Sound Event Recognition-Based Classification Model for Automated Emergency Detection in Indoor Environment

63

Kyungjun Min, Minhyuk Jung, Jinwoo Kim, and Seokho Chi

Abstract

Prompt emergency detection and response in indoor environments is a significant issue due to the difficulties in detecting indoor emergency events. However, current indoor monitoring tasks are mainly carried out by manual observations of occupants and such human-dependent methods generally have limitations in taking actions against emergency events. Many researchers have made much effort to develop automated indoor monitoring systems using wearable sensing device technologies and computer vision. While these methods have various advantages, there still remain challenges to be addressed for detecting indoor emergency events; for instance, wearable sensors need to be attached to a human body and occlusions make it hard to recognize the emergencies. To overcome those deficiencies, this paper proposes a sound event recognition (SER)-based indoor event classification (e.g., emergency and normal event) method with a convolutional neural network (CNN). The research consists of four main steps. First, the sound types of indoor events are determined as four emergency sounds (explosion, gunshot, glass break, and scream) and one normal sound (sleeping). Second, 692 sound data of identified events are collected from online sound data sharing services, and the preprocessing is performed. Third, SER model is developed through CNN algorithm with log-scaled mel-spectrogram features. Finally, model performance is evaluated using 5-fold cross validation. The experimental results showed that the sounds caused by indoor emergency events could be automatically recognized by the proposed method with F-score of 77.32%, which demonstrates its applicability for real emergency situations.

Keywords

Indoor environment • Emergency event • Sound event recognition • Convolutional neural network

63.1 Introduction

Emergency detection and response system has been demanded due to the spatial characteristics of indoor environments. As the indoor environment usually involves a confined area separated from the outside, numerous indoor incidents are usually identified by occupants in the building. Moreover, the interior space consists of building components (e.g., floors, walls, windows, and doors), which obstruct the line of sight [1] and emergency detection can be delayed or neglected as a result.

K. Min · M. Jung · J. Kim · S. Chi (✉)
Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, Seoul, 08826, South Korea
e-mail: shchi@snu.ac.kr

K. Min
e-mail: pen1206@snu.ac.kr

M. Jung
e-mail: archidea914@snu.ac.kr

J. Kim
e-mail: jinwoo92@snu.ac.kr

I. Mutis and T. Hartmann (eds.), *Advances in Informatics and Computing in Civil and Construction Engineering*,
https://doi.org/10.1007/978-3-030-00220-6_63

However, the indoor emergency monitoring is largely dependent on manual observations and responses by occupants. As this approach is subject to missing or overlooking the accidents in indoor environments [2], there remains a need for developing indoor monitoring systems that automatically recognize and respond to emergencies.

A number of supplementary methods have been studied such as wearable sensing device technologies (e.g., radio frequency identification and inertial measurement unit) [1, 3, 4] and computer vision [2, 5, 6]. Although the information (e.g., location and action type) of occupants tagged with sensors can be obtained by wearable device-based methods, detection of non-tagged people is not available. Moreover, the occlusions by indoor obstacles make it hard to detect emergency events with computer vision.

To overcome such deficiencies, sound recognition from the indoor events can be an alternative way for the following reasons. First, it is possible to recognize various events in a space with a few acoustic sensors, not to be tagged on every occupant. Second, within an effective detection range, the events can be detected without occlusions. Finally, it is possible to classify indoor events by extracting distinct sound patterns from each event. Thus, the purpose of this paper is to develop a sound event recognition (SER)-based classification model to distinguish emergency events from normal daily events and classify emergency types using convolutional neural network (CNN).

63.2 Preliminary Study

63.2.1 Sound Types of Indoor Emergencies

In recent years, firearm accidents and terrorism have become global issues. According to national vital statistics reports in U. S. [7], 36,562 people died from injuries caused by firearms in 2015. Moreover, 25,621 people were killed by terrorism and 54% of attack types were bombing and explosion around the world in 2016 [8]. While these events can occur both in indoor and outdoor environments, there are practical challenges in the indoor emergency detection caused by spatial characteristics of indoor environments. For instance, indoor obstacles, such as floors and walls, obscure the view for emergency detection in occlusion area. To mitigate the problems, it is required to identify the emergency events and analyze the corresponding characteristics (e.g., distinct sounds) in indoor emergencies.

The emergencies can be divided into primary events and secondary events. The primary events are directly involved with dangerous situations and they cause the secondary events. In the case of firearm accidents and terrorism, gunshot and bomb explosion are categorized into the primary events and people's screaming and window break, which are subsequently caused by the primary events, can be classified into the secondary events.

Since these events have unique sound patterns distinct from other events, the sounds of indoor events can be classified as emergency sounds, including primary and secondary events, and normal sounds. In previous research, 7 types of sound events were analyzed [9], 3 for alarming sounds (glass break, screams, and dishes sounds) and 4 for usual sounds (door clapping, ringing phone, step sounds, and door lock). In the study of [10, 11], the categories of sounds were comprised of normal sounds and critical sounds related to distress situation. In this paper, the sound types of indoor events were determined as emergency sounds (explosion, gunshot, glass break, and scream) and normal sound.

63.2.2 Sound Event Recognition

Sound event recognition (SER), the main methodology of this paper, has been recently studied in the field of sound recognition behind speech recognition and music recognition. In contrast to the other two research fields, SER aims to capture non-stationary and random sounds in daily life [12, 13] which have large variations of frequencies.

In order to recognize specific events using SER, distinct sound characteristics are extracted in terms of frequency and magnitude, which is called features. Also, various features can be selected depending on the types of the sounds (e.g., long-term/short-term and stationary/nonstationary). Then these features are fed into classification models and the models are optimized by machine learning techniques.

Previous research on the SER has been addressed for the purpose of monitoring such as surveillance system [14]. In addition, for health monitoring of elderly people, falling accident was detected using one-versus-all classifiers with nearest neighbor, support vector machine, and Gaussian mixture model [15] and sound classification and localization methods were

studied [16]. However, few researchers have focused on emergency monitoring in indoor environments. To the best of the authors' knowledge, this research is the first attempt to apply a sound recognition method to emergency monitoring in construction domain.

63.3 Research Framework

The research framework illustrated in Fig. 63.1 is comprised of four main steps: preprocessing, feature extraction, model development, and evaluation. First, the preprocessing is carried out to improve quality of the original sound data by eliminating non-sound intervals, applying normalization, and splitting into several sound clips with same durations. Second, log-scaled mel-spectrograms are extracted from the preprocessed data as features which represents the patterns of sound signal across time and frequency. Third, the SER model is developed by training the extracted features with CNN. Finally, the proposed model is evaluated using 5-fold cross validation method. The details of each step are explained in the following sections.

63.3.1 Preprocessing

The data preprocessing was carried out to improve quality of original sound data before extracting the features. First, non-sound intervals irrelevant to the characteristics of sound events were manually trimmed in order to prevent performance degradation of the model. Second, to address scale effects (i.e., different amplitude ranges) resulting from different recording conditions (e.g., environments), the amplitudes of sound data were normalized between −1 to 1. Finally, the original sound data was split into several sound clips that have same duration in order to fit a specific size of spectrogram in the next step.

63.3.2 Feature Extraction

Log-scaled mel-spectrogram was extracted from a sound clip through following steps [17]. First, short time Fourier transform was applied to convert a sound clip from time domain to frequency domain. Second, the transformed sound clip was represented as spectrogram with mel-scaled frequency bands and log-scaled magnitudes that are perceptual scales of the human auditory system. Finally, the corresponding delta of log-scaled mel-spectrogram was added as a feature, building two channels. An example of the original sound wave and extracted feature is illustrated in Fig. 63.2.

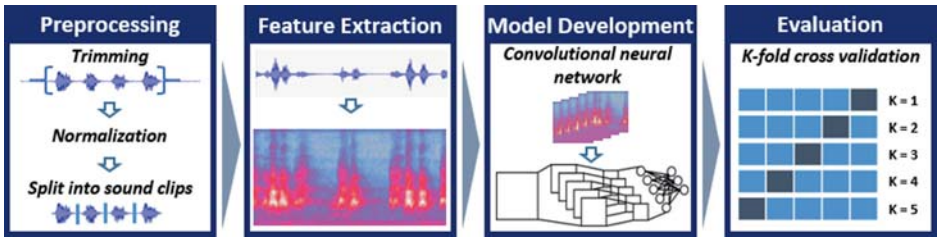

Fig. 63.1 Research framework

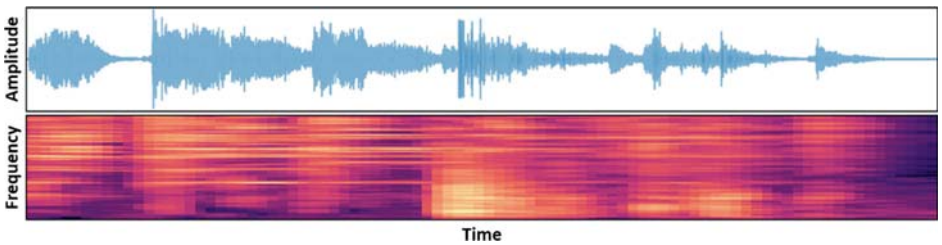

Fig. 63.2 An example of sound wave and log-scaled mel-spectrogram

63.3.3 Model Development

Convolutional neural network (CNN) was proposed as the classification model in this study, which has demonstrated outstanding performance in image recognition [18]. The typical architecture of CNN consists of convolutional layers, pooling layers, and fully-connected layers as well as input and output layers.

As it is possible to transform a sound signal to the form of two-dimensional spectrogram through the feature extraction method described in the Sect. 63.3.2, sound data can be fed as input features into CNN; a shape of input in CNN usually consists of three-dimensional data, which is a 2D matrix with channels. In addition, previous research has shown high performance of CNN in the field of SER [17, 19, 20]. The authors concluded that CNN would be an effective classifier in the sound recognition field as well as image recognition.

63.3.4 Evaluation

The procedure of model training and testing was implemented using 5-fold cross validation and the test results were represented as accumulated confusion matrix. To quantify the performance of the proposed classification model, F-score was calculated based on the precision and recall. The evaluation metrics are defined as Eq. (63.1–63.3), where the true positive (TP) is the number of cases that types of events are correctly predicted; the false positive (FP) is the number of cases that the other events are incorrectly predicted as the target events; and the false negative (FN) is the number of cases that the target events are incorrectly predicted as the other events. The precision indicates the reliability of predictions and the recall represents the ratio of correct predictions without omission. The F-score is harmonic mean of the precision and the recall. For obtaining the average precision, recall, and F-score, the micro-average method was applied to address imbalance of the number of sound clips in each class.

$$Precision = TP/(TP + FP) \tag{63.1}$$

$$Recall = TP/(TP + FN) \tag{63.2}$$

$$Fscore = 2 \cdot Precision \cdot Recall/(Precision + Recall) \tag{63.3}$$

63.4 Experiment and Result

63.4.1 Data Collection

In order to train and evaluate the SER model, total five classes were selected, including two classes (explosion and gunshot) for representing primary events in emergencies (e.g., terrorism and firearm accident), other two classes (glass break and scream) for secondary events corresponding to the primary events, and the other one class for normal event (sleeping) irrelevant to emergency events. Thus, four emergency sounds and one normal sound were determined as the classes for the model.

Then, 692 sound data was collected from online sound data sharing services, including Freesound.org and Youtube. These sound data were evenly distributed into five folds of each class for cross validation, three folds for training, one for validation, and the other one for test. Total numbers and durations of each class are as follows.

- Explosion: 54 data (287.98 s) was split into 1155 clips
- Gunshot: 354 data (447.19 s) was split into 1407 clips
- Glass Break: 140 data (471.06 s) was split into 1822 clips
- Scream: 89 data (398.24 s) was split into 1583 clips
- Normal: 55 data (508.44 s) was split into 2103 clips.

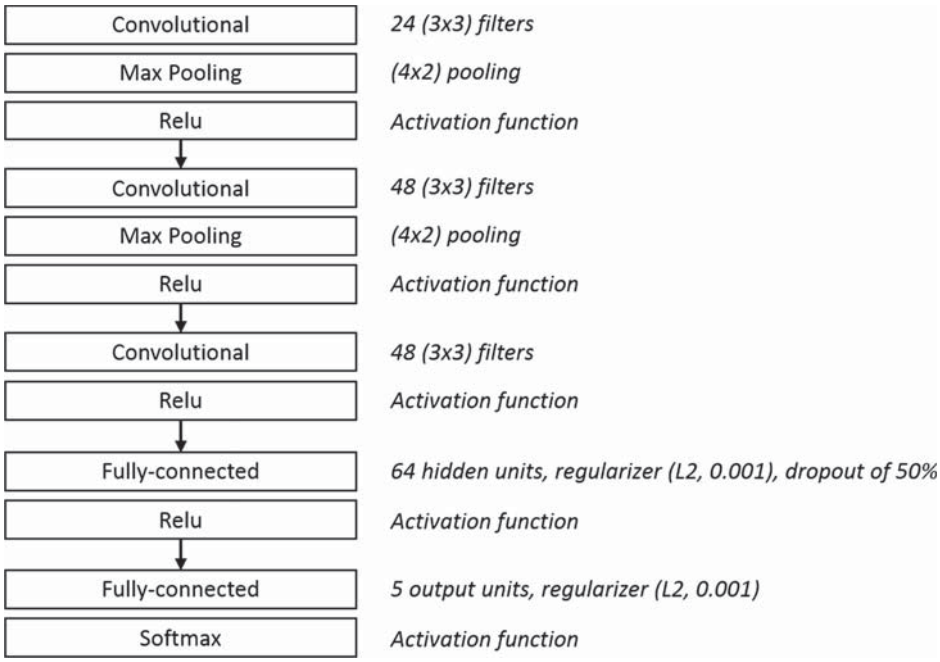

Fig. 63.3 Model configuration

63.4.2 Experiment Setup

Features. In this study, each sound clip has 0.46 s duration with the sampling rate of 22050 Hz, which has 41 overlapping frames with window size of 2048 and hop length of 256. As described in the Sect. 63.3.2, log-scaled mel-spectrogram features were extracted with three-dimensional array ($120 \times 41 \times 2$), which means 120 bands, 41 frames, and 2 channels respectively.

Model Configuration and Implementation. The model is comprised of five layers: three convolutional layers with pooling layers and two fully-connected layers. For the implementation, the research team built upon an open source code [21], which refers to the research paper [17] and [20]. Detailed information of each layer and parameters for training setup are described in Fig. 63.3. The proposed model was trained with 50 epochs and batch size of 32.

63.4.3 Results and Discussions

The accumulated confusion matrix is represented in Table 63.1. From the confusion matrix, the precision, recall, and F-score of each class and micro-average of total classes were obtained as shown in Table 63.2.

The lowest F-score of 68.67% was measured from the classification results of the normal sound class, which might have been caused by the insufficient amount of normal sound data to reflect large variation of its sound patterns. In the case of the

Table 63.1 Accumulated confusion matrix from 5-fold cross validation (unit: sound clips)

Event		Predicted label				
		Explosion	Gunshot	Glass break	Scream	Normal
True label	Explosion	1366	130	477	7	133
	Gunshot	70	1823	85	44	80
	Glass break	326	111	1432	154	86
	Scream	169	119	300	1549	33
	Normal	236	213	317	102	1261

Table 63.2 Results of precision, recall and F-score

Event	Precision (%)	Recall (%)	F-score (%)
Explosion	81.62	79.22	80.40
Gunshot	75.76	90.41	82.44
Glass break	77.82	71.08	74.30
Scream	78.29	89.77	83.64
Normal	74.65	63.58	68.67
Micro-average	77.32	77.32	77.32

explosion and glass break classes, 41% of FP in explosion class was related to the false predictions that predicted glass break class as explosion class. Similarly, 40% of FP in glass break class was caused by false predictions that predicted explosion class as glass break class. The reason of these results can be inferred that the explosion sounds that are often accompanied with glass broken sounds made it difficult to distinguish between the two classes.

Based on the evaluation and results analyses, the trained model provided high performance with the average F-score of 77.32%, while the error rate of emergency detections that classifies the emergency events as normal events need to be decreased as it would be fatal in real emergencies. Additionally, the developed model can be utilized to detect the emergency sounds from various sound mixtures with environmental noise; for example, several event classes can be selected if they have higher presence probabilities than predetermined threshold values.

63.5 Conclusions

This study proposes a SER-based emergency classification method using CNN. To implement the proposed method, emergency sounds and normal sounds data were collected and preprocessed, log-scaled mel-spectrograms were extracted as features, and the model was trained by CNN. For evaluating the proposed model, 5-fold cross validation with 5 classes was implemented and the average F-score of 77.32 was obtained. The experiment results showed that the emergency events in indoor environments can be automatically identified with the proposed method and demonstrated its potential to real life applications. Nevertheless, the proposed method still has improvement opportunities. This study assumed that one type of sound event occurs at a time, which is called single-label classification. The approach of multi-label classification is required to consider real emergencies in which various sound events can occur simultaneously. Future study will be conducted focusing on additional conditions for emergency detection such as polyphonic sound event detection with multi-labels to be applied in real life.

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (2017R1E1A2A01077468).

References

1. Cillis, F. De, Simio, F. De, Faramondi, L., Inderst, F., Pascucci, F., Setola, R.: Indoor positioning system using walking pattern classification. In: 22nd Mediterranean Conference on Control and Automation. pp. 511–516. IEEE, Palermo, Italy (2014)
2. Gómez, M.J., García, F., Martín, D., de la Escalera, A., Armingol, J.M.: Intelligent surveillance of indoor environments based on computer vision and 3D point cloud fusion. Expert Syst. Appl. **42**(21), 8156–8171 (2015). https://doi.org/10.1016/J.ESWA.2015.06.026
3. Dirafzoon, A., Lokare, N., Lobaton, E.: Action classification from motion capture data using topological data analysis. In: 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP). pp. 1260–1264. IEEE, Washington, DC (2016)
4. Spadacini, M., Savazzi, S., Nicoli, M.: Wireless home automation networks for indoor surveillance: technologies and experiments. EURASIP J. Wirel. Commun. Netw. **2014**, 6 (2014). https://doi.org/10.1186/1687-1499-2014-6
5. Tran, D.-D., Le, T.-L., Tran, T.-T.-H.: Abnormal event detection using multimedia information for monitoring system. In: 2014 IEEE Fifth International Conference on Communications and Electronics (ICCE). pp. 490–495. IEEE, Danang, Vietnam (2014)
6. Foroughi, H., Aski, B.S., Pourreza, H.: Intelligent video surveillance for monitoring fall detection of elderly in home environments. In: 2008 11th International Conference on Computer and Information Technology. pp. 219–224. IEEE, Khulna, Bangladesh (2008)
7. Murphy, S.L., Xu, J., Kochanek, K.D., Curtin, S.C., Arias, E.: National vital statistics reports. (2017). https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_06.pdf
8. Statista Homepage. https://www.statista.com/. Last accessed 17 Apr 2018

9. Istrate, D., Vacher, M., Castelli, E., Nguyen, C.-P.: Sound processing for health smart home. In: 2nd International Conference on Smart homes and health Telematic. Amsterdam, Singapore (2004)
10. Abdoune, L., Fezari, M.: Everyday life sounds database: telemonitoring of elderly or disabled. J. Intell. Syst. **25**(1), 71–84 (2016). https://doi.org/10.1515/jisys-2014-0110
11. Abdoune, L., Fezari, M.: A sound database for health smart home. In: 2014 World Congress on Computer Applications and Information Systems (WCCAIS). pp. 1–5. IEEE, Hammamet, Tunisia (2014)
12. Chachada, S., Kuo, C.-C.J.: Environmental sound recognition: a survey. APSIPA Trans. Signal Inf. Process. **3**, 1–15 (2014). https://doi.org/10.1017/ATSIP.2014.12
13. Cowling, M., Sitte, R.: Comparison of techniques for environmental sound recognition. Pattern Recognit. Lett. **24**(15), 2895–2907 (2003). https://doi.org/10.1016/S0167-8655(03)00147-8
14. Nguyen, Q., Choi, J.S.: Matching pursuit based robust acoustic event classification for surveillance systems. Comput. Electr. Eng. **57**, 43–54 (2017). https://doi.org/10.1016/j.compeleceng.2016.11.007
15. Popescu, M., Mahnot, A.: Acoustic fall detection using one-class classifiers. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 3505–3508. IEEE, Minneapolis, MN, USA (2009)
16. Goetze, S., Schroder, J., Gerlach, S., Hollosi, D., Appell, J.-E., Wallhoff, F.: Acoustic monitoring and localization for social care. J. Comput. Sci. Eng. **6**(1), 40–50 (2012). https://doi.org/10.5626/JCSE.2012.6.1.40
17. Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, Boston, MA, USA (2015)
18. 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)
19. Zhang, H., McLoughlin, I., Song, Y.: Robust sound event recognition using convolutional neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 559–563. IEEE, Brisbane, QLD, Australia (2015)
20. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. **24**(3), 279–283 (2017). https://doi.org/10.1109/LSP.2017.2657381
21. Github of Jaron Collis. https://github.com/jaron. Last accessed 17 Apr 2018