Security Threat Sounds Classification using Neural Network

Shivam Agarwal
BML Munjal University
School of Engineering & Technology
Gurgaon,India
shivam.agarwal.16cse@bmu.edu.in

Kiran Khatter
BML Munjal University
School of Engineering & Technology
Gurgaon,India
kiran.khatter@bmu.edu.in

Devanjali Relan
BML Munjal University
School of Engineering & Technology
Gurgaon,India
devanjali.relan@bmu.edu.in

2021 8th International Conference on Computing for Sustainable Global Development (INDIACom) | 978-93-80544-43-4/21/$31.00 ©2021110.1109/INDIACom51348.2021.00122

***Abstract*—Sound plays a key role in human life and therefore sound recognition system has a great future ahead. Sound classification and identification system has many applications such as system for personal security, critical surveillance, etc. The main aim of this paper is to detect and classify the security sound event using the surveillance camera systems with integrated microphone based on the generated spectrograms of the sounds. This will enable to track security events in cases of emergencies. The goal is to propose a security system to accurately detect sound events and make a better security sound event detection system. We propose to use a convolutional neural network (CNN) to design the security sound detection system to detect a security event with minimal sound. We used the spectrogram images to train the CNN. The neural network was trained using different security sounds data which was then used to detect security sound events during testing phase. We used two datasets for our experiment training and testing datasets. Both the datasets contain 3 different sound events (glass break, gun shots and smoke alarms) to train and test the model, respectively. The proposed system yields the good accuracy for the sound event detection even with minimum available sound data. The designed system achieved accuracy was 92% and 90% using CNN on training dataset and testing dataset. We conclude that the proposed sound classification framework which using the spectrogram images of sounds can be used efficiently to develop the sound classification and recognition systems.**

Keywords— Convolutional Neural Network , Sound recognition and classification, Mel Frequency Cepstral Coefficient

I. INTRODUCTION

Audio classification is a task to classify audio recordings into different classes. There are many important applications related to sound and audio processing and classification such as real Environment Sound Classification (ESC), bird sound classification etc.. The sound detection and classification task are very different as compared to Automatic Speech Recognition (ASR). For example, in ASR, speech is converted to text, whereas in environment sound classification system there is no speech associated with it and it's just a sound. As sound features differ drastically from speech sounds thus sound detection system is different and at the same time complex as compared to ASR models.

The main objective of the paper is to provide intelligence to the camera so that it can identify the sounds in real time and classify the emergency and security sound into one of various categories like a gun shot, smoke, fire alarm, screams of a person or window breaking. A security sound event can have multiple or long repeating sound events such as fire alarm or burglary alarm, or a single short sound event such as glass breaking, gun shots or explosions. In these sound events where a detectable sound clip is present is called security sound event. This study is conducted to show that it is possible to equip the cameras with both the video and audio intelligence.

At instances, the intelligent camera identifies emergency and/or security sounds and thus able to classifies it into one of the above categories. On detection and classification of sound into correct category, it can able to send an alert to the appropriate authority so that depending on the situation an appropriate action can be taken.

In real life these security sounds happen simultaneously due to multiple reasons. Moreover, there can be multiple sources of sound such as due to noise in background or echo. Moreover, during a threatening event, there might be different or multiple sound events all are overlapping each other [1]. For instance, there might be gunfire, glass breaking as well as screams at the same time and in another event, there might be fire alarms, smoke alarms as well as screams and glass breaking. In another example there might be sounds from thunderstorm and heavy winds, which overlaps with security/ emergency sounds. Moreover, during a burglary there might be glass breaking, screams and objects breaking noises all overlaps. Thus, when the sound from multiple sources overlaps, the detection and classification task becomes even more complex. Humans can correctly recognize and categorize the sound event even in the presence of noisy environment. So, the present security sound detection system requires human intervention for labeling and classifications of sound events. Thus, designing a security sound detection system is still a difficult and complex task specially when the sounds event are mixed with noisy background and need to be solved. Once sound event is detected it becomes easier to identify the event that might be happening and appropriate action can be taken.

Any sound classification system consists of mainly two components - audio feature extraction techniques and classification system. The most widely used audio feature

extraction technique is the Mel Frequency Cepstral Coefficient (MFCC) [2]

The main aim of the paper is to design accurate security sound event classification, as well as correctly identify overlapping sound events. The initial data set are single sound event data and the system is trained with this data. It is then tested with single as well as multiple sound events together. The aim of this paper is to create a system that is trained on a small set of training sound and it is possible to continuously add new data to the training sound to increase accuracy as well as successfully classify more events. The rest of the paper consists of Section 2 which introduces the paper and background work, Section 3 contains the methodology which explains the method used, Section 4 contains the details of the dataset used. Section 5 gives the overview of the algorithm used whereas Section 6 consists of conclusion and future scopes.

II. Review of Literature

In recent years, many authors have proposed different automatic sound recognition system in multidisciplinary fields such as environmental sounds detection [3], multimedia [4, 5], bioacoustics and multimedia sound monitoring [4, 5], intruder recognition in wildlife areas [6], audio surveillance [7] . In recent paper [8] author classified environmental sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. They used two datasets for the experiment: ESC-10 and ESC-50 and achieved the accuracy of 77% and 49% in CNN and 56% in TDSN trained on the ESC-10.

The author in [9] proposed detection system for detecting abstract signal representations using unsupervised learning. They integrated deep learning autoencoder based technique with extreme machine learning models for designing the detector system. In another paper [10], the authors develop the system to predict the bird species in the presence of background noise from audio recordings from the BirdCLEF2019 dataset of 50,000 audio recordings of 659 bird species. They adopted CNN architectures — the ResNet [11] and the inception model [12]. In similar study [13], the author performed study for the identification of bird species in audio recordings. They used BirdCLEF 2017 training dataset which consist of 36,496 audio recordings containing 1500 different bird species. They used CNN to generate features extracted from visual representations of field recordings.

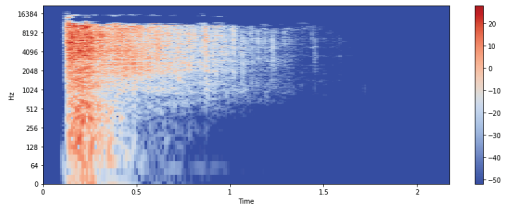
III. Proposed Methodology

The sounds are divided into one second each and converted into Mel Frequency Cepstral Coefficient (MFCC) single layer features represented in vector format. The feature vector is then used to train the convolutional neural network.

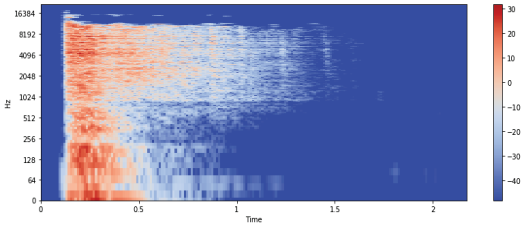
Most of the surveillance systems that is used today use facial and image recognition systems for detecting threats and raising alarms. But none of the solutions use the sound data provided by the microphones present in cameras. The solution proposed aims to use the microphones present in the cameras and intelligently classify the sound into different context such as fire, storm, theft burglary etc. The user himself can add an audio clip to detect and identify the sound that he wants to detect using his surveillance system.

The paper aims to develop a way of identifying the security sounds events based on the sound's frequency, intensity, and loudness. Every material has different characteristics and vibrates with a different frequency. This results in a distinct spectrograph for different materials can be used to identify materials and used for classification. Different objects made from the same material can show minor differences in the spectrograph, but the underlying frequency range remains the same for it. Data that is sample sounds of different objects of different materials can be fed to a Neural Network for creating a model that can identify the security sounds in the future events.

The training the system is based on data of few security sounds such as gun shot, glass breaking and fire alarms. Once the model is developed it can be applied to the security surveillance system to make it smarter and able to detect security threats not only using video but audio also. The spectrographs of car glass breaking and gun fire is shown in Fig. 1 and 2 respectively.



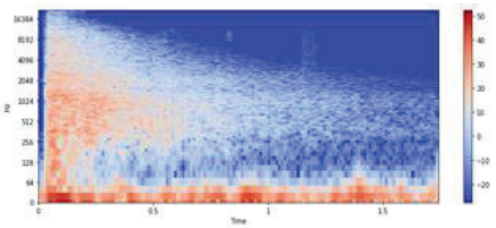
(a)



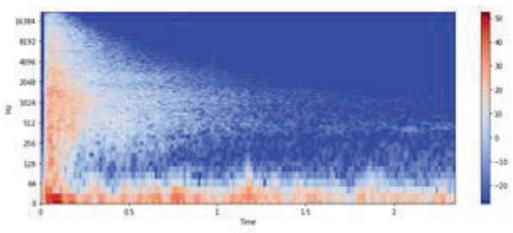
(b)

Fig. 1. (a) Spectrogram of car glass break 1 (b) Spectrogram of car glass break 2

As it is illustrated in the Figure 1 both the spectrograms are very similar to each other in terms of loudness and frequency. The fundamental shape of the spectrogram is also very similar which can be used to categorize them into one category [14].



(a)



(b)

Fig. 2. (a) Spectrogram of gunshot 1 (b) Spectrogram of gunshot 2

Figure 2. shows the spectrograms produced by the sound of two different gun shots from different guns. Figure 2 also shows that both the spectrograms are very similar to each other in terms of frequency and fundamental shape of the spectrogram is also very similar which can be used to categorize them into one category. Furthermore, the spectrographs of gunshot and glass breaking are fundamentally different in nature. The similar plot can be generated from different sounds as such alarms or explosions. In this way sound data can be converted to matrix data which can be then fed to a neural network for training it. Each row of the spectrogram can be flattened and added to its column to make a single row representing the whole sound data event in a single row which then can be used to train the neural network model as well as identify the sound event in the future event.

The training data consist of 1.05 minutes of glass break noise, 1.20 minutes of gun shots sounds and 1.39 minutes of smoke alarm sound. Each of these sound clips is then divided into a second sound each and features extraction is done using MFCC [15]. The Mel-frequency cepstral (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. [16] . The data is then flattened and converted into single rows for training the model. The model used for training is convolutional neural network with 3 hidden layers. Model used 1 batch size and 72 number of epochs in convolutional neural network for training the model.

IV. Dataset

The paper uses labeled data that was collection of the 3 sound events (glass break, gun shots and smoke alarms) to train the model. The trained model was then tested against labeled data of the same 3 sound events. The MFCC algorithm for feature extraction of sound data was used. The training dataset consist of 1.05 minutes of glass break noise, 1.20 minutes of gun shots sounds and 1.39 minutes of smoke alarm sound. Each of these sound clips is then divided into a second sound each and features extraction is done using MFCC for training purpose. The testing dataset consist of 0.40 minutes of glass break noise, 0.33 minutes of gun shots sounds and 0.5 minutes of smoke alarm sound. Each of these sound clips is then divided into a second sound each and features extraction is done using MFCC for testing purpose.

V. Convolutional Neural Network

Using the Keras library running over TensorFlow, a sequential model was built with the following specification. The convolutional neural network here was a 3-layer deep architecture with a final fully connected layer and an output prediction layer (see Figure 3). The pseudocode of proposed implementation method is shown in Figure 4. Complete workflow procedure of CNN is described in Figure 5. The first convolutional layer contained 24 filters of 5×5 size with ReLU activation. Max pooling of size 4×2 was used to reduce the dimensionality of the data and filter out the unnecessary data in the feature maps. The next layer contained the 48 filters of 5×5 size with ReLU activation. Max pooling of size 4×2 was again used in second hidden layer. To avoid the over fitting of data, dropout learning with 50% dropout probability was used to create high co-adaption among hidden layer units. The next layer contained the 48 filters of 5×5 size with ReLU activation. Max pooling of size 4×2 was again used in third hidden layer. To avoid the over fitting of data, dropout learning with 50% dropout probability was used. Before passing the information to fully connected layer, flattened the features to form a one-dimensional feature vector. A fully connected layer with 64 ReLU activations is used to process the features vector. The prediction layer was a softmax layer to predict the final class. The loss function used in the model was categorical cross entropy [17].

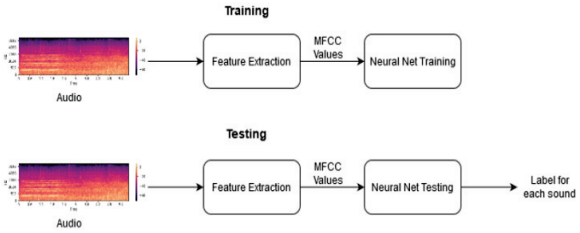


Fig. 3. Architecture

 # Defining the model

Sequential Model (Convolution 2D, Pooling, Dense, Activation)

- Input audio shape for different convolutions
- Add 2D convolutions with 24 filters and size 5*5 until filter reaches 64
- Perform sigmoid and relu activations till dense reached 64 layers.
- Involve mask of size 4*2 through max pooling operations until dense layer not reached
- Hyper-tune model with drop rate of 0.5 to avoid overfitting
- Add fully connected layer with Dense classes
- Perform Softmax activation through SGD optimizer to avoid model decay
- Define the accuracy metric and compile the model

 # Viewing model configuration

Sequential Model (get_config, get_weights)

- View generated model configuration
- View model input and output shape
- Assign the weights to individual layers
- Model is ready to train once weights get updated

 # Training

- Perform model fitting using batch size and epochs
- Validation of data also tested till training completed

#Visualizing losses and accuracy

- Visualize model loss and validation loss with obtained accuracy until epochs

Fig. 4. Pseucode

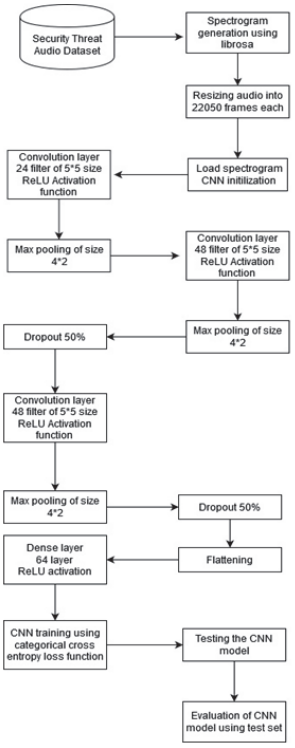


Fig. 5. *Workflow Procedure*

VI. Results and Discussions

The MFCC algorithm for feature extraction of sound data was used. The paper uses convolutional neural network algorithm to train the model. The trained model was tested against labeled data that was collected of the 3 sound events. As evaluation metric F1 score is used to measure to test the accuracy of the model.

$$F_1\ score = 2 * \frac{precision * recall}{precision + recall}$$

$$\text{precision} = \frac{\text{correct}}{\text{correct + false alarm}}$$

$$\text{recall} = \frac{\text{correct}}{\text{correct + missed}}$$

The model was successfully able to detect and classify the sounds into respective categories. The accuracy of 90% and value loss of 0.30 was recorded on the test data. Figure 6 and 7 shows validation vs. training accuracy and validation vs. training loss respectively. The precision of the model is recorded as 0.857 and the recall as 0.84 when the testing of the model was done using test data. The calculated micro F1 score was 0.848, macro F1 score as 0.815 and weighted F1 score to be 0.844. The model has reached a reasonable high accuracy and is able to classify the security sounds successfully.

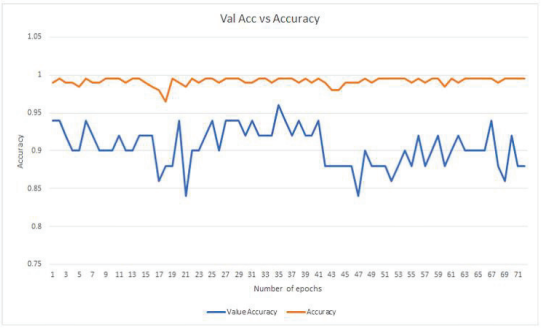


Fig. 6. Validation Accuracy vs Training Accuracy for ReLU

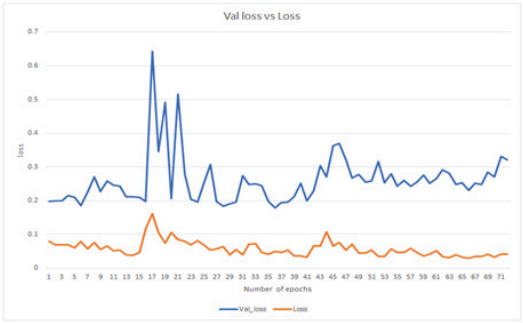


Fig. 7. Validation loss vs Training loss for ReLU

VII. Conclusion

Convolutional neural network was successfully implemented on test data sets. More classes will be added for successful sound events for detection. The system was successfully detecting and classify security sounds using trained model with a decent accuracy. The primary aim of the paper was to successfully detect security sound events with the use of minimum training data and information and the paper shows promising results.

References

[1] Agarwal, A. (2016). Sound Event Detection using a Neural Network. Jaipur: ResearchGate.

[2] Md. Sahidullah and Goutam Saha. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. Speech Communication, 54(4):543 – 565, 2012.

[3] A. S. Chu, S. Narayanan and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features", IEEE Trans. Audio Speech Language Process., vol. 17, no. 6, pp. 1142-1158, Aug. 2009.

[4] E. Wold, T. Blum, D. Keislar and J. Wheaten, "Content-based classification search and retrieval of audio", IEEE Multimedia, vol. 3, pp. 27-36, Jun. 1996.

[5] A. Rabaoui, M. Davy, S. Rossignol and N. Ellouze, "Using one-class SVMs and wavelets for audio surveillance", IEEE Trans. Inf. Forensics Security, vol. 3, no. 4, pp. 763-775, Dec. 2008.

[6] F. Weninger and B. Schuller, "Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations", Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 337-340, May 2011.

[7] M. V. Ghiurcau, C. Rusu, R. C. Bilcu and J. Astola, "Audio based solutions for detecting intruders in wild areas", Signal Process., vol. 92, no. 3, pp. 829-840, 2012.

[8] F. Khamparia, Aditya, et al. "Sound classification using convolutional neural network and tensor deep stacking network." IEEE Access 7 (2019): 7717-7727.

[9] K. Sun, J. Zhang, C. Zhang and J. Hu, "Generalized extreme learning machine autoencoder and a new deep neural network", Neurocomputing, vol. 230, pp. 374-381, Mar. 2017.

[10] Koh, Chih-Yuan, et al. "Bird Sound Classification Using Convolutional Neural Networks." CLEF (Working Notes). 2019.

[11] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 770–778 (2016)

[12] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)

[13] Kahl, Stefan, et al. "Large-Scale Bird Sound Classification using Convolutional Neural Networks." CLEF (Working Notes). 2017.

[14] Khunarsala, P. (2013). Very short time environmental sound classification based on spectrogram pattern matching. Bangkok, Thailand: Information Sciences.

[15] Parwinder Pal Singh, P. R. (2014). An Approach to Extract Feature using MFCC. Chandigarh: IOSR Journal of Engineering (IOSRJEN).

[16] Tejal Chauhan, H. S. (2013). A Review of Automatic Speaker Recognition System. Blue Eyes Intelligence Engineering.

[17] Khamparia, A. (January 2019). Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. Phagwara: IEEE.