Robust Sound Classification for Surveillance using Time Frequency Audio Features

Saleet Ul Hassan, Muhammad Zeeshan Khan, Muhammad Usman Ghani Khan, Summra Saleem
Al-Khwarizmi Institute of Computer Science UET, Lahore [1,2,4]
Computer Science Department UET, Lahore[3]
{*saleet.hassan, zeeshan.khan, usman.ghani, summra.saleem*}@kics.edu.pk

***Abstract*—Over the years, technology has reformed the perception of the world related to security concerns. To tackle security problems, we proposed a system capable of detecting security alerts. System encompass audio events that occur as an outlier against background of unusual activity. This ambiguous behaviour can be handled by auditory classification. In this paper, we have discussed two techniques of extracting features from sound data including: time-based and signal based features. In first technique, we preserve time-series nature of sound, while in other signal characteristics are focused. Convolution neural network is applied for categorization of sound. Major aim of research is security challenges, so we have generated data related to surveillance in addition to available datasets such as UrbanSound 8k and ESC-50 datasets. We have achieved 94.6% accuracy for proposed methodology based on self-generated dataset. Improved accuracy on locally prepared dataset demonstrates novelty in research.**

***Index Terms*—MFCC, Mel Spectrogram, Sound, Convolution Neural Network, Surveillance**

I. Introduction

In last few years, computer systems have progressed exponentially as they perform more complex tasks in seconds. Machines have surpassed humans in the field of machine perceptron. One of the major achievement is a visual recognition with deep learning approaches. On the other hand, the role of sound is very crucial in success of artificial Intelligence. Machines should be able to interpret the sound the way humans do. Classifying urban and environmental sound is elementary, sound signal processing problem. Applications of sound classification varies from surveillance [1], context aware computing [2], noise mitigation to large scale content based retrieval.

In todays world, security is a major concern due to terrorism and insurgency issues. Security is balancing parameter between invader and protector, so its a sort of tradeoff, which is never been static. When technology changes it affects both sides. As we know, during the last few decades, one of the major problem that Pakistan is facing is terrorism. It creates frustration not only for federation but it is also a nightmare for public. Though, it effects the whole world, but Pakistan is among the countries who have suffered a lot due to this menace. As Pakistan is an under developed country, so it is difficult to hire a full-fledge force only against a few number of terrorists, with guns. So, to tackle this problem, cameras have been installed at the various public locations for keeping an eagle eye on the situation. The streams of the cameras are stored in repository. Consequently, if the law enforcement agencies have to search some specific suspicious and tragic happening from the past for forensic investigation, they have to manually go through the repository of videos to find specific one. Currently, most of the approaches for video retrieval system using the visual features [3] [4]. Whereas the major aim of this research is not only to enhance the performance of video classification but also to build an application, capable of retrieving video on the basis of defined security related sounds.

One of the major challenges faced by many researchers in sound classification is feature extraction. Features of images and text can be expressed in vector form, but features of sound data cannot be expressed in vector. In this paper, we have discussed two techniques of feature extraction. First one focusses on signal characteristics of sound, and the second one focuses on time-series nature of sound. In signal-based feature extraction important characteristics of each sample are expanded and isolated, following are the characteristics that are used in this technique, MFCC (Mel Frequencies Cepstral Coefficients), Tonal centroid features, Spectral contrasts, and Spectrograms. In the second technique of time series feature extraction, log filterbanks and filterbank methods are used. As sound is segmented into smaller chunks, these methods filter the significant information from each chunk. In this way it preserves time series nature of the data.

Well labelled datasets available for this purpose are very limited and rare. In 2014 Justin Salamone, Juan Pablo Bello, and Christopher Jacoby [5] created a free large-scale dataset (Urban Sound 8k). Urban Sound 8K consists of 8732 sound from 10 classes. The length of the audios is less than 5 seconds. Some researchers have worked on ESC-50 dataset [6], it is also a free dataset having 2000 environmental sounds. The length of each recording is 5 second and these recordings are organized into 50 classes. Since our focus is to retrieve the videos on the basis of the security related sounds. So we have elected categories that are related to security from above described two datasets and rest of the data we generated by

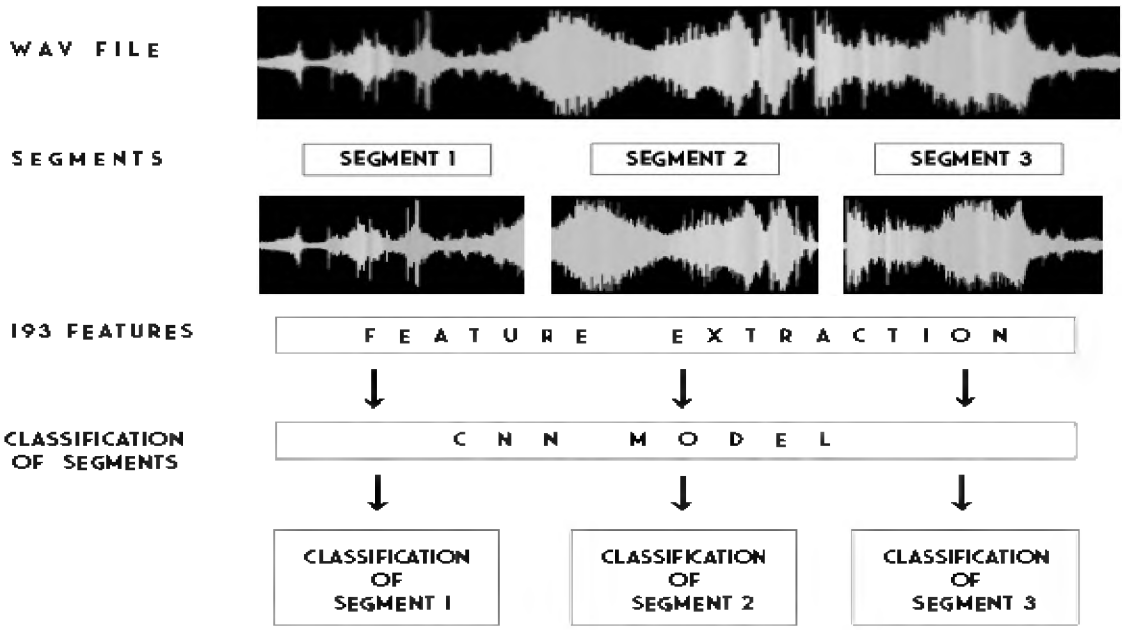

Fig. 1. Framework Diagram

ourselves. We have applied the novel deep learning approach CNN (Convolution Neural Network) for sound classification purpose. We have also compared the results on DNN (Deep Neural Network), and RNN (Recurrent Neural Network) using the same dataset and shown that proposed CNN architecture achieved the better results on generated dataset. Related work will be explained in detail in section II. Section III describes the proposed system and methodology, section IV describes the dataset, section V based on the results and section VI will conclude the paper.

II. Related Work

Recently, urban and environmental sound classification has received a little attention from researchers around the globe. Till now only a few number of techniques of machine learning and signal processing have been proposed and built to inspect the urban sound classification on raw audio data as compared to image and video. These techniques include the matrix factorization [6], dictionary learning [7] [8], wavelet filter banks [9] and most recently deep neural networks. Antti J. Eronen et al. [10] worked on Audio based context recognition, as there was a high variance and randomness in environmental sound, their recognition rate was effected with increased number of classes. As they increase the number of classes their recognition rate falls rapidly. They achieved almost 92% accuracy on 5 classes, as they increased the classes their accuracy fell from 92% to 77% on 11 classes, and 60% on 13 classes. Machine learning clustering techniques have also been applied for sound classification purpose. Features derived from the audio representation are used in machine learning techniques like SVM [11], KNN [12].

Justin Salamon et al. [5], used machine learning based spherical k-mean algorithm for classifying urban sound. They worked on MFCC (Mel-Frequency Cepstral Coefficients) and Mel spectrograms. They created urban sound dataset, it was a huge contribution for the research community. Urban Sound 8K consists of 8732 sounds. For the validation purpose they used 10-fold cross validation. They achieved 5% accuracy improvement over baseline. As our task is related to the deep neural network, so here we focussed the most of the literature related to the deep learning for sound classification purpose.

Karol J. Piczak et al. [13] proposed deep learning-based CNN consisting of 2 convolution layers with max pooling and 2 fully connected layers. They worked on both ESC-50 and urban sound 8K dataset. They achieved 64.5% accuracy on ESC-50 dataset and 73.6% accuracy on urban sound 8K dataset. To improve the accuracy Yuji Tokozume et al. [14] used deep learning-based CNN. Firstly, they applied convolution layer two times with small filter size to raw waveform to extract the local features. Their architecture was based on two convolution layers followed by pooling layers. At the end they have 2 fully connected layers for sound classification. For experimental purpose they used ESC-50 dataset and they achieved 71.0% accuracy. Whereas Hardik B. Sailor et al. [15] used deep learning-based CNN with filterbanks learned using convolution RBM and fusion with GTSC and mel-energies. They used ESC-50 Dataset which consists of 2000 short (5 seconds) environmental Sound recordings. They achieved the accuracy of 86.50%.

The audioset dataset have also been utilized for sound classification purpose in deep learning. Anurag Kumar [16] proposed Deep learning-based CNN for classification of Sound events

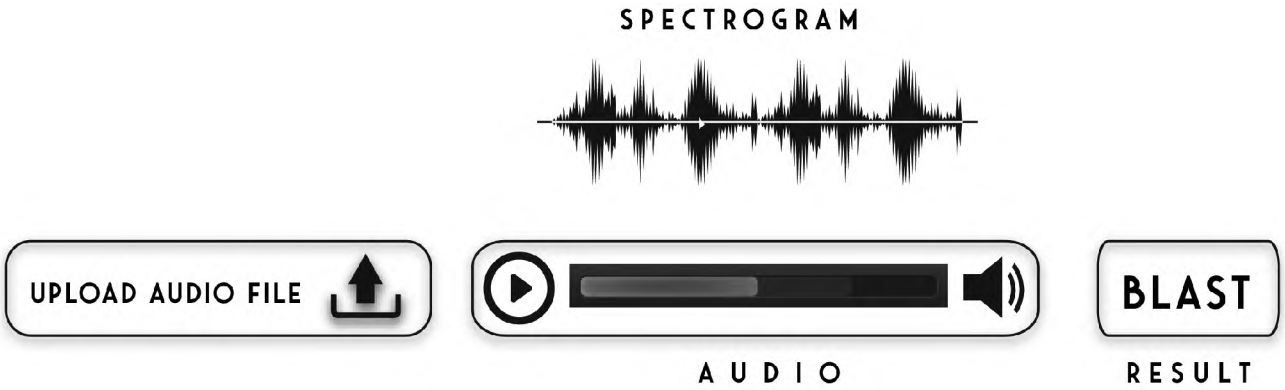

Fig. 2. Graphical User Interface

and scenes. They used audioset dataset which consists of weak labels for 527 sound events of YouTube videos. The total dataset of audioset [17] consists of over two million audio recordings. They used balanced training which provides a total of 22000 training audio sounds with minimum of 59 examples per class. Around 20000 test audio sounds again with minimum 59 examples per class. They achieved the accuracy of 83.5%. Since one of the major problem in deep learning is the availability of large number of dataset for defined categories. To tackle this problem Justin Salamon et al. [18], proposed deep learning-based CNN and data augmentation techniques for environmental sound classification. Firstly, they proposed deep CNN architecture for environmental sound recognition and used audio data augmentation for overcoming the problems of data scarcity. They explored the influence of different augmentations on the performance of proposed CNN Architecture. They worked on urban sound 8K dataset, they are actually the owner of Urban Sound 8k dataset they achieved the accuracy of 79%. Computer vision techniques have also been applied for sound classification purpose. Haomin Zhang et al. [19] used convolution neural network for robust sound event recognition. They first created the spectrograms of the audio files using Fourier transformation. Spectrograms are passed to the neural network for classification purpose. This method was used to reduce the noise and achieved the accuracy of 85% on Sound Scene database in real acoustic environments. As far as we know most of the work either maintained the time series nature of sound signal or worked signal based features of sound data. However, we used both of two techniques of feature extraction. We extracted both types of features and worked on them simultaneously. After collecting the features we applied convolution neural network on evaluated features for classification of sound. The framework diagram of our proposed methodology have been shown in figure 1.

III. Proposed System

Our proposed system works in the following order: It accepts the input audio file, convert the audio file into chunks and after processing each chunk it returns the label (from predefined classes) at the end. The Interface of our system is shown figure 2.

A. Feature Extraction

Features extraction from audio data is more complicated then other formats of the data. For each second the audio data contains the 11025 values. We first extract the characteristics from each sample in our feature extraction approach, to make fixed shape for each sample present in the raw data. MFCC is the most common feature extraction technique which is used in speech recognition and environmental sound analysis. Following steps are used to extract the MFCC features;
1: Input signal is broken into the short frames of 20-40 miliseconds. The frame is 50 % overlapped with its neighbouring frames. 2: Periodogram estimate of the power spectrum have been calculated to record the frequencies present in each frame. We first apply the discrete fourier transform on the particular frame using the following equations 1 and 2.

$$F_x(l) = \sum_{m=1}^{M} f_x(m) g(m) e^{\frac{-y 2\pi l m}{M}} \quad (1)$$

Here f_x is the time domain signal where x is in ranges number of frames. $F_x(l)$ represents the discrete Fourier transform of the particular frame, where x donates the frames number related to the time domain frame. In above equation 1 g(m) is the analysis window of the M sample long.

$$S_x(l) = \frac{1}{M} |F_x(l)|^2 \quad (2)$$

Periodogram estimate of the power spectrum is represented by $S_x(l)$, and k represented the length of DFT. 3: To sum up the energy present in each filter, Mel filter bank have been applied to power spectra of frames. By doing so, the number of features have been reduced because of the summation of periodogram bins. Humans are much better to perceive the small changes in pitch at low frequency as compare to high frequency. So to implement this attribute on our features we have used Mel Scale. Mel scale is used to relate perceived

frequency of a tone with the actual frequency. Frequency is converted to Mel using the below equation 3.

$$\text{M(freq)= } 1125ln(1 + \frac{freq}{700}) \quad (3)$$

4: As human does not perceive the loudness in the linear scale, so logarithm of filter bank energies have been taken. 5: To decorrelate the overlapping frames the discrete cosine transformation have been taken of log filter bank energies. This autocorrelation creates a problem for some classifiers like hidden markov chain but we apply neural network classifiers, so autocorrelation is not a big issue for this. 6: The MFCC are extracted from the amplitudes of the resulting spectrum. We have also extracted the Mel-scaled spectrograms, chromagram, spectral-contrast and tonal centroid features. We have applied filter banks in our second category of feature extraction. This approach is used to keep time-series attribute of the data. At the end we have 193 features which is the combination of the above described methods.

B. Network Architecture and Training Parameters

After getting 193 features from each audio file of dataset, we have developed deep neural network classifier based on the convolution neural network. Our architecture comprises

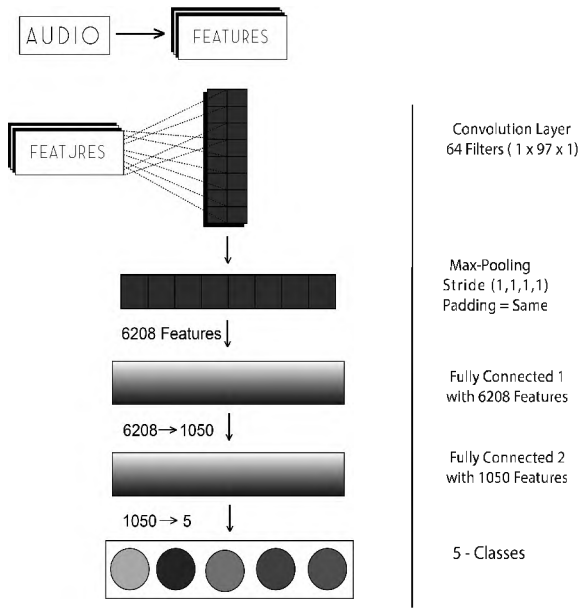

Fig. 3. Network Architecture

1 convolution layer and 2 fully connected layers as shown in figure 3. The input to the first convolution layer have (1x10x1x64) form of input data. Where (1x10x1) shows 1D data of first row, contains 10 columns of data and having 1 number of channel, and 64 shows the number of filters applied on training data. After applying the first convolution layer we get 6208 audio features. These features are passed to first fully connected layer through which we get 1050 output features. Input to our second fully connected layer is 1050 features through which we get the 5 neurons. After the third fully connected layer we put the softmax normalization layer to get the probabilistic score against each category using equation 4.

$$\text{Softmax score: } P_y = \frac{e^{p_y}}{\sum_x e^{p_x}} \forall\ x\ in\ \{1, 2..L\} \quad (4)$$

Here P is probability score against each label L present in our model. After getting the exponent of each class score, the sum of obtained score have been divided to get the probability against each category. After that by applying ArgMax we get the class label with highest probability.

IV. Dataset Collection

Most of the researchers have used UrbanSound 8k and ESC-50 for classification of urban sound. Urban Sound 8K dataset consists of 8732 sounds from 10 classes of different sources of sounds: dog bark, air conditioner, gun shot, drilling children playing, engine idling, car horn, jackhammer, siren and street music. ESC consists of 2000 sound recordings from 50 classes. These 50 classes are generalized into 5 main categories: animals, human, interior sounds, natural sounds and exterior or urban sounds. As our work is related to security and surveillance so, we have collected data on our own. We have total 5 classes, which are horn, gunshot, blast, crowd, siren. The total audio files for each category are approximately 100. So, we have total of 500 audio files, their length varies from 5 to 20 seconds. The type of audio files we are processing is .wav. We have sub grouped the 500 audio recordings into 5 folds. Each fold contains 100 audio recordings, 20 from each category. For Siren and Horn we took some recordings from urban sound and Esc-50 dataset. For crowd, gunshot and blast we have downloaded the videos from YouTube, containing sounds of the blast, gunshot or crowd. Then we trimmed the videos containing sound and finally we converted the videos into wav format. Data distribution of generated dataset have been shown in Table I.

Category	No of Samples	Duration
Gun Shot	100	5 to 20 seconds
Blast	100	5 to 20 seconds
Horn	100	5 to 20 seconds
Siren	100	5 to 20 seconds
Crowd	100	5 to 20 seconds

TABLE I: Dataset Distribution

V. Results

For training our proposed model tensor flow deep learning frame work have been utilized. The training have been done on the NVIDIA 1080 Ti Gpu which have the memory capacity of 11 Gb. The system took approximately 2 hours for complete training of the architecture. As mentioned earlier, dataset has been divided into 5 folds. For a particular iteration we have a 400 training audio files and 100 validation files as depicted in

Training	Validation
400	100

TABLE II: Training and validation data for each iteration

Table II. The main reason to use the cross validation is to avoid the model from over-fitting test set by find the error rate over test set. One of the major problem in training of deep learning models is to avoid the over-fitting. The overall accuracy of a model will also be effected if the model is over fitted. The over fitting happens when model is captures noise from training set of the dataset. Here noise are data points which do not represent true properties of your dataset, but have random chances to occur. By learning those parameters we are able to create more flexible model, but on the risk of over fitting. So to discourage the process of learning complex data points and construct the model based on these parameters, regularization have been utilized to avoid process of over fitting. A simple representation of linear regression is represented in below equation 5.

$$\text{X: } \approxeq \beta 0 + \beta 1X1 + \beta 2X2... + \beta_n X_n \tag{5}$$

X is used to represent learned relation, where β is used for estimation of coefficients on predictor X. As the fitting process involves loss function, so we have used mean square error loss function as depicted in equation 6.

$$\text{Mean Square Error: } \frac{1}{i}\sum_{j=1}^{i}(\hat{X}_j - X_j)^2 \tag{6}$$

Here X^- represents the vectors after adding the β value for n number of prediction and X presents the true value against each sample. It is subjected to calculate loss based on predictions compared to actual value. We have set the 0.01 learning rate and 0.7 beta for regularization. Adam Optimization have been used for the optimization of the weights. We have trained our model till the 5000 epochs. Training is stopped at that point because we get the maximum loss convergence. Figure 4 illustrates this statement clearly. Since our model is trained till 5000 epochs, so we calculate the loss and validation accuracy after each 1000 epochs. Table III shows the accuracy on validation data after every 1000 epochs. We have achieved maximum 94.6% accuracy on our

Epoch	Accuracy
100	25.81 %
1000	67.54 %
2000	79.34 %
3000	85.12 %
4000	91.34 %
5000	94.6 %

TABLE III: Accuracy on different epochs

validation data. The total processing time of our system is approximately one to two seconds. We have also evaluated

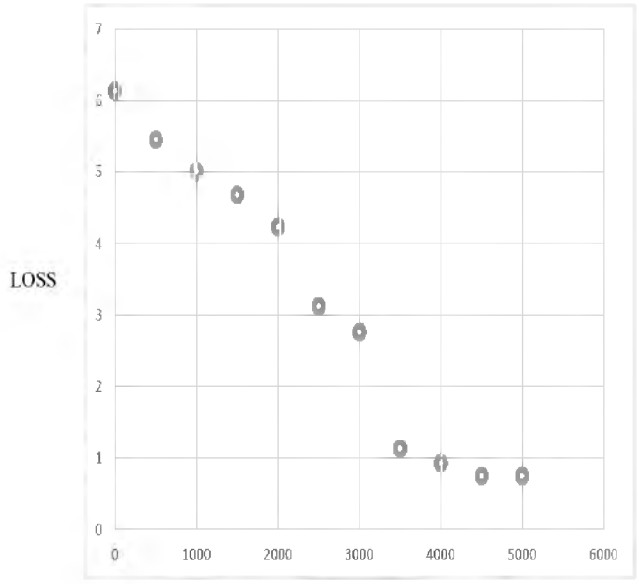

Fig. 4. Loss convergence

our dataset on different architectures of deep neural network. We have trained the simple deep neural network model as well as the recurrent neural network on our generated dataset. The results have been shown Table IV. Overall training and

Approaches	Accuracy
Deep Neural Network	64.37 %
Recurrent Neural Network	57.23 %

TABLE IV: Accuracy on different neural network approaches

validation accuracy of our trained model has been shown in table V. Table VI shows that our methodology achieved much

Training	Validation
94.6 %	92.2 %

TABLE V: Accuracy during training and validation

better results as compared to the other techniques using the same dataset. We have also evaluated our model using the

Reference	Accuracy	Methodlogy
[18]	79	Data augmentation along CNN
[19]	85	Image spectrogram along CNN
Ours	94.6	Time, Frequency features along CNN

TABLE VI: Accuracy comparison on different methodologies

confusion matrix. Figure 5 shows the the confusion matrix of our trained model at our self generated dataset.

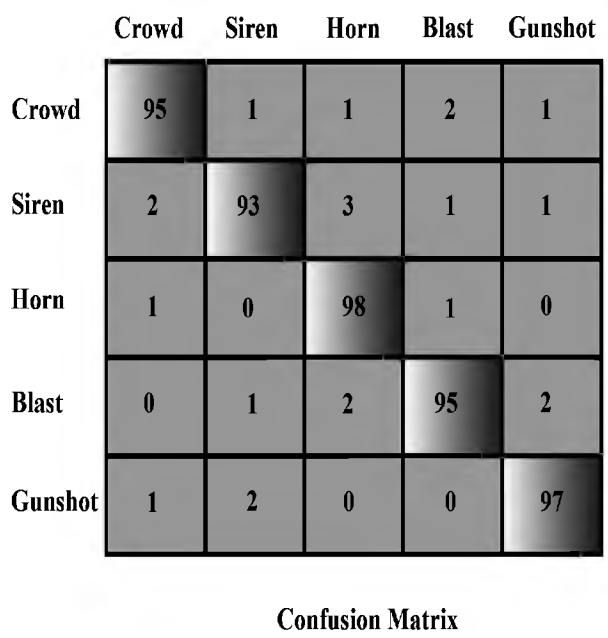

Fig. 5. Confusion matrix of validation data

VI. Conclusion

In this paper, we have proposed a CNN based model which classifies suspicious sound events. Distinct audio features including: MFCC, chroma, melspectro-gram, spectral contrast and torrentz techniques have been applied to attain state of the art results. Proposed system extracts time and signal based characteristics which are fed to CNN for classification. Proposed CNN network comprise one convolution layer and two fully connected layers. As our work is security related, so we collected relevant data on our own. Commonly used sound categories i.e. siren, horn, gun shot are elected from available datasets in addition to data generated for bomb and crowd. Its applications vary from surveillance, sound retrieval, music recommendation to music transcription.

ACKNOWLEDGMENT

We would like to express our earnest gratitude to National Center of Artificial Intelligence Fund for full supporting our research work. The authors would also like to appreciate full team (colleagues & management) and organization (KICS) for their support, dedication, technical sessions and knowledge sharing effort.

References

[1] Radhakrishnan R, Divakaran A, Smaragdis A. Audio analysis for surveillance applications. InApplications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on 2005 Oct 16 (pp. 158-161). IEEE.

[2] Chu S, Narayanan S, Kuo CC. Environmental sound recognition with timefrequency audio features. IEEE Transactions on Audio, Speech, and Language Processing. 2009 Aug;17(6):1142-58.

[3] Khan MZ, Hassan MA, Hassan SU, Khan MU. Semantic Analysis of News Based on the Deep Convolution Neural Network. In2018 14th International Conference on Emerging Technologies (ICET) 2018 Nov 21 (pp. 1-6). IEEE.

[4] Jabeen S, Khan G, Naveed H, Khan Z, Khan UG. Video Retrieval System Using Parallel Multi-Class Recurrent Neural Network Based on Video Description. In2018 14th International Conference on Emerging Technologies (ICET) 2018 Nov 21 (pp. 1-6). IEEE.

[5] Salamon J, Jacoby C, Bello JP. A dataset and taxonomy for urban sound research. InProceedings of the 22nd ACM international conference on Multimedia 2014 Nov 3 (pp. 1041-1044). ACM.

[6] A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Brisbane, Australia, Apr. 2015, pp. 151155.

[7] J. Salamon and J. P. Bello, Unsupervised feature learning for urban sound classification, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Brisbane, Australia, Apr. 2015, pp. 171175.

[8] J. Salamon and J. P. Bello, Feature learning with deep scattering for urban sound analysis, in Proc. 2015 23rd Eur. Signal Process. Conf., Nice, France, Aug. 2015, pp. 724728.

[9] J. T. Geiger and K. Helwani, Improving event detection for audio surveillance using gabor filterbank features, in Proc. 23rd Eur. Signal Process. Conf., Nice, France, Aug. 2015, pp. 714718.

[10] Piczak KJ. ESC: Dataset for environmental sound classification. InProceedings of the 23rd ACM international conference on Multimedia 2015 Oct 13 (pp. 1015-1018). ACM.

[11] Guo G, Li SZ. Content-based audio classification and retrieval by support vector machines. IEEE transactions on Neural Networks. 2003 Jan;14(1):209-15.

[12] Dennis J, Tran HD, Chng ES. Image feature representation of the subband power distribution for robust sound event classification. IEEE Transactions on Audio, Speech, and Language Processing. 2013 Feb;21(2):367-77.

[13] Piczak KJ. Environmental sound classification with convolutional neural networks. InMachine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on 2015 Sep 17 (pp. 1-6). IEEE.

[14] Tokozume Y, Harada T. Learning environmental sounds with end-to-end convolutional neural network. InAcoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on 2017 Mar 5 (pp. 2721-2725). IEEE.

[15] Sailor HB, Agrawal DM, Patil HA. Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification. Proc. Interspeech 2017. 2017:3107-11.

[16] Kumar A, Khadkevich M, Fgen C. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018 Apr 15 (pp. 326-330). IEEE.

[17] Gemmeke JF, Ellis DP, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, Ritter M. Audio set: An ontology and human-labeled dataset for audio events. InAcoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on 2017 Mar 5 (pp. 776-780). IEEE.

[18] Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters. 2017 Mar;24(3):279-83.

[19] Zhang H, McLoughlin I, Song Y. Robust sound event recognition using convolutional neural networks. InAcoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on 2015 Apr 19 (pp. 559-563). IEEE.