A Scalogram-Based CNN Approach for Audio Classification

Subjects: Computer Science, Interdisciplinary Applications

Contributor: Michele Scarpiniti, Raffaele Parisi, Yong-Cheol Lee

The automatic monitoring of activities in construction sites through the proper use of acoustic signals is a recent field of research that is currently in continuous evolution. In particular, the use of techniques based on Convolutional Neural Networks (CNNs) working on the spectrogram of the signal or its mel-scale variants was demonstrated to be quite successful. However, the spectrogram has some limitations, which are due to the intrinsic trade-off between temporal and spectral resolutions. In order to overcome these limitations, CNNs can employ the scalogram representation, as a proper time–frequency representation of the audio signal.

Keywords: automatic construction site monitoring (ACSM) ; environmental sound classification (ESC) ; deep learning ; convolutional neural network (CNN)

1. Introduction

In recent years significant research efforts have been made in the field of Environmental Sound Classification (ESC) [1], allowing significant results to be obtained in practical sound classification applications. This initiative has been enabled by the use of Convolutional Neural Networks (CNNs), which allowed a superior performance in image processing problems [2] to be obtained. In order to extend the use of CNNs to the field of audio processing, the audio input signal is usually transformed into suitable bi-dimensional image-like representations, such as spectrograms, mel-scale spectrograms, and other similar methods [3][4].

Recently, the approaches employed in ESC have been transferred to advancing the construction domain by converting vision-based work monitoring and management systems into audio-based ones [5][6][7]. In fact, audio-based systems not only are more cost-effective than video-based ones, but they also work more effectively in a construction field when sources are far from the light of sight of sensors, making these systems very flexible and appropriate for combining other sensor-based applications or Artificial Intelligence (AI)-based technologies [7]. Furthermore, the amount of memory and data flow needed to handle audio data is much smaller than the one needed for video data. In addition, audio-based systems outperform accelerometer-based ones since there is no need to place sensors onboard, thus promoting 360-degree-based activity detection and surveillance without having an illumination issue [8].

Such audio-based systems can be successfully used as Automatic Construction Site Monitoring (ACSM) tools [7][9][10][11], which can represent an invaluable instrument for project managers to promptly identify severe and urgent problems in fieldwork and quickly react to unexpected safety and hazard issues [12][13][14][15][16].

ACSM systems are usually implemented by exploiting both machine learning (ML) and deep learning (DL) techniques [17]. Specifically, several ML approaches, including Support Vector Machines (SVMs), the k-Nearest Neighbors (k-NN) algorithm, the Multilayer Perceptron (MLP), random forests, Echo State Networks (ESN), and others, have already demonstrated their effectiveness in properly performing activity identification and detection in a construction site [5][16]. However, DL approaches generally outperform ML-based solutions providing much improved results [6]. It is expected that DL techniques including CNNs, Deep Recurrent Neural Networks (DRNNs) implemented with the Long Short-Term Memory (LSTM) cell, Deep Belief Networks (DBNs), Deep ESNs, and others can produce more suitable and qualified performances than ML ones for robustly managing construction work and safety issues.

Approaches based on CNNs have demonstrated good flexibility and considerably convincing performance in these applications. In fact, CNNs exhibit advanced accuracy in image classification [18]. In order to meet the bi-dimensional format of images, the audio waveform can be transformed into a bi-dimensional representation by a proper time–frequency transformation. The main time–frequency representation used in audio applications is the spectrogram, i.e., the squared magnitude of the Short Time Fourier Transform (STFT) [19][20]. The spectrogram is very rich in peculiar information that can be successfully exploited by CNNs. Instead of using the STFT spectrogram, in audio processing, it is very common to use some well-known variants, such as the constant-Q spectrogram, which uses a log-frequency mapping, and the mel-scale spectrogram, which uses the mel-scale of frequency to better capture the intrinsic characteristic of the human ear. Similarly, the Bark and/or ERB scales can be used, producing other variants of the spectrogram [21].

Although the spectrogram representation and its variants provide an effective way to extract features from audio signals, they entail some limitations due to the unavoidable trade-off between the time and frequency resolutions. Unfortunately, it is hard to provide an adequate resolution in both domains: a shorter time window provides a better time resolution, but it reduces the frequency resolution, while using longer time windows improves the frequency resolution but obtains a worse time resolution. Even if some solutions have been proposed to mitigate such an unwanted effect (such as the time–frequency reassignment and synchrosqueezing approach [22]), the problem can still affect the performance of deep learning methods. Moreover, the issue is also complicated by the fact that sound information is usually available at different time scales that cannot be captured by the STFT.

The scalogram was defined as the squared magnitude of the Continuous Wavelet Transform (CWT) [23]. By overcoming the intrinsic time–frequency trade-off, the scalogram is expected to offer an advanced and robust tool to improve the overall accuracy and performance of ACSM systems. In addition, the wavelet transform allows to it work at different time scales, which is a useful characteristic for the processing of audio data.

2. A Scalogram-Based CNN Approach for Audio Classification in Construction Sites

In the digital era, great and increasing attention has been devoted to research on automated methods for real-time monitoring of activities in construction sites [15][24][25]. These modern approaches are able to offer better performance with respect to the most traditional techniques, which are typically based on manual collection of on-site work data and human-based construction project monitoring. In fact, these activities are typically time-consuming, inaccurate, costly, and labor-intensive [13]. In the last years, the literature related to applications of deep learning techniques to the construction industry has been continuously increasing [26][27]. In particular, many works have been published describing proper exploitation of audio data [5][16].

The work of Cao et al. in [28] was one of the first attempts in this direction. They introduced an algorithm based on the processing of acoustic data for the classification of four representative excavators. This approach is based on some acoustic statistical features. Namely, for the first time the short frame energy ratio, concentration of spectrum amplitude ratio, truncated energy range, and interval of pulse (i.e., the time interval between two consecutive peaks) were developed in order to characterize acoustic signals. The obtained results were quite effective for this kind of source; however, no other types of equipment were considered.

Paper [29] proposed the construction of a dataset of four classes of equipment and tested several ML classifiers. The results obtained in this work were aligned to those shown in [5], which compared and assessed the accuracy of 17 classifiers on nine classes of equipment. These two papers work on both temporal and spectral features extracted from audio signals. Similarly, Ref. [30] compared some ML approaches on five input classes by using a single in-pocket smartphone, obtaining similar numerical results.

Akbal et al. [14] proposed an SVM classifier. After an iterative neighborhood component analysis selector chooses the most significant features extracted from audio signals, this classifier produces an effective accuracy on two experimental scenarios. Moreover, Kim et al. [7] proposed a sound localization framework for construction site monitoring able to work in both indoor and outdoor scenarios.

Maccagno et al. [31] proposed a deep CNN-based approach for the classification of five pieces of construction site machinery and equipment. This customized CNN is fed by the STFT spectrograms extracted from different-sized audio chunks. Similarly, Sherafat et al. [32] proposed an approach for multiple-equipment activity recognition using CNNs, tested on both synthetic and real-world equipment sound mixtures. Different from [31], this work implements a data augmentation method to enlarge the used dataset. Moreover, this model uses a moving mode function to find the most frequent labels in a period ranging from 0.5 to 2 s, which generates an acceptable output accuracy. The idea to join different output labels inside a short time period was also exploited in [33][34], which implement a Deep Belief Network (DBN) classifier and an Echo State Network (ESN), respectively.

Kim et al. in [35] applied CNNs and RNNs to spectrograms for monitoring concrete pouring work in construction sites, while Xiong et al. in [6] used a convolutional RNN (CRNN) for activity monitoring. Moreover, Peng et al. in [36] used a similar DL approach for a denoising application in construction sites. On the other hand, Akbal et al. [37] proposed an approach, called DesPatNet25, which extracts 25 feature vectors from audio signals by using the data encryption standard cipher and adopts a k-NN and an SVM classifier to identify seven classes.

Additionally, some other approaches also fused information from two different modalities. For example, the work in [38] used an SVM classifier by combining both auditory and kinematics features, showing an improvement of about 5% when compared to the use of only individual sources of data. Similarly, Ref. [39] exploited visual and kinematic features, while [40] utilized location data from a GPS and a vision-based model to detect construction equipment. Finally, a multimodal audio–video approach was presented in [41], based on the use of different correlations of visual and auditory features, which has shown an overall improvement in detection performance.

In addition, Elelu et al. in [42] exploited CNN architectures to automatically detect collision hazards between construction equipment. Similarly, the work in [43] presented a critical review of recent DL approaches for fully embracing construction workers' awareness of hazardous situations in construction sites by the employment of auditory systems.

Most of the DL approaches described in this section work on the spectrogram extracted from audio signals or some variants, such as the mel-scaled spectrogram. However, the idea of exploiting different time scales (which is an intrinsic property of audio signals) can be used to improve the overall accuracy of such methodologies. For this purpose, the use of scalograms can be recommended. In fact, while spectrograms are suitable for the analysis of stationary signals providing a uniform resolution, the scalogram is able to localize transients in non-stationary signals.

References

1. Bansal, A.; Garg, N.K. Environmental Sound Classification: A descriptive review of the literature. Intell. Syst. Appl. 2022, 16, 200115.
2. Zaman, K.; Sah, M.; Direkoglu, C.; Unoki, M. A Survey of Audio Classification Using Deep Learning. IEEE Access 2023, 11, 106620–106649.
3. Demir, F.; Abdullah, D.A.; Sengur, A. A New Deep CNN Model for Environmental Sound Classification. IEEE Access 2020, 8, 66529–66537.
4. Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP 2015), Boston, MA, USA, 17–20 September 2015; pp. 1–6.
5. Lee, Y.C.; Scarpiniti, M.; Uncini, A. Advanced Sound Classifiers and Performance Analyses for Accurate Audio-Based Construction Project Monitoring. ASCE J. Comput. Civ. Eng. 2020, 34, 1–11.
6. Xiong, W.; Xu, X.; Chen, L.; Yang, J. Sound-Based Construction Activity Monitoring with Deep Learning. Buildings 2022, 12, 1947.
7. Kim, I.C.; Kim, Y.J.; Chin, S.Y. Sound Localization Framework for Construction Site Monitoring. Appl. Sci. 2022, 12, 783.
8. Sanhudo, L.; Calvetti, D.; Martins, J.; Ramos, N.; Mêda, P.; Gonçalves, M.; Sousa, H. Activity classification using accelerometers and machine learning for complex construction worker activities. J. Build. Eng. 2021, 35, 102001.
9. Jungmann, M.; Ungureanu, L.; Hartmann, T.; Posada, H.; Chacon, R. Real-Time Activity Duration Extraction of Crane Works for Data-Driven Discrete Event Simulation. In Proceedings of the 2022 Winter Simulation Conference (WSC 2022), Singapore, 11–14 December 2022; pp. 2365–2376.
10. Sherafat, B.; Ahn, C.R.; Akhavian, R.; Behzadan, A.H.; Golparvar-Fard, M.; Kim, H.; Lee, Y.C.; Rashidi, A.; Azar, E.R. Automated Methods for Activity Recognition of Construction Workers and Equipment: State-of-the-Art Review. J. Constr. Eng. Manag. 2020, 146, 03120002.
11. Rao, A.; Radanovic, M.; Liu, Y.; Hu, S.; Fang, Y.; Khoshelham, K.; Palaniswami, M.; Ngo, T. Real-time monitoring of construction sites: Sensors, methods, and applications. Autom. Constr. 2022, 136, 104099.
12. Zhou, Z.; Wei, L.; Yuan, J.; Cui, J.; Zhang, Z.; Zhuo, W.; Lin, D. Construction safety management in the data-rich era: A hybrid review based upon three perspectives of nature of dataset, machine learning approach, and research topic. Adv. Eng. Inform. 2023, 58, 102144.

13. Navon, R.; Sacks, R. Assessing research issues in Automated Project Performance Control (APPC). Autom. Constr. 2007, 16, 474–484.

14. Akbal, E.; Tuncer, T. A learning model for automated construction site monitoring using ambient sounds. Autom. Constr. 2022, 134, 104094.

15. Meng, Q.; Peng, Q.; Li, Z.; Hu, X. Big Data Technology in Construction Safety Management: Application Status, Trend and Challenge. Buildings 2022, 12, 533.

16. Rashid, K.M.; Louis, J. Activity identification in modular construction using audio signals and machine learning. Autom. Constr. 2020, 119, 103361.

17. Jacobsen, E.; Teizer, J. Deep Learning in Construction: Review of Applications and Potential Avenues. J. Comput. Civ. Eng. 2022, 36, 1010.

18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 18–22 June 2015; pp. 1–9.

19. Wyse, L. Audio Spectrogram Representations for Processing with Convolutional Neural Networks. In Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN, Anchorage, AK, USA, 17–18 May 2017; pp. 37–41.

20. Dörfler, M.; Bammer, R.; Grill, T. Inside the spectrogram: Convolutional Neural Networks in audio processing. In Proceedings of the 2017 International Conference on Sampling Theory and Applications (SampTA), Bordeaux, France, 8–12 July 2017; pp. 152–155.

21. Traunmüller, H. Analytical expressions for the tonotopic sensory scale. J. Acoust. Soc. Am. 1990, 88, 97–100.

22. Auger, F.; Flandrin, P.; Lin, Y.T.; McLaughlin, S.; Meignen, S.; Oberlin, T.; Wu, H.T. Time-Frequency Reassignment and Synchrosqueezing: An Overview. IEEE Signal Process. Mag. 2013, 30, 32–41.

23. Mallat, S. A Wavelet Tour of Signal Processing: The Sparse Way, 3rd ed.; Academic Press: Cambridge, MA, USA, 2009.

24. Sacks, R.; Brilakis, I.; Pikas, E.; Xie, H.; Girolami, M. Construction with digital twin information systems. Data-Centric Eng. 2020, 1, e14.

25. Deng, R.; Li, C. Digital Intelligent Management Platform for High-Rise Building Construction Based on BIM Technology. Int. J. Adv. Comput. Sci. Appl. 2022, 13, 1057–1067.

26. Mansoor, A.; Liu, S.; Ali, G.; Bouferguene, A.; Al-Hussein, M. Scientometric analysis and critical review on the application of deep learning in the construction industry. Can. J. Civ. Eng. 2023, 50, 253–269.

27. Garcia, J.; Villavicencio, G.; Altimiras, F.; Crawford, B.; Soto, R.; Minatogawa, V.; Franco, M.; Martínez-Muñoz, D.; Yepes, V. Machine learning techniques applied to construction: A hybrid bibliometric analysis of advances and future directions. Autom. Constr. 2022, 142, 104532.

28. Cao, J.; Wang, W.; Wang, J.; Wang, R. Excavation Equipment Recognition Based on Novel Acoustic Statistical Features. IEEE Trans. Cybern. 2017, 47, 4392–4404.

29. Jeong, G.; Ahn, C.R.; Park, M. Constructing an Audio Dataset of Construction Equipment from Online Sources for Audio-Based Recognition. In Proceedings of the 2022 Winter Simulation Conference (WSC), Singapore, 11–14 December 2022; pp. 2354–2364.

30. Wang, G.; Yu, Y.; Li, H. Automated activity recognition of construction workers using single in-pocket smartphone and machine learning methods. In Proceedings of the IOP Conference Series: Earth and Environmental Science; IOP Publishing: Bristol, UK, 2022; Volume 1101, p. 072008.

31. Maccagno, A.; Mastropietro, A.; Mazziotta, U.; Scarpiniti, M.; Lee, Y.C.; Uncini, A. A CNN Approach for Audio Classification in Construction Sites. In Progresses in Artificial Intelligence and Neural Systems; Esposito, A., Faudez-Zanuy, M., Morabito, F.C., Pasero, E., Eds.; Springer: Singapore, 2021; Volume 184, pp. 371–381.

32. Sherafat, B.; Rashidi, A.; Asgari, S. Sound-based multiple-equipment activity recognition using convolutional neural networks. Autom. Constr. 2022, 135, 104104.

33. Scarpiniti, M.; Colasante, F.; Di Tanna, S.; Ciancia, M.; Lee, Y.C.; Uncini, A. Deep Belief Network based audio classification for construction sites monitoring. Expert Syst. Appl. 2021, 177, 1–14.

34. Scarpiniti, M.; Bini, E.; Ferraro, M.; Giannetti, A.; Comminiello, D.; Lee, Y.C.; Uncini, A. Leaky Echo State Network for Audio Classification in Construction Sites. In Applications of Artificial Intelligence and Neural Systems to Data Science; Esposito, A., Faudez-Zanuy, M., Morabito, F.C., Pasero, E., Eds.; Springer: Singapore, 2023; Volume 360.

35. Kim, I.; Kim, Y.; Chin, S. Deep-Learning-Based Sound Classification Model for Concrete Pouring Work Monitoring at a Construction Site. Appl. Sci. 2023, 13, 4789.

36. Peng, Z.; Kong, Q.; Yuan, C.; Li, R.; Chi, H.L. Development of acoustic denoising learning network for communication enhancement in construction sites. Adv. Eng. Inform. 2023, 56, 101981.

37. Akbal, E.; Barua, P.D.; Dogan, S.; Tuncer, T.; Acharya, U.R. DesPatNet25: Data encryption standard cipher model for accurate automated construction site monitoring with sound signals. Expert Syst. Appl. 2022, 193, 116447.

38. Sherafat, B.; Rashidi, A.; Lee, Y.C.; Ahn, C.R. A Hybrid Kinematic-Acoustic System for Automated Activity Detection of Construction Equipment. Sensors 2019, 19, 4286.

39. Kim, J.; Chi, S. Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles. Autom. Constr. 2019, 104, 255–264.

40. Soltani, M.M.; Zhu, Z.; Hammad, A. Framework for Location Data Fusion and Pose Estimation of Excavators Using Stereo Vision. J. Comput. Civ. Eng. 2018, 32, 04018045.

41. Jung, S.; Jeoung, J.; Lee, D.E.; Jang, H.; Hong, T. Visual–auditory learning network for construction equipment action detection. Comput. Aided Civ. Infrastruct. Eng. 2023, 38, 1916–1934.

42. Elelu, K.; Le, T.; Le, C. Collision Hazard Detection for Construction Worker Safety Using Audio Surveillance. J. Constr. Eng. Manag. 2023, 149.

43. Dang, K.; Elelu, K.; Le, T.; Le, C. Augmented Hearing of Auditory Safety Cues for Construction Workers: A Systematic Literature Review. Sensors 2022, 22, 9135.

Retrieved from https://encyclopedia.pub/entry/history/show/120219