Android Application based practical implementation of Scream Recognition System

Gwanghyun Yu[1], Linh Van Ma[1], Jonghyun Jang[1], Jin-Young Kim[1], Jinsul Kim[1]

[1] Dept. of Electronics and Computer Engineering Chonnam National University, Gwangju, Rep. of Korea
{sayney1004@naver.com}

Abstract. This study proposes an effective scream recognition system for android phones, capable to monitor emergency circumstances at the user end. Unlike traditional surveillance systems, the proposed system could be carried in pocket and while connected to the internet, being capable of detecting screams among a continuous stream of sound, it could alert the relevant authorities for any emergency situations. Given the fact that scream is usually the first response of any victim under emergency situation, this suggested automated system not only eliminates the need of continuous manual monitoring of CCTV videos but also can be more efficient because practically there is no blind spot in this, unless the user leaves his phone away. The proposed model works in real time, by capturing a stream of audio in android phones and transmitting it to server over the internet, where a deep network based system analyses the audio and detects if there is a scream. It can alert the authorities or the respective 'to be contacted' person, in case of an affirmative detection. Mel-frequency cepstral coefficients (MFCC) are adopted as audio features to create Mel-Spectrogram images, while Convolutional Neural Network (CNN) is exploited for robust detection. Various audio classes are included in the dataset for exact analysis of the event and the system is extensively tested under various indoor and outdoor environments, results of which are discussed in this research work.

Keywords: Scream-Recognition, Mel-Frequency, CNN, Android

1 Introduction

These days, due to increasing violence and the resulting potential dangers, the need of an efficient surv eillance system is inevitable. Mostly, citizens rely on traditional CCTV systems to prevent unsolicited socia l events like theft, robbery, kidnapping, rape, home violence or accidents. But many of the existing CCTV systems are not automated and requires continuous monitoring of the videos by a human and in case of any mistake or a miss there is no way to identify a crime scene. In addition, these CCTV systems are pr one to various outdoor environmental conditions. Given all these circumstances, a smartphone based surveil lance system seems a promising solution [1].

Android-based phones have got widely popular during the last decade. More than 60% of the populati on uses smartphones in developed countries like USA, Korea, Australia, Germany and UK [2]. The time smartphones were introduced, the constraints of battery and data usage were extremely high. But nowaday s, these limitations are overcome by the advancements like rapid charge system (RCS) and the supplement ary power banks. The two mentioned points are fundamental requirements of the android application in ou r proposed system.

Audio classification in the field of neural networks leads to the creation of Deep Networks [3]. With Convolutional neural network (CNN) [4] and Recurrent Neural Network (RNN) [5], these two networks a re surprisingly capable of extremely outstanding performances. Generally, CNN is the representative manne r to analyze an image and RNN is very powerful way about recurrent-sequential data like language, video, and audio. After spectrogram was discovered of audio features, these images can be trained by CNN.

Keeping given facts under consideration, an android based scream recognition system is presented whic h is convenient in its mobility and ease of use (one button start/stop), capable of exploiting Mel-spaced fre quency bins spectrogram images to be classified on CNN and proficient to provide runtime detections, wh ich can then be used to follow the appropriate actions. The system has an obvious superiority of being ha ndy, mobile and easy to use. In addition, the usage is possible in extreme physical conditions like weather, isolated space, low light conditions etc.

In order to make an efficient and detailed system which could fully exploit the capabilities of CNN, e xtensively large dataset has been used to perform the training. This dataset, recorded in various indoor and outdoor environments, contains twenty-one classes other than scream and hence could give a detailed ove rview of the under surveillance scenario, rather than mere binary classification.

2. Architecture

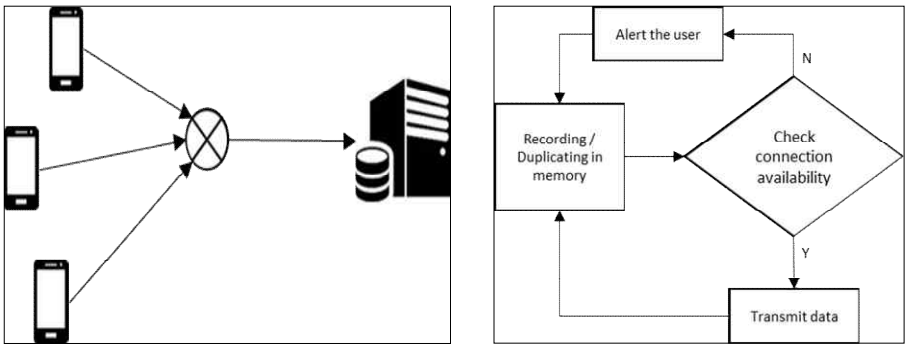

Fig. 1. Architecture of the system and User-end Application Flowchart.

The proposed system is a client-server model, in which mobile phone, as a client, can connect to the s erver using the internet. In general, the performance of any system is constrained by its hardware and opti mized by its software. So various physical constraints are involved in this architecture including unavailabil ity of the internet, keep data usage minimal hence use low-quality audio etc. Whereas, at the server end, a lthough deep networks are involved, detection must be performed runtime with a capability to alert the aut horities swiftly in case of any mishap. A proficient algorithm is proposed in this study to execute all these tasks smoothly while overcoming the constraints. The overall architecture of the system, shown in the left of Figure 1, can be divided into the main categories of smartphone, network and server system.

2.1 Smart-phone

Providing mobility and handiness to the proposed system meanwhile being able to collect and transmit audio stream to the server, the smartphone is a key element in the given system. One of the motivations behind using smartphones is their abundant number of users nowadays. Hence, use of smartphones as da ta collectors ensures no additional hardware requisite. In addition, traditional surveillance systems show hig h performance but lack mobility. Whereas the use of smartphones could easily tackle this problem.

A custom android application has been developed for this system, flowchart of which is shown in the r

ight of figure 1. With a simple GUI and user-friendly interface, this application can run as a background s ervice on the phone and can capture and transmit the audio. All this data is transmitted to the server as w ell as duplicated in phone memory, which can later be deleted manually or by the system automatically af ter a specified time.

In the proposed system, Samsung Note IV Edge has been used for testing purposes. Acoustic data spe cifications include sample rate of 48 kHz and bit rate of 256 kpbs, which are supported by this model.

2.2 Network

Socket communication is implemented between smart phone and server using the internet. As fixed IP is used for the server socket, the phone can connect via mobile data or use Wi-Fi. This part of architectur e could be a bottleneck to the system because of connection interruptions or low-speed internet, but due to the small sampling rate and truncated quality of audio, it is not likely.

2.3 Server computer

The server computer, receiving data from a mobile phone, is responsible for audio processing and class ification. Also in case of any unwanted event detection, the system is made capable to transmit a warning signal to relevant authorities or a specified person.

Initial system training has been performed in Python used TensorFlow, particulars of which are discuss ed in Section 4. Whereas, Four major steps of the testing-only part which are communication with the cli ent, spectrogram extraction of the incoming audio chunks, classification and finally take classification decisi ons for generating an alert to the authorities or a specified person.

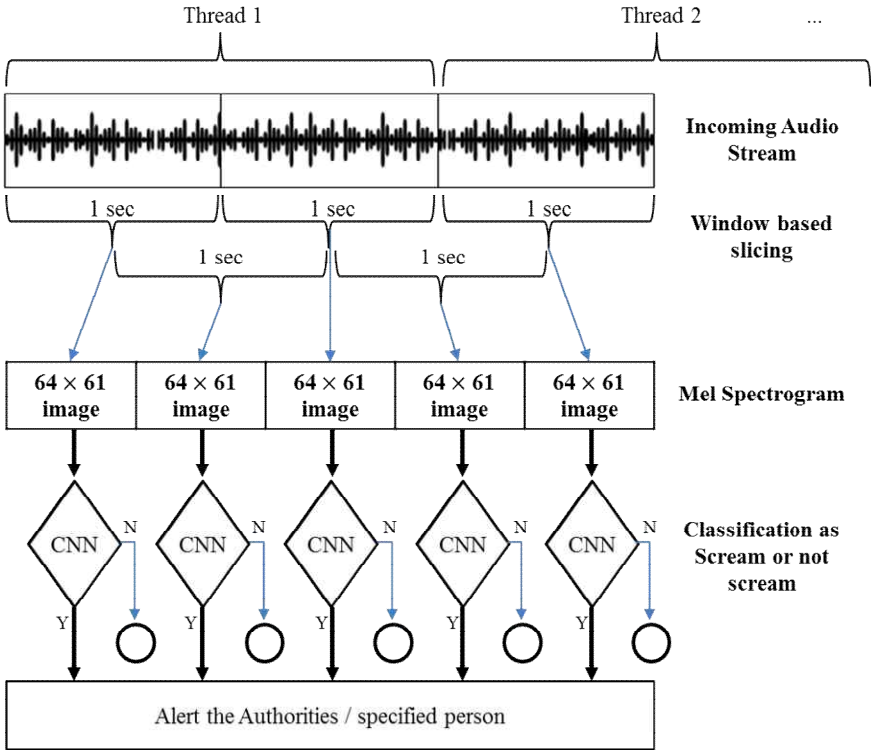

Fig. 2. Architecture of Server-end Application.

The step by step procedure carried out by server system is shown in Figure 2. The received audio stre

am is first sliced into 1-second windows with the overlapping size of 0.5 seconds. The system works on multithreading concept.

For spectrogram extraction, it is integrated into 64 mel-spaced frequency bins, and the magnitude of ea ch bin is log-transformed after getting like human earing system. While Mel-frequency cepstral coefficients (MFCC) has been handled, taking spectrogram image of each bin after using 64 of filters.

In this proposed system, classification is done by the Convolutional neural network (CNN).

3. Dataset

In order to attain higher level analysis of the under surveillance area, the dataset in this proposed syste m has been equipped with 21 classes representing various possible audio events from daily life such as gl ass shattering, phone ringing, door opening etc. Including 2798 total number of training and 280 testing so unds by ourselves. It contains 142 scream sounds for training, recorded by more than 100 men and wom en of miscellaneous ages. The application, of this proposed system as indoor as well as an outdoor applic ation is kept in consideration while recording this database hence, the sounds are recorded in varied enviro nments including classrooms, playgrounds, parking areas and sidewalks. In addition, the distance of scream ing person from the recording device is also kept variable.

4. Training

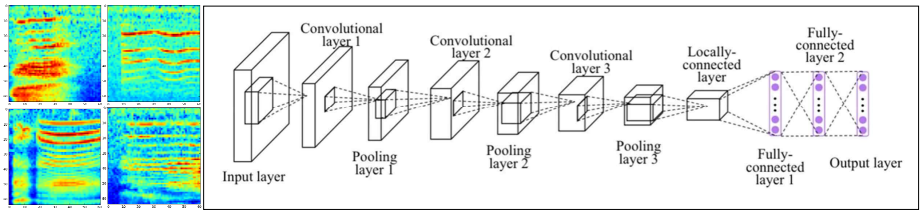

Fig. 3. 64 Mel-spaced frequency bins spectrogram images and Convolutional Neural Network.

Each 1 second sound in the dataset transformed into the image of Mel-Spectrogram [6] (Left of Figure 3), Convolutional Neural Network (CNN) has been used for classification of the audio signal.

Our baseline contains three convolutional layers and two fully connected Deep Neural network (DNN), which we compared to several architectures that had been controlled with various models, closely modele d on good image classifiers (Right of Figure 3).

Our model used Python with TensorFlow [7] on GPU. Using the Adam [8] optimizer, appropriate lear ning rate, batch sizes and ReLU activations [9]. In addition, for classification to exploit softmax and cross entropy algorithm. There are X depth with Convolutional layer and N layers, M nodes with DNN. We c hanged over X = [16, 32, 64], N = [1, 2, 3] and M = [256, 512, 1024]. Our best model had X = [32, 6 4, 128] with each convolutional layers, N = 2, M = [512, 256] units each fully connected layers and lear ning rate of 0.001.

5. Experimentation and Result

The research has been conducted on the scream data, along with other 20 other classes of sounds whic h are likely to be present in real environments, moreover, to make a practical system the dataset has been

obtained in actual physical indoor and outdoor environments like lab, home, walkways, parking lots etc. All the experiments have been carried out on NVIDIA Tesla 300 GPU with 16 GB of RAM.

Firstly, the CNN has been trained and tested fully in Python. For the whole process of wise training a nd fine-tuning, it takes around less than 1 hours. In testing, Mel-Spectrogram images have been calculated, the effectiveness of the approach has been evaluated parameters have been optimized. As shown, out of t otal 280 test cases, 259 were accurately classified (92.5 %). Contingency table summarizes the accuracy of the proposed system, in terms of scream recognition, as can be seen in Table 1. All the test screams wer e classified accurately by the system, while out of 280 non-scream test cases, only 2 (0.3 %) fell into scre am class.

Table 1. Successful scream recognition

%	Scream	Non-Scream
Scream	**98.2**	**1.8**
Non-Scream	**0.3**	**99.7**

After the analysis and evaluation, the second step would be to evaluate the live working system in real environments. In this assessment process, to capture the audio stream, Galaxy Note IV has been used. The phone uses 2 mics and can record 2 channel 256kbps audio stream in .wav format which is transferred to the server in chunks of 4-5 seconds. The evaluation has been performed in various physical environments including classrooms, computer labs, hallways, home, and walkways. While testing the system, various background noises were induced, in addition to those already naturally present. Multiple hours of testing, along with the variation in distance of phone from the user, validated the practical applications of the proposed system to be used as a real-time screaming and danger alert system.

The experimental results showed that suggested android phone based method is successful even if there is strong background noise, like music playing or people talking closely to the phone. All the screams were successfully classified by the system and as the hop size of the analysis window is 0.5 second (Figure 2) so even short length screams were exactly perceived by the system. Incorporating this approach leads to the robust detection of screams with almost no false alarm.

6. Conclusion

This paper presents an android phone based scream recognition system which exploits the powerful classification strength of Convolutional Neural Network (CNN) to present a robust and practical solution towards the problem. Mel Spectrogram image has been used as feature extraction technique whereas CNN, trained and extensively tested on a pre-recorded dataset, has been imported in Python application for live audio stream classification coming from Android phone. In both setups, experimental as well physical environment, the system performed 98.2 % accuracy satisfactorily with few misclassifications and less than 1 % false alarms. The extremely small misclassification and false alarm rates ensure the promising prospects of proposed system's application.

The system can be practically implemented as emergency identification and cell phone location be used to localize the position. In our future work, a plan has been chalked out to incorporate more feature extraction techniques to specifically enhance the system for audio coming from Android phones. It would also help to overcome misclassifications of classes and give a more accurate insight of the event happening at under surveillance area.

7. Acknowledgment

"This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2016-R2718-16-0011) supervised by the IITP(Institute for Information & communications Technology Promotion)"

8. References

1. Seo, T. W., Lee, S. R., Bae, B. C., Yoon, E., & Kim, C. S.: An analysis of vulnerabilities and performance on the CCTV security monitoring and control. *Journal of Korea Multimedia Society*, *15*(1), 93-100. (2012)
2. Poushter, J.: Smartphone ownership and internet usage continues to climb in emerging economies. *Pew Research Center*, *22*. (2016)
3. Zaheer, M. Z., Kim, J. Y., Kim, H. G., & Na, S. Y.: A Preliminary Study on Deep-Learning Based Screaming Sound Detection. In *IT Convergence and Security (ICITCS), 2015 5th International Conference on* (pp. 1-4). IEEE. (2015, August)
4. Krizhevsky, A., Sutskever, I., & Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105). (2012)
5. Graves, A., Mohamed, A. R., & Hinton, G.: Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on* (pp. 6645-6649). IEEE. (2013, May)
6. Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Slaney, M.: CNN architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* (pp. 131-135). IEEE. (2017, March)
7. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Ghemawat, S.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*. (2016)
8. Kingma, D., & Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. (2014)
9. Nair, V., & Hinton, G. E.:Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807-814). (2010)