electronics

MDPI

Article

Automatic Scene Recognition through Acoustic Classification for Behavioral Robotics

Sumair Aziz [1], Muhammad Awais [2,*], Tallha Akram [2], Umar Khan [1], Musaed Alhussein [3] and Khursheed Aurangzeb [3,*]

1 Department of Electronic Engineering, University of Engineering and Technology Taxila, Taxila 47080, Pakistan; sumair.aziz@uettaxila.edu.pk (S.A.); umar.khan@uettaxila.edu.pk (U.K.)
2 Department of Electrical and Computer Engineering, COMSATS University Islamabad—Wah Campus, Wah Cantt 47040, Pakistan; tallha@ciitwah.edu.pk
3 Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia; musaed@ccis.ksu.edu.sa
* Correspondence: muhammadawais@ciitwah.edu.pk (M.A.); kaurangzeb@ksu.edu.sa (K.A.); Tel.: +92-346-5316595 (M.A.)

Received: 27 March 2019; Accepted: 23 April 2019; Published: 30 April 2019

Abstract: Classification of complex acoustic scenes under real time scenarios is an active domain which has engaged several researchers lately form the machine learning community. A variety of techniques have been proposed for acoustic patterns or scene classification including natural soundscapes such as rain/thunder, and urban soundscapes such as restaurants/streets, etc. In this work, we present a framework for automatic acoustic classification for behavioral robotics. Motivated by several texture classification algorithms used in computer vision, a modified feature descriptor for sound is proposed which incorporates a combination of 1-D local ternary patterns (1D-LTP) and baseline method Mel-frequency cepstral coefficients (MFCC). The extracted feature vector is later classified using a multi-class support vector machine (SVM), which is selected as a base classifier. The proposed method is validated on two standard benchmark datasets i.e., DCASE and RWCP and achieves accuracies of 97.38% and 94.10%, respectively. A comparative analysis demonstrates that the proposed scheme performs exceptionally well compared to other feature descriptors.

Keywords: feature extraction; sound classification; support vector machine; sound processing; robotics; MFCC

1. Introduction

Robotics is the branch of artificial intelligence which is concerned with designing robots that can perform tasks and interact with the environment, without the aid of human intervention. Although the mechanical control technology of robots has been remarkably well developed in recent years. The ability of robots to perceive and analyse their surrounding environment, especially the auditory scenes still requires a significant research effort. Acoustic-based classification complements the vision based classification in a number of ways. First, considering the field of view, microphones are more nearly omni-directional than even wide-angle camera lenses. Second, audio signals require a significantly smaller bandwidth and low processing power. Third, acoustic classification is more reliable as the parameters of image/video processing algorithms are affected by variations in light intensity, thus, increasing the probability of false

alarms. Detection and classification of acoustic scenes can help to facilitate the human-robot interaction and increase the application domain of behavioral and assistive robotics.

One of the key aspects of designing an acoustic classification system is the selection of proper signal features that could achieve an effective discrimination between different sound signals. Sounds coming from a general environment are considered neither music nor speech, but a collection of some audio signals that resemble noise signals. While sufficient research has focused on music and speech analysis, very little work has been done on concrete analysis of feature selection for classification of environmental sounds. One of the main objectives of this research is to investigate the effect of multiple features on the efficiency of an environmental scene classification system.

The state-of-the-art for acoustic scene classification features a number of approaches. Table 1 presents a summary of some considerable works in this domain which are discussed as follows. In [1], an approach based on local binary patterns (LBP) is adopted to construct the spectrogram image of environmental sounds. The LBP features are enhanced by incorporating local statistics, normalized and finally classified by a linear SVM. The accuracy is validated against RWCP dataset. In [2], the authors studied sound classification in a non-stationary noise environment. At first, probabilistic latent component analysis (PLCA) is performed for noise separation. Further, regularized kernel fisher discriminant analysis (KFDA) is adopted for multi-class sound classification. The method is validated on RWCP dataset. In [3], acoustic classification is performed using large-scale audio feature extraction. First, a large number of spectral, cepstral, energy and voice related features are extracted from highly variable recordings. Then, a sliding window approach is adopted with SVM to classify short recordings. Finally, a majority voting is employed to classify large recordings. The work further proposes Mel spectra as the most relevant features.

Table 1. Summary of published works on acoustic scene classification.

Work	Features	Classifier	Dataset	Accuracy
[1]	ID-LBP	Linear SVM	RWCP	98%
[2]	PLCA, temporal–spectral patterns of sound spectrogram	FDA	RWCP	91.04%
[3]	MFCC, Spectral and energy features	SVM	DCASE	73%
[4]	Multichannel LBP	SVM	RWCP, NTU-SEC	99.85%, 96.29%
[5]	Matching Pursuit and MFCC	GMM	BBC sound effects	98.4%
[6]	Thresholds based pre-processing, FFT	SVM	Self collected 250 recordings of dropping and hitting sounds	87%
[7]	LFCC	GMM	self collected dataset using a microphone set up on cleaning robot platform	90%
[8]	HOG	pooling	DCASE-challenge, Litis Rauin, EA	70%
[9]	MP decomposition using Gabor function with time frequency histogram	Random Forest	Combination of self collected sounds, Sound Idea database [10], Free sound project [11]	
[12]	Deep neural network based transfer learning	Softmax	DCASE	85.6%
[13]	MFCC	CNN	UrbanSoundK	77%
[14]	Multiple	Hierarchical	Self collected	92.6%
[15]	MFCC, ZC, LAR etc.	KNN	Self Collected	99%
[16]	average peak, height & width, no. of half-wavelengths of music wave	Regression analysis	self collected	77%

In [4], features based on LBP from the logarithm of the Gammatone-like spectrogram are proposed. However, LBP is sensitive to noise and discards important information. Therefore, a two-projection-based LBP feature descriptor is also proposed that captures the texture information of the spectrogram of sound events. In [5], a matching pursuit (MP) algorithm is used to extract effective time-frequency features from sounds. The MP technique uses a dictionary of atoms for feature selection, resulting in a set of features that are flexible and physically interpretable. In [6], Fast Fourier Transform (FFT) is used to extract spectral power and duration of event based sounds. A number of features are extracted which include time-domain zero crossings, spectral centroid, roll off, flux and MFCC. Further, sound classification is done using SVM and multi-layer perceptron (MLP). In [7], a combination of log frequency cepstral coefficient (LFCC), Gaussian mixture models (GMMs) and a maximum likelihood criterion is employed to recognize various sound events for a cleaning robot. Experimental results demonstrate that LFCC based approach performs better than MFCC under low signal to noise ratio (SNR) environment. Human classification accuracy in performing similar classification tasks is also evaluated by experiments.

In [8], a feature extraction pipeline is proposed for analyzing audio scene signals. Features are computed from a histogram of gradients (HOG) of constant Q-transform followed by an appropriate pooling scheme. The performance of the proposed scheme is tested on several datasets including Toy, East Anglia (EA) and another dataset named Litis Rouen collected by the authors. In [9], MP algorithm is used to extract useful Gabor atoms from input audio stream. MP is applied over the whole duration of acoustic event. The time-frequency features are constructed from atoms in order to capture temporal and spectral information of a sound event. Further, the classification is done using a random forest classifier. Deep neural network (DNN) based transfer learning is proposed in [12] for acoustic classification. First, the DNN is trained on source domain task that performs mid-level feature extraction. Then, the pre-trained model is re-used on the DCASE target task. In [13], the authors proposed that dilated CNN architecture performs better environmental sound classification as compared to CNN with max pooling. The effect of dilation rate and number of layers on performance is also investigated. The work in [14] proposes a hierarchical approach to classify different sound events such as silence, non-silence, speech, non-speech, music and noise. In contrast to a classical one-step classification scheme, a different set of effective features is selected at each level. In [15], a hearing aid system is proposed for real time recognition of various sounds. The system is based on generating audio finger print i.e., a brief summary of audio file which collects a number of features including spectrogram zero crossings (ZC), MFCCs, linear prediction coefficients (LPCs) and log area ratio (LAR). The recognition is done on self collected sound samples using a K nearest neighbors (KNN) classifier. The system achieves a maximum accuracy of 99%. In [16], the authors propose automatic emotion classification system for music sounds. The work utilizes several features of sound wave, i.e., peak value, average height, the number of half wavelengths, average width and beats per minutes. Finally, regression analysis is perform to recognize various emotions from the sound. The system achieves an average accuracy of 77%. In [17], sound identification method for a mobile robot in home and office environment is proposed. A simple sound database called Pitch-Cluster-Maps (PCMs) based on vector quantization technique is constructed and its codebook is generated using binarized frequency spectrum. The works in [18,19] demonstrate that acoustic local ternary patterns (LTPs) show better performance as compared to MFCCs for fall detection problem. In the literature, various convolutional neural network (CNN) architectures are used to classify soundtracks from a dataset of 70 million training videos (5.24 million hours) with 30,871 video-level labels [20]. Experiments are performed using fully connected DNNs, VGG [21], AlexNet [22], Inception [23] and ResNet [24] etc.

The acoustic scene classification approach proposed in this work has the following contributions.

- An extended feature descriptor is proposed which takes advantage of modified 1-D LTP in combination with MFCC.
- A feature fusion methodology is opted, which exploits the complementary strengths of both MFCC and modified 1-D features to generate a serial vector.
- To provide a better insight, a set of classifiers are tested on two standard benchmark datasets. This action supports researchers in selecting the best classifiers for this application.

The rest of the paper is organized as follows. In Section 2, the proposed method of acoustic scene classification is discussed. Section 3 discusses the experimental setup and datasets. The performance results and discussions are presented and discussed in Section 4 and finally, Section 5 concludes the paper.

2. Proposed Method

2.1. System Overview

Figure 1 shows the overall architecture of the proposed acoustic scene classification system. The sound signal is captured from environment through a microphone. It is digitized using an ADC in the preprocessing step and fed into the feature extraction stage. The MFCC and 1D-LTP features are extracted from the digital sound signal, they are fused together in a joint feature vector and finally classified using an SVM classifier. The main processing steps of the proposed system are discussed as follows.

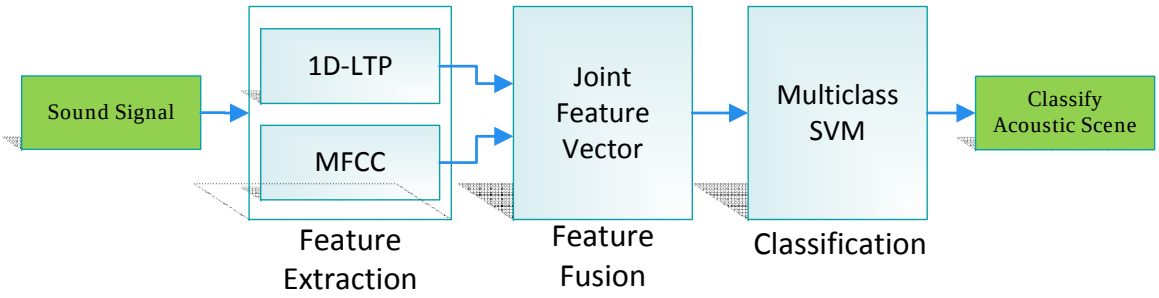

Figure 1. System Architecture for Acoustic Scene Classification.

2.2. Feature Extraction

2.2.1. 1-D Local Ternary Patterns

The local binary patterns (LBPs) have been investigated as among the most prominent feature descriptors in the field of computer vision and image analysis [25]. The basic idea behind LBP is to compare each pixel of an image with its neighborhood. Each compar ison of an image pixel with its neighbors results in binary values '0' or '1'. This helps to summarize a local structure in an image and obtains powerful feature descriptors for a number of promising applications such as face recognition [26] and texture analysis [27]. LBPs are invariant to monotonic grey scale changes and have low computational cost [28]. Applying the LBP method for 1-D signals such as sound, helps to obtain useful information about local temporal dynamics of sound. The LBPs achieve discriminative features of several sounds, as exhibited by the works on music genre recognition [29] as well as environmental sound classification [1]. However, LBPs are highly affected by noise and fluctuations in acoustic samples [1]. In order to further improve the discriminative power of LBP, LTPs were proposed for face recognition in 2010 [30], and later on applied in a number of works [31–33]. In contrast to the LBPs which encode the relationships of 'greater than' or 'less than' between the pixel and its neighbor, the LTPs reflect the 'greater than', 'equal to' or 'less

than' relationships. Under the same sampling conditions, LTPs help to achieve more discriminative and sophisticated sound features as compared to 1D-LBPs.

Analog audio signal is first digitized with sampling frequency F_s to form a discrete signal $X[i]$ having N number of samples. The 1D-LTPs of sampled signal $X[i]$ are computed using a sliding window approach. Consider a signal sample $x[i]$ with amplitude α is placed at the center of window with size $P+1$. Defining the upper and lower values of amplitude threshold as $(\alpha+t)$ and $(\alpha-t)$ respectively, where t is arbitrary constant. From the amplitudes of signal samples that lie in the window, a ternary code vector F of size P is obtained whose individual values are computed as;

$$F[j] = Q(x[i + \frac{P}{2} - r]),\ \forall j \in \{0, \cdots, P-1\}, \tag{1}$$

$$r = \left\{ \begin{array}{cc} j & j < \frac{P}{2} \\ j+1 & j \geq \frac{P}{2} \end{array} \right\}, \tag{2}$$

where $Q(x[i])$ is defined as;

$$Q(x[i]) = \left\{ \begin{array}{cc} 1, & x[i] > (\alpha + t) \\ 0, & (\alpha - t) \leq x[i] \leq (\alpha + t) \\ -1, & x[i] < (\alpha - t) \end{array} \right\}. \tag{3}$$

From the ternary code vector, the upper and lower local ternary patterns are computed as;

$$LTP_{upper}[i] = \sum_{k=0, k \neq i}^{P-1} s_u(F[k]) \cdot 2^k, \tag{4}$$

$$LTP_{lower}[i] = \sum_{k=0, k \neq i}^{P-1} s_l(F[k]) \cdot 2^k, \tag{5}$$

where,

$$s_u(F[k])) = \left\{ \begin{array}{cc} 1 & F[k] = 1 \\ 0 & otherwise \end{array} \right\}, \tag{6}$$

$$s_l(F[k])) = \left\{ \begin{array}{cc} 1 & F[k] = -1 \\ 0 & otherwise \end{array} \right\}, \tag{7}$$

Figure 2 illustrates the extraction of 1D-LTP features for one sample of a discrete audio signal.

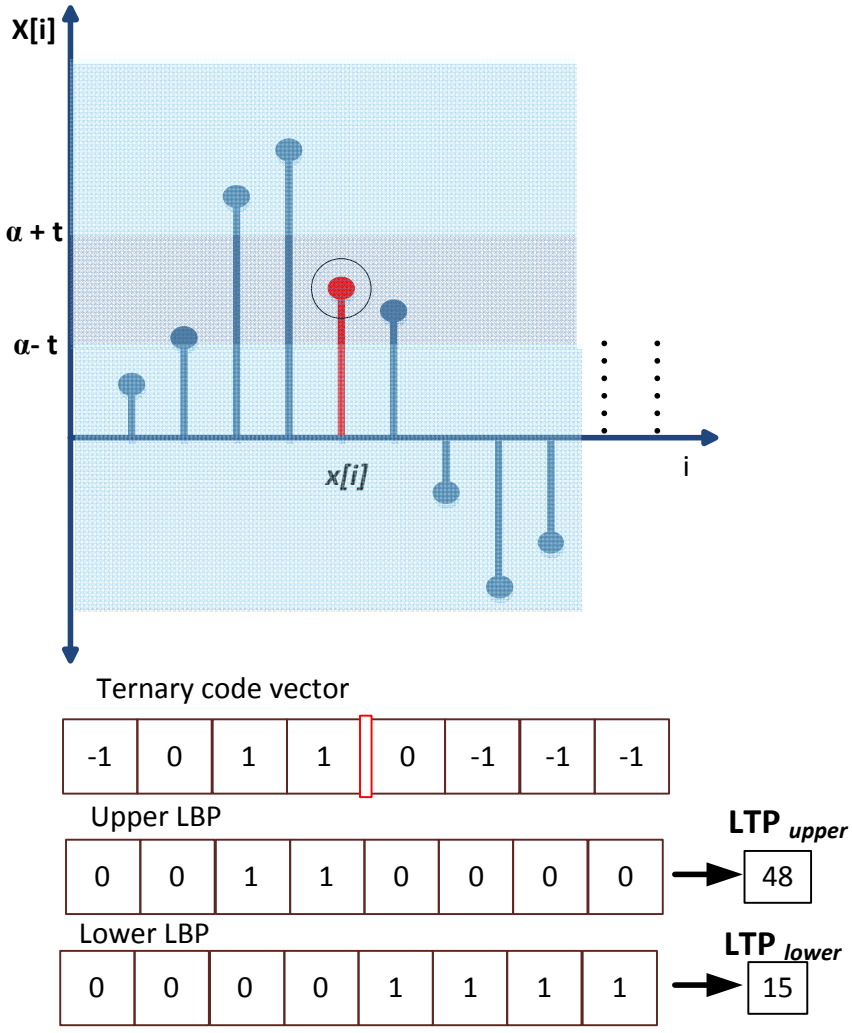

Figure 2. Extraction of 1D-LTP features.

2.2.2. Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs are a baseline method that has been widely used in the analysis of audio signals. Although primarily designed for speech recognition [34,35], they have been a popular feature of choice in the automatic scene classification [36,37]. The MFCCs are the coefficients that collectively make up the Mel Frequency Cepstrum (MFC), a representation of short term power spectrum of sound based on linear cosine transform of a log power spectrum on a non linear Mel scale of frequency. The MFCCs are linearly spaced on the Mel frequency scale which closely approximates the human auditory system's response. Such a representation of sound signal extracts discriminant features which help to achieve environmental sound classification with good accuracy.

Figure 3 shows a standard pipeline for the extraction of MFCC features. In the first step, the digitized sound signal is segmented in to short frames each having N samples. Next, the periodogram-based power spectrum is estimated for each frame. Let $s_i(n)$ denote the time domain signal (of N samples) that belongs to frame i, its Discrete Fourier Transform (DFT) is calculated as;

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j2\pi kn/N}, \ 1 \leq k \leq K \tag{8}$$

where K denotes the length of DFT and $h(n)$ denotes the N sample long analysis window. In this work, Hamming window is used to realize a high-pass FIR filter to emphasize the high frequency part of the signal and remove DC content. In the next step, the output of complex Fourier transform is magnitude squared and power spectral estimate of frame i is computed as;

$$P_i(k) = \frac{1}{N}|S_i(k)|^2, \ 1 \leq k \leq K. \tag{9}$$

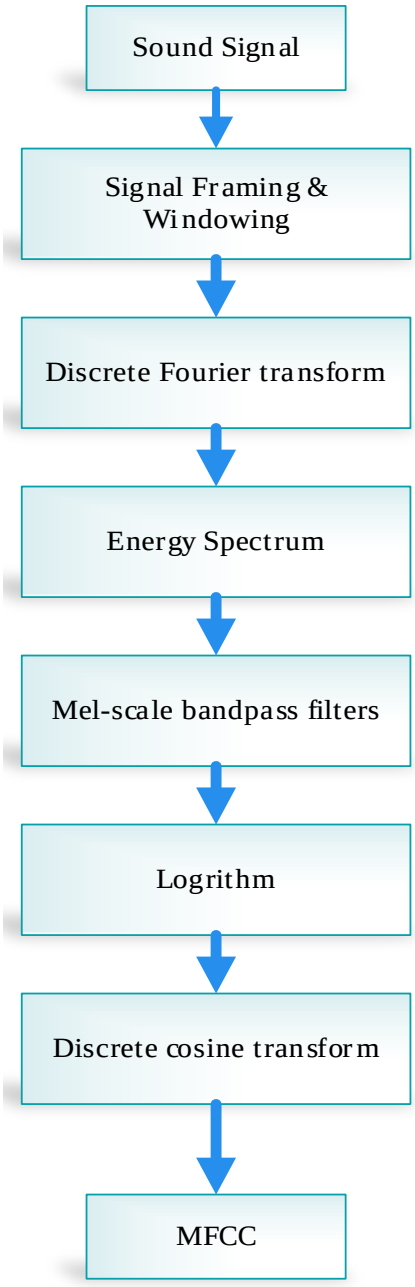

Figure 3. MFCC Feature Extraction Pipeline.

Then, a set of Mel-scaled filter banks is computed and applied to power spectrum of each frame. The Mel-scale is linear for frequencies lower than 1000 Hz and a logarithm above it. To compute the filter bank energy spectrum, each filter is multiplied by the power spectrum computed above and coefficients are added up. The Mel-filtered spectrum of frame i is computed as;

$$E_i(l) = \sum_{k=0}^{N-1} P_i(k) H_l(k), \ \forall l = 1, \cdots, L \tag{10}$$

where L denotes the total number of filters and H_l denotes the transfer function of lth filter. Next, the *logarithm* of Mel-filtered energy spectrum is computed and Discrete Cosine Transform (DCT) is applied to it. Mathematically,

$$E_l^{'}(l) = \log(E_i(l)), \ \forall l = 1, \cdots, L \tag{11}$$

$$c_i(n) = \sum_{l=1}^{L} E_l^{'}(l) cos(n(l - 0.5)/\pi/L) \tag{12}$$

where $n = 1, \cdots, L$ is the cepstral coefficient number. In the proposed frame work, initial 13 MFCCs are used for scene classification.

2.3. Feature Fusion

The 1D-LTP and MFCC features extracted above are fused together to form a joint feature vector for classification. The fusion of 1D-LTP and MFCC features helps to obtain a more sophisticated feature representation which has better discriminative properties as well as an accurate representation in frequency domain. The fusion process is a simple serial concatenation of 1D-LTP and MFCC feature vectors.

$$\mathbb{F}^{(c,s)} = c_\kappa || s_\kappa \tag{13}$$

2.4. Classification

The classification stage employs a multiclass SVM. The basic idea of SVM is to find a hyperplane that separates D-dimensional data into its two classes [38]. SVM is a discriminative model for classification that principally depends on two basic assumptions. First, complex classification problems can be classified through simple linear discriminative functions by transforming data into a high-dimension space. Second, the training samples for SVMs consist only of those data points that lie close to the decision surface, with the supposition that they provide the most relevant information for classification [39]. SVMs were originally proposed as binary classifiers. However, in real scenarios, data is to be classified into multiple classes. This is done by using multiclass SVM. Either a one-against-one (OAO) or one-against-all (OAA) approach can be used [40]. For acoustic scene classification setup proposed in this work, the joint feature vector extracted from previous stage is used to train the multiclass SVM OAO classifier.

3. Experiments

3.1. Setup

Experiments were performed using MATLAB 2016a software on 2.2 GHz Intel i7 processor with 8 GB RAM. The extracted features are MFCC (13 coefficients) and 1D-LTPs (13 bins) with threshold $t = 0.0002$. The classification is being done by applying various SVM kernels, and by finalizing quadratic and cubic kernels because of their best performance [41]. Training/testing percentage is fixed to be 80/20 (80% for training, and 20% for testing) for both datasets. The performance of classifier is measured through

classification accuracy averaged over k-fold cross validation. The value of $k = 10$ has been selected based on experimentation to generally result in best accuracy with low bias, modest variance and low correlation. The classifier accuracy is measured as,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (14)$$

where TP stands for true positive, TN for true negative, FP for false positive and FN for false negative. The performance of the proposed approach is also compared with several state-of-the-art audio feature representation techniques i.e., MFCC, ID-LBP and linear prediction cepstral coefficients (LPCC).

3.2. Datasets

An important challenge in acoustic scene classification for robotics is the collection of proper environmental sound database. Since there is an infinite number of sounds, no single database can cover all of them. Therefore, no robotic system is capable of recognizing all the sounds. Instead, the scene recognition capability is limited by the application domain and set of tasks performed by the particular robot. In order to have an initial reference for comparison, two standard benchmark datasets are selected, i.e., (a) real world computing partnership (RWCP) sound scene dataset [42] and (b) DCASE challenge dataset [43].

RWCP is one of the first datasets which are collected for scene understanding. It contains sounds of various audio sources which were moved using a mechanical device. Recordings were done using a linear array of 14 microphones and a semi-spherical array of 54 microphones with a DAT recorder at 48 KHz frequency and 16 bit resolution. The average length of sound sample is about 1 s. A proposed feature descriptor was tested on experimental dataset consisting of 17 different environmental sounds shown in Table 2 (a) along with the number of samples for each class.

The DCASE challenge dataset consists of a set of recorded sounds in fifteen different urban environments. The duration of each sound clip is 30 s and recording is performed in London. The DCASE dataset consists of 15 different classes of urban sounds; each class contains 78 sound samples as given in Table 2 (b). The RWCP and DCASE databases contain a variety of sound classes that accurately model the general indoor or outdoor environment. We believe that verifying the performance of our proposed solution on these databases can help to realize intelligent systems for advanced applications such as sound localization [44] and human–robot interaction [45,46].

As discussed earlier, 1D-LTP features are discriminative. The scatter plots of Figures 4 and 5 show the distribution of 1D-LTPs for several classes of RWCP and DCASE datasets. These plots demonstrate that the 1D-LTP feature values that belong to the same class are spaced close to each other, whereas the features belonging to different classes are spaced relatively far on the scatter plot. Features having these strong discriminative properties result in a good classification accuracy.

Table 2. Details of Individual Classes of RWCP and DCASE Datasets.

(a) RWCP Dataset	
Class	**No. of Samples**
Aircap	100
Bells	400
Bottle	200
Buzzer	100
Case	300
Clap	400

Table 2. *Cont.*

Cup	200
Drum	100
Phone	200
Pump	100
Saw	200
Spray	100
Stapler	100
Tear	100
Toy	200
Whistle	300
Wood	300
Total	**3400**
(b) DCASE Dataset	
Class	**No. of Samples**
Beach	78
Bus	78
Cafe	78
Car	78
City Center	78
Forest	78
Grocery Store	78
Home	78
Library	78
Metro Station	78
Office	78
Park	78
Residential area	78
Train	78
Tram	78
Total	**1170**

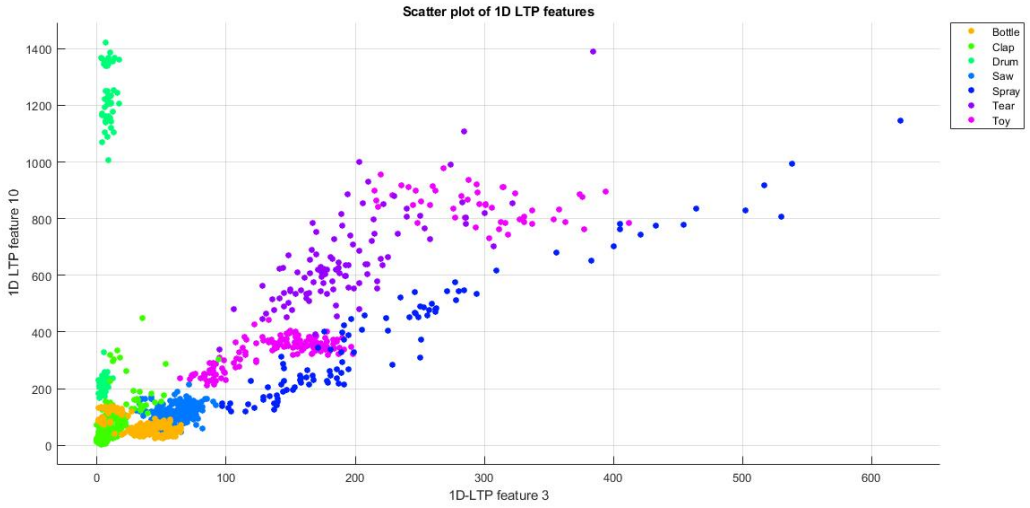

Figure 4. Scatter plot of ID-LTPs of RWCP dataset.

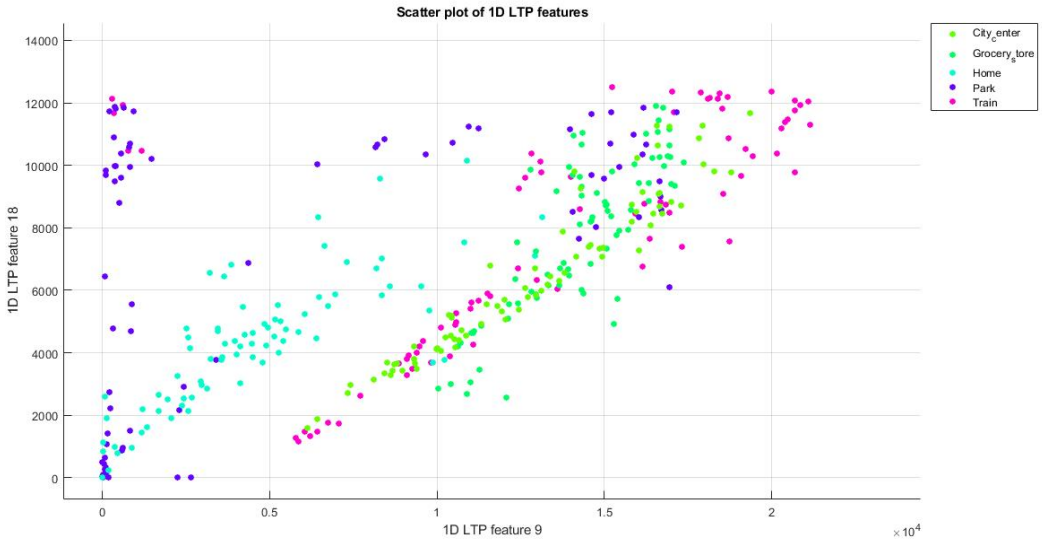

Figure 5. Scatter plot of ID-LTPs of DCASE dataset.

4. Results and Discussion

The accuracy trend for both datasets is demonstrated in Figure 6. Table 3 presents the overall classification accuracy of the proposed and existing methods along with their computational time in seconds. It can be comfortably observed from the stats that the proposed method (i.e., ID-LTP + MFCC) outperforms shows a better accuracy with computational time smaller or comparable to other approaches.

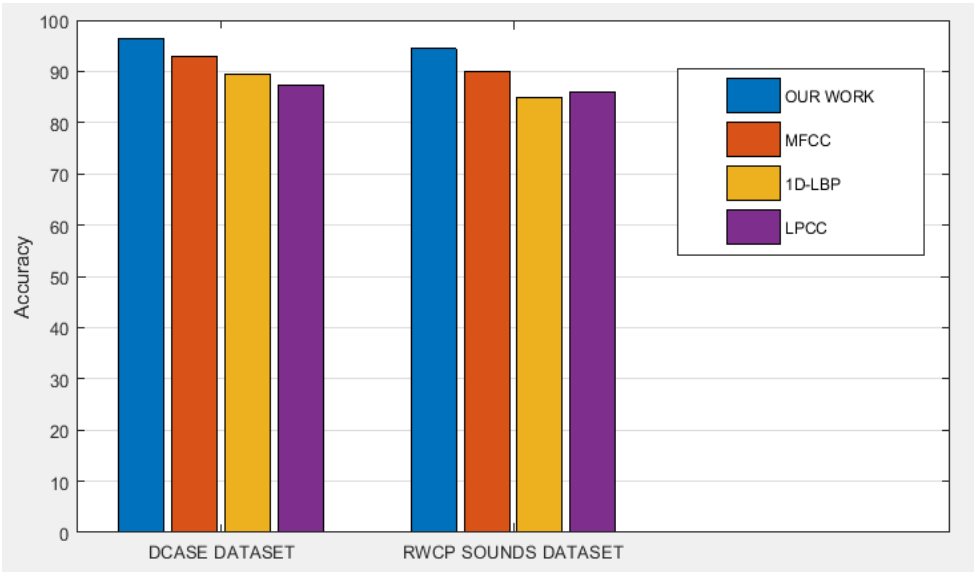

Figure 6. Classification performance of the proposed ID-LTP and several other features over DCASE and RWCP dataset.

To get a better insight, few other performance metrics are also investigated including sensitivity, specificity, and error rate. Moreover, for a fair comparison, two classifier families, i.e., SVM and KNN are contemplated due to their greater number of variants. Table 4 provides a comparison of seven classifiers on the DCASE dataset. The SVM with quadratic kernel (SVM-Q) shows better results in terms of accuracy,

specificity and error rate while SVM with cubic kernel (SVM-C) and KNN weighted (KNN-W) show better sensitivity. In Table 5, the performance results are demonstrated for RWCP dataset. The SVM-Q classifier achieves a high accuracy and error rate while better sensitivity and specificity values are achieved by the KNN medium (KNN-M) and SVM-C, respectively.

Table 3. Performance results for DCASE and RWCP datasets.

Feature Descriptor	Accuracy		Time (s)
	DCASE Dataset	RWCP Sound Dataset	
MFCC	92.9%	90%	1.2
ID-LBP	89.5%	85%	0.75
LPCC	87.3%	86%	0.92
ID-LTP + MFCC	**97.38**%	**94.10**%	**0.81**

Table 4. Performance of various classifiers for proposed feature extraction approach for DCASE dataset.

DCASE Dataset				
Classifier	Performance			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Error Rate
SVM-L	89.49	83.33	99.54	0.1051
SVM-Q	**94.10**	**91.03**	**99.91**	**0.0590**
SVM-C	93.85	93.59	99.91	0.0615
SVM-G	93.16	92.31	99.82	0.0684
KNN-M	85.04	92.31	98.81	0.1496
KNN-W	90.26	93.59	99.36	0.0974
KNN-C	82.56	84.62	98.35	0.1744

Table 5. Performance of various classifiers for proposed feature extraction approach for RWCP dataset.

RWCP Dataset				
Classifier	Performance			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Error Rate
SVM-L	93.97	98.50	99.93	0.0603
SVM-Q	**97.38**	**99.0**	**99.83**	**0.0262**
SVM-C	97.26	99.25	99.97	0.0274
SVM-G	94.44	98.75	99.57	0.0556
KNN-M	97.26	99.50	99.83	0.0274
KNN-W	96.85	99.00	99.80	0.0315
KNN-C	96.35	99.25	99.80	0.0365

Classification results of individual classes for the DCASE dataset are shown by a confusion matrix of Figure 7. The figure shows that all classes except the *city center* class have an accuracy of more than 90%. The confusion matrix of the proposed approach for RWCP dataset is shown in Figure 8. Here, the *phone* class has an accuracy of 89% whereas, all the remaining classes have accuracy above 90%. The classification results of Figure 7 and 8 confirm the accuracy and validity of the proposed feature classification technique. To reveal the authenticity and robustness of our proposed method, confidence intervals against both datasets are also provided for two state-of-the-art classifiers. Figure 9 demonstrates the confidence interval showing min, max and average classification values of both classifiers. From the stats, its quite obvious that SVM-Q can be formally selected as a standard classifier for this application.

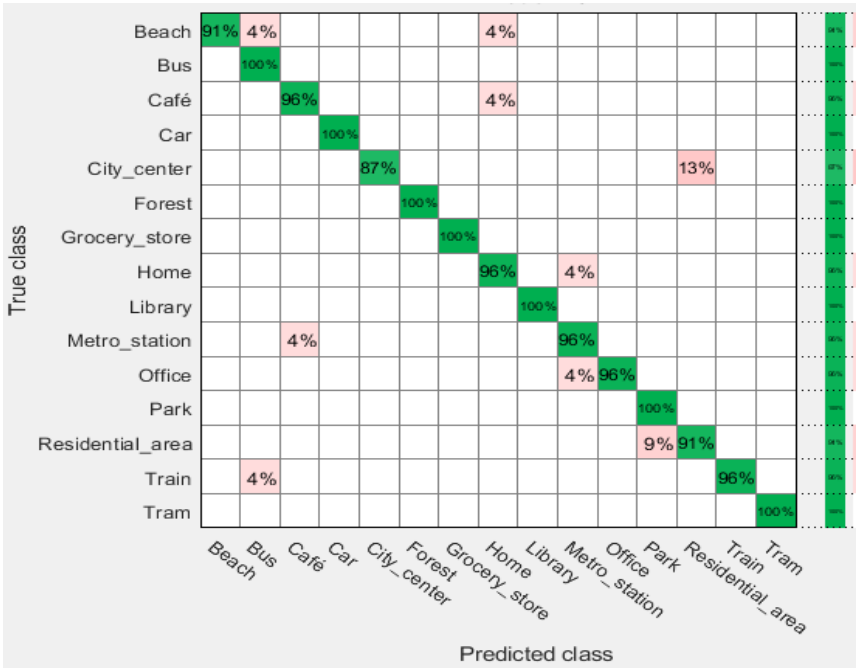

Figure 7. Confusion matrix of the proposed approach for DCASE dataset.

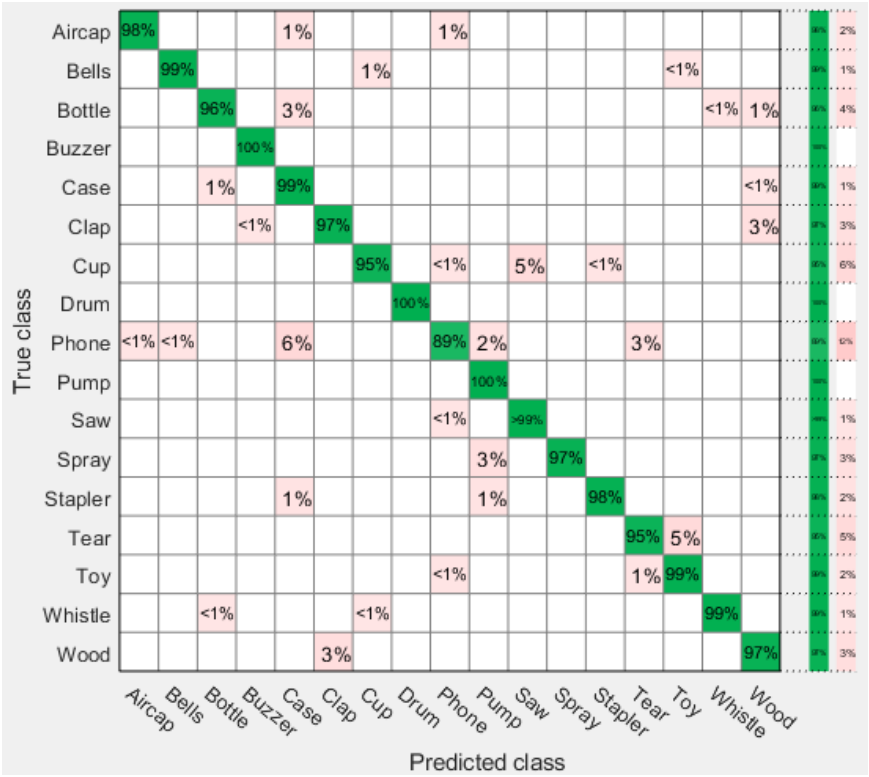

Figure 8. Confusion matrix of the proposed approach for RWCP dataset.

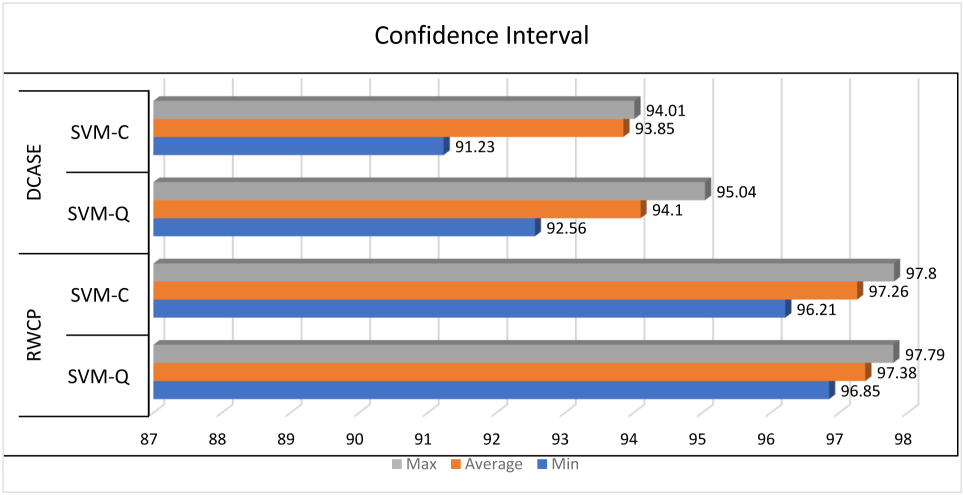

Figure 9. Confidence interval against two selected classifiers on benchmark datasets.

5. Conclusions

Scene classification is an important task in behavioral robotics. Using acoustic signals for environmental scene classification complements the visual-based classification in many ways. This study aimed to select the image texture classification features and investigate their effect on the classification of sound signals. In particular, the work proposes a modified feature descriptor as a combination of 1D-LTPs and MFCCs. Our analysis and simulation results for the two reference datasets i.e., DCASE and RWCP show that 1D-LTPs exhibit good discriminative properties for sound signals. On the other hand, the MFCCs as the baseline method, approximates the behavior of the human auditory system. Fusing 1D-LTPs with MFCCs achieves a more sophisticated and discriminative feature representation of environmental sounds. The proposed fused feature vector is classified with various kernels of multi-class SVM. Results demonstrate that SVM with quadratic kernel achieves high accuracy as compared to other feature representations. The proposed system can be applied to a number of practical indoor and outdoor robotic scenarios.

6. Materials

Two publicly available datasets are utilized in this research are RWCP and DCASE. The RWCP dataset is available at [42] and DCASE is available at: http://dcase.community/challenge2018/index.

Author Contributions: Conceptualization, S.A.; Data curation, M.A. (Muhammad Awais) and T.A.; Funding acquisition, M.A. (Musaed Alhussein) and K.A.; Investigation, M.A. (Muhammad Awais) and U.K.; Methodology, S.A., M.A. (Muhammad Awais) and U.K.; Project administration, T.A.; Resources, S.A. and K.A.; Software, M.A. (Muhammad Awais); Validation, M.A. (Musaed Alhussein) and K.A.; Writing—original draft, M.A. (Muhammad Awais).

Funding: The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group NO (RG-1438-034).

Acknowledgments: We greatly acknowledge Department of Electrical & Computer Engineering, COMSATS University Islamabad, Wah Campus and Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh for their technical support throughout the duration of this research work.

Conflicts of Interest: The authors declare that they have no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

LBP	Local Binary Patterns
LTP	Local Ternary Patterns
MFCC	Mel Frequency Cepstral Coefficients
SVM	Support Vector Machine
PCLA	Probabilistic Component Latent Analysis
KFDA	Kernel Fisher Discriminant Analysis
HOG	Histogram of Gradients
DNN	Deep Neural Networks
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GMM	Gaussian Mixture Model
KNN	K-Nearest Neighbour
SVM-C	SVM with Cubic kernel
SVM-Q	SVM with Quadratic kernel
SVM-G	SVM with mean Gaussian kernel
KNN-M	K Nearest Neighbors-Medium
KNN-W	K Nearest Neighbors-Weighted
KNN-C	K Nearest Neighbors-Cubic
OAO	One Against One
OAA	One Against All

References

1. Kobayashi, T.; Ye, J. Acoustic feature extraction by statistics based local binary pattern for environmental sound classification. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 3052–3056.
2. Ye, J.; Kobayashi, T.; Murakawa, M.; Higuchi, T. Robust acoustic feature extraction for sound classification based on noise reduction. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 5944–5948.
3. Geiger, J.T.; Schuller, B.; Rigoll, G. Large-scale audio feature extraction and SVM for acoustic scene classification. In Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20–23 October 2013; pp. 1–4.
4. Ren, J.; Jiang, X.; Yuan, J.; Magnenat-Thalmann, N. Sound-Event Classification Using Robust Texture Features for Robot Hearing. *IEEE Trans. Multimed.* **2017**, *19*, 447–458. [CrossRef]
5. Chu, S.; Narayanan, S.; Kuo, C.J. Environmental Sound Recognition With Time-Frequency Audio Features. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 1142–1158. [CrossRef]
6. Saltali, I.; Sariel, S.; Ince, G. Scene Analysis Through Auditory Event Monitoring. In Proceedings of the International Workshop on Social Learning and Multimodal Interaction for Designing Artificial Agents, Tokyo, Japan, 16 November 2016; pp. 5:1–5:6.
7. Park, S.; Rho, J.; Shin, M.; Han, D.K.; Ko, H. Acoustic feature extraction for robust event recognition on cleaning robot platform. In Proceedings of the 2014 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 10–13 January 2014; pp. 145–146.
8. Rakotomamonjy, A.; Gasso, G. Histogram of Gradients of Time–Frequency Representations for Audio Scene Classification. *IEEE-ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 142–153.
9. Nguyen, Q.; Choi, J. Matching pursuit based robust acoustic event classification for surveillance systems. *Comput. Electr. Eng.* **2017**, *57*, 43–54. [CrossRef]

10. Sehili, M.A.; Lecouteux, B.; Vacher, M.; Portet, F.; Istrate, D.; Dorizzi, B.; Boudy, J. Sound Environment Analysis in Smart Home. In *Ambient Intelligence*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 208–223.
11. Wang, J.; Lin, C.; Chen, B.; Tsai, M. Gabor-Based Nonuniform Scale-Frequency Map for Environmental Sound Classification in Home Automation. *IEEE Trans. Autom. Sci. Eng.* **2014**, *11*, 607–613. [CrossRef]
12. Mun, S.; Shon, S.; Kim, W.; Han, D.K.; Ko, H. Deep Neural Network based learning and transferring mid-level audio features for acoustic scene classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 796–800.
13. Chen, Y.; Guo, Q.; Liang, X.; Wang, J.; Qian, Y. Environmental sound classification with dilated convolutions. *Appl. Acoust.* **2019**, *148*, 123–132. [CrossRef]
14. Saki, F.; Kehtarnavaz, N. Real-time hierarchical classification of sound signals for hearing improvement devices. *Appl. Acoust.* **2018**, *132*, 26–32. [CrossRef]
15. Yağanoğlu, M.; Köse, C. Real-Time Detection of Important Sounds with a Wearable Vibration Based Device for Hearing-Impaired People. *Electronics* **2018**, *7*, 50. [CrossRef]
16. Seo, Y.S.; Huh, J.H. Automatic Emotion-Based Music Classification for Supporting Intelligent IoT Applications. *Electronics* **2019**, *8*, 164. [CrossRef]
17. Sasaki, Y.; Kaneyoshi, M.; Kagami, S.; Mizoguchi, H.; Enomoto, T. Daily sound recognition using Pitch-Cluster-Maps for mobile robot audition. In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 10–15 October 2009; pp. 2724–2729.
18. Irtaza, A.; Adnan, S.M.; Aziz, S.; Javed, A.; Ullah, M.O.; Mahmood, M.T. A framework for fall detection of elderly people by analyzing environmental sounds through acoustic local ternary patterns. In Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; pp. 1558–1563.
19. Adnan, S.M.; Irtaza, A.; Aziz, S.; Ullah, M.O.; Javed, A.; Mahmood, M.T. Fall detection through acoustic Local Ternary Patterns. *Appl. Acoust.* **2018**, *140*, 296–300. [CrossRef]
20. Hershey, S.; Chaudhuri, S.; Ellis, D.P.W.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135.
21. Karen, S.; Andrew, Z. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 3–8 December 2012.
23. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Ojala, T.; Pietikainen, M.; Harwood, D. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognit.* **1996**, *29*, 51–59. [CrossRef]
26. Zhang, B.; Gao, Y.; Zhao, S.; Liu, J. Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. *IEEE Trans. Image Process.* **2010**, *19*, 533–544. [CrossRef] [PubMed]
27. Liu, L.; Fieguth, P.; Guo, Y.; Wang, X.; Pietikäinen, M. Local binary features for texture classification: Taxonomy and experimental study. *Pattern Recognit.* **2017**, *62*, 135–160. [CrossRef]
28. Thwe, K.Z. Sound event classification using bidirectional local binary pattern. In Proceedings of the 2017 International Conference on Signal Processing and Communication (ICSPC), Tamil Nadu, India, 28–29 July 2017; pp. 501–504. [CrossRef]

29. Costa, Y.M.; Oliveira, L.; Koerich, A.L.; Gouyon, F.; Martins, J. Music genre classification using LBP textural features. *Signal Process.* **2012**, *92*, 2723–2737. [CrossRef]
30. Tan, X.; Triggs, W. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.* **2010**, *19*, 1635–1650. [PubMed]
31. Raja, M.; Sadasivam, V. Optimized local ternary patterns: A new texture model with set of optimal patterns for texture analysis. *J. Comput. Sci.* **2013**, *9*, 1–15. [CrossRef]
32. Wu, S.; Yang, L.; Xu, W.; Zheng, J.; Li, Z.; Fang, Z. A mutual local-ternary-pattern based method for aligning differently exposed images. *Comput. Vis. Image Underst.* **2016**, *152*, 67–78. [CrossRef]
33. Zhang, Y.; Li, S.; Wang, S.; Shi, Y.Q. Revealing the traces of median filtering using high-order local ternary patterns. *IEEE Signal Process. Lett.* **2014**, *21*, 275–279. [CrossRef]
34. Han, W.; Chan, C.F.; Choy, C.S.; Pun, K.P. An efficient MFCC extraction method in speech recognition. In Proceedings of the 2006 IEEE International Symposium on Circuits and Systems, Island of Kos, Greece, 21–24 May 2006.
35. Ittichaichareon, C.; Suksri, S. Speech Recognition using MFCC. In Proceedings of the International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012), Pattaya, Thailand, 28–29 July 2012; pp. 28–29.
36. Mesaros, A.; Heittola, T.; Virtanen, T. TUT database for acoustic scene classification and sound event detection. In Proceedings of the 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 29 August–2 September 2016; pp. 1128–1132.
37. Shaukat, A.; Ahsan, M.; Hassan, A.; Riaz, F. Daily sound recognition for elderly people using ensemble methods. In Proceedings of the 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Xiamen, China, 19–21 August 2014; pp. 418–423.
38. Amarappa, S.; Sathyanarayana, S. Data classification using Support vector Machine (SVM), a simplified approach. *Int. J. Electron. Comput. Sci. Eng.* **2014**, *3*, 435–445.
39. Faziludeen, S.; Sabiq, P.V. ECG beat classification using wavelets and SVM. In Proceedings of the 2013 IEEE Conference on Information Communication Technologies, Thuckalay, India, 11–12 April 2013; pp. 815–818.
40. Jonathan, M.; Mohamed, C.; Robert, S. "One Against One" or "One Against All": Which One is Better for Handwriting Recognition with SVMs? In Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule, France, 23–26 October 2006.
41. Lee, S.W.; Verri, A. (Eds.) *Pattern Recognition with Support Vector Machines*; Springer: Berlin, Germany, 2002.
42. Nakamura, S.; Hiyane, K.; Asano, F.; Nishiura, T.; Yamada, T. Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition. In Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece, 31 May–2 June 2000.
43. Giannoulis, D.; Stowell, D.; Benetos, E.; Rossignol, M.; Lagrange, M.; Plumbley, M.D. A database and challenge for acoustic scene classification and event detection. In Proceedings of the 21st European Signal Processing Conference (EUSIPCO 2013), Marrakech, Morocco, 9–13 September 2013; pp. 1–5.
44. Rascon, C.; Meza, I. Localization of sound sources in robotics: A review. *Robot. Auton. Syst.* **2017**, *96*, 184–210. [CrossRef]
45. Toyoda, Y.; Huang, J.; Ding, S.; Liu, Y. Environmental sound recognition by multilayered neural networks. In Proceedings of the Fourth International Conference on Computer and Information Technology, Wuhan, China, 16 September 2004; pp. 123–127. [CrossRef]
46. Yamakawa, N.; Takahashi, T.; Kitahara, T.; Ogata, T.; Okuno, H.G. Environmental sound recognition for robot audition using matching-pursuit. In Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Syracuse, NY, USA, 29 June–1 July 2011; pp. 1–10.