vector machine. Furthermore, score level fusion provides better results than functional level fusion. As a result of the classification, the accuracy was about 17% and 9% relatively higher compared to the baseline system for dataset development and evaluation. [14]. Frame level statistics supplied to the spectrum function of the support vector machine. Furthermore, score level fusion provides better results than functional level fusion. As a result of the classification, the accuracy was about 17% and 9% relatively high compared to the baseline system of dataset development and evaluation. [15].

III. PROPOSED METHODOLOG

Acoustic Scene classification is one of the challenging task in digital signal processing. The main steps of Acoustic Scene classification are: Audio pre-processing, feature extraction and classification. In pre-processing steps, the input audio file is sampling and windowing to get the modified audio signal for feature extraction. In feature extraction, maximal overlap wavelet packet transform (MODWPT) in level 5, energy (E) 25 is 32 features, relative energy (RE) 25 is 32 features. Therefore, 64 features are extracted from (MODWPT) and then spectral features are energy entropy, short time energy, spectral roll off, spectral centroid, spectral flux and zero crossing rates over six statistic values such as average and standard deviation. Hence the length of the proposed feature is 70. The overview of the Acoustic Scene Classification is shown in figure1.

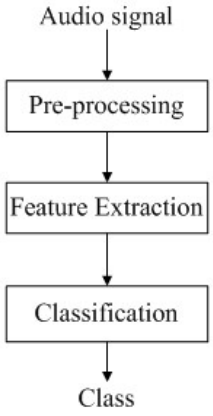

Fig. 1.Overview of Acoustic Scene Classification

A. Maximal Overlap Discrete Wavelet Packet Transform (MODWPT) Extraction

To the best of our knowledge, this is the first study applying MODWPT in acoustic scene classification. Decomposing the signal to the MODWPT bases in each level can be performed by a two-band (high and low-pass) filter bank. The decimation (down-sampling) step is excluded in the process of MODWPT method comparing to the standard discrete wavelet packet transform (DWPT) in which the decimation with the factor of 2 is always applied. Perfect reconstruction of the original signal is achievable by applying the reverse composition filters. The following formula describes a two-band decomposition of the signal in the j[th] level.

$$S_j^{2z+1}(k) = \frac{1}{\sqrt{2}} \sum_{n=-\infty}^{n=\infty} h(n) S_{j-1}^{z}(k-n) \quad (1)$$

$$S_j^{2z}(k) = \frac{1}{\sqrt{2}} \sum_{n=-\infty}^{n=\infty} g(n) S_{j-1}^{z}(k-n) \quad (2)$$

Where $s^0{}_0$ means the original signal z = 2m (m € N; m≤ 2j - 1-1; scale of j) is the node number, m ≤ 2j-1-1; for z=0; S0j (k) corresponds to the packet coefficients in the lower frequency band of the j[th] scale, while for (z≠0), $S^z{}_j$ (k) is defined as the coefficients of packet in the higher frequency band of the j[th] scale. The properties of g and h, as the scaling and wavelet filters, respectively, are as following:

$$\sum_{-\infty}^{\infty} g(n) = \sqrt{2}, \sum_{-\infty}^{\infty} g^2(n) = \sqrt{2} \sum_{-\infty}^{\infty} g(n)h(n) = 0 \quad (3)$$

$$\sum_{-\infty}^{\infty} h(n) = 0, \sum_{-\infty}^{\infty} h^2(n) = 0, \sum_{-\infty}^{\infty} g(n)h(n) = 0 \quad (4)$$

The Sym4 was selected as the best performing mother wavelet based on the classification performance in this algorithm.

B. Time and Frequency domain features

Six sets of time domain and frequency domain features are energy entropy, short time energy, spectral roll off, spectral centroid, spectral flux and zero crossing rate over statistic values average and standard deviation.

1. Energy Entropy

The short-term entropy of energy can be interpreted as a measure of abrupt changes in the energy level of an audio signal. The entropy, H(i) of the sequence ej is computed according to the equation:

$$\mathrm{H(i)} = -\sum_{j=1}^{K} \mathrm{e}_j \log_2(e_j) \qquad (5)$$

2. Short Time Energy

xi(n), n = 1, . . . , WL be the sequence of audio samples of the ith frame, where WL is the length of the frame. The short-term energy is computed according to the equation:

$$E(i) = \sum_{n=1}^{W_L} \left| x_i(n) \right|^2 \qquad (6)$$

3. Spectral roll off

This feature is defined as the frequency below which a certain percentage, if the mth DFT coefficient corresponds to the spectral roll off of the ith frame, then it satisfies the following equation:

$$\sum_{k=1}^{m} X_i(k) = C \sum_{k=1}^{W_{f_L}} X_i(k) \qquad (7)$$

4. spectral centroid

The spectral centroid is the center of 'gravity' of the spectrum. The value of spectral centroid, C_i, of the i^{th} audio frame is defined as:

$$C_i = \frac{\sum_{k=1}^{wf_L} kX_i(k)}{\sum_{k=1}^{Wf_L} X_i(k)} \qquad (8)$$

5. Spectral Flux

Spectral flux measures the spectral change between two successive frames and is computed as the squared difference between the normalized magnitudes of the spectra of the two successive short-term windows:

$$Fl_{(i,i-1)} = \sum_{k=1}^{Wf_L} ((EN_i(k) - EN_{i-1}(k)))^2 \qquad (9)$$

6. Zero Crossing Rate

The Zero-Crossing Rate (ZCR) of an audio frame is the rate of sign-changes of the signal during the frame. The ZCR is defined according to the following equation:

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} \left| \mathrm{sgn}[x_i(n)] - \mathrm{sgn}[x_i(n-1)] \right| \qquad (10)$$

C. Machine Learning Classification Algorithms

Classification is a supervised data mining technique that assigns labels to a collection of data in order to get more accurate predictions and analysis. The ASC is the task to assigns label to audio data to know the label of audio by using trained classifier. The labels for unknown audio data may different according to the application domain. In this proposed work, K-nearest neighbors (KNN), Support Vector Machine (SVM), Decision Tree (ID3) plane.

D. K-nearest neighbors (KNN)

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure distance functions. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor. The Euclidean distance between X = [x1, x2, x3,..,xn] and Y = [y1, y2, y3,...,yn], D(X,Y) is defined as:

$$D(X,Y) = \sqrt{\sum_{n=1}^{N} ((xi - yi)^{\wedge 2})} \qquad (11)$$

E. Support Vector Machines

SVM also used in regression for some kind of application and problems but rarely used for regression. There are two kind of SVM: binary class SVM and multi-class SVM. The main purpose of SVM is to find the separation line to classify the label of the data. The two ways to separate the data according to their labels are: hyper plane based approach and kernel based approach. According to the type of kernel, there are many type of SVM classifier such as Liner SVM, Quadratic SVM, Cubic SVM and Gaussian SVM.

F. Decision Tree

Decision Tree (DT) is a tree where the root and each internal node are labeled with a question. Decision trees have been around for a long time and also known to suffer from bias and variance. The DT needs to consider issue in over fitting, rectangular partition and pruning while construction of DT.

G. DCASE 2016Dataset

TUT Acoustic scenes 2016 dataset will be used for the task. The dataset consists of recordings from various acoustic scenes, all having distinct recording locations. For each recording location, 3-5 minute long audio recording was captured. The original recordings were then split into 30-second segments for the challenge. The 15 labels of the DCASE dataset are: bus, Cafe / Restaurant, Car, City center, Forest path, Grocery store, Home, Lakeside beach, Library, Metro station, Office, Residential area, Train, Tram, Urban park.

IV. EXPERIMENTAL RESUTLS

In this system, the evaluation of proposed feature is perform by measuring the average classification accuracy in the structure of k-fold cross validation. In k-fold cross validation, the validation is performed for k times. For each validation, the dataset is divided into k subsets, k-1 subsets are used for training and the remaining one is used for testing. The average classification accuracy is calculated over these k validations. The value of k is 4. The average classification accuracy is calculated over this four validation time. D1, D2, D3 and D4 are the randomly divided subset of the dataset.

A. Results for machine learning technique

Proposed system shows its properties and advantages in DCASE 2016 dataset for four-fold cross validation.

TABLE I. Comparison of Average Classification Accuracy for MODWPT Feature at (Level 5) With Different Classifiers

Classifier Name	Classification Accuracy
Ensemble(Bagged Trees)	62.7%
KNN	55.6%
SVM	60.90%

TABLE II Classification Accuracy for Combining Features (MODWPT) Level 5 and Statistics Features of SVM Classifer on Decase 2016 Dataset (4-FOLD)

Kernel function	Classification Accuracy
Linear	63.2%
Quadratic	73.5%
Cubic	77.6%
Fine Gaussian	45.8%
Medium Gaussian	65.5%
Coarse Gaussian	47.8%

TABLE III Classification Accuracy for Combining Features (MODWPt) Level 6 and Statistics Features of SVM Classifer on Decase 2016 Dataset (4-FOLD)

Kernel function	Classification Accuracy
Linear	66.2%
Quadratic	71.2%
Cubic	73.5%
Fine Gaussian	45.8%
Medium Gaussian	67.5%
Coarse Gaussian	46.4%

TABLE IV Classification Accuracy for Combining Features (MODWPT) Level 7 and Statistics Features of SVM Classifer on Decase 2016 Dataset (4-FOLD)

Kernel function	Classification Accuracy
Linear	67.2%
Quadratic	73.2%
Cubic	74.5%
Fine Gaussian	45.8%
Medium Gaussian	66.5%
Coarse Gaussian	44.4%

In Table I, II, III, IV average classification accuracy in the structure of 4-fold cross validation. Combining two features MODWPT in level 5 and statistic values for Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian, Coarse Gaussian. These SVM kernel functions cubic SVM leads to the better classification accuracy.

TABLE V. Classification Accuracy for Combining Features (MODWPT) Level 5 and Statistics Features of SVM Classifer on Decase 2016 Dataset (10-FOLD)

Kernel function	Classification Accuracy
Linear	66.6%
Quadratic	74.7%
Cubic	80.1%
Fine Gaussian	48.3%
Medium Gaussian	68.6%
Coarse Gaussian	48.1%

TABLE VI Classification Accuracy for Combining Features (MODWPT) Level 6 and Statistics Features of SVM Classifer on Decase 2016 Dataset (10-FOLD)

Kernel function	Classification Accuracy
Linear	65.5%
Quadratic	72.5%
Cubic	77.9%
Fine Gaussian	45.2%
Medium Gaussian	68.3%
Coarse Gaussian	43.4%

TABLE VII Classification Accuracy for Combining Features (MODWPt) Level 7and Statistics Features of SVM Classifer on Decase 2016 Dataset (10-FOLD)

Kernel function	Classification Accuracy
Linear	64.8%
Quadratic	70.4%
Cubic	75.2%
Fine Gaussian	38.6%
Medium Gaussian	67.3%
Coarse Gaussian	42.1%

In Table V, VI, VII 10-fold cross validation, Cubic SVM is highest classification accuracy among the SVM classifier. Comparison of 4-fold and 10-fold, Classification accuracy 10-fold is higher than 4- fold.

V. CONCLUSION

In this paper, the combination of MODWPT at level 5, 6, 7 Energy E and Relative Energy RE and spectral feature computes statistic values are extracted from audio file. In level 5, E (25) is 32, RE (25) is 32 and statistic values 6. Total feature set is 70. For level 6, feature set is 134. In level 7, feature set is 262. An overall feature of information used to represent a typical acoustic scene of an audio signal to improve the classification accuracy of the acoustic scene. In level 5, the classification results of the proposed feature reach 80.1% for cubic SVM (10- fold) cross-validation. This accuracy is higher than other kernel function. In k-fold cross-validation, the classification accuracy of the proposed features did not change significantly. By combining MODWPT and spectral features, the proposed features are successfully classified as acoustic scene labels, except for certain conditions. Although the suggested feature has an acceptable rating accuracy in the rating of audio scenes, we need to learn other image and signal processing method in order to competently signify audio information.

REFERENCES

[1] S. H., Bae, I., Choi, and Kim, N. S., 2016, September. Acoustic scene classification using parallel combination of LSTM and CNN. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016) (pp. 11-15).

[2] Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M. and Plumbley, M.D., 2015. Detection and classification of acoustic scenes and events. IEEE Transactions on Multimedia, 17(10), pp.1733-1746.

[3] Phan, H., Hertel, L., Maass, M., Koch, P. and Mertins, A., 2016. CNN-LTE: a class of 1-X pooling convolutional

neural networks on label tree embeddings for audio scene recognition. arXiv preprint arXiv:1607.02303.

[4] Han, Y. and Lee, K., 2016. Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events.

[5] Ahmadi, A., Tafakori, S., Shalchyan, V. and Daliri, M. R., 2017, October. Epileptic seizure classification using novel entropy features applied on maximal overlap discrete wavelet packet transform of EEG signals. In 2017 7th International Conference on Computer and Knowledge Engineering (ICCKE) (pp. 390-395). IEEE.

[6] Alves, D. K., Costa, F. B., de Araujo Ribeiro, R. L., de Sousa Neto, C.M. and Rocha, T.D.O.A., 2016. Real-time power measurement using the maximal overlap discrete wavelet-packet transform. IEEE Transactions on Industrial Electronics, 64(4), pp.3177-3187.

[7] Alves, D. K., Costa, F.B., de Araujo Ribeiro, R.L., de Sousa Neto, C. M. and Rocha, T.D.O.A., 2016. Real-time power measurement using the maximal overlap discrete wavelet-packet transform. IEEE Transactions on Industrial Electronics, 64(4), pp.3177-3187.

[8] Khalighi, S., Sousa, T., Oliveira, D., Pires, G. and Nunes, U., 2011, August. Efficient feature selection for sleep staging based on maximal overlap discrete wavelet transform and SVM. In 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 3306-3309). IEEE.

[9] Jondya, A. G. and Iswanto, B. H., 2017. Indonesian's Traditional Music Clustering Based on Audio Features. Procedia computer science, 116, pp.174-181.

[10] Ghodasara, V., Waldekar, S., Paul, D. and Saha, G., 2016. Acoustic Scene Classification Using Block Based MFCC Features. Detection and Classification of Acoustic Scenes and Events.

[11] Santoso, A., Wang, C. Y. and Wang, J. C., 2016. Acoustic scene classification using network-in-network based convolutional neural network. DCASE2016 Challenge, Tech. Rep.

[12] Salamon, J. and Bello, J.P., 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3), pp.279-283.

[13] Xie, J. and Zhu, M., 2019. Investigation of acoustic and visual features for acoustic scene classification. Expert Systems with Applications, 126, pp.20-29.

[14] Waldekar, S. and Saha, G., 2018. Classification of audio scenes with novel features in a fused system framework. Digital Signal Processing, 75, pp.71-82.

[15] Souli, S. and Lachiri, Z., 2018. Audio sounds classification using scattering features and support vectors machines for medical surveillance. Applied Acoustics, 130, pp.270-282.