AI-based Situation Awareness for Smart Environment

Von der Fakultät für Ingenieurwissenschaften,
Abteilung Informatik und Angewandte Kognitionswissenschaft
der Universität Duisburg-Essen

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
(Dr. rer. nat.)

genehmigte kumulative Dissertation

von

Yang Yu

aus

Henan, China

Gutachter: Prof. Dr.-Ing. Torben Weis
Gutachter: Prof. Dr.-Ing. Weiyan Hou
Tag der mündlichen Prüfung: 5.12.2022

Abstract

In this dissertation, we investigate the requirements for situation awareness applications and the selected existing situation awareness technologies, including pedestrians positioning, pedestrians traffic detection, pedestrians walking direction detection, pedestrians number counting, emergency detection, explosion detection, detection of calling for help, detection of seeking rescue. And we innovatively proposed a situation awareness system with a set of approaches and a mechanism that balances privacy protection and high-reliability detection. The system includes two layers which are the perception layer and the representation layer. The perception layer handles the data from the multi-mode sensors and perceives the situation of the environment. The representation layer receives the perception results from the perception layers and visualizes this information. The system considers the privacy protection requirement from the system architecture level, because the functions and the data or devices used by each layer can be controlled according to the level of crisis and the level of privacy.

In the perception layer, the low-cost piezoelectric sensors, audio sensors, and cameras deployed in the environment are used according to the level of the crisis of the environment. Our piezoelectric sensors-based approach can detect if an emergency happens in the smart environment. The audio sensors or cameras are not allowed to turn on while there is no emergency. Only with piezoelectric sensors data, emergency detection, pedestrians number counting, pedestrian positioning, and pedestrians walking direction detection functions can be completed. If an emergency event is detected, the audio sensors and camera will turn on to confirm the credibility of the emergency. If indeed it is an emergency event, these additional sensors will increase the reliability of each detection function to support actions of responses to crises. In the representation layer, a 3D virtual environment is built to visualize the results of the perception layer. According to the crisis and privacy levels, the system can show pedestrians with anonymous synthetic images or real-time authentic camera images.

The proposed system fulfills the requirements of both privacy protection and situation awareness functions.

Zusammenfassung

In dieser Dissertation untersuchen wir die Anforderungen an Situational-Awareness-Anwendungen und ausgewählte existierende Technologien, einschließlich der Erkennung von Fußgängerverkehr, der Erkennung der Laufrichtung, Zählung und Positionierung von Fußgängern, sowie der Erkennung von Notfällen, Explosionen, Hilferufen und Rettungsaktionen. Wir schlagen ein innovatives Situationserkennungssystem mit einer Reihe von Ansätzen und einem Mechanismus vor, das den Schutz der Privatsphäre und eine hochzuverlässige Erkennung in Einklang bringt. Das System umfasst zwei Schichten: die Wahrnehmungsschicht und die Darstellungsschicht. Die Wahrnehmungsschicht verarbeitet die Daten der Multimode-Sensoren und nimmt die Umgebungssituation wahr. Die Darstellungsschicht empfängt die Ergebnisse der Wahrnehmungsschichten und visualisiert diese Informationen. Das System berücksichtigt die Anforderungen an den Schutz der Privatsphäre bereits in der Systemarchitektur, da die Funktionen und die Daten oder Geräte, die von den einzelnen Schichten verwendet werden, dem Grad der Krise und dem Grad der Privatsphäre angemessen gesteuert werden können.

In der Wahrnehmungseschicht werden kostengünstige piezoelektrische Sensoren, Audiosensoren und Kameras, die in der Umgebung vorhanden sind, entsprechend der Schwere der Krise verwendet. Unser auf piezoelektrischen Sensoren basierender Ansatz kann erkennen, ob in der intelligenten Umgebung ein Notfall eintritt. Die Audiosensoren oder Kameras dürfen sich nicht einschalten, wenn kein Notfall vorliegt. Bereits mit den Daten der piezoelektrischen Sensoren können die Funktionen Notfallerkennung, Fußgängerzählung, Fußgängerpositionierung und Erkennung der Laufrichtung von Fußgängern ausgeführt werden. Wenn so ein Notfall erkannt wird, schalten sich die Audiosensoren und die Kamera ein, um das Vorhandensein des Notfalls zu bestätigen. Wenn es sich tatsächlich um einen Notfall handelt, erhöhen diese zusätzlichen Sensoren die Zuverlässigkeit der einzelnen Erkennungsfunktionen, um Maßnahmen zur Krisenbewältigung zu unterstützen. In der Darstellungsschicht wird eine virtuelle 3D-Umgebung aufgebaut, um die Ergebnisse der Wahrnehmungsebene zu visualisieren. Je nach Risiko und Datenschutzanforderungen kann das System Fußgänger als anonyme synthetische Bilder oder authentische Kamerabilder in Echtzeit zeigen.

Das vorgeschlagene System erfüllt sowohl die Anforderungen an den Schutz der Privatsphäre als auch an Situational Awareness.

Contents

List of Figures

List of Tables

List of Acronyms

SA	Situation Awareness
AI	Artificial Intelligence
PAC	Principal Components Analysis
RF	Random Forests
DNN	Deep Neural Network
MF	Manual Feature
SNR	Signal-to-Noise Ratio
TDOA	Time Difference of Arrival
SoI	Time Difference of Arrival
SE	Step Event
CV	Cross Validation
MPAA	Multi Peaks Average Algorithm
MPAF	Multi Peaks Averaged Feature
DNNC	Deep Neural Network-based Classifier
IOT	Internet Of Things
CSI	Channel State Information
OFDM	Orthogonal Frequency Division Multiplexing
RSSI	Received Signal Strength Indicator
SCPI	Standard Commands for Programmable Instruments

TOA	Time Of Arrival
AOA	Angle Of Arrival
FDOA	Frequency Difference Of Arrival
RSS	Received Signal Strength
GPS	Global Positioning System
ATDOA	Angle constrained Time Difference Of ArrivalsP76
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
TP	True Positive
FP	False Positive
FN	False Negative
AED	Audio Event Detection
SVM	Support Vector Machine
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
FNN	Feed-forward Neural Network
PETs	Privacy-enhancing technologies

CHAPTER 1

Introduction

With the development of computing power of electronic computer hardware in the last decade, the algorithms like neural networks, gradient descent, and backpropagation which were considered difficult to complete tasks in a meaningful time in the 1900s that consume numerous computing resources, have regained life today. The deep learning-based[1] approaches have made the breakthroughs of the solving of images classification problems[2], object detection problems[3], target tracking problems[4], semantic segmentation problems[5], natural language processing problems[6] and so on. Furthermore, the game of Go AI, AlphaGo[7, 8], beat the human world champion in the most challenging of classic games for artificial intelligence, and it further raises people's expectations of the possibility of AI.

Society has entered an era of artificial intelligence. At the same time, it also stimulated people's demand for artificial intelligence technology to ease the work of human beings and make functions automated and intelligent to increase productivity further. Security is always a prerequisite for all activities, whether in social life or industrial production. In the AI era, people require the situation in daily life environment or industrial production environment can be perceived understood for security purposes. Moreover, the potential security threats in the environment can be predicted. Thereby, this information can be used to support decision-making and assist in protecting the safety of people's lives and property. The implementation of perception, prediction functions, and how the perception information is organized and coordinated with decision making and the actions constitute a decision model. The work of this dissertation refers to the concept of "situation awareness" and "aircrew decision model"[9].

The concept "situation awareness"[9, 10] firstly proposed in the 1980s by U.S. Air Force, which was initially invented for military purposes, has been introduced into civilian fields related to people's daily works and lives. This dissertation proposes an AI-based situation awareness system for intelligent environments. It can cover the industrial production environment and daily public living environment. This proposed system can perceive the

situation while considering protecting people's privacy in the monitored area.

The proposed situation awareness system includes a perception layer and a representation layer. The perception layer includes modules for pedestrian positioning, pedestrians traffic detection, pedestrian walking direction detection, pedestrians number counting, emergency detection, explosion detection, detection of calling for help, detection of seeking rescue, and the visualization for this information. The perception is constructed with piezoelectric sensors, audio sensors, and cameras. The use of these sensors is strictly organized and moderated according to the privacy level and privacy level. The representation layer visualizes the output of the perception layer for assisting decision-making and actions. Similarly, the image that the representation layer can express is also controlled by the privacy level. In this dissertation, a crisis-privacy status model is defined. This model is used to measure the level of privacy and crisis and to moderate the function of privacy violation sensors.

1.1 Motivation

People require that a system can detect and understand the situation in the specific area [11, 12]. With the help of sensors and AI algorithms, a system can sense the environment and classifies according to patterns. Although, as a hot spot in the research field and popular technology in the industry application, computer version-based approaches have achieved satisfactory results in environmental perception. However, this kind of technology depends on the camera's use, which introduces a problem of privacy violation. This kind of technology is like a double-edged sword. On the one hand, it solves the problems that plague people, but on the other hand, it poses a threat to people's privacy. Is that possible for a system to achieve both points that protect people's privacy and reliably percept the envelopment situation? Are privacy protection and reliable detection functions irreconcilable contradictions? Under what circumstances one of the conflicting factors can compromise with the other? How does a system know that one factor should compromise the other if the current circumstance matches the conditions?

We consider the following use case and application scenarios for the system described in this paper. We considered two scenarios: chemical factory buildings and shopping mall buildings. We considered three kinds of emergencies: hijack, fire, and explosion. The system should detect if there is an emergency that happens regardless of what type of the emergency is. We require that the safety of the people in the scenarios obtain protection when an emergency happens. Therefore, rescue personnel should arrive promptly to people (locations) who need help or are in danger when encountering an emergency. In order to achieve this goal, we need a perception system that can perceive the event that a person encounters danger. For example, when a person yells for help, falls and then knocks on

the floor, or falls and becomes inactive, it means the person is in danger and urgent need of help. Thus we require the system to detect these behavior or activity events, including falls, knocking on the floor, and yelling for help. Meanwhile, it requires our system to recognize the location of each person in the scene. Also, this monitoring system should count the number of people in the scenario and record the location of each person. While meeting the above conditions, we require people's privacy in the scenario to be protected. It requests that as long as the emergency does not occur, the use of the camera is prohibited. The acquired data should be anonymous. For data collected by sensors involving personal information, desensitization should be performed. Personal information is only permitted to be disclosed in the event of an emergency and for the purpose of saving lives.

To sum up, we want to build a system that takes both privacy protection and the implementation of the environmental situation perception function into consideration. Find a reasonable approach to balance the conflict of these two factors, and under special conditions, one factor could be compromising the other. We need to define this condition, and the system should be able to detect if this condition is fulfilled.

1.2 Contribution

The contribution of this thesis can be summarised as follows:

- A crisis-privacy status model is designed and implemented.

  The integrated situation awareness system will decide which sensors and function modules can be used according to the crisis-privacy level output by the crisis-privacy status model.

- We innovatively designed and implemented an approach that can detect the emergency depending on vibration signal, which can be used in the high-privacy situation.

  Our method can detect explosions, fires, and robbery emergencies. However, the emergency event detection method is based on the characteristics of the emergency event and thus is not limited to the above specific event types.

  This approach can trigger the crisis-privacy status model from "high-privacy" status to "high-crisis" status.

- We innovatively designed and implemented an approach to count the number of pedestrians based on vibration signals. This approach can be used on "high-privacy" status.

- We innovatively designed and implemented an approach to detect the walking direction of the pedestrians only based on vibration signals.

- We innovatively designed and implemented an approach to detect the position of the pedestrians only based on vibration signals.
- We implement subsystems based on audio and video signals to detect the situation and monitor the pedestrians that work on high-crisis status. We combined the vibration signal-based approaches and the existing video and audio-based approaches and designed an integrated system that considers both privacy-protecting and smart environment situation awareness.

1.3 Thesis Organization

The first 2 chapters introduce this thesis's background, motivation, and fundamentals. Then the system architecture is explained. Chapter 4 to chapter 10 of this thesis describes the design and implementation of each module of the situational awareness system. The approaches described in Chapter 4 to Chapter 7 are based on piezoelectric sensors. Because the vibration signal is a non-privacy violation, these approaches work on both high privacy level conditions and high crisis level conditions.

In Chapter 4, we analyze the characteristics of emergencies and propose an approach to detect if there is an emergency in a public place. This work uses piezoelectric sensors to collect data that is privacy-protecting. In particular, the system introduced in this section, serving as an emergency detecting module in the situation awareness system, will decide the crisis level. The work in this chapter is presented on *ACM International Joint Conference on Pervasive and Ubiquitous Computing and ACM International Symposium on Wearable Computers* [12] in September 2020. The crisis-privacy status model will be defined in Chapter 2.

In Chapter 5, we proposed a pedestrians number counting approach with piezoelectric sensors. Without cameras or audio sensors, it can work in both high privacy level and high crisis level conditions. The approach can detect the number of pedestrians in a 3 meters by 3 meters zone. Its function is an essential precondition of pedestrians tracking function. The work in this chapter is published in *Applied Sciences* [13] in January 2022.

In Chapter 6, we proposed a pedestrian walking direction detecting approach. The approach can detect if the pedestrian walks left or right. With the number counting approach, the situation awareness system will be able to track the pedestrians' traffic even at a high privacy level without using a camera.

In Chapter 7, we proposed a personnel positioning system. This approach can locate persons in an environment with piezoelectric sensors deployed. This work was published on *IEEE Access* in November 2020.

In Chapter 8, we describe function modules which use audio sensor (microphone) to detect explosion and distress. The explosion detection function works in high privacy level

conditions, while distress detecting works in high crisis conditions. The explosion detecting module can also trigger the awareness system into a high crisis level if an explosion is detected. The distress detecting module only works under high crisis level conditions. The work in this chapter relies on previous research, but its implementation is an important component of the proposed system. I do not claim a scientific contribution to the work in this chapter.

In Chapter 9, we describe an approach that uses a camera to track and locate pedestrians, detect falls, and count pedestrians' traffic. This approach only works under high-crisis level conditions. When a crisis or emergency happens, the priority of using all methods to ensure people's life, health, and safety is higher than all relatively unimportant factors, including privacy protection. The use of a camera in the SA system offers the most intuitive information for rescue or evacuation. Nevertheless, the use of the camera is restricted by the crisis-privacy status model. The work in this chapter relies on previous research, but its implementation is an important component of the proposed system. I do not claim a scientific contribution to the work in this chapter.

In Chapter 10, we describe a privacy-providing situation information visualization approach. This approach constitutes the representation layer for the situation awareness system. It receives the output from the perception layer and presents the information to related persons, such as the rescue team. This approach can output anonymous synthesis images in high-privacy-level or real-time reality images in high-crisis conditions. The work in this chapter is more about engineering implementation than scientific research, but the implementation is an important component of the proposed system. I do not claim a scientific contribution to the work in this chapter.

In Chapter 11, we conclude this work and provide an outlook for future development based on this thesis's findings.

CHAPTER 2

Fundamentals

2.1 Situation Awareness

The works involved in this dissertation introduce the concept of "Situation Awareness". Situation awareness(SA) is a term that initially emerged in aviation, which first introduced by Endsley [9, 14]. According to Endsley's definition, a SA system includes three key aspects: *"perception, comprehension, and projection"*, as shown in Fig. 2.1.

Nowadays, the SA has been widely referenced in a wide variety of domains, including air traffic control, driving, military operations, cyber security, etc. [15, 16].

2.1.1 Definition of Situation Awareness

The earliest definition for SA is as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future"[9, 14].

At the broadest level, "Situation Awareness" involves an environment including dynamic elements. When an intelligent object wants to recognize and understand the environment and thus make good decisions for the actions, it includes three phases: 1, perceive the basic facts in the environment; 2, understand the facts in the environment combined with knowledge and experience; 3, forecast for the future. After the intelligent agent has completed the perception, understanding, and prediction regarding the environment, it will make decisions and execute the decision, thus impacting the environment. In this way, the perception-decision-practice cycle is formed, and the whole process is repeated.

The SA provides a perception-understanding-decision framework for using artificial intelligence to solve practical problems. Under the SA framework, the first step, "perceiving the basic facts in the environment," is to perceive the physical signal in the experiment, such as humidity, temperature, shock wave amplitude, and frequency. The wide variety of

sensors presented in the market allows us to use engineering methods to perceive physical signals of the surrounding environment. After the first step, the acquired information can be used to support the following process. In the second step, the intelligent object will try to understand the environmental situation subjectively, referring to knowledge and experience. In this process, an example of how a human makes a judgment regarding the weather is given. When people feel the air pressure decreases, the humidity increases, and see the heavy clouds(step 1), they will know that it will probably rain soon(step 2 understanding and step 3 forecasting). Then he will bring an umbrella if he goes out(decision and actions).

Similarly, if an algorithm can make judgment depending on the current situation, and help people to achieve some goals, then the perception phases of the SA is completed.

In 1995, Endsley picked up a model which describes the decision-making process with a SA system [10]. As shown in Fig. 2.1, this model considers the objective factors of the environment, the situation awareness, decision, action, and feedback to the environment. It also considers the subjective factors as goals and expectations, ability or experience of the relevant person. Meanwhile, how the different factors interact and their relationships are described.

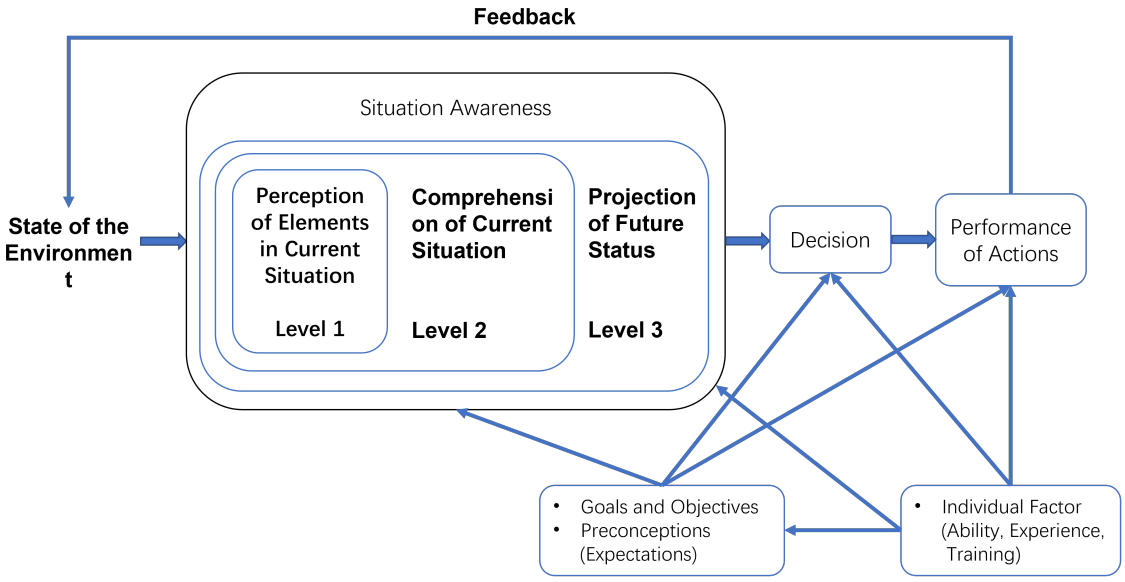


Figure 2.1: SA dynamic decision making model [10].

The output of the SA is actions. What kind of action to take is based on the understanding of the environment. After the action is executed, it will eventually affect and change the environment. At the same time, each process is cyclical and real-time. SA, serving as a model, is used to make good decisions for specific goals. According to different purposes, usage scenarios, and action execution subjects, SA has different manifestations in different use cases.

2.1.2 SA Related Works

The differences in the application of SA models in the different domains are reflected in the different meanings of the SA's three key aspects.

Initially, researchers used the SA model to analyze the aircraft pilots' behavior. Its purpose is to study what decision is optimal for a pilot in a complex real-time situation [17]. It describes the pilot's understanding of the current environment. Thus he can make good decisions or carry out appropriate actions to achieve specific goals. In this domain, the "perception aspect" is that the pilot should perceive the basic elements, including the situations of other aircraft, the terrain, the signal from the aircraft instrument, and the commander's order. These perceptions information are obtained by relying on the pilot's sense of sight, hearing, and touch. As for the "comprehension aspect," a trained pilot can understand the current situation based on the signal he or she perceives. For example, when the pilot notices the various signal on the aircraft instrument panel and combined with the sound of the aircraft engines, he can know if a malfunction is happening. In contrast, a person who does not know how to fly an aircraft can also see the signal on the instrument panel and hear the engine's sound, but it is difficult for this person to judge whether the aircraft is malfunctioning. In the "projection aspect," a well-trained pilot can predict the near future situation based on the current situation. For example, when a pilot discovers an engine failure, he can project that if it is not properly handled, the failure will expand and even cause a crash.

In around the 1990s, Endsley proposed the SA concept by studying the behavior of humans. Later, the researchers hope that the machine, rather than a human, can automatically complete some specific tasks according to the SA model. The SA system with various algorithms as the core came into being.

In recent years, the situation in cyberspace has become increasingly complex, and various types of threats and attacks flood cyberspace. People hope to monitor, capture and analyze these threats and thus take actions to avoid disruption to systems in cyberspace. In order to achieve this purpose, researchers introduced SA in the cyber security domain, which is called cyber situation awareness. Cyber SA collects, analyzes, processes, and evaluates a specific system's data to understand the system's environment in the cyber security domain. Meanwhile, it predicts the current and future situation and tries to make decisions to take action to respond to potential cyber threats [16]. In this domain, the "perception aspect" of SA is to recognize the attack occurrences, their types, attack source, and attack target. The "comprehension aspect" of SA is to know *"the attack's objectives, action, tendency, the impact on the current and future situation, and understand the rationale of the current situation."* The "projection aspect" of SA is to be aware of how the situation evolves and the possible effects [16]. Thus cyber SA system provides necessary information support for the deciders.

In ubiquitous computing, heterogeneous devices constitute distributed IoT networks and systems. Almeida proposed an SA-based system to process the IoT event logs, thus providing functions such as rapid detection and blocking of attacks, detection of whether the entire system needs to add more nodes for scalability purposes, and autonomy demand [18, 19]. Similarly, as a SA system, the three key aspects of Almeida's SA system are as follows. The "perception aspect" of this SA is to collect event logs and statistical status. This log information is generated based on log systems on each distributed node, including *"Filebeat, Winlogbeat, Metricbeat, Logstash"* [19]. The "comprehension aspect" of Almeida's ubiquitous computing SA system implements the decision tree algorithms and rule-based correlation algorithm to match the stream of incoming events against a pattern. Thus, the SA system can propose corresponding processing strategies to respond to these events. The "projection aspect" in this SA system is to avoid unwanted situations which already been handled cases during the comprehension step.

2.1.3 The Proposed Situation Awareness for Smart Environment

In this thesis, the proposed SA system perceives and comprehensives human and scenario-related information and events and provides privacy protection features. Our system's "perception aspect" uses vibration sensors, audio sensors, and cameras to collect floor vibration signals, audio signals, and video signal data. Our system's "comprehension aspect" uses AI-based algorithms to detect emergencies, pedestrians' walking direction, the position of the pedestrians, the behaviors of the pedestrians, the person in distress, and count the number of pedestrians. Our system's "projection aspect" is to evaluate the threat to the life of an individual or group of people in distress.

The proposed SA system controls the global situation of the monitoring area in real-time. When an emergency happens, it can provide adequate information for the rescue staff to support efficient and rapid rescue operations. Furthermore, as an "AI-based Situation Awareness for Smart Environments" with a privacy protection mechanism, this system innovatively balances the contradiction between privacy protection and effective monitoring. The vibration signal-based perception modules of the system fill the gap in the research field in terms of function and performance. As a whole, the proposed SA system fits into the state-of-the-art.

2.2 Privacy

Privacy is a fundamental human right. Generally, Westin [20] in 1967 defined privacy as *"the ability of an individual to control the terms under which personal information is acquired and used."* According to Internet Security Glossary [21], privacy is defined as *"The right of an entity (normally a person), acting in its own behalf, to determine the degree to which*

it will interact with its environment, including the degree to which the entity is willing to share information about itself with others." The EU General Data Protection Regulation (GDPR) [22] as well as the OECD privacy framework [23] consider the information which can be used to identify a natural person as "personal information". According to the law, personal information can not be recorded or used without the authorization of its owner. Meanwhile, the GDPR also classifies personal biometric information and personal location information as the category of privacy that should be protected. In the scope of this thesis, we will discuss the definition of privacy, the means of privacy protection and the methods we use in accordance with the requirements of the GDPR. Generally, when personal information is used in a non-permissible way, there is an invasion of privacy.

Privacy-enhancing technologies (PETs) provide privacy protection based on technology, and it can be considered the implementation of the policy from the view of the technology [24, 25] The privacy metrics [24] are used to measure the level of privacy in a system quantitatively. Anonymity is an important indicator in various privacy metrics. In an information system involving user data, information encryption [26] and necessary privilege management are also means to protect privacy to ensure that users can control their personal information.

"Privacy is also its own field of research in statistical databases where information about a population are to be disclosed while at the same time the privacy of individuals should be protected [27]." The privacy protection in such a scenario belongs to the *differential privacy* [28], and it is also referred to as disclosure control. In this case, it is required that by accessing such databases from statistical information, the personal information and identities of the participants in the statistics cannot be obtained [29]. In other words, if an adversary has acquired the data, he cannot determine whether an individual is included in this dataset. In order to achieve differential privacy security, previous studies have verified that this problem can be solved to a certain extent by introducing noise into the dataset [28]. Meanwhile, in some specific use cases, some researchers use desensitization methods to remove privacy-sensitive information while allowing the dataset to serve its purpose [30, 31].

Whether the scientific research or market-oriented products, privacy-protecting will be an issue that cannot be avoided when it comes to data collection. Considering the situational awareness system perception requirements, deployments of various sensors are necessary. The deployment of sensors in public areas must consider the protection of the privacy of people in the monitoring area. Although video data can provide rich information support for perception algorithms, It is straightforward for people to recognize a person's identity according to the face image in the data stream. Consequently, according to the laws, it is not globally possible to deploy cameras in public places. In scenarios with adequate reasons to deploy cameras, the deployment practice should be compliant and legal. Avoid

introducing cameras in the products can be a direct and effective way.

The privacy protection method involved in this paper is mainly the anonymization of personal information. This thesis's proposed privacy protection approaches are based on vibration data. The vibration sensor is a kind of low-profit sensor. By watching the ground vibration signal caused by footsteps, humans cannot identify the person who caused the signal data. Without the necessary support of identity-related data, even with the aid of computers and algorithms, the identity of the observed pedestrian cannot be identified. Thus the perception approaches based on vibration sensors provide sensor-level privacy protection and differential privacy security. The vibration signal data can be considered a sensor-level desensitization approach for privacy protection.

2.3 Machine Learning and Neural Network

2.3.1 Machine Learning

Machine Learning Definition

It is recognized in the academic that Arthur Lee Samuel first defined the term "machine learning" in 1959.

"Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed." [32].

Tom Mitchell gave out the definition for well-posed learning problem in 1997 [33] as *"Well Posed Learning Problem: A computer program is said to learn from experience E in context to some task T and some performance measure P, if its performance on T, as was measured by P, upgrades with experience E."*

Nowadays, the term "machine learning" is widely used to describe scenarios involving tasks such as classification, regression, clustering, perception, and pattern recognition.

Machine Learning Algorithms

Machine learning algorithms are generally been categorized into 3 classes according to the learning approach:

- Supervised learning [34]
- Unsupervised learning [35]
- Reinforcement learning [36]

Supervised learning is an approach to making a machine or a model to learn the patterns and relationships between the input data and the labeled output. Thus when new data are

fetched into the model, the trained model can output accurate or correct results. Supervised learning works well with classification problems and regression problems.

Supervised learning requires the training data to include labels. The so-called label means that the data input to the model during the training process has the corresponding correct output. Like a student solving a problem, the correct answer is known. The model's training is supervised by the correct answers (labeled data). Unsupervised learning is a kind of algorithm that learns patterns from unlabeled data.

The work of this thesis involves the use of deep learning, random forests, principal components analysis (PCA), Deep learning and random forests belong to the supervised learning algorithms. PCA is an unsupervised learning algorithm.

2.3.2 Neural Network

Artificial neural networks are inspired by biomimicry. It originated when scientists wanted to design an algorithm to mimic the working mechanism of the human brain. Although the academic understanding of the working mechanism of the human brain is still at a superficial level so far, some theories about the human brain are assumed to have been successfully applied in engineering practice.

The "one learning algorithm" hypothesis

As the basic unit of the brain, neurons play an essential role in the realization of brain functions. The working mechanism of the neuron has been deeply studied by academia. In brain science, there is evidence that the brain uses only one "learning algorithm" to achieve. Scientists hypothesize that the complex and rich functions implemented by the brain are based on "one same algorithm" to achieve various functions. Some of these studies support this hypothesis from a certain angle. Roe's research found that the auditory cortex is able to learn to see [37]. Metin's research found that the somatosensory cortex is able to learn to see [38]. This feature is called "neuroplasticity". There are many more experiments serving as evidence for it. Nagel et al.'s research [39] use an electronic compass and belt with vibrators to train people to learn the sense of directions. Danilov et al.'s research [40] shows that the visual information can be transmitted via tactile nerves to the cerebral cortex, which is responsible for vision. Evidence shows that multiple neurons linked by specific structures and transmitting signals by certain mechanisms can form units (neural networks) that achieve specific functions.

Neural Model Logistic Unit

The currently widely used neuron model is the point-like model that treats each neuron as a node and determines the output by calculating whether the sum of the inputs exceeds

a threshold, and it is named as perceptron [41]. Although the actual working mechanism of human brain neurons is still in the stage of understanding in academia [42], the neural network model based on perceptron has been widely used.

Later, researchers evolved logistic regression [43] by improving the perceptron's activation function and optimization method. Accordingly, logistic regression has a probabilistic explanation for the final classification result compared with the perceptron.

The composition of nerve cells includes dendrites and axons. The dendrites that receive signals from other neurons can be considered "input wires". The axons output the signals to other neurons serving as "output wires". According to the neuron model, when the signal strength received by a neuron is more significant than a certain threshold, the neuron will be activated and transmit the signal out to the connected neuron through the axon as shown in Fig. 2.2 and equation 2.7. A logistic regression unit can be thought of as a neuron in an artificial neural network.

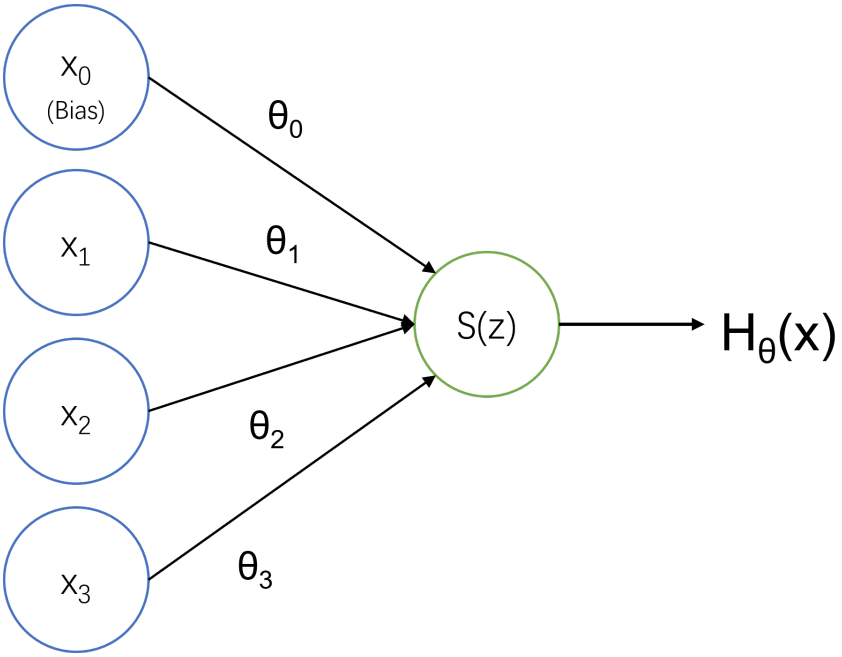


Figure 2.2: Neural model logistic unit.

As shown in Fig. 2.2, the neuron model, which has 3 connected dendrites (3 input wires and 1 bias), is represented. The neuron has 3 input: x_1, x_2, x_3. The x_0 is the bias unit, which is a constant normally $x_0 = 1$. θ is the weights of each inputs. θ_0 is the weight of the bias, normally $\theta_0 = 1$. H_θ is the output. Mathematically, it can be represented with equations from 2.1 to 2.6. $S(z)$ is the sigmoid function. $H_\theta(x)$ is the output.

$$x = \begin{bmatrix} x_0 & x_1 & x_2 & x_3 \end{bmatrix}^{\mathrm{T}} \qquad (x_0 = 1) \tag{2.1}$$

$$\theta = \begin{bmatrix} \theta_0 & \theta_1 & \theta_2 & \theta_3 \end{bmatrix}^{\mathrm{T}} \tag{2.2}$$

$$S(z) = \frac{1}{1 + e^{-z}} \tag{2.3}$$

$$z = \theta^{\mathrm{T}} x \tag{2.4}$$

$$= \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \tag{2.5}$$

$$H_\theta(x) = \frac{1}{1 + e^{-\theta^{\mathrm{T}} x}} \tag{2.6}$$

Generally, a neuron with n inputs will be represented as equations 2.7. In the equations, b stand for the bias.

$$x = \begin{bmatrix} b \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \qquad H_\theta(x) = \frac{1}{1 + e^{-\theta^{\mathrm{T}} x}} \tag{2.7}$$

The Bias in the Neural Networks

In the logistic regression or the mathematical model of neurons, there is always a bias parameter, which cannot be omitted. The mathematics of the logistic regression model is essentially to use the function $y = wx + b$ to draw the decision boundary. The w is the model's parameters, and the b is the intercept. Take the two-dimensional data classification task as an example, as shown in Fig. 2.3. If there is no bias term in two-dimensional space, we can only draw a straight line through the origin. For data that is not 0-centered distributed, a straight line through the origin cannot be used for classification tasks. In this case, the model will fail to converge during training. However, typically the data used for training is always not 0-centered. Similarly, in a high-dimensional space, the decision boundary will be a hyperplane.

In a neural network, the existence of the bias term enables the neurons in this layer to fit the input data of which distribution is not 0-centered. In contrast, if the input of a layer of the neural network is normalized, it can be fitted even without bias.

However, in the practical design of the neural network, we always keep this bias term. A neural network with the bias term can fit the data faster. The bias term makes the neural network with the bias term converge faster than the neural network without the bias term.

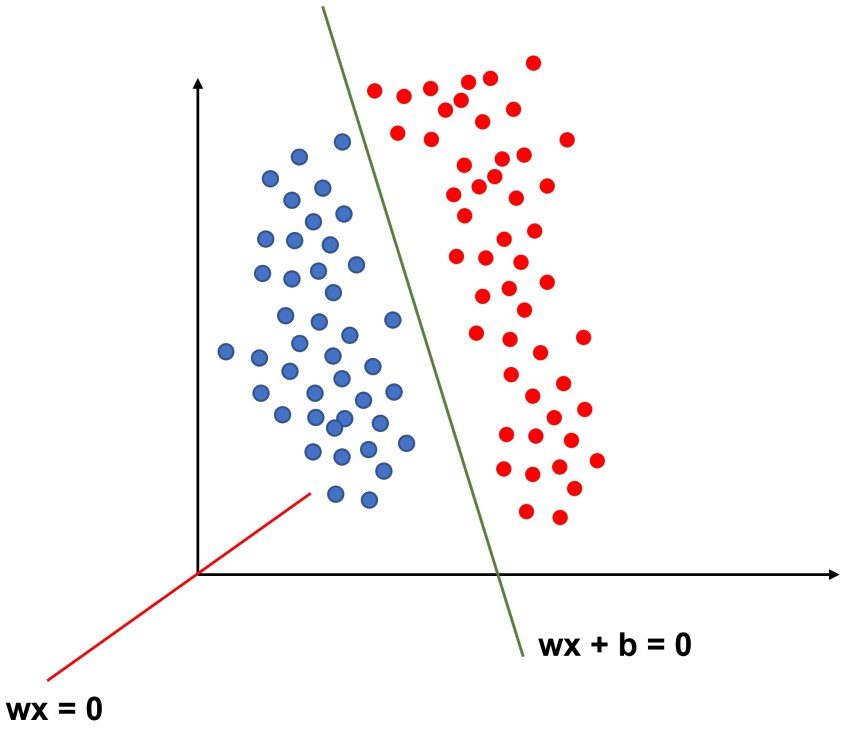


Figure 2.3: 2D non-zero-centered distribution data.

Meanwhile, the bias term makes the neuron more flexible and improves the neuron's fitting ability. Furthermore, the bias can be understood as the control over the activation state of neurons. Referring to the activation function Equation 2.3, when the bias is relatively small, it shows the inhibition of a neuron. Conversely, the neuron will be easier to activate.

Model Representation of Neural Networks

The neural network comprises multiple neurons organized according to a particular structure. Information is passed between neurons according to specific rules. Multilayer feedforward neural network [44] is a typical neural network architecture. The feedforward neural network groups neurons into groups according to the order in which they receive information. Each group is considered a neural layer. The neurons in each layer receive the previous layer's output and output to the neurons in the next layer. Information in the entire network is propagated in one direction, and there is no reverse information propagation.

Taking the three-layer feedforward neural network as an example, Fig. 2.4 is a 3 layers feedforward neural network. Layer 1 is the input layer, layer 2 is the hidden layer, and layer 3 is the output layer. Layer 1 has 3 units, and layer 2 has 3 hidden units. x_0 and a_0 are the "bias units" which are constants.

As shown in Fig. 2.4 and equation 2.8, $a_i^{(j)}$ is the "action" of unit i in layer j. $\theta^{(j)}$ is

the matrix of weights from layer j to layer $j+1$. When the neural network has s_j units in layer j, s_{j+1} units in layer $j+1$, then $\theta^{(j)}$ is of dimension $s_{j+1} \times (s_j + 1)$. The information is computed and forward propagation as shown in equation 2.8. Normally $x_0 = 1, a_0^{(2)} = 1$.

$$\begin{aligned} a_1^2 &= S(\theta_{10}^{(1)} x_0 + \theta_{11}^{(1)} x_1 + \theta_{12}^{(1)} x_2 + \theta_{13}^{(1)} x_3) \\ a_2^2 &= S(\theta_{20}^{(1)} x_0 + \theta_{21}^{(1)} x_1 + \theta_{22}^{(1)} x_2 + \theta_{23}^{(1)} x_3) \\ a_3^2 &= S(\theta_{30}^{(1)} x_0 + \theta_{31}^{(1)} x_1 + \theta_{32}^{(1)} x_2 + \theta_{33}^{(1)} x_3) \\ S(z) &= \frac{1}{1+e^{-z}} \\ H_\theta(x) &= a_1^{(3)} = S(\theta_{10}^{(2)} a_0^{(2)} + \theta_{11}^{(2)} a_1^{(2)} + \theta_{12}^{(2)} a_2^{(2)} + \theta_{13}^{(2)} a_3^{(2)}) \end{aligned} \tag{2.8}$$

In the equations 2.8, during the initialization of the forward propagation, the biases x_0 and $a_0^{(2)}$ are set to 1. The value of the elements of the weight matrix $\theta^{(j)}$ is initialized to random numbers. Typically, the weight matrix is marked as w, and the bias matrix is marked as b. The parameters w and b will be updated in the training process.

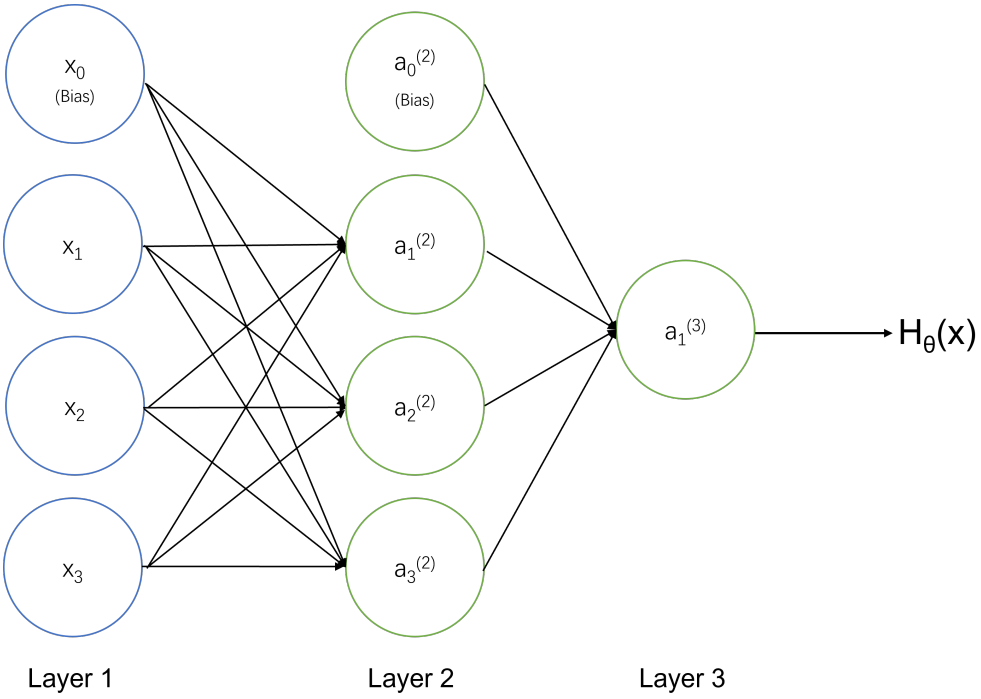


Figure 2.4: 3 layers feedforward neural network.

Convolutional Neural Networks

Convolutional neural networks are also mentioned as convolutional networks[45] or CNNs. *"Convolutional networks are simply neural networks that use convolution in place of general*

matrix multiplication in at least one of their layers" [44].

$$f(x) * g(x) = \int_{-\infty}^{+\infty} f(\tau)g(x-\tau)d\tau \tag{2.9}$$

As shown in equation 2.9, $f(x)$ and $g(x)$ are two integrable functions on $\mathbb{R}$. The convolution of $f(x)$ and $g(x)$ is witten $f(x) * g(x)$.

For discrete convolution operations, we can rewrite the discrete convolutional operation as equation 2.10:

$$f(n) * g(n) = \sum_{m=-\infty}^{\infty} f(m)g(n-m) \tag{2.10}$$

In equation 2.10, $f(n)$ and $g(n)$ are complex-valued functions on integers set $\mathbb{Z}$.

Signal theory shows that if a signal is finite in the time domain, it is infinite in the frequency domain. A signal with an infinite time domain must be band-limited in the frequency domain, and a signal with an infinite frequency domain is finite in the time domain. Suppose a signal is a non-periodic signal in the time domain and a time-limited signal. An approximation has been made in engineering. In that case, it is assumed by default in engineering signal processing that the signal is a periodic signal with this signal segment as a complete cycle. In the scenario of signal processing by the computer, the signal saved by the computer is usually a digital signal obtained by sampling and quantizing an analog signal according to the sampling theorem. In engineering, the processing of a signal, whether in the time or frequency domain, is regarded as finite signal processing.

The convolutions can be used over more than one axis. When the input data I is a two-dimensional matrix, the kernel K is also a two-dimensional matrix, and the convolutional operation can be denoted as equation 2.11

$$\begin{aligned} S(i,j) = (I * K)(i,j) &= \sum_m \sum_n I(m,n)K(i-m, j-n) \\ &= \sum_m \sum_n I(i-m, j-n)K(m,n) \end{aligned} \tag{2.11}$$

Many machine learning libraries use cross-correlation, as shown in equation 2.12, instead of convolution but still call it convolution. The difference between cross-correlation $(f \star g)$ and convolution operation $(f * g)$ is that either $f(x)$ or $g(x)$ is reflected about the y-axis.

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n) \tag{2.12}$$

The convolutional layer includes an input matrix and a kernel matrix. The output of the convolutional layer is the convolutional operation between the input matrix and the kernel matrix. The weight parameters in the kernel matrixes are learned during the training process. Meanwhile, usually, a pooling layer is cascaded after the convolution layer. Then

no matter cross-correlation operation or the convolutional operation will not influence the final results [44].

Classic Deep Convolutional Neural Network Practices

Usually, convolutional neural networks are not used alone. Fig. 2.5 is the deep convolutional neural network architecture used in the positionning task in this thesis.

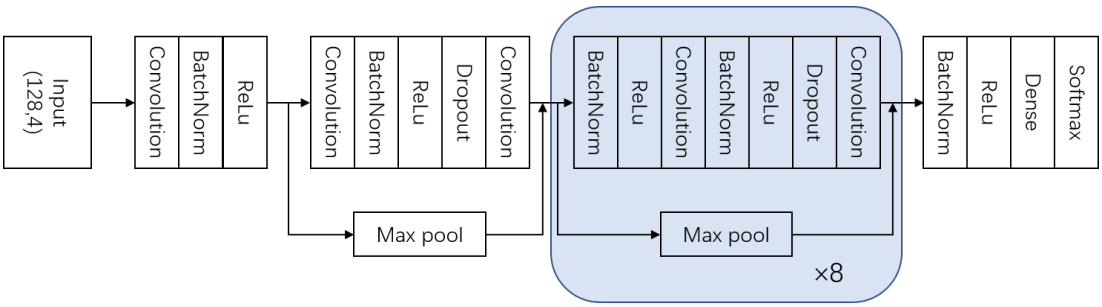


Figure 2.5: Deep CNN architecture example.

A pooling layer [44] is typically cascaded after the convolution layer. The pooling layer can decrease the size of the feature map [44] thus reducing computational volume. Meanwhile, it can increase the receptive field [46]. Also, as pooling is a nonlinear operation, it can enhance the expressiveness of the network. In practice, a batch normalization layer[44, 47] is cascaded after each convolution layer to maintain the mean output to 0 and output standard deviation to 1. This can solve the gradient dispersion problem and gradient explosion problem. Because convolution is a linear operation, a "ReLu" layer is cascaded after each batch normalization layer. The "Relu" layer has an activation function, enabling the neural network to have nonlinear learning capabilities. Furthermore, "Dropout" technology [48] will be inevitably used in the deep neural network architecture. The "Dropout" technology makes a neuron's activation function stop working with a certain probability in the forward propagation. Dropout can suppress overfitting because it makes the entire neural network less dependent on some local features. Meanwhile, it can reduce the computational costs of neural network inference and training. Not only that, the deep neural network in Fig. 2.5 is of residual architecture (Residual Network, ResNet). The ResNet use shortcut connections to make the information can jump some layers of the neural network in the forward propagation. The residual architecture solves the degradation problem [49] and the shattering gradient problem [50]. Finally, the softmax layer typically used as the output layer can map multiple scalars into a probability distribution, each of which outputs a value in the range $(0, 1)$. In multi-classification tasks, the neural network with softmax as the output layer can output the prediction of each category in the form of probability.

Deep neural network architectures based on convolutional neural networks emerge in an

endless stream, constantly breaking the performance limit. DNN-based technologies have proliferated and continue to push the performance limits[2, 49, 51, 52, 53, 54]. Generally, there is a trend for neural networks to become deeper and deeper. The deeper network exhibits better feature extraction performance.

2.4 Feature Selection and Extraction

In machine learning tasks, machine learning algorithms learn patterns from input data. However, the unprocessed data is redundant. Usually, the dimensionality of the raw data is very high. For example, a human can recognize the identity of a person only base on the face image rather than the image of the whole body. That is to say, for a face recognition task, picture data other than faces is redundant. In order to improve the computational efficiency of the machine learning algorithm, shorten the training time, improve the versatility, and suppress the overfitting problem, the data is usually processed for feature selection and feature extraction before the machine learning algorithm is trained. Feature selection returns a subset of the original features. Feature extraction extracts the parts useful for the task by processing the original data. Feature extraction is to create new features from the original features.

2.4.1 Principal Components Analysis

The principal component analysis is a classic feature extraction method, and it can reduce the dimensionality of the data. It uses an orthogonal transformation to linearly transform the observations of a series of potentially related variables, thereby projecting them into a series of linearly uncorrelated variables [55]. These uncorrelated variables are called principal components (Principal Components).

As shown in Fig. 2.6, the blue points are data points in the 2-dimensional space. The red points are data points in the 1-dimensional space. The projecting blue points to red points is dimensionality reduction. The purpose of PCA is to find a coordinate system with a lower dimension than the original data dimensionality so that the loss of information is minimized after the data is reduced in dimension. Reduce from n-dimension to k-dimension: Find k vectors $u^{(1)}, u^{(2)}, \cdots, u^{(k)}$ onto which to project the data, so as to minimize the projection error [55].

The PCA algorithm to reduce the data of n-dimension to k-dimension is as follows. Assume the data has m samples. Firstly, we calculate the mean normalization of the j dimensional raw data as shown in equation 2.13. $x_j'^i$ is the element of the raw data. x_j^i is the mean normalization data. μ_j is the mean, and s_j is the standard deviation.

Then calculate the covariance matrix of the mean normalization data as shown in equation 2.14. The convariance matrix is presented with Σ.

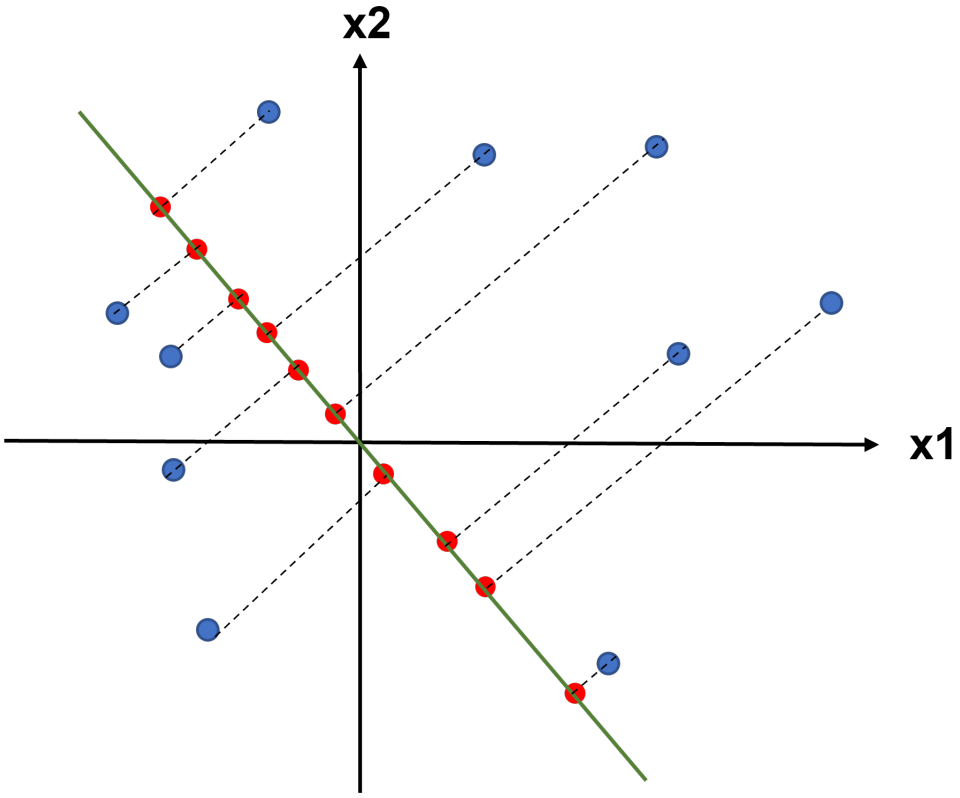


Figure 2.6: 2 dimensinality data reduced to 1 dimensinality example.

$$x_j^i = \frac{x_j'^i - \mu_j}{s_j} \tag{2.13}$$

$$\Sigma = \frac{1}{m}\sum_{n-1}^{n}(x^{(i)})(x^{(i)})^{\mathrm{T}} \tag{2.14}$$

$$[U, S, V] = SVD(\Sigma) \tag{2.15}$$

$$U = [u^1, u^2, u^3, \cdots, u^n] \qquad U \in \mathbb{R}^{n\times n} \tag{2.16}$$

$$\begin{aligned} z^{(i)} &= U'^{\mathrm{T}} x^{(i)} \\ &= [u^1, u^2, u^3, \cdots, u^k]^{\mathrm{T}} x^{(i)} \end{aligned} \tag{2.17}$$
$$z \in \mathbb{R}^k, \qquad U' \in \mathbb{R}^{n\times k}$$

Next, calculate the eigenvectors of the matrix Σ. As shown in equation 2.15, we use singular value decomposition to calculate the eigenvectors of the covariance matrix. The matrix U is the eigenvectors of matrix Σ as shown in equation 2.16. Finally, we can calculate the k-dimension data $z^{(i)}$ as shown in equation 2.17. The U' is a matrix with the first k columns of matrix U.

PCA can be thought of as a data compression algorithm. In general, PCA can speed up machine learning algorithms and reduce the memory to store data. PCA is also a means of visualizing high-dimensional data. When the high-dimensional data is reduced to three dimensions and below by PCA, we can visualize the data.

2.4.2 Deep Neural Network Feature Extraction

In the field of image recognition, it is well known that deep convolutional networks can automatically extract features[56, 57]. Not only that, the features learned through deep learning performed exceptionally well in the classification task. The output data of each convolution layer is called feature map [44]. When the neural network goes deeper, the receptive field increases. Meanwhile, the feature extracted becomes more abstract.

The deep one-dimensional convolutional network can also automatically extract features in the processing of vibration signals. In this thesis, deep convolutional neural networks are also used as feature extraction methods.

CHAPTER 3

System Architecture

The architecture of the proposed system is presented as shown in Fig. 3.1. The SA system includes a representation layer and a perception layer. Each layer has several modules. The crisis-privacy status model controls the modules in both layers.

The motivation, the functions, and the constraints to be satisfied for these layers will be explained. The functions of the modules included in each layer will be explained one by one. How each module is coordinated with each other will be discussed. However, The specific implementation details of each module will be discussed in Part 2 of the thesis.

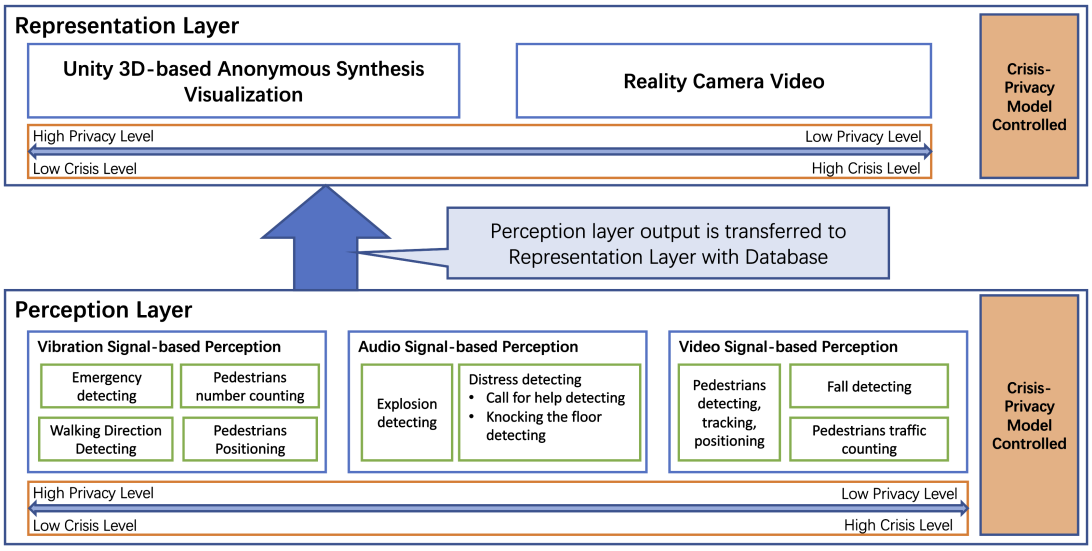


Figure 3.1: System architecture.

3.1 Crisis-Privacy Status Model

Most data of modality is not as rich and intuitive as image data collected by the camera. At the same time, considering the outstanding progress in machine vision in recent years in academia and industry, rather than completely discarding the use of cameras, it is better to restrict them and formulate policies that meet both privacy protection and perception requirements. To this end, the crisis-privacy status model is proposed as an engineering implementation of this strategy.

Compared with the safety of people's lives, the importance of privacy comes second. The principle of this model is as follows. When there are no incidents that endanger people's lives (no crisis), prioritize protecting people's privacy. On the contrary, once there is a crisis in the environment, privacy protection is ignored, and the safety of people's lives is given priority.

We define two factors in the crisis-privacy status model. The first factor is the crisis level, and the second is the privacy level. The crisis level is defined subject to the crisis severity of the situation in the detected area. The crisis level determines the privacy level.

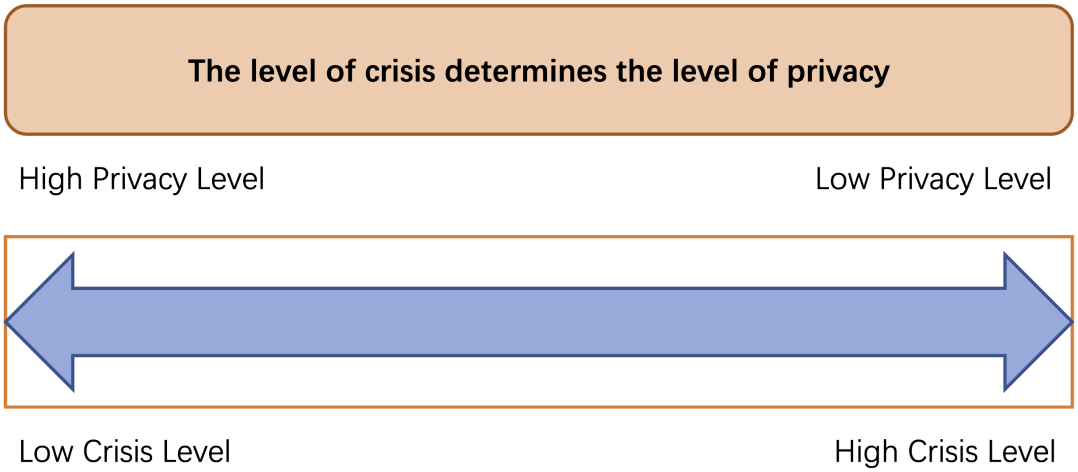


Figure 3.2: Crisis-privacy status model.

When an event in the environment may infringe on people's life safety, the crisis level is defined as a high level. Otherwise, it is a low crisis level.

The SA system presented in the thesis does not deprecate the cameras. The system only prohibits the use of cameras at the low crisis level (high privacy protection level). Only data collection methods that do not violate privacy protection can be used at this level. When the crisis level is triggered to high, the privacy level will be adjusted to low.

A system constrained by the crisis-privacy status model should meet the following conditions.

1. The system needs to be able to adjust the level of privacy based on the level of crisis in the environment.

2. The kinds of sensors that the system activates the usage status of being constrained by the privacy level.

3. The system should be able to switch between high crisis status and low crisis status autonomously.

The system should rely on sensors permitted only at high privacy levels to detect and understand environmental information and identify crisis events. When the crisis is resolved or confirmed that the crisis is a false alarm, it will automatically adjust back to the low crisis-high privacy level.

3.2 Two Layer SA System

Considering perception and expression are two relatively independent functions, and considering the storage requirements for perception results, we layer the system into two functional layers as shown in Fig. 3.1.

The perception layer depends on the perception modules of this layer to perceive the environment and detects the target events. The perception layer will record the results in a database. The representation layer expressions the perception results by querying the result of the database.

The representation layer should be able to express the perception results in a human-friendly manner.

3.2.1 Perception Layer

In order to accurately perceive events in the environment from multiple perspectives, the perception functions should be implemented with sensors of different modalities. However, according to privacy protection restrictions, there should be at least one sensor that does not violate privacy. Meanwhile, algorithms should identify whether there is a crisis event in the environment based on the data of this sensor alone.

Furthermore, the system solution is kept the low cost as much as possible, thus providing advantages for engineering. Considering all of the requirments, the vibration sensor, audio sensor, and cameras are selected to construct the perception layer.

The perception layer is divided into three parts based on the different types of sensors used.

Four modules belong to the vibration signal-based perception part. These modules are the emergency detecting module, pedestrian number counting module, walking direction

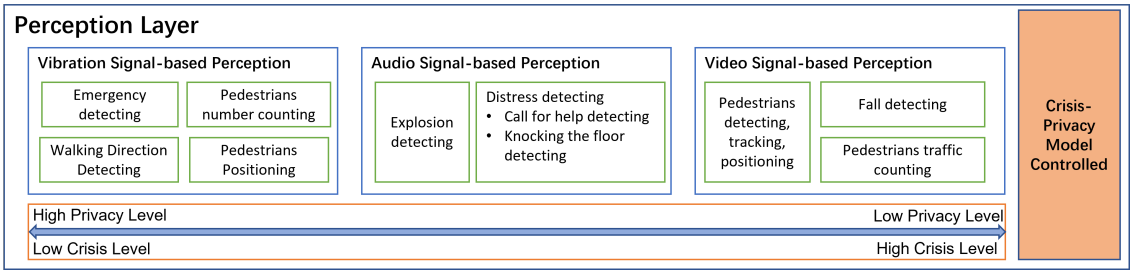


Figure 3.3: Perception layer diagram.

detecting module, and pedestrians positioning module. These modules only depend on the vibration signal, and they are privacy-protecting approaches. These modules can work under both high privacy mode and low privacy mode.

The emergency detection module can detect if there is an emergency happening in the environment based on the vibration data. When the crisis status is under the low crisis level, if the emergency detection module detects an emergency event, it will set the crisis status to high (low privacy mode).

The pedestrian number counting module can detect the number of pedestrians in a 3 meters by 3 meters zone. Usually, it is deployed near the entrance of an area.

The walking direction detecting module can detect if a pedestrian is walking left or right. It is also commonly deployed near the entrance of an area and is typically used together with the pedestrian number counting module. With this module, we can know a pedestrian comes in through a door or go out through the door.

The pedestrian positioning module can detect the location of a pedestrian.

There are 2 modules in the audio signal-based perception part. These modules are the explosion detecting module and distress detecting module. The explosion detection module can work under both high and low privacy modes. The distress detecting module only works on high crisis mode (low privacy mode).

The explosion detection module is able to detect if there is an explosion. When the explosion is detected, it will set the crisis status to high no matter the current crisis status.

The distress detecting module can detect if someone is calling for help or knocking on the floors. This module depends on the audio signal and only works under high crisis mode.

Three modules belong to the video signal-based perception part. These modules are the pedestrians detecting, tracking, positioning module, the fall detection module, and the pedestrians' traffic counting module. The modules of this part depend on camera data and only work under high crisis mode (low privacy mode).

The pedestrians detecting, tracking, and positioning module can detect persons according to the input video data. It can identify whether the person's identity appearing in the image is the same. It can output the coordinates of the world coordinate system of the

person appearing in the image screen and track the walking path of the pedestrians.

The fall detection module is able to detect if there is a person in the video falling. The pedestrians' traffic counting module is able to detect the pedestrian s throughput by counting the number of pedestrians in a specified area.

3.2.2 Representation Layer

The presentation layer presents the perception results in a human-friendly way. The representation layer involves two privacy model-constrained expression approaches with different privacy properties. One approach is based on the Unity 3d-based anonymous synthesis visualization module, and the other approach is reality camera video.

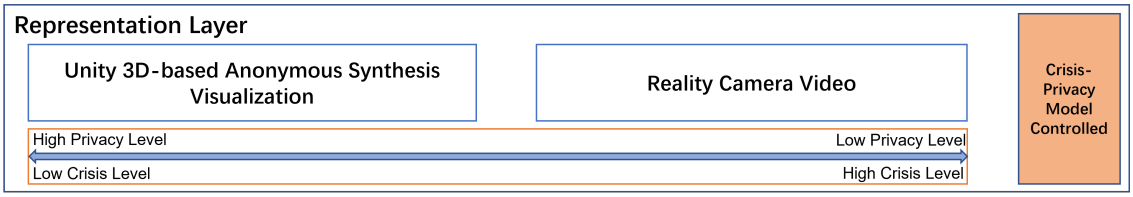


Figure 3.4: Representation layer diagram.

The Unity 3d-based anonymous synthesis visualization module will anonymously present pedestrians' position and status information in the constructed virtual 3D scenes based on the information of the perception module. The images of pedestrians are represented by virtual 3D human figures anonymously. The virtual 3D scene is constructed according to the monitored area, which corresponds to the area of the real environment. This representation approach can work under both high crisis and low crisis modes.

The reality camera video expression approach direct shows the video stream to the relevant person in charge, such as the commander of firefighters, police, or rescue teams. This expression approach only works under the high crisis level.

3.2.3 Model Constraints and State Transitions

The crisis-privacy status model constrains the active status of all the modules or approaches in the system. The principle is that privacy protection is not compromised unless it is a last resort. As shown in Fig. 3.5 and Fig. 3.6, when a crisis is detected, all technical means will be activated and used to maximize the protection of people's lives. In this case, privacy protection is compromised.

As shown in Fig. 3.7 and Fig. 3.8, when the crisis is low, The system will prioritize privacy protection, and some features will be disabled. Even in this case, the system still provides a wealth of perception intelligence that can serve some potential intelligent applications.

Fig. 3.9 shows how the crisis status is shifted. There are two statuses: low crisis status (high privacy level) and high crisis status (low privacy level). The explosion detection

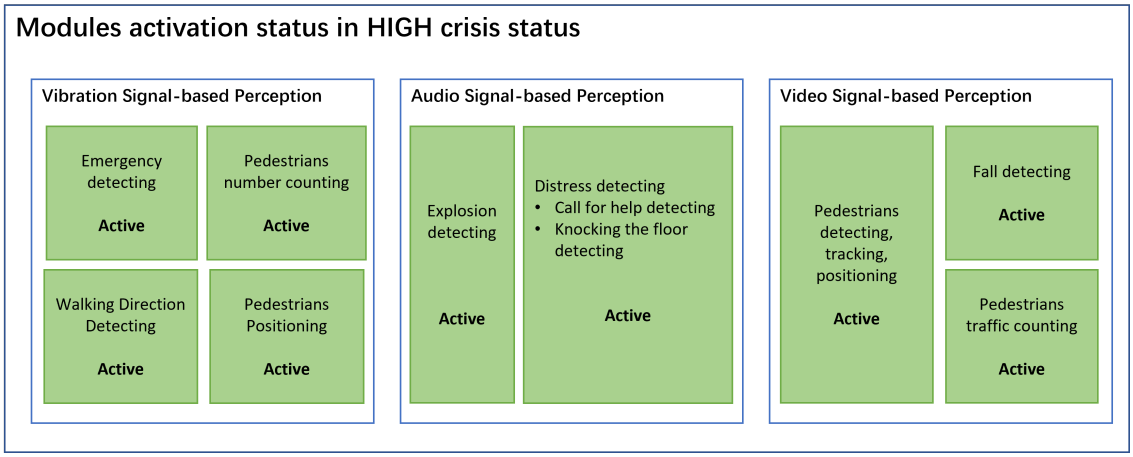


Figure 3.5: High crisis scene modules activation status of the perception layer.

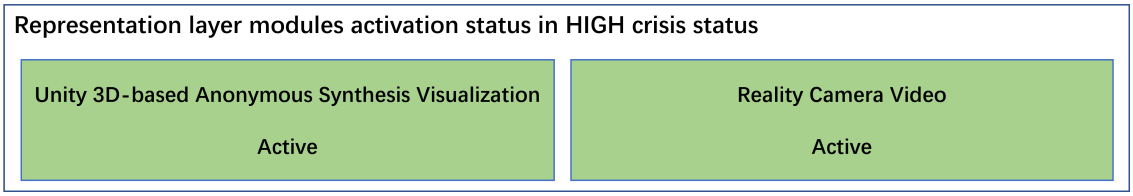


Figure 3.6: High crisis scene modules activation status of the representation layer.

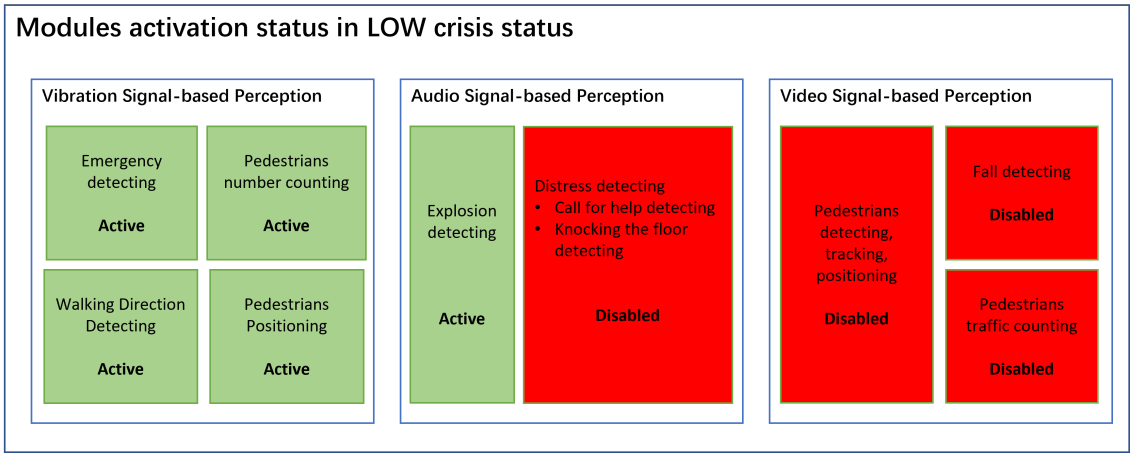


Figure 3.7: Low crisis scene modules activation status of the perception layer.

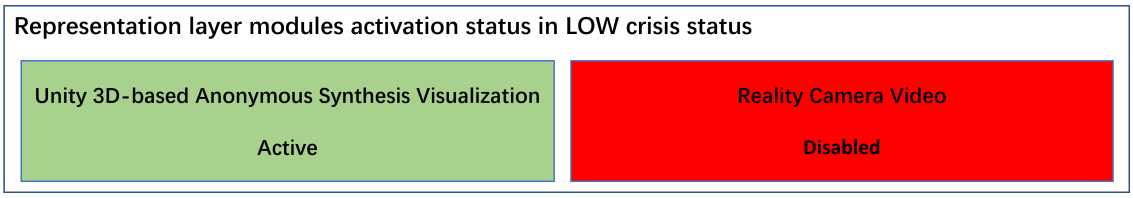


Figure 3.8: Low crisis scene modules activation status of the representation layer.

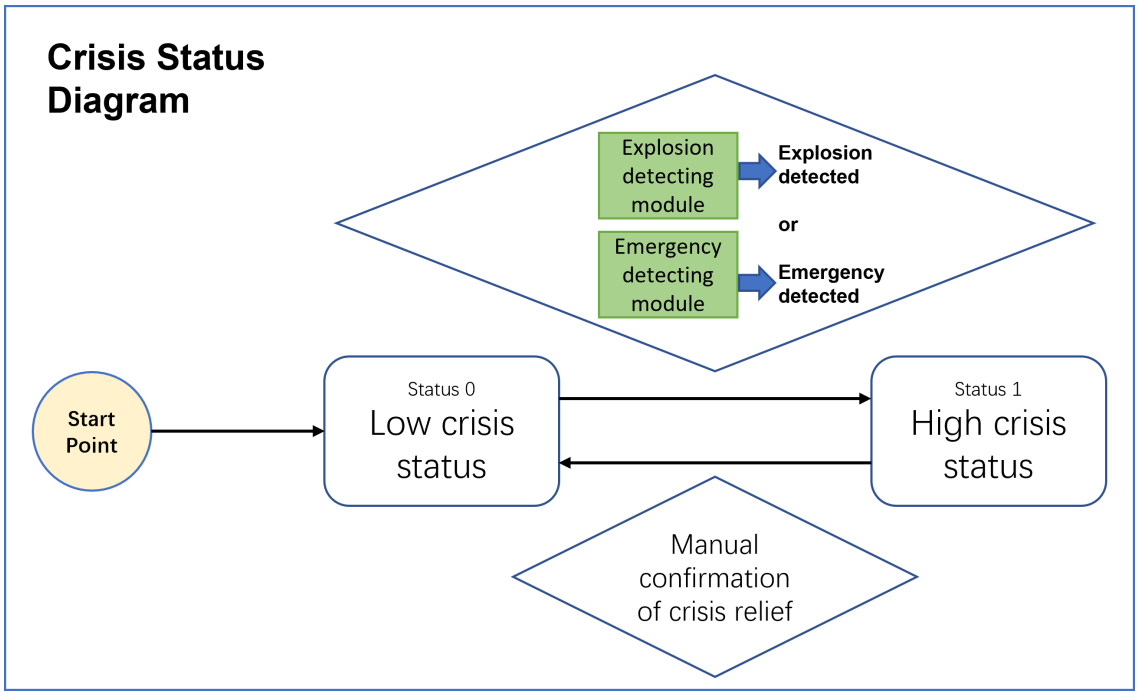


Figure 3.9: Crisis status diagram.

module and the emergency detection module can trigger the shifting of the crisis status. When an explosion is detected, or an emergency is detected, the system will shift the status to high crisis status. When the relevant persons in charge confirmed that the crisis is resolved or it was a false alarm, they can manually set the crisis status to low.

CHAPTER 4

Emergency Monitoring

In this work we present an indoor emergency context monitoring system based on ground vibration caused by persons in the target area. The system is designed for production plants and large buildings to perceive the safety status of this area. Our approach is privacy-protecting, because it requires neither video nor sound. Instead, piezo sensors on the floor measure vibrations, which are analyzed with machine learning to compute the safety status of the covered area. This way our system can determine whether an emergency occurred, but it is not straight forward possible to attach names to the detected persons. We compare the impact of different feature extraction methods and different types of classifiers on the classification results. Our experiments show that we can determine an emergency event with an average F1 score of 0.97 [12].

This part of the work constitutes the emergency detection module of the system described in this dissertation. This module only depends on the vibration signal. Thus it can run on the low crisis status. The crisis-privacy status is updated according to the result of this module.

The research work described in this Chapter is addressed in the following refereed paper that constitutes part of this cumulative dissertation:

- Yang Yu and Torben Weis, "A privacy-protecting indoor emergency monitoring system based on floor vibration," in Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers, New York, NY, USA, Sep. 2020, pp. 164–167. doi: 10.1145/3410530.3414423.

The content is included in Section 12.1.

CHAPTER 5

Pedestrians Number Counting

Pedestrian counting has attracted much interest of academic and industry community for its widespread application in many real-world scenarios. While many recent studies have focused on computer vision-based solutions for the problem, the deployment of cameras brings up concerns about privacy invasions. This chapter proposes a novel indoor pedestrian counting approach, based on footstep-induced structural vibration signals with piezoelectric sensors. The approach is privacy-protecting because no audio or video data is acquired. Our approach analyzes the space-differential features from the vibration signals caused by pedestrian footsteps and outputs the number of pedestrians. The proposed approach supports multiple pedestrians walking together with signal mixture. Moreover, it makes no requirement about the number of groups of walking people in the detection area. The experimental results show that the averaged F1-score of our approach is over 0.98, which is better than the vibration signal-based state-of-the-art methods [13].

This part of the work constitutes the pedestrians counting module of the system in the perception layer described in this dissertation. This module only depends on the vibration signal. Thus it can run on the low crisis status.

The research work described in this Chapter is addressed in the following refereed paper that constitutes part of this cumulative dissertation:

- Yang Yu, Xiangju Qin, Shabir Hussain, Weiyan Hou, and Torben Weis, "Pedestrian Counting Based on Piezoelectric Vibration Sensor," Applied Sciences, vol. 12, no. 4, Art. no. 4, Jan. 2022, doi: 10.3390/app12041920.

The content is included in Section 12.2.

CHAPTER 6

Pedestrians Walking Direction Detecting

We present an algorithm and measurement system to detect the walking direction of persons based on ground vibrations. The approach is privacy-preserving, because it solely relies on piezoelectric sensors built into the floor. Therefore, our system can be used in areas where cameras are not allowed or cannot capture the entire area. To analyze the ground vibrations caused by footsteps, we present a multi-peaks average algorithm (MPAA), which already offers a robust detection. MPAA judges the walking direction of pedestrians by analyzing the time-space relationship of at least two consecutive footstep vibration signals from multiple sensors. In addition, we show that the multi-peak averaging feature used by the MPAA can be used as an input feature to a deep neural network classifier. This multi-peaks averaged feature with deep neural network-based classifier (MPAF-DNNC) approach further improves the detection quality and especially reduces the number of footsteps necessary to determine the walking direction.

This part of the work constitutes the pedestrians walking direction detecting module of the system in the perception layer described in this dissertation. This module only depends on the vibration signal. Thus it can run on the low crisis status.

The research work described in this Chapter is addressed in the following refereed paper that constitutes part of this cumulative dissertation:

- Yang Yu, Oskar Carl, Shabir Hussain, Weiyan Hou, and Torben Weis, "A Privacy-Protecting Step-Level Walking Direction Detection Algorithm based on Floor Vibration," IEEE Sensors Journal, 2022, Accepted.

The content is included in the Section 12.3.

CHAPTER 7

Pedestrians Positioning

In this work, we present a person localization system based on ground vibration caused by walking persons. The system is designed for production plants and large buildings to track the movement of workers. Position and movement in these settings are especially safety-relevant in emergencies. Our approach is privacy-preserving, because it requires neither video nor sound. Instead, piezo sensors on the floor measure vibrations, which are analyzed with machine learning to derive a person's position from the vibration signals. This way, our system can determine where a person is moving, but it is not straightforward to attach names to the detected persons. Due to the anisotropic characteristic of the ground vibration wave, classical analysis methods are not applicable. We show that a deep learning-based approach is feasible. Our experiments show that we can determine the position with an average F1 score of 0.95 [11].

This part of the work constitutes the pedestrians positioning module of the system in the perception layer described in this dissertation. This module only depends on the vibration signal. Thus it can run on the low crisis status.

The research work described in this Chapter is addressed in the following refereed paper that constitutes part of this cumulative dissertation:

- Yang Yu, Marian Waltereit, Viktor Matkovic, Weiyan Hou, and Torben Weis, "Deep Learning-Based Vibration Signal Personnel Positioning System," IEEE Access, vol. 8, pp. 226108–226118, 2020, doi: 10.1109/ACCESS.2020.3044497.

The content is included in the Section 12.4.

CHAPTER 8

Audio signal-based Perception system

In this chapter, two microphone-based modules are introduced. One module is the exploration detecting module, and the other is the distress detecting module.

The work in this chapter relies on existing research, but its implementation is an important component of the proposed system. I do not claim the scientific contribution to the work in this chapter.

8.1 Introduction

Hearing is an important way of perception for humans. Similarly, we use the acoustic sensor for our situational awareness system to perceive audio events. The bomb blasts and public shooting incidents are vicious incidents endangering people's safety in public places. Meanwhile, automated explosion detection technology can be helpful to improve the response speed of police or rescue personnel to control the hazard [58]. Thus explosion detection module is implemented in our SA system. When an explosion occurs, there is an emergency happening in the place, and the proposed situation awareness system considers the explosion event as a sign to trigger the high-crisis status.

Concerning the various sound events that may occur in public places, people's cry for help is an important event that requires rescuers' attention. When people are in danger, they call out "help" for aid. However, in a fire or explosion, typically, there will be thick smoke. In such an environment, people in distress will also tap the ground with their hands to ask for help. Therefore, the proposed SA system also includes the call for help detecting and knocking on the floor detecting module.

8.2 Related Works

8.2.1 Explosion Detecting

Explosion detecting belongs to the Audio Event Detection (AED) tasks. As explosion sounds are relatively rare in everyday life scenarios, the detection of explosion sound task, more precisely, belongs to rare sound event detection [59].

By studying the existing Audio Event Detection research of recent years, it is founded that the most approaches with competitive performance are data-driving approaches and share the similar processing steps[59, 60, 61, 62, 63, 64, 65, 66, 67, 68]. These approaches can be summarized in the following steps: dataset preparation, feature extraction, and classifier training. Different feature extraction methods also have various performances in specific sound event detection tasks. In terms of the choice of classifier, these papers show that the deep learning-based method is superior to the classical Support Vector Machine (SVM), nearest neighbor, and decision tree classifiers.

In particular, regarding the explosion detecting task, the performance of Shukla's research [58] is outstanding. This research has considered the characteristics of explosion sounds and uses three methods to extract features from raw audio sample data. The first feature is spectral contrast [69, 70]. The second feature is Log-mel spectrograms [61]. Moreover, the third feature is the 1D convolution feature extracted by a 1D convolutional neural network from the raw data. The spectral contrast features and Log-mel spectrograms are fetched into a ResNet [49] like residual neural network blocks individually. These two residual blocks and the 1D-convolutional feature-extracting network run parallel. Then a fully connected neural network receives the output of the three parallel subnetworks to fusion the data and make the classification. In other words, this research uses a three-input one-output neural network to detect explosion audio.

The explosion detecting module of the SA system is implemented according to Shukla's paper [58].

8.2.2 Distress Detecting

The audio-based distress detecting module should implement two functions. Firstly, the module should detect if someone is yelling "Help". Secondly, the module should detect if someone is knocking on the floor. Considering the scenes the SA system is running, the events that these two functions should detect are rare acoustic events.

Recently deep learning-based approach have presented promising performance in speech recognition tasks [71, 72, 73]. Although both the speech recognition task and the distress detecting task are pattern recognition tasks based on audio signals, there are apparent differences between these two kinds of tasks.

In the speech recognition task, there is a presupposition. The speech recognition approach supposes that the scene of the input audio is one person speaking in a quiet environment. If this assumption is not satisfied, such as multiple people speaking alternately or simultaneously in a quiet place or one person speaking in a noisy place, the speech recognition system will not be able to show satisfactory recognition accuracy. Therefore, in the proposed SA system, using the speech recognition approach to detect keywords like "help" thus to detect distress is not feasible.

After studying the working scenario of the situational awareness system, the rare sound event detection methods [74, 75] are investigated. The above researches use the dataset published in DCASE 2017 Challenge [59]. The dataset includes samples of 3 categories: a baby crying, glass breaking, and gunshot, which belong to the rare sound event that may happen in daily life.

Amiriparian's approach [75] converts the audio data into mel-spectrogram as features. Then fed the sliced mel-spectrogram frames into the 2D convolutional network to extract high-level features. Afterward, the high-level features are fed into the Long Short-Term Memory-based recurrent network (LSTM-RNN). Finally, the predictions are outputted by the feed-forward network (FNN), cascaded after the LSTM-RNN. Amiriparian's approach achieved top performance, and the distress detecting module of the SA system is implemented according to this approach.

CHAPTER 9

Video signal-based Perception System

In this chapter, the camera-based perception modules are intorduced. These modules are: pedestrian detecting, tracking, positioning module, fall detecting module, and pedestrians traffic counting module.

The work in this chapter relies on existing research, but its implementation is an important component of the proposed system. I do not claim the scientific contribution to the work in this chapter.

9.1 Introduction

The video data collected by the camera can provide rich information. It can provide timely assistance for the accurate judgment of the perception system and the rescue work of the staff in the disaster. Although camera-based methods involve privacy violations, in order to ensure the reliability and robustness of the SA system described in this thesis, video-based perception modules are also introduced. Controlled by the crisis-privacy status model, the perception module based on video data will only be enabled in high-crisis status.

By studying the crisis in the public places and the rescue scenarios, it is helpful for the SA system to detect pedestrians, the position of the pedestrians, the behavior of the pedestrians, and tracks their routine. Pedestrian falling is a hallmark behavior that should be paid attention to by the perception system and the rescue staff. Meanwhile, the number of people in an area is also essential information rescuers need to know in a crisis. Thus, the proposed SA system implemented the pedestrians detecting, tracking, position module, fall detecting module, and the pedestrian traffic counting module.

In recent years, deep neural networks presented impressive performance regarding computer vision-related tasks. The practice has verified the advantages of deep learning-based

methods in object detection, target recognition, and target tracking tasks based on video data. Cao's paper [76, 77] proposed a system, called "OpenPose", to detect the anatomical keypoints of human basing on 2D images. This approach can detect the persons in images and mark these persons' anatomical key points. It can support multiple people in the input images and output the key points of the persons in the input images in real-time. Our SA system pedestrians detecting, positioning, and fall detecting functions are based on this approach. Wojke's paper proposed approaches to re-identify person [78] and track person [79] and implemented these approaches in a system called "DeepSORT". The DeepSort can re-identify each person in the video stream and track each person. Our pedestrian tracking function and traffic counting function are based on these approaches.

9.2 Implementation

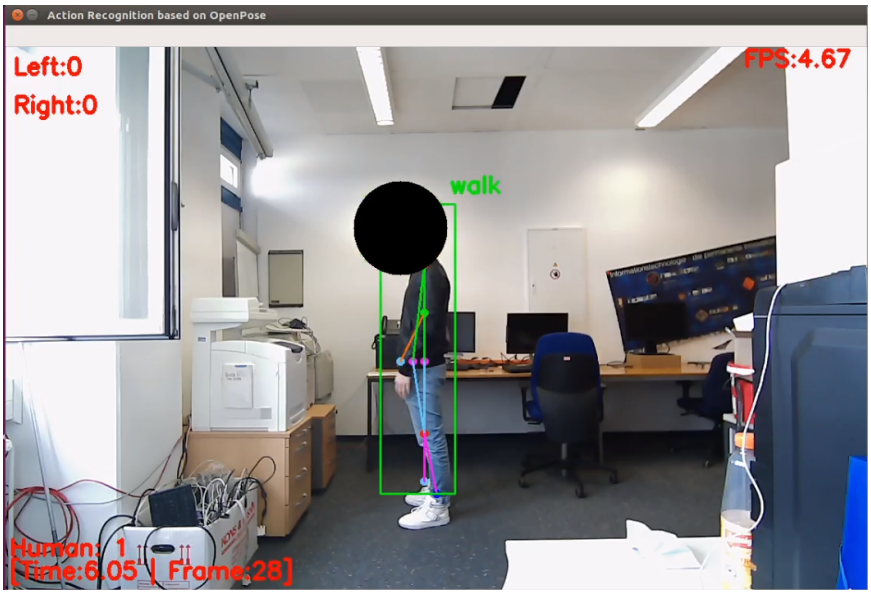


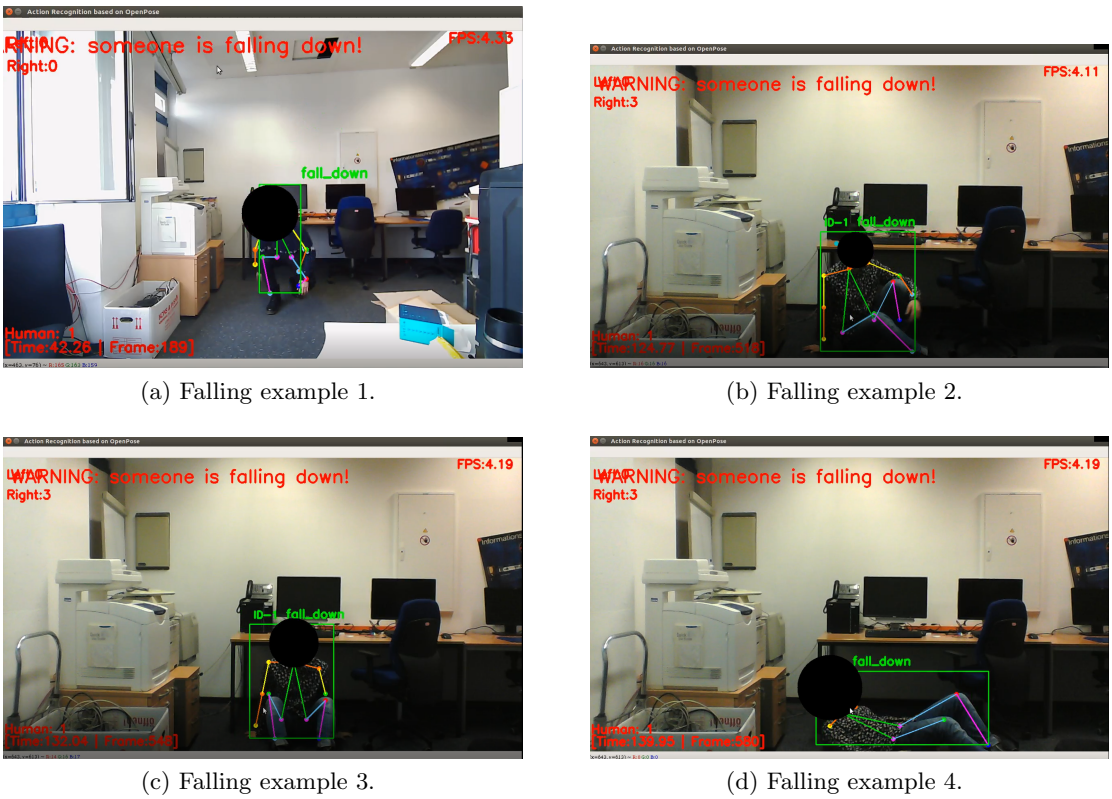
Figure 9.1: Video perception module screen example

9.2.1 Pedestrians Detecting, Tracking, and Positioning Module

The OpenPose [76, 77] can recognize the human from the input images and output the skeleton information of human body. DeepSort [78, 79] is able to recognize if the person in the different video frames is the same, thus tracking the person in the video. The pedestrians detecting, tracking, and positioning module combine the output of the OpenPose, the DeepSort, and the physical position of the cameras to locate and track the pedestrians in the surveillance area.

As shown in Fig. 9.1, a person appeares in the surveillance area. This person is marked with a green box and a skeleton. The person's face area has been automatically cut out of the image by the system for desensitization purposes. The system detects that the person is in a normal state (walking state).

9.2.2 Fall Detection Module



(a) Falling example 1.

(b) Falling example 2.

(c) Falling example 3.

(d) Falling example 4.

Figure 9.2: Falling examples

In Chen's paper [80], the posture of people in the act of falling has been studied, and an approach to detect human falling based on OpenPose is proposed. This approach extracts the skeleton information of the human body with OpenPose based on video images and detects falling according to the following indexes: *"(1) The angle between the centerline of the body and the floor; (2) The angle between the centerline of the human and the ground; (3) The width to height ratio of the human body external rectangular [80]."* Our fall detecting module implemented this approach.

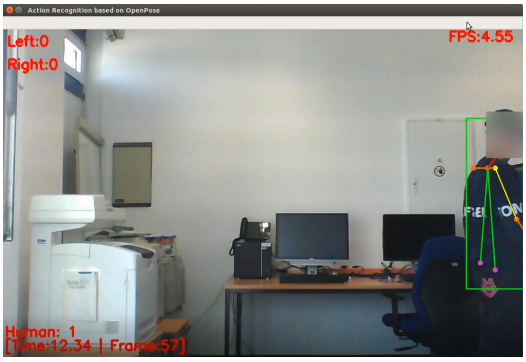
As shown in Fig. 9.2, the system detects that a person has fallen. There are various postures when a person falls. No matter what the posture of the person who fell, our system can recognize it. The system has issued a warning and marked the warning sign in the upper left corner of the field of view.

9.2.3 Pedestraits Traffic Counting

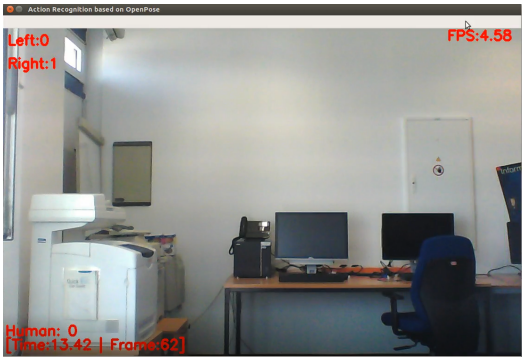
Figure 9.3: The pedestraits traffic counting method

The pedestrian traffic counting module receives data from cameras at the area's entrance. Then it combines the output data of DeepSort to count the number of persons entering and exiting the specified scene to measure the traffic of people in the area. The system has a pedestrian identification re-identification function. The identities of pedestrians in the same frame are uniquely identified. As shown in Fig. 9.3, the field of view of the camera is divided into three parts: left, middle and right. When the system detects a pedestrian moving from the green area to the yellow area on the left side of the field of view, the left counter is incremented by 1. Similarly, when a pedestrian moves from the green area to the right yellow area, the right counter is incremented by 1.

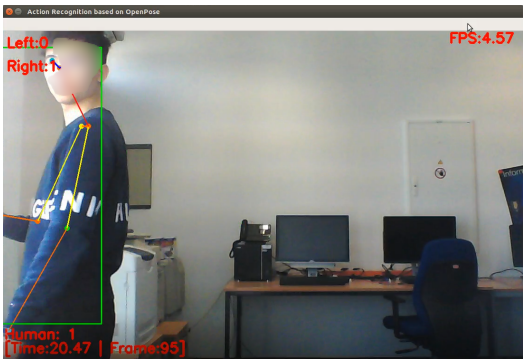
Fig. 9.4a shows a pedestrian moving to the right of the field of view. Fig. 9.4b shows that the pedestrian has moved out of the right side of the field of view. The upper left corner of the image shows the number of people disappearing to the right of the field of view. Similarly, Fig. 9.4c and Fig. 9.4d showns the pedestrian moves to the left of the field ov view and disappears. The lower-left corner of the figure shows the number of pedestrians in the current field of view. We assume that the camera is deployed at the entrance of the building. Moreover, the entrance is to the right of the camera's field of view. Whenever the record moves a pedestrian to the right, we know that a person walked out of the building. Whenever the record moves a pedestrian to the left, we know that a person has entered the building. Assuming cameras are deployed at all entrances, the system can count the number of people in the building by summing up the pedestrians' throughput of all entrances.



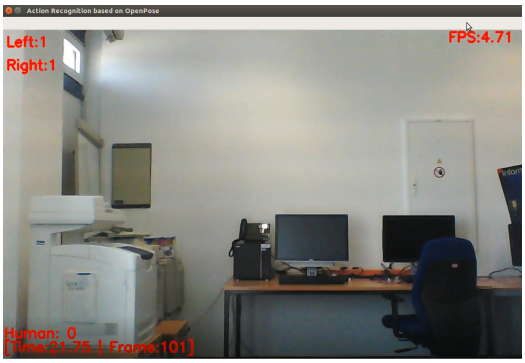
(a) Move to the right before disappearing to the right edge of the field of view.



(b) Move to the right after disappearing to the right edge of the field of view.



(c) Move to the left before disappearing to the left edge of the field of view.



(d) Move to the left after disappearing to the left edge of the field of view.

Figure 9.4: Pedestraits Traffic Counting

CHAPTER 10

Representation Layer

In this chapter, the representation layer is discussed. The representation layer of the proposed SA system includes the Unity 3D-based anonymous synthesis visualization representation methods and the reality camera video representation methods. Because the latter method simply transmitted the camera signal to the relevant staff, which does not involve scientific innovation, this chapter will only explain the Unity 3D-based anonymous synthesis visualization representation methods.

The work in this chapter relies on existing research or engineering practice, but its implementation is an important component of the proposed system. I do not claim a scientific contribution to the work in this chapter.

10.1 Introduction

The representation layer of the SA proposed system delivers the perception results to the related personnel like mall security and rescue workers intuitively. The representation layer includes two methods to represent the perceived results. The first method is the Unity 3D-based anonymous synthesis visualization representation method. This method is anonymous because it uses avatars to represent the pedestrians in the monitoring area that works under the low crisis status. The monitoring places are represented using Unity 3D-based game engine-based 3D virtual scene. With this method, the viewer can only see the anonymous avatar moving in the specified virtual scene. The location of the pedestrians will be displayed on the screen with an avatar so that the rescue staff will know the exact number of people in the area who need to evacuate when a crisis happens.

Unity 3D [81] is a game engine. People can build interactive real-time 3D content with Unity 3D. Unity 3D game engine can render and display controlled instances of the 3D model. Our SA perception layer records the perception results in a database. The Unity 3D-based anonymous synthesis visualization module queries the database, including the

(a) Example 1. (b) Example 2.

(c) Example 3. (d) Example 4.

Figure 10.1: Anonymous avatars in synthetic scenes.

number and location of pedestrians in the monitoring area, through an HTTP request. It renders them in the virtual 3D scene. This module can work both on the high-privacy status and the high-crisis status.

The "reality camera video representation method" will be enabled when the SA system sets the crisis level to high. Related personnel can acquire visualized information both from real-time video streams of the cameras and the virtual 3D scene.

10.2 Unity 3D-based Anonymous Synthesis Visualization

With the help of the Unity 3D engine, we can visualize the scene of the smart environment and the moving pedestrians in the scene. Fig. 10.1 is an example of a smart virtual environment we have implemented. As we can see from the figure, the person in the scene is a virtual character and is not binding to the identity of any actual pedestrian in the scene. This method anonymizes the characters acting in the scene.

The number of pedestrians in the virtual scene and the real-time location of pedestrians are updated in real-time by querying the database. The number and location information of these pedestrians are acquired by the system perception layer and stored in the database. Our virtual scene system will ask for configuration server information during the startup phase, as shown in the Fig. 10.2

The virtual scene in Fig.10.1 includes furniture such as tables and chairs. These virtual

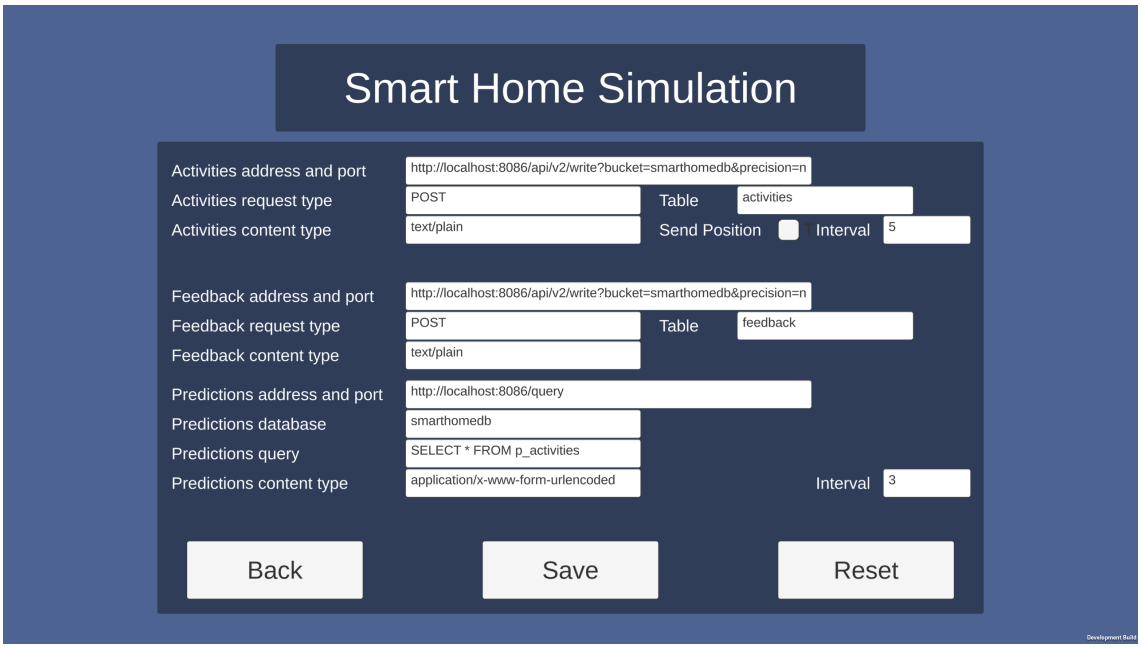


Figure 10.2: Server configuration.

objects' 3D modeling and placement require manual design and implementation. However, considering the practical requirements for situational awareness in a smart environment, visualization of this furniture is not necessary. We are more concerned with the number and location of pedestrians in the monitored area. So in actual use, we will use automatically generated simple scenes to represent the environment, as shown in Fig. 10.3.

The area size of the virtual scene can be automatically generated according to the specified size. This saves the work of manual modeling. Fig. 10.4 shows an example when there is one person and four people in the scene. Whenever a pedestrian appears in the monitoring area of the real world, a virtual character corresponding to it will appear in the virtual scene. Our unity3d visualization system periodically pulls the updated perception data from the database and displays it in the virtual scene.

10.3 Camera Video with Privacy Protection

In high crisis status (low privacy status), the video representation unit based on camera data will be activated. Usually, at this time, the system perception module senses the emergency and adjusts the privacy status to high crisis and low privacy. Video from cameras in the smart environment will be shown to security personnel.

As shown in Fig. 10.5, the video representation module of the system will detect the people's faces and erase the faces from each frame of the video stream. Fig. 10.5 shows examples of the erasure effect when the pedestrian is at different distances from the camera

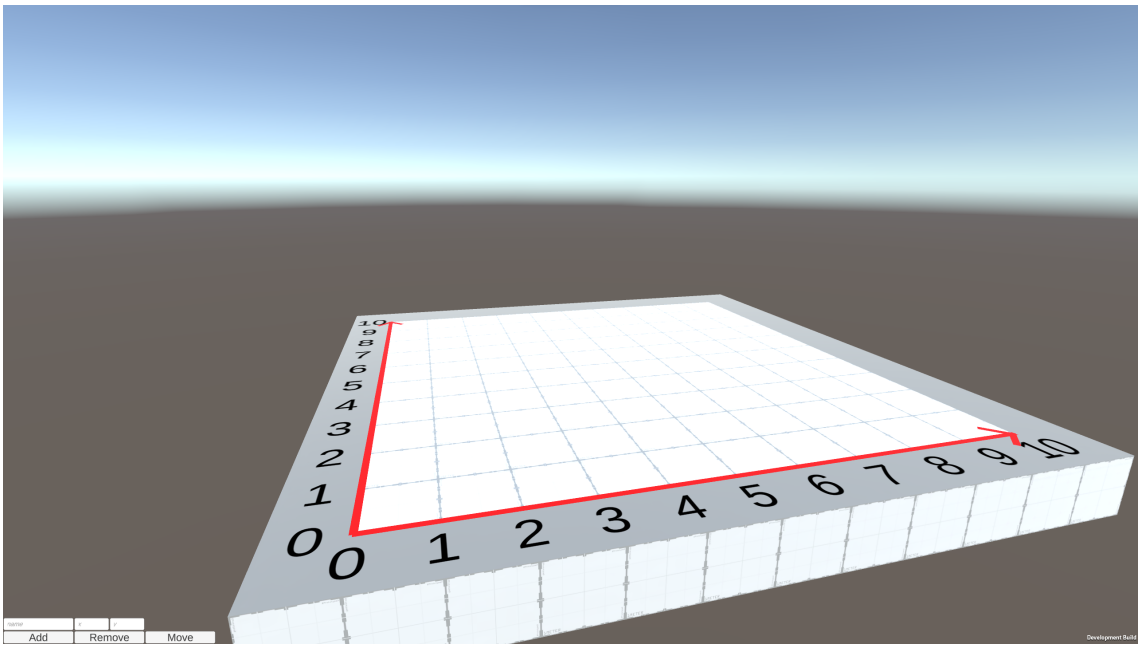
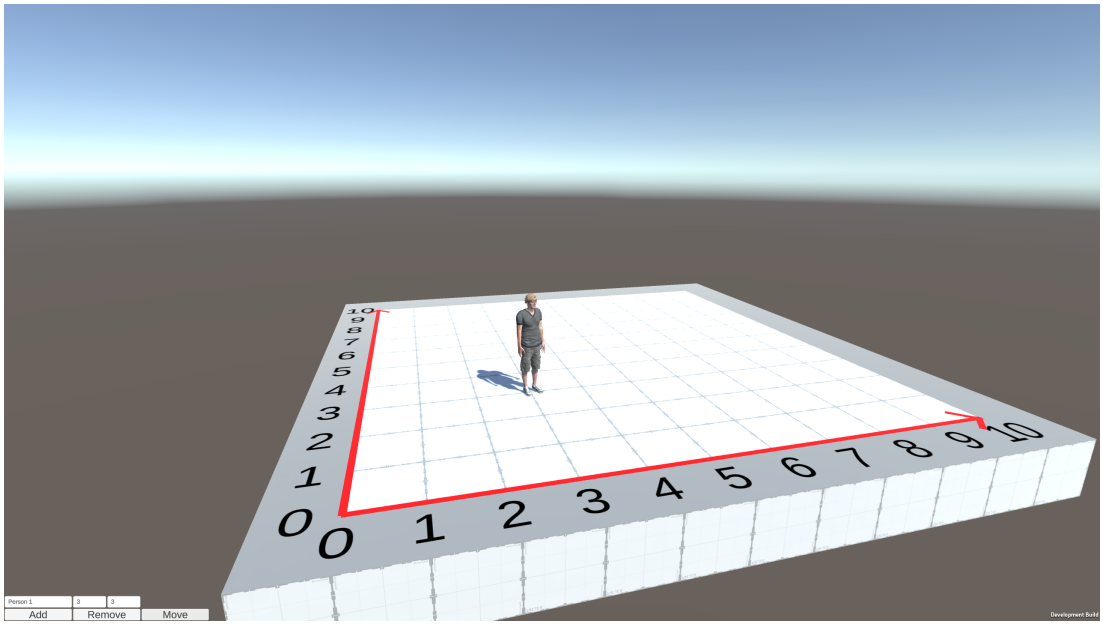


Figure 10.3: Automatically generate simplified scenarios.

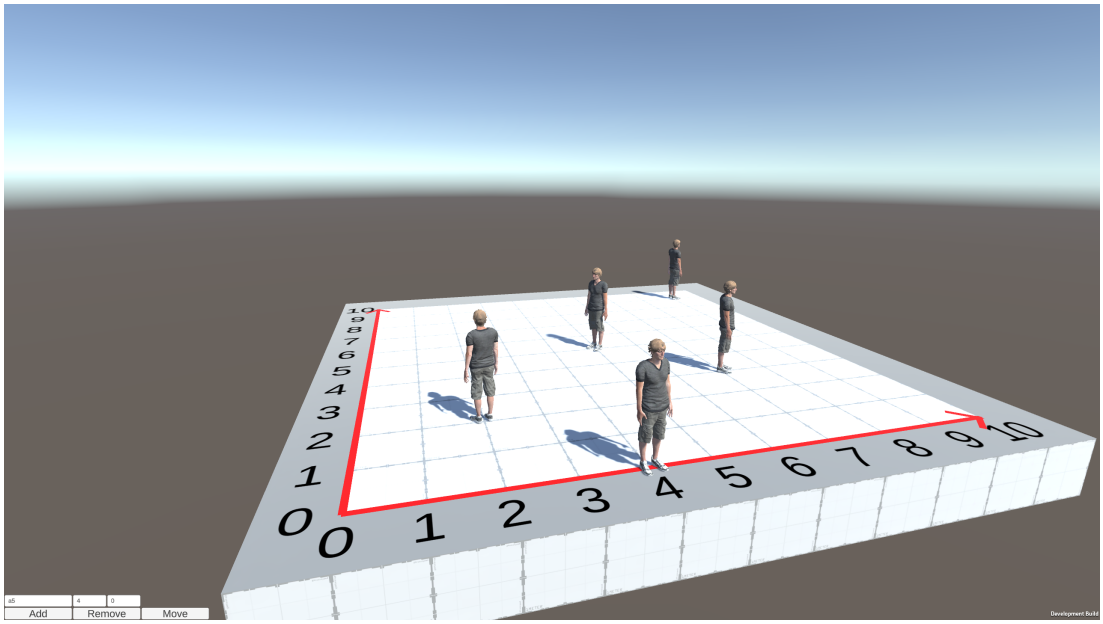
and the pedestrian is facing different directions.

The recognition of the face area is realized based on the pedestrian body recognition function rather than face recognition. Even if the pedestrian is facing away from the camera, the system will still erase the image area above the person's shoulder.

Usually, simply erasing the facial image will satisfy the requirements for anonymization of personal image data. However, to avoid identifying human identity through video information of clothing, shoes, and hats, our system also supports erasing the entire human body area, as shown in Fig. 10.6.



(a) 1 Person in the scene.



(b) 5 Persons in the scene.

Figure 10.4: Pedestrians visualization in the automatically generated scene.

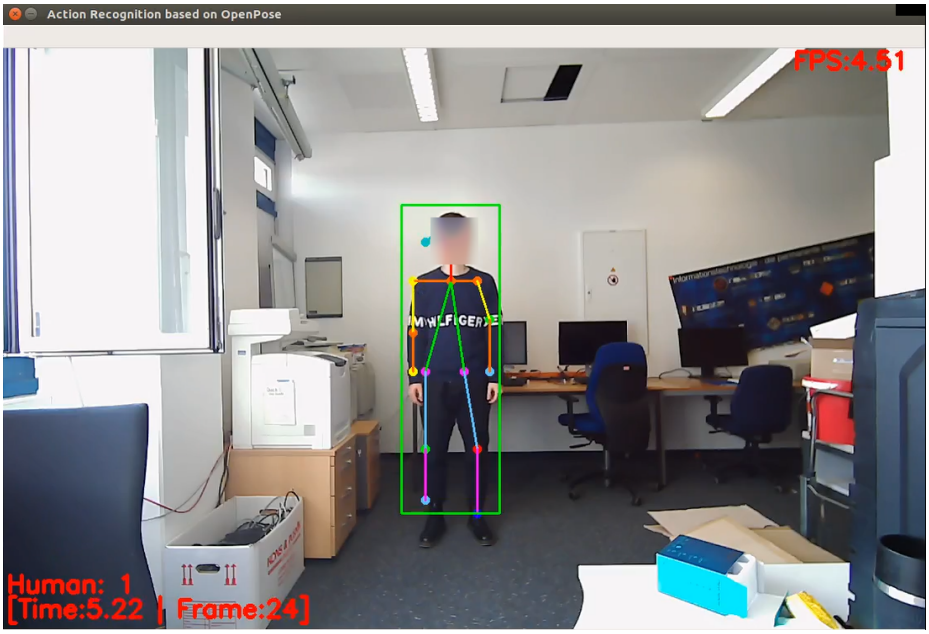
(a) Face erasing example 1.

(b) Face erasing example 2.

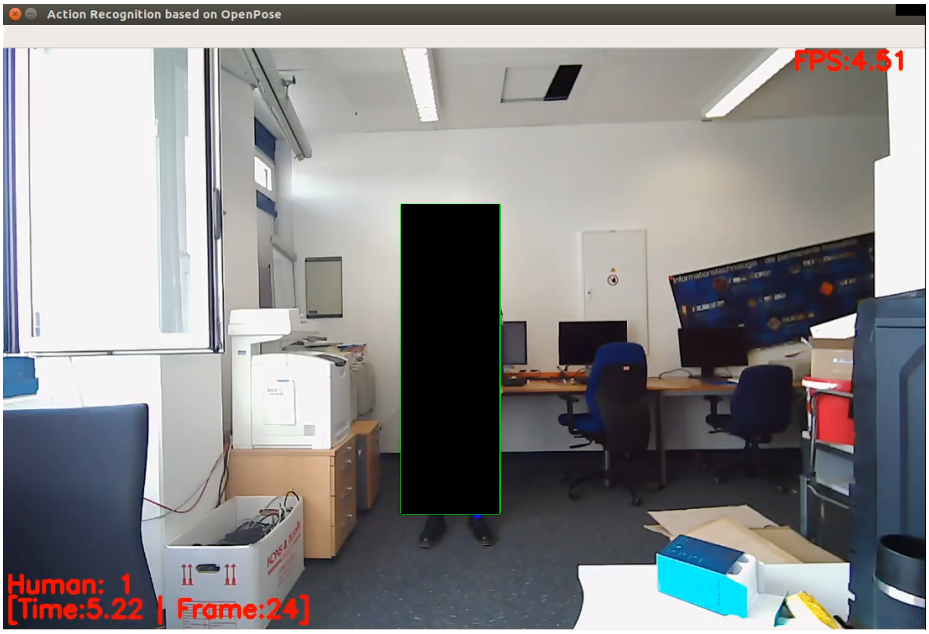
(c) Face erasing example 3.

(d) Face erasing example 4.

Figure 10.5: Video frames without face.



(a) Body image before erasing.



(b) Body image after erasing.

Figure 10.6: Video frames without body.

CHAPTER 11

Conclusion and Outlook

11.1 Conclusions

This dissertation proposed a situational awareness system that provides privacy-protecting features. The situational awareness system perceives the crisis status of the monitored area, pedestrians' behaviors, traffic, and pedestrians' position. The system receives the data from vibration sensors, audio sensors, and cameras and perceives the situation according to the input data. The emergency detecting modules adjust the crisis status. The crisis-privacy status model restricts the usability of the above sensors. When the monitored area is under high crisis status, audio and camera will be enabled, and the real-time video stream will be enabled. In contrast, only the vibration sensors-based perception module and the audio-based exploration detecting module are enabled under low emergency status. The proposed system offers two representation methods for the perception results. Similarly controlled by the crisis-privacy status model, the Unity 3D-based anonymous synthesis visualization representation method works on both high crisis and high privacy statuses. In contrast, the reality camera video representation method only works on the high-crisis model.

We innovatively proposed a set of approaches and algorithms that perceive the monitored area situation from several critical perspectives in a privacy-protecting manner. These approaches and algorithms are only based on vibration sensor data, thus providing the privacy-protecting feature. The implementation of these approaches is able to detect the emergency, count the number of the pedestrians, locate and track the pedestrians, and detect the walking direction of the pedestrians. These vibration sensor-based approaches can fulfill the perception tasks and protect the people's privacy in the monitored area, which reached the state-of-the-art. In addition to our vibration sensors-based approach, we also implemented the audio-based exploration detection module, audio-based distress detection module, the video signal-based pedestrians detecting, tracking, positioning module, fall detecting module, and the pedestrians' traffic counting module. These perception

modules using a variety of modal sensors input signals, serving as essential components of the SA system, provide the system with higher reliability and robustness. Meanwhile, we implemented a Unity 3D-based 3D scene for the anonymous synthesis visualization of the perception results. To sum up, the proposed solution in this dissertation solved the perception and representation problems when privacy-protecting is concerned.

Moreover, the perception modules in the proposed SA system depend on the infrastructures needed to implement additional sensors in the monitoring area. Recently, some researchers have used WiFi devices as radar to solve perception problems. As in the current society, WiFi devices belong to the infrastructures that most buildings will deploy. The WiFi devices-based pedestrian positioning, tracking, and walking direction detecting approaches will be studied in future works. Finnaly, we have integrated all the vibration signal-based functional modules [11, 12, 13] into one system, which includes situational awareness and visualization. This system simultaneously accomplishes two tasks of situational awareness and privacy protection while reconciling the conflict between protecting privacy and perception algorithms relying on identity-sensitive data.

To sum up, the scitics contributions can be summarized as follows.

- This thesis innovatively contributed a smart environment situation perception and visualization solutions that support privacy protection.

- This thesis innovatively contributes an approach for emergency detection in public areas. This emergency detection approach is privacy-preserving. Because it only relies on the floor vibration signal, which is identity desensitized.

- This thesis contributes a privacy-preserving pedestrian counting approach. This pedestrian counting method only relies on ground vibration signals. Compared with the similar existing methods, it has smaller usage restrictions and higher precision.

- This thesis innovatively contributes a step-level pedestrian walking direction detection approach. With competitive accuracy, this approach can derive the movement direction in only three steps.

- This thesis innovatively contributes a pedestrian localization method based on vibration signals.

- This thesis innovatively proposes a situational awareness and visualization framework controlled by a privacy crisis state model while considering the actual requirements of privacy protection and perception. This framework reconciles the conflict between protecting privacy and perception algorithms relying on identity-sensitive data.

11.2 Outlook

In order to narrow the individual-to-individual differences of each sensor, improve signal quality, and reduce deployment costs, we will customize the floor tiles embedded with piezoelectric sensors. In the current system, the vibration sensor-based perception modules used piezoelectric sensors, which are fixed on the floor's surface. This affects the everyday use of the area by people. We can build bricks with embedded piezoelectric sensors in the future, thus allowing our approach to better transition to the final product. Meanwhile, the noise of the vibration signal sensor is reduced, and the consistency of signal characteristics between different sensors is improved. Furthermore, the versatility of the related deep learning models is increased. Additionally, we will introduce the domain adaptation [82] technology into our vibration signal-based perception module to further improve the adaptability of the neural network.

The emergency detecting module setup is deployed in a 3-meter by 3-meter area. In the future, we will extend the setup to a larger area using significantly more sensors. When sensor node networks are deployed on a large scale, cooperative communication between different sensor units and cooperative prediction can be a new exploration direction. At the same time, the solution for low-cost and high-precision time synchronization between sensors in distributed sensor networks is also an exciting research topic. In the future, we will explore the possibility of synergy between the output data of the perception layer and the metaverse [83, 84], thus towards constructing a privacy-protecting smart city.

CHAPTER 12

Included Publications

This chapter provides bibliographic information and reprints of the publications included in this cumulative dissertation.

12.1 A Privacy-Protecting Indoor Emergency Monitoring System based on Floor Vibration

Title	A Privacy-Protecting Indoor Emergency Monitoring System based on Floor Vibration
Authors	Yang Yu, Torben Weis
Publication Venue	Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp/ISWC '20 Adjunct), September 12-16, 2020, Virtual Event, Mexico
Publication Type	Extended Abstract
Publication Status	Published
Research Topic	Emergency Detecting
DOI	https://doi.org/10.1145/3410530.3414423

Table 12.1: Bibliographic information of publication A.

A Privacy-Protecting Indoor Emergency Monitoring System based on Floor Vibration

Yang Yu
Distributed Systems Group, University of Duisburg-Essen
Duisburg, Germany
yang.yu@uni-due.de

Torben Weis
Distributed Systems Group, University of Duisburg-Essen
Duisburg, Germany
torben.weis@uni-due.de

ABSTRACT

In this work we present an indoor emergency context monitoring system based on ground vibration caused by persons in the target area. The system is designed for production plants and large buildings to perceive the safety status of this area. Our approach is privacy-protecting, because it requires neither video nor sound. Instead, piezo sensors on the floor measure vibrations, which are analyzed with machine learning to compute the safety status of the covered area. This way our system can determine whether an emergency occurred, but it is not straight forward possible to attach names to the detected persons. We compare the impact of different feature extraction methods and different types of classifiers on the classification results. Our experiments show that we can determine an emergency event with an average F1 score of 0.97.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Human-centered computing** → **Ubiquitous computing**.

KEYWORDS

vibration signal, pattern recognition, deep learning, privacy protection, emergency detecton, context recognition, piezo sensor

ACM Reference Format:
Yang Yu and Torben Weis. 2020. A Privacy-Protecting Indoor Emergency Monitoring System based on Floor Vibration. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp/ISWC '20 Adjunct), September 12–16, 2020, Virtual Event, Mexico.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3410530.3414423

UbiComp/ISWC '20 Adjunct, September 12–16, 2020, Virtual Event, Mexico

1 INTRODUCTION

This work wants to lay a foundation stone for a new intelligent indoor context perception system, which mainly focuses on emergency detection.

Many emergency systems in buildings rely on the people inside to recognize an emergency situation and report it, for example by pressing an alarm button or calling the police or fire brigade. The problem is that not everyone immediately calls for help and reports such a situation, maybe because he forgets or is unable to do so, and just leaves the danger zone first. This behavior delays the time until an emergency is reported, bringing other people in danger by preventing a fast intervention. Although there are often sensors installed that can detect smoke or gas, there are numerous other emergencies caused by failing machines or building constructions, fighting or the appearance of a gunman that cannot be detected by these sensors. Our system focuses on measuring human behavior to detect an emergency, and at the same time.

During an emergency, people try to escape either running or moving with fast and small steps. Then the vibration signal is generated by the pressure people put on the ground when walking. The state of mind of the people in this environment is reflected by the vibration signal. By analyzing the vibration signal and recognizing the pattern, we can infer the context situation in a specific area.

The advantage of our approach is that we do not rely on persons wearing any senders or receivers. In addition, our approach uses very cheap piezo sensors rather than expensive geophones[7]. From a cost perspective, this allows for large-scale commercial deployment. Furthermore, our system is less privacy-invasive than cameras and it does not detect who is walking. In addition, our system can work in smoked areas where cameras do not work anymore.

The key contribution of this paper is as follows. We present a novel machine learning-based indoor emergency monitoring system that relies on the reaction of people rather than trying to detect the cause of the emergency. Thus, our approach is more general than special purpose gas or smoke detectors. Furthermore, our system relies on cheap vibration sensors only. Being a passive approach, it does not require

people to carry a device. We evaluated and validated the classification performance of our system with lab experiments.

2 RELATED WORKS

Researchers before have used acceleration sensors of smartphones to recognize activity[1, 3], or used Bluetooth Low energy beacons[2] to detect occupant movement patterns. These approaches could be used to deduce the situation in an indoor environment. However, this kind of method requires people to carry a mobile device and to install special software. This renders these approaches impractical. Also some research use video-based methods[5, 8], to detect the movement of persons. However, an environment with ubiquitous video surveillance is very privacy invasive.

In contrast, our approach uses cheap piezo sensors deployed in the area that, and it does not require anybody to carry a special device. In addition, our system does not rely on audio or video recordings and thus our system design respects people's privacy.

3 APPROACH

Our emergency monitoring function is based on human behavior, because this allows for sensing a wide range of emergency situations independently from the cause of the emergency. The proposed approach considers four kinds of situations: "none", "walking normally", "walking fast", and "panic". When persons are walking normally or even a bit faster, we classify it as the normal walking class or fast walking class. We consider both classes to indicate a normal situation. When persons are moving in panic, we classify it as an emergency situation. The "none" state means there is no person in this area. In our lab experiments, we tested 1 to 4 people moving in the $9m^2$ area covered by four sensors. Larger areas can be covered by deploying more sensors in a grid.

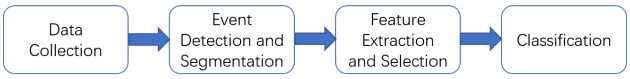


Figure 1: Emergency Detector Processing Flow.

As shown in Figure 1, our emergency monitoring includes the steps data collection, event detection, data segmentation, feature extraction, feature selection, and classification. The experiment shows that a deep learning-based [4] end to end emergency detection system has the best performance. In several comparative experiments we combined principal components analysis-based (PCA) automatic feature extraction methods and manual feature selection methods with a random forest classifier. In addition, we compared the classification performance to that of a deep neural network. The performance analysis for each scheme is discussed in section 4. In this section, the design details of each procedure are discussed in depth.

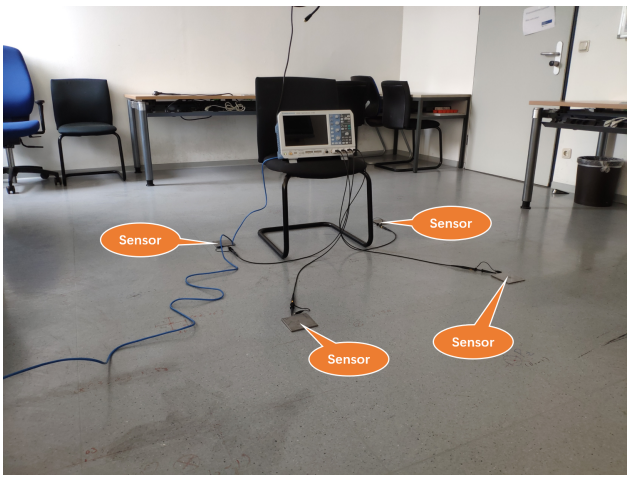


Figure 2: Experiment setup.

Data Collection

The floor vibration signal is sensed with a sensor matrix of four EPZ-27MS44W piezoelectric sensors as shown in Figure 2. This sensor can detect a frequency band ranging from 0 Hz to 4400 Hz. The signal was amplified, sampled, quantized, and recorded with a R&S-RTB2000 oscilloscope. The sampling rate is 10000Sam/s, which is more than double the maximum bandwidth of the sensor.

Event Detection and Data Segmentation

To detect an event, we split the vibration sensor input data into windows using a sliding window algorithm. In each step, the window is sliding forward by one value. If a certain amount of values in this window is greater than the mean of the amplitude of the peaks multiplied by an arbitrary factor, the current window is further analyzed, because it might contain a significant event. All other windows are dismissed as being non-significant. When a significant window has been found and processed by the classifier, the window slides forward a given overlap percentage and the continuous sliding

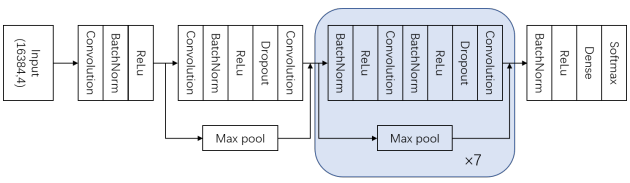


Figure 3: Deep neural network architecture.

forward value by value. This way, we reduce the number of windows that have to be passed through the classifier. Equation (1) shows the maximum number of windows that have to be passed through the classifier for a given number of input values.

$$windows = \frac{number\ of\ values}{window\ size - window\ size \times \frac{overlap\ percentage}{100}} \quad (1)$$

Based on the results of our experiments, we choose 16384 as the window size, with a time length of 1.638 seconds and an overlap of 50%.

Feature Extraction and Selection

The first three schemes designed in this paper use a random forest as a classifier combined with one of three feature extraction methods. The fourth scheme uses an end to end deep learning based approach, which conducts the feature extraction and classification process completely inside the deep neural network.

In the first two schemes we use principal components analysis(PCA) as an automatic feature extraction method. We used PCA 512 and PCA 64, which reduce the sample dimension from 16384 to 512 and 64 respectively.

In addition, we implemented manual feature extraction. We have chosen the maximum value, the "range", the 90% quantile, the mean, the "bin18" value and the amplitude and frequency of the first 3 peaks in the frequency domain from each sample as features. The results in 44 features for 4 sensors and has proven to yield good performance. The "range" value is computed as the maximum value minus the minimum value. The "bin18" means the 18th bin probability value of the distribution, where the sampling point values are normalized to -1 to 1 and the space between -1 and 1 is divided into 20 equally-sized bins.

Classification

In the first three schemes, we use a random forest[6] as the classifier. Based on the repeated experiments, the number of estimators of the random forest is set to 91 and 30 maximum depth. The input of this classifier are PCA feature values or manually selected features values.

In the fourth scheme, we use a convolutional neural network with residual structure as an end to end approach to analyze the data sample and conduct classification. In this scheme, the deep neural network receives raw event data set as input and outputs one of the four classes "walking", "fast walking", "panic", and "none". The deep neural network architecture is shown in Figure 3.

Table 1: Classification Performance of the Emergency Detection Classifiers

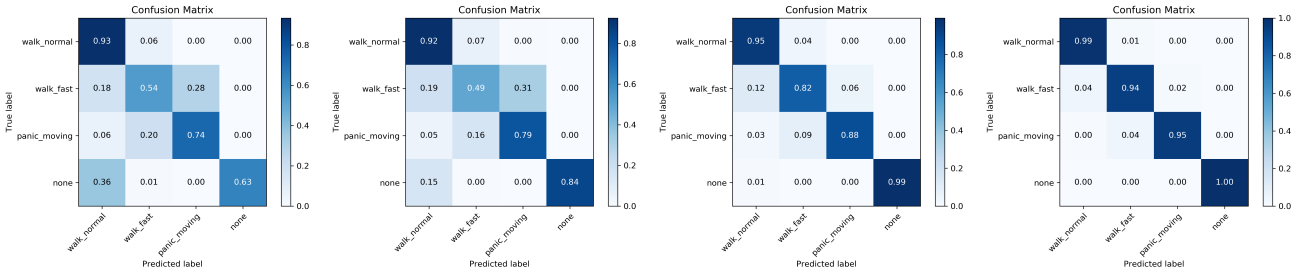
Method	Class	Precision	Recall	F1-score
PCA64 + RF	Walk Normal	0.7670	0.9333	0.8420
	Walk Fast	0.6358	0.5399	0.5839
	Panic Moving	0.7261	0.7351	0.7306
	None	0.9937	0.6318	0.7724
	Weighted Avg.	0.7543	0.7482	0.7420
PCA512 + RF	Walk Normal	0.8120	0.9242	0.8645
	Walk Fast	0.6460	0.4935	0.5596
	Panic Moving	0.7174	0.7914	0.7526
	None	0.9976	0.8431	0.9138
	Weighted Avg.	0.7719	0.7753	0.7685
MF + RF	Walk Normal	0.9031	0.9544	0.9280
	Walk Fast	0.8339	0.8168	0.8253
	Panic Moving	0.9356	0.8790	0.9065
	None	1.0000	0.9920	0.9960
	Weighted Avg.	0.9071	0.9068	0.9064
DNN	Walk Normal	0.9690	0.9874	0.9781
	Walk Fast	0.9357	0.9407	0.9382
	Panic Moving	0.9849	0.9510	0.9676
	None	0.9980	1.0000	0.9990
	Weighted Avg.	0.9687	0.9685	0.9685

4 EVALUATION

In this section, the classification performance of the emergency detection system is discussed. The classifier can detect the classes "none", "walking", "fast walking", and "panic". When the system detects a "panic", we conclude that persons are behaving as expected due to an emergency situation. We analyze the classification performance of our four schemes in this section.

Data organization

The data sets include measurements of persons walking, fast walking, running in panic, and measurements of an empty room. In the experiment for "panic moving" data collection, the experimenters imitated the movement as during the emergency, running fast, or moving with fast pace and small steps. The data set for each class include measurements with one, two, three and four persons. We totally recorded 19048 data samples including the four classes and one to four persons moving in the same area. We labeled each class of data samples and then mixed them randomly. We divided the data samples into a training set, a verification set, and a test set according to a 60% : 20% : 20% ratio. The training set is used to train the classifier models, the verification set is used to



(a) PCA64 + RF. (b) PCA512 + RF. (c) MF + RF. (d) DNN.

Figure 4: Figure 4(a) to Figure 4(d) are the confusion matrixes of the four schemes.

test and optimize the parameters of the classifiers and the test set is used for evaluation.

The performance results shown in this section are based on the test set, in which the data had never been used in the research and exploration stage. For the evaluation purpose, a total of 3809 samples of the test set were used, which included 497 none class samples, 1425 walking normal class samples, 928 walking fast class samples, and 959 panic moving class samples.

Classification Performance

The Table 1 shows the classification performance. In the table, PCA64 means the PCA is used to extract 64 feature values from each sample. PCA 512 means the PCA is used to extract 512 values from each sample. MF means manual feature extraction. RF means random forest, which is always combined with a feature extraction method. DNN means the end to end deep learning-based scheme.

The average F-measures are 0.75 for "PCA64 + RF" , 0.77 for "PCA512 + RF", 0.91 for "MF + RF" and 0.97 for "DNN". We can see that for the PCA feature extraction method, more feature values slightly improved the classification performance. For manual features, the random forest classification algorithm has the least number of input features and obtains better classification performance than PCA based methods. The end to end deep learning based method obtains the best classification performance among all the methods.

Considering the emergency detection purpose of our system, the classification performance metric for panic moving is the most important. The four schemes have the capability of classification of panic moving class, with F1-score 0.73, 0.75, 0.91, 0.97 for "PCA64 + RF", "PCA512 + RF", "MF + RF" and "DNN" schemes.

5 CONCLUSION

In this paper, we presented a system for identifying an environment by analyzing the movement of persons. Our system can detect an emergency using piezoelectric sensors only, which is protected people's privacy. Furthermore, our system is not limited by the cause of the emergency, since we measure people's reaction rather than the cause.

Our classifier distinguishes between normal walking, fast walking, panic moving and an empty area. We compared different feature extracting method and classifiers. In the end, the deep neural network yielded the best results. As an additional advantage, this scheme does not require handpicked features.

In the future, we will extend the setup to a larger area using significantly more sensors.

REFERENCES

[1] Ling Bao and Stephen S Intille. 2004. Activity recognition from user-annotated acceleration data. In *International conference on pervasive computing*. Springer, 1–17.

[2] Avgoustinos Filippoupolitis, William Oliff, and George Loukas. 2016. Occupancy detection for building emergency management using BLE beacons. In *International Symposium on Computer and Information Sciences*. Springer, 233–240.

[3] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. *Nature* 521, 7553 (May 2015), 436–444. https://doi.org/10.1038/nature14539

[5] Jun Lee, Se-Jong Lee, Jeong-Sik Park, and Yong-Ho Seo. 2012. Emergency Detection Method using Motion History Image for a Video-based Intelligent Security System. *Journal of Advanced Smart Convergence Vol* 1, 2 (2012), 39–42.

[6] Andy Liaw and Matthew Wiener. 2002. Classification and Regression by randomForest. 2 (2002), 6.

[7] Mostafa Mirshekari, Shijia Pan, Pei Zhang, and Hae Young Noh. [n.d.]. Characterizing wave propagation to improve indoor step-level person localization using floor vibration. In *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2016* (2016-04-20), Vol. 9803. International Society for Optics and Photonics, 980305. https://doi.org/10.1117/12.2222136

[8] Robert Tomastik, Yiqing Lin, and Andrzej Banaszuk. 2008. Video-based estimation of building occupancy during emergency egress. In *2008 American Control Conference*. IEEE, 894–901.

12.2 Pedestrian Counting Based on Piezoelectric Vibration Sensor

Title	Pedestrian Counting Based on Piezoelectric Vibration Sensor
Authors	Yang Yu, Xiangju Qin, Shabir Hussain, Weiyan Hou, Torben Weis
Publication Venue	Applied Sciences, February 12, 2022
Publication Type	Article
Publication Status	Published
Research Topic	Pedestrian Counting
DOI	https://doi.org/10.3390/app12041920

Table 12.2: Bibliographic information of publication B.

applied sciences

MDPI

Article

Pedestrian Counting Based on Piezoelectric Vibration Sensor

Yang Yu [1,2], Xiangju Qin [3], Shabir Hussain [2], Weiyan Hou [2,*] and Torben Weis [1]

[1] Distributed Systems Group, University of Duisburg-Essen, 47057 Duisburg, Germany; yang.yu@uni-due.de (Y.Y.); torben.weis@uni-due.de (T.W.)
[2] School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China; shabir@gs.zzu.edu.cn
[3] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, 00014 Helsinki, Finland; xiangju.qin@helsinki.fi
* Correspondence: houwy@zzu.edu.cn

Abstract: Pedestrian counting has attracted much interest of the academic and industry communities for its widespread application in many real-world scenarios. While many recent studies have focused on computer vision-based solutions for the problem, the deployment of cameras brings up concerns about privacy invasion. This paper proposes a novel indoor pedestrian counting approach, based on footstep-induced structural vibration signals with piezoelectric sensors. The approach is privacy-protecting because no audio or video data is acquired. Our approach analyzes the space-differential features from the vibration signals caused by pedestrian footsteps and outputs the number of pedestrians. The proposed approach supports multiple pedestrians walking together with signal mixture. Moreover, it makes no requirement about the number of groups of walking people in the detection area. The experimental results show that the averaged F1-score of our approach is over 0.98, which is better than the vibration signal-based state-of-the-art methods.

Keywords: vibration signal; pedestrian counting; pattern recognition; deep learning; privacy protection; piezoelectric sensor

Citation: Yu, Y.; Qin, X.; Hussain, S.; Hou, W.; Weis, T. Pedestrian Counting Based on Piezoelectric Vibration Sensor. *Appl. Sci.* **2022**, *12*, 1920. https://doi.org/10.3390/app12041920

Academic Editor: Mayank Kejriwal

Received: 24 January 2022
Accepted: 10 February 2022
Published: 12 February 2022

1. Introduction

Detecting the number of people in a specific area is of great importance in many real-world scenarios. It can support intelligent building and security monitoring applications, such as search and rescue after disasters, pedestrian traffic monitoring, energy-consuming optimization, indoor space management, marketing, and infection spread control for epidemic scenarios [1]. Meanwhile, people are very concerned about protecting their privacy when an intelligent monitoring system is deployed. Based on how the data is sensed, the existing approaches are categorized as device-based and device-free approaches (e.g., infrastructure-based approaches). The device-based approaches [2–4] require individuals in the monitored area to carry a special device or a smartphone. The infrastructure-based approaches [5–16] deploy sensors such as cameras, infrared sensors, and piezoelectric sensors where no requirement to carry any devices is needed.

However, previous studies have the following limitations. Firstly, the application scenarios of the device-based approaches are restricted, and such approaches are not appropriate to deploy in public places, such as shopping malls. As it is unrealistic to hand out a device to every individual in an open space or require everyone who enters the area to install a smartphone app. Secondly, although recently many studies have focused on camera-based crowd counting approaches (e.g., [17–21]), such approaches do not protect privacy and the deployed devices are easy to be destroyed. Besides that, the camera or infrared sensor-based approach will not work well in an extreme environment, such as areas with heavy smoke or low visibility. This greatly limits the deployment of the approach in certain real-life situations, such as rescue after disasters and security monitoring in a restricted area.

In this paper, we present a novel infrastructure-based approach based on piezoelectric sensors. The piezoelectric sensor is much cheaper than the geophone sensor [22] used in previous studies [14,15,23], which is advantageous from a cost perspective. Meanwhile, our approach does not require high-density sensor deployment [24]. Furthermore, different from the existing studies, our approach can be applied to many real-life scenarios where multiple people are in the same room.

While the identity authentication-based approaches [23,25,26] requires that the signals should not mix or there is only one person in the area, our approach can handle scenarios where signals from different people are mixed together. Our approach does not require that only one group of people should be in the monitored area [15]. Our system can detect the number of pedestrians in many possible cases where groups of people may walk with different walking speed, frequency, and directions. In this work, we consider the cases from 0 person to 4 persons in a 3 m by 3 m area. Our system can be treated as a minimal functional unit and be scaled out to support more people in a larger area. In future products, sensors can be embedded in floor tiles to unify the physical transmission characteristics of vibration signal.

The major contributions of our work are as follows:

- We propose a novel approach that can count the number of people with vibration signals from the piezoelectric sensors while protecting privacy.
- Our approach supports the situations where multiple people walk together with the signals mixed.
- Our approach does not require that only one group of people should be in the detection area.
- Different from the room-level approach [14], our approach is a step-level pedestrian counting approach, making it more appropriate for many real-world applications.
- Our approach uses piezoelectric sensors, which are much cheaper than geophone sensors, making our solution economically viable.
- Experimental evaluation shows that our approach outperforms the vibration signal based state-of-the-art methods in accuracy for similar pedestrian counting task.

This paper is structured as follows. In Section 2, we discuss previous works regarding infrastructure-based pedestrian counting approaches from different perspectives, which motivates our approach. In Sections 3 and 4, we introduce and present our systematic approach and methodology. In Section 5, we present the experimental evaluation of our system. We conclude the work in Section 6 and discuss potential future directions.

2. Related Work

The vibration signal not only contains rich environmental information but also causes no invasion of privacy. Vibration signal-based device-free situation awareness detecting approaches have attracted much attention from the academic and industry community, which shows great potential in pervasive computing applications [27–29].

2.1. Sensor Selection

In general, recent studies regarding vibration signal-based ubiquitous computing applications mainly use geophones (triaxial seismic sensor) [14,15,23,29,30] and piezoelectric sensors [27,28,31,32].

A geophone [33] is deployed on the floor. It detects the velocity of movement of the floor and outputs a voltage signal. In contrast, the piezoelectric sensor measures changes in the pressure it bears. Although geophones can detect signals from three orthogonal axes in space, they are significantly larger than piezoelectric sensors in physical size. Furthermore, the price of a geophone [22] is 100 times higher than that of piezoelectric sensor [34]. In addition, the piezoelectric sensor has a simpler structure, higher sensitivity, wider frequency band, and larger dynamic range [35]. However, when it is used to detect floor vibration signals caused by pedestrian walking, there are issues with poor signal quality and low signal-to-noise ratio (SNR) [28]. The characteristics and physical parameters of

different piezoelectric sensors are not strictly consistent and usually present significant individual-to-individual differences. The measurement error between different sensors is varied, and the SNR is not uniform. Furthermore, piezoelectric can only detect the signals perpendicular to the floor. Previous research [23] showed that using a triaxial seismometer can achieve an increased accuracy in a localization task by introducing signal arrive angle to a TDOA (time-difference-of-arrival) system. However, piezoelectric sensors provide less information than triaxial geophones sensors. Thus it is more challenging to achieve good results only with the signals perpendicular to the floor.

To summarize, from a cost perspective the piezoelectric sensors are advantageous, especially in scenarios that require large-scale sensor deployment. However, there are more considerable challenges to get good results with the piezoelectric sensor-based approach. Our approach uses the piezoelectric sensors with a novel system design to overcome the limitations of this kind of sensors.

2.2. Vibration Signal-Based Approaches

Although the approaches based on the vibration signal present the advantage of protecting privacy compared with camera-based approaches [18,19,21,36], there are significant challenges to detect the number of pedestrians.

The vibration signal pedestrian identification-based approaches for pedestrian counting [23,25,26] require that the step event (SE) signals must not be mixed. This means that the system can only work when a maximum of one person is walking at the same time. In indoor multi-person scenarios, it is quite common for multiple people to move together at the same time. Such approaches cannot handle the use cases of normal daily life and only work in a well-defined environment such as a lab experiment. This dramatically limits their practical applications.

The room-level pedestrian counting approach [14] requires that there is a maximum of one person in the detection area. When the person leaves the detected area and enters a room, the counter will increase by one. This kind of approach can count the number of pedestrians when people go into and leave the detection area one by one. This works well in experimental scenarios, but is not suitable for practical use cases such as pedestrian counting in a large shopping mall.

The studies [15,16] support multiple people walking simultaneously in the same area where the signals between people can be mixed. Nevertheless, these studies require that there is only one group of pedestrians in the detected area, and this group of people should walk close together. In addition, the distance between each individual in the group should not be too far. In Pan et al.'s work [15], the signal of interest (SoI) is defined as the ambient vibration signal induced by occupant footsteps. In other words, a SoI is a piece of signal from the sensor when someone passes the sensor. The four features used in [15] are given in Table 1.

Table 1. Feature selection of previous work [15]. The features capture the information in vibration signals for footstep events from different perspectives.

	Features for Pedestrian Counting
(1)	Space-differential: Cross-correlation between SoIs from different sensors for the same footsteps.
(2)	Time-differential: Cross-correlation between SoIs for consecutive footsteps from the same sensor.
(3)	SoI duration.
(4)	Energy-specific: SoI signal entropy.

However, in real-world scenarios such as shopping malls, it is more likely that more than one group of pedestrians walk with different walking patterns in the same area. Because the method proposed in [15] deployed the sensors sparsely in a room, the problem

cannot be solved by deploying sensors with a different grid. Furthermore, when there is more than one group of people in the area, features (2) and (3) in Table 1 will not be available.

2.3. Overview of Our Approach

In this work, we propose a novel pedestrian counting approach based on footstep-induced structural vibration signals with piezoelectric sensors, which overcomes the limitations of previous approaches [16,23,26]. Specifically:

- Our approach can detect the number of pedestrians in an area while making no strict requirement about the number of groups of walking people in the detected area.
- Our approach supports the use cases where multiple people walk together with their signals mixed.
- Our approach uses the piezoelectric sensor, which is much cheaper than the geophone sensor used in previous approaches, making our approach economically viable.

Overall, our approach shows better performance than the existing work. Table 2 compares the capabilities of our approach with previous approaches.

Table 2. Capabilities of different approaches.

Approaches	Support Extreme Environment	Support More than One Person in the Detected Area	Support More than One Group of People	Device-Free	Privacy Protection	Resilient to Destruction
Camera-based [21,36]	-	✓	✓	✓	-	-
Device-based [2–4]	-	✓	✓	-	-	-
Li et al. [23]	✓	-	-	✓	✓	✓
Pan et al. [14,25,26]	✓	-	-	✓	✓	✓
Pan et al. [15,16]	✓	✓	-	✓	✓	✓
Our approach	✓	✓	✓	✓	✓	✓

3. Problem Formulation

In this section, we first define the problem, then discuss the possible solution based on the observations made by Pan et al. [15], which further motivates our solution to the problem.

3.1. Problem Definition

This paper focuses on counting the number of people based on floor vibration signals from piezoelectric sensors. Our system is designed to detect up to four pedestrians in an area of 3 m by 3 m, which can cover most indoor scenarios where multiple pedestrians walk in parallel with different stepping patterns, frequency, and directions. The piezoelectric sensors are deployed in the area of 3 m by 3 m as shown in Figure 1. The layout of the sensor deployment should guarantee that the signal from any vibration source in this area could be detected by any of the sensors. Previous studies showed that this particular layout works with good performance [27,28]. Overall, our system can detect the number of walking pedestrians who step in this area. The system supports the real-life scenarios where multiple persons are walking together at the same time and different people may walk in different directions. Our system is designed to handle the following stepping patterns [15]:

1. Footsteps from different pedestrians are fully synchronized in terms of striking timing.
2. Footsteps from different pedestrians are off-sync, but induced vibration signals presents temporal overlapping.
3. Footsteps are temporally staggered.

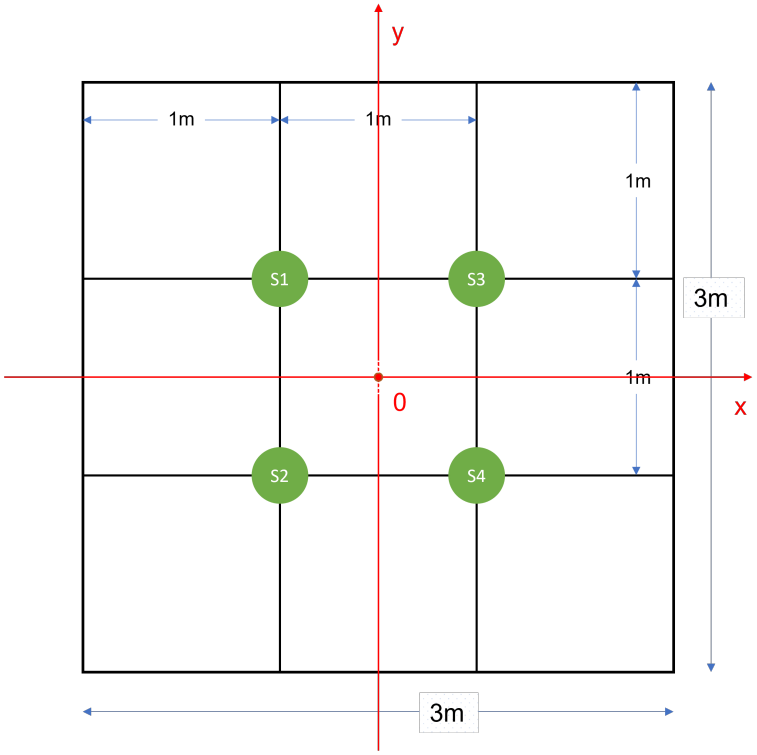


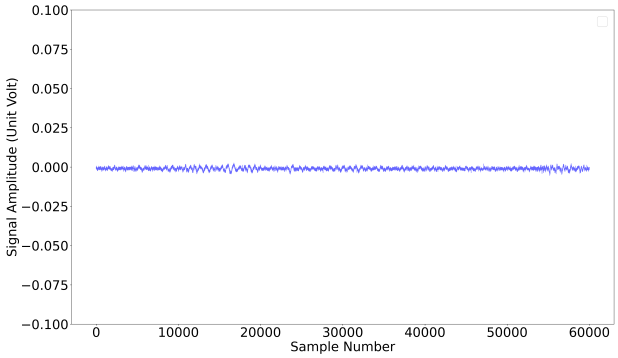
Figure 1. Data acquisition devices and experiment setup.

3.2. Problem Analysis

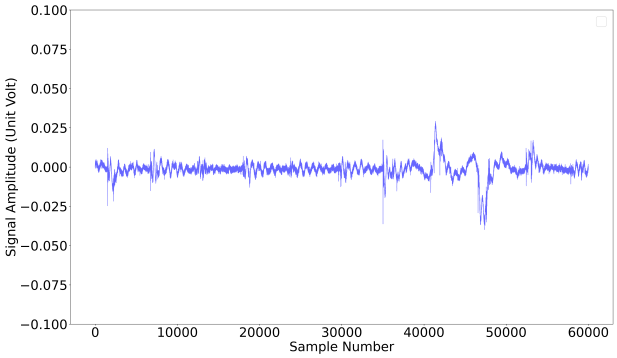
Figure 2 visualizes the original signals before denoising for the cases from 0 to 4 persons. These figures show the characteristics of time-specific signals captured by sensors. Intuitively, when the number of pedestrians is increased from 0 to 2, or from 3 to 4, the waveforms look different. However, it is not easy to differentiate whether the vibration signals are generated by 2 or 3 persons with only time domain information.

As discussed in Section 2.2, previous work [15] investigated the validation of the vibration features in Table 1 for counting the number of walking people. Figure 3 presents the results of the impulse load test experiment conducted in [15], where Figure 3a–d show the predictive capability of the vibration features in Table 1, respectively. Figure 3a shows the cross-correlation of the same SE from different sensors, representing the predictive capability of space-differential features. Figure 3b shows the cross-correlation of the same trace from the same sensor, representing the predictive capability of time-differential features. Figure 3c shows the step event duration. Figure 3d shows the step event signal entropy, representing the predictive capability of energy-specific features. Pan et al. [15] showed in Figure 3b,c that features (2) and (3) are only appropriate if the detection task is to distinguish whether the number of the pedestrians is 1 or more than 1, but they are uninformative to determine the exact number of pedestrians if there are 2 or more than 2 people. Similarly, feature (4) is only useful in the cases where there is more than one individual, making it difficult to distinguish the number of people when there are less than 2. Furthermore, the generation of features (2) and (3) required that there should be a maximum of one group of pedestrians in the detected area. Intuitively, when there are two groups of people, the signal from the first group of people may be mixed with the signal from the second group. The detected signal from the sensor is a mixture of both groups of people. As a result, it is challenging to differentiate whether the “SoI" is from the footsteps of the first or the second group of pedestrians. Similarly, the “SoI duration" is meaningless when the signals from two groups of pedestrians are mixed. Meanwhile, different groups of people may move at different speeds, where some groups may run while others walk at an average speed. These different moving events may occur simultaneously and the corresponding signals may be mixed up. On the other hand, when pedestrians

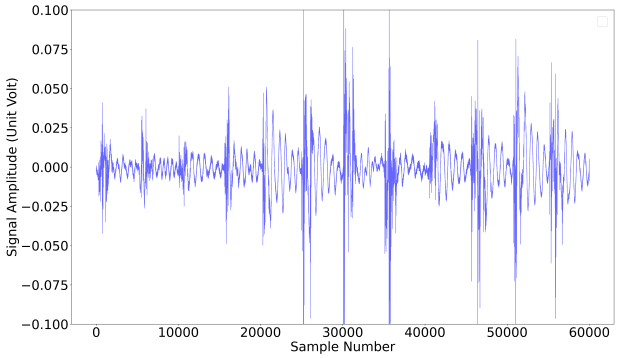
walk in the monitored area, the spatial difference of the floor vibration signal source from each individual is clear, which can be captured and detected with multiple sensors. This information is encoded in the space-differential feature (1) in Table 1. To summarize, only feature (1) can be effectively used to predict the number of walking people in a more practical scenario.



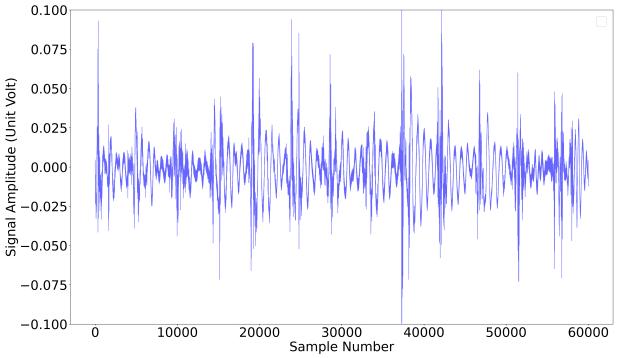
(**a**) 0 person case.



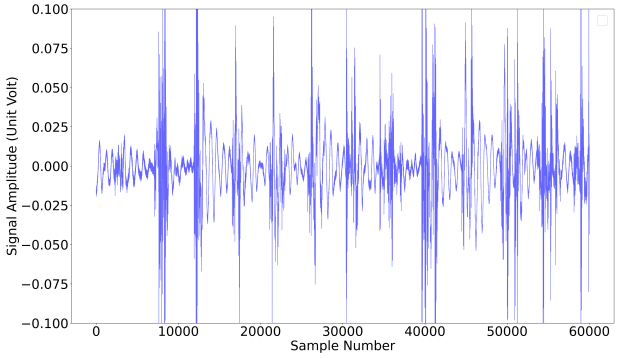
(**b**) 1 person in the detected area.



(**c**) 2 persons in the detected area.



(**d**) 3 persons in the detected area.



(**e**) 4 persons in the detected area.

Figure 2. (**a**–**e**) present the vibration signal in the time domain. The figures show samples of time-specific signal fragments from one of the sensors.

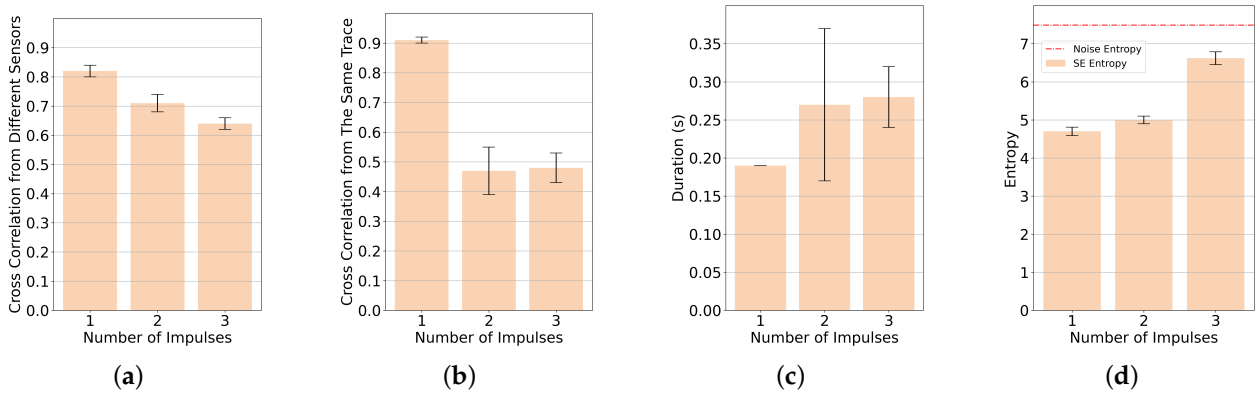


Figure 3. (**a**–**d**) are the results of the impulse load test experiment done in Pan et al.'s work [15]. This small ball hitting experiment treats ball hit impulse, an analogy to footstep event, as vibration source to study the predictive ability of each feature for pedestrian counting task. The x-axis represents the number of impulses. These figures only consider the 3 vibration source case, and can be interpreted in this way: the larger the difference among the height of the three bars, the more informative the corresponding feature. (**a**) Cross Correlation from Different Sensors. (**b**) Cross Correlation from The Same Trace. (**c**) Step Event Duration. (**d**) Step Event Signal Entropy.

Similar to cross correlation computing in [15], convolutional computation shares a similar symbolic calculation form. We assume that the convolutional computation of the same walking step event signal among different sensors will extract useful spatial features for pedestrian counting tasks. Our approach uses a deep neural network with convolutional layers to extract features from step event data and detect the number of pedestrians. The experimental results are presented in Section 5.

4. System Design

In this section, technical details about our approach are presented. The pedestrian counting architecture is shown in Figure 4. We regard the pedestrian counting task as a classification task. The vibration signal data with either none or up to four persons has been recorded and labeled. We used a deep neural network to extract features and perform the classification. Figure 5 shows the different steps our approach uses to determine the number of pedestrians.

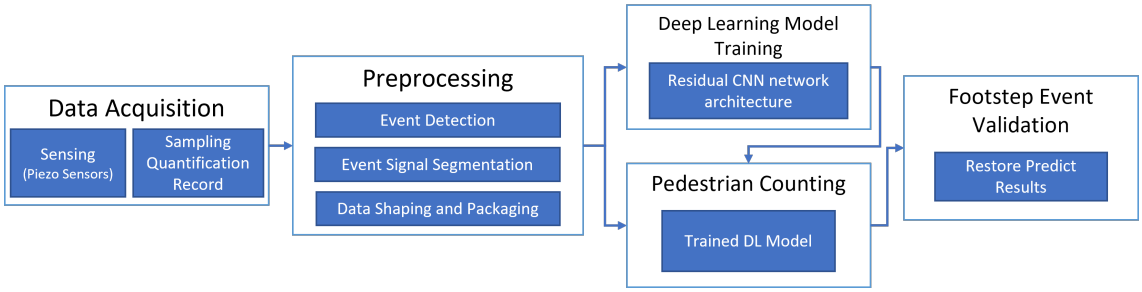


Figure 4. System architecture. The footstep event validation is as shown in Figure 5.

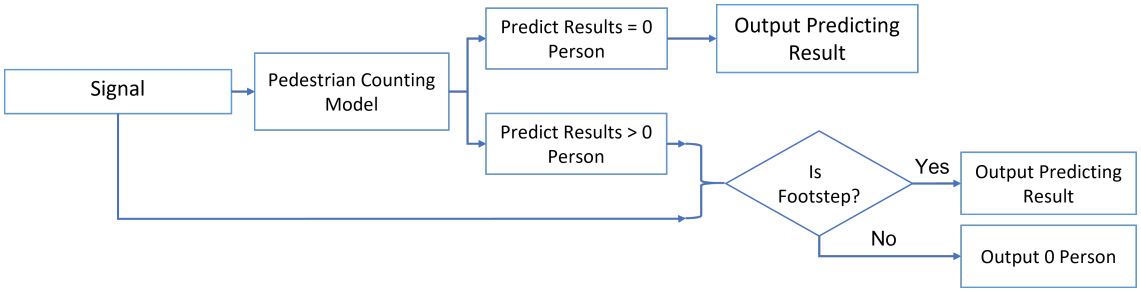


Figure 5. The logic about making prediction for the system. If the system predicts that there is more than one person in the area due to a vibration event detected, it will further check that the signal/vibration event is caused by footsteps. If it is, the system outputs the results. Otherwise, the system outputs 0 person as a result.

4.1. Data Acquisition

The previous works [27,28] showed that if four piezoelectric sensors (marked as S1, S2, S3, S4) are deployed spatially as shown in Figures 1 and 6, the vibration event in this 3 m by 3 m area can be detected.

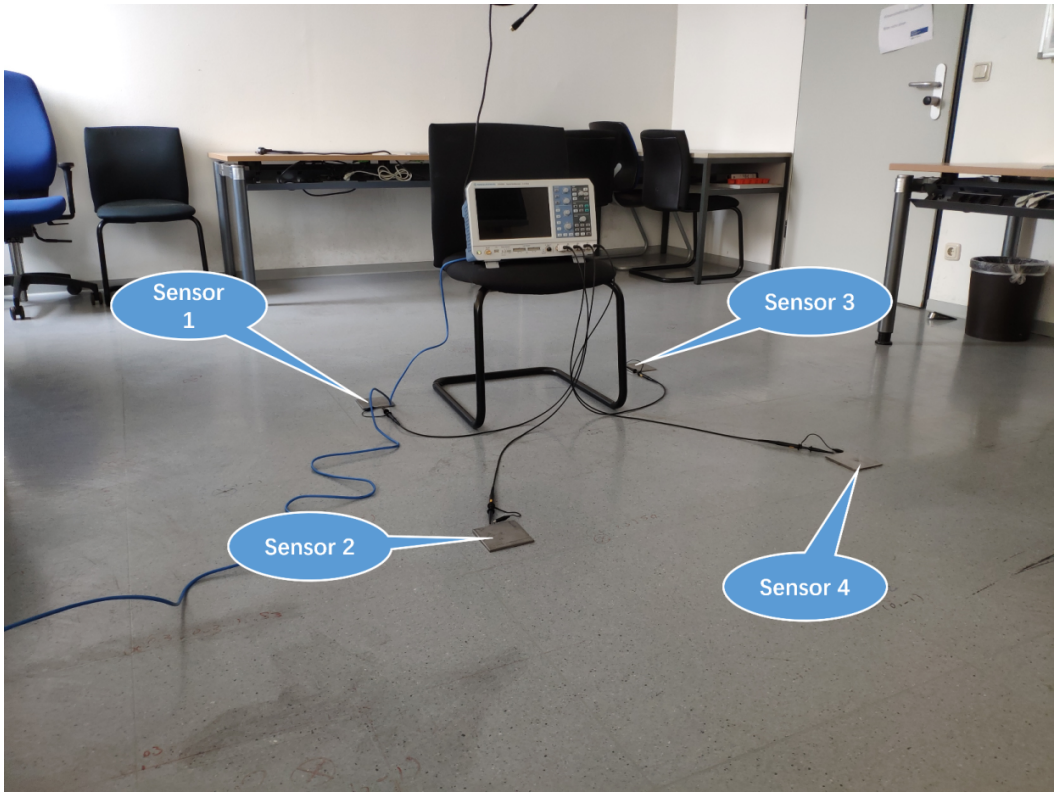


Figure 6. Data acquisition devices and experiment setup.

Although the piezoelectric sensors are not able to measure the signals statically over a long period of time, this feature will not have a real impact on the performance of the proposed approach. Intuitively, the proposed approach makes use of the space-differential and time-differential relationship features but not each sensor's absolute signal amplitude value, to detect the number of people. Similarly, the poor signal quality from the piezoelectric sensor will not affect the approach's feasibility.

Figure 6 shows the data acquisition devices and the experimental setup. Four EPZ-27MS44W piezoelectric sensors are deployed in the detected area to sense the vibration signals on the floor. The bandwidth of the sensor ranges from 0 Hz to 4400 Hz. The signal is amplified, sampled, quantize, and recorded with an R&S-RTB2000 oscilloscope. Time

synchronization of the signals is handled by the oscilloscope. The sampling rate is set to 10 kHz. The maximum value of the sampling point amplitude is 0.1 V. Thus, the waveforms higher than 0.1 V or lower than −0.1 V will be cut off.

4.2. Preprocessing

Traditional signal-based pattern recognition tasks need to filter the signals and denoise them during preprocessing and extract features manually. However, we do not filter or denoise the signals manually. Intuitively, the number of walking people is related to the signal's energy. Thus, filtering can lead to the loss of valuable information. Furthermore, we used an end-to-end deep learning-based approach in which feature extraction can be learned. Previous studies have shown that deep neural networks can tolerate modest amounts of noise in the training data [28,37]. Processing raw vibration data without denoising will not only avoid downgrading the accuracy of the deep neural network, but also increase its prediction performance, because valuable information in the data is preserved.

Overall, the preprocessing workflow of the system includes value normalization, event detection, event signal segmentation, and data shaping and packaging, as shown in Figure 4.

4.2.1. Normalization and Downsampling

Our approach is based on a deep neural network. The training procedure of deep neural network requires the values of training samples to range from −1 to +1. The ranges of the values of raw data are $[-0.1, 0.1]$. To make the data fitted into deep neural network, we divide the raw values by 0.1.

The signal data is downsampled to 2000 samples per second. The experiment shows that the informative signal is distributed in the frequency range of 0 to 1000 Hz. Thus, it is sufficient to provide data sampled at 2000 as the input of the neural network.

4.2.2. Signal Selection and Event Detection

A sliding window selects signal samples from the signal stream. The window size is 2048 samples. The sliding window shifts 64 samples each time. The data of the four sensors share the same sliding window.

We used the first-order second-moment method [38] to detect the beginning of a vibration event. By analyzing the change of Gaussianity, this method can be used to differentiate the vibration event from Gaussian noise. The first-order second-moment is defined in Equation (1), where N is the window length of the first-order second-moment method, μ is the mean of the values in this window, x_i is each value in the window. Empirically we determined a window size of 64 samples, which equals 32 milliseconds of sampling. We set the variance of ambient background noise as the threshold. When the m_2 values are larger than the threshold, the system detected a vibration event. If a vibration event is detected, the window will shift 2048 samples. If the number of data values in the shift window is less than 2048, all the data in the window will be dropped.

$$m_2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{1}$$

We used the signal of Sensor 1 to detect vibration events. The layout of sensor deployment in Figure 1 guarantees that any sensor can detect even the weakest signal generated in its area. Regarding the isolation of step events, the choice of the reference sensor does not make any difference regarding the detection of vibration events. When a vibration event is detected, the data of four sensors in the sliding window will be recorded.

Figure 7 shows an example of vibration event segmentation from the signal stream. The data values between each solid black line and red line will be extracted and packaged as input samples. The solid black lines denote the beginnings of an input sample, and the

red lines the corresponding ends. After each shift of the window, all the data in the shift window will be packaged as an input sample to the neural network for pedestrian counting.

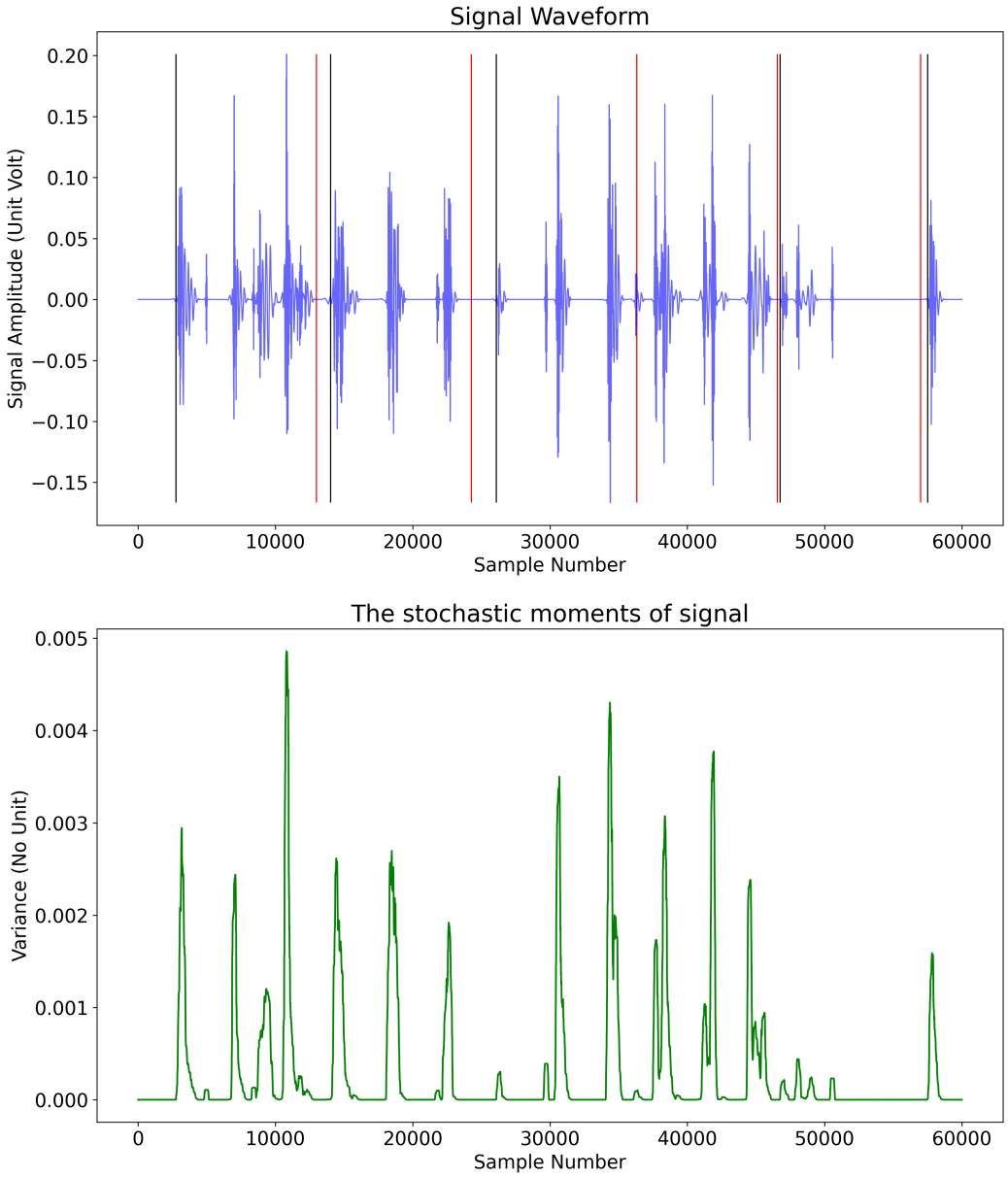


Figure 7. An example of signal selection and event detection from the signal stream. The black solid lines denote the beginning of each input sample and red lines the end.

4.3. Data Set Collection and Deep Learning Model

We used a deep learning model to extract features and detect the number of walking pedestrians. In Section 3, we discussed and analyzed the effectiveness of convolutional computing of data samples between different sensors for feature extraction. The extracted features are used as input for the deep learning model to predict the number of walking pedestrians. In this subsection, the data set collection and model architecture for training are presented.

4.3.1. Data Collection

We generated a vibration signal data set ranging from 0 to 4 persons in the experiment. Two males and two females participated in the data collection process. When there is no one in the area, the data acquired by our devices are labeled as “0 Person" or “P0". In each

turn of the experiment, there is a known and fixed number of participants in the monitored area. The number of participants is marked as the label of each data sample, which is further used as class label for deep learning model. The data set includes cases that cover most practical scenarios, in which the participants may walk, walk fast, and run. After preprocessing, we collected a total of 20,955 data samples. The statistics about each class is presented in Table 3.

Table 3. Statistics about the dataset.

	P0	P1	P2	P3	P4
#Samples	1954	3440	4752	5181	5628

4.3.2. Deep Learning Model

As shown in Figure 8, our deep learning model has 13 one-dimensional convolutional layers with residual structure [39–42]. The dropout rate is 0.3. The learning rate is 0.001. The batch size is set to 32. The input size is 2048 rows and 4 columns.

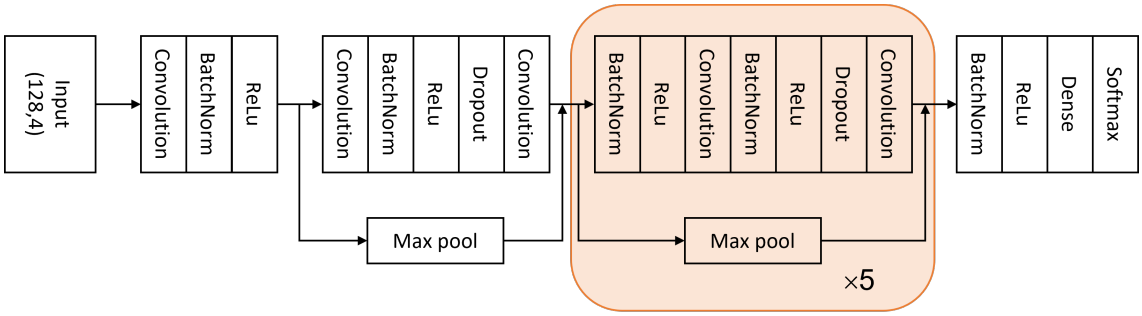


Figure 8. Architecture of deep learning network.

4.4. Prediction Output Judgment Logic

Sometimes a non-footstep vibration event [23] may trigger the model to output a result which makes no sense. Once the prediction result of deep learning model is not "0 Person", our system further validates whether the vibration event is a footstep event or not. Existing research regarding footstep detection [14,43] presents methods to judge whether a vibration event is caused by a footstep or some other events, for example a door closing or a bus driving outside the building. The entire detection logic including the mentioned footstep detection is shown in Figure 5.

If the input signal is a superposition of step vibrations and non-footsteps vibration, and if the non-footsteps signal is not too strong or does not persist too long, the system will treat the non-footsteps signal as noise. If the non-footsteps signal is very strong or persists for a long time, the risk will increase a lot that the footstep detecting module will block the proposed system. The current approach can only guarantee that the system will work in the most common scenarios, such as in an office building where the interference is not that strong or persist for a long time.

5. Evaluation

In this section, we present the performance of our system for the prediction task. We conducted a 5-fold cross-validation (CV) [44,45]. For our results we computed the average of all cross-validation folds.

5.1. Data Preparation for K-Fold Cross-Validation

The data is divided evenly and randomly into five folds exclusively as shown in Figure 9. For each fold, we train the deep learning model with the training set and evaluate the performance of the model with the test set. We repeat this training and evaluation

process five times for a 5-fold cross-validation. The deep neural network classifies the number of pedestrians according to the input samples.

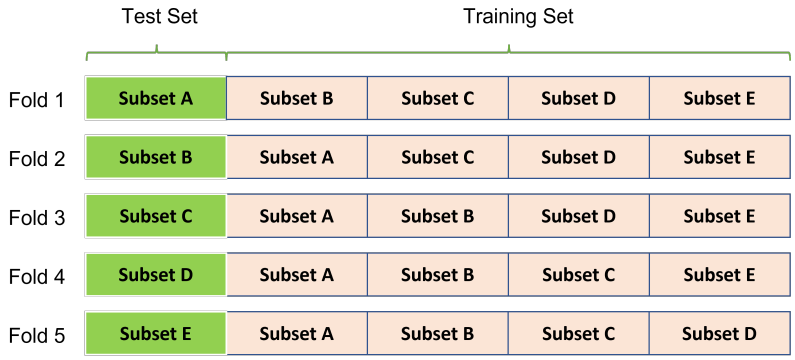


Figure 9. Diagram of K-fold cross-validation with K=5. We split the whole dataset into 5 non-overlapping subsets. For the *i*-th fold, the *i*-th subset is used as test set and the remaining subsets as training set.

5.2. Performance

Precision, recall, and F1-score are used as performance metrics. Intuitively, the precision quantifies the ability of the classifier not to label a sample as positive that is actually negative. The recall represents the ability of the classifier to retrieve all positive samples. The F1-score can be interpreted as a weighted harmonic mean of the precision and recall, which is useful to balance the trade-off between the two quantities and tends to give more weight to lower values [46]. An F1-score reaches its best value at 1 and the worst score at 0. The calculation of macro and micro average can refer to [28]. The larger the metric values, the better the classification of the system.

The averaged classification performance over the 5-fold cross-validation is presented in Table 4. The confusion matrix in Figure 10 is calculated and normalized according to the prediction for each test sample in the 5-fold cross-validation experiment. We observe from Table 4 that: (i) the averaged precision, recall, and F1-score for each of the 5 classes are over 0.95; (ii) the averaged macro and micro for the three metrics are over 0.98 for the 5-class classification task; (iii) except for the standard deviation of the precision for the 0 Person class, the standard deviation for all performance metrics are relatively low (less than 0.1). These observations suggest that our classifier presents outstanding performance for the prediction task. Moreover, the high values of macro and micro F1-score (over 0.98) indicate that the classifier shows excellent performance on all classes over the entire data set. Meanwhile, it can be observed from Figure 10 that most of the off-diagonal values in the confusion matrix are close to 0 and all the diagonal values are larger than 0.96, suggesting that our approach can predict the number of walking pedestrians with high accuracy. On the other hand, Pan et al. [15,16] performed a similar classification task, but only achieved an averaged accuracy of 0.6875 for a 4-class classification task. The averaged accuracy is obtained by calculating the arithmetic mean of the accuracy for the four classes from Table 1 in [15], i.e., $(0.8333 + 0.6667 + 0.3333 + 0.9167)/4 = 0.6875$. This suggests that our approach is significantly better than Pan et al.'s method [15,16].

Table 4. Classification performance of the DNN. The average and standard deviation (stdev) for each metric are calculated using results from the 5-fold cross-validation.

	Precision	Recall	F1-Score
0 Person	0.9508 ± 0.1042	0.9990 ± 0.0014	0.9717 ± 0.0589
1 Person	0.9988 ± 0.0019	0.9632 ± 0.0806	0.9793 ± 0.0440
2 Persons	0.9966 ± 0.0060	0.9977 ± 0.0018	0.9971 ± 0.0038
3 Persons	0.9900 ± 0.0178	0.9685 ± 0.0476	0.9785 ± 0.0247
4 Persons	0.9732 ± 0.0393	0.9898 ± 0.0175	0.9810 ± 0.0206
Accuracy			0.9827 ± 0.0253
Micro Average	0.9819 ± 0.0293	0.9836 ± 0.0256	0.9815 ± 0.0301
Macro Average	0.9847 ± 0.0212	0.9827 ± 0.0253	0.9828 ± 0.0252

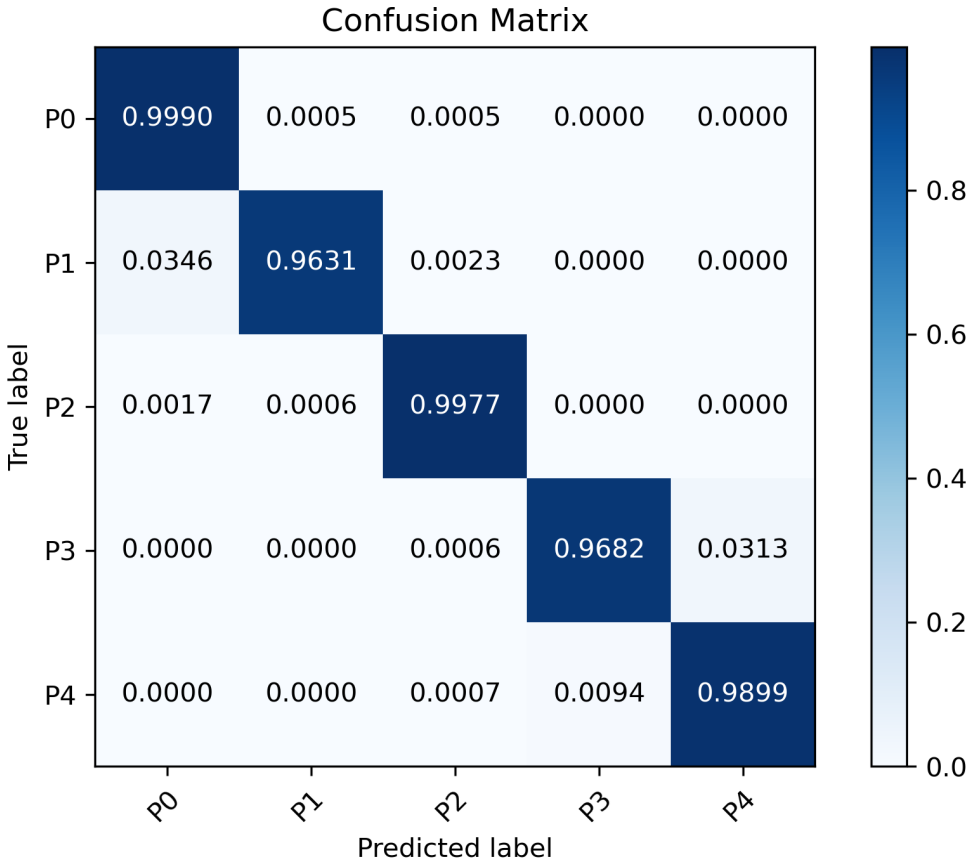


Figure 10. Confusion matrix normalized over the 5-fold cross-validation evaluation results.

6. Conclusions and Future Work

In this paper we presented a novel device-free walking pedestrian counting approach based on piezoelectric sensors. Our approach can protect the privacy of the pedestrians, because only vibration signals are acquired. The sensors used in our work are much cheaper than the geophone sensors used in previous studies, making our approach more economically viable. Furthermore, our approach does not require a high-density sensor deployment. This means that our system can be easily expanded to cover large areas. Our approach supports that multiple people are walking at the same time with the signals mixed together. Unlike previous approaches [15,16], it makes no strict requirement about the number of groups of walking people in the detection area. Our approach can detect the

number of walking people (up to a maximum of four persons within a 3 m by 3 m area) with an averaged F1-score of over 0.98.

In the future, we will integrate all the vibration signal-based functional modules [27,28] into one system. As a whole, the vibration-based system, together with the audio and video-based system, will serve as a perception layer for a privacy-protecting smart city.

Author Contributions: Conceptualization, Y.Y. and T.W.; methodology, Y.Y.; software, Y.Y.; validation, Y.Y. and T.W.; formal analysis, Y.Y.; investigation, Y.Y.; resources, Y.Y.; data curation, Y.Y.; writing—original draft preparation, Y.Y. and T.W.; writing—review and editing, Y.Y., X.Q., S.H., W.H., T.W.; visualization, Y.Y.; supervision, T.W. and W.H.; project administration, T.W.; funding acquisition, T.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Evonik Digital. The APC was funded by the Open Access Publication Fund of the University of Duisburg-Essen.

Acknowledgments: We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen. We acknowledge and thank Evonik Digital for this research work's financial support. We thank Hao Ma for his help in the experiment and data collection.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hussain, S.; Yu, Y.; Ayoub, M.; Khan, A.; Rehman, R.; Wahid, J.A.; Hou, W. IoT and Deep Learning Based Approach for Rapid Screening and Face Mask Detection for Infection Spread Control of COVID-19. *Appl. Sci.* **2021**, *11*, 3495. [CrossRef]
2. Fierro, G.; Rehmane, O.; Krioukov, A.; Culler, D. Zone-Level Occupancy Counting with Existing Infrastructure. In *BuildSys '12: Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, Toronto, ON, Canada, 6 November 2012*; Association for Computing Machinery: New York, NY, USA, 2012; pp. 205–206. [CrossRef]
3. Corna, A.; Fontana, L.; Nacci, A.A.; Sciuto, D. Occupancy Detection via iBeacon on Android Devices for Smart Building Management. In Proceedings of the 2015 Design, Automation Test in Europe Conference Exhibition (DATE), Grenoble, France, 9–13 March 2015; pp. 629–632. [CrossRef]
4. Conte, G.; De Marchi, M.; Nacci, A.A.; Rana, V.; Sciuto, D. BlueSentinel: A First Approach Using iBeacon for an Energy Efficient Occupancy Detection System. In Proceedings of the BuildSys @ SenSys, Memphis, TN, USA, 3–6 November 2014; pp. 11–19.
5. Krumm, J.; Harris, S.; Meyers, B.; Brumitt, B.; Hale, M.; Shafer, S. Multi-Camera Multi-Person Tracking for EasyLiving. In Proceedings of the Third IEEE International Workshop on Visual Surveillance, Dublin, Ireland, 1 July 2000; pp. 3–10. [CrossRef]
6. Hnat, T.W.; Griffiths, E.; Dawson, R.; Whitehouse, K. Doorjamb: Unobtrusive Room-Level Tracking of People in Homes Using Doorway Sensors. In *SenSys '12: Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems, Toronto, ON, Canada, 6–9 November 2012*; ACM Press: Toronto, ON, Canada, 2012; p. 309. [CrossRef]
7. Narayana, S.; Prasad, R.V.; Rao, V.S.; Prabhakar, T.V.; Kowshik, S.S.; Iyer, M.S. PIR Sensors: Characterization and Novel Localization Technique. In *IPSN '15: Proceedings of the 14th International Conference on Information Processing in Sensor Networks, Seattle, WA, USA, 13–16 April 2015*; ACM: Seattle, DC, USA, 2015; pp. 142–153. [CrossRef]
8. Song, J.; Dong, Y.F.; Yang, X.W.; Gu, J.H.; Fan, P.P. Infrared Passenger Flow Collection System Based on RBF Neural Net. In Proceedings of the 2008 International Conference on Machine Learning and Cybernetics, Kunming, China, 12–15 July 2008; Volume 3, pp. 1277–1281. [CrossRef]
9. Xia, L.; Chen, C.C.; Aggarwal, J.K. Human Detection Using Depth Information by Kinect. In Proceedings of the CVPR 2011 WORKSHOPS, Colorado Springs, USA, 21–23 June 2011; pp. 15–22. [CrossRef]
10. Xu, C.; Firner, B.; Moore, R.S.; Zhang, Y.; Trappe, W.; Howard, R.; Zhang, F.; An, N. SCPL: Indoor Device-Free Multi-Subject Counting and Localization Using Radio Signal Strength. In *IPSN '13: Proceedings of the 12th International Conference on Information Processing in Sensor Networks, , Philadelphia, PA, USA, 8–11 April 2013*; Association for Computing Machinery: New York, NY, USA, 2013; pp. 79–90. [CrossRef]
11. Xu, C.; Firner, B.; Zhang, Y.; Howard, R.; Li, J.; Lin, X. Improving RF-Based Device-Free Passive Localization in Cluttered Indoor Environments through Probabilistic Classification Methods. In Proceedings of the 2012 ACM/IEEE 11th International Conference on Information Processing in Sensor Networks (IPSN), Beijing, China, 16–19 April 2012; pp. 209–220. [CrossRef]
12. Zhang, D.; Liu, Y.; Guo, X.; Ni, L.M. RASS: A Real-Time, Accurate, and Scalable System for Tracking Transceiver-Free Objects. *IEEE Trans. Parallel Distrib. Syst.* **2013**, *24*, 996–1008. [CrossRef]
13. Zhang, P.; Martonosi, M. LOCALE: Collaborative Localization Estimation for Sparse Mobile Sensor Networks. In Proceedings of the 2008 International Conference on Information Processing in Sensor Networks (Ipsn 2008), St. Louis, MI, USA, 22–24 April 2008; pp. 195–206. [CrossRef]
14. Pan, S.; Bonde, A.; Jing, J.; Zhang, L.; Zhang, P.; Noh, H.Y. BOES: Building Occupancy Estimation System Using Sparse Ambient Vibration Monitoring. In Proceedings of the SPIE Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring, San Diego, California, USA, 9–13 March 2014; Lynch, J.P., Wang, K.W., Sohn, H., Eds.; p. 90611O. [CrossRef]

15. Pan, S.; Mirshekari, M.; Zhang, P.; Noh, H.Y. Occupant Traffic Estimation through Structural Vibration Sensing. In *SPIE Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring*; Lynch, J.P., Ed.; 2016; p. 980306. [CrossRef]
16. Pan, S.; Mirshekari, M.; Fagert, J.; Ruiz, C.; Noh, H.Y.; Zhang, P. Area Occupancy Counting Through Sparse Structural Vibration Sensing. *IEEE Pervasive Comput.* **2019**, *18*, 28–37. [CrossRef]
17. Hafeezallah, A.; Al-Dhamari, A.; Abu-Bakar, S.A.R. U-ASD Net: Supervised Crowd Counting Based on Semantic Segmentation and Adaptive Scenario Discovery. *IEEE Access* **2021**, *9*, 127444–127459. [CrossRef]
18. Liu, L.; Jiang, J.; Jia, W.; Amirgholipour, S.; Wang, Y.; Zeibots, M.; He, X. DENet: A Universal Network for Counting Crowd With Varying Densities and Scales. *IEEE Trans. Multimed.* **2021**, *23*, 1060–1068. [CrossRef]
19. Hafeezallah, A.; Abu-Bakar, S. Crowd Counting Using Statistical Features Based on Curvelet Frame Change Detection. *Multimed. Tools Appl.* **2017**, *76*, 15777–15799. [CrossRef]
20. Wang, Z.; Li, W.; Shen, Y.; Cai, B. 4-D SLAM: An Efficient Dynamic Bayes Network-Based Approach for Dynamic Scene Understanding. *IEEE Access* **2020**, *8*, 219996–220014. [CrossRef]
21. Zhang, C.; Li, H.; Wang, X.; Yang, X. Cross-Scene Crowd Counting via Deep Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 833–841.
22. Geophone—SM-24—SEN-11744—SparkFun Electronics. Available online: https://www.sparkfun.com/products/11744 (accessed on 20 January 2022).
23. Li, F.; Clemente, J.; Valero, M.; Tse, Z.; Li, S.; Song, W. Smart Home Monitoring System via Footstep-Induced Vibrations. *IEEE Syst. J.* **2020**, *14*, 3383–3389. [CrossRef]
24. Al-Naimi, I.; Wong, C.B. Indoor Human Detection and Tracking Using Advanced Smart Floor. In Proceedings of the 2017 8th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 4–6 April 2017; pp. 34–39. [CrossRef]
25. Pan, S.; Yu, T.; Mirshekari, M.; Fagert, J.; Bonde, A.; Mengshoel, O.J.; Noh, H.Y.; Zhang, P. FootprintID: Indoor Pedestrian Identification through Ambient Structural Vibration Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2017**, *1*, 89:1–89:31. [CrossRef]
26. Pan, S.; Wang, N.; Qian, Y.; Velibeyoglu, I.; Noh, H.Y.; Zhang, P. Indoor Person Identification through Footstep Induced Structural Vibration. In *HotMobile '15: 16th International Workshop on Mobile Computing Systems and Applications, Santa Fe, NM, USA, 12–13 February 2015*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 81–86. [CrossRef]
27. Yu, Y.; Weis, T. A Privacy-Protecting Indoor Emergency Monitoring System Based on Floor Vibration. In *UbiComp-ISWC '20: Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers , Virtual Event, Mexico, 12–17 September 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 164–167. [CrossRef]
28. Yu, Y.; Waltereit, M.; Matkovic, V.; Hou, W.; Weis, T. Deep Learning-Based Vibration Signal Personnel Positioning System. *IEEE Access* **2020**, *8*, 226108–226118. [CrossRef]
29. Clemente, J.; Valero, M.; Li, F.; Wang, C.; Song, W. Helena: Real-Time Contact-Free Monitoring of Sleep Activities and Events around the Bed. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom), Austin, TX, USA, 23–27 March 2020; pp. 1–10. [CrossRef]
30. Valero, M.; Clemente, J.; Li, F.; Song, W. Health and Sleep Nursing Assistant for Real-Time, Contactless, and Non-Invasive Monitoring. *Pervasive Mob. Comput.* **2021**, *75*, 101422. [CrossRef]
31. Kashimoto, Y.; Fujimoto, M.; Suwa, H.; Arakawa, Y.; Yasumoto, K. Floor Vibration Type Estimation with Piezo Sensor toward Indoor Positioning System. In Proceedings of the 2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Alcala de Henares, Spain, 4–7 October 2016; pp. 1–6.
32. Akiyama, S.; Yoshida, M.; Moriyama, Y.; Suwa, H.; Yasumoto, K. Estimation of Walking Direction with Vibration Sensor Based on Piezoelectric Device. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 23–27 March 2020; pp. 1–6. [CrossRef]
33. Krohn, C.E. Geophone Ground Coupling. *GEOPHYSICS* **1984**, *49*, 722–731. [CrossRef]
34. Team (webmaster@reichelt.de), r.e.G..C.K.I. EPZ-27MS44W—Piezoelement, 4,4 kHz, 200 Ohm, Bedrahtet. Available online: https://www.reichelt.de/piezoelement-4-4-khz-200-ohm-bedrahtet-epz-27ms44w-p145918.html (accessed on 20 January 2022).
35. Hou, T.; Liu, H.; Zhu, J.; Liu, T.; Liu, L.; Li, Y.; Qian, C.; Xin, Y. Piezoelectric Geophone: A Review from Principle to Performance. *Ferroelectrics* **2020**, *558*, 27–35. [CrossRef]
36. Liu, N.; Long, Y.; Zou, C.; Niu, Q.; Pan, L.; Wu, H. ADCrowdNet: An Attention-Injective Deformable Convolutional Network for Crowd Understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3225–3234.
37. Hannun, A.Y.; Rajpurkar, P.; Haghpanahi, M.; Tison, G.H.; Bourn, C.; Turakhia, M.P.; Ng, A.Y. Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. *Nat. Med.* **2019**, *25*, 65–69. [CrossRef] [PubMed]
38. Clemente, J.; Li, F.; Valero, M.; Song, W. Smart Seismic Sensing for Indoor Fall Detection, Location, and Notification. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 524–532. [CrossRef] [PubMed]
39. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]

40. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
42. Bouvrie, J. Notes on Convolutional Neural Networks. 2006. Available online: http://cogprints.org/5869/ (accessed on 20 January 2022).
43. Mirshekari, M.; Fagert, J.; Pan, S.; Zhang, P.; Noh, H.Y. Step-Level Occupant Detection across Different Structures through Footstep-Induced Floor Vibration Using Model Transfer. *J. Eng. Mech.* **2020**, *146*, 04019137. [CrossRef]
44. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 112.
45. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Prentice Hall: Hoboken , NJ, USA, 2002.
46. Grandini, M.; Bagli, E.; Visani, G. Metrics for Multi-Class Classification: An Overview. *arXiv* **2020**, arXiv:abs/2008.05756.

12.3 A Privacy-Protecting Step-Level Walking Direction Detection Algorithm based on Floor Vibration

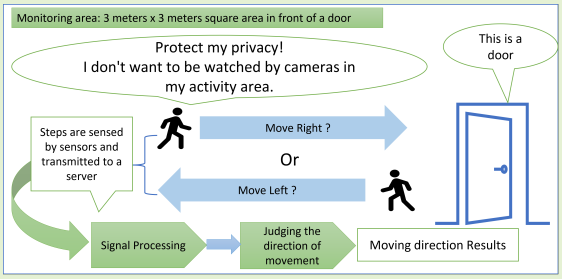
Title	A Privacy-Protecting Step-Level Walking Direction Detection Algorithm based on Floor Vibration
Authors	Yang Yu, Oskar Carl, Shabir Hussain, Weiyan Hou, Torben Weis
Publication Venue	IEEE Sensors Journal
Publication Type	Article
Publication Status	Accepted
Research Topic	Walking Direction Detection

Table 12.3: Bibliographic information of publication C.

A Privacy-Protecting Step-Level Walking Direction Detection Algorithm based on Floor Vibration

Yang Yu, Oskar Carl, Shabir Hussain, Weiyan Hou, and Torben Weis

***Abstract*— We present an algorithm and measurement system to detect the walking direction of persons based on ground vibrations. The approach is privacy-preserving because it solely relies on piezoelectric sensors built into the floor. Therefore, our system can be used in areas where cameras are not allowed or cannot capture the entire area. We present and compare our two innovative methods to analyze the ground vibrations caused by footsteps: the multi-peaks average algorithm (MPAA) and the multi-peaks averaged feature with a deep neural network-based classifier (MPAF-DNNC). MPAA judges the walking direction of pedestrians by analyzing the time-space relationship of at least two consecutive footstep vibration signals from multiple sensors. MPAF-DNNC receives multi-peaks averaged feature as input and uses a deep neural network-based classifier to judge walking direction. Our experiments and evaluation show that our system can correctly determine the walking direction based on only 3 input step events and provides an average F1 score of 0.97. When more than 5 step events are inputted, the proposed system can correctly determine the walking direction with an average F1 score of 1.00.**



***Index Terms*— Walking direction detection, vibration signal, piezoelectric sensor, privacy protection, algorithm, pattern recognition, data fusion.**

I. Introduction

In recent years, industries and consumers put forward requirements for intelligent monitoring and situational awareness of the environment [1], [2]. Specifically, operators of malls or chemical plants want to monitor human traffic to know where people are headed in case of an emergency. Emergencies like hostage-taking in malls or explosions in chemical plants demonstrate the need to send security staff to the appropriate areas quickly. Ideally, a fully automatic system can guide security and rescue staff to reduce the response time.

Pedestrian moving direction detecting is an essential technology for a situational perception system. Pedestrian walking direction detecting technology can be used to observe the human traffic in a building when working together with the pedestrian number counting technology (as shown in paper [3]–[6]). Suppose we record the number of the pedestrians who pass by each entrance of a building and the moving direction of the pedestrians(enter the entrance or exit the entrance). In that case, we can calculate the traffic of the pedestrians. Not only that, detecting a single pedestrian's location and the moving direction can support many services, such as monitoring a person's activities, predicting the tendencies of the residents, and supporting the intelligent control of appliances [7].

Deploying sensors is required for all the implementation. However, the deployment of sensors in public areas must consider the protection of the privacy of people in the monitoring area. It must not violate the laws of local regions on privacy protection. The EU Geral Data Protection Regulation (GDPR) [8] as well as the OECD privacy framework [9] consider the information which can be used to identify a natural person as "personal information". Personal information can not be recorded or used without the authorization of its owner. Meanwhile, the GDPR also classifies personal biometric information and personal location information as the category of privacy that should be protected. Therefore, a privacy-preserving approach to pedestrian movement direction detection is needed.

This paper contributes a privacy-protecting approach for pedestrians walking direction detection task based on vibration sensors. The system based on ground vibrations can trace the

We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen. We acknowledge and thank Evonik Digital for this research work's financial support.

Yang Yu is with the University of Duisburg-Essen, Duisburg, Germany and Zhengzhou University, Zhengzhou, China(e-mail: yang.yu@uni-due.de).

Oskar Carl and Torben Weis are with the University of Duisburg-Essen, Duisburg, Germany (e-mail: {oskar.carl, torben.weis}@uni-due.de).

Shabir Hussain is with the Zhengzhou University, Zhengzhou, China (e-mail: shabir@gs.zzu.edu.cn).

Corresponding author: Weiyan Hou is with the Zhengzhou University, Zhengzhou, China (e-mail: houwy@zzu.edu.cn).

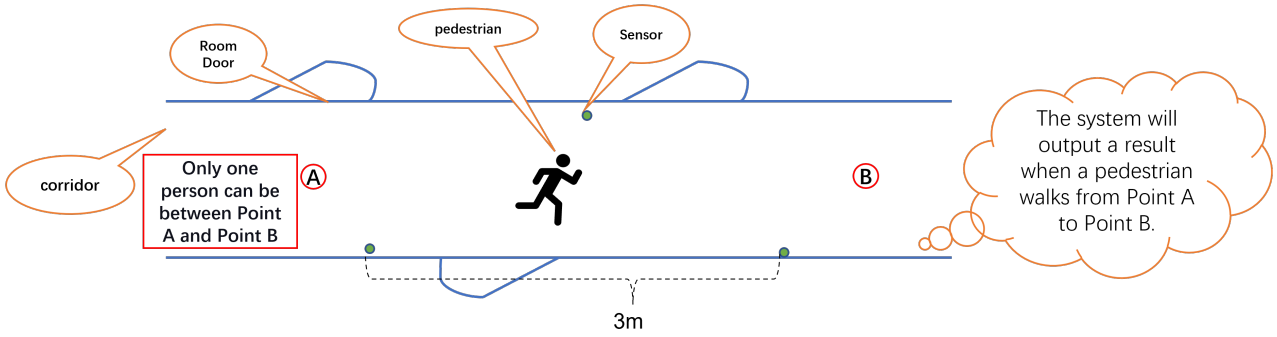


Fig. 1. Schematic diagram of the existing technical solution [10]–[12].

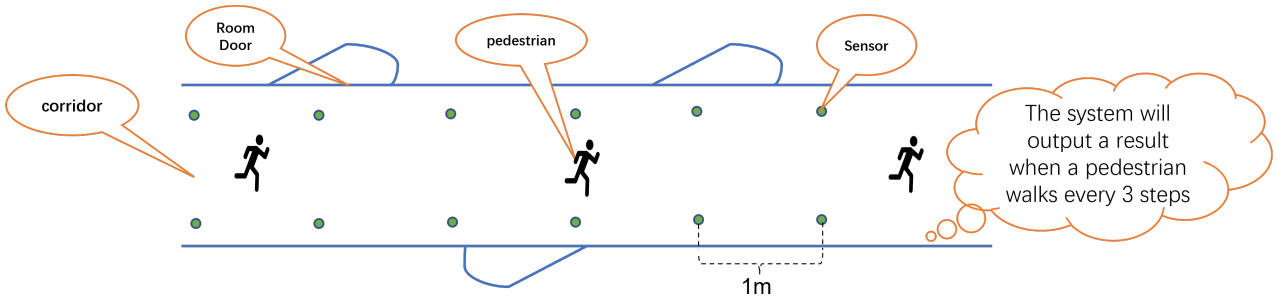


Fig. 2. Schematic diagram of the our proposed approach.

whereabouts of walking people by detecting the vibrations caused by footsteps. Our system uses a sensor matrix consisting of 4 sensors to collect vibration signal data. The algorithm presented in this paper relies on the data of multiple sensors in the sensor matrix as the data source. It uses data fusion technology based on multiple data sources to improve the signal's measurement accuracy and information quality. The proposed data fusion method reduces the required minimum deployment distance between sensors, improving the system's usability and feasibility in practical application scenarios. However, the sensors are less privacy-invasive because neither video nor audio is being captured. The proposed approach can be scaled out to support more areas Our approach does not require people to carry any wearable devices as a passive detection method. The proposed approach can tolerate the interference of smoke. In detail, the main innovation and contribution of this paper are as follows.

- Our research is a privacy protection approach because it uses sensors to measure ground vibrations to compute the direction of persons walking. Without the help of other technologies and data, it is impossible to intuitively identify the pedestrian's identity only based on the ground vibration signal caused by the pedestrian's walking.
- Our approach is step level. The prediction procedure only depends on the step events to make a walking direction judgment. Thus, it has no constraints on pedestrians' relative walking position to the sensors.
- Our system uses data fusion technology based on multiple data sources, which overcomes the disadvantages brought by the anisotropy of the vibration signal propagation medium. The approach reduces the deployment distance between sensor units, thus making it possible to deploy multi-sensor arrays in adjacent areas. Therefore, it can support the detection of the moving direction of multiple people in an adjacent area, improving the system's feasibility.
- Our approach uses cheap sensors and data fusion technology to achieve performance that rivals solutions using expensive sensors.
- We compared the proposed two innovative methods, Multi-Peaks Average Algorithm (MPAA) and multi-peaks averaged feature with the deep neural network-based classifier (MPAF-DNNC). We reduce the required SEs to 3 with an F1 score of 0.97.

II. Related Work

In recent years, machine learning has achieved impressive results in various fields [13]–[15], especially useful in situational awareness tasks in intelligent environments [1], [16]. Additionally, the application of data fusion technology [17], [18] allows the environmental monitoring tasks to be better solved [19]. The above studies demonstrate that the combination of artificial intelligence and data fusion technology promotes effective information extraction, providing more room for improvement in the performance of data-based perception tasks.

On the other hand, the problem of walking direction detection is not sufficiently solved by existing research for deploying it in our use cases.

Camera-based methods [20], [21] do not protect people's privacy because the use of cameras often means more than

necessary information is collected which can be used to identify a nature person.

Wu et al. [22] propose an approach based on Channel State Information (CSI) data from WIFI devices. This approach infers walking direction by analyzing the difference in the phase change of different sub-carriers of orthogonal frequency-division multiplexing (OFDM) channels between receivers and transmitters. However, considering the phase difference of two waveforms changes periodically between 0 to 2π, and if the phase difference is bigger than π, we cannot know which waveform comes first. Thus, their approach limits the maximum allowable phase delay of the reflected waveform to $\pi/2$ and does not work in a big room or an open scene with complex obstacles. Meanwhile, limited by the number of OFDM subcarriers of the WIFI card, this approach can not scale-out without limitations.

There are approaches based on floor vibration with geophones [11], [12]. As shown in Fig. 1, in these approaches, sensors are deployed sparsely in a room, and they are limited to at most one person in the room. This limits these approaches can only work in limited scenarios or experimental settings. However, multiple people may walk in the same room in typical scenarios simultaneously. Thus this approach shows low feasibility.

There is work based on floor vibration with piezoelectric sensors [10]. However, the approach presented is not precise on an individual step level. Similarly, as shown in Fig. 1, it can only work when the pedestrian walks from one end of the line segment formed by the connection of the two sensors to the other end. Their approach limits the relative position of the pedestrian to the sensors. The versatility and applicability in actual use scenarios are thus limited.

The existing piezoelectric sensor-based approaches [10] or geophone-based (seismic sensor) approaches [11], [12] do not support the step-level resolution, or their prerequisites are challenging to meet in real-life scenarios because it has constraints on pedestrians' relative walking position to the sensors. The above vibration sensor-based approaches require that a person walks in the same direction for several meters, because it does not analyze individual footsteps. Additionally, the price of a geophone is more than 100 times that of a piezoelectric sensor.

All the above mentioned vibration sensor-based approaches deploy different sensors with a far distance between each other. This is technically simpler to implement [10]–[12]. However, in doing so, there are many restrictions on usage scenarios. For example, the sensor monitoring area can only have one person. When multiple people appear, the results will be meaningless. Undoubtedly this limits the feasibility of the method in practical use.

In comparison, the sensors are deployed closer together in our method, as shown in Fig. 2. The deployment distance between sensors is 1 meter. In this way, multi-sensor units can be deployed in an area, supporting practical scenarios of multiple people walking simultaneously.

However, the more significant challenge arises for our detection purposes when the sensors are placed closer together. Our previous research showed that vibration signals propagate anisotropically on the ground [16]. The closer the different sensors are deployed, the more difficult it is to identify a signaling event. This challenge is even more formidable when inexpensive piezoelectric sensors are used, because the signal quality of such inexpensive sensors varies significantly from individual to individual. To this end, the approach proposed in this paper innovatively uses data fusion to overcome the challenge while achieving performance that rivals expensive sensors deployed over longer distances in our sensor placement scheme. At the same time, our method provides higher feasibility in practical scenarios.

Wearable-based approaches [23]–[26] require each person to carry a device hence this solution is not applicable in many scenarios. Consider a use case scenario of a shopping mall in an open building or a restricted zone: It is impractical to distribute a special device to each pedestrian or force all people to install some special apps on the smartphone.

The papers [27], [28] propose approaches to detect the position based on audio data from wearable devices. Similarly, these approaches are also wearable-based. The audio signal-based approaches are easily disturbed by environmental noise and can easily be interfered with maliciously. Also, large-scale deployments will be impossible because the audio signals of different units will interfere with each other. Thus, it is difficult to use in potential adversarial scenarios, like bank security monitoring systems, military restricted area monitoring use cases, or open scenarios.

In contrast to some approaches described above, our approach is privacy-protecting because it is only based on floor vibration signal data. It does not require people to carry any wearable devices. Our approach has broader applicability that does not limit the size and shape of the room compared with the WiFi signal-based approach. We used data fusion techniques to overcome the signal quality degrading of inexpensive sensors when deployed close to each other. Obstacles in the zone have no negative impact on the system's effectiveness. When the density of sensors for an area is increased, sensor units will not interfere with each other. As a step-level approach, it does not require that pedestrians walk along the specified route as Akiyanma's research [10]. It is more difficult to destroy than cameras, and its performance is unaffected by visibility impairments. Hence, our approach works even in smoky areas impacted by a fire.

III. Approach and System

In this section, we first define the problem to be solved, followed by a presentation of our approach. Then we explain and analyze the implementation.

A. Problem definition and analysis

The aim of this paper is bidirectional walking detection of pedestrians based on floor vibration. In this paper, we use "Move Right" and "Move Left" to define the 2 directions that persons move, as shown in Fig. 8.

To detect the walking direction, at least two piezoelectric sensors are needed, because piezoelectric sensors can only perceive the amplitude of a ground vibration. When a step

event occurs closer to a sensor, it causes the signal received from this sensor to be larger. Inversely, if the distance where a step event occurs is further away from the sensor, the amplitude of the signal received from sensor will be smaller. Unfortunately, we cannot simply compute the position based on signal strength comparable to triangulation of radio signals, because the ratio between distance and signal strength is not proportional and not even isotropic. The reason for this is that the ground floor is not homogeneous [16].

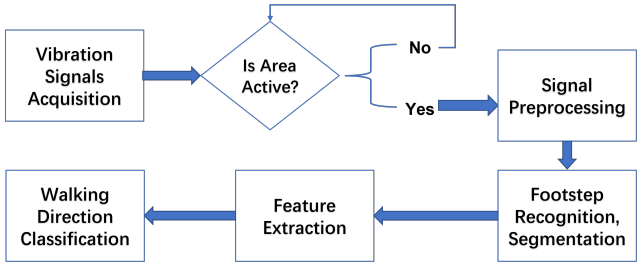


Fig. 3. System architecture.

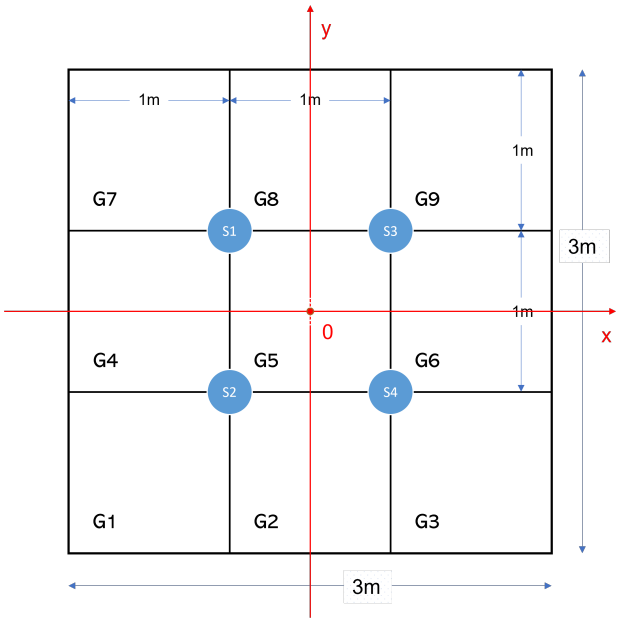


Fig. 4. Detecting area and sensors setup.

As shown in Fig. 3, the proposed approach includes acquiring data from the sensor, active area judgment, data preprocessing, footstep recognition and segmentation, feature extraction, and walking direction classification.

B. Data acquisition

Fig. 4 shows the detecting area and sensors setup. In the system prototype implementation, four EPZ-27MS44W piezoelectric sensors are used to detect the vibration signal, deployed on the position of S1 to S4 as shown in Fig. 4. Previous researches [1], [16] shows that in this way, a 3-meter by 3-meter square area can be well monitored. This kind of sensor can detect a frequency band ranging from 0 Hz to 4400 Hz. The signal is amplified, sampled, quantized, and recorded with a R&S-RTB2000 oscilloscope. Time synchronization of the signals from all four sensors is done by the oscilloscope. Sampling of the signals is done at a rate of 10 kHz. The maximum value of the sampling point is 0.1 V, and any waveforms higher than 0.1 V are cut off. The oscilloscope is controlled, programmed, and configured by a connected laptop using SCPI (Standard Commands for Programmable Instruments) [29] commands.

C. Area active judgment

In the proposed system, the incoming signal will first be recorded in a buffer, which stores 10000 points (1 second). Meanwhile, a sliding window mechanism [30] is used. To prevent missing interesting events, the sliding window size has been empirically determined to be set to 400 points and an overlap of 50%. We define the *active index* as the number of values in the window bigger than the mean of the amplitude of the peaks multiplied by an arbitrary factor (empirically 4 here). If the active index is greater than 10, the area will be considered *active* and data in the current buffer will be further processed. Thus, when the area is active, the data will be fetched to the signal prepossessing module and processed further. If no vibration event is detected (no activity), the system will keep on checking the active status until an activity is detected. This module guarantee that computing resources are not wasted.

D. Signal preprocessing

Wavelet denoising can sufficiently suppress non-stationary noises present in the surroundings [31]. Additionally, the characteristics of the circuit introduce high-frequency noise interference to the signal [32]. Additionally, the characteristics of the circuit introduce high-frequency noise interference to the signal. Our system uses both wavelet denoising techniques and a high-pass filter to enhance the signal. As shown in Fig. 5, after denoising, the signal becomes much clearer.

Afterwards, normalization is used. The range of values of the original data is $[-0.1, 0.1]$. Thus we normalize the signal to $[-1, 1]$ by multiplying all signal values by 10.

E. Step event detection and segmentation

To detect the pedestrian walking direction, we firstly recognize the step events from the vibration events. We used a probabilistic-based first-order second-moment method [33], [34] to isolate the beginning and the end of each vibration event:

$$m_2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 \tag{1}$$

In equation 1, N is the size of the window to be processed, μ is the mean of the values in the window, and x_i are the values in the window. Empirically we determined to set the

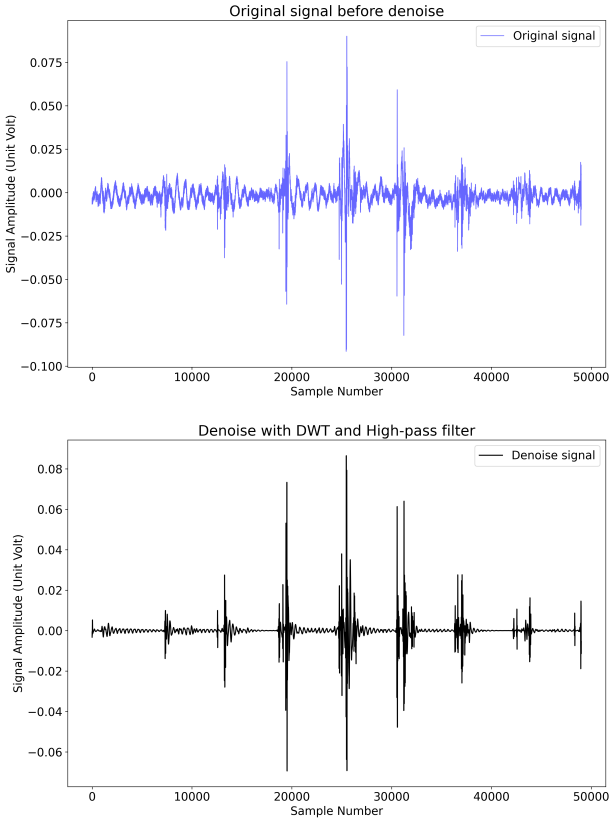


Fig. 5. Signal before and after denoise in time domain. The blue curve shows the original signal before denoising and the black curve describes the signal after denoising.

window size to 30 milliseconds.[1] When the m_2 values become greater than the variance of the ambient background noise, as a threshold, the moment where a SE sample occurred is determined as the beginning of the vibration event. Also, we empirically determined that when the vibration event is longer than 30ms and the m_2 becomes less than 0.3 times the threshold (the variance of the ambient background noise), this moment is considered the end of the event. We use the signal from one of the four sensors to compute the event start and end time. Then the signal of the four sensors in between is the entire SE. If a vibration event is no longer than 30 milliseconds and the gap between it and the next vibration event is less than 10 milliseconds, the two vibration events will be merged and considered as one vibration event.

The event is selected by a sliding window recorded as a event sample. After we get the segmented vibration events, we detect footstep events [11], [35]. Each footstep event is recorded as a SE sample. Fig. 6 shows an example for this SE segmentation.

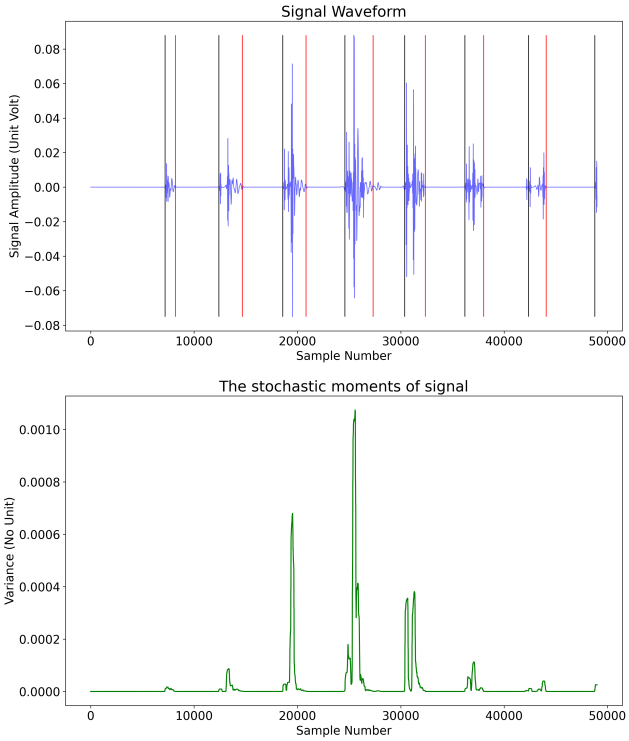


Fig. 6. An example of vibration event segmentation from the signal stream. The black solid lines denote the beginnings of vibration events and red lines the ends.

F. Multi-Peaks Average Walking Direction Detection Algorithm

The sensor 1 and sensor 2 are marked as a "Group A", and sensor 3 and sensor 4 are marked as "Group B", as shown in Fig. 8. We mark the averaged absolute value of the biggest 8 signal peaks from the sensor as $MPAF$ (multi-peak averaging feature). We denote the averaged signal with the group A as $MPAF_A$ and group B as $MPAF_B$. The $MPAF$ represents the energy of a SE. Two consecutive MPAFs are defined as a Multi Step Event (MSE).

If we only consider the *move right* case as shown in Fig. 8, when the A group signal detects that the signals of several adjacent steps are gradually weakening and the B signal detects that the signals of several adjacent steps are gradually increasing, we can assume that the person is located between the sensors, moving away from the A sensor and approaching the B sensor. We label this as the person going right. If both the A signal and B signal detect that several adjacent SE signals are gradually increasing or decreasing, we assume that the person is approaching or moving away from the A and B sensors. In this case, after we compare the magnitude of the signals of the same SE detected by the A and B sensor groups, we determine that the person is on the side with a stronger signal, thus acquiring the walking direction. This is described

[1]This window has no relationship with the window in the active area detecting module. If the area is active, the active judgment module will continuously forward the data stream. If the area is not active, the signal will not be forwarded to the prepossessing module to reduce meaningless computing costs.

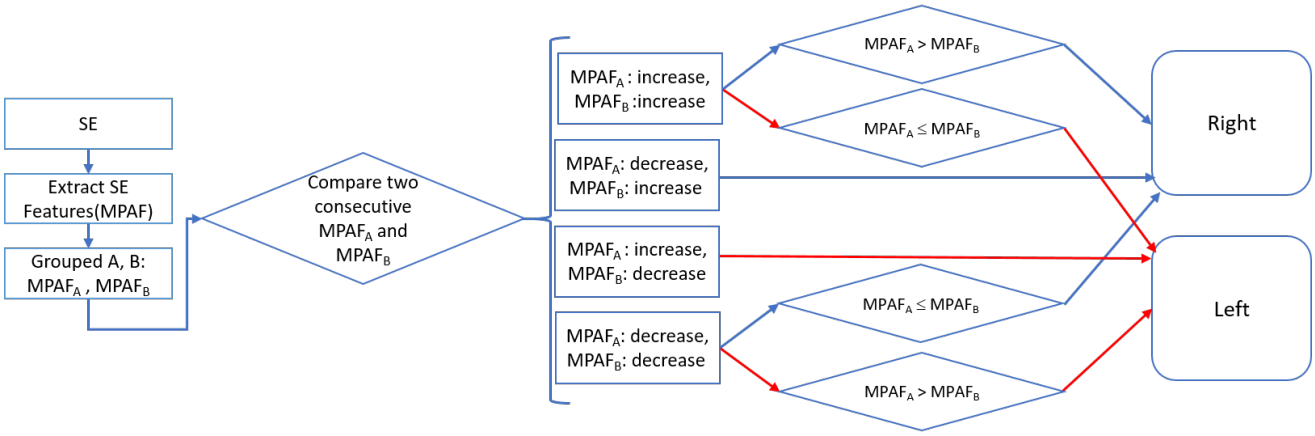


Fig. 7. Multi-peak averaging algorithm. The input of this algorithm is footstep events, and the output results are represented with "left" and "right", referring to the two detected directions.

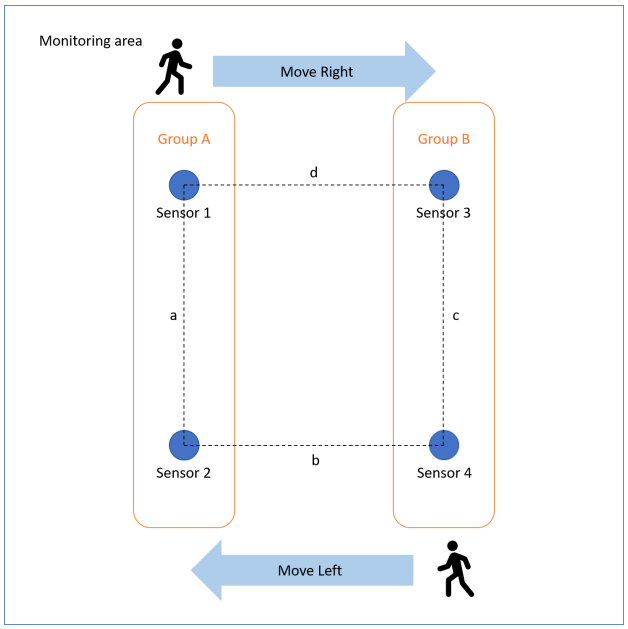


Fig. 8. Walking people in the monitoring area. When people are walking in the monitoring area, the vibration signal will be recorded by the oscilloscope and transmitted to the computer with a USB cable or ethernet directly after the recording procedure.

in Fig. 7.

1) Multi-peak averaging feature: The *multi-peak averaging feature* is obtained by the following method: We use s_{ij} to present the normalized signal (ranging from -1 to 1) of a SE. Here i represents the sensor. i is an integer ranging from 1 to 4. The j represents the order of the sampling point. The range of j can vary according to the time length of a SE. Then the multi-peak averaging feature from sensor i is represented by $MPAF_i$. We use the symbol "$getPeaks()$" to represent the function to get all the peak values of an input signal. "$Max()$" denotes the function computing an array sorted from the largest values to the smallest.

To calculate the $MPAF_i$, we first calculate the absolute value of the normalized SE signal data from sensor i. Then we extract the biggest 8 peak values and calculate the mean value

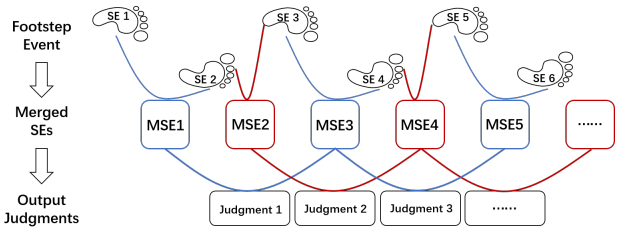


Fig. 9. Multi Steps Events (MSEs) are setten from continuous Step Events. The Multi-peak average algorithm makes a judgment every other MSE, as the MSEs shown in the blue and red colors.

of the 8 peaks. This mean value is the multi-peak averaging feature of sensor i, marked as $MPAF_i$ shown in equation 2.

$$MPAF_i = \frac{1}{8}\sum_{p=1}^{8}\{Max\left[getPeaks(|s_{ij}|)\right]\}_p \tag{2}$$

$$MPAF_A = \frac{1}{2}\left(MPAF_1 + MPAF_2\right) \tag{3}$$

$$MPAF_B = \frac{1}{2}\left(MPAF_3 + MPAF_4\right) \tag{4}$$

2) Multi-peak averaging algorithm design: The next step in tackling the issues found in reality is the *multi-peak averaging algorithm*. It uses the $MPAF$ to represent each SE and compute the walking direction as discussed in section III-A and Fig. 7.

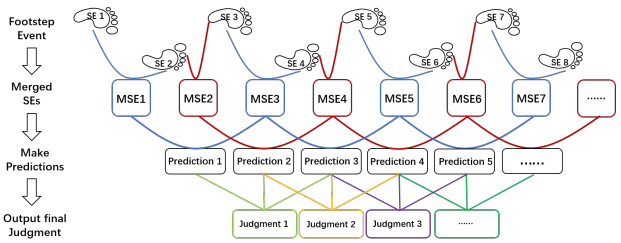


Fig. 10. MPAA with 6 SEs makes a judgment based on every 3 consecutive "predictions". The predictions are made with the same methods as MPAA with 4 SEs.

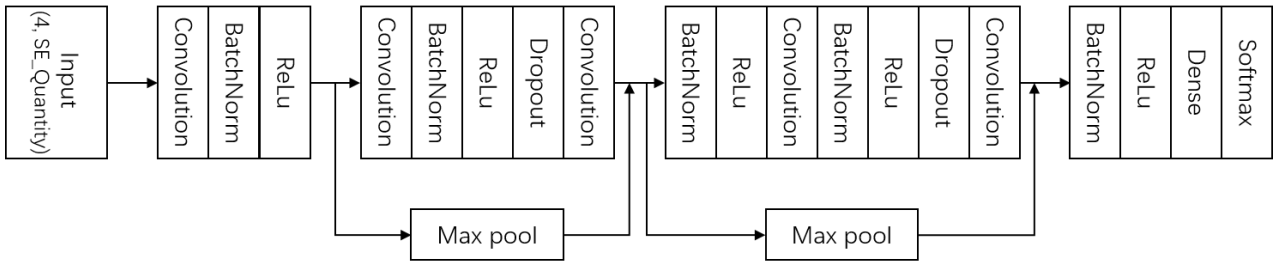


Fig. 11. Deep neural network classifier used in this paper. "SE_Quantity" means the quantity of SE features used. In the implementation, we have trained 6 models with different SEs and different response times. The experiments show that the model with more input SEs has higher classification accuracy but with a longer response time. The model with fewer input SEs has lower classification accuracy but with shorter response time.

We merge 2 consecutive SE into a *Multi Steps Event* (MSE). As shown in Fig. 9, the algorithm makes judgments based on every other MSE, as the MSEs shown in blue and red colors. For example, the first judgment is based on MSE1 and MSE3, and the second judgment on MSE2 and MSE4. We implemented two versions of the algorithm: One is based on 4 SEs, the other one uses 6 SEs. We mark these 2 versions as $MPAASE4$ and $MPAASE6$, respectively. The difference between these 2 versions is that the MPAASE6 introduces a majority voting mechanism. After 6 SEs are fetched, MPAASE6 outputs 3 predictions, and the majority prediction will be the final judgment for the detection.

There is a window shifting one SE each time. When there are enough SEs in the window (either 4 or 6), they are fed to the algorithm. Then the algorithm will output a single judgment. For each time the window shifts, the algorithm will always output a judgement.

3) Deep learning and Multi-peak averaged feature-based classifier: The multi-peak averaging algorithm has verified our hypothesis's correctness that by analyzing the spatiotemporal relationship of the SEs signal, it is feasible to determine the walking direction.

Besides the Multi-peak averaging algorithm approach, we implement another approach which is based on a multi-peak averaged feature of each SE and deep neural network. Moreover, this method further reduces the response time while improving the prediction accuracy. We use the multi-peak averaged feature ($MPAF$), as discussed in section III-F.1, as the input of the neural network. A Softmax is used in the last layer of the neural network serving as a classifier. The architecture of this neural network is as shown in Fig. 11. The proposed neural network includes a residual constructure [36]. There are a total 5 convolutional layers in the architecture. The batch normalization and dropout are introduced [36]–[39]. We use the Relu functions as the activation function in the neural network. There are 32 convolution filters in each convolutional layer, with dropout rates of each at 0.3. The learning rate is set to 0.001, and the batch size 32. The size of the input data sample is 4 rows and "$SE_Quantity$" columns (4, $SE_Quantity$). The "$SE_Quantity$" means the number of SEs to use when one pedestrian passes by the sensors.

We fully trained 6 models with input SEs from 2 to 7. The more input SEs are used, the higher the classification accuracy becomes at the cost of longer response times. Fewer SEs inputted to the classifier achieve lower classification accuracy but shorter response times. The more footsteps needed to make a judgment, the longer a person has to move constantly in one direction, which limits applicability. Thus, the fewer footsteps needed, the better, which is also a key feature of our approach. We implemented 3 to 7 SEs as input and compared the performance in the evaluation section IV.

G. Discussion

Using our data processing scheme, sensors can be deployed at a distance of 1m from each other. Such deployment density allows the system to track multiple persons moving in a room.

The signal of each sensor is preprocessed by signal denoising and filtering. These measures are all conducive to ensuring reliable and stable signal quality. At the same time, our algorithm adopts a data fusion design based on multiple data sources, which further improves the system's measurement accuracy.

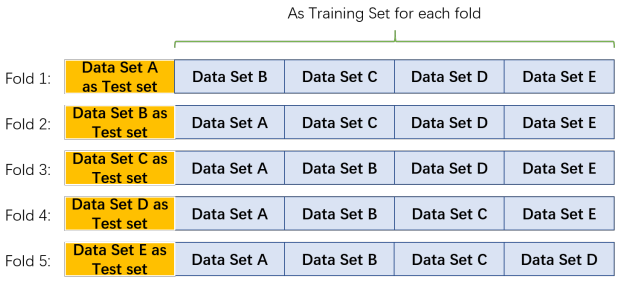
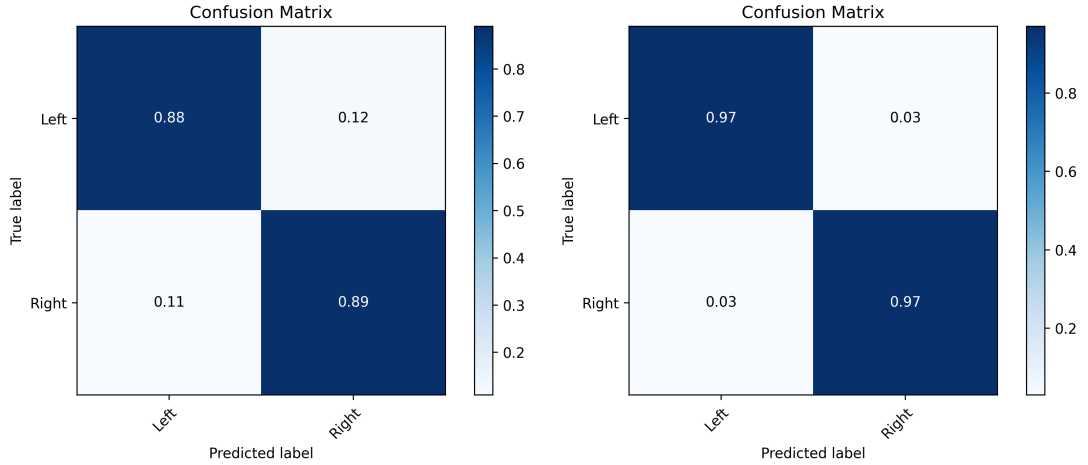


Fig. 12. For each SE2 to SE7 model, the samples are randomly shuffled and evenly split into 5 subset of the data. In the 5-fold cross-validation, each time we pick out one subset as test set and the rest subsets bind together as training set.

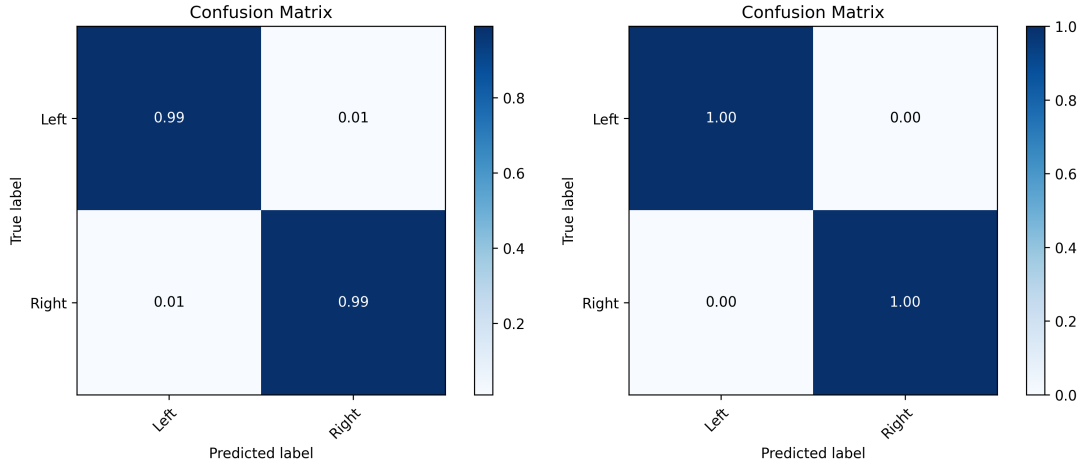
IV. Experiment and Evaluation

In this section statistical methods are used to evaluate the results. As shown in Table. I, Table. II, Table. III, the statistical methods metrics including accuracy, recall, f1-score are used to evaluate the performance of our classification algorithms. As shown in Fig. 13, the statistical method confusion matrix is used to present the DNN-based classifier evaluation results.



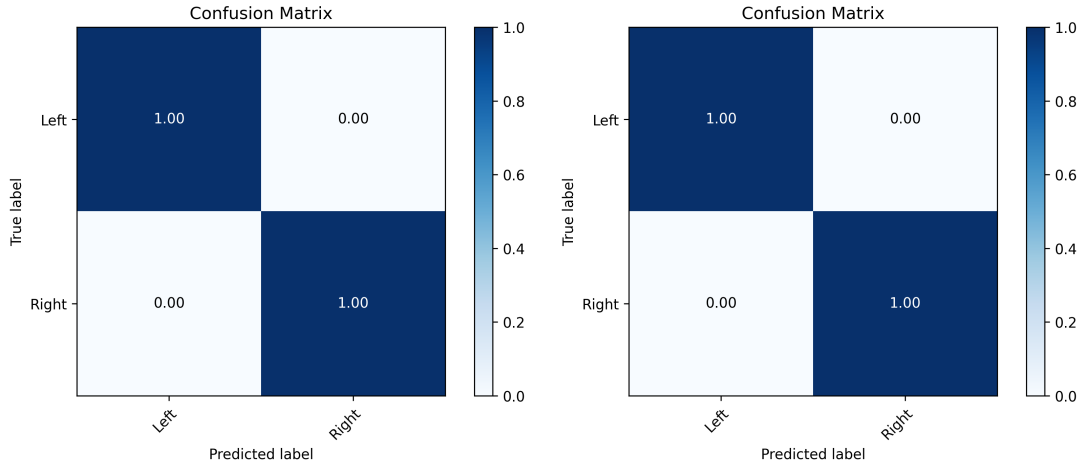
(a) 5-Folds Confusion Matrix of DNN-based classifier with 2 SEs.

(b) 5-Folds Confusion Matrix of DNN-based classifier with 3 SEs.



(c) 5-Folds Confusion Matrix of DNN-based classifier with 4 SEs.

(d) 5-Folds Confusion Matrix of DNN-based classifier with 5 SEs.



(e) 5-Folds Confusion Matrix of DNN-based classifier with 6 SEs.

(f) 5-Folds Confusion Matrix of DNN-based classifier with 7 SEs.

Fig. 13. 4 sensors-based 5-Folds cross-validation confusion matrix of DNN-based classifier with 2 SEs to 7 SEs.

	Precision	Recall	F1-score
2 sensors			
Left	0.81	0.87	0.84
Right	0.87	0.82	0.84
Accuracy			0.84
Micro Average	0.84	0.84	0.84
Macro Average	0.84	0.84	0.84
4 sensors			
Left	0.91	0.93	0.92
Right	0.94	0.91	0.93
Accuracy			0.92
Micro Average	0.92	0.92	0.92
Macro Average	0.92	0.92	0.92

TABLE I
CLASSIFICATION PERFORMANCE OF THE MPAA WITH 4 SES.

	Precision	Recall	F1-score
2 sensors			
Left	0.89	0.95	0.92
Right	0.95	0.89	0.92
Accuracy			0.92
Micro Average	0.92	0.92	0.92
Macro Average	0.92	0.92	0.92
4 sensors			
Left	0.98	0.99	0.98
Right	0.99	0.98	0.98
Accuracy			0.98
Micro Average	0.98	0.98	0.98
Macro Average	0.98	0.98	0.98

TABLE II
CLASSIFICATION PERFORMANCE OF THE MPAA WITH 6 SES WITH MAJORITY VOTING.

A. Data collection

We totally collected 1080 pieces of *Left* walking data and 1175 *Right* walking data. The sampling rate is 10000KHz. The data set was generated by 2 experiment participants with 2 pairs of shoes for each person.

During the walking direction data collection experiment, the participant was instructed to walk normally, as in daily life. The experiment devices recorded the signal from the floor vibration. For each round of walking, the person walks in the experimental room from one end to the other. Each walking round takes less than 6 seconds.

B. Algorithm performance of MPAA

Table I shows the classification performance by the MPAA with 4 SEs. Table II is the performance of the MPAA with 6 SEs with majority voting. Both tables contain the experiment results when 2 sensors and 4 sensors are used. For the "2 sensors" data in the table, the averaged value of the experimental data is presented when only 2 sensors are used (sensor 1 and sensor 3, or sensor 2 and sensor 4). The tables show that the 4 sensors approach is better than the 2 sensors approach under the same input SEs. When 4 SEs are inputted, the F1-score achieved 0.92. The prevision of our MPAA 4SEs approach has already greatly exceeded the best performance of Akiyama's research (0.756) [10]. Considering that with average human walking speed four steps take about 3 seconds and six steps 4.5 seconds, our system features a faster response time. Our 6 SEs approach obtains a F1-score as high as 0.98.

Furthermore, our approach is a step event-based step level approach, different from Akiyama's research, that does not limit the relative location of the people to the sensor. This makes the system more practical in high-traffic scenarios.

C. MPAF-DNNC approach performance

We conducted 5-fold cross-validation [40], [41] to evaluate the performance of our model. We used the Left and Right walking recording data totally 2255 pieces to generated 17585 samples for 2 SEs, 13530 samples for 3 SEs, 11275 samples for 4 SEs, 9020 samples for 5 SEs, 6765 samples for 6 SEs and 4510 sampels for 7 SEs. As shown in Fig. 12, for each 2 SEs model to 7SEs model, we randomly shuffled the data samples and evenly split them into 5 subsets. For each fold of the validation, we use one subset as test set and the rest subsets binding together as training set. We train the model 5 times and evaluate it 5 times. As shown in TABLE III, the performance with data only from 2 sensors and 4 sensors are compared. The experiment shows that with 4 sensors, the system can output satisfied judgment only with 3 SEs. In contrast, if only with 2 sensors we can get satisfied judgment at least 4 SEs. The 4 sensors-based method can offer faster response and better results than the 2 sensors-based method.

As shown in the TABLE III and Fig. 13, the deep nerual network-based classifier 5-folds cross-validation averaged performance metrics are presented. In the TABLE III, values accurate to 2 decimal places. Our DNN-based classifier has shown very good performance when 3 SEs are used. Comparing with Akiyama's research [10], even our 2 SEs based classifier is beyond their best k-fold cross-validation performance. The MPAF-DNNC approach only requires 3 SEs while having a 0.97 F1 score.

Our approachs further reduces the number of the required footstep events while significantly improving the judgment accuracy. Since the proposed system is step-level, thus this approach dramatically improves the usability of this technology in practical application scenarios.

Compared with existing researches [10], [11], [11], [12], Our method achieves 100% accuracy in the highest-accuracy configuration, numerically outperforming or matching existing methods. However, our method overcomes many impractical limitations of existing methods on usage scenarios, making it more feasible for practical use.

V. CONCLUSION

This paper presents a floor vibration signal-based bi-directional walking direction detection approach. The proposed approach uses the piezoelectric sensor to detect the floor vibration signal, which is a privacy-protecting approach. The contribution of our approach is that we reduce the number of required step events to 3 SEs regarding the vibration signal-based walking direction detection solution. At the same time, keep the F1 score better than 0.97. Our approach achieved an advantage in performance and usability as a step-level solution over existing research. Moreover, the characteristic of

TABLE III
5-FOLDS CLASSIFICATION PERFORMANCE OF THE DNN-BASED CLASSIFIER WITH 2 SENSORS AND 4 SENSORS.

Number of Step Event	Expected Response time		4 sensors			2 sensors		
			Precision	Recall	F1-score	Precision	Recall	F1-score
2 SEs	1.5 second	Left	0.87	0.89	0.88	0.72	0.78	0.74
		Right	0.91	0.89	0.89	0.79	0.73	0.76
		Accuracy			0.89			0.75
		Micro Average	0.88	0.88	0.88	0.75	0.75	0.75
		Macro Average	0.88	0.88	0.88	0.75	0.75	0.75
3 SEs	2.25 second	Left	0.97	0.97	0.97	0.88	0.88	0.88
		Right	0.97	0.97	0.97	0.89	0.88	0.88
		Accuracy			0.97			0.88
		Micro Average	0.97	0.97	0.97	0.88	0.88	0.88
		Macro Average	0.97	0.97	0.97	0.88	0.88	0.88
4 SEs	3 second	Left	0.99	0.99	0.99	0.95	0.95	0.95
		Right	0.99	0.99	0.99	0.95	0.95	0.95
		Accuracy			0.99			0.95
		Micro Average	0.99	0.99	0.99	0.95	0.95	0.95
		Macro Average	0.99	0.99	0.99	0.95	0.95	0.95
5 SEs	3.75 second	Left	0.99	1.00	1.00	0.99	0.97	0.98
		Right	1.00	1.00	1.00	0.97	0.99	0.98
		Accuracy			1.00			0.98
		Micro Average	1.00	1.00	1.00	0.98	0.98	0.98
		Macro Average	1.00	1.00	1.00	0.98	0.98	0.98
6 SEs	4.5 second	Left	1.00	1.00	1.00	0.99	0.99	0.99
		Right	1.00	1.00	1.00	0.99	0.99	0.99
		Accuracy			1.00			0.99
		Micro Average	1.00	1.00	1.00	0.99	0.99	0.99
		Macro Average	1.00	1.00	1.00	0.99	0.99	0.99
7 SEs	5.25 second	Left	1.00	1.00	1.00	1.00	0.99	1.00
		Right	1.00	1.00	1.00	0.99	1.00	1.00
		Accuracy			1.00			1.00
		Micro Average	1.00	1.00	1.00	1.00	1.00	1.00
		Macro Average	1.00	1.00	1.00	1.00	1.00	1.00

our approach makes it feasible for the vibration signal-based walking direction detection technology to come from the lab to daily life.

Because the proposed approach is step-level and the system's input only regards step events, it is able to support multiple pedestrians walking use cases when combined with the people re-identify technology. Moreover, it is a wearable-free approach that does not need pedestrians to carry any wearable devices. Furthermore, it does not limit the relative position of the pedestrian to the sensors and is easy to scale out. Our approach does not limit the size or shape of the room where it is deployed.

The oscilloscope used in the experiment can be replaced with a custom integrated circuit in products, thereby further reducing costs in the future.

ACKNOWLEDGMENTS

The author would like to thank Evonik Digital to access their production plant for ground wave measurements. We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen. We acknowledge and thank Evonik Digital for this research work's financial support.

REFERENCES

[1] Y. Yu and T. Weis, "A privacy-protecting indoor emergency monitoring system based on floor vibration," in *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, ser. UbiComp-ISWC '20. New York, NY, USA: Association for Computing Machinery, Sep. 2020, pp. 164–167.

[2] S. Hussain, Y. Yu, M. Ayoub, A. Khan, R. Rehman, J. A. Wahid, and W. Hou, "IoT and Deep Learning Based Approach for Rapid Screening and Face Mask Detection for Infection Spread Control of COVID-19," *Applied Sciences*, vol. 11, no. 8, p. 3495, Jan. 2021.

[3] Y. Yu, X. Qin, S. Hussain, W. Hou, and T. Weis, "Pedestrian Counting Based on Piezoelectric Vibration Sensor," *Applied Sciences*, vol. 12, no. 4, p. 1920, Jan. 2022.

[4] F. Li, J. Clemente, M. Valero, Z. Tse, S. Li, and W. Song, "Smart Home Monitoring System via Footstep-Induced Vibrations," *IEEE Systems Journal*, vol. 14, no. 3, pp. 3383–3389, Sep. 2020.

[5] S. Pan, M. Mirshekari, P. Zhang, and H. Y. Noh, "Occupant traffic estimation through structural vibration sensing," in *SPIE Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring*, J. P. Lynch, Ed., Las Vegas, Nevada, United States, Apr. 2016, p. 980306.

[6] S. Pan, N. Wang, Y. Qian, I. Velibeyoglu, H. Y. Noh, and P. Zhang, "Indoor Person Identification through Footstep Induced Structural Vibration," in *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, ser. HotMobile '15. New York, NY, USA: Association for Computing Machinery, Feb. 2015, pp. 81–86.

[7] S. Savazzi, S. Sigg, M. Nicoli, V. Rampa, S. Kianoush, and U. Spagnolini, "Device-Free Radio Vision for Assisted Living: Leveraging wireless channel quality information for human sensing," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 45–58, Mar. 2016.

[8] P. Voigt and A. von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed. Springer Publishing Company, Incorporated, 2017.

[9] H. Gassmann, "OECD guidelines governing the protection of privacy and transborder flows of personal data," *Computer Networks (1976)*, vol. 5, no. 2, pp. 127–141, Apr. 1981.

[10] S. Akiyama, M. Yoshida, Y. Moriyama, H. Suwa, and K. Yasumoto, "Estimation of Walking Direction with Vibration Sensor based on Piezo-

electric Device," in *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, Mar. 2020, pp. 1–6.
[11] S. Pan, A. Bonde, J. Jing, L. Zhang, P. Zhang, and H. Y. Noh, "BOES: Building Occupancy Estimation System using sparse ambient vibration monitoring," in *SPIE Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring*, J. P. Lynch, K.-W. Wang, and H. Sohn, Eds., San Diego, California, USA, Apr. 2014, p. 90611O.
[12] S. Pan, M. Mirshekari, J. Fagert, C. Ruiz, H. Y. Noh, and P. Zhang, "Area Occupancy Counting Through Sparse Structural Vibration Sensing," *IEEE Pervasive Computing*, vol. 18, no. 1, pp. 28–37, Jan. 2019.
[13] S. Hussain, M. Ayoub, G. Jilani, Y. Yu, A. Khan, J. A. Wahid, M. F. A. Butt, G. Yang, D. P. Moller, and H. Weiyan, "Aspect2Labels: A novelistic decision support system for higher educational institutions by using multi-layer topic modelling approach," *Expert Systems with Applications*, vol. 209, p. 118119, Dec. 2022.
[14] V. Matkovic, M. Waltereit, P. Zdankin, and T. Weis, "Towards Bike Type and E-Scooter Classification With Smartphone Sensors," in *MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, ser. MobiQuitous '20. New York, NY, USA: Association for Computing Machinery, Dec. 2020, pp. 395–404.
[15] V. Matkovic, M. Waltereit, and T. Weis, "Towards Predictive Safety Maintenance for IoT Equipped Bikes," in *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops)*, Mar. 2021, pp. 320–323.
[16] Y. Yu, M. Waltereit, V. Matkovic, W. Hou, and T. Weis, "Deep Learning-Based Vibration Signal Personnel Positioning System," *IEEE Access*, vol. 8, pp. 226 108–226 118, 2020.
[17] F. Castanedo, "A Review of Data Fusion Techniques," *The Scientific World Journal*, vol. 2013, p. e704504, Oct. 2013.
[18] S. A. Kashinath, S. A. Mostafa, A. Mustapha, H. Mahdin, D. Lim, M. A. Mahmoud, M. A. Mohammed, B. A. S. Al-Rimy, M. F. M. Fudzee, and T. J. Yang, "Review of Data Fusion Methods for Real-Time and Multi-Sensor Traffic Flow Analysis," *IEEE Access*, vol. 9, pp. 51 258–51 276, 2021.
[19] Y. Himeur, B. Rimal, A. Tiwary, and A. Amira, "Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives," *Information Fusion*, vol. 86–87, pp. 44–75, Oct. 2022.
[20] L. Chen, Chi-Ren Chen, and Da-En Chen, "VIPS: A video-based indoor positioning system with centimeter-grade accuracy for the IoT," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, Mar. 2017, pp. 63–65.
[21] G. Zhao, M. Takafumi, K. Shoji, and M. Kenji, "Video based estimation of pedestrian walking direction for pedestrian protection system," *Journal of Electronics (China)*, vol. 29, no. 1, pp. 72–81, Mar. 2012.
[22] D. Wu, D. Zhang, C. Xu, Y. Wang, and H. Wang, "WiDir: Walking direction estimation using wireless signals," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, Sep. 2016, pp. 351–362.
[23] M. F. Shaikh, Z. Salcic, and K. I.-K. Wang, "A Novel Accelerometer-Based Technique for Robust Detection of Walking Direction," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 8, pp. 1740–1747, Aug. 2018.
[24] N. Roy, H. Wang, and R. Roy Choudhury, "I am a smartphone and i can tell my user's walking direction," in *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '14. New York, NY, USA: Association for Computing Machinery, Jun. 2014, pp. 329–342.
[25] J. W. Kim, H. J. Jang, D.-H. Hwang, and C. Park, "A Step, Stride and Heading Determination for the Pedestrian Navigation System," *Journal of Global Positioning Systems*, vol. 3, no. 1&2, pp. 273–279, Dec. 2004.
[26] J. Á. B. Link, P. Smith, N. Viol, and K. Wehrle, "FootPath: Accurate map-based indoor navigation using smartphones," in *2011 International Conference on Indoor Positioning and Indoor Navigation*, Sep. 2011, pp. 1–8.
[27] S. I. Lopes, J. M. N. Vieira, J. Reis, D. Albuquerque, and N. B. Carvalho, "Accurate smartphone indoor positioning using a WSN infrastructure and non-invasive audio for TDoA estimation," *Pervasive and Mobile Computing*, vol. 20, pp. 29–46, Jul. 2015.
[28] S. Cao, X. Chen, X. Zhang, and X. Chen, "Effective Audio Signal Arrival Time Detection Algorithm for Realization of Robust Acoustic Indoor Positioning," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7341–7352, Oct. 2020.
[29] I. G. o. I. . Measurement, A. N. S. Institute, and I. S. Board, *IEEE Standard Digital Interface for Programmable Instrumentation*. IEEE, 1983.
[30] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International Conference on Pervasive Computing*. Springer, 2004, pp. 1–17.
[31] W. Chen, M. Guan, L. Wang, R. Ruby, and K. Wu, "FLoc: Device-free passive indoor localization in complex environments," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.
[32] S. Durand and J. Froment, "Artifact free signal denoising with wavelets," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 6, May 2001, pp. 3685–3688 vol.6.
[33] F. Li*, J. Rich, K. J. Marfurt, and H. Zhou, "Automatic event detection on noisy microseismograms," in *SEG Technical Program Expanded Abstracts 2014*, ser. SEG Technical Program Expanded Abstracts. Society of Exploration Geophysicists, Aug. 2014, pp. 2363–2367.
[34] J. Clemente, F. Li, M. Valero, and W. Song, "Smart Seismic Sensing for Indoor Fall Detection, Location, and Notification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 524–532, Feb. 2020.
[35] M. Mirshekari, J. Fagert, S. Pan, P. Zhang, and H. Y. Noh, "Step-Level Occupant Detection across Different Structures through Footstep-Induced Floor Vibration Using Model Transfer," *Journal of Engineering Mechanics*, vol. 146, no. 3, p. 04019137, Mar. 2020.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[37] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
[38] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
[39] J. Bouvrie, "Notes on Convolutional Neural Networks," http://cogprints.org/5869/, Nov. 2006.
[40] S. Russell and P. Norvig, "Artificial intelligence: A modern approach," 2002.
[41] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013, vol. 112.

Yang Yu is currently pursuing the Ph.D. degree with the Distributed Systems Research Group, University of Duisburg-Essen, Duisburg, Germany. His research interests include pervasive computing, AI-based context recognition, AI-based context reasoning, pattern recognition. He received the M.Sc. degree in Embedded System Engineering from the University of Duisburg-Essen and the B.Eng. degree in Communication Engineering from Zhengzhou University

Oskar Carl is currently pursuing a Ph.D. at the Distributed Systems Research Group, University of Duisburg-Essen, Duisburg, Germany. His primary research interests include communication in distributed systems, decentralized systems and cloud computing. He received M.Sc. and B.Sc. degrees in Applied Computer Science at the University of Duisburg-Essen.

Shabir Hussain is currently pursuing his Ph.D. degree majoring in Software engineering from the School of Information Engineering, Zhengzhou University, Henan, China. He received an M.S. degree in Computer Science from NCBA&E, Lahore, Pakistan, in 2016. He has been a lecturer with the Computer Science Department of NCBA&E, Rahim Yar Khan campus, from 2016 to 2018. His research interests include machine learning, natural language processing, IoT Systems, and data analytics, and Complex network analysis.

Weiyan Hou is a Professor in the school of information engineering of University Zhengzhou, China. He has been engaged in the field of network control, time performance modeling of industrial IoT and infrastructure of wired/wireless communication integration, obtained his Bachelor and Master degree in 1986, 1998 in China and doctor degree from Shanghai University in a Sandwich Ph.D. program between Germany and China in 2004.

Torben Weis is a Professor with the University Duisburg-Essen, Duisburg, Germany, where he leads the Distributed Systems Research Group. His research interests include cyber-physical systems, cloud computing as well as security and privacy in distributed systems. He received the Ph.D. degree in computer science from the Technical University of Berlin, Berlin, Germany.

12.4 Deep Learning-Based Vibration Signal Personnel Positioning System

Title	Deep Learning-Based Vibration Signal Personnel Positioning System
Authors	Yang Yu, Marian Waltereit, Viktor Matkovic, Weiyan Hou, Torben Weis
Publication Venue	IEEE Access, December 14, 2020
Publication Type	Article
Publication Status	Published
Research Topic	Personnel Positioning
DOI	https://doi.org/10.1109/ACCESS.2020.3044497

Table 12.4: Bibliographic information of publication D.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 26, 2020, accepted December 8, 2020, date of publication December 14, 2020,
date of current version December 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3044497

Deep Learning-Based Vibration Signal Personnel Positioning System

YANG YU[1,2], MARIAN WALTEREIT[1], VIKTOR MATKOVIC[1], WEIYAN HOU[2], AND TORBEN WEIS[1]

[1]Distributed Systems Group, University of Duisburg-Essen, 47057 Duisburg, Germany
[2]School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China

Corresponding author: Yang Yu (yang.yu@uni-due.de)

We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen. We acknowledge and thank Evonik Digital for this research work's financial support.

ABSTRACT In this work, we present a person localization system based on ground vibration caused by walking persons. The system is designed for production plants and large buildings to track the movement of workers. Position and movement in these settings are especially safety-relevant in emergencies. Our approach is privacy-preserving, because it requires neither video nor sound. Instead, piezo sensors on the floor measure vibrations, which are analyzed with machine learning to derive a person's position from the vibration signals. This way, our system can determine where a person is moving, but it is not straightforward to attach names to the detected persons. Due to the anisotropic characteristic of the ground vibration wave, classical analysis methods are not applicable. We show that a deep learning-based approach is feasible. Our experiments show that we can determine the position with an average F1 score of 0.95.

INDEX TERMS Vibration signal, localization, pattern recognition, deep learning, privacy protection, robustness, piezo sensor.

I. INTRODUCTION

Localization technology has attracted attention from industry and academics [1]. It is the cornerstone for a large number of person localization services [2]–[6]. Person localization data can be used in scenarios like post-disaster rescue, situational awareness, security defense, and intrusion detection. At the same time, people also hope that the data applied while obtaining location information does not violate their privacy.

However, there are many studies on surveillance and situation awareness, but rarely the issue of privacy is considered. In some countries or regions, such as Europe, privacy is extremely important due to legal restrictions. At the same time, people's need for intelligent situational awareness is growing. The motivation of our research is to balance these two contradictory points in order to enable intelligent situation awareness while preserving privacy.

In this paper, we introduce an approach that is based on the analysis of structural vibration waves caused by footsteps, measured with cheap piezo sensors on the floor. The main research challenge is to determine the position of individual footsteps on the floor. This information can be used to determine a higher-level context, such as the walking direction or the walking path of a person.

The advantage of our approach is that we do not rely on persons wearing any senders or receivers. Meanwhile, our approach uses very cheap piezo sensors rather than expensive geophones [7], which makes a large-scale commercial deployment possible. Furthermore, our system is less privacy-invasive than cameras, and it does not detect who is walking. Especially for rescue scenarios in large production plants, it is imperative to know the location of people and their direction of movement. Besides, our system works in smoked areas where cameras do not work anymore.

Our approach is based on the observation that the steps of a walking person cause a mechanical vibration of the ground. At this time, the position of the person is the position of the vibration source. Nevertheless, the time difference and amplitude of the recorded vibration signals are related to the vibration source position. Thus, we assumed that the vibration source could be determined by an analysis of the vibration signals' characteristics.

However, due to the non-uniform nature of the ground structure materials and factors caused by the ground construction process, the ground substances' property is anisotropic. Also, ground waves show multipath effects caused by various obstacles like walls, etc. in the indoor space. Furthermore,

The associate editor coordinating the review of this manuscript and approving it for publication was Lefei Zhang.

the recorded vibration signal has the characteristics of low signal-to-noise ratio (SNR) [7], which leads to an inevitable large measurement error in signal processing. Due to these characteristics, person localization based on mechanical ground waves is challenging. The traditional time differences of arrival (TDOA), time-of-arrival (TOA), angle-of-arrival (AOA), Doppler shift frequency-difference-of-arrival (FDOA) as well as received signal strength (RSS) based methods are widely used in wireless radio-based localization approaches with sensor arrays [7]–[10]. Due to mechanical wave propagation characteristics on the indoor floor, these methods cannot achieve good performance in our scenario.

Meanwhile, there are methods based on time difference distribution models [11] to recognize the ball impact localization on table tennis rackets using piezo-electric sensors. In this research, the sensors are distributed relatively close to each other, and the vibration is spreading on wood. The starting point of the wave signal can be identified relatively clearly with low measurement errors. In contrast, as mentioned above, in our scenario the quality of the signal propagation on the concrete floor is much worse than on the wood floor, and the measurement error of the waveform arrival time is immense. So we need to find a different method to overcome this problem.

Neural networks are capable of learning complex nonlinear relationships [12]. Convolutional neural networks show good performance in feature extraction from data [13]. Meanwhile, residual connections make deep networks easier to train without increasing the complexity of the network [14]. There is existing research that uses deep residual convolutional networks to analyze time serial data with good performance [15]. Therefore, in this paper, we show that the analysis of ground vibration signals is feasible with deep residual neural networks.

Our approach's starting point is a customized positioning application in a specific area of a chemical plant with the feature of privacy protection. In the future, our approach can be used on vibration-based positioning applications in general scenarios. The cost of one set of the sensor matrix is lower than 2 Euro. With custom circuits and mass production, the hardware cost involved in our approach can be greatly reduced.

Meanwhile, our sensor deployment requires one unit every $9m^2$. The function of each unit is relatively independent. When there is a larger area, more units can be deployed. When pedestrians walk across different units, the entire integrated system can give global positioning information to pedestrians.

Our approach's sensor can be embedded into floor tiles to unify the distribution characteristics of training data and data from actual application scene, thereby improving the feasibility of our approach in practical applications. Meanwhile, there is ongoing research about domain adaptation [16], [17]. The domain adaptation technique makes the deep learning-based approach work on the target domain data, which has different distribution characteristics than the training data. Furthermore, the dimensionality reduction technology can further reduce the dimensionality of the input data, thereby reducing the computing cost [18], [19]. The introduction of domain adaptation methods and dimensionality reduction methods to our system will be further studied in future work.

The contribution of this paper is as follows. We present a novel fine-grained deep learning [20] based localization technique using cheap vibration sensors, which is a passive detection that does not need the pedestrian to carry a device. The proposed technique can well tolerate noise in the indoor environment. We evaluated and validated the accuracy and robustness of the approach with experiments in a real scenario.

II. RELATED WORK

Localization systems for outdoor scenarios are often realized with GPS. However, GPS does not work indoor [21]. Pedestrian Dead Reckoning-based methods use inertial sensors, and they cannot avoid the error accumulation problem [22]–[24].

The radiofrequency or WiFi fingerprint-based positioning methods [25], [26] calculate the location with the signal from the transmitter carried by the pedestrian. All the above methods require each person to carry a device, not the passive localization method, without carrying equipment that our target scenario pursues.

Video-based localization techniques do not support privacy protection for people in the given scenario. In extreme environments, such as explosions, fires, etc., where visibility is reduced, the location functionality of video-based techniques and visible light communication-based positioning techniques [27] will no longer be available.

There are floor vibration-based indoor localization methods which introduced TDOA [7], [28]. The TDOA method assumed that the speed of wave propagation is constant. Due to the ground's anisotropy, the ground substances do not satisfy this condition when a mechanical wave is conducted. In other words, the propagation velocity on the indoor floor is not constant. So if the TDOA method is used, the constant speed assumption should be eliminated. However, even if the speed is constant, the simple TDOA method cannot obtain satisfactory positioning performance. To improve the positioning accuracy, researchers introduced an ATDOA [29] method that combines the TDOA method and the AOA method by the weighted average method. Although this ATDOA method reduces the positioning error to a certain extent, the authors had to use a triaxial seismic sensor, which increases deployment costs, creating obstacles to commercial deployment from cost considerations. Furthermore, this method's localization performance on the concrete floor is far from satisfactory compared to the wooden floor.

Meanwhile, because the TDOA algorithm's essence is to find the coordinate of the intersection of more than three hyperbolae, this algorithm has the well known "no solu-

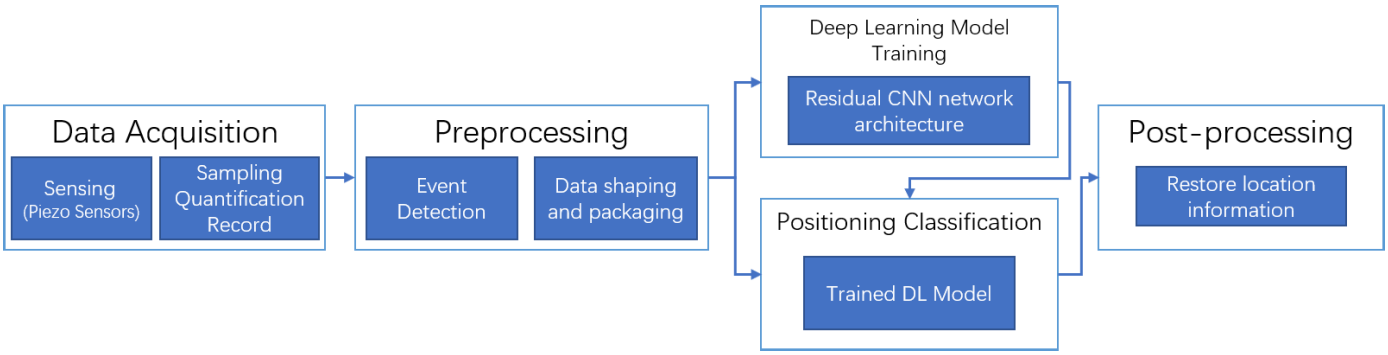


FIGURE 1. The system includes data acquisition, data preprocessing, deep learning model training, positioning classification, and post-processing.

tion'' and ''multiple solutions'' problems. This has led to restrictions on where sensors can be deployed. To avoid these problems, all the above TDOA-based research deploys the sensors on the four vertices of the rectangle, and the area inside the rectangle is the positionable area. However, this is very unfavorable for large-scale deployment, especially in large areas, because the signal decays and thus the size of the supported area is strictly limited.

As a comparison, there is research using the cheap piezo sensor to conduct indoor localization [30]. Nevertheless, this method requires reference positioning objects with a known position, which can hardly be recognized as a real positioning system. The positioning accuracy depends on the density of the reference position objects. This method's limitations are significant, and it can hardly be used in the most indoor localization scenario.

In contrast to the existing localization systems, our approach is a passive localization method that does not require persons to carry any device. This feature is an added benefit in chemical plants where, by default, mobile electronic devices are not allowed unless they are certified not to cause explosions in combination with potentially explosive gas. Furthermore, our approach does not require a clear line of sight and therefore works in heavily smoked rooms. Our approach uses cheap piezo sensors, which has a cost advantage, and there is no restriction on the relative position of the sensor position and the located area like the TDOA-based method. This makes it more feasible in large-scale deployments, especially in large areas. Our method still has good accuracy and robustness on the concrete floor.

III. METHODOLOGY AND SYSTEM

As shown in Fig. 1, the proposed system is a deep learning-based indoor and outdoor passive pedestrian localization system. In this section, the procedure is discussed.

A. PROBLEM FORMULATION AND ANALYSIS

In our system, four piezo sensors on the floor can detect footsteps in a square of $9m^2$. In our experiments, we want to determine the position of individual footsteps in this square. This three meters by three meters square is divided into nine zones, each one meter by one meter in size (see Fig. 2).

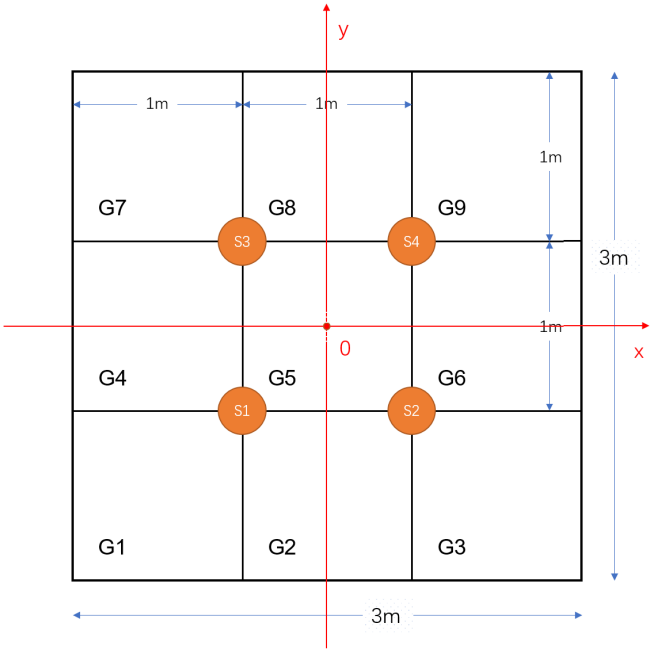


FIGURE 2. System model. A space of 3 meters by 3 meters is divided into nine grids named G1 to G9. Four sensors, S1 to S4, are deployed in the position points, as shown in this figure. When the sensor-matrix detects a vibration source in the observed zone, the zone name should be output.

For each footstep, we want to determine the zone in which the footstep occurred. These zones are labeled as G1 to G9. As the signal decays quickly in thick concrete floors, we can test by amplitude which arrangement of four sensors is closest to the pedestrian. Thus, we investigate how to detect a person inside these nine zones and the approach scales to arbitrarily larger floors nevertheless.

Concerning our machine learning approach, each zone is represented by a class. Thus, we designed a classifier that can determine the zone based on four channels of vibration signals. The sensors producing these signals are labeled as S1 to S4. In order to cover larger areas, the same arrangement of zones and sensors can be repeated in both directions.

The location information of the vibration source has a relationship with the maximum amplitude of the four channels and the relative peak position in time. The waveform of the signal of the channel nearest to the vibration source will start to rise first. Meanwhile, because the amplitude of the wave decays with the distance in the direction of the

wave propagation, the waveform of the sensor nearest to the vibration source will have the maximum amplitude among all four signals. To achieve localization, we used a deep residual convolutional neural network to learn the nonlinear relationship between the point location and the signal of the four sensors.

B. METHODOLOGY DESCRIPTION

In this research, the aim is to detect persons walking across the floors of large scale buildings, for example in a chemical plant.

Our methodology is to deploy an array of four sensors every 9 m^2. When a person is stepping inside such a 9 m^2 area the four sensors will sense the vibration signal and we can such determine the 9 m^2 area in which a footstep occurred. Using an AI-approach we can then determine the position of each footstep inside the 9 m^2 area to get a precision of 1m in both directions. This way we can sense and track footsteps across the entire factory store.

An advantage of our approach is that we can concentrate on solving the positioning problem for the 9 m^2 area and our AI has to process 4 sensor signals only. However, our system can then be scaled out to larger areas.

Our methodology is privacy-friendly in contrast to video surveillance. Furthermore, our approach works in dark or heavily smoked areas, too. Thus, our approach is ideal to track the movements of persons especially in the case of an emergency.

C. DATA ACQUISITION

Fig. 4 shows our data acquisition devices setup. We used four EPZ-27MS44W piezoelectric sensors to detect the vibration signal. This sensor can detect a frequency band ranging from 0 Hz to 4400 Hz. As shown in Fig. 2, we established a plane rectangular coordinate system on the ground. The four sensors S1, S2, S3, S4 were deployed according to Fig.2. and the corresponding signals recorded are labeled as $s_1(t)$, $s_2(t)$, $s_3(t)$, $s_4(t)$.

A square steel plate of 10 cm by 10 cm is pressed onto each sensor to guarantee that ground waves cause pressure on the sensor from below due to the steel plate's inertia above. In the user's actual application environment, the sensor can be embedded in the floor tiles to ensure that the sensing device will not affect normal walking. The signal was amplified, sampled, quantized, and recorded with a R&S-RTB2000 oscilloscope. The sampling rate is 23.8 kHz, which is more than double the sensor's maximum bandwidth. Hence, a higher sampling rate would not improve the results anymore. Each sampling point's possible maximum value is manually set to 0.1 V, and waveforms higher than 0.1 V are cut off.

The same person with the same shoes generated the ground vibration by walking on points evenly distributed in each zone. We repeated this procedure for all nine zones. This way, we can easily label the data since we know how samples and zones are related. The recorded data includes four columns, which refer to the S1, S2, S3, S4 signals, and the values in each row are the sampled data values of which the sampling period is 0.042 ms. Our approach is time-based, and therefore all signals must be synchronized. We achieved this by sampling all four signals with one oscilloscope.

D. PREPROCESSING

It is necessary to detect the starting point of the effective signal when there is vibration. After analyzing the existing threshold-based method [31], [32], a customized shift window with a grouped frame threshold-based method is used in this research. The sampling points are grouped into windows of 32 points, which means that each window covers 1.344 ms. When the maximum of the absolute value of the median of any channel in a group becomes bigger than the threshold, the first sampling point in the group before this group will be considered as the starting point of the signal segment. The group with the starting point is named the activated group. We chose the previous group of the activated group to guarantee all the relevant signal sampling points are taken into consideration by the system. The trigger threshold is $V_t = 25mv$. Starting from the activated group, four consecutive groups are grouped as a big group, which serves as a valid data frame. Furthermore, a valid data frame, totally with 128 sampling points, is served as one sample to the neural network.

One sample includes the waveform of 5.376 ms, which contains all the valid information for a one-time localization task for one shock. Considering that the step frequency time of elite professional sprinters is always longer than 200 ms [33], and the fact that the signal collected by the piezoelectric sensor will decay after 80ms, the data frame contains all the necessary information for one-time foot shock without any signal caused by adjacent steps. Meanwhile, after one data frame is detected, the next frame will be detected after 224 groups of about 300ms. There are no coincident points between the adjacent data frames.

After we got the data samples for an event, the valid data frame should be normalized before training the deep neural network. As the maximum value of the signal envelope is 0.1V, we scaled all the points from the four channels by dividing the values by 0.1 to guarantee all the values are between -1 and 1. After scaling, we got the training samples. We do not conduct the denoising in the preprocessing procedure and still got good results, as shown in the experiment and evaluation section. Deep neural networks can, to some extent, tolerate noise. This result is consistent with the literature [15].

E. CLASSIFIER DESIGN AND DEEP LEARNING MODEL TRAINING

Considering that existing works [15], [34], [35] suggest that deep learning is more suitable to analyze time-series of sensor data, our approach is based on CNNs as well. The works [34] compared CNN, LSTM, perceptron, and random forest and conclude that CNN provides the best results. The existing works [35] use the same sensors on a pattern recognition task and compare a deep learning approach with random forest

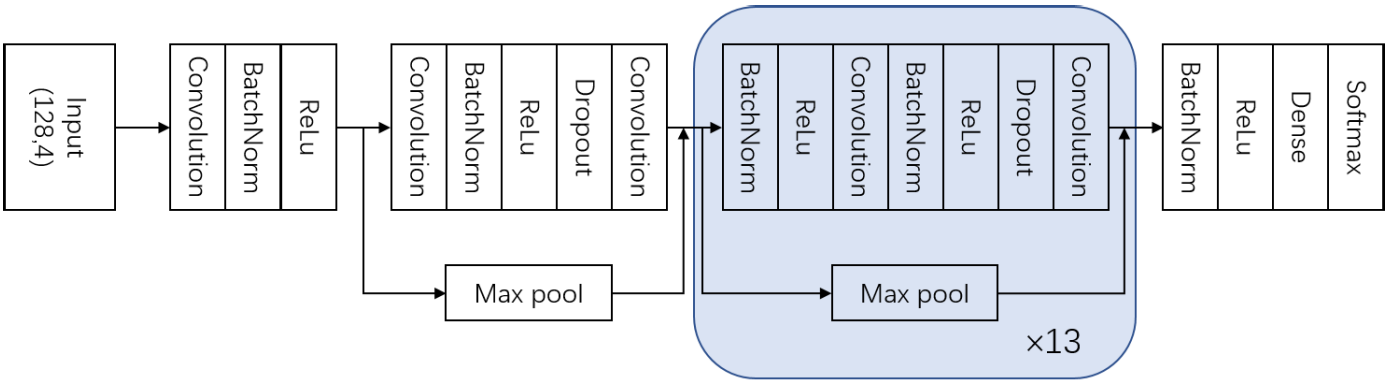


FIGURE 3. Deep neural network architecture. Our DNN model comprises 29 convolution layers, followed by a linear output layer that ended with a softmax. The network accepts packaged vibration signal as input(sampled at 23800Hz, or 23800 samples per second). The output of the DNN is a probability vector, which is the prediction of one of 9 possible location-related classes every consequent 128 points. The residual structure is utilized to gain accuracy from the increasing depth of the neural network.

classifiers in the experiment. The results also show that the deep learning approach works better than the classical classifier random forest. Also, deep residual networks show good performance in the fixed-length series pattern recognition task [15]. Therefore, we excluded classic classifiers and used a CNN-based deep learning approach instead.

We collected a data set, marked with "Data Part 1", and split it into a training set for training model and validation set for model hyperparameters optimization. We finally determined the best model architecture and hyperparameters, as shown in this paper.

Fig. 3 shows the architecture of the deep neural network designed in this experiment, which presents a good performance in the classification task, which we discuss in the next section.

Our deep neural network consists of 29-dimensional convolution layers followed by a linear output layer into a softmax layer. The network accepts packaged effective vibration signal data as input and outputs a prediction of one out of 9 possible location-related classes every one input sample.

The first convolution layer in the deep neural network contains 48 filters, and the number of filters used in the convolution layer increases according to the increasing of the order of the layer. From the 2nd convolution layer, the filter number doubles every eight convolution layers, and finally, in the last convolution layer, 384 filters have been used. Due to the convolution operations and the max pooling operations, the data frame's length becomes the half after the 3rd convolution layer. From the 4th convolution layer, the data frame length becomes half of the output length of the previous layer. Finally, the size of the data frame becomes 1 row and 384 columns. The dropout rate was set to 0.5, learning rate 0.0005, and batch size 32. The 1D convolution kernel size is 16 [36]. The initial stopping criteria are introduced to restrain the over-fitting issues, which are based on performance validation. After eight consequent training epochs have been conducted without improvement in the performance validation result, the training procedure has been stopped.

To objectively reflect the performance of our approach, four-fold cross-validation has been conducted on totally new

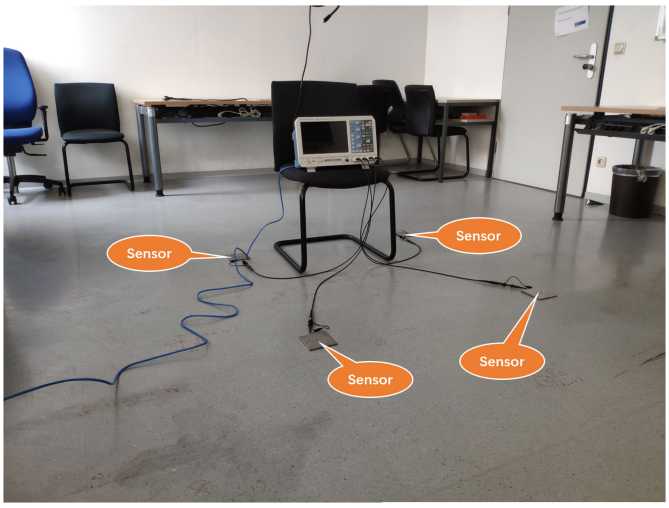


FIGURE 4. Experiment setup.

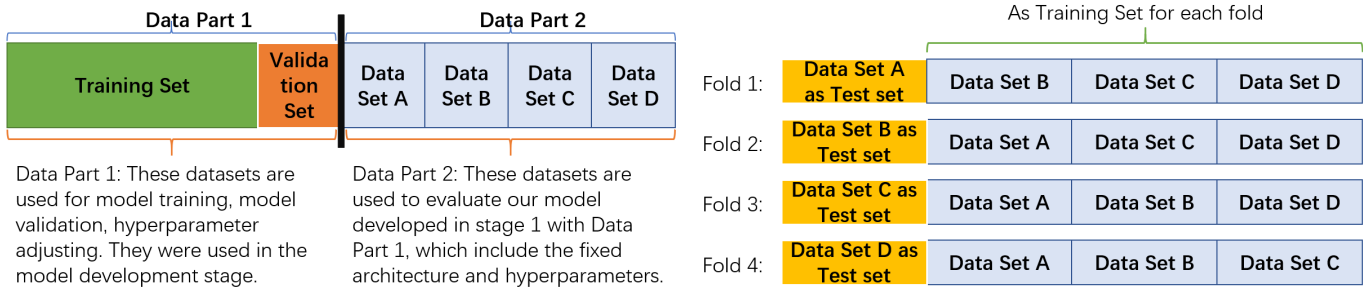
data, marked as data part 2, as shown in Fig. 5(a). We discuss this in detail in the next section.

IV. EXPERIMENT AND EVALUATION

In this section, firstly, we discuss how we organized the data for the evaluation. Secondly, we show the location-related classification performance. The ROC figure and the AUC index are used to represent the classification performance. Then the error analysis was conducted, and the results are shown as in Table. 2. In the error analysis part, Euler distance and Manhattan distance are introduced. We calculate the error distance with each fold, and the average values for the four folds evaluation procedures were calculated, as shown in Table. 2.

A. DATA ORGANIZATION

To evaluate our approach, model architecture, and hyperparameters, we conducted four experiments to get four data sets, marked as "Data Part 2", as shown in Fig. 5(a). The "Data part 2", including data set A, data set B, data set C, and data set D, a total of 30576 samples, are from four independent experiments A, B, C, and D. This "Data Part 2" has never been used in the model hyperparameter optimization procedure.



(a) Data Sets of 2 Parts. (b) 4-Folds Cross Validation

FIGURE 5. As shown in Fig. 5, we totally collected two parts of data sets. For data set part 1, we randomly shuffle the data samples and split them into a training set and a validation set according to a rate of 4:1. We developed our model with the data set part 1, finally fixing the architecture and the hyperparameters. We evaluate our approach, including the developed model architecture and hyperparameters with data set part 2. We conducted four independent experiments: experiment A, experiment B, experiment C, experiment D, and collected the data set A, data set B, data set C, data set D individually. We conducted a "4-folds cross-validation" evaluation, as shown in Fig. 5(b). The evaluation results are shown in Chapter IV.

In the evaluation, all data samples of "Data Part 2" were used for four-fold cross-validation. The typical k-folds cross-validation method [37], [38] splits the entire data set into k mutually exclusive subsets as folds with approximately equal size. Every time, one set of data is used as the test set and the rest as the training set. Our evaluation procedure made minor adjustments to the typical cross-validation method and implemented it as a variant. We independently collected four sets of data in four times. We did not merge all the 4 data sets or randomly split it into folds. Each test set in each fold can be considered as online data, and the training set as offline training data. In this way, it is even more challenging for our system to make good classification results, because no data point of the measurement session has been used for training. In comparison, imagine that we shuffled all data of "Data Part 2" and imagine we performed four-fold cross-validation on this shuffled data (which is the default procedure for cross-validation). In this case, each training set contains data points from all four measurement sessions. Thus, CNN is trained with data from all measurement sessions and would have an easier time to classify the test set, which contains data from all measurement sessions. In our paper, however, CNN is trained with three measurement sessions and must classify data from another measurement session that it has not seen during training. Thus, our cross-validation is even more rigorous than the default approach, and the evaluation experiments can better simulate the system's situation in actual use, that it can objectively show that our approach works.

As shown in Fig. 5(b), for evaluating purpose, we trained the model four times and then tested each model instance. The "Data Part 2" has never been used in the model creation procedure. In the "four-fold evaluation", all the data samples have been used for training and testing. The values of each evaluation metric were summed up, and the average values of them were calculated. Precision, recall, F1-score were used as evaluation metrics, as shown in Table. 1.

B. CLASSIFICATION PERFORMANCE ANALYSIS

Table. 1, Fig. 6 and Fig. 7 show the classification performance presented by our DNN model. Precision, recall, and F1-score [40] are used as metrics for the evaluation and analysis approach. Intuitively, the precision represents the classifier's ability not to label as positive a sample that is negative. The recall shows the ability of the classifier to find all the positive samples. The F1 score can be interpreted as a weighted harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and the worst score at 0.

TABLE 1. Classification performance of the DNN.

	Precision	Recall	F1-score
G1	0.9465	0.9479	0.9471
G2	0.9066	0.9362	0.9209
G3	0.9202	0.9211	0.9206
G4	0.9684	0.9616	0.9650
G5	0.9609	0.9628	0.9617
G6	0.9457	0.9136	0.9291
G7	0.9741	0.9696	0.9718
G8	0.9644	0.9595	0.9618
G9	0.9371	0.9508	0.9437
Micro Average	0.9472	0.9472	0.9472
Macro Average	0.9471	0.9470	0.9469

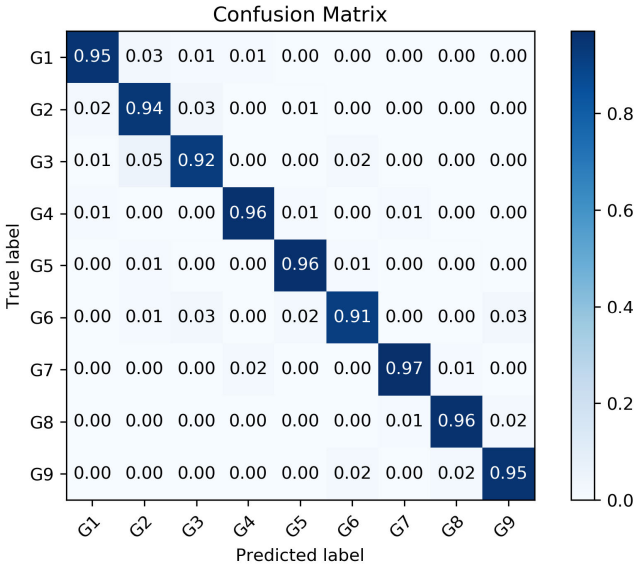
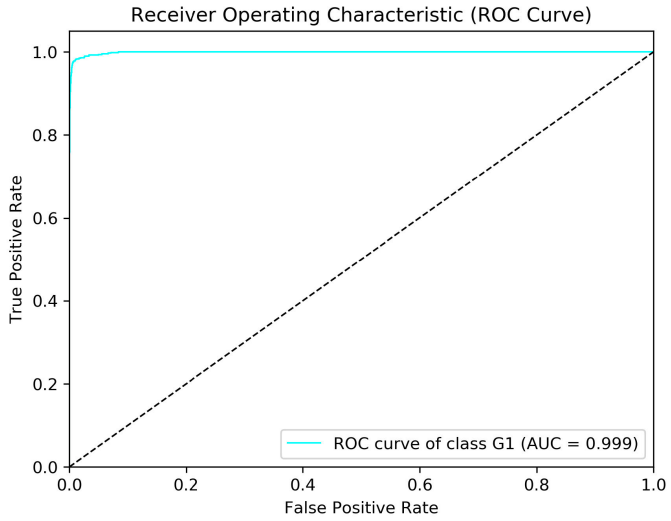
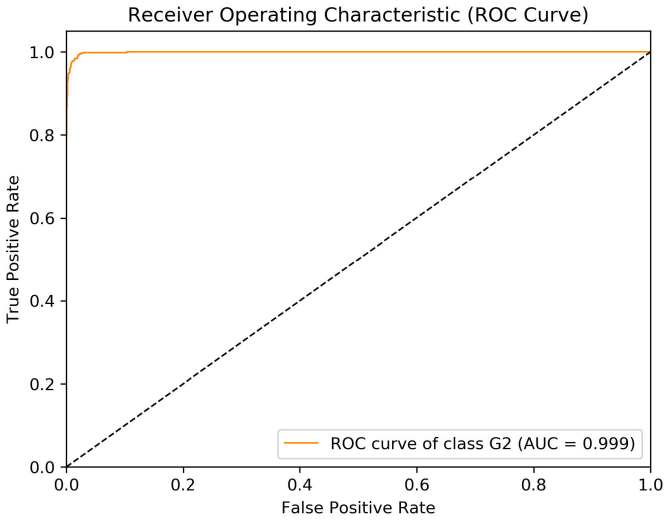


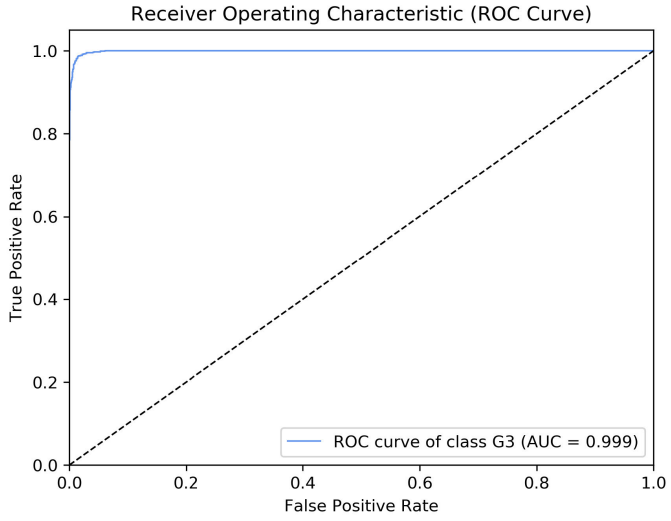
FIGURE 6. Confusion matrix. Fig. 6 is the confusion matrix for the location-related class prediction of the classifier versus the sample labels. As part of the four-fold cross-validation, this confusion matrix's values are the average of the corresponding values from the four tests.



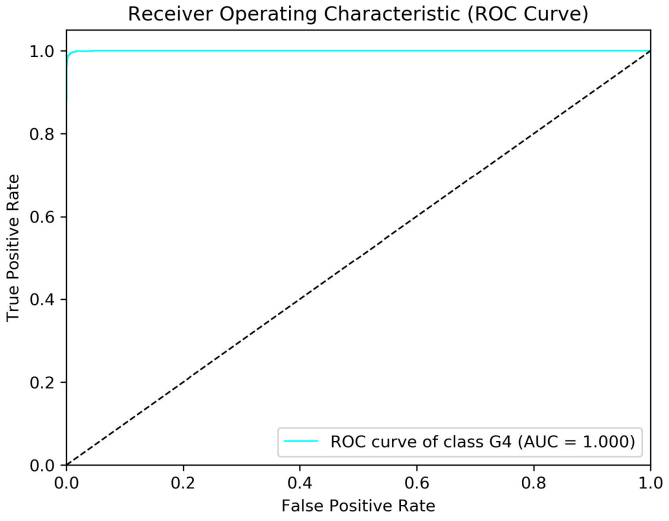
(a) ROC G1.



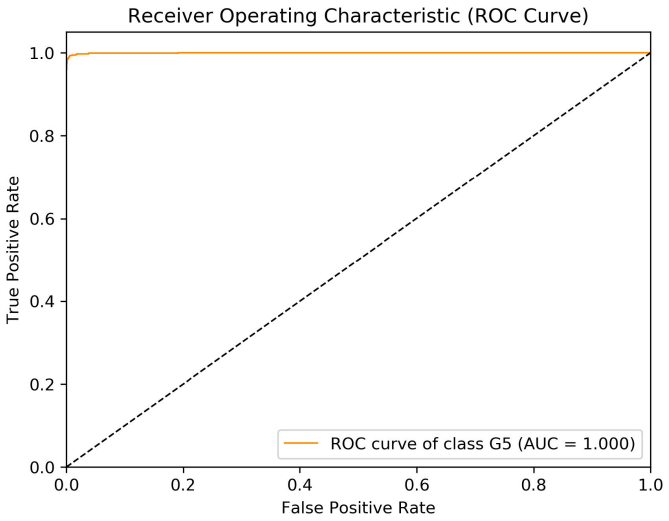
(b) ROC G2.



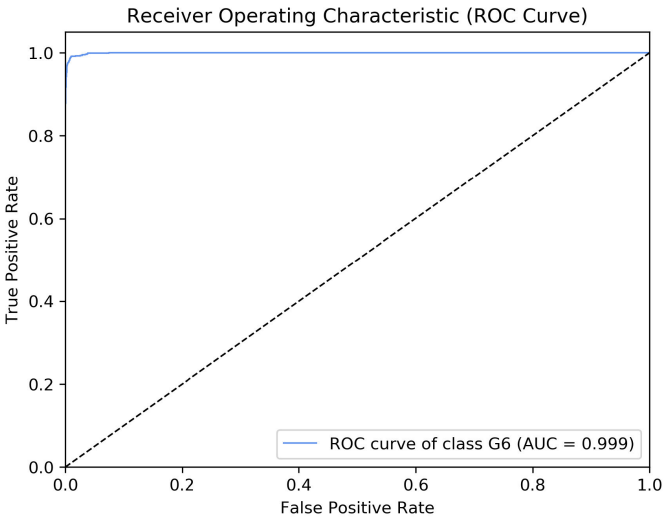
(c) ROC G3.



(d) ROC G4.

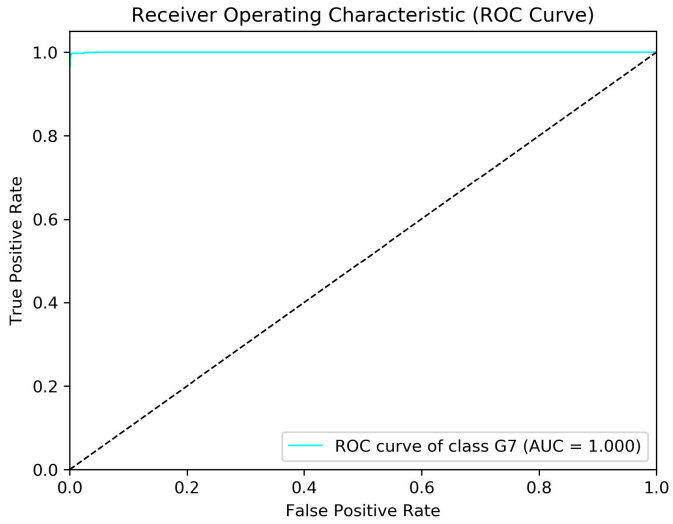


(e) ROC G5.

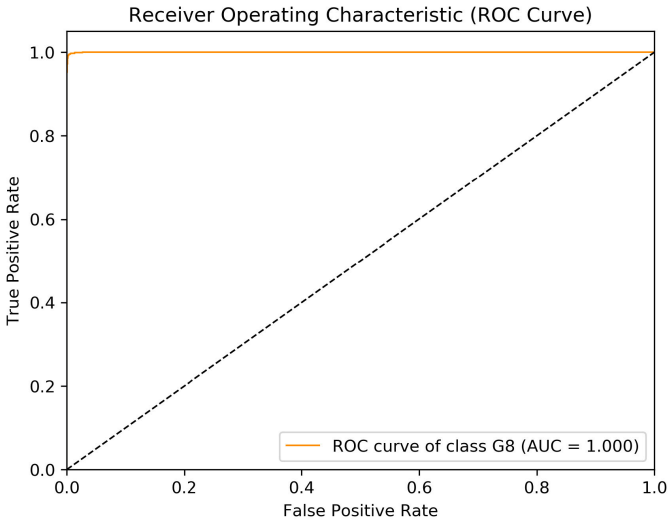


(f) ROC G6.

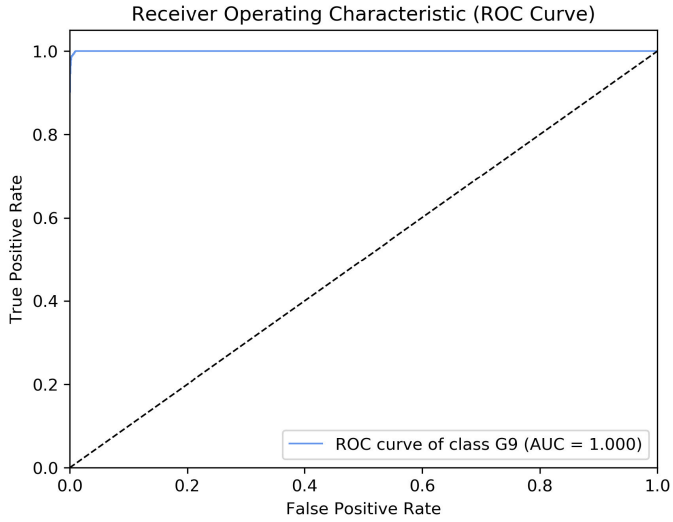
FIGURE 7. Fig. 7(a) to Fig. 7(k) are the ROC curves for deep neural network predictions on nine location-related classes. The macro-average ROC curve and micro-average ROC curve [39] reflect the multi-class classifier performance.



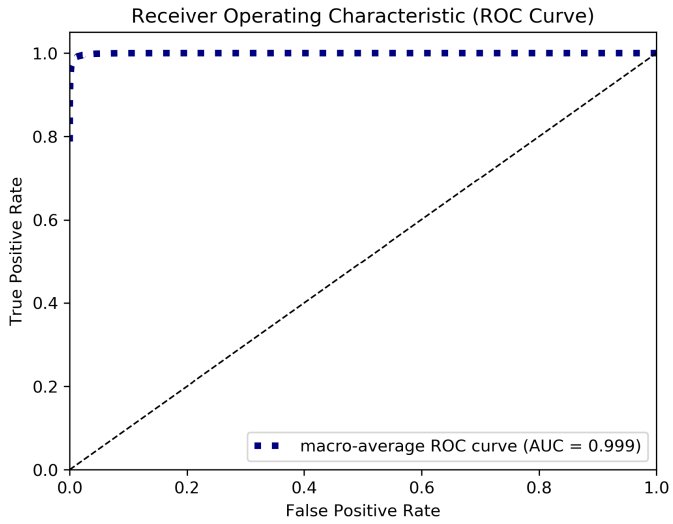
(i) ROC G7.



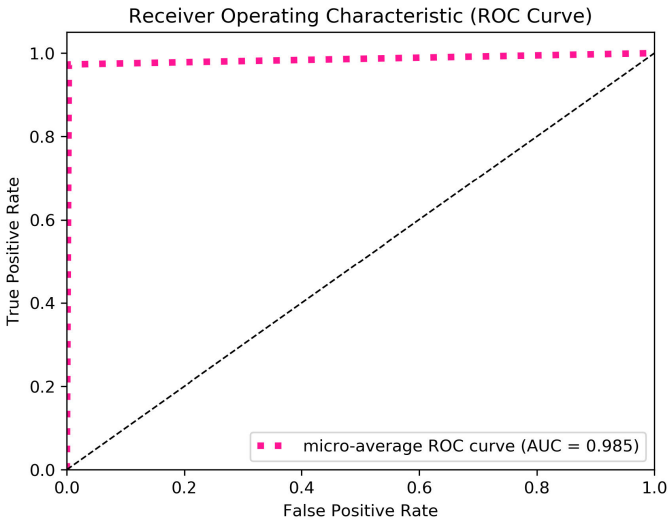
(j) ROC G8.



(k) ROC G9.



(l) ROC macro.



(m) ROC micro.

FIGURE 7. ***(Continued)*** **Fig. 7(a) to Fig. 7(k) are the ROC curves for deep neural network predictions on nine location-related classes. The macro-average ROC curve and micro-average ROC curve [39] reflect the multi-class classifier performance.**

The F1 score in our evaluation considered recall and precision as equally important.

$$Micro - average\ Precision = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FP_c} \quad (1)$$

$$Micro - average\ Recall = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FN_c} \quad (2)$$

The metrics used in Table. 1 include the macro average of the metrics and the micro average of the metrics. Table. 1 shows the metrics precision, recall, and F1-score. The macro average precision and the macro average recall are the arithmetic average of each class's precision and recall. The macro average F1-score is the harmonic mean of macro average precision and macro average recall. Micro average precision and micro average recall are as described in the formulas (1) and (2). In formula (1) and formula (2), TP means true positive; FP means false positive; FN means false negative; c is the class label. The micro average F1-score is the harmonic mean of micro-average of precision and micro-average of recall, which means averaging the unweighted mean per label [41], [42]. The micro and macro average performance metrics reflect the system's average level of performance concerning each class. In our evaluation setting, the number of instances in each class is similar. So the micro average figures are similar to the macro average figures.

Fig. 7 shows the ROC curves of the classifier. A ROC curve is a visual indicator of the trade-off between true-positive and false-positive cases. The area under the ROC curve (AUC) will always be between 0 and 1. According to the ROC curves in Fig. 7, our classifier shows outstanding performance.

Fig. 6 shows that most of the evaluation samples were correctly classified. Considering the practical walking people location scenario, the positioning-related grid can be predicted correctly in most cases. 5% of G3 samples have been classified as G2, which is the highest classification error. Meanwhile, 3% G1 were predicted as G2; 3% G2 samples were predicted as G3; 3% G6 samples are predicted as G3; 3% G3 samples were predicted as G6; 3% G6 samples were predicted as G9. However, all the above mentioned misclassified samples are location-adjacent. As G8 is next to G5, G3 is next to G2. That is to say, although some samples are wrongly classified, in most cases, the wrongly classified results are still with practical meaning. The remaining misclassified samples are less than 2% of the total sample number of its class.

C. ERROR DISTANCE ANALYSIS

From the perspective of walking people positioning, the error distance analysis has been conducted, as shown in Table. 2. For error distance analysis, Euler distance and Manhattan distance [43] were counted.

In Table. 2, the error Euler distance and the error Manhattan distance of a sample, which wrongly classified G_x as G_y, is the Euler distance and Manhattan distance of the geometric center point of G_x to the Euler distance and the Manhattan distance of the geometric center point of G_y respectively. The unit is meter. The average error Euler distance and the average error Manhattan distance are 0.0560 0.0596, respectively, and the medians of that respectively are 0 and 0. The error distance evaluation results show that the proposed system can implement localization function with good performance and low error.

TABLE 2. Error distance description of misclassified samples.

	Error Euler Distance	Error Manhattan Distance
Mean	0.059	0.062
Standard Deviation	0.255	0.280
Min	0	0
Median	0	0
Max	2.532	3.500

V. CONCLUSION

We presented a novel approach to determine the walking persons' position by analyzing ground vibrations caused by individual footsteps. Due to the anisotropic nature of the ground waves, we used a machine learning approach to determine a footstep position. We divided a $9m^2$ area into nine squares of $1m^2$ size. The learned classifier can assign a ground vibration signal to one of these nine areas. The mean Euler error of our approach, as defined in our paper, is 0.059m.

Our approach can identify the position of an individual step. Therefore, for a walking person, we can identify their position and the direction of their movement. This is especially useful in production plants, where safety demands that firefighters know where people are and in which direction they are running. In contrast to video surveillance, our approach works in smoked areas. Furthermore, our approach is less privacy-invasive than cameras. Since many companies are not allowed to install always-on cameras for privacy reasons, our approach can deliver more safety without invading workers' privacy.

ACKNOWLEDGMENT

The author would like to thank Evonik Digital to access their production plant for ground wave measurements.

REFERENCES

[1] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2568–2599, 3rd Quart., 2019.

[2] A. Serra, D. Carboni, and V. Marotto, "Indoor pedestrian navigation system using a modern smartphone," in *Proc. 12th Int. Conf. Hum. Comput. Interact. With Mobile Devices Services - MobileHCI*, 2010, pp. 397–398.

[3] A. E. Fano, "Shopper's eye: Using location-based filtering for a shopping agent in the physical world," in *Proc. 2nd Int. Conf. Auton. Agents*. New York, NY, USA: Association for Computing Machinery, 1998.

[4] K.-H. Park, Z. Bien, J.-J. Lee, B. K. Kim, J.-T. Lim, J.-O. Kim, H. Lee, D. H. Stefanov, D.-J. Kim, J.-W. Jung, J.-H. Do, K.-H. Seo, C. H. Kim, W.-G. Song, and W.-J. Lee, "Robotic smart house to assist people with movement disabilities," *Auto. Robots*, vol. 22, no. 2, pp. 183–198, Jan. 2007.

[5] Y. Tajika, T. Saito, K. Teramoto, N. Oosaka, and M. Isshiki, "Networked home appliance system using Bluetooth technology integrating appliance control/monitoring with Internet service," *IEEE Trans. Consum. Electron.*, vol. 49, no. 4, pp. 1043–1048, Nov. 2003.

[6] J. Scott, A. J. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, and N. Villar, "PreHeat: Controlling home heating using occupancy prediction," in *Proc. 13th Int. Conf. Ubiquitous Comput. - UbiComp*, 2011, pp. 281–290.

[7] M. Mirshekari *et al.*, "Characterizing wave propagation to improve indoor step-level person localization using floor vibration," in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems* (International Society for Optics and Photonics), vol. 9803. Las Vegas, NV, USA: SPIE, 2016. Accessed: Jan. 6, 2020. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9803/980305/Characterizing-wave-propagation-to-improve-indoor-step-level-person-localization/10.1117/12.2222136.short

[8] X. Tang, M.-C. Huang, and S. Mandal, "An 'Internet of ears' for crowd-aware smart buildings based on sparse sensor networks," in *Proc. IEEE Sensors*, Oct. 2017, pp. 1–3.

[9] W. Chen, M. Guan, L. Wang, R. Ruby, and K. Wu, "FLoc: Device-free passive indoor localization in complex environments," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[10] J. D. Poston, R. M. Buehrer, and V. Tech, "Vibration sensing in smart buildings," in *Proc. Int. Conf. Indoor Position. Indoor Navig. (IPIN)*, 2015, p. 4.

[11] P. Blank, T. Kautz, and B. M. Eskofier, "Ball impact localization on table tennis rackets using piezo-electric sensors," in *Proc. ACM Int. Symp. Wearable Comput.*, New York, NY, USA: Association for Computing Machinery, 2016, pp. 72–79, doi: 10.1145/2971763.2971778.

[12] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[15] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, no. 1, p. 65, 2019.

[16] Z. Wang, B. Du, and Y. Guo, "Domain adaptation with neural embedding matching," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–11, Sep. 2019.

[17] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107173.

[18] F. Luo, H. Huang, Y. Duan, J. Liu, and Y. Liao, "Local geometric structure feature for dimensionality reduction of hyperspectral imagery," *Remote Sens.*, vol. 9, no. 8, p. 790, Aug. 2017. [Online]. Available: https://www.mdpi.com/2072-4292/9/8/790

[19] G. Shi, H. Huang, and L. Wang, "Unsupervised dimensionality reduction for hyperspectral imagery via local geometric structure feature learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1425–1429, Aug. 2020.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[21] F. Capezio, A. Sgorbissa, and R. Zaccaria, "GPS-based localization for a surveillance UGV in outdoor areas," in *Proc. 5th Int. Workshop Robot Motion Control RoMoCo*, Jun. 2005, pp. 157–162.

[22] W. Kang and Y. Han, "SmartPDR: Smartphone-based pedestrian dead reckoning for indoor localization," *IEEE Sensors J.*, vol. 15, no. 5, pp. 2906–2916, May 2015.

[23] C. Fischer, K. Muthukrishnan, M. Hazas, and H. Gellersen, "Ultrasound-aided pedestrian dead reckoning for indoor navigation," in *Proc. 1st ACM Int. Workshop Mobile Entity Localization Tracking GPS-Less Environments MELT*, 2008, pp. 31–36.

[24] S. Beauregard and H. Haas, "Pedestrian dead reckoning: A basis for personal positioning," in *Proc. 3rd Workshop Positioning, Navigat. Commun.*, 2006, pp. 27–35.

[25] B. Wang, X. Liu, B. Yu, R. Jia, and X. Gan, "An improved WiFi positioning method based on fingerprint clustering and signal weighted Euclidean distance," *Sensors*, vol. 19, no. 10, p. 2300, May 2019.

[26] S. Xia, Y. Liu, G. Yuan, M. Zhu, and Z. Wang, "Indoor fingerprint positioning based on Wi-Fi: An overview," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 5, p. 135, Apr. 2017.

[27] H.-S. Kim, D.-R. Kim, S.-H. Yang, Y.-H. Son, and S.-K. Han, "An indoor visible light communication positioning system using a RF carrier allocation technique," *J. Lightw. Technol.*, vol. 31, no. 1, pp. 134–144, Jan. 2013.

[28] J. Clemente, W. Song, M. Valero, F. Li, and X. Liy, "Indoor person identification and fall detection through non-intrusive floor seismic sensing," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2019, pp. 417–424.

[29] F. Li, J. Clemente, M. Valero, Z. Tse, S. Li, and W. Song, "Smart home monitoring system via footstep-induced vibrations," *IEEE J. Mag.*, vol. 14, no. 3, pp. 3383–3389, Sep. 2020, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8840889/versions

[30] Y. Kashimoto, M. Fujimoto, H. Suwa, Y. Arakawa, and K. Yasumoto, "Floor vibration type estimation with piezo sensor toward indoor positioning system," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Oct. 2016, pp. 1–6.

[31] S. Pan, A. Bonde, J. Jing, L. Zhang, P. Zhang, and H. Y. Noh, "BOES: Building occupancy estimation system using sparse ambient vibration monitoring," *Proc. SPIE*, vol. 9061, Apr. 2014, Art. no. 90611O.

[32] M. Lam, M. Mirshekari, S. Pan, P. Zhang, and H. Y. Noh, "Robust occupant detection through step-induced floor vibration by incorporating structural characteristics," in *Dynamics of Coupled Structures*, vol. 4. New York, NY, USA: Springer, 2016, pp. 357–367.

[33] R. F. Chapman, A. S. Laymon, D. P. Wilhite, J. M. Mckenzie, D. A. Tanner, and J. M. Stager, "Ground contact time as an indicator of metabolic cost in elite distance runners," *Med. Sci. Sports Exercise*, vol. 44, no. 5, pp. 917–925, May 2012.

[34] V. Matkovic, M. Waltereit, P. Zdankin, M. Uphoff, and T. Weis, "Bike type identification using smartphone sensors," in *Proc. Adjunct Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, Sep. 2019, pp. 145–148.

[35] Y. Yu and T. Weis, "A privacy-protecting indoor emergency monitoring system based on floor vibration," in *Proc. Adjunct Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, Sep. 2020, pp. 164–167.

[36] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, Nov. 2016.

[37] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. London, U.K.: Pearson, 2009.

[38] G. James *et al.*, *An Introduction to Statistical Learning*, vol. 112. New York, NY, USA: Springer, 2013.

[39] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[40] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn. - ICML*, 2006, pp. 233–240.

[41] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Berlin, Germany: Springer, 2004.

[42] H. Narasimhan, W. Pan, P. Kar, P. Protopapas, and H. G. Ramaswamy, "Optimizing the multiclass F-Measure via biconcave programming," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 1101–1106.

[43] S. Craw, "Manhattan distance," in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds. Boston, MA, USA: Springer, 2017, pp. 790–791.

YANG YU received the B.Eng. degree in communication engineering from Zhengzhou University and the M.Sc. degree in embedded system engineering from the University of Duisburg-Essen, Duisburg, Germany, where he is currently pursuing the Ph.D. degree with the Distributed Systems Research Group. His research interests include AI-based context recognition, AI-based context reasoning, machine learning, and explainable AI.

MARIAN WALTEREIT received the M.Sc. degree in computer science from the University of Duisburg-Essen, Duisburg, Germany, where he is currently pursuing the Ph.D. degree with the Distributed Systems Research Group. His research interest includes privacy and digital forensics in automotive systems.

VIKTOR MATKOVIC received the M.Sc. degree in computer science from the University of Duisburg-Essen, Duisburg, Germany, where he is currently pursuing the Ph.D. degree with the Distributed Systems Research Group. His research interest includes bike-aware pervasive applications, such as bike type-related navigation services.

WEIYAN HOU received the bachelor's and master's degrees, in China, in 1986 and 1998, respectively, and the Ph.D. degree from Shanghai University, in a Sandwich Ph.D. Program between Germany and China, in 2004. He is currently a Professor with the School of Information Engineering, Zhengzhou University, China. His research interests include network control, time performance modeling of the industrial IoT, and infrastructure of wired/wireless communication integration.

TORBEN WEIS received the Ph.D. degree in computer science from the Technical University of Berlin, Berlin, Germany. He is currently a Professor with the University Duisburg-Essen, Duisburg, Germany, where he leads the Distributed Systems Research Group. His research interests include cyber-physical systems, cloud computing, and security and privacy in distributed systems.

Bibliography

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[4] Lijun Wang, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. Visual Tracking With Fully Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3119–3127, 2015.

[5] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1925–1934, 2017.

[6] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624, February 2021.

[7] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016.

[8] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, October 2017.

[9] M.R. Endsley. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, pages 789–795 vol.3, May 1988.

[10] Mica R. Endsley. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors*, 37(1):32–64, March 1995.

[11] Y. Yu, M. Waltereit, V. Matkovic, W. Hou, and T. Weis. Deep Learning-Based Vibration Signal Personnel Positioning System. *IEEE Access*, 8:226108–226118, 2020.

[12] Yang Yu and Torben Weis. A privacy-protecting indoor emergency monitoring system based on floor vibration. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp-ISWC '20, pages 164–167, New York, NY, USA, September 2020. Association for Computing Machinery.

[13] Yang Yu, Xiangju Qin, Shabir Hussain, Weiyan Hou, and Torben Weis. Pedestrian Counting Based on Piezoelectric Vibration Sensor. *Applied Sciences*, 12(4):1920, January 2022.

[14] Mica R. Endsley. Design and Evaluation for Situation Awareness Enhancement. *Proceedings of the Human Factors Society Annual Meeting*, 32(2):97–101, October 1988.

[15] Francis T Durso and Scott D Gronlund. Chapter 10 Situation Awareness. page 33.

[16] Hooman Alavizadeh, Julian Jang-Jaccard, Simon Yusuf Enoch, Harith Al-Sahaf, Ian Welch, Seyit A. Camtepe, and Dan Dongseong Kim. A Survey on Cyber Situation Awareness Systems: Framework, Techniques, and Insights. *ACM Computing Surveys*, April 2022.

[17] Mica R. Endsley. Measurement of Situation Awareness in Dynamic Systems. *Human Factors*, 37(1):65–84, March 1995.

[18] João Lopes, Rodrigo Souza, Cláudio Geyer, Cristiano Costa, Jorge Luis Victória Barbosa, Ana Pernas, and Adenauer Yamin. A Middleware Architecture for Dynamic Adaptation in Ubiquitous Computing. *JUCS - Journal of Universal Computer Science*, 20(9):1327–1351, September 2014.

[19] Ricardo Borges Almeida, Victor Renan Covalski Junes, Roger da Silva Machado, Diórgenes Yuri Leal da Rosa, Lucas Medeiros Donato, Adenauer Corrêa Yamin, and Ana Marilza Pernas. A distributed event-driven architectural model based on situational awareness applied on internet of things. *Information and Software Technology*, 111:144–158, July 2019.

[20] Alan F Westin. Privacy And Freedom. page 6.

[21] R. Shirey. Internet Security Glossary. Technical Report RFC2828, RFC Editor, May 2000.

[22] Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated, 1st edition, 2017.

[23] H.P Gassmann. OECD guidelines governing the protection of privacy and transborder flows of personal data. *Computer Networks (1976)*, 5(2):127–141, April 1981.

[24] Isabel Wagner and David Eckhoff. Technical Privacy Metrics: A Systematic Survey. *ACM Computing Surveys*, 51(3):57:1–57:38, June 2018.

[25] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA, October 2016. Association for Computing Machinery.

[26] Dr Prerna Mahajan and Abhishek Sachdeva. A Study of Encryption Algorithms AES, DES and RSA for Security. *Global Journal of Computer Science and Technology*, December 2013.

[27] Lorenz Schwittmann. *Privacy Threats in the Mobile Web & Social Media*. PhD thesis, DuEPublico: Duisburg-Essen Publications online, University of Duisburg-Essen, Germany, June 2019.

[28] Cynthia Dwork. Differential Privacy: A Survey of Results. In Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li, editors, *Theory and Applications of Models of Computation*, Lecture Notes in Computer Science, pages 1–19, Berlin, Heidelberg, 2008. Springer.

[29] Tore Dalenius. Towards a methodology for statistical disclosure control. *statistik Tidskrift*, 15(429-444):2–1, 1977.

[30] Yun Li, Ningfeng Li, and Yiyu Xia. Research on a Data Desensitization Algorithm of Blockchain Distributed Energy Transaction Based on Differential Privacy. In *2019 IEEE 8th International Conference on Advanced Power System Automation and Protection (APAP)*, pages 980–985, 2019.

[31] Xinzao Jiang, Yubo Song, Rui Song, and Aiqun Hu. Data desensitization mechanism of Android application based on differential privacy. In *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, pages 1–5, September 2021.

[32] Arthur L Samuel. Some studies in machine learning using the game of checkers. page 13.

[33] Tom M. Mitchell. Machine learning, 1997.

[34] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised Learning. In Matthieu Cord and Pádraig Cunningham, editors, *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*, Cognitive Technologies, pages 21–49. Springer, Berlin, Heidelberg, 2008.

[35] H.B. Barlow. Unsupervised Learning. *Neural Computation*, 1(3):295–311, September 1989.

[36] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning, Second Edition: An Introduction*. MIT Press, November 2018.

[37] A. W. Roe, S. L. Pallas, Y. H. Kwon, and M. Sur. Visual projections routed to the auditory pathway in ferrets: Receptive fields of visual neurons in primary auditory cortex. *Journal of Neuroscience*, 12(9):3651–3664, September 1992.

[38] C. Métin and D. O. Frost. Visual responses of neurons in somatosensory cortex of hamsters with experimentally induced retinal projections to somatosensory thalamus. *Proceedings of the National Academy of Sciences*, 86(1):357–361, January 1989.

[39] Saskia K. Nagel, Christine Carl, Tobias Kringe, Robert Märtin, and Peter König. Beyond sensory substitution—learning the sixth sense. *Journal of Neural Engineering*, 2(4):R13–R26, 2005.

[40] Yuri Danilov and Mitchell Tyler. Brainport: An alternative input to the brain. *Journal of Integrative Neuroscience*, 04(04):537–550, 2005.

[41] Marvin Minsky and Seymour Papert. *Perceptrons*. Perceptrons. M.I.T. Press, Oxford, England, 1969.

[42] Albert Gidon, Timothy Adam Zolnik, Pawel Fidzinski, Felix Bolduan, Athanasia Papoutsi, Panayiota Poirazi, Martin Holtkamp, Imre Vida, and Matthew Evan Larkum. Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science*, 367(6473):83–87, 2020.

[43] David G. Kleinbaum and Mitchel Klein. *Logistic Regression*. Statistics for Biology and Health. Springer New York, New York, NY, 2010.

[44] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, Cambridge, Massachusetts, November 2016.

[45] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, December 1989.

[46] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[47] Aston Zhang, Zachary C Lipton, Mu Li, and Alexander J Smola. Dive into Deep Learning. page 1027.

[48] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors, July 2012.

[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[50] David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The Shattered Gradients Problem: If resnets are the answer, then what is the question? In *Proceedings of the 34th International Conference on Machine Learning*, pages 342–350. PMLR, July 2017.

[51] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, April 2015.

[52] Min Lin, Qiang Chen, and Shuicheng Yan. Network In Network. *arXiv:1312.4400 [cs]*, March 2014.

[53] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper With Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[54] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[55] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, August 1987.

[56] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, October 2012.

[57] Manjunath Jogin, Mohana, M S Madhulika, G D Divya, R K Meghana, and S Apoorva. Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning. In *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, pages 2319–2323, 2018.

[58] Vishwajeet Shukla and Mayank Singour. Multimodal Learning for Early Detection of Explosive Sounds using Relative Spectral Distribution. In *2020 Sensor Signal Processing for Defence Conference (SSPD)*, pages 1–5, September 2020.

[59] Annamaria Mesaros, Aleksandr Diment, Benjamin Elizalde, Toni Heittola, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. Sound Event Detection in the DCASE 2017 Challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(6):992–1006, June 2019.

[60] Turab Iqbal, Yong Xu, Qiuqiang Kong, and Wenwu Wang. Capsule Routing for Sound Event Detection. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2255–2259, September 2018.

[61] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, March 2017.

[62] Guodong Guo and S.Z. Li. Content-based audio classification and retrieval by support vector machines. *IEEE Transactions on Neural Networks*, 14(1):209–215, 2003.

[63] Justin Salamon and Juan Pablo Bello. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*, 24(3):279–283, March 2017.

[64] C. Clavel, T. Ehrette, and G. Richard. Events Detection for an Audio-Based Surveillance System. In *2005 IEEE International Conference on Multimedia and Expo*, pages 1306–1309, July 2005.

[65] Izabela Freire and José Jr. Gunshot detection in noisy environments. In *Anais de VII International Telecommunications Symposium*. Sociedade Brasileira de Telecomunicações, 2010.

[66] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 21–26, September 2007.

[67] Qiuqiang Kong, Yong Xu, Iwona Sobieraj, Wenwu Wang, and Mark D. Plumbley. Sound Event Detection and Time–Frequency Segmentation from Weakly Labelled Data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4):777–787, April 2019.

[68] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech 2019*, pages 2613–2617, September 2019.

[69] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 1, pages 113–116 vol.1, 2002.

[70] Chang-Hsing Lee, Jau-Ling Shih, Kun-Ming Yu, and Jung-Mau Su. Automatic Music Genre Classification using Modulation Spectral Contrast Feature. In *2007 IEEE International Conference on Multimedia and Expo*, pages 204–207, July 2007.

[71] Jinxi Guo, Tara N. Sainath, and Ron J. Weiss. A spelling correction model for end-to-end speech recognition. *arXiv:1902.07178 [cs, eess]*, February 2019.

[72] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*, pages 5036–5040. ISCA, October 2020.

[73] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S. Chang, K. Rao, and A. Gruenstein. Streaming End-to-end Speech Recognition for Mobile Devices. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6381–6385, May 2019.

[74] Hyungui Lim, Jeongsoo Park, Kyogu Lee, and Yoonchang Han. RARE SOUND EVENT DETECTION USING 1D CONVOLUTIONAL RECURRENT NEURAL NETWORKS. page 6, 2017.

[75] S Amiriparian, S Julka, N Cummins, and B Schuller. Deep Convolutional Recurrent Neural Network for Rare Acoustic Event Detection. page 4, 2018.

[76] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[77] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(01):172–186, January 2021.

[78] Nicolai Wojke and Alex Bewley. Deep Cosine Metric Learning for Person Re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 748–756, March 2018.

[79] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, September 2017.

[80] Weiming Chen, Zijie Jiang, Hailin Guo, and Xiaoyang Ni. Fall Detection Based on Key Points of Human-Skeleton Using OpenPose. *Symmetry*, 12(5):744, May 2020.

[81] Unity Technologies. Unity Real-Time Development Platform — 3D, 2D VR & AR Engine. https://unity.com/.

[82] Zengmao Wang, Bo Du, and Yuhong Guo. Domain Adaptation With Neural Embedding Matching. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11, 2019.

[83] Stylianos Mystakidis. Metaverse. *Encyclopedia*, 2(1):486–497, March 2022.

[84] Matthew Sparkes. What is a metaverse. *New Scientist*, 251(3348):18, August 2021.

Refereed Papers

1. Yang Yu and Torben Weis. A privacy-protecting indoor emergency monitoring system based on floor vibration. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp-ISWC '20, pages 164–167, New York, NY, USA, September 2020. Association for Computing Machinery.

2. Y. Yu, M. Waltereit, V. Matkovic, W. Hou, and T. Weis. Deep Learning-Based Vibration Signal Personnel Positioning System. *IEEE Access*, 8:226108–226118, 2020.

3. Yang Yu, Xiangju Qin, Shabir Hussain, Weiyan Hou, and Torben Weis. Pedestrian Counting Based on Piezoelectric Vibration Sensor. *Applied Sciences*, 12(4):1920, January 2022.

4. Yang Yu, Oskar Carl, Shabir Hussain, Weiyan Hou, and Torben Weis. A Privacy-Protecting Step-Level Walking Direction Detection Algorithm based on Floor Vibration. *IEEE Sensors Journal*, Accepted.

5. Shabir Hussain, Yang Yu, Muhammad Ayoub, Akmal Khan, Rukhshanda Rehman, Junaid Abdul Wahid, and Weiyan Hou. IoT and Deep Learning Based Approach for Rapid Screening and Face Mask Detection for Infection Spread Control of COVID-19. *Applied Sciences*, 11(8):3495, January 2021.

6. Shabir Hussain, Muhammad Ayoub, Ghulam Jilani, Yang Yu, Akmal Khan, Junaid Abdul Wahid, Muhammad Farhan Ali Butt, Guangqin Yang, Dietmar P.F. Moller, and Hou Weiyan. Aspect2Labels: A novelistic decision support system for higher educational institutions by using multi-layer topic modelling approach. *Expert Systems with Applications*, 209:118119, December 2022.

7. Oskar Carl, Peter Zdankin, Matthias Schaffeld, Viktor Matkovic, Yang Yu, Timo Elbers, and Torben Weis. Persistent Streams: The Internet With Ephemeral Storage. 2021.

DOI: 10.17185/duepublico/78270
URN: urn:nbn:de:hbz:465-20230420-084528-8