2020 8th International Conference on Information Technology and Multimedia (ICIMU)

Detection of impulsive sounds in stream of audio signals

Azizah Suliman
College of Computing & Informatics
Putrajaya Campus
Universiti Tenaga Nasional
Kuala Lumpur, Malaysia
azizah@uniten.edu.my

Batyrkhan Omarov
International Information Technology University,
Al-Farabi Kazakh National University
Khoja Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan
batyahan@gmail.com

Zhandos Dosbayev
Satpayev Kazakh National Research Technical University
Almaty, Kazakhstan
d_jandos_93@mail.ru

***Abstract* —Video analysis has become a standard feature of many security cameras. However, built-in audio analytics continues to be quite rare despite the presence of both the audio channel itself in the devices and the available computing power for processing audio data. Audio analytics has some advantages over video analytics such as cheaper devices and maintenance costs. Furthermore, when the system is running in real-time, the audio data stream is significantly smaller in volume than the data stream from video cameras and makes it more loyal requirements for the bandwidth of the data channel. Audio analytics systems can be particularly useful for urban surveillance with the start of automated broadcasting live video to the police console from the scene of an explosion and shooting. Audio analytics technologies can also be used to study video recordings and determine events. This article proposes a method for automatic detection of pulse sounds that signifies critical situation in audio signals based on Support Vector Machine learning models. The models were able to classify sounds from events such as gunshot, broken glass, explosion, siren, cry and dog barking with accuracy ranges from 95% to 81%.**

Keywords—audio analytics, impulsive sound detection, machine learning, support vector machine, audio signals

I. INTRODUCTION

Recently, automatic systems that control daily human activity have become more common [1, 2]. Their main goal is to ensure civil security, which is achieved by monitoring in public places and recognizing potentially dangerous situations. Research in the field of automatic surveillance systems is mainly focused on detecting events using video analytics [3]. In turn, acoustic monitoring can be used as an additional source of information, and its integration with video surveillance systems would increase the effectiveness of event detection [4, 5]. Audio analysis has features that in some situations would assist with solving monitoring tasks more effectively than video analysis systems, such as: a) low computational needs, b) independence from visibility conditions (for example, the presence of fog or insufficient lighting).

The aim of this work is to develop a method for detecting sounds of critical situations in the audio stream based on SVM (support vector method). In this paper, the term "critical situation" refers to an event whose characteristic sound signs may indicate acoustic artifacts (a shot, a scream, a glass crash, an explosion, a siren, etc.).

As part of the work, the minimum set of features for the machine learning model was determined, small training and test samples were formed, and a method was developed that allows determining critical situations in the audio signal.

II. LITERATURE REVIEW

To ensure the safety of the world's population, we are actively planning to implement the "safe city" system, which uses a network of video cameras and video analysis functions to quickly recognize and respond to various emergency situations and cases of law enforcement violations.

Recently, given that an increasing number of video cameras are equipped with built-in microphones, such a direction of recognizing abnormal or emergency situations as audio analytics is actively developing.

One of the most well-known commercial developments in audio analytics is the American ShotSpotter system [6]. This system has been installed in disadvantaged areas of Washington since 2006 and over the past years has localized 39,000 shots from firearms, quickly alerting police to the onset of this event.

Another example of implementing such system is the development of Audio analytical Ltd from the UK [7]. This company's sensors, designed to create a "smart home", are able to register events such as gunshots, aggressive screams, crying children, sounds of car alarms, broken glass, etc. After the event is registered, the system sends notifications to the user and security agencies for further response.

The AudioAnalytics project [8], based in the UK, provides several solutions for various use cases at once. The architecture of the proposed solutions is as follows: the CoreLogger program, running on the end user's device, allows user to receive and display/save alarm events. It works in conjunction with another part of the overall system – Sound Packs – which is nothing more than a set of different audio analytics modules [9]. The main features of these modules are detection of the following audio events:

- aggression (high-pitched conversation, shouting)
- car alarm system;
- breaking glass;
- search for keywords ("police"," help" , etc.).
- shots fired;
- cry / cry of a child.

Knowing the location of microphones and using triangulation methods [9-12], these systems accurately determine the location of the event. The undoubted advantages of audio analytics systems include the low cost of microphones compared to video cameras, the absence of "blind spots", and lower density of territory coverage compared to video surveillance. The audio stream from the microphone takes up less space than the video stream, which means it is easier to transport and process.

978-1-7281-7310-8/20/$31.00 ©2020 IEEE

An important point in the work of audio analytics systems is the means of registering signals, their geographical distribution (coverage of a wide area, because early notification for time reserve is needed), the availability of stable data transmission channels, ease of communication and interaction with users of the system [13-15].

Currently, the ideal tools that can perform these tasks are smartphones connected to data channels formed by mobile operators' cellular networks. The main advantage of this approach is that smartphones are increasingly popular in Ukraine (as well as in the world). According to the latest data, more than 30 % of the population aged 18 to 50 years use smartphones in Ukraine. Each smartphone has a microphone, an information exchange channel, a location sensor, and can be equipped with various software [16-17].

This makes the smartphone a potential candidate as a means of personal notification of the owner about the occurrence of an emergency, as well as a means of early detection and identification of an emergency. That is, if the smartphone is equipped with special software and a working data transmission channel, which is the mobile Internet, it would be able to detect sounds that identify an emergency, transmit the characteristics of these sounds, data about its location and exact time to a remotely located analytical system.

Detected emergencies may include situations involving a terrorist threat, public order violations, or various man-made accidents that are accompanied by loud explosions, sirens, and other acoustic artifacts.

There are some research gaps in detecting impulsive sounds:

1. The absence of a database of impulse sounds. To address this research gap we developed a dataset of impulsive sounds. The dataset contains 10 000 sounds, split into eight categories: gunshot, broken glass, fire, siren, explosion, cry, dog barking, fire alarm bell. Sounds of the given categories allow to train machine learning models and detect impulsive sounds immediately with high precision.

2. Noises in input data. In order to get high precision in detection of impulsive sounds, the input data must not contain noises. We removed all the noises from our dataset to improve the quality of the detection process.

3. Machine Learning Model. Proposed machine learning model is essential to get high precision in detection.

III. Materials and Methods

The system continuously records audio from the microphone. The recording is limited by a certain time limit. When this time limit is reached, the user subsystem saves the newly recorded file for possible analysis, deletes the previously saved file, and begins writing the next file.

To determine the expected location of the source of intense short-term sound, it is performed by sending requests about the current location of the client subsystems that registered the occurrence of the event. The received data about the location of client subsystems and the time of the registered event are processed using triangulation methods for determining the coordinates of the sound source.

The audio identification system accepts audio files with the identified intense short-term sound and performs the audio event recognition procedure. The procedure includes operations for buffering the signal with overlap, preprocessing the signal, extracting signal features (markers), postprocessing them, classifying them, and training the subsystem. After identifying the source of the signal, its tactical and technical characteristics are transmitted to the analysis and decision-making subsystem for predicting the affected areas. Such characteristics as the speed and range of propagation of striking factors of the source of intense short-term sound, the area of possible damage are used.

The proposed system architecture is illustrated in Figure 1. The system consists of several parts as preprocessor, first stage frame classifier, and second stage frame classifier.

Preprocessor generates a sequence of overlapped audio frames from the original audio signal and extracts, from each frame, a set of features. First Stage Frame Classifiers gives a set of label to each feature vector (frame). It is worth noting that it is possible to assign more than one label to each frame. Eventually, it aggregates frames into intervals. Second Stage Interval Classifier performs a classification at interval-level assigning to each interval a final prediction made through a Weighted Majority Voting (WMV) strategy among the frames that composed it.

The task of recognizing various disturbing audio events is divided into 2 sub-tasks [18]:

- detection (selection) of sharp pulse signals from background noise in the audio data stream;
- classification (recognition) of the detected signal to one of the types of audio events.

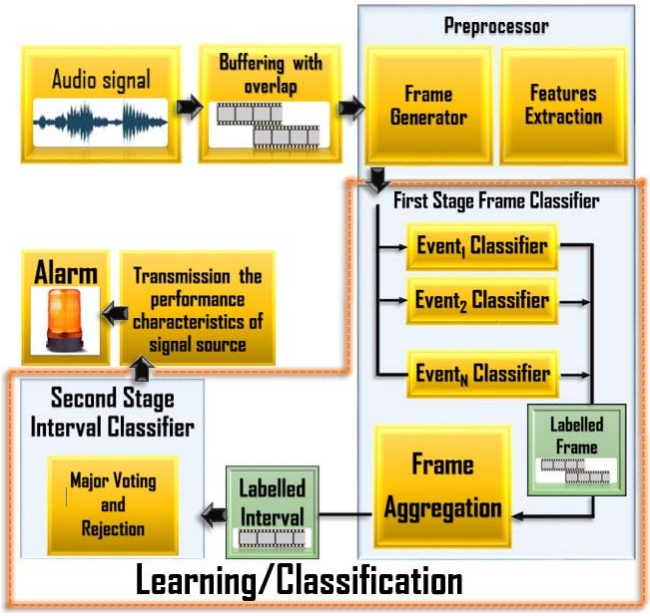


Figure 1. Architecture of audio event detection

The preprocessor consists of two main components: a frame generator, which splits the input audio signal into a sequence of overlapping audio frames of a given length, and a frame extractor, which extracts from each frame a set of features belonging to some class of those discussed in the previous section.

There are several problems with selecting the frame size and the most appropriate functions, both depending on the recognized events and the application requirements. In addition, the frame generator can only consider individual audio frames, such as those with an SPL greater than the specified W.r.t. background threshold, thereby reducing the amount of data being analyzed.

A. First Stage Frame Classifier

At this stage, audio frames represented as vectors of particular features are classified in parallel through an ensemble of classifiers of the same class: each individual classifier is trained to distinguish only one specific event. The proposed architecture should provide two basic information for adding a new classifier: the classification model (in terms of internal parameters obtained at the training stage) and an assessment of its own reliability. In our model, reliability is a measure (expressed as a floating number in the interval [0, 1], where if its value is 0, the classifier is not reliable and useless within the ensemble, otherwise, if the value is 1, the classifier can be considered completely reliable), useful for "weighing" the outputs of classifiers during the second stage of interval classification. After classification, all frames that belong to an interval (the length of which must depend on the specific events to be recognized) are aggregated through the frame aggregator component. It is worth noting that the number of tags received may differ from the number of frames involved, since more than one classifier may provide different tags for each frame.

B. Some Common Mistakes

This stage performs classification at the interval level, assigning each interval a final forecast based on a weighted majority voting strategy (WMV). For each sound class we are interested in, we link the following membership score:

$$\sigma(i) = \frac{\sum_{j=1}^{\lambda} C_j^i \cdot \alpha_i}{\lambda}, i = \{Event_1, Event_2, \dots Event_n\} \quad (1)$$

λ is the total number of classification labels in the range under consideration, Ci j is a Boolean variable that is true if the j-th label was recognized as related to an event, and ai is the reliability of a single class classifier for recognition-this is a proper event of interest.

Thus, the predicted class is estimated as:

$$\hat{C} = arg \max_{i \in \{Event_1, Event_2, \dots Event_n\}} (\sigma(i)) \quad (2)$$

If the estimate for the predicted class Cˆ is below the specified threshold, the forecast is rejected.

This classification of the second stage at the interval level allows for higher accuracy based on a simple majority voting strategy, depending on the reliability of each classifier obtained at the training stage.

In the next section, we represent the results of the current research, actually, we demonstrate samples of the collected dataset of impulsive audio events and classification of impulsive audio events.

IV. Results

A. Dataset

The performance of the proposed system was evaluated for an automated surveillance application that should be able to recognize the following events (considered " abnormal" in the observed environment): shots, screams, broken windows.

For this purpose, we built a data set from several audio samples recorded in various railway station scenarios.

The data set consists of background noise signals: pick, shot, and broken glass. Background noise was obtained indoors and outdoors to account for the characteristics of various application scenarios.

For our experiments, the signals were divided into intervals of one second (the average time length of each event of interest), and then each interval was divided into frames of 200 MS, overlapping by 50%: in particular, each interval consists of nine frames.

The composition of the data set in terms of signals, frames, and intervals is summarized in Table 1.

TABLE I. SAMPLES OF DIFFERENT IMPULSIVE SOUNDS

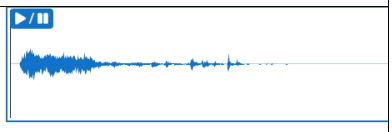
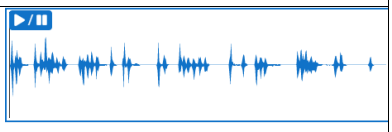
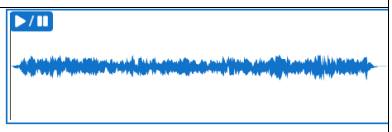
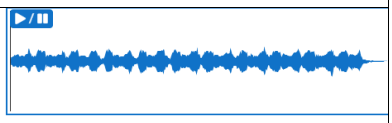
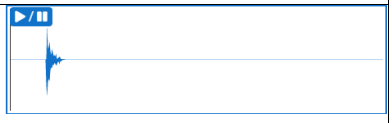
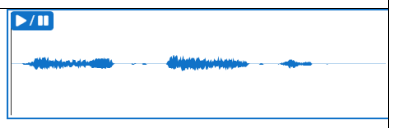
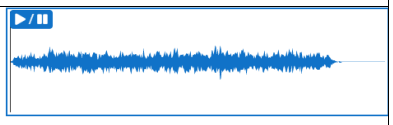
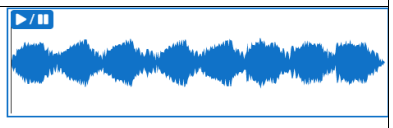
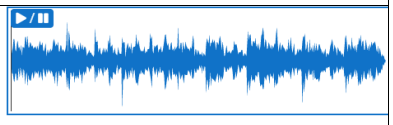
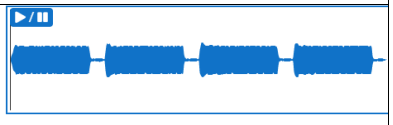
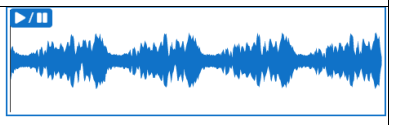
Table Column Head		
Types of impulsive sounds	***Time (sec)***	***Image***
Automobile glass shattering	3.84	
Dog barking	22.15	
Police siren	24.19	
Ambulance siren	15.41	
Constant Wail from Police Siren	56.87	
Single gun shot	3.84	
Explosion	7.78	
Artillery shell explosion	4	

Table Column Head		
Types of impulsive sounds	***Time (sec)***	***Image***
Baby crying	6.66	
Burglar alarm	11.13	
Fire alarm beeping	1.41	
Fire alarm bell	1.59	
Smoke alarm	0.99	
Fire alarm yelp	2.3	

We promptly divided the data set into two separate sections as 80% to 20%, and used the first section for the training process, and the second for the sequential evaluation phase.

B. Audio Event Classification

Second stage interval classifier: this stage performs classification at the interval level, assigning each interval a final forecast based on a weighted majority voting strategy (WMV).

As a measure of reliability for each Classifier, we chose measure F, considering it as a good compromise between accuracy and recall:

$$recision = \frac{tp}{tp + fp} \tag{3}$$

Here tp is true positive classified samples, fp is false negative classified samples.

$$Recall = \frac{tp}{tp + fn} \tag{4}$$

$$F_{measure} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

Finally, as for the deviation threshold, it was determined at the setup stage using ROC analysis aimed at maximizing the accuracy of predictions.

TABLE II. RESULTS OF IMPULSIVE AUDIO EVENT TYPE DETECTION

Event type	***Accuracy***	***Precision***	***Recall***	***F1 score***	***AUC ROC***
Gunshot	0.9178	0.9245	0.9427	0.8945	0.9748
Broken glass	0.9372	0.9765	0.9215	0.9154	0.9578
Fire	0.9435	0.9346	0.9215	0.9345	0.9576
Siren	0.9537	0.9462	0.9876	0.9642	0.9623
Explosion	0.8132	0.8254	0.8352	0.8124	0.9348
Cry	0.8635	0.8524	0.8864	0.8754	0.9467
Dog barking	0.8456	0.8325	0.8571	0.8254	0.9425
Fire alarm bell	0.8654	0.8452	0.8576	0.8457	0.9472

V. DISCUSSION

The fight against crime is one of the most important conditions for the functioning of all cities. Video surveillance plays a crucial role in this area, but it only provides a visual component. A more complete solution should take into account environmental sounds that can increase situational awareness.

Audio analytics can be used to process both archived files and online streams. In some situations, the technology is used as an alternative to video surveillance: the technology recognizes sounds in complete darkness, and microphones are much cheaper than cameras and do not require special conditions for placement and maintenance [19-21]. Sound recognition technology can be used in a variety of scenarios: recognizing individual sounds in the audio stream (screams, gunshots, footsteps, sounds of broken glass, crying), clearing audio recordings of noise, identifying people by their voices, increasing the clarity of the speaker's voice, identifying problems in the operation of mechanisms [22].

Many cities rely on video surveillance systems to fight crime. However, the report says that video surveillance alone cannot be a sufficiently effective solution for detecting and preventing crimes [23-25].

Urban security solutions most important components of today includes motion sensors, thermal imaging systems, and license plate recognition software. However, this software focused mainly on visual factors. Experts had recommended that the urban security solutions should include sound detection technology [26].

A security solution with audio transmission will allow operators to hear if a person is in trouble, give them instructions, or scare off criminals by alerting them over a loudspeaker.

Current research indicates that in 90 percent of cases of physical aggression, it is preceded by verbal aggression [27]. The value of the aggression sound detection system is that it allows security personnel to identify tension in voices and other sounds associated with anger, fear and verbal aggression. Audio analytics will help security and law enforcement officials determine which sounds are of interest and which are not related to them. The aggression sound detection software analyzes the resulting noises based on advanced algorithms and matches them with patterns. If the sound is recognized as noteworthy, the software immediately sends an alert to the security staff.

The two categories of sounds that are most crucial for analysis to ensure safety in cities are the sounds of aggression (for example, verbal abuse) and the sounds of firearms. Systems for detecting the sounds of aggression and the use of

firearms will help law enforcement agencies more effectively deal with crime.

CONCLUSION

The research demonstrated the validity and prospects of an approach to automatic detection of impulsive sounds in audio files based on the combined use of amplitude-time, spectral parameters of the signal. Further work is related to a more thorough selection and statistical analysis of low-level signal features, as well as to research the possibilities of using deep machine learning models in the task of detecting impulsive sounds.

ACKNOWLEDGMENT

The publication of this paper was funded by Innovation & Research Management Centre (iRMC), UNITEN.

REFERENCES

[1] Tharwat, A., Mahdi, H., Elhoseny, M., & Hassanien, A. E. (2018). Recognizing human activity in mobile crowdsensing environment using optimized k-NN algorithm. Expert Systems With Applications, 107, 32-44.

[2] Vanus, J., Belesova, J., Martinek, R., Nedoma, J., Fajkus, M., Bilik, P., & Zidek, J. (2017). Monitoring of the daily living activities in smart home care. Human-centric Computing and Information Sciences, 7(1), 30.

[3] Bux, A., Angelov, P., & Habib, Z. (2017). Vision based human activity recognition: a review. In Advances in Computational Intelligence Systems (pp. 341-371). Springer, Cham.

[4] Leo, M., Medioni, G., Trivedi, M., Kanade, T., & Farinella, G. M. (2017). Computer vision for assistive technologies. Computer Vision and Image Understanding, 154, 1-15.

[5] Muhammad, K., Ahmad, J., Lv, Z., Bellavista, P., Yang, P., & Baik, S. W. (2018). Efficient deep CNN-based fire detection and localization in video surveillance applications. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(7), 1419-1434.

[6] Goldenberg, A., Rattigan, D., Dalton, M., Gaughan, J. P., Thomson, J. S., Remick, K., ... & Hazelton, J. P. (2019). Use of ShotSpotter detection technology decreases prehospital time for patients sustaining gunshot wounds. Journal of Trauma and Acute Care Surgery, 87(6), 1253-1259.

[7] Weiss, A., Halevi, O., Manus, H., & Springer, D. (2018). U.S. Patent No. 10,021,457. Washington, DC: U.S. Patent and Trademark Office.

[8] http://www.audioanalytic.com/

[9] Virtanen, T., Plumbley, M. D., & Ellis, D. (Eds.). (2018). Computational analysis of sound scenes and events (pp. 3-12). Berlin: Springer.

[10] Omarov, B. (2017, October). Exploring uncertainty of delays of the cloud-based web services. In 2017 17th International Conference on Control, Automation and Systems (ICCAS) (pp. 336-340). IEEE.

[11] Gabriel, D., Kojima, R., Hoshiba, K., Itoyama, K., Nishida, K., & Nakadai, K. (2019). 2D sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system. Advanced Robotics, 33(7-8), 403-414.

[12] Omarov, B. (2017, October). Applying of audioanalytics for determining contingencies. In 2017 17th International Conference on Control, Automation and Systems (ICCAS) (pp. 744-748). IEEE.

[13] Morehead, A., Ogden, L., Magee, G., Hosler, R., White, B., & Mohler, G. (2019, December). Low Cost Gunshot Detection using Deep Learning on the Raspberry Pi. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 3038-3044). IEEE.

[14] Altayeva, A., Omarov, B., & Im Cho, Y. (2018, January). Towards smart city platform intelligence: PI decoupling math model for temperature and humidity control. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 693-696). IEEE.

[15] Alsina-Pagès, R. M., Navarro, J., Alías, F., & Hervás, M. (2017). homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring. Sensors, 17(4), 854.

[16] Wang, K., Yang, L., & Yang, B. (2017, September). Audio Event Detection and classification using extended R-FCN Approach. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017) (pp. 128-132).

[17] Choi, I., Bae, S. H., & Kim, N. S. (2019). Deep Convolutional Neural Network with Structured Prediction for Weakly Supervised Audio Event Detection. Applied Sciences, 9(11), 2302.

[18] Romanov, S. A., Kharkovchuk, N. A., Sinelnikov, M. R., Abrash, M. R., & Filinkov, V. (2020, January). Development of an Non-Speech Audio Event Detection System. In 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus) (pp. 1421-1423). IEEE.

[19] Bello, J. P., Mydlarz, C., & Salamon, J. (2018). Sound analysis in smart cities. In Computational Analysis of Sound Scenes and Events (pp. 373-397). Springer, Cham.

[20] Tseng, S. Y., Li, J., Wang, Y., Szurley, J., Metze, F., & Das, S. (2017). Multiple instance deep learning for weakly supervised small-footprint audio event detection. arXiv preprint arXiv:1712.09673.

[21] Cao, Y., Iqbal, T., Kong, Q., Galindo, M., Wang, W., & Plumbley, M. (2019). Two-stage sound event localization and detection using intensity vector and generalized cross-correlation. DCASE2019 Challenge, Tech. Rep.

[22] Cerutti, G., Prasad, R., Brutti, A., & Farella, E. (2019). Neural Network Distillation on IoT Platforms for Sound Event Detection. In Proc. Interspeech (Vol. 2019, pp. 3609-3613).

[23] Zinemanas, P., Cancela, P., & Rocamora, M. (2019). MAVD: A Dataset for Sound Event Detection in Urban Environments.

[24] Wu, D. (2019, January). An Audio Classification Approach Based on Machine Learning. In 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) (pp. 626-629). IEEE.

[25] Alías, F., & Alsina-Pagès, R. M. (2019). Review of Wireless Acoustic Sensor Networks for Environmental Noise Monitoring in Smart Cities. Journal of Sensors, 2019.

[26] McFee, B., Salamon, J., & Bello, J. P. (2018). Adaptive pooling operators for weakly labeled sound event detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(11), 2180-2193.

[27] Sammarco, M., & Detyniecki, M. (2018). Car Accident Detection and Reconstruction Through Sound Analysis with Crashzam. In Smart Cities, Green Technologies and Intelligent Transport Systems (pp. 159-180). Springer, Cham.