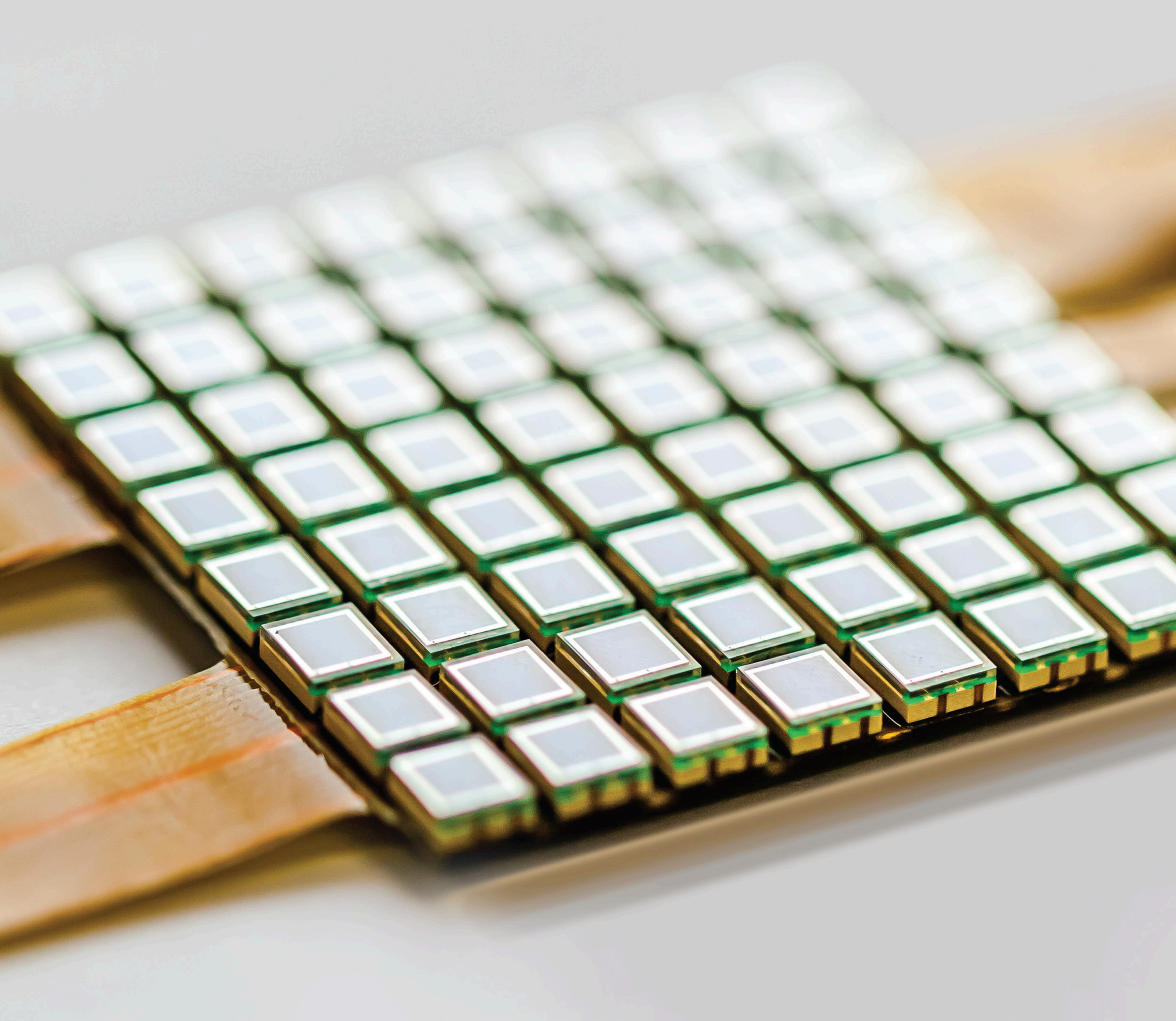
Journal of Sensors

Sensor Networks for Structural Health Monitoring

Lead Guest Editor: Samir Mustapha

Guest Editors: Ching-Tai Ng, Ye Lu, and Pawel Malinowski



Sensor Networks for Structural Health Monitoring

Journal of Sensors

Sensor Networks for Structural Health Monitoring

Lead Guest Editor: Samir Mustapha
Guest Editors: Ching-Tai Ng, Ye Lu, and Pawel Malinowski

Chief Editor

Editorial Board

Contents

Hindawi
Journal of Sensors
Volume 2020, Article ID 3060672, 2 pages
https://doi.org/10.1155/2020/3060672

Editorial

Sensor Networks for Structural Health Monitoring

Samir Mustapha,[1] **Ching-Tai Ng**,[2] **Ye Lu**,[3] **and Pawel Malinowski**[4]

[1] *American University of Beirut, Beirut, Lebanon*
[2] *University of Adelaide, Adelaide, Australia*
[3] *Monash University, Clayton, Australia*
[4] *Institute of Fluid Flow Machinery, Polish Academy of Sciences, Gdansk, Poland*

Correspondence should be addressed to Samir Mustapha; sm154@aub.edu.lb

Received 8 January 2020; Accepted 8 January 2020; Published 17 January 2020

Billions of dollars are spent globally every year on public infrastructure, automotive, and aerospace structures to keep up with the world's population growth. Managing those assets is of high significance, and to do that efficiently is an exhausting task that requires the need for continuous monitoring, maintenance, and rehabilitation. For instance, considering the bridge network in Europe, a recent survey showed that around 31% age between 50 and 100 years [1]. In-service and aging steel bridges entail a great attention to ensure a high level of safety, to maintain them in a good shape, and to extend their lifespan. The increase in the traffic flow and the size of the vehicles, environmental pollution, poor quality of construction, and inadequate maintenance necessitate the development of robust methods of inspection, in particular when dealing with complex and extremely large structures. Another example is the steel pipeline systems that are very crucial in transporting dangerous substances, such as crude oil and petroleum products, due to their practicality, efficiency, and cost-effectiveness. With many kilometres of the pipeline, carrying large volumes of hazardous substances, there are many reasons that may engender the pipeline's failure (mechanical, operational, corrosion, and natural failure or intrusion from a third party), rupture, leaks, and spillages. According to Concawe's report [2] having data collected on European cross-country oil pipelines from 1971 till 2016, 135 mechanical failures occurred from which 49 were due to construction faults and 86 due to design or material faults. As such, structural integrity monitoring and management are very crucial to maintain the high level of safety and prevent failure-related incidents and consequences.

Continuous structural health monitoring (SHM) systems would form a major establishment in the field of damage detection, assessment, and failure prediction. Knowing the integrity of in-service structures, on a continuous real-time basis, is crucial for manufacturers, maintenance teams, and operators. SHM is an area of growing interest and worthy of new and innovative approaches. A typical SHM system requires the constant collection of data from sensors that are embedded within the structure. The data can then be analyzed to detect the presence of any possible flaws; moreover, the remaining life of the monitored system can be estimated. The advancement in sensor technology, in its various forms, as well as hardware, leads to major developments of smart systems in many fields including manufacturing, automotive, aerospace, and civil engineering. The presence of a wide range of sensors, at a reduced cost, resulted in significant work in real-time monitoring of components and structures, in the last two decades, aiming at extending their lifetime, reducing the associated maintenance costs, and ensuring the public safety [3].

When designing a sensor network, many considerations must be taken into account that may impact every component of the SHM system, starting with the type of measurements (sensor selection or development), number and location of sensors, communication and data transmission, and finally data storage [4].

In this special issue, we aimed to shed some light on a very important component of an SHM system, which is related to sensing technologies and sensor networks. With no doubt, sensors' development and design of sensor networks are a fundamental step for an efficient and robust SHM system.

Although there have been a lot of focuses on sensor technologies and design of sensor networks, the implementation is not very smooth in particular when dealing with sensitive or critical components in automotive and aerospace structures. Retrofitting sensors on structures is not always plausible, and therefore, novel approaches for sensor installation or sensor embedment during manufacturing are of a high demand. Researchers and engineering scientists, in the future, should focus more on development of smart sensing systems that are effective for practical applications such as smart skins, smart paint, or miniature advanced sensing nodes.

Conflicts of Interest

The editors declare that they have no conflicts of interest regarding the publication of this special issue.

Acknowledgments

We wish to thank all the researchers who have contributed in their great work to the success of this special issue. Also, we would like to thank the reviewers that have been involved in the reviewing process. We would like to acknowledge the editorial board members for approving the publication of this special issue.

Samir Mustapha
Ching-Tai Ng
Ye Lu
Pawel Malinowski

References

[1] P. P. Reza Haghan and J. Patino, "Needs for maintenance and refurbishment of bridges in urban environments," in *Low Disturbance Sustainable Urban Construction*, Chalmers University, 2011.

[2] M. Cech, P. Davis, F. Gambardella et al., *Performance of European cross-country oil pipelines–statistical summary of reported spillages in 2016 and since 1971, report no. 6/18*, Concawe, 2018.

[3] P. Runcie, S. Mustapha, and T. Rakotoarivelo, "Advances in structural health monitoring system architecture," in *Proceedings of the the fourth International Symposium on Life-Cycle Civil Engineering, IALCCE*, vol. 14, pp. 1064–1071, 2014.

[4] W. Ostachowicz, R. Soman, and P. Malinowski, "Optimization of sensor placement for structural health monitoring: a review," *Structural Health Monitoring*, vol. 18, no. 3, pp. 963–988, 2019.

Hindawi
Journal of Sensors
Volume 2019, Article ID 9614630, 10 pages
https://doi.org/10.1155/2019/9614630

Research Article

A Real-Valued Genetic Algorithm for Optimization of Sensor Placement for Guided Wave-Based Structural Health Monitoring

Rohan Soman and Pawel Malinowski

Institute of Fluid Flow Machinery, Polish Academy of Sciences, 14 Fiszera Street, Gdansk 80-231, Poland

Correspondence should be addressed to Rohan Soman; rsoman@imp.gda.pl

Received 13 August 2019; Accepted 5 November 2019; Published 10 December 2019

Academic Editor: Antonio Lazaro

The paper presents a novel implementation of the genetic algorithm (GA) to improve the coverage of the sensor network for damage detection using guided waves. The implementation allows depiction of sensor locations with real values which is closer to the real-life situation. Also, additional features such as proximity checks and node insertions have been implemented in order to improve the convergence of the GA as well as the thoroughness of the search space. For the traditional integer-based implementation, the size of the problem is large but finite. For the real-valued implementation, the problem size can indeed be infinitely large. So added measures have been introduced such as a two-step optimization process for the reduction in size and improved convergence.

1. Introduction

Guided wave- (GW) based structural health monitoring (SHM) in one of the most widely used techniques for large plate or pipe-like structures. The propagating wave may be used to cover a large area and through the processing of the time of flight (TOF) allows damage isolation. The GW have been shown to be sensitive to extremely small levels of damage and have been employed for detection of damage due to impact, corrosion, and fatigue [1–4].

The research in the area of GW in metallic structures is quite extensive, but the work in the area of sensor placement is quite limited. Ostachowicz et al. [5] present an excellent review of the techniques used in the optimization of sensor placement with a special section dedicated to the optimization of sensor placement for GW-based SHM. The literature can be divided into primarily 3 areas. The first work in the area of sensor placement optimization was based on improving the probability of detection (POD). Staszewski et al. [6] used it in conjunction with artificial neural networks for improving the probability of impact localization and detection. Markmiller and Chang [7] used a metric dependent on the POD which was computed based on the response reconstruction of the impact event. Staszewski et al. and Markmiller et al. both used GA for the optimization. Flynn and Todd [8] used the probabilistic approach as well in the form of Bayes' risk. The aim is to minimize the false-positive and false-negative errors caused by the sensor network. Haynes [9] built on the Bayes' risk framework and included the cost of the SHM system in the decision-making process. Similar approaches based on the false alarms were also proposed by Vanli et al. [10] and Coelho et al. [11].

The second philosophy of the optimization is to improve the sensitivity of the network to damage. The work done in this area is largely finite element model based, where different damage scenarios are numerically simulated and used to determine the sensor locations such as the one by Lee and Staszewski [12]. Venkat et al. [13] and Ewald et al. [14] also present a method for locating the sensors at the maxima of the differential image of the healthy and damaged condition of the structure. These methods are useful for optimization of placement for the known hotspots in the structure.

Another philosophy of the optimization is to maximize the coverage of the sensor network. Soni et al. [15] developed a sensor placement algorithm based on the minimal sensing distance. The sensing range was determined based on the

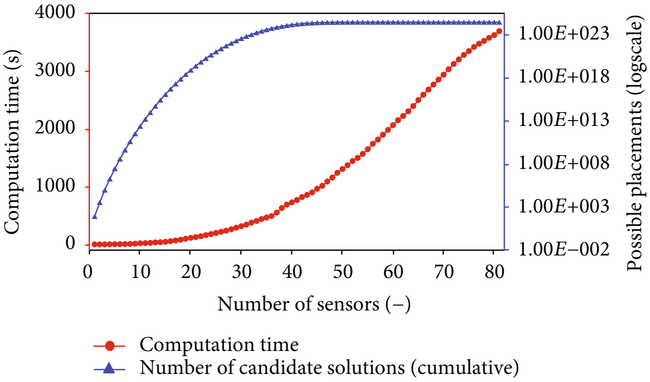


Figure 1: Problem size and computation time for increasing number of sensors.

signal to noise ratio (SNR) and the attenuation of the waves. The minimum sensor range was a circle of fixed radius determined experimentally. The backward sequential sensor placement (BSSP) was used in order to remove the redundant sensors in the network. Coelho et al. also developed an approach based on maximizing the coverage area by minimizing the probability of false alarms. Thiene et al. [16] proposed maximizing of coverage area based on a pixelated approach. The sample of interest was divided into pixels, and the coverage of the sensor network was calculated for each pixel. The different wave propagation features such as attenuation, line of sight, and shape of the sample can be incorporated based on different multiplication factors for obtaining the fitness function. The number of candidate locations are restricted in the study in order to limit the problem size. But this unnatural constraint may limit the performance of the optimization algorithm. The number of possible sensor locations was increased by Soman et al. [17] through the use of an analytical approach which is computationally more efficient than the pixel-based approach. Soman et al. [17] then extended the optimization cost function also to improve the quality of the damage isolation. The damage isolation in the GW-based SHM is carried out by the triangulation technique. Soman et al. included the area covered by at least 3 sensors as an additional metric. The multiobjective optimization problem was scalarized using weighing functions in order to simplify the optimization using the GA. Tarhini et al. [18] too used coverage of the specimen as a optimization objective. They developed a mixed integer nonlinear program which does not constrain the optimization search to a limited number of possible sensor locations and is a motivation for the current research.

In the present paper, the authors build on the defined cost function with 3 optimization objectives, namely, coverage by at least 1 sensor-actuator pair, coverage by 3 sensor-actuator pairs and the number of sensors. The implementation of the GA is changed from an integer GA to a real-valued GA. In order to restrict the size of the optimization problem, the number of sensors is limited to a range of values. This range is determined based on the sensor densities required for the SNR to allow reliable damage detection. The cost function computation is the most computationally demanding step, and hence, the number of unnecessary computations needs to be reduced. In order to limit this number, some features such as node insertion and the proximity detection have been added to the implementation.

The rest of the paper is organized as follows: the next section explains the methodology for defining the optimization problem. Section 3 presents the additional functionalities such as node insertion and proximity check implemented for improving the performance of the GA. Section 4 covers the results of the optimization and the comparison of the improved GA with the earlier work. The last section draws some conclusions and presents areas of future work.

2. Methodology

The increase in number of sensors deployed on a structure leads to an increase in the deployment costs as well as secondary costs related to the extra weight of the sensors and the wiring as well as the processing and storage of the data. Hence, one of the objectives of the optimization of sensor placement should be the minimization of sensors used. This minimization can be implemented in the cost function or as a constraint in the allowed placements. If it is incorporated in the cost function as by Soman et al. [17], the number of possible sensor placements increases. The optimization problem becomes very large, and as a result, the time needed for convergence is very large. Also, the time consumed for the computation of the cost function increases with the increase in number of sensors as shown in Figure 1. The computations are based on the implementation of the GA reported in [17].

Also, the sensor placements with the large number of sensors are not feasible due to the availability of the resources. Thus, in order to reduce the size of the optimization problem, constraints on the number of sensors may be imposed right at the implementation stage. This constraint must be imposed in an objective way in order to ensure that the sensor performance is within the acceptable range. Thus, this section discusses a formal method for determining the maximum number of sensors.

2.1. Sensor Number Determination. The number of sensors is determined based on sensor densities using the concept developed by Croxford et al. [19]. They provide an excellent discussion and step by step process for calculating the different parameters for determining the sensor densities. For completeness, the equations for calculating the sensor pitch and all the factors are provided here without the derivation which can be found in [19]. The minimum pitch of the sensors is given by

$$p = \left(\frac{3^{3/4} R_{\text{damage}}}{\sqrt{2} S \beta \delta T} \right)^{(2/3)}, \tag{1}$$

where R_{damage} is the reflection coefficient of the damage (defined in terms of the scattered wave amplitude at unit distance from the damage), S is the minimum SNR required for reliable damage assessment, β is the coefficient corresponding to the post subtraction noise between the baseline signal and the signal at the present time, and δT is the change in temperature. The factor β is dependent on the type of subtraction carried out as well as the wave mode. In the paper by Croxford et al., the value for RF subtraction is given by

$$\beta_{\text{RF}} = 2\pi f \frac{1}{\text{v}_{\text{ph}}} \left(\alpha - \frac{\text{k}_{\text{ph}}}{v_{\text{ph}}} \right), \tag{2}$$

where v_{ph} is the phase velocity, k_{ph} is the coefficient relating the sensitivity of the phase velocity to temperature, and α is the coefficient of expansion.

The factor R_{damage} is dependent on the type of damage considered. For a hole in the plate considered as a cylindrical scatterer, the coefficient can be analytically given by

$$R_{\text{damage}} = 0.55\sqrt{d}, \tag{3}$$

where d is the diameter of the hole in m.

Knowing the values for all the parameters in equation (1), the pitch of sensors can be calculated which in turn may be used for determining the minimum number of sensors. The maximum number of sensors then can be determined by introducing some redundancy in the system. As shown in Figure 1, the problem size and the computation time increase with the increase in the maximum number of sensors. So care should be taken in defining the maximum number. For the purpose of the study and to ensure some redundancy, the maximum number of sensors was identified as 50% more than the minimum number of sensors required.

The sample of interest was an aluminium plate with dimensions 1 m × 1 m × 1 mm shown in Figure 2. Added mass was used to simulate damage. The backscatter profile of the added mass was obtained based on the full-field measurements from the laser Doppler vibrometer as shown in Figure 3. The result is for the centrally located mass shown in Figure 2.

As can be seen, the minimum value for the backscatter was 0.073 which is taken as the back-scatter R_{damage}. Key points to note is that the backscatter is more or less symmetrical (within reasonable errors). The small error can be attributed to the fact that the sampled points were in a rectangular grid as opposed to a radial grid. Hence, the distances at the point of measurement were approximately equal. The maximum backscatter occurs at 45° to the incident angle. The minimum value is in the area just beyond the mass as is expected. So the worst case will be when the sensor is at the other side of the actuator which is considered in computing the minimum number of sensors. The backscattering index obtained is equivalent to 17 mm hole in the sample according to equation (3) which is a reasonable assumption for a scattering object. The β value for the aluminium plate S0 wave based on equation (2) is given as 0.0962. In the author's team, methods have been developed for temperature compensation which allow -14 dB change in the SNR for 10°C change in temperature [20]. The 14 dB change results in $\beta = 0.0192$. The SNR = 2 (similar to [19]) has been assumed to be necessary for ensuring reliable damage detection. Based on these inputs, the p calculated based on equation (1) is 0.454 which relates to the minimum number of sensors as 6. Taking into

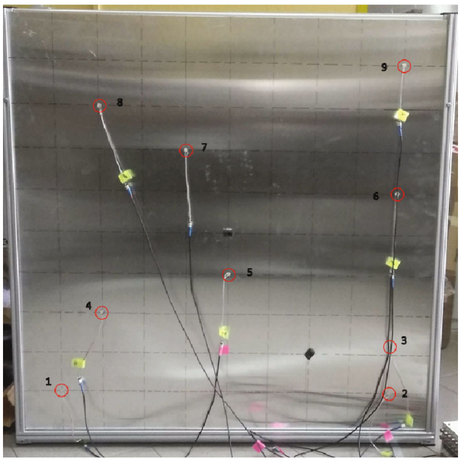


Figure 2: Aluminium plate under investigation [17].

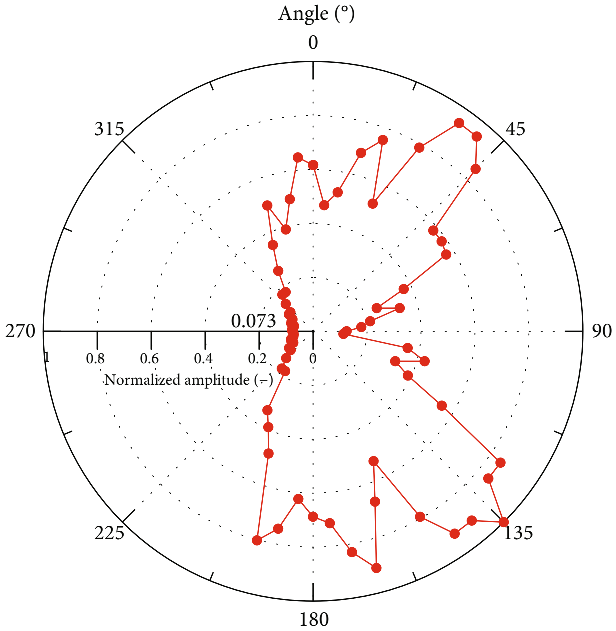


Figure 3: Backscatter of the waves due to discontinuity (added mass)—excitation from 90°.

consideration the proposed redundancy in the system, the maximum number of sensors is calculated as 9. This allows comparison of the method with the older method as the sensor optimization carried out previously and reported in [17] was on a network of 9 sensors.

2.2. Sensor Location Optimization. Once the number of sensors is known, an optimization scheme can be implemented by restricting the number of sensors between the lower and the upper limits. The criterion for the optimization is given by the cost function. As has been mentioned in [17], the three demands from the application are as follows:

(1) coverage with at least 1 sensor-actuator pair (coverage1)

(2) coverage with at least 3 sensor-actuator pairs (coverage3)

(3) number of sensors (s)

Based on these demands, a scalarized cost function can be developed by using weighing factors as shown in

$$\text{cost} = -1 \times \left(\omega \frac{\text{coverage3}}{s^{\gamma}} + (1-\omega) \frac{\text{coverage1}}{s^{\delta}} \right), \qquad (4)$$

where coverage3 is the % of points of the grid which lie within the sensing range of 3 or more sensor-actuator pairs; coverage1 is the % of points which lie in the sensing range of a single sensor-actuator pair. ω, γ, and δ are weighting values to determine the relative merit for each of the parameters, and s is the number of sensors. The parameters γ and δ can be treated as independent of each other or dependent based on the choice. The two parameters were introduced to show the different correlations of the coverage3 and coverage1 values to the number of sensors.

For the two-stage optimization implementation illustrated in this paper, the choice of the weighing parameters is even more sensitive. As the change in the number of sensors is limited, the range of values for the parameter too are limited and do not show much change. Hence if the weighing values for γ and δ are too low (e.g., 0), the algorithm will choose solutions with maximum number of sensors while if they are too high (e.g., 1), the number of sensors will have a very high bearing the sensor placement, and as such, the placements with lower number of sensors will be preferred. This value depends on the contribution of each sensor to the coverage of the network. For metallic structure without any structural features such as stiffeners, each sensor contributes significantly; hence, the low values for γ and δ need to be chosen. Sensitivity studies were carried out with evenly spaced sensor placements (Figure 4) for different number of sensors as shown in Table 1. It is acknowledged that the evenly spaced sensor placements may or may not be optimal. The aim of Table 1 and Figure 4 is to show the contribution of each sensor towards the coverage1 and coverage3 values and their bearing on the choice of γ and δ values.

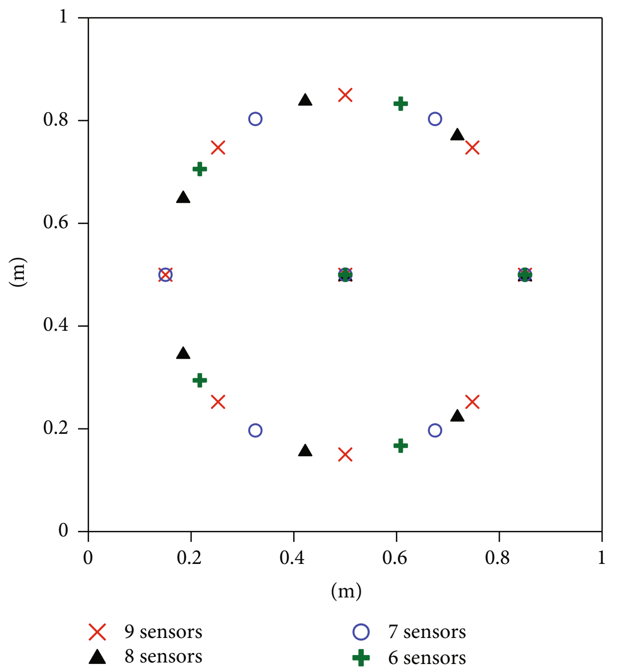


Figure 4: Even distribution of sensors to show sensor contribution.

Table 1: Change in metrics with different number of evenly distributed sensors.

s	coverage1	coverage3	Contribution per sensor to coverage1	% contribution per sensor to coverage3
	%	%	%	%
6	84.1	74.5	14.0	12.4
7	85.2	77.4	12.2	11.1
8	86.6	78.5	10.8	9.8
9	90.2	82.8	10.0	9.2

As can be seen in Table 1, the contribution per sensor reduces but the overall coverage increases with the increase in the sensor number. In order to obtain similar cost for sensor placement with 6 sensors and 9 sensors, the δ value needs to be 0.17. Similarly, the γ value needs to be 0.26. As mentioned, the evenly placed sensor placement is suboptimal; as a result, the sensor contribution too is suboptimal. For optimized sensor placements, the values for γ and δ should be significantly lower. Hence, for the purpose of the study, values of γ and δ were taken as 0.15.

The optimization of the locations was carried out using a real-valued implementation of the GA with special tools and routines incorporated for improved convergence which have been described in the next section.

3. Implementation of the GA

The main innovation of the paper is the implementation of real-valued GA as opposed to the commonly used integer GA for sensor placement optimization. The underlying motivation for this is the observation that the more realistic the encoding of the optimization, the better the performance of

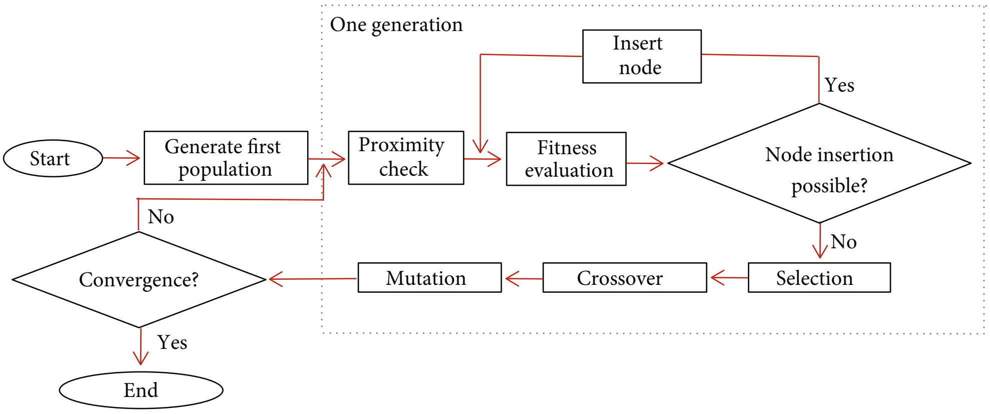
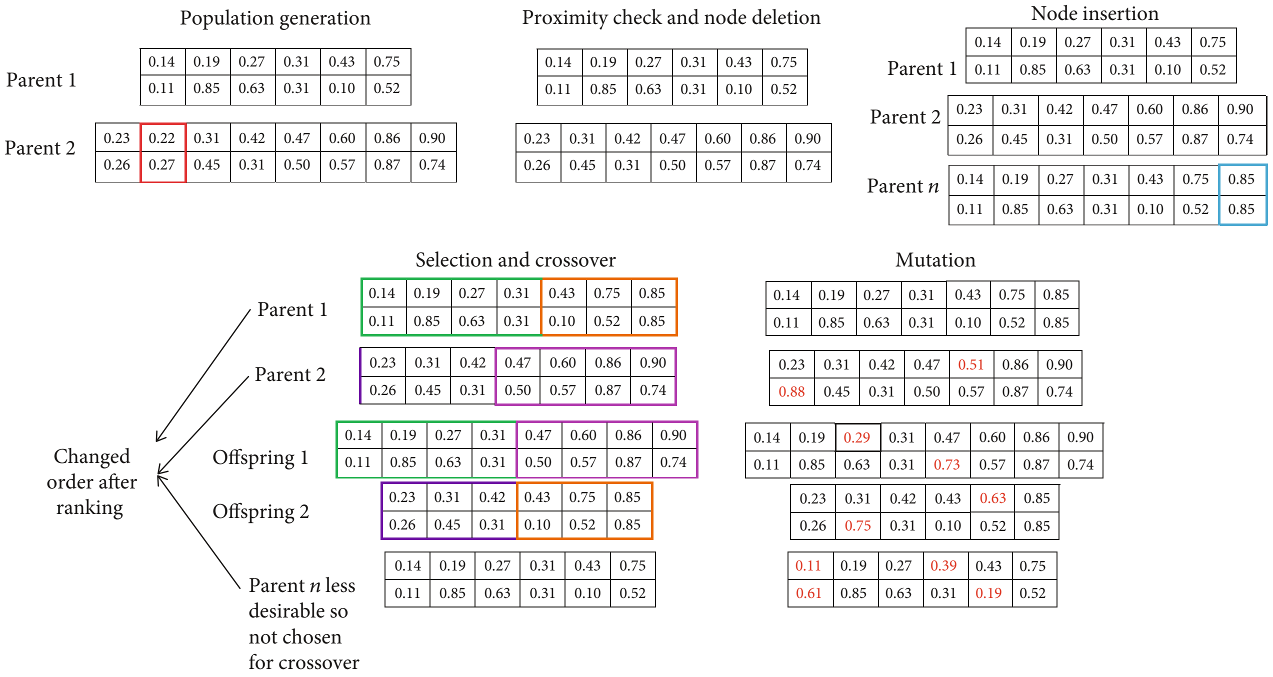


Figure 5: GA flowchart.


Population generation
Proximity check and node deletion
Node insertion
Selection and crossover
Mutation
Changed order after ranking
Parent n less desirable so not chosen for crossover

Population generation

Parent 1

0.14	0.19	0.27	0.31	0.43	0.75
0.11	0.85	0.63	0.31	0.10	0.52

Parent 2

0.23	0.22	0.31	0.42	0.47	0.60	0.86	0.90
0.26	0.27	0.45	0.31	0.50	0.57	0.87	0.74

Proximity check and node deletion

0.14	0.19	0.27	0.31	0.43	0.75
0.11	0.85	0.63	0.31	0.10	0.52

0.23	0.31	0.42	0.47	0.60	0.86	0.90
0.26	0.45	0.31	0.50	0.57	0.87	0.74

Node insertion

Parent 1

0.14	0.19	0.27	0.31	0.43	0.75
0.11	0.85	0.63	0.31	0.10	0.52

Parent 2

0.23	0.31	0.42	0.47	0.60	0.86	0.90
0.26	0.45	0.31	0.50	0.57	0.87	0.74

Parent *n*

0.14	0.19	0.27	0.31	0.43	0.75	0.85
0.11	0.85	0.63	0.31	0.10	0.52	0.85

Selection and crossover

Parent 1

0.14	0.19	0.27	0.31	0.43	0.75	0.85
0.11	0.85	0.63	0.31	0.10	0.52	0.85

Parent 2

0.23	0.31	0.42	0.47	0.60	0.86	0.90
0.26	0.45	0.31	0.50	0.57	0.87	0.74

Offspring 1

0.14	0.19	0.27	0.31	0.47	0.60	0.86	0.90
0.11	0.85	0.63	0.31	0.50	0.57	0.87	0.74

Offspring 2

0.23	0.31	0.42	0.43	0.75	0.85
0.26	0.45	0.31	0.10	0.52	0.85

0.14	0.19	0.27	0.31	0.43	0.75
0.11	0.85	0.63	0.31	0.10	0.52

Mutation

0.14	0.19	0.27	0.31	0.43	0.75	0.85
0.11	0.85	0.63	0.31	0.10	0.52	0.85

0.23	0.31	0.42	0.47	0.51	0.86	0.90
0.88	0.45	0.31	0.50	0.57	0.87	0.74

0.14	0.19	0.29	0.31	0.47	0.60	0.86	0.90
0.11	0.85	0.63	0.31	0.73	0.57	0.87	0.74

0.23	0.31	0.42	0.43	0.63	0.85
0.26	0.75	0.31	0.10	0.52	0.85

0.11	0.19	0.27	0.39	0.43	0.75
0.61	0.85	0.63	0.31	0.19	0.52

Figure 6: Example of the real-encoded GA with additional features.

the algorithm. Also, by changing the implementation from the integer to real GA, the difference in the phenotype for a unit change in the sensor values is significantly reduced thus allowing better search in the sample space. On the downside, the size of the problem is no longer finite but infinite. Thus, there is no way for checking the validity of the optimization tool with brute-force methods. The flow chart for the GA is provided in Figure 5.

The population is generated with each individual sensor placement depicted by $2 \times N$. The first row corresponds to the x-coordinate while the second row corresponds to the y -coordinate. The x and y coordinates are treated as independent in the population generation, fitness evaluation, node insertion, and mutation phases while in the crossover and selection phase, y-coordinate is treated as dependent variable. The number N corresponds to number of sensors and can take any value in the chosen range determined by the method outlined in Section 2.1. The different features incorporated in the GA are shown through an example in Figure 6.

3.1. Proximity Check. This feature is introduced to avoid concentration of the nodes at a point or ensure the feasibility of the sensor placement. In the first step, nodes which are too close to the boundaries are omitted as it will be difficult to

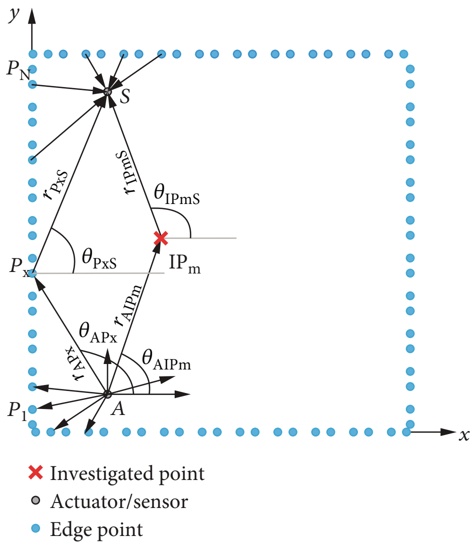


Figure 7: Schematic explaining the ray-tracing approach.

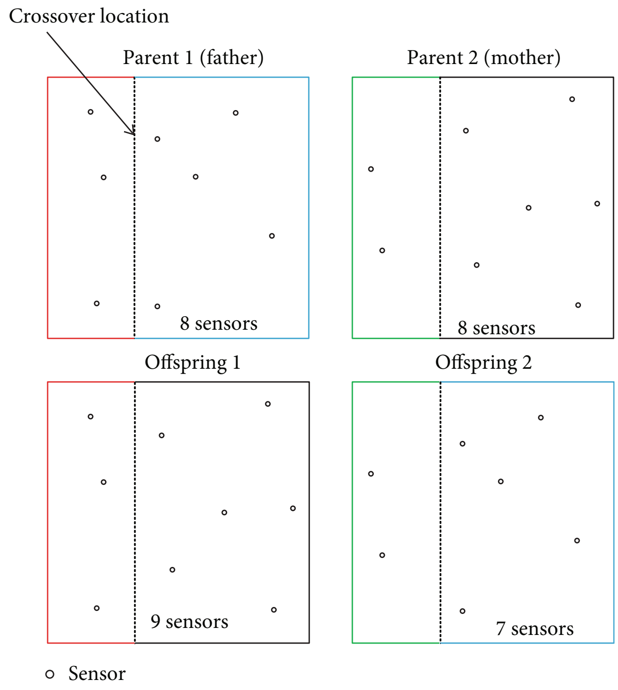


Figure 8: Mirror crossover schematic.

distinguish between direct signals and the reflection from the boundary. In the next stage, redundancy in the system because of 2 closely spaced sensors is reduced by deleting the sensor. The limit for the proximity check was taken as the diameter of the sensors used which was 0.01 m. This constraint ensures that the optimized network is possible to be realized physically.

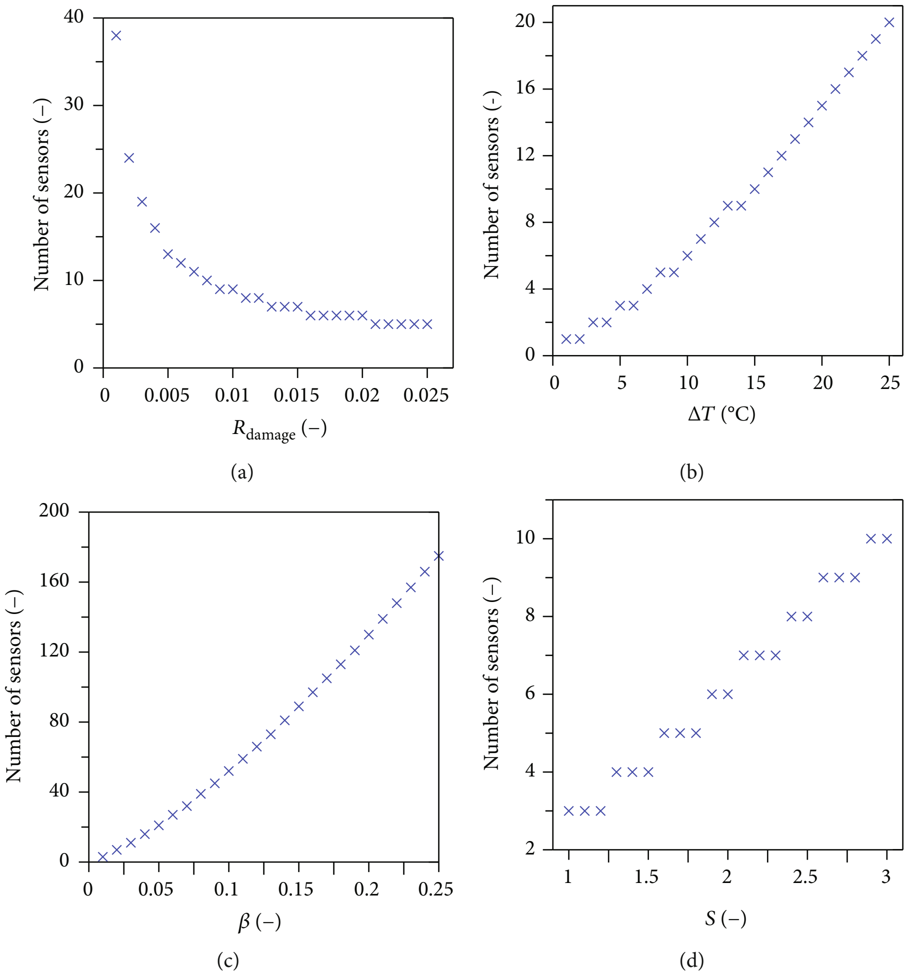
3.2. Node Insertion. In case the proximity check removes a gene from the chromosome, there is a possibility to increase the number of sensor in the chromosome by adding a node at the location with the poorest coverage. The node is inserted if it provides an advantage over the existing sensor placement in terms of the coverage3, coverage1, and the scalarized cost function. The node insertion is repeated until the insertion is possible and desirable. The chromosome with the added gene replaces the lowest ranked chromosome in the selection process. The node insertion allows a better local search but at the cost of possible entrapment in the local minima. This entrapment is caused as there are two copies of very similar chromosomes which are very desirable in the population. So in order to avoid their domination in the subsequent generations, the number of chromosomes in each generation is increased as compared to the previous implementation of the GA reported in [17].

3.3. Fitness Evaluation. In the previous work by the authors, the analytical approach based on the largest ellipse fitting inside the plate was employed for determining the coverage of each sensor-actuator pair. This approach is simple to implement for simple structures and is computationally efficient. For problems where the propagation is direction dependent (anisotropic structures, or structures with damage backscatter with an angle-dependent profile), the ellipse approach is not valid. Hence, the ray-tracing approach [21] explained in Figure 7 was employed. In the ray-tracing approach, a ray is extended from the actuator to the location under investigation and another ray is extended between the investigated point and the sensor. The attenuation, velocity, or backscatter can be incorporated based on the angle of the rays with the coordinate axes and the distance between the points. The maximum allowed TOF is determined from the edge points. This TOF is then used to construct a limiting ellipse with the major axis equal to the product of maximum velocity and the TOF. The points within the ellipse are then checked individually with the ray-tracing approach to determine the coverage of the sensor-actuator pair. The fitness value evaluated for the sensor network is the superposition of the coverage for each sensor-actuator pair.

3.4. Crossover. The standard crossover techniques used in the GA are the single-point crossover, the multipoint crossover, the arithmetic crossover, etc. [22]. They are simplistic to implement but often are not exactly aligned with the implementation and the physical nature of the problem. As mentioned earlier, the closer the encoding of the problem to reality, the better is the performance of the optimization. Hence, the mirror crossover [23] was implemented for the optimization problem. The method for the mirror crossover is shown in Figure 8 and is as follows: two parents are selected randomly similarly to the other crossover techniques.

Then, a random value x_{cross} of the x-coordinate is generated. All genes with $x > x_{\text{cross}}$ in the father are transferred to offspring 1 and $x \leq x_{\text{cross}}$ in the father to offspring 2. The genes with $x > x_{\text{cross}}$ from the mother are transferred to offspring 1 and $x \leq x_{\text{cross}}$ from the mother are transferred to offspring 2.

As can be seen, the number of sensors in both parents is 8. By the use of mirror crossover, it is possible to obtain a sensor placement with fewer number of sensors (7 in



(a)

(b)

(c)

(d)

Figure 9: Change in minimum number of sensors with different parameters: (a) R_{damage}; (b) ΔT; (c) β; (d) S.

offspring 2). Thus, increasing the search capability of the optimization algorithm.

3.5. Algorithm Inputs. For obtaining the optimized sensor placement, several variables need to be determined based on the sensitivity analysis. For the problem size at hand, the number of chromosomes was taken as 256. This is to minimize the domination of the gene pool by a few genes due to the node insertion phase. The elitism was 50%. The mutation rate was 25%, and number of generations was 5000. The next section compares the results of the optimization from the real-valued GA with the integer GA.

4. Results and Discussion

4.1. Sensor Number. As shown in Section 2, the pitch of the sensors and in turn the sensor density is dependent on the values of R_{damage}, S, β, and ΔT. The parameter R_{damage} depends on the backscatter characteristics of the damage while the parameter β depends on the material and the central frequency used for the excitation. Figure 9 shows the change in the number of sensors with unilateral change in any of the 4 variables.

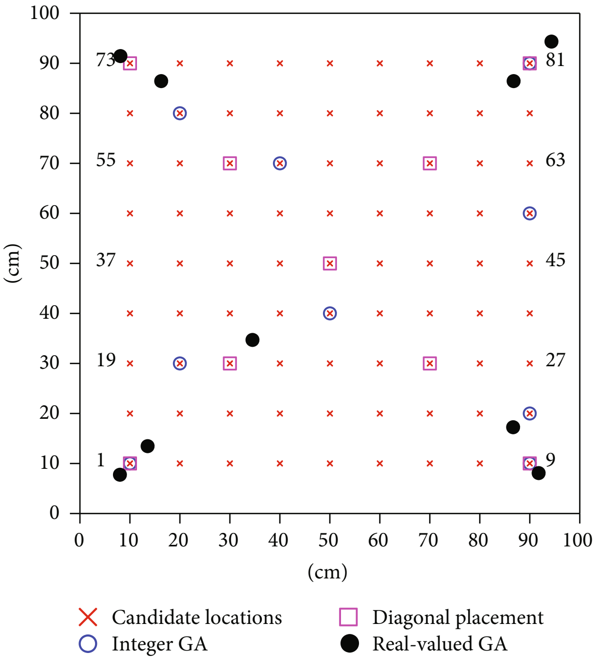
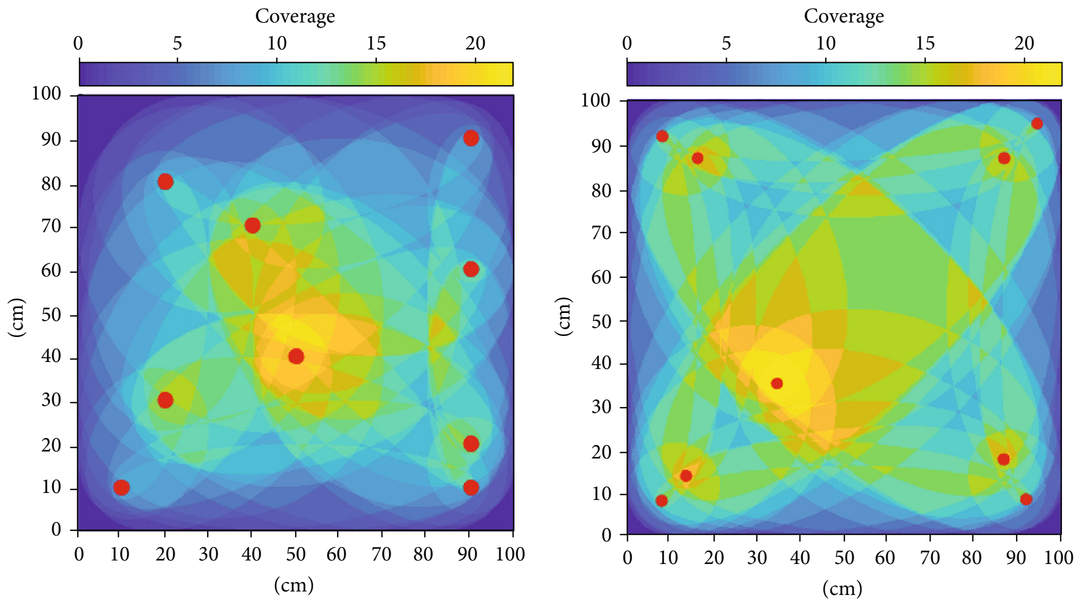


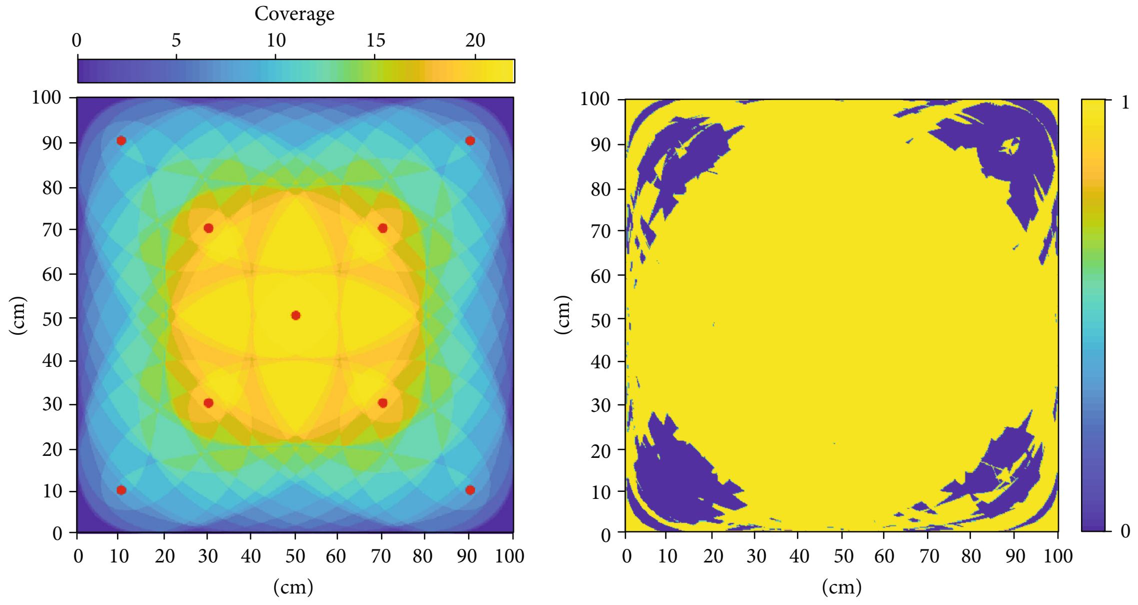
Figure 10: Sensor placement for different runs.



(a)

(b)

Figure 11: Surface plot showing coverage: (a) integer GA-based placement; (b) real GA-based placement.



(a)

(b)

Figure 12: Surface plot showing coverage: (a) diagonal placement; (b) difference in coverage for real-coded GA and diagonal placement (yellow shows area with improvement).

The factor R_{damage} can be changed with the size and type of defect which is to be detected. The factor ΔT depends on the uncertainty in ambient conditions expected during the application. The value of β depends on the frequency of excitation as well as the material properties. The material properties affect the phase and group velocity as well as the dependence of the material on the change in temperature. The factor S depends on the quality of the signal processing and noise cancellation algorithm. It can also be used to introduce the effects of attenuation which is significant in composites. The value of S can be increased in case the attenuation is high in order to determine the sensor density.

4.2. Sensor Location. The real-valued implementation of the optimization eliminates the unnecessary constraint on the locations of sensors imposed due to the integer-based implementation. As a result, better coverage3 and coverage1 and in turn better fitness value may be achieved. Figure 10 shows the

Table 2: Performance parameters for different sensor placements.

Run	Generations	coverage1	coverage3
Integer GA	5000	96.1%	88.7%
Real GA	5000	98.2%	96.0%
Diagonal	—	97.7%	94.8%

optimal sensor placement achieved for the integer placement and for the real-valued optimization. Figures 11(a), 11(b), and 12(a) show the coverage plot for the three sensor placements. Figure 12(b) also shows the improved coverage achieved through the real-valued implementation. The objective values for the optimization are quantitatively compared in Table 2.

5. Conclusions

The paper outlines a two-step methodology for optimization of sensor placement for GW-based damage detection. In the first step, the minimum number of sensors needed is calculated based on the quality of the signal processing algorithm. Once the number of sensors is determined, the location of the sensors is optimized through improved implementation of the GA. The optimization problem is posed using real values rather than constraining the search with the use of integer-based implementation. In order to account for the increase in the search space for optimal solution and improve the computational performance of the algorithm, some key features have been introduced in the GA such as proximity checking, node insertion, and use of mirror crossover scheme. The use of these features allows the improvement in the search capability as well as the computational efficiency of the search algorithm.

The paper presents sensitivity studies for the different parameters in determining the number of sensors. The paper also shows that through the use of real-valued implementation improved coverage using the same number of sensors can be achieved. Also, the computational efficiency for the real-valued GA is better than the integer GA. Based on the presented results, the use of real-valued GA is recommended. Incorporation of backscatter profiles from different damage scenarios, use of the technique for composite structures with structural features such as stiffener, rivets, and the experimental validation of the proposed methodology are identified as the areas of further research.

Data Availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Disclosure

The opinions expressed in this paper do not necessarily reflect those of the sponsors.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

The conceptualization, methodology, validation, optimization, writing, review and editing, and visualization for the manuscript were carried out by R.S. The resources, LDV measurements and data processing, supervision, project administration, and funding acquisition were carried out by P.M.

Acknowledgments

The authors are grateful to TASK-CI for allowing the use of their computational resources. This research was funded by the National Science Center, Poland, under the Sonata-Bis 6 call with the grant number: 2016/22/E/ST8/00068.

References

[1] M. Salmanpour, Z. Sharif Khodaei, and M. Aliabadi, "Impact damage localisation with piezoelectric sensors under operational and environmental conditions," *Sensors*, vol. 17, no. 5, p. 1178, 2017.

[2] J. He, Y. Ran, B. Liu, J. Yang, and X. Guan, "A fatigue crack size evaluation method based on lamb wave simulation and limited experimental data," *Sensors*, vol. 17, no. 9, p. 2097, 2017.

[3] S. Sikdar and S. Banerjee, "Identification of disbond and high density core region in a honeycomb composite sandwich structure using ultrasonic guided waves," *Composite Structures*, vol. 152, pp. 568–578, 2016.

[4] W. Li, C. Xu, and Y. Cho, "Characterization of degradation progressive in composite laminates subjected to thermal fatigue and moisture diffusion by lamb waves," *Sensors*, vol. 16, no. 2, p. 260, 2016.

[5] W. Ostachowicz, R. Soman, and P. Malinowski, "Optimization of sensor placement for structural health monitoring: a review," *Structural Health Monitoring*, vol. 18, no. 3, pp. 963–988, 2019.

[6] W. J. Staszewski, K. Worden, R. Wardle, and G. R. Tomlinson, "Fail-safe sensor distributions for impact detection in composite materials," *Smart Materials and Structures*, vol. 9, no. 3, pp. 298–303, 2000.

[7] J. F. Markmiller and F.-K. Chang, "Sensor network optimization for a passive sensing impact detection technique," *Structural Health Monitoring*, vol. 9, no. 1, pp. 25–39, 2010.

[8] E. B. Flynn and M. D. Todd, "Optimal placement of piezoelectric actuators and sensors for detecting damage in plate structures," *Journal of Intelligent Material Systems and Structures*, vol. 21, no. 3, pp. 265–274, 2010.

[9] C. M. Haynes, *Effective health monitoring strategies for complex structures, [Ph.D. thesis]*, UC San Diego, 2014.

[10] O. A. Vanli, C. Zhang, A. Nguyen, and B. Wang, "A minimax sensor placement approach for damage detection in composite structures," *Journal of Intelligent Material Systems and Structures*, vol. 23, no. 8, pp. 919–932, 2012.

[11] C. K. Coelho, S. B. Kim, and A. Chattopadhyay, "Optimal sensor placement for active guided wave interrogation of complex

metallic components," in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2011*, vol. 7981, p. 79813O, San Diego, CA, USA, 2011.

[12] B. Lee and W. Staszewski, "Sensor location studies for damage detection with lamb waves," *Smart Materials and Structures*, vol. 16, no. 2, pp. 399–408, 2007.

[13] R. S. Venkat, C. Boller, N. Ravi et al., "Optimized actuator/sensor combinations for structural health monitoring: simulation and experimental validation," in *Structural Health Monitoring 2015*, Stanford, USA, 2015.

[14] V. Ewald, R. M. Groves, and R. Benedictus, "Transducer placement option of lamb wave SHM system for hotspot damage monitoring," *Aerospace*, vol. 5, no. 2, p. 39, 2018.

[15] S. Soni, S. Das, and A. Chattopadhyay, "Optimal sensor placement for damage detection in complex structures," in *Volume 2: Multifunctional Materials; Enabling Technologies and Integrated System Design; Structural Health Monitoring/NDE; Bio-Inspired Smart Materials and Structures*, pp. 565–571, Oxnard, CA, USA, 2009.

[16] M. Thiene, Z. S. Khodaei, and M. Aliabadi, "Optimal sensor placement for maximum area coverage (MAC) for damage localization in composite structures," *Smart Materials and Structures*, vol. 25, no. 9, article 095037, 2016.

[17] R. Soman, P. Kudela, K. Balasubramaniam, S. K. Singh, and P. Malinowski, "A study of sensor placement optimization problem for guided wave-based damage detection," *Sensors*, vol. 19, no. 8, p. 1856, 2019.

[18] H. Tarhini, R. Itani, M. A. Fakih, and S. Mustapha, "Optimization of piezoelectric wafer placement for structural health-monitoring applications," *Journal of Intelligent Material Systems and Structures*, vol. 29, no. 19, pp. 3758–3773, 2018.

[19] A. J. Croxford, P. D. Wilcox, B. W. Drinkwater, and G. Konstantinidis, "Strategies for guided-wave structural health monitoring," *Proceedings Mathematical, Physical and Engineering Sciences*, vol. 463, no. 2087, pp. 2961–2981, 2007.

[20] C. A. Dan and P. Kudela, "Temperature compensation methods for elastic wave based SHM," in *Recent Progress in Flow Control for Practical Flows*, P. Doerffer, G. Barakos, and M. Luczak, Eds., pp. 483–497, Springer, Cham, 2017.

[21] R. Soman, P. Kudela, and P. Malinowski, "Improved damage isolation using guided waves based on optimized sensor placement," in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2019*, vol. 10970, p. 109700B, Denver, CO, USA, 2019.

[22] A. Umbarkar and P. Sheth, "Crossover operators in genetic algorithms: a review," *ICTACT Journal on Soft Computing*, vol. 6, no. 1, pp. 1083–1092, 2015.

[23] L. Dai and B. Wang, "Sensor placement based on an improved genetic algorithm for connected confident information coverage in an area with obstacles," in *2017 IEEE 42nd Conference on Local Computer Networks (LCN)*, pp. 595–598, Singapore, 2017.

Hindawi
Journal of Sensors
Volume 2019, Article ID 2607893, 12 pages
https://doi.org/10.1155/2019/2607893

Research Article

A Three-Dimensional Strain Rosette Sensor Based on Graphene Composite with Piezoresistive Effect

Zhiqiang Wu,[1] Jun Wei,[1] Rongzhen Dong,[1] and Hao Chen[2]

[1]*School of Civil Engineering, Central South University, Changsha 410075, China*
[2]*China Construction Second Engineering Bureau Co. Ltd, Beijing 100071, China*

Correspondence should be addressed to Zhiqiang Wu; 124801026@csu.edu.cn

Received 7 June 2019; Revised 24 September 2019; Accepted 10 October 2019; Published 22 November 2019

Academic Editor: Ching-Tai Ng

Obtaining the internal stress and strain state of concrete to evaluate the safety and reliability of structures is the important purpose of concrete structural health monitoring. In this paper, a three-dimensional (3D) strain rosette sensor was designed and fabricated using graphene-based piezoresistive composite to measure the strains in concrete structures. The piezoresistive composite was prepared using reduced graphene oxide (RGO) as conductive filler, cellulose nanofiber (CNF) as dispersant and structural skeleton, and waterborne epoxy (WEP) as polymer matrix. The mechanical, electrical, and electromechanical properties of RGO-CNF/WEP composite were tested. The results show that the tensile strength, elastic modulus, and conductivity of the composite are greatly improved by the addition of RGO and CNF. The relative resistance change of composite films demonstrates high sensitivity to mechanical strain with gauge factors of 16-52. Within 4% strain, the piezoresistive properties of composites are stable with good linearity and repeatability. The sensing performance of the 3D strain rosette was tested. The measured strains are close to the actual strains of measure point in concrete, and the error is small. The RGO-CNF/WEP composite has excellent mechanical and piezoresistive properties, which enable the 3D strain rosette to be used as embedded sensor to measure the internal strain of concrete structures accurately.

1. Introduction

Concrete is the most widely used material in civil engineering. The service periods of concrete structures are usually several decades or even longer. In the long-term process, combined actions of multiple factors, such as load effect, environmental erosion, and material aging, will lead to damage accumulation and resistance attenuation of concrete structures [1, 2]. If the crisis situation cannot be timely warned and repaired, it will easily lead to catastrophic accidents. Therefore, structural health monitoring is necessary to determine the stress and strain states of key points in concrete and to evaluate the safety and reliability of the structure.

Sensing system is the base of concrete structural health monitoring. Traditional sensors include strain gauges, fiber Bragg grating [3, 4], and piezoelectric materials [5–7]. However, these sensors have some problems, such as single testing direction, high cost, bad durability, and poor compatibility with concrete. The emergence of cement-based smart composites [8–14] provides a new sensing mean for structural health monitoring, and it has been well applied in civil engineering as embedded sensors. However, although the cement-based smart sensor has high sensitivity, good linearity, and well compatibility with concrete, it is greatly affected by environmental factors, especially temperature and humidity, and has poor plasticity. Therefore, it is mostly used to measure one-dimensional stress and strain in concrete. Polymer-based intelligent sensor [15, 16] can avoid the above problems, because of the good corrosion resistance and waterproofing, simple molding, and well plasticity of polymer materials.

Due to the unique crystal structures and properties, nanomaterials can be used to enhance and modify polymer, which provides a theoretical basis for the preparation of new sensors [17–20]. At present, most polymer-based intelligent composites are based on flexible materials such as rubber. Because of the low elastic modulus of rubber, the mechanical properties of concrete will be greatly reduced

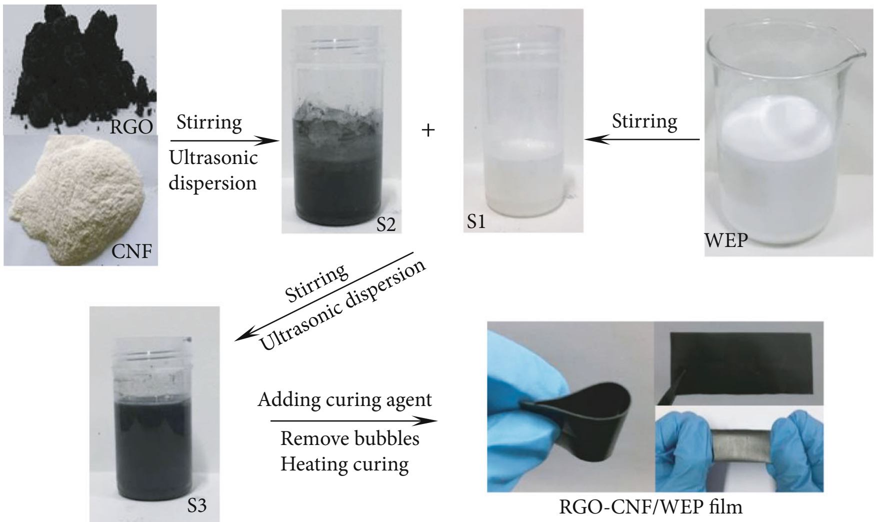


Figure 1: The fabrication process of the RGO-CNF/WEP composite.

when embedded in the component. In addition, the composites using nanocarbon black or carbon nanotube as conductive fillers have low sensitivity, high filling content, and poor repeatability. Therefore, to make new sensors, epoxy resin with high mechanical strength [21] is used as the matrix material, and graphene, a two-dimensional nanomaterial, with excellent mechanical and electrical properties [22] is selected as the conductive filler.

In this paper, a graphene-based composite material with piezoresistive effect was prepared first. RGO was used as the conductive filler of the composite. CNF with good dispersion in water [23] was used as dispersant and carries RGO to evenly disperse in WEP to form a stable and continuous 3D reinforcing and conductive network. The mechanical, electrical, and electromechanical properties of the composite were tested. Then, based on the strain theory, a 3D strain rosette was designed and fabricated to obtain the strain state of a point inside the concrete structure. The sensing elements of the strain rosette are six one-dimensional strain sensors formed by the RGO-CNF/WEP composite and arranged in six directions of an epoxy cube. At last, the sensing performance of the 3D strain rosette was tested and analyzed.

2. Fabrication of the RGO-CNF/WEP Composite

2.1. Materials and Instruments. Reduced graphene oxide (RGO) was purchased from Suzhou Tanfeng Graphene Technology Co., Ltd, China (purity > 98 wt%, thickness 1-3 nm, flake size 0.5-5 μm). Cellulose nanofiber (CNF) was purchased from Guilin Qihong Technology Co., Ltd, China (purity > 99 wt%, diameter 4-10 nm, length 1-3 μm). Waterborne epoxy (WEP, product number F0704) and curing agent (product number F0705) were purchased from Shenzhen Jitian Chemical Co., Ltd, China. The materials were all used as received. Deionized (DI) water was prepared in our lab. A horn-type sonicator (JY92, Ningbo Scientz Biotechnology Co., Ltd, China) and a magnetic stirrer (LC-TN-1, LICHEN Instrument, China) were used to assist the dispersion of RGO.

2.2. Preparation of the RGO-CNF/WEP Composite. Figure 1 shows the fabrication process of the composite. WEP has high viscosity which is not conducive to uniform dispersion of RGO. So, as a first step, WEP is mixed with water, and the mixture is magnetic stirred at a speed of 1000 r/min for 10 min. This yields an aqueous solution of WEP (S1) with low viscosity. Second step, the RGO and CNF powders are added in DI water, magnetic stirred at a speed of 2000 r/min for 5 min and then ultrasonic dispersed at 200 W for 1 h. Due to less surface group, low chemical activity, and high specific surface area of RGO, agglomeration phenomenon often occurs when RGO is dispersed in water. There are a large number of hydrophilic hydroxyl groups on the surface of CNF, which enable CNF to form stable and uniform suspension in water. Some unreduced hydroxyl and carboxyl groups are distributed on the surface of RGO. They can interact with the hydroxyl groups on the surface of CNF by hydrogen bonding. CNF acted as a template to help RGO disperse in water. Then, RGO-CNF suspension solution (S2) is obtained. CNF, on the one hand, acts as a dispersant for RGO, on the other hand, acts as a framework for supporting RGO, to promote the formation of three-dimensional reinforcement and conductive network in polymer matrix. The two prepared solutions, S1 and S2, are mixed in one container with magnetic stirring at a speed of 1000 r/min for 5 min and followed ultrasonic dispersion at 200 W for 3 h. The solution (S3) with RGO-CNF uniformly dispersed in WEP is obtained. Then, a curing agent is added in S3 at a mass ratio of 1 : 2. After 10 min of magnetic stirring at a speed of 1000 r/min, it is placed in the vacuum box for 30 min to remove bubbles. Subsequently, it is put in the oven to complete curing at 40°C for 3 hours and at 60°C for 24 hours. The choice of curing temperature is the key to the formation of a stable reinforcement and conductive network. If the temperature is high and the water evaporates too quickly, the network between graphene flakes will be destroyed, thus affecting the mechanical and electrical properties of the composites [24]. Table 1 is the filler percentage of samples, and the values of RGO and CNF are mass percent to WEP.

Table 1: Filler percentage of samples.

Sample number	RGO	CNF	WEP
WEP	0	0	100
A2/0	2	0	100
A4/0	4	0	100
B2/2	2	2	100
B4/2	4	2	100
B6/2	6	2	100
B8/2	8	2	100
B10/2	10	2	100

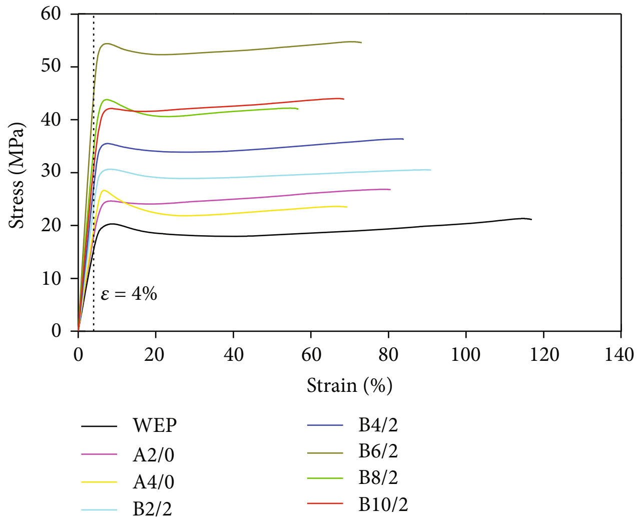


Figure 2: The stress-strain curves of the composites.

2.3. Test Methods of Composite Properties. The mechanical property of the composite was tested using a tensile instrument (HP-500, LANB Instrument, China) at a speed of 1 mm/min. The electrical resistance of the composite was measured by a 6.5-digit source meter (DMM6500, Keithley Instruments, USA) and the adopted voltage was AC 220 V. Copper foil electrodes were preembedded in the composite during fabrication process, and 2-prode method was used. The composite specimen was cut into a rectangular shape, and the distance between two electrodes was 100 mm. The width of the specimen was 10 mm and its thickness was 1 mm. At least three effective specimens were tested for each sample.

3. Properties of the RGO-CNF/WEP Composite

3.1. Mechanical and Electrical Properties. Figure 2 shows the stress-strain curve of the composites. It shows that the addition of RGO and CNF can significantly improve the tensile strength of epoxy. With the increase of RGO content, the tensile strength of the composites increases firstly and then decreases. The composites are in the elastic deformation stage within 4% strain range, and stress is proportional to strain. Their elastic moduli are calculated and the results can be seen in Table 2. When the content of RGO and CNF is 6 wt% and 2 wt%, respectively, the elastic modulus of the composite reaches the largest value of 12.02 GPa, which is in the same magnitude order with the elastic modulus of concrete.

With the help of CNF, the conductive filler RGO is evenly dispersed in epoxy. When the distance between RGO flakes is small enough, the "tunnel effect" occurs, which makes the RGO-CNF/WEP composite becomes electrically conductive. Table 2 shows the conductivity of the composites. The maximum conductivity is 3.4×10^{-1} S/m, and the content of RGO and CNF is 6 wt% and 2 wt%, respectively, which corresponds to the filler content of the composite with the maximum elastic modulus.

The improvement of mechanical and electrical properties of the RGO-CNF/WEP composites mainly depends on the binding state of RGO and CNF, the dispersion level of RGO-CNF in WEP, and the combination of RGO-CNF with WEP in micro level. As a green and renewable one-dimensional nanomaterial, CNF acts as dispersant and structural skeleton in the composite. RGO combines with CNF through hydrogen bond. The overlapping CNF can carry RGO to disperse evenly in the WEP matrix. It helps form a stable and continuous conductive network of RGO in the WEP matrix. Moreover, owing to the high strength and modulus of RGO and CNF, a cross-linking enhanced network is also constructed in the WEP matrix, which significantly improves the mechanical property of the composite. However, if RGO or CNF is added excessively, CNF could not carry redundant RGO to be uniformly dispersed into the WEP matrix, which destroyed the balance among the three components in the composite. The redundant RGO and CNF agglomerate in the WEP matrix, resulting in stress concentration which would decrease the mechanical property of the composite, meanwhile affecting the electrical property.

3.2. Electromechanical Properties of the RGO-CNF/WEP Composite. When the composite material is deformed by external force, the internal conductive network also deforms and the distance between RGO flakes changes, which makes the composite resistance change. Figure 3 shows the relative resistance change versus the strain of the RGO-CNF/WEP composite film. It can be seen that except for sample B2/2, the relative resistance change of other films has a good linear relationship with the strain. This was because the electrical network of sample B2/2 was imperfect at low RGO content. The phenomenon of resistance varying with strain belongs to piezoresistive effect. Gauge factor (GF) is usually used to evaluate this property of materials, and it associates resistance change rate with external strain, as shown in equation (1). The slope of fitting line of the curve in Figure 3 equals the GF of the film. The GFs of B4/2-B10/2 are 16-52, which are all obviously larger than the GF of traditional metal strain gauges (~2). Cycle tensile tests on specimens B6/2 and B10/2 showed good repeatability and GF remained basically unchanged, as shown in Figure 4. The results indicate that

Table 2: Elastic modulus and conductivity of samples.

Sample number	A2/0	A4/0	B2/2	B4/2	B6/2	B8/2	B10/2	WEP
Elastic modulus (GPa)	4.74	4.98	6.13	6.91	12.02	9.39	8.26	4.07
Conductivity (S/m)	6.2×10^{-8}	2.7×10^{-5}	2.1×10^{-3}	3.9×10^{-2}	3.4×10^{-1}	9.6×10^{-2}	1.5×10^{-1}	Nonconductive

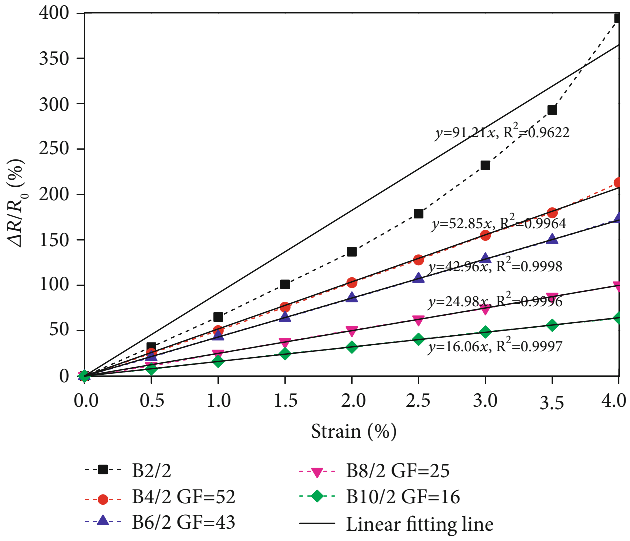


Figure 3: The relative resistance change versus the strain of the RGO-CNF/WEP composite films.

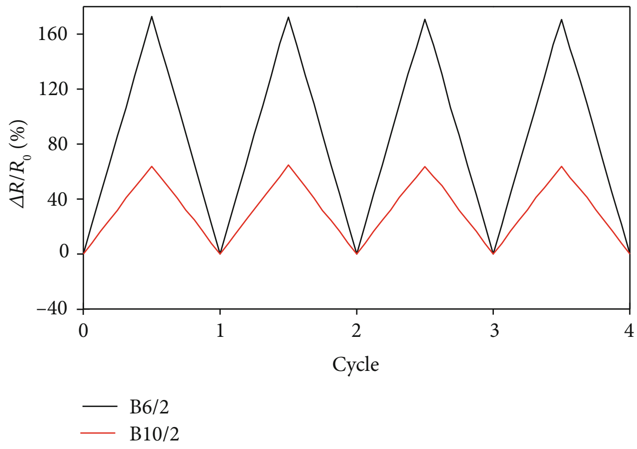


Figure 4: Cycle test of the RGO-CNF/WEP composite films.

the RGO-CNF/WEP composite has good strain sensing property and can be used to structural health monitoring.

$$\mathrm{GF} = \frac{\Delta R / R_0}{\varepsilon}, \qquad (1)$$

where ΔR is the relative resistance change and R_0 is the initial resistance of the composite film.

Strain sensor is widely used in structural damage detection and health monitoring. The traditional resistance strain sensor is mainly fabricated by metal or semiconductor materials. They have some defects such as small range, poor toughness, and easy to damage, which are unable to meet the needs in complex structures and large strain monitoring. From the above results, it can be seen that the films made of the RGO-CNF/WEP composite have good strain sensing performance with high sensitivity, good stability, and large measurement range. And the excellent mechanical property and corrosion resistance of epoxy can protect the film from environmental impact. In addition, it also has well plasticity to be made into the desired shape. Therefore, the RGO-CNF/WEP composite can be made into film strain gauges instead of metal or semiconductor strain gauges, which can be used to measure the strain on the surface of concrete structures. Film strain gauges with different GF can be obtained by adjusting the filler content of the composite to the needs of individual application, for example, high GF for low-strain applications and low GF for high deformation applications [25].

However, there is still a lack of effective means to measure the internal strain state of concrete. At present, most of the researches are to bury smart cement or polymer blocks into the structure and can only obtain strain data in a single direction. However, the internal stress state of concrete structure is complex, and it is difficult to get accurate results from one-dimensional sensors. Therefore, it is necessary to develop a new sensor which can measure 3D strain.

4. Principle and Design of a 3D Strain Rosette

4.1. Principle of a 3D Strain Rosette. The strain of a point in concrete under 3D state can be described by three normal strains and three shear strains, as shown in Figure 5. If the strain state of point A in Figure 5(b) is $(\varepsilon_x, \varepsilon_y, \varepsilon_z, \gamma_{xy}, \gamma_{yz}, \gamma_{zx})$, according to the strain theory, the linear strain ε in any direction through point A can be expressed as [26–28]

$$\varepsilon = \varepsilon_x l^2 + \varepsilon_y m^2 + \varepsilon_z n^2 + \gamma_{xy} lm + \gamma_{yz} mn + \gamma_{zx} nl, \qquad (2)$$

$$l = \sin\delta \cos\varphi, \qquad (3)$$

$$m = \sin\delta \sin\varphi, \qquad (4)$$

$$n = \cos\delta, \qquad (5)$$

where l, m, and n are directional cosines of line AB on x-, y-, and z-axes, respectively, δ is the angle between line AB and z-axis, and φ is the angle between x-axis and the projection of line AB on planar xAy.

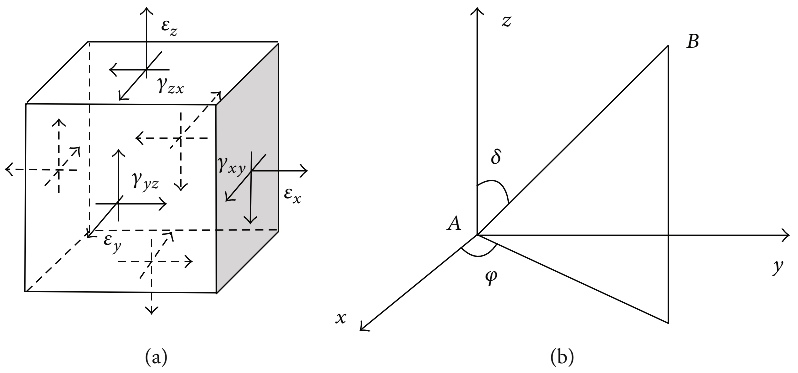


Figure 5: The strain of a point under three-dimensional state.

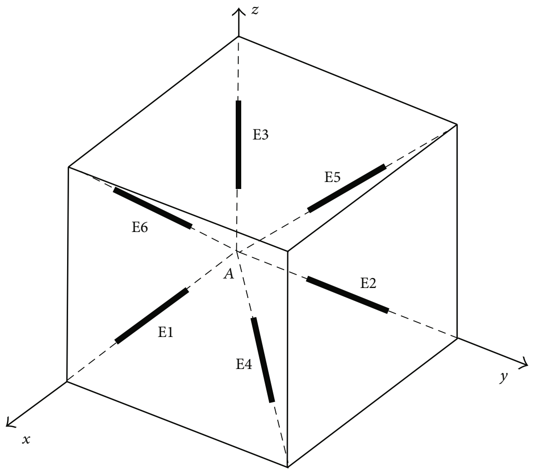


Figure 6: The structure of a 3D strain rosette.

For the selected direction, δ and φ are known quantities, then l, m, and n can be obtained by equations (3)–(5). Therefore, for equation (2), in order to solve the six unknown variables $(\varepsilon_x, \varepsilon_y, \varepsilon_z, \gamma_{xy}, \gamma_{zx}, \gamma_{yz})$, at least six different linear strains at this point need to be known. In theory, there can be multiple layouts for several strain gauges to form a 3D strain rosette. Figure 6 shows the structure of a sample 3D strain rosette. E1-E6 are six sensing elements (SE) through point A.

If the strains of E1-E6 are ε_i ($i = 1, 2, 3, 4, 5, 6$), the following can be obtained from equation (2):

$$\begin{Bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{Bmatrix} = \begin{bmatrix} l_1^2 & m_1^2 & n_1^2 & l_1 m_1 & m_1 n_1 & n_1 l_1 \\ l_2^2 & m_2^2 & n_2^2 & l_2 m_2 & m_2 n_2 & n_2 l_2 \\ l_3^2 & m_3^2 & n_3^2 & l_3 m_3 & m_3 n_3 & n_3 l_3 \\ l_4^2 & m_4^2 & n_4^2 & l_4 m_4 & m_4 n_4 & n_4 l_4 \\ l_5^2 & m_5^2 & n_5^2 & l_5 m_5 & m_5 n_5 & n_5 l_5 \\ l_6^2 & m_6^2 & n_6^2 & l_6 m_6 & m_6 n_6 & n_6 l_6 \end{bmatrix} \begin{Bmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \\ \gamma_{xy} \\ \gamma_{yz} \\ \gamma_{zx} \end{Bmatrix}. \tag{6}$$

Table 3: The directional cosines of the six sensing elements.

Sensing element	δ	φ	l	m	n
E1	90	0	1	0	0
E2	90	90	0	1	0
E3	0	0	0	0	1
E4	90	45	0.707	0.707	0
E5	45	90	0	0.707	0.707
E6	45	0	0.707	0	0.707

Setting

$$[K] = \begin{bmatrix} l_1^2 & m_1^2 & n_1^2 & l_1 m_1 & m_1 n_1 & n_1 l_1 \\ l_2^2 & m_2^2 & n_2^2 & l_2 m_2 & m_2 n_2 & n_2 l_2 \\ l_3^2 & m_3^2 & n_3^2 & l_3 m_3 & m_3 n_3 & n_3 l_3 \\ l_4^2 & m_4^2 & n_4^2 & l_4 m_4 & m_4 n_4 & n_4 l_4 \\ l_5^2 & m_5^2 & n_5^2 & l_5 m_5 & m_5 n_5 & n_5 l_5 \\ l_6^2 & m_6^2 & n_6^2 & l_6 m_6 & m_6 n_6 & n_6 l_6 \end{bmatrix}, \tag{7}$$

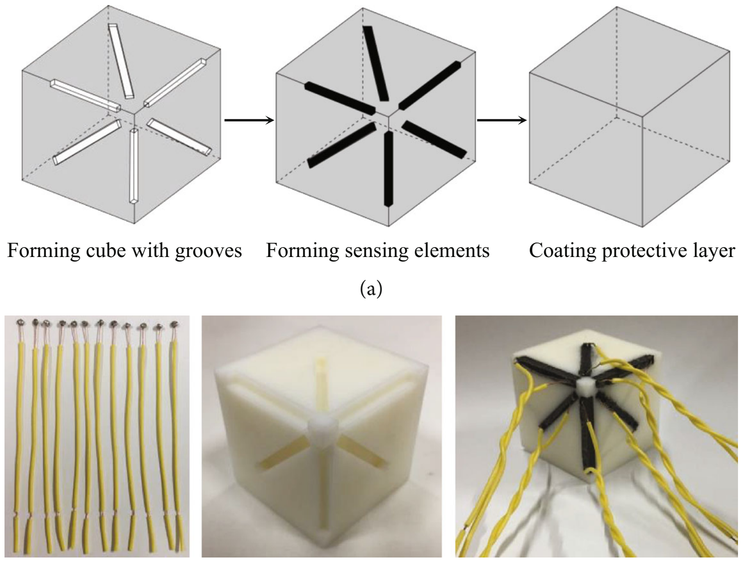
then

$$\{\varepsilon_i\} = [K]\{\varepsilon_j\}, \tag{8}$$

where $\{\varepsilon_i\} = \{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_6\}^T$, $\{\varepsilon_j\} = \{\varepsilon_x, \varepsilon_y, \varepsilon_z, \gamma_{xy}, \gamma_{yz}, \gamma_{zx}\}^T$.

Table 3 is the directional cosines of the six sensing elements in Figure 6. From equation (7), it can be obtained as follows:

$$[K] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 & 0.5 \end{bmatrix}. \tag{9}$$


Forming cube with grooves Forming sensing elements Coating protective layer

(a)

Electrodes Cube with grooves Cube with sensing elements

(b)

Figure 7: Fabrication of the 3D strain rosette sensor. (a) Fabrication process. (b) Some components.

Then

$$\{\varepsilon_j\} = [K]^{-1}\{\varepsilon_i\}. \tag{10}$$

4.2. Design and Fabrication of a 3D Strain Rosette. A 3D strain rosette sensor based on the RGO-CNF/WEP composite was fabricated. Figure 7 shows the fabrication process of the 3D strain rosette sensor. Using the good plasticity of WEP, six grooves were reserved as the layout of Figure 6 when forming a WEP cube with the size of 40 mm × 40 mm × 40 mm. The grooves were formed by strong magnetic strips attracted at the designed location of the mould. The size of the grooves is 30 mm × 3 mm × 3 mm. Then, the RGO-CNF/WEP composite was filled in the grooves. After the composite curing, six sensing elements were formed. The composite can be well joint with the WEP cube which ensures sensing elements and cube to deform together. At last, a WEP layer of about 1 mm thick was coated on the surface of the cube. The epoxy protective layer has good corrosion resistance and waterproof performance, which can prevent the damage of the internal sensing elements. It avoids the influence of external environment on the sensor and ensures the stability of the sensor performance.

4.3. Error Analysis. The total error of the sensor includes systematic error and random error. The errors that existed before the measurement, which will always affect the accuracy of the measurement results inevitability, are systematic errors, for example, the errors caused by the angle deviation between the sensing elements of the three-dimensional strain gauge. If the systematic error of each sensing element is $\Delta\varepsilon_i$, the systematic error of strain component can be obtained as follows:

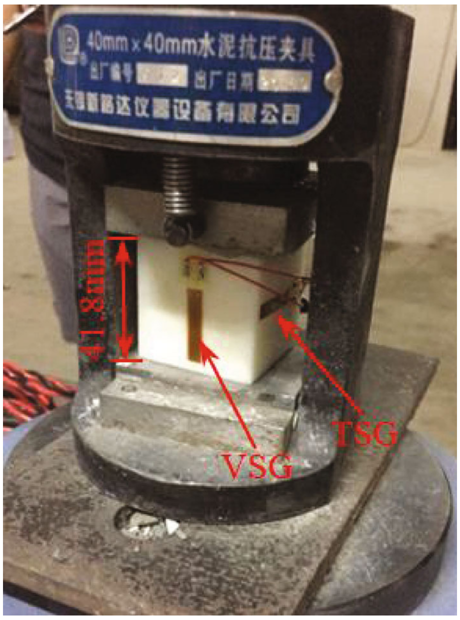
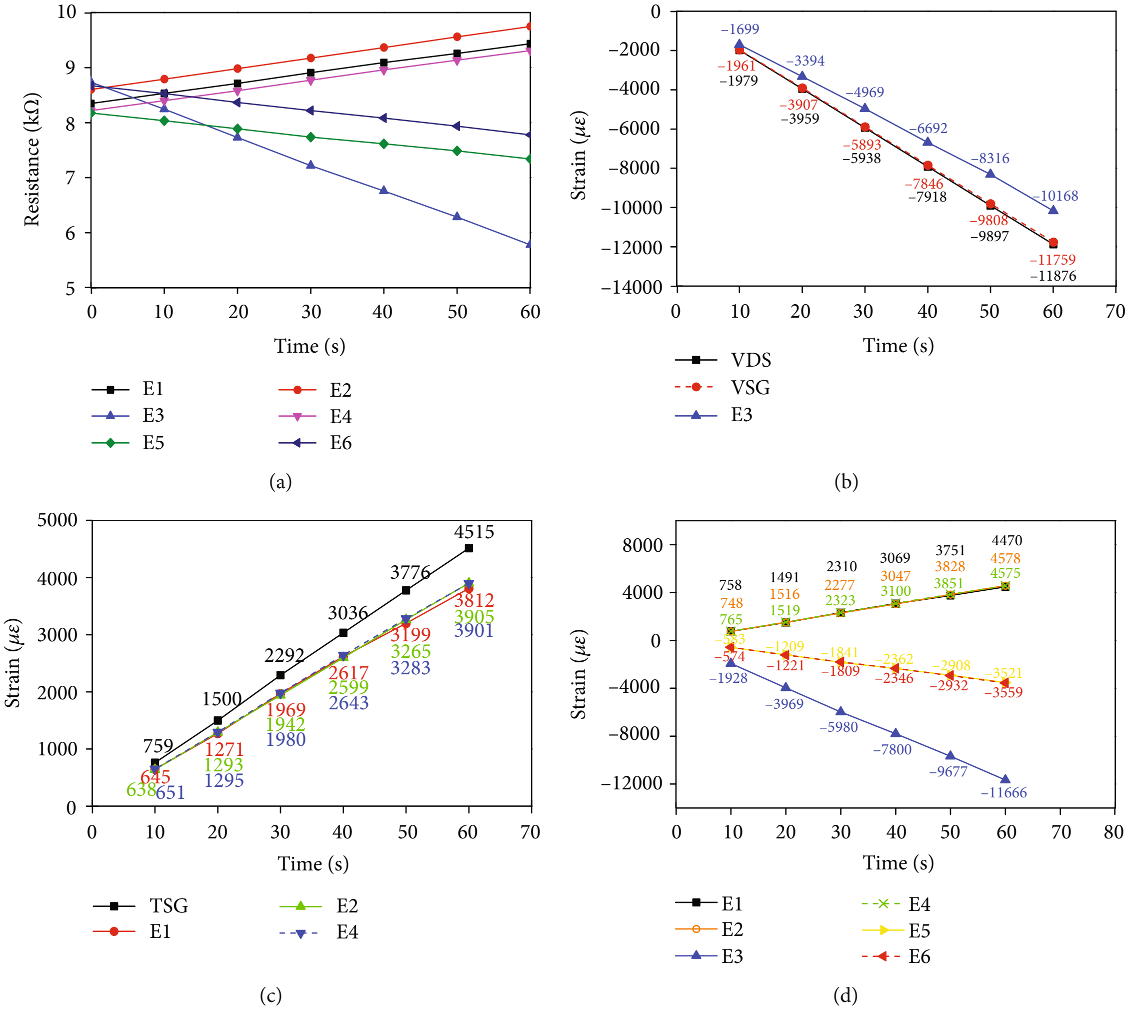


Figure 8: Test of the 3D strain rosette sensor.

$$\Delta\varepsilon_j = \sum_{i=1}^{6} k_{ji}\Delta\varepsilon_i, \tag{11}$$

where k_{ji} is the value of the matrix $[K]^{-1}$ at row j and column i.



(a)

(b)

(c)

(d)

FIGURE 9: Resistances and strains of SEs and strain gauges. (a) Resistances of SEs. (b) Vertical strains. (c) Transverse strains. (d) Strains of SEs.

TABLE 4: Slope of fitting line of strain.

Slope	Vertical strain	Transverse strain		
SL1	167.73 (E3)	64.2 (E1)	65.05 (E2)	65.47 (E4)
SL2	196.08 (VSG)	75.56 (TSG)		
SL1/SL2	0.86	0.85	0.86	0.87
		0.86 (average value)		

When repeated measurements of the same sensor are carried out with equal precision, a series of different results are obtained, and the deviations with actual value are random errors. If the random error of each sensing element is $R(\varepsilon_i)$, the random error of strain component can be obtained as follows:

$$R\left(\varepsilon_j\right) \le R\left(\varepsilon_i\right)\sqrt{\sum_{i=1}^{6} k_{ji}^2}. \tag{12}$$

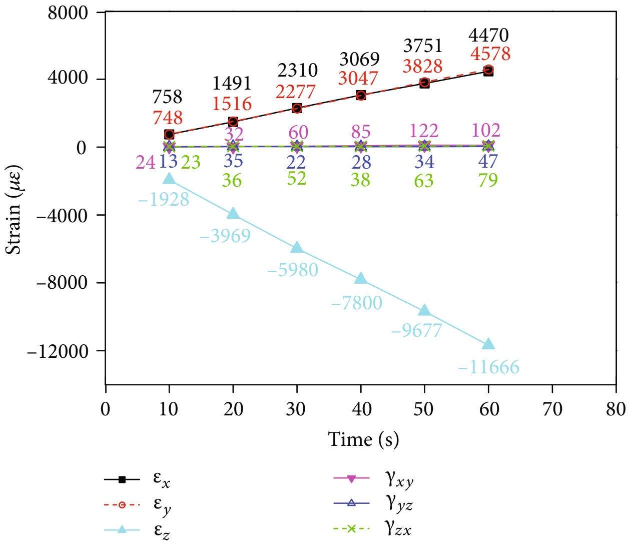


FIGURE 10: Three-dimensional strain state of point A.

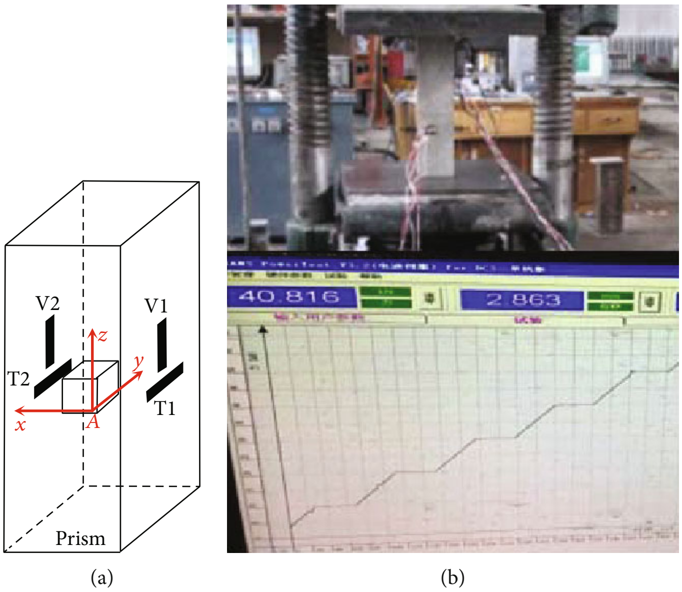


Figure 11: Sensor arrangement and load test. (a) Sensor arrangement. (b) Load test.

5. Performance Testing and Analysis of the 3D Strain Rosette Sensor

5.1. Test of the 3D Strain Rosette. Because of the good linearity and high sensitivity of the B6/2 RGO-CNF/WEP composite, it was selected to fabricate SEs of the 3D strain rosette sensor. As shown in Figure 8, the performance of the prepared sensor was tested. Axial pressure was applied by a pressure testing machine (TYA-300B, Wuxi Xinluda Instrument Equipment, China), and the loading speed was 0.5 mm/min. Two PTFE films were placed on the upper and lower surfaces of the machine to reduce friction between the sensor and the machine. The resistances of SEs were measured by the DMM6500 source meter. A vertical strain gauge (VSG) and a transverse strain gauge (TSG) were attached to one side of the sensor, and a static resistance strain indicator (JM3841, Yangzhou Jingming Technology Co., Ltd, China) was used to measure vertical and transverse strains of VSG and TSG. During the loading process, displacement, resistance, and strain were collected synchronously every 10 seconds for a total of 60 seconds. The coordinate system of the strain rosette is set as shown in Figure 6.

Figure 9(a) shows the resistance of SEs over time. The initial resistances of SEs vary slightly due to the difference in fabrication. GF of sample B6/2 is 34, and the strain of E3 can be obtained according to equation (1), as shown in Figure 9(b). The vertical displacement of the sensor was the same as that of the instrument. Then, the displacement was transformed into strain (VDS), and its change over time was shown in Figure 9(b). The strain of VSG was also shown in Figure 9(b). As can be seen in Figure 9(b), the line of VSG and VDS almost coincide, while the line of E3 deviates from them but has the same trend. It can be inferred that the strain of VSG approximates to the true vertical strain, while the strain of E3 needs to be modified. The strains of E1, E2, and E4 transformed from their resistances by equation (1) were shown in Figure 9(c). The strain of TSG was also shown in Figure 9(c). They all represent the transverse strain of the sensor. The strains of E1, E2, and E4 are close, but they have deviations from the strain of TSG. It is similar to the analysis of vertical strain. The difference between the measured strains of SEs and strain gauges is mainly caused by the GF value which is used to calculate strain from resistance. It is because that GF of the composite is shape-dependent. The length-thickness ratio of SEs is different with that of the tested specimens. So $\mathrm{GF}_c = 34$ is not the true value (GF_t) of SEs. According to equation (1), the ratio of calculated strain (ε_c) to true strain (ε_t) is as follows:

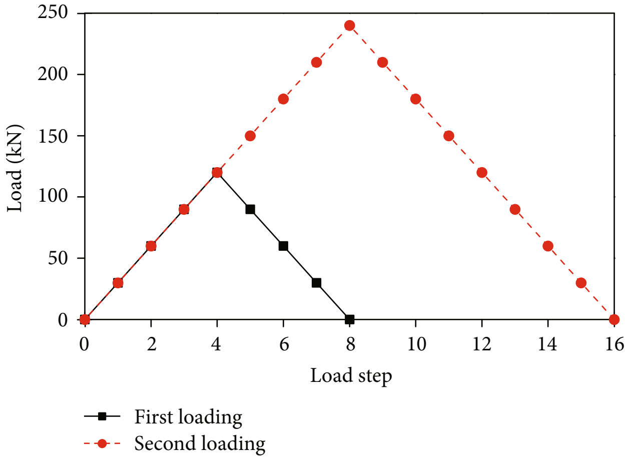
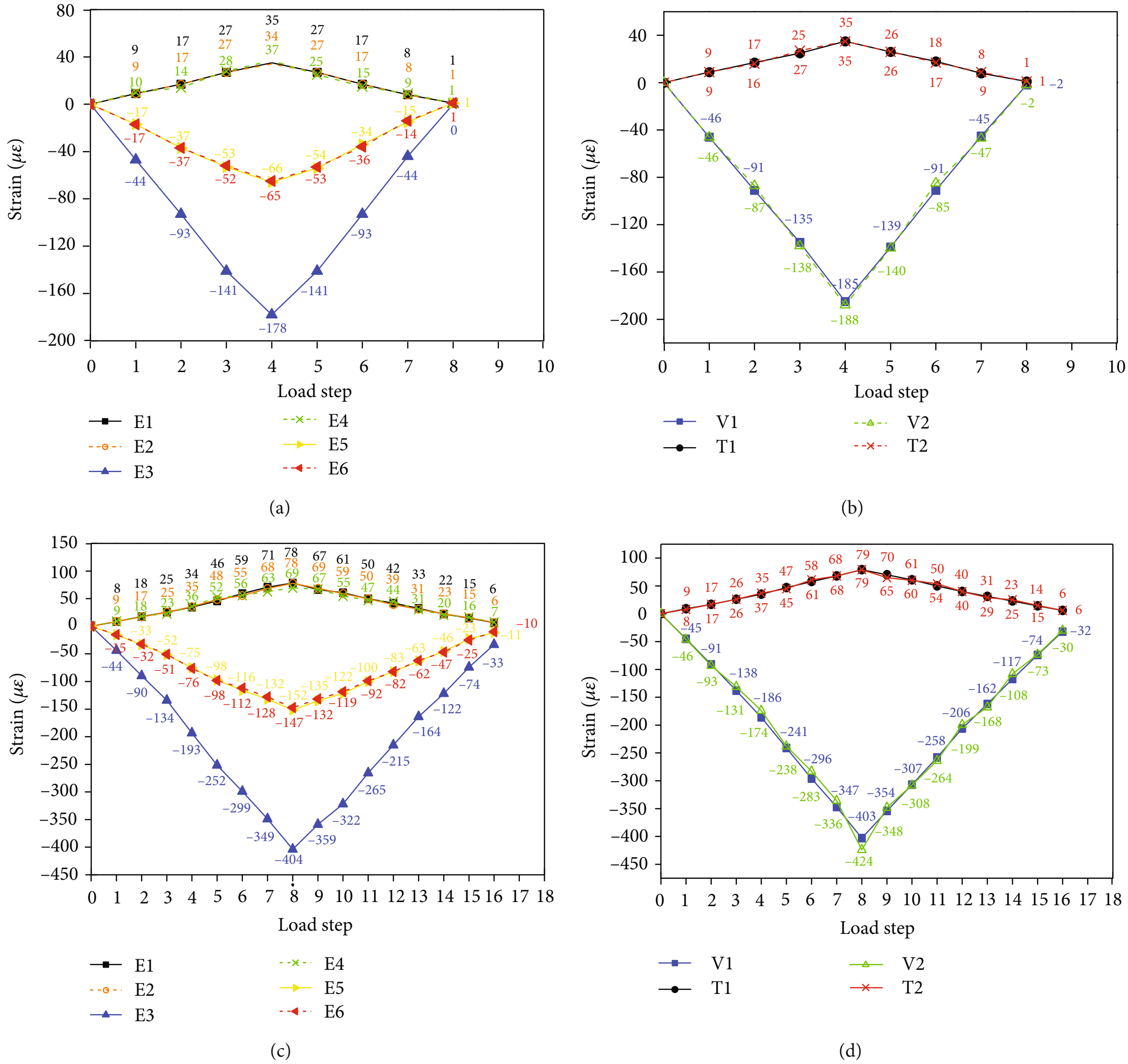


Figure 12: Loading history.

$$\frac{\varepsilon_c}{\varepsilon_t} = \frac{(\Delta R/R_0)/\mathrm{GF}_c}{(\Delta R/R_0)/\mathrm{GF}_t}, \tag{13}$$

$$\mathrm{GF}_t = \frac{\varepsilon_c}{\varepsilon_t} \cdot \mathrm{GF}_c. \tag{14}$$



(a)

(b)

(c)

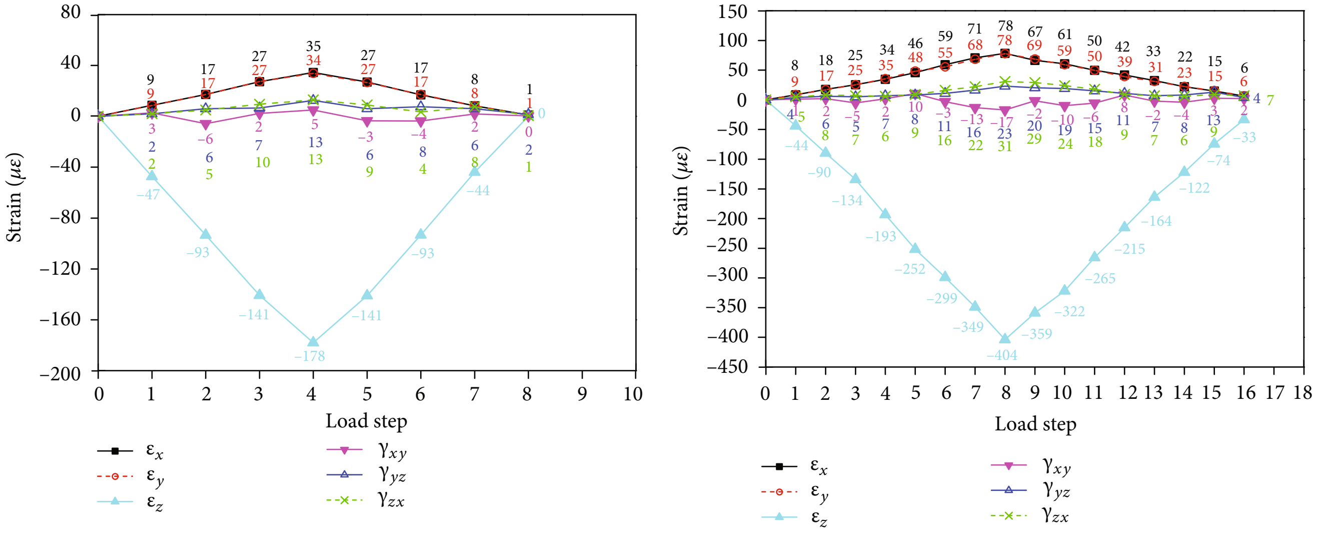
(d)

Figure 13: Measured strains of the 3D strain rosette and strain gauges. (a) Strains of rosette at first loading. (b) Strains of V1, T1, V2, and T2 at first loading. (c) Strains of rosette at second loading. (d) Strains of V1, T1, V2, and T2 at second loading.

The strains of VSG and TSG can be considered as true strains, and their values can be used to modify the GF of SEs. In Figures 8(b) and 8 (c), the strains of VSG, TSG, and SEs almost change linearly, and Table 4 shows the slope of the fitting line of the corresponding data. According to equation (14), $\mathrm{GF}_t = 0.86 \times 34 = 29$. In addition, it can be seen from the table that the ratio of TSG to VSG is 0.385, which is consistent with Poisson's ratio of the WEP matrix (0.39, provided by the material company).

Figure 9(d) shows the strain of SEs calculated using the modified GF_t. Figure 10 shows the 3D strain state of point *A* calculated from the strain of SEs according to equation (10). It can be seen that the vertical and transverse strains of point *A* coincide with those of VSG and TSG, and its shear strains are very small, relatively. The measured 3D strain state of point *A* is consistent with its actual strain state, which verifies the correctness and feasibility of the 3D strain rosette sensor.

5.2. Test of the 3D Strain Rosette Embedded in Concrete. To test the performance of the 3D strain rosette in concrete strain monitoring, a prism with a size of 150 mm × 150 mm × 500 mm was poured using C30 concrete (standard value of axial compressive strength $f_{\mathrm{ck}} = 20.1$ MPa). The 3D strain rosette was embedded in the prism, and its surface was polished so that it can bond well with the concrete. After standard curing, a loading test was performed on the prism by a pressure testing machine (WAW-600, Shanghai Hualong Testing Instrument, China). A vertical strain gauge (V1) and a transverse strain gauge (T1) were pasted on one



(a)

(b)

Figure 14: 3D strain state of point A. (a) First loading. (b) Second loading.

side of the prism. A vertical strip film (V2) and a transverse strip film (T2) made of the B6/2 composite were applied on the opposite side. The sensor arrangement and load test are shown in Figure 11. The axial pressure was applied step by step and unloaded step by step after reaching the predetermined value. Two times of loading were carried out. Figure 12 shows the loading history. The JM3841 static resistance strain indicator was used to measure the strains of the 3D strain rosette and strain gauges. The modified $\mathrm{GF}_t = 29$ was used for the 3D strain rosette.

Figures 13(a) and 13(c) show the strains of SEs of the 3D strain rosette. Figures 13(b) and 13(d) show the strains of V1, T1, V2, and T2. During the two loading processes, for longitudinal strain, the strain of E3 is close to that of L1, while for transverse strain, the strains of E1, E2, and E4 are close to that of T1. This indicates that there is a good bond between SEs and the WEP matrix and between the 3D strain rosette and the concrete. So they can deform synchronously.

According to equation (2), the 3D strain state of point A in concrete can be obtained, which is shown in Figure 14. It can be seen that the three normal strains are close to the corresponding strains of L1 and T1, and the three shear strains are small, which is consistent with the actual strain state of point A. But there are also some tiny differences, which may be caused by the following reasons: (i) Concrete is not a fully homogeneous material, so the stress transfer is not completely uniform. (ii) There are errors in manufacturing, such as deviations in the angle and size between SEs. There may also be deflection in the placed angle of the strain rosette. (iii) Poisson's ratio of the matrix material (0.39) of the strain rosette is different from that of the concrete (0.19). This will result in a difference in deformation of SE when the strain rosette is embedded in concrete or not. However, because of the large slenderness ratio of SE, the deformation of the width direction of SE has little effect on the deformation of the length direction.

From the two loading processes, the stress of the first loading is small, the concrete is in the stage of elastic deformation, the stress-strain relationship is close to a straight line, and the strain almost recovers to zero after unloading; the stress of the second loading is larger, the microcracks expand, and there is plastic strain after unloading. This is consistent with the measured results of the 3D strain rosette, which shows that the 3D strain rosette can be used to measure the internal strain of concrete. In addition, the strains of L1 and T1 are close to the strains of L2 and T2, respectively. It indicates that the composite film can also be used to measure the surface strain of concrete.

6. Conclusions

Modified polymer material with piezoresistive effect can be used to prepare new sensors for concrete structural health monitoring. Based on strain theory, a 3D strain rosette sensor was designed and fabricated using the RGO-CNF/WEP composites to obtain the strains in concrete structures. Firstly, the composite was prepared using RGO as the conductive filler, CNF as the dispersant and structural skeleton, and WEP as the polymer matrix. Then, a WEP cube was formed and six grooves were reserved in its six different directions. Subsequently, the RGO-CNF/WEP composite was filled in grooves to form six one-dimensional sensing elements. At last, a protective layer was coated on the cube surface and the 3D strain rosette was finished. Nanomaterials RGO and CNF with high strength and modulus form a stable and continuous three-dimensional reinforcing and conductive network in the WEP matrix, which significantly improves the mechanical and electrical properties of the composite. The GFs of the composites are 16-52, which are larger than the GF of traditional metal strain gauges. Within 4% strain, the sensing performance of the composites is stable with good linearity and repeatability. The 3D strain rosette was tested, and the measured strains are close to the actual strain state of measure point. The RGO-CNF/WEP composite has excellent mechanical and piezoresistive properties, which enable the fabricated 3D strain rosette to be used as an embedded sensor to measure the internal strain of concrete structures accurately. Moreover, the composite with good plasticity also can be made into film sensors to replace the traditional metal or semiconductor strain gauges for strain measurement on concrete surface.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The work was funded by the National Natural Science Foundation of China (Project Nos. 51778628 and 51578547).

References

[1] J. L. Humar and M. S. Amin, “Structural health monitoring,” *Structural Engineering Mechanics & Computation*, vol. 6531, no. 8, pp. 1185–1193, 2001.

[2] J. M. Ko and Y. Q. Ni, “Technology developments in structural health monitoring of large-scale bridges,” *Engineering structures*, vol. 27, no. 12, pp. 1715–1725, 2005.

[3] Y. Du, B. Sun, J. Li, and W. Zhang, “Cable Stress Monitoring Technology Based on Fiber Bragg Grating,” in *Optical Fiber Sensing and Structural Health Monitoring Technology*, pp. 249–269, Springer, Singapore, 2019.

[4] D. Kinet, P. Mégret, K. W. Goossen, L. Qiu, D. Heider, and C. Caucheteur, “Fiber Bragg grating sensors toward structural health monitoring in composite materials: challenges and solutions,” *Sensors*, vol. 14, no. 4, pp. 7394–7419, 2014.

[5] B. Dong, Y. Liu, L. Qin et al., “In-situ structural health monitoring of a reinforced concrete frame embedded with cement-based piezoelectric smart composites,” *Research in Nondestructive Evaluation*, vol. 27, no. 4, pp. 216–229, 2016.

[6] M. Sun, Q. Liu, Z. Li, and Y. Hu, “A study of piezoelectric properties of carbon fiber reinforced concrete and plain cement paste during dynamic loading,” *Cement and Concrete Research*, vol. 30, no. 10, pp. 1593–1595, 2000.

[7] S. Makireddi and K. Balasubramaniam, “A 1–3 piezoelectric fiber reinforced carbon nanotube composite sensor for crack monitoring,” *Journal of The Institution of Engineers (India): Series C*, vol. 97, no. 3, pp. 345–356, 2016.

[8] H. Shifeng, X. Dongyu, C. Jun, X. Ronghua, L. Lingchao, and C. Xin, “Smart properties of carbon fiber reinforced cement-based composites,” *Journal of Composite Materials*, vol. 41, no. 1, pp. 125–131, 2007.

[9] B. Han, X. Yu, K. Zhang, E. Kwon, and J. Ou, “Sensing properties of cnt-filled cement-based stress sensors,” *Journal of Civil Structural Health Monitoring*, vol. 1, no. 1-2, pp. 17–24, 2011.

[10] L. Zhang, S. Ding, L. Li et al., “Effect of characteristics of assembly unit of cnt/ncb composite fillers on properties of smart cement-based materials,” *Composites Part A: Applied Science and Manufacturing*, vol. 109, pp. 303–320, 2018.

[11] K. Loamrat, M. Sappakittipakorn, and P. Sukontasukkul, “Application of cement-based sensor on compressive strain monitoring in concrete members,” *Advanced Materials Research*, vol. 931-932, pp. 446–450, 2014.

[12] J. Olivera, M. González, J. Fuente, R. Varga, A. Zhukov, and J. Anaya, “An embedded stress sensor for concrete shm based on amorphous ferromagnetic microwires,” *Sensors*, vol. 14, no. 11, pp. 19963–19978, 2014.

[13] M. Schulz, Y. Song, A. Hehr, and V. Shanov, “Embedded carbon nanotube thread piezoresistive strain sensor performance,” *Sensor Review*, vol. 34, no. 2, pp. 209–219, 2014.

[14] F. Ubertini, A. L. Materazzi, A. D’Alessandro, and S. Laflamme, “Natural frequencies identification of a reinforced concrete beam using carbon nanotube cement-based sensors,” *Engineering Structures*, vol. 60, pp. 265–275, 2014.

[15] I. Kang, M. J. Schulz, J. H. Kim, V. Shanov, and D. Shi, “A carbon nanotube strain sensor for structural health monitoring,” *Smart Materials and Structures*, vol. 15, no. 3, pp. 737–748, 2006.

[16] A. R. Burton, J. P. Lynch, M. Kurata, and K. H. Law, “Fully integrated carbon nanotube composite thin film strain sensors on flexible substrates for structural health monitoring,” *Smart Materials and Structures*, vol. 26, no. 9, p. 095052, 2017.

[17] J. Herrmann, K. H. Müller, T. Reda et al., “Nanoparticle films as sensitive strain gauges,” *Applied Physics Letters*, vol. 91, no. 18, p. 183105, 2007.

[18] T. Yamada, Y. Hayamizu, Y. Yamamoto et al., “A stretchable carbon nanotube strain sensor for human-motion detection,” *Nature Nanotechnology*, vol. 6, no. 5, pp. 296–301, 2011.

[19] X. Xiao, L. Yuan, J. Zhong et al., “High-strain sensors based on ZnO nanowire/polystyrene hybridized flexible films,” *Advanced Materials*, vol. 23, no. 45, pp. 5440–5444, 2011.

[20] Z. Jing, Z. Guang-Yu, and S. Dong-Xia, “Review of graphene-based strain sensors,” *Chinese Physics B*, vol. 22, no. 5, 2013.

[21] X. Ji, H. Li, D. Hui, K. T. Hsiao, J. Ou, and A. K. T. Lau, “_I -V_ characteristics and electro-mechanical response of different carbon black/epoxy composites,” *Composites Part B: Engineering*, vol. 41, no. 1, pp. 25–32, 2010.

[22] K. S. Kim, Y. Zhao, H. Jang et al., “Large-scale pattern growth of graphene films for stretchable transparent electrodes,” *Nature*, vol. 457, no. 7230, pp. 706–710, 2009.

[23] H. P. S. A. Khalil, A. H. Bhat, and A. F. I. Yusra, “Green composites from sustainable cellulose nanofibrils: a review,” *Carbohydrate Polymers*, vol. 87, no. 2, pp. 963–979, 2012.

[24] Z. Wu, J. Wei, R. Dong, and H. Chen, “Epoxy composites with reduced graphene oxide–cellulose nanofiber hybrid filler and their application in concrete strain and crack monitoring,” *Sensors*, vol. 19, no. 18, p. 3963, 2019.

[25] Y. Liu, D. Zhang, K. Wang, Y. Liu, and Y. Shang, “A novel strain sensor based on graphene composite films with layered structure,” *Composites Part A: Applied Science and Manufacturing*, vol. 80, pp. 95–103, 2016.

[26] W. E. Baker and R. C. Dove, “Construction and evaluation of a three-dimensional strain rosette,” *Experimental Mechanics*, vol. 3, no. 9, pp. 201–206, 1963.

[27] S. Rossetto, A. Bray, and R. Levi, "Three-dimensional strain rosettes: pattern selection and performance evaluation," *Experimental Mechanics*, vol. 15, no. 10, pp. 375–381, 1975.

[28] E. G. Little, D. Tocher, D. Colgan, and P. O'Donnell, "An Analysis of the Factors Influencing the Data Derived from a Plug Type Three-Dimensional Strain Rosette under Compression and Torsion," *Strain*, vol. 41, no. 4, pp. 193–202, 2005.

Hindawi
Journal of Sensors
Volume 2019, Article ID 3140980, 13 pages
https://doi.org/10.1155/2019/3140980

Research Article

Detecting Gear Surface Defects Using Background-Weakening Method and Convolutional Neural Network

Liya Yu,[1] **Zheng Wang**,[1] **and Zhongjing Duan**[2]

[1] *School of Mechanical Engineering, Guizhou University, Guiyang 550025, China*
[2] *Key Laboratory of Advanced Manufacturing Technology of Ministry of Education, Guizhou University, Guiyang 550025, China*

Correspondence should be addressed to Zheng Wang; zhengwang0216@163.com

Received 15 July 2019; Accepted 17 September 2019; Published 19 November 2019

Academic Editor: Pawel Malinowski

A novel, efficient, and accurate method to detect gear defects under a complex background during industrial gear production is proposed in this study. Firstly, we first analyzed image filtering and smoothing techniques, which we used as a basis to develop a complex background-weakening algorithm for detecting the microdefects of gears. Subsequently, we discussed the types and characteristics of gear manufacturing defects. Under the complex background of image acquisition, a new model S-YOLO is proposed for online detection of gear defects, and it was validated on our experimental platform for online gear defect detection under a complex background. Results show that S-YOLO has better recognition of microdefects under a complex background than the YOLOv3 target recognition network. The proposed algorithm has good robustness as well. Code and data have been made available.

1. Introduction

In recent years, the demand for online quality inspection of mechanical parts under high-efficiency, high-precision manufacturing conditions has continued to grow with the rapid development of the manufacturing industry. Considering that a gear is a transmission part with a wide range of applications in the machinery industry, gear quality is particularly important in production. The development of the gear industry currently faces great challenges. Complex backgrounds, such as oil stains and dust particles, cannot be avoided in the gear manufacturing line. Identifying ways to accurately and efficiently identify gear surface defects in complex backgrounds and improve the quality inspection accuracy and production efficiency of gear production lines is important to advance the level of the manufacturing industry.

Traditional testing standards mainly detect the appearance size [1] and shape error [2] of parts, among which the error is maintained between 0.12 mm and 0.23 mm. In this paper, the gear defect is located by a deep learning algorithm, which lays a foundation for more precise quality inspection such as the subsequent dimension measurement. The traditional detection of gear manufacturing defect detection is based mainly on machine vision [3, 4], in which the contour extraction algorithm is often used to extract the image features of a single gear. After extracting the features, the gear is detected and checked via template matching. This method not only processes the image at a slow speed but also has low detection efficiency because only one gear sample can be detected in each feature image. In the case of insufficient illumination or complex background, the traditional visual detection method relies heavily on the light source, and the background-weakening effect is poor. As a result, detection accuracy is greatly reduced.

With the rapid development of deep learning in daily life [5–7] and industrial fields [8, 9], many scholars attempt to apply deep learning methods for detecting part defects [10]. To ensure the quality of online defect detection, the network must exhibit fast positioning speed and high classification accuracy. At present, the mainstream target recognition networks include You Only Look Once (YOLOv3) [11], FAST-RCNN [12, 13], SSD, and FPN [14]. No complicated computation is required because the YOLOv3 target detection network uses an end-to-end method to regress features. Previous research shows that YOLOv3 is faster than SSD, FPN, and other target recognition networks. However, the direct application of the YOLOv3 method to detect gear defects cannot satisfy the high accuracy requirements in industrial production. Therefore, the *k*-means clustering method is adopted to obtain the most suitable anchor to improve the positioning and detection accuracy of YOLOv3 for detecting gear defects. The background is weakened and denoised via image filtering and smoothing under the complex background of gear manufacturing, thereby improving the accuracy and detection efficiency of online gear defect detection. The proposed algorithm provides reference for the gear manufacturing industry to improve production efficiency, enhance product quality, and strengthen quality control capabilities.

A defect detection algorithm based on the deep learning algorithm of YOLOv3 for 62 gear line surface manufacturing is proposed in this study, which has the following main contributions:

(1) By analyzing image filtering and smoothing technology aimed at the microdefects of gears under a complex background, this study proposes a complex background-weakening algorithm based on image filtering and smoothing, which weakens the background noise of oil and dust, among others

(2) This study designs and opens source gear defect datasets for common defects, including missing teeth, broken teeth, surface scratches, and normal gear

(3) This study proposes an improved network for online gear defect detection called S-YOLO. This network is created by combining the types and characteristics of defects during the actual manufacturing of gears under the complex background of image acquisition on the factory production line. S-YOLO improves detection accuracy

The main structure of this paper is as follows. The second section mainly describes the related works on gear manufacturing defects and sorts the techniques of gear running fault and fatigue damage defect detection. The third section proposes a background-weakening algorithm for the complex background in gear manufacturing. The fourth section introduces the deep learning network target detection model, which is based on the YOLOv3 model for improvement and model training. The fifth section designs and manufactures an online detection platform for industrial defects. The sixth section designs and makes the gear defect dataset and compares and analyzes the experimental results. The final section summarizes the research content.

2. Related Work

In the research on fault diagnosis during gear runtime, Mączak and Jasiński [15] discussed the simulation model of the helical gearbox and analyzed a phenomenon during the tooth-meshing process in the presence of manufacturing and assembly errors. This work proposed a kind of gear fault diagnosis method based on the model. The detection method is simple, and the detection speed is fast. However, the effect of gear detection in large-volume motion on the production line is unknown. Gandarias et al. [16] took pressure reading as a standard image processing technique with the new high-resolution pressure sensor. It connects the tactile sensor with the robot detector with high resolution and realizes the image recognition of the contact object via a convolutional neural network (CNN) and migration learning. Lu et al. [17] applied the improved CNN model to an embedded system composed of signal acquisition and processing circuits and proposed a method for on-site motor fault diagnosis. A heterogeneous computing framework was proposed, and an integrated embedded system was designed based on the analysis of different motor signals. This method uses artificial intelligence technology to provide a solution for the field motor fault diagnosis on small, flexible, and convenient handheld devices. Cheng and Hu [18] proposed a method based on a physical model to detect the damage quantification of the planetary gear set. The performance of the feature in the damage evolution tracking was analyzed via the double-sample test method, and the state monitoring of the planetary gear transmission system was realized. Nabih et al. [19] experimentally verified the dynamic model of the single-stage gear transmission system and analyzed the effect of the perforation on TE. The results proved that a simple perforation model can reproduce the actual vibration caused by the failure of the perforation surface. Younes et al. [20] proposed a vibration acoustic signal analysis theory. The theory uses the feature extraction and classification of acoustic signals to accurately identify the defects of gears and bearings, but its algorithm cannot identify the exact location of the defects.

In research on gear defect detection through data acquisition and signal processing during gear operation, Zhao et al. [21] proposed a gearbox health evaluation framework based on R/C (run-up/coast-down) signal analysis by studying the mechanical vibration information. A feature enhancement scheme based on sparse guidance was proposed to extract the weak phase jitter associated with gear defects and detect the damage position of the gear. Kidar et al. [22] provided the crack characteristics in the vibration signal through the numerical model of the data. The analysis of the phase estimated using the Hilbert method and the signal parameters estimated via the sliding window-based rotation invariant technique were compared to achieve the detection of gear cracks. A sensor position optimization method based on finite element analysis and spectrum analysis was proposed in [23]. The existing two nonlinear models of mechanical rotating parts were solved, and the dynamic response of the whole system under defect excitation was used to determine the predictive maintenance for defect detection in the optimal sensor location. The defect of mechanical rotating parts was accurately detected. Moreno et al. [24] proposed various signal processing strategies for the detection and quantification of early gear defects. A comparison among the early detection capabilities of the microphone, accelerometer, and LDV sensors verified that the acoustic signal was the first method to detect the initial progressive crack of the gear (detecting a 1.3 mm long crack). Using a microphone signal had obvious advantages, but the result was sensitive to speed and torque. The pitting of gears was tested, and the vibration data was recorded in [25]. The application of vibration-based time, frequency, cepstrum, wavelet transform, and other methods in each set of experimental data, pitting fault, and the progress of pitting failure in gears were reviewed as well.

In research on detecting small defects of gears, Liu et al. [26] aimed to address the high cost, low efficiency, slow speed, and low precision of manual detection of automobile bevel gear surface defects and dimensional measurement. They studied and analyzed the three effective algorithms—neighborhood means difference method, circular approximation method, and fast rotation positioning method. A comprehensive bevel gear quality detection system was developed based on multicamera vision technology, which could simultaneously detect and measure the size of bevel gear surface defects. Fedala et al. [27] aimed to improve the detection and recognition ability of gear defects by extracting the features of the angular frequency domain of angular acceleration sampling, transmission error, and instantaneous angular velocity. SVM was then used to classify and realize gear fault detection under normal and nonstationary states. To isolate the defect signal from the measured signal, Djebala et al. [28] proposed a gear defect detection method based on wavelet multiresolution analysis and Hilbert transform. Experiments show that, in contrast with the commonly used analysis tools, this new method can isolate defect frequency, which enables the detection of small or combined defects. Focusing on the internal meshing gear defects, Zhang and Fan [29] proposed a universal formula for the identification and conducted the closed defects of the N-lobed noncircular gears (N-LNG) positioning function. The closed condition of the positioning function was satisfied by introducing two correction parameters: proportional and controllable. The controllable correction parameters were further verified and improved on the basis of the relationship between the inner pitch curve and the curvature radius of the outer pitch curve of the inner meshing of N-lobed noncircular gears. The method was applied in several numerical examples, and the simulation results showed that the method can effectively identify and conduct the closed defects of the N-LNG positioning function.

In the field of gear defect detection, many scholars conducted relevant theoretical research on gear operation faults, surface defects, and other aspects. However, research on surface manufacturing defects during the manufacturing of gears and high-speed online defect detection with numerous parts requires further improvement.

3. Complex Background-Weakening Algorithm

Substantial oil, dust, and other debris accumulate on the conveyor during gear production, and they complicate the background of the gear image sample to be tested. Accurately identifying the minor manufacturing defects on the gears, such as scratches and pinion broken teeth, is difficult. Such defects are called background noise. The images collected by the camera also generate noise due to the randomness of the photon flux and the fact that the gears are in motion on the conveyor belt. If the real pixel value $g_{r,c}$ is disturbed by the noise $n_{r,c}$, the gray value obtained is as follows:

$$\widehat{g}_{r,c} = g_{r,c} + n_{r,c}. \tag{1}$$

Noise $n_{r,c}$ is assumed to be smooth in the whole picture; that is, the noise is independent of the position of the pixels on the image. This noise, which is called stationary noise, is equally distributed for each pixel in the picture.

Two methods are commonly used to weaken the two kinds of noise in the picture collected during gear production: time-domain average denoising [30] and spatial average denoising [31]. Time-domain averaging captures and averages multiple images of the same scene. If n images are collected, then time-domain average is obtained as follows:

$$g_{r,c} = \frac{1}{n}\sum_{i=1}^{n} \widehat{g}_{r,c;i}, \tag{2}$$

where $\widehat{g}_{r,c;i}$ denotes the grayscale value at position (r, c) on the i image. The time-domain average method effectively reduces noise, and the variance of the noise is reduced to original $1/n$. To suppress noise, the method must collect images in the same scene. For online defect detection, the acquisition of multiple images in the same scene improves the accuracy of defect identification. However, it greatly increases the running time of the algorithm, thereby reducing the overall detection efficiency.

Therefore, the spatial average is used for denoising by taking a filter with a pixel of $(2n + 1) \times (2m + 1)$ and traversing the same image. Depending on the operation and the filter, the filtering algorithm includes meaning filtering, block filtering, Gaussian filtering, and median filtering. Among them, mean filtering and Gaussian filtering are the most commonly used filtering algorithms. Mean filtering can be expressed as

$$h_{r,c} = \begin{cases} \dfrac{1}{(2n+1)\times(2m+1)} |r| \le n \wedge |c| \le m \\ 0 \text{ Other} \end{cases}, \tag{3}$$

where (r, c) denotes the pixel position of the image and m and n are the parameters that determine the length and width of the filter.

If the original image matrix is

$$A = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,k} \\ \vdots & & \ddots & \vdots \\ \alpha_{j,1} & \alpha_{j,2} & \cdots & \alpha_{j,k} \end{bmatrix}, \tag{4}$$

then the filtered matrix is

$$B = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,k} \\ \vdots & & \ddots & \vdots \\ \beta_{j,1} & \beta_{j,2} & \cdots & \beta_{j,k} \end{bmatrix}. \tag{5}$$

In the actual operation process, the input image is usually a square, so $m = n$. The pixel $\alpha_{p,q}$ in matrix A is then processed through the mean filter with the size of $(2n+1) \times (2n+1)$ to obtain $\beta_{p,q}$ in B:

$$\beta_{p,q} = \begin{bmatrix} \alpha_{p-n,q-n} & \cdots & \alpha_{p-n,q+n} \\ \vdots & \ddots & \vdots \\ \alpha_{p+n,q-n} & \cdots & \alpha_{p+n,q+n} \end{bmatrix}_{(2n+1)(2n+1)} \times \frac{1}{(2n+1)^2} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{(2n+1)(2n+1)}. \tag{6}$$

As shown in Formulas (3) and (6), the averaging filter actually averages the pixels in the effective calculation range and assigns them to the middle value of the filtering window. For the oil stain and dust background of the gear production workshop, the mean filter averages pixel values, such as oil and dust, with the surrounding background pixels. It blurs the oil, dust, and other small particles. It also highlights the position and feature information of the gear in the whole image to prepare for subsequent feature extraction.

Although the mean filter weakens small particles, such as oil stains and dust in the background, most of the stationary noises in the image due to the principle of lens imaging appear in the form of high-frequency fluctuation of gray value. The suppression of high-frequency noise via filtering is not satisfactory. Therefore, to maximize the suppression of the influence of high-frequency stationary noise, the Gaussian filter is used for secondary image smoothing. As such, the eigenvalue of the processed image becomes easy to extract. The 1D Gaussian filter can be expressed as

$$g_\sigma(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/(2\sigma^2)}. \tag{7}$$

The two-dimensional Gaussian filter applied to image processing can be expressed as

$$g_\sigma(r, c) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(r^2+c^2)/(2\sigma^2)} = g_\sigma(r) g_\sigma(c). \tag{8}$$

After the first mean filtering of complex background images, the effect of oil, dust, and other abrupt noises in the background is weakened. The high-frequency noise in the complex background of the image is weakened after the second Gaussian filtering, and the gear body in the relative image is highlighted, allowing for the easy extraction of the gear body's features.

	Type	Filters	Size	Output
	Convolutional	32	3 x 3	256 x 256
	Convolutional	64	3 x 3 /2	128 x 128
1 x	Convolutional	32	1 x 1	
	Convolutional	64	3 x 3	
	Residual			128 x 128
	Convolutional	128	3 x 3 /2	64 x 64
2 x	Convolutional	64	1 x 1	
	Convolutional	128	3 x 3	
	Residual			64 x 64
	Convolutional	256	3 x 3 /2	32 x 32
8 x	Convolutional	128	1 x 1	
	Convolutional	256	3 x 3	
	Residual			32 x 32
	Convolutional	512	3 x 3 /2	16 x 16
8 x	Convolutional	256	1 x 1	
	Convolutional	512	3 x 3	
	Residual			16 x 16
	Convolutional	1024	3 x 3 /2	8 x 8
4 x	Convolutional	512	1 x 1	
	Convolutional	1024	3 x 3	
	Residual			8 x 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure 1: YOLOv3 basic network Darknet-53 [11].

4. Improved Construction and Training of YOLOv3 Network

4.1. Characteristics of YOLOv3 Network Structure. The YOLOv3 network model uses an end-to-end network architecture implemented in a CNN. The basic network structure is shown in Figure 1.

Its network first divides the input image into $S \times S$ grids and the image by clustering. If the center point of an object in the image falls in the YOLO-divided grid, then the grid is responsible for predicting the object. Each grid is responsible for predicting B bounding boxes and the confidence of the bounding boxes. The confidence reflects the probability of containing objects in the bounding box predicted by the network model and the accuracy of the predicted position of the bounding box, which can be expressed as

$$\text{Confidence} = \Pr(\text{Object}) \times \text{IOU} \frac{\text{truth}}{\text{pred}}, \tag{9}$$

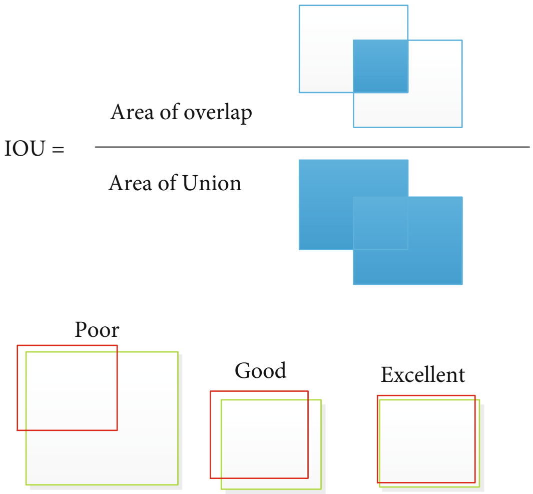


Figure 2: IOU evaluation diagram.

where IOU (Intersection over Union) represents the intersection ratio of the real target bounding box and the predicted target bounding box, which can be represented by Figure 2. If an object exists in the grid, $\Pr(\text{Object}) = 1$, then

$$\text{Confidence} = \text{IOU}\frac{\text{truth}}{\text{pred}}. \tag{10}$$

Otherwise, $\Pr(\text{Object}) = 0$, that is,

$$\text{Confidence} = 0. \tag{11}$$

In the YOLOv3 network, each bounding box predicts five values, including (x, y, w, h) and confidence, where x, y represents the coordinates of the center point of the predicted bounding box and w, h are the width and the height of the bounding box. Confidence is the IOU that predicts the bounding and the real bounding boxes.

Each grid predicts the probability of C condition categories, that is, the probability of the mesh containing objects belonging to a certain category. $\Pr(\text{Class}_i \mid \text{Object})$. Finally, the conditional probability is multiplied by confidence, and the probability that a certain type of object appears in the box and the degree of fit of the bounding box to the object are obtained:

$$\begin{aligned}\Pr(\text{Class}_i \mid \text{Object}) \times \Pr(\text{Object}) \times \text{IOU}\frac{\text{truth}}{\text{pred}} \\ = \Pr(\text{Class}_i) \times \text{IOU}\frac{\text{truth}}{\text{pred}}.\end{aligned} \tag{12}$$

In the design of loss function, the YOLO network takes the form of a weighted summation of the partial loss functions. By weighing the coordinate error, IOU error, and classification error and summing them, the total loss function is calculated and can be expressed as

$$\text{loss} = \sum_{i=0}^{s^2} \text{coordErr} + \text{iouErr} + \text{clsErr}. \tag{13}$$

The loss of the predicted center coordinates is expressed as

$$\lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \ell_{ij}^{\text{obj}} \left[(x_i - x_i\wedge)^2 + (y_i - y_i\wedge)^2 \right]. \tag{14}$$

The loss of the width and the height of the predicted bounding box is expressed as follows:

$$\lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \ell_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{w_i\wedge}\right)^2 + \left(\sqrt{h_i} - \sqrt{h_i\wedge}\right)^2 \right] \tag{15}$$

where λ_{coord} denotes the weight factor of the coordinate error in the overall loss function.

The loss made to the forecast category is expressed as

$$\sum_{i=0}^{s^2} \ell_i^{\text{obj}} \sum_{j=0}^{B} \left[(p_i(c) - p_i\wedge(c))^2 \right]. \tag{16}$$

The loss of confidence in the prediction is expressed as follows:

$$\sum_{i=0}^{s^2} \sum_{j=0}^{B} \ell_{ij}^{\text{obj}} \left[(c_i - c_I\wedge)^2 \right] + \lambda_{\text{noobj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \ell_{ij}^{\text{obj}} \left[(c_i - c_I\wedge)^2 \right], \tag{17}$$

where C is the confidence score; $\widehat{C}$ is the intersection of the predicted bounding box and the basic fact, when an object exists in a cell; and ℓ_{ij}^{obj} is equal to 1; otherwise, it is 0; λ_{noobj} represents the confidence weight when no object exists in the bounding box [10].

4.2. Improved YOLOv3 Network. The original YOLOv3 network uses a CNN, so the image is extracted through multiple convolutional layers for abstract feature extraction. Finally, the image is classified and predicted. Combining the types and characteristics of defects during actual gear manufacturing and the complex background of image acquisition on the factory production line, an improved online defect detection network for YOLOv3 gear is proposed. This network is called S-YOLO, which stands for smoothing-YOLOv3. The network structure is shown in Figure 3.

In the network structure of S-YOLO, the end-to-end Darknet-53 convolutional network formed in YOLOv3 is maintained. Moreover, an image-smoothing layer is added at the front end of the network to weaken the background noise of gear image collection during production.

In the smoothing layer, an average filter with pixel 8 × 8 is used to filter and smoothen the collected image for the first time. This process is aimed at weakening the influence of impurities, such as oil and fine dust particles, in the image. A Gaussian filter with a pixel of 3 × 3 is then used for the secondary smoothing of the image. This filter mainly reduces

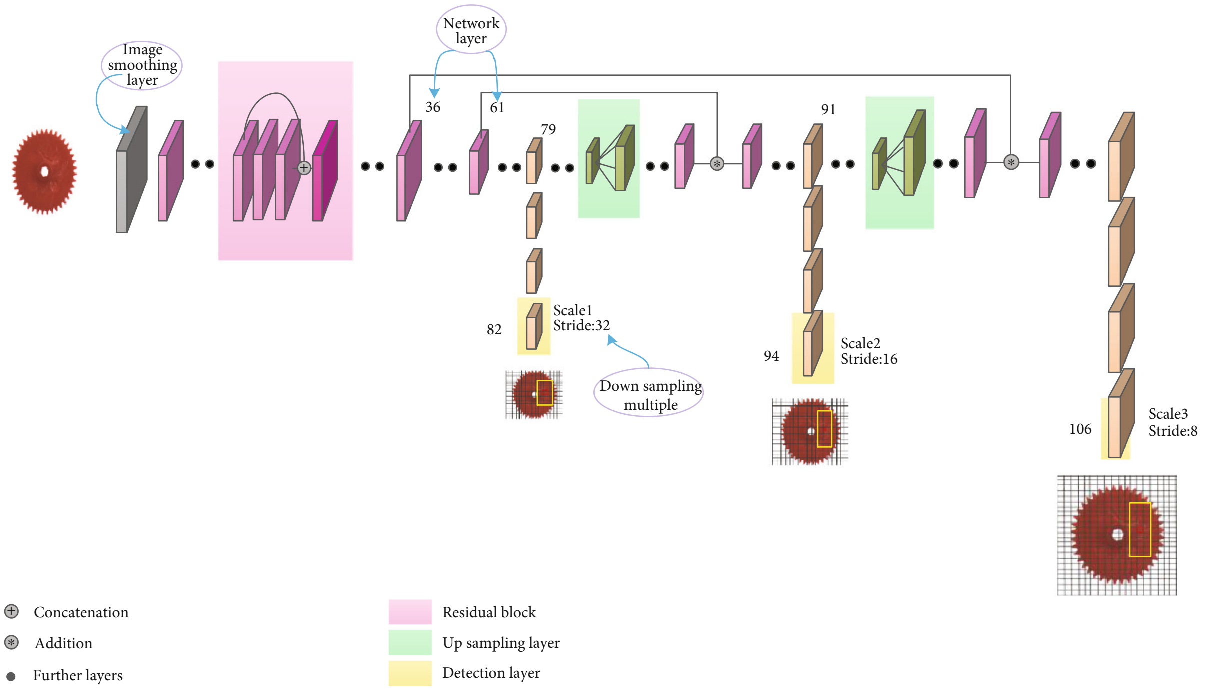


Figure 3: S-YOLO detection principle.

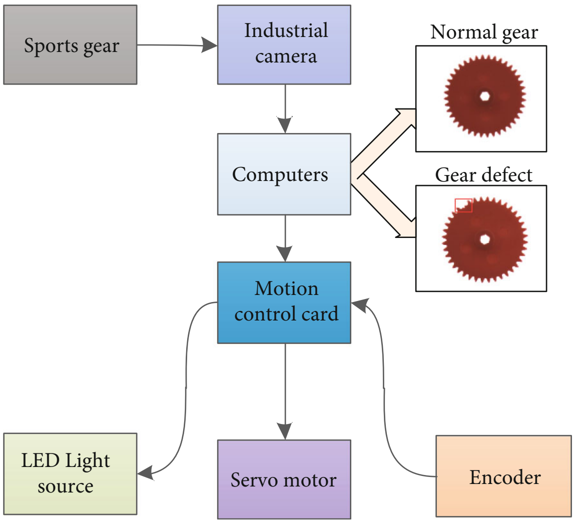


Figure 4: Flow chart of the experimental platform.

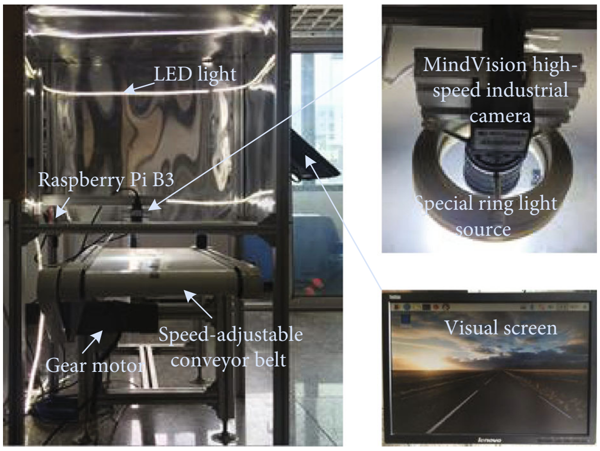
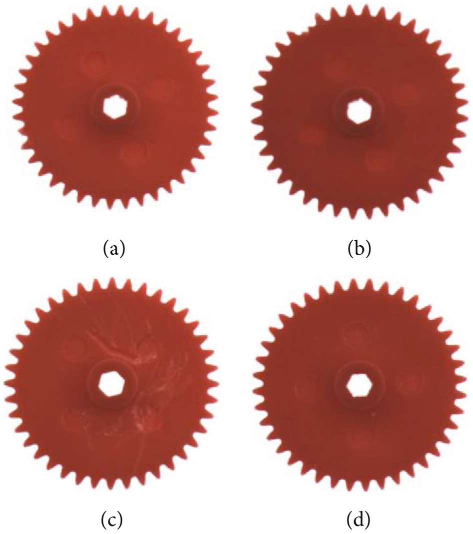

LED light
Raspberry Pi B3
Gear motor
Speed-adjustable conveyor belt
MindVision high-speed industrial camera
Special ring light source
Visual screen

Figure 5: Experimental platform.

the high-frequency noise in the collected image and further reduces the influence of oil, dust particles, and other impurities in the gear production workshop.

After passing the smooth layer, the pixel size and gear defect characteristics remain unchanged. The following YOLOv3 network uses three different scale feature maps for defect detection. As shown in Figure 3, a scale detection result is obtained through several yellow convolution layers after the 79th convolutional layer. The input image size during the experiment is 416 × 416 pix. Hence, the feature image pixel size at this time is 13 × 13 pix. The receptive field of the feature map is relatively large at this time because the downsampling factor is high, which is suitable for detecting relatively large defect size in the image. The network starts upsampling from the feature map of the 79th layer. It then fuses with the 61st layer feature map to obtain the 91st layer of the finer-grained feature map. After several convolution layers, the feature map 16 times of the input image is obtained. It has a medium-scale receptive field and is suitable for detecting objects with medium defect size. Finally, the 91st layer feature map is again upsampled and merged with the 36th layer feature map to obtain a feature map that is downsampled for eight times from the input image. It has the smallest receptive field and is suitable for detecting small defect sizes.

4.3. k-Means Clustering-Based A Priori Box Acquisition. Although the YOLO network itself can improve the value



(a) (b)

(c) (d)

Figure 6: Common defects in the gear manufacturing process. (a) Break. (b) Lack. (c) Scratch. (d) Normal.

Table 1: Distribution of common manufacturing defect datasets for gears.

Defect type	Broken tooth		Missing tooth		Scratch		Normal	
Number of images	Test	Train	Test	Train	Test	Train	Test	Train
	100	900	100	900	100	900	100	900

of the IOU and constantly adjust the size of the bounding box via training, allowing the network to modify through a large amount of data will slow down the network training and prevent the value of the IOU from gaining substantial improvement. With the gear training dataset as a basis, the k-means method is used to find the anchors of the a priori box that best fits the size of the gear defect. The standard k-means method uses Euclidean distance, and this usage will result in large boxes that generate more errors than small boxes. Therefore, Formula (18) is used to represent the distance and obtain a large IOU value in network prediction:

$$d(\text{box}, \text{centroid}) = 1 - \text{IOU}(\text{box}, \text{centroid}). \tag{18}$$

5. Online Platform for Industrial Defect Detection

Figure 4 depicts the system flow chart of the online testing platform for gear manufacturing defects designed by the research group. Figure 5 is an online test platform for gear manufacturing defects built by the research team [10, 32]. This platform includes the conveyor belt, data processor, data acquisition sensor, light source, and other mechanical supports, wherein the touch display for inputting and displaying data is the 32-inch industrial touch screen. The vision sensor device uses the MindVision high-speed industrial camera with an electronic rolling shutter, which can collect high-speed moving samples for real-time testing. The data processor is the Raspberry Pi B3. To ensure sufficient light in the system box, a band-shaped ambient light source LED with adjustable brightness is installed. A dedicated circular light source of Microscope LED Ring Light is installed outside the industrial camera to fill the test sample with light and to obtain a clear sample image. The device uses a variable speed motor to drive the conveyor belt. The outside of the box is equipped with a display for visualizing the test results. The Dell workstation of GPU1080 graphics card, which is mainly used for data analysis, is used to reduce the computational load of data processor. At the same time, Raspberry PI B3 has a wireless communication module, which can realize end-to-end communication between the test experimental platform and the workstation. The SQL SERVER 2008 R2 database is installed on the workstation to realize real-time local data capturing and automatic real-time data storage in the cloud.

The gear is transported to the field of view of the industrial camera's lens through the conveyor belt. After detecting the gear passing, the fiber optic sensor sends a trigger pulse to the image acquisition part. The image acquisition part then sends a start pulse to the industrial camera and the illumination system according to the preset program and delay. Industrial cameras begin to capture images, and the Microscope LED Ring Light's dedicated ring light source provides illumination that matches the exposure time of the industrial cameras. After capturing the image, the image acquisition of the camera receives the analog signal and digitizes it via an analog to digital conversion. The image acquisition part stores the digital image in the processor or computer memory. The processor then processes, analyzes, and recognizes the collected gear image. It then obtains and saves the detection result.

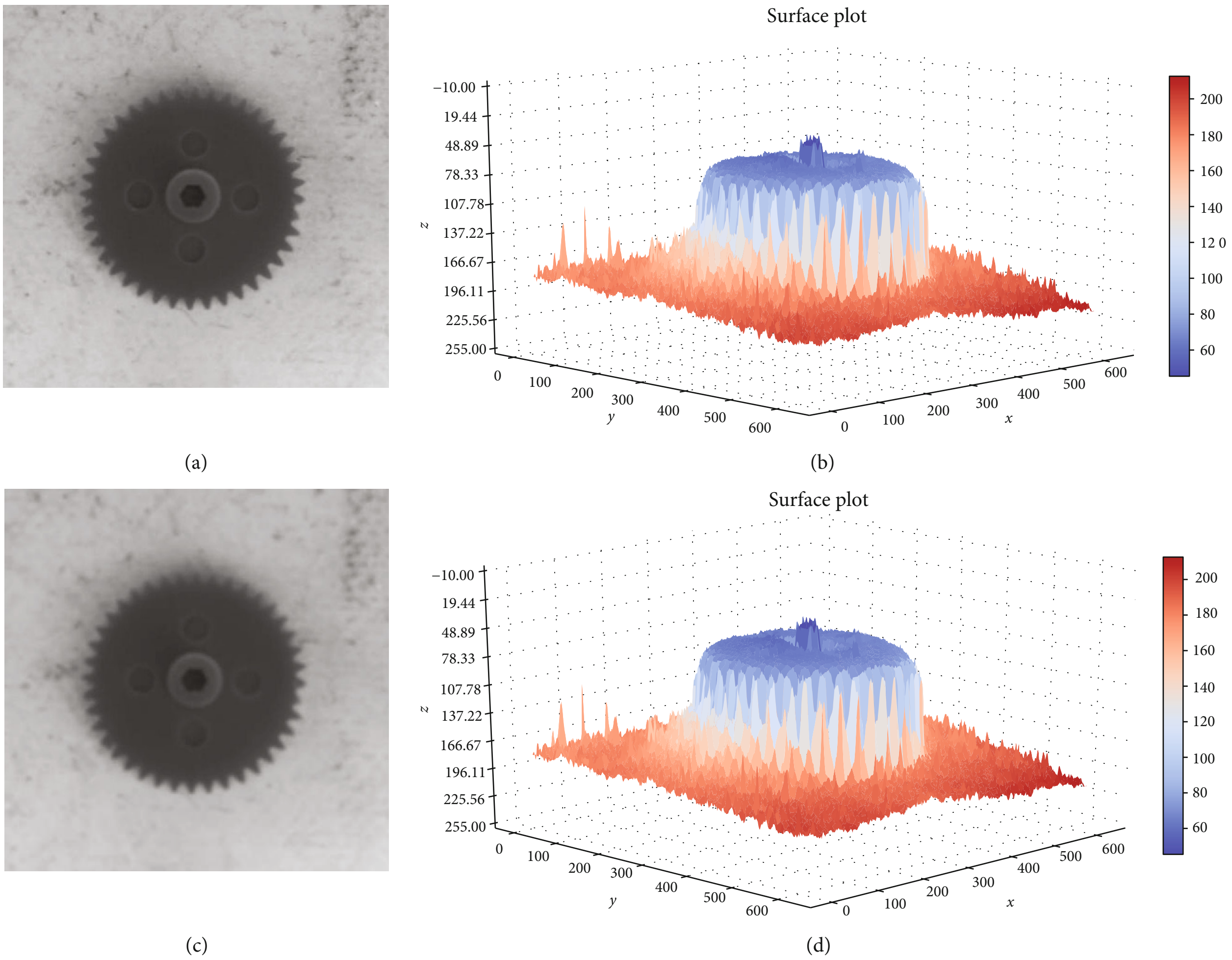


Figure 7: Comparison of image smoothness after mean filtering. (a) Original grayscale image. (b) The pixel gradient of the original grayscale image. (c) Image after mean filtering. (d) Image pixel gradient after mean filtering.

6. Experimental Results and Analysis

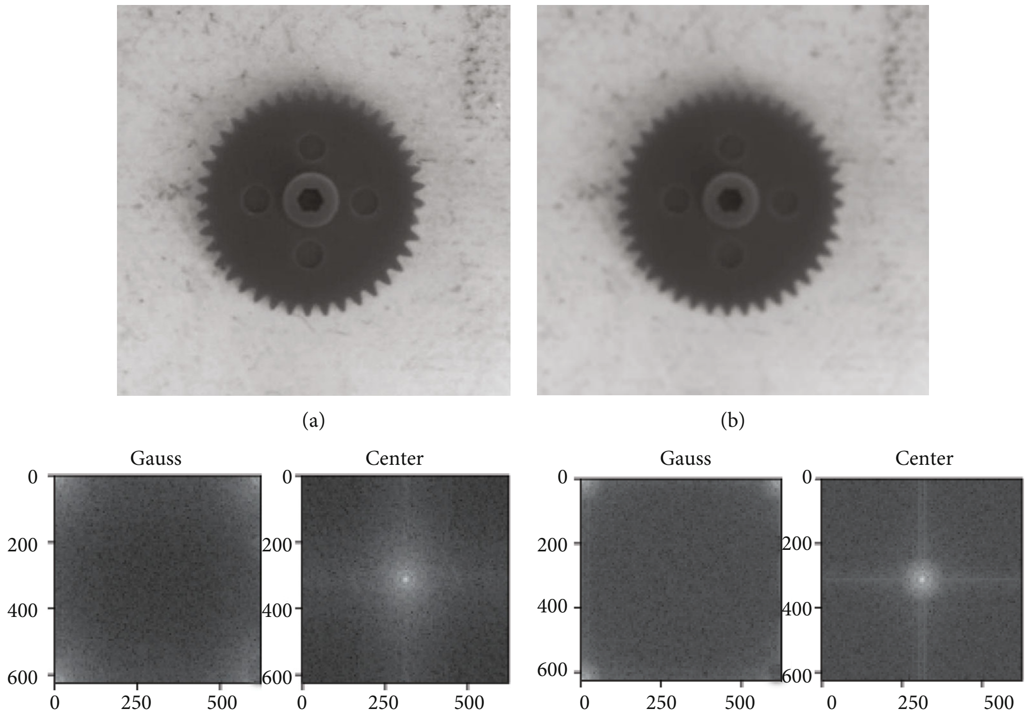
6.1. Production of Gear Datasets. During gear manufacturing, the bluntness of the turbine hob or the uneven material of the gear billet often causes gear tooth surface tear, tooth fracture, and gear surface scratches, among others, as shown in Figure 6.

Gear defect datasets $P = \{L, B, S, N\}$ are collected according to the types of defects commonly found in gear production. The four types of datasets are broken tooth image set $B = \{B_1, B_2, \cdots, B_{300}\}$, missing tooth image set $L = \{L_1, L_2, \cdots, L_{300}\}$, gear surface scratch image set $S = \{S_1, S_2, \cdots, S_{300}\}$, and normal image set $N = \{N_1, N_2, \cdots, N_{300}\}$.

Data enhancements can enrich small datasets or poorly diverse datasets. Common data enhancement methods include color jittering, PCA jittering, random scale, random crop, and horizontal/vertical flip. After collecting 300 pieces of gear data for each type of gear through industrial cameras, the images are rotated at random angles to achieve data enhancement. Finally, 1000 pieces of image data for each type are obtained, thereby collecting a total of 4000 pieces of gear image data. The specific data distribution is shown in Table 1.

6.2. Double Filtering Background Weakening. The effect of mean filtering on image noise removal in the complex background is considered. As shown in Figure 7, the original grayscale image has background noises, such as dust and oil stains, as illustrated in Figure 7(a). These noises have a certain influence on the later gear feature extraction. After the mean filtering operation, as shown in Figure 7(c), the background noise is partially weakened, and the degree of weakening depends on the convolution kernel size of the mean filter. After the mean filtering operation, the entire part of the gear still has all the features required for defect detection. As indicated in the comparison between Figures 7(b) and 7(d), the smoothness of the image increases after mean filtering. Moreover, the overall pixel gradient tends to be smooth, which is a good data condition for defect recognition and classification via the deep learning algorithm.

The comparison in Figure 8 shows that the high-frequency noise in the image is suppressed after the secondary filtering by the Gaussian filter, and the low-frequency part of the image is highlighted. Thus, the effect of the main part of the gear in the protruding image is achieved, which lays the foundation for the following feature extraction.



(a)

(b)

(c)

(d)

Figure 8: Comparison with the Fourier transform of the Gaussian filtered image. (a) Original. (b) Gaussian filtered. (c) The original Fourier transform. (d) The gaussian filtered Fourier transform.

Table 2: Training part of the main parameter settings.

Parameter	Numerical value	Parameter	Numerical value	Parameter	Numerical value	Parameter	Numerical value
Batch	64	Angle	0	Burn_in	2000	Scales	0.1, 0.1
Subdivisions	32	Saturation	1.5	Max_batches	50000	Learning_rate	0.001
Momentum	0.9	Exposure	1.5	Policy	Steps	Random	1
Decay	0.0008	Hue	0.1	Steps	4000, 4500	Jitter	3

Table 3: The effect of different k values on the clustering effect of datasets.

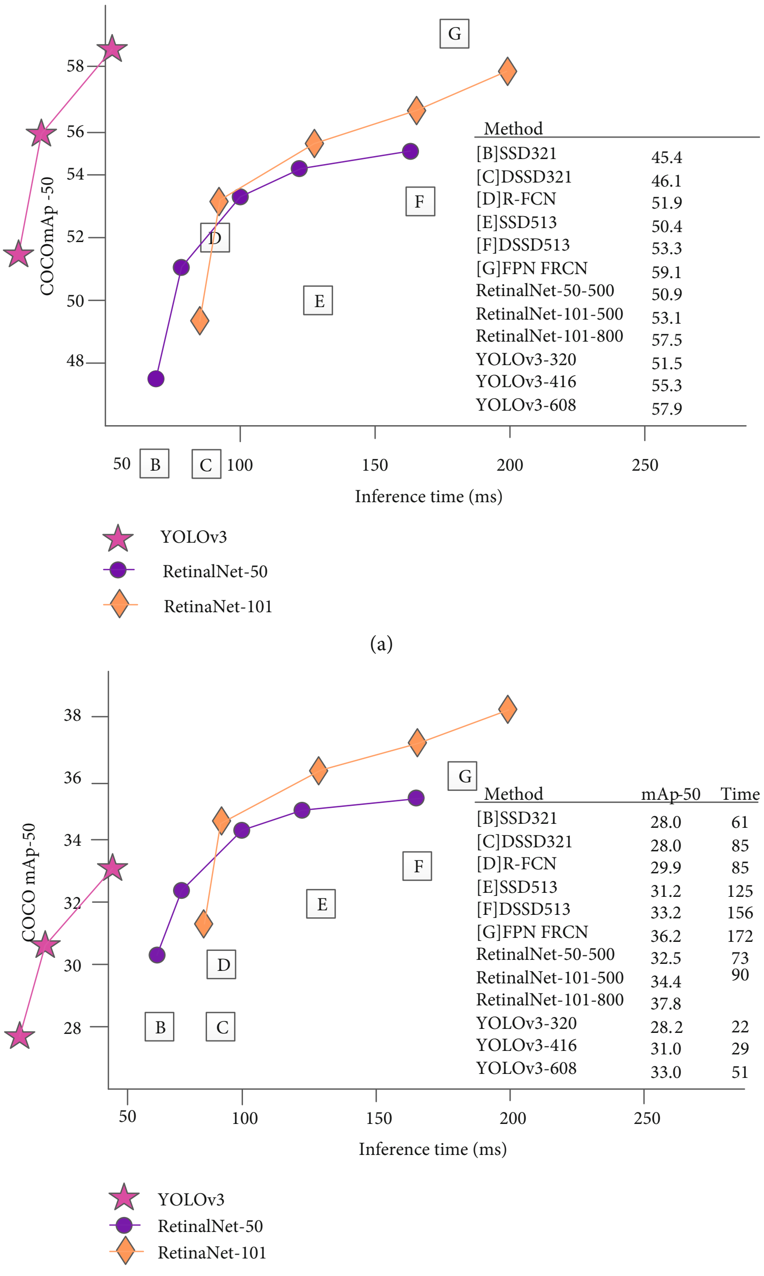
k value	5
Accuracy	82.32%
Boxes	(47, 42), (68, 45), (40, 60), (37, 25), (31, 43)
Ratios	[0.67, 0.72, 1.12, 1.48, 1.51]
k value	7
Accuracy	82.32%
Boxes	(72, 37), (37, 25), (46, 36), (39, 60), (66, 52), (42, 44), (50, 43)
Ratios	[0.65, 0.95, 1.16, 1.27, 1.28, 1.48, 1.95]
k value	9
Accuracy	86.06%
Boxes	(39, 61), (58, 53), (54, 38), (33, 25), (43, 45), (41, 26), (31, 41), (46, 40.5), (73, 38)
Ratios	[0.64, 0.76, 0.96, 1.09, 1.14, 1.32, 1.42, 1.58, 1.92]

6.3. Experimental Results under Different k-Means. To constantly adjust the size of the bounding box, the value of the IOU must be increased. Under the parameter settings in Table 2, the clustering effect of different k values on training data in different k-means algorithms is tested. The experimental results are listed in Table 3.

The clustering effect is conducive to the gear defect situation. S-YOLO allocates three different sizes of a priori boxes

Table 4: A priori boxes for different receptive field assignments.

Feature map	13 × 13	26 × 26	52 × 52
Receptive field	Big	Medium	Small
Prior box	(58, 53) (54, 38) (73, 38)	(39, 61) (43, 45) (46, 40.5)	(33, 25) (31, 41) (41, 26)



Method	
[B]SSD321	45.4
[C]DSSD321	46.1
[D]R-FCN	51.9
[E]SSD513	50.4
[F]DSSD513	53.3
[G]FPN FRCN	59.1
RetinalNet-50-500	50.9
RetinalNet-101-500	53.1
RetinalNet-101-800	57.5
YOLOv3-320	51.5
YOLOv3-416	55.3
YOLOv3-608	57.9

(a)

Method	mAp-50	Time
[B]SSD321	28.0	61
[C]DSSD321	28.0	85
[D]R-FCN	29.9	85
[E]SSD513	31.2	125
[F]DSSD513	33.2	156
[G]FPN FRCN	36.2	172
RetinalNet-50-500	32.5	73
RetinalNet-101-500	34.4	90
RetinalNet-101-800	37.8	
YOLOv3-320	28.2	22
YOLOv3-416	31.0	29
YOLOv3-608	33.0	51

(b)

Figure 9: Comparison between the detection speed and accuracy of YOLOv3 and other algorithms [11].

for each scale when performing three-scale feature detection. When the k value is equal to 9, nine kinds of a priori boxes are available for allocation. Hence, when assigning, three a priori boxes may be assigned for each scale feature. Details are shown in Table 4.

At the smallest feature map 13 × 13 (larger receptive field), the larger priority box (58, 53) (54, 38) (73, 38) is applied to the feature map, which is suitable for detecting surface scratches with large defect sizes. Medium feature map 26 × 26 (medium receptive field) applies a medium

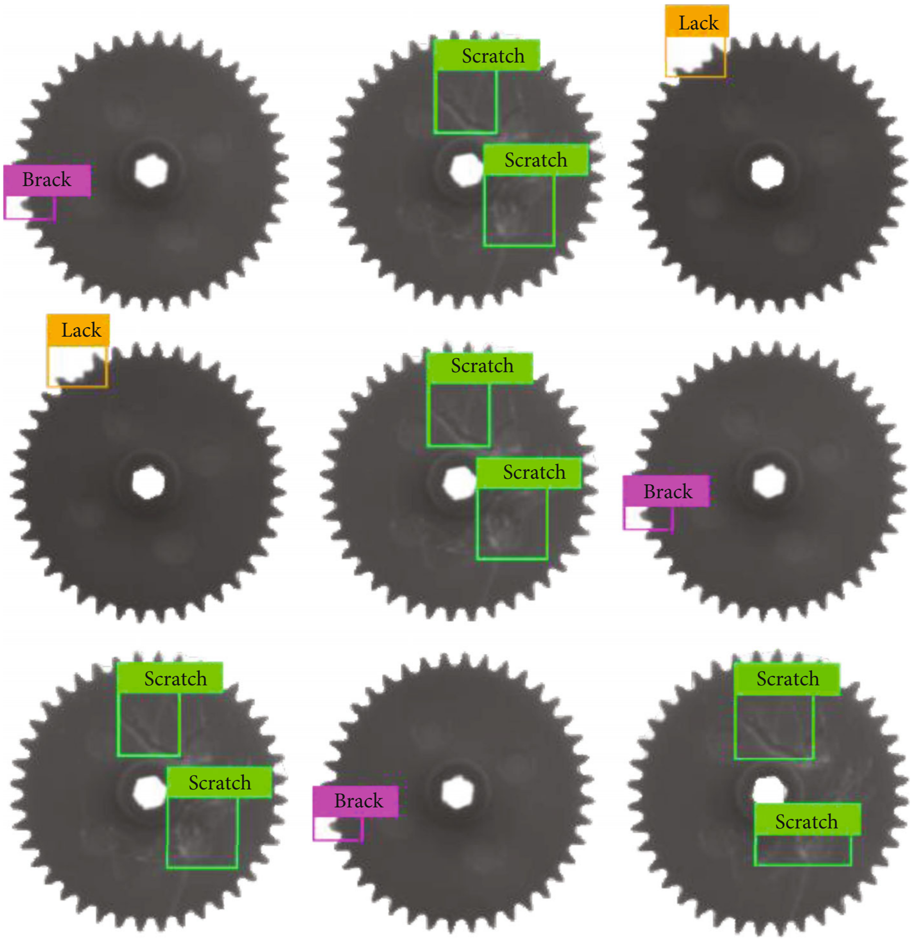


Figure 10: S-YOLO network test results without background interference.

priority box (39, 61) (43, 45) (46, 40.5), which is suitable for detecting objects of medium-size defects. A smaller priority box (33, 25) (31, 41) (41, 26) is applied on the larger feature map 52×52 (small receptive field), which is suitable for detecting objects with small defect sizes, such as broken and missing teeth. When training, the model training using the cluster generated by $k = 9$ can significantly shorten the model training time and improve the model IOU value.

6.4. Analysis of Gear Defect Detection Results. Figure 9 shows the combined performance of the YOLOv3 object detection network and other mainstream networks on the COCO datasets. After modifying the YOLOv3 model, the S-YOLO target detection model is trained. Through model training, the gear defect detection verification is finally performed on the detection platform. Figure 10 shows the detection of the S-YOLO model in the absence of complex background conditions, such as oil stains and dust particles. Figure 11 depicts the testing situation of the S-YOLO model when oil and dust particles are filled in the background in the simulation of the actual factory production on the platform for high-speed gear manufacturing defect testing. The experimental test results are provided in Table 5.

A comparison between Table 5 and Figure 11 shows that proposed network S-YOLO increases the complex background of gear manufacturing while retaining the advantages of traditional YOLOv3, which are detection speed and multiscale prediction. The image-smoothing layer and k -means clustering method are used to assign the most priority box to multiscale detection, which greatly inhibits the influence of the complex background on the detection effect of the model. It also makes the model lose stability and improves the average IOU value during training. S-YOLO is applied to the high-speed gear manufacturing defect detection experimental platform. Its classification effect reaches 100% accuracy, and the average confidence reaches 93.96%. The algorithm has good robustness.

7. Summary

The manufacturing defects in the gear manufacturing process were analyzed and studied. A dual-filtering background-weakening algorithm was proposed to address oil pollution, dust, and other complex backgrounds during production. Combined with the deep learning algorithm and target detection network model of YOLOv3, the network model of S-YOLO for gear manufacturing defect detection was proposed. Nine optimal anchor values were obtained via k-means clustering, which reduced the declining fluctuation of loss during model training and improved the

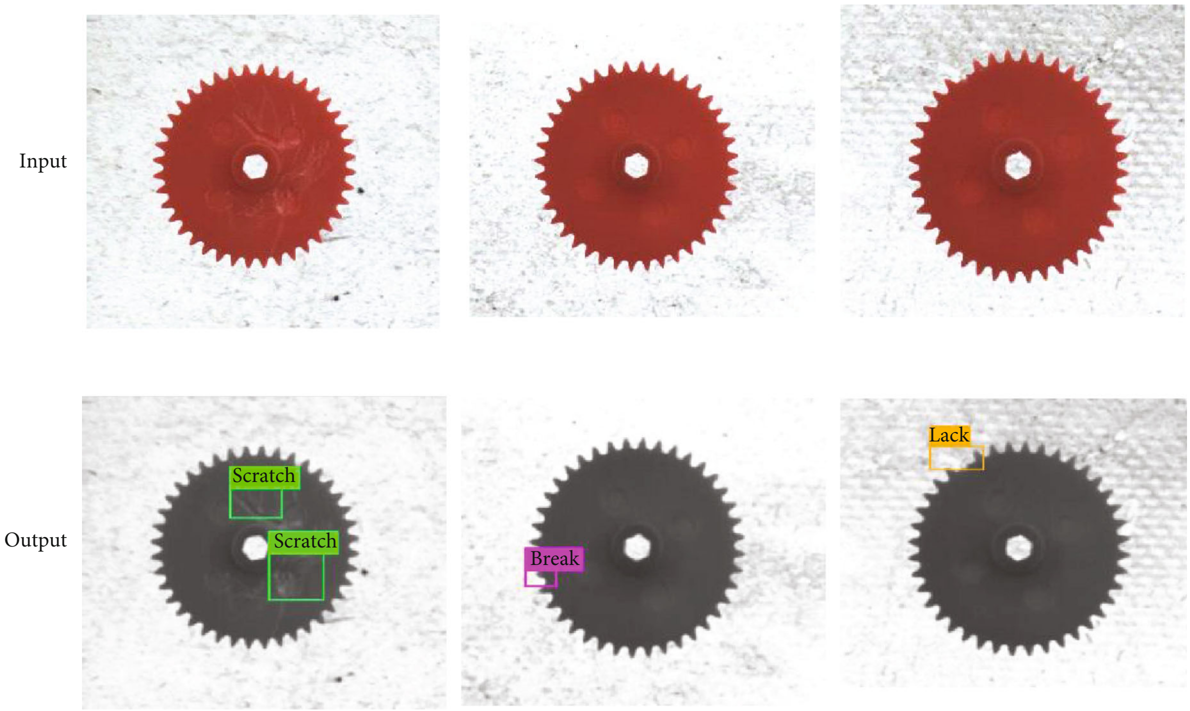


Figure 11: S-YOLO network defect detection when background interference exists on the detection platform.

Table 5: S-YOLO network detection gear defect situation.

Defect type	Number of defects in the test set	S-YOLO classification accuracy	YOLOv3 classification accuracy	Average confidence of each type of defect in S-YOLO	YOLOv3 average confidence of each type of defect
Broken tooth	300	100%	98.2%	98.30%	88.4%
Missing tooth	300	100%	98.7%	97.05%	84.9%
Surface scratch	300	100%	98.6%	86.54%	80.2%

average IOU value of the model. The gear manufacturing defect dataset was established using the data enhancement method. The application of the proposed algorithm and model was verified by building an online platform for industrial defect detection. The results showed that the proposed algorithm can meet actual production requirements.

Data Availability

Code and data have been made available at https://github.com/Yuli-Ya/Detecting-Gear-Surface-Defects.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

L.Y. and Z.W. worked on conceptualization and data curation. Z.W. performed the methodology. L.Y. worked with software and resources and did writing (original draft preparation and review and editing) and funding acquisition. Z.W, L.Y., and Z.D. carried out the validation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 91746116 and 51741101, the Science and Technology Project of Guizhou Province under Grant Nos. [2017]2308, [2015]4011, and [2016]5013, and the Collaborative Innovation of Guizhou Province (YJSCXJH[2018]052).

References

[1] M. J. Robinson, J. P. Oakley, and M. J. Cunningham, "The accuracy of image analysis methods in spur gear metrology," *Measurement Science and Technology*, vol. 6, no. 7, pp. 860–871, 1995.

[2] M. A. Ayub, A. B. Mohamed, and A. H. Esa, "In-line inspection of roundness using machine vision," *Procedia Technology*, vol. 15, pp. 807–816, 2014.

[3] Q. Guo, C. Zhang, H. Liu, and X. Zhang, "Defect detection in tire X-ray images using weighted texture dissimilarity," *Journal of Sensors*, vol. 2016, Article ID 4140175, 12 pages, 2016.

[4] J. Yuan, Q. Wang, and B. Li, "A flexile and high precision calibration method for binocular structured light scanning system," *The Scientific World Journal*, vol. 2014, no. 8, Article ID 753932, 8 pages, 2014.

[5] J. Yang and G. Yang, "Modified convolutional neural network based on dropout and the stochastic gradient descent optimizer," *Algorithms*, vol. 11, no. 3, pp. 28–43, 2018.

[6] G. Yang, J. Yang, W. Sheng, F. Junior, and S. Li, "Convolutional neural network-based embarrassing situation detection under camera for social robot in smart homes," *Sensors*, vol. 18, no. 5, pp. 1530–1553, 2018.

[7] J. Yang, W. Sheng, and G. Yang, "Dynamic gesture recognition algorithm based on ROI and CNN for social robots," in *2018 13th World Congress on Intelligent Control and Automation (WCICA)*, pp. 389–394, Changsha, China, China, July 2018.

[8] Y. Xu, C. Yang, J. Zhong, N. Wang, and L. Zhao, "Robot teaching by teleoperation based on visual interaction and extreme learning machine," *Neurocomputing*, vol. 275, pp. 2093–2103, 2018.

[9] M. Trobe and M. D. Burke, "The molecular industrial revolution: automated synthesis of small molecules," *Angewandte Chemie International Edition*, vol. 57, no. 16, pp. 4192–4214, 2018.

[10] J. Yang, S. Li, Z. Gao, Z. Wang, and W. Liu, "Real-time recognition method for 0.8 cm darning needles and KR22 bearings based on convolution neural networks and data increase," *Applied Sciences*, vol. 8, no. 10, p. 1857, 2018.

[11] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, http://arxiv.org/abs/1804.02767.

[12] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, Santiago, Chile, December 2015.

[13] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," in *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, volume 9905*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., pp. 21–37, Springer, Cham, 2016.

[14] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, Honolulu, HI, USA, July 2017.

[15] J. Mączak and M. Jasiński, "Model-based detection of local defects in gears," *Archive of Applied Mechanics*, vol. 88, no. 1-2, pp. 215–231, 2018.

[16] J. M. Gandarias, A. J. García-Cerezo, and J. M. Gómez-De-Gabriel, "CNN-based methods for object recognition with high-resolution tactile sensors," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6872–6882, 2019.

[17] S. Lu, G. Qian, Q. He, F. Liu, Y. Liu, and Q. Wang, "Insitu motor fault diagnosis using enhanced convolutional neural network in an embedded system," *IEEE Sensors Journal*, p. 1, 2019.

[18] Z. Cheng and N. Hu, "Quantitative damage detection for planetary gear sets based on physical models," *Chinese Journal of Mechanical Engineering*, vol. 25, no. 1, pp. 190–196, 2012.

[19] F. Nabih, C. Jérôme, V. Fabrice, and V. Philippe, "Detection of gear tooth pitting based on transmission error measurements," in *Design and Modeling of Mechanical Systems. Lecture Notes in Mechanical Engineering*Springer, Berlin, Heidelberg.

[20] R. Younes, N. Ouelaa, N. Hamzaoui, and A. Djebala, "Experimental study of combined gear and bearing faults by sound perception," in *Advances in Acoustics and Vibration. Applied Condition Monitoring, vol 5*, T. Fakhfakh, F. Chaari, L. Walha, M. Abdennadher, M. Abbes, and M. Haddar, Eds., vol. 76, no. 5-8pp. 927–940, Springer, Cham, 2015.

[21] M. Zhao, J. Lin, Y. Miao, and X. Xu, "Feature mining and health assessment for gearboxes using run-up/coast-down signals," *Sensors*, vol. 16, no. 11, p. 1837, 2016.

[22] T. Kidar, M. Thomas, M. Elbadaoui, and R. Guilbault, "Phase monitoring by ESPRIT with sliding window and Hilbert transform for early detection of gear cracks," in *Advances in Condition Monitoring of Machinery in Non-Stationary Operations. Lecture Notes in Mechanical Engineering*, G. Dalpiaz, Ed., pp. 287–299, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[23] K. Debray, F. Bogard, and Y. Q. Guo, "Numerical vibration analysis on defect detection in revolving machines using two bearing models," *Archive of Applied Mechanics*, vol. 74, no. 1-2, pp. 45–58, 2004.

[24] R. Moreno, J. M. Chicharro, and P. Pintado, "Comparison of minimum detectable crack size in a geared system from three different vibration transducer types," *Experimental Techniques*, vol. 38, no. 1, pp. 76–87, 2014.

[25] H. Ozturk, I. Yesilyurt, and M. Sabuncu, "Detection and advancement monitoring of distributed pitting failure in gears," *Journal of Nondestructive Evaluation*, vol. 29, no. 2, pp. 63–73, 2010.

[26] R. Liu, D. Zhong, H. Lyu, and J. Han, "A bevel gear quality inspection system based on multi-camera vision technology," *Sensors*, vol. 16, no. 9, p. 1364, 2016.

[27] S. Fedala, D. Rémond, R. Zegadi, and A. Felkaoui, "Gear fault diagnosis based on angular measurements and support vector machines in normal and nonstationary conditions," in *Advances in Condition Monitoring of Machinery in Non-Stationary Operations. CMMNO 2014. Applied Condition Monitoring, vol 4*, F. Chaari, R. Zimroz, W. Bartelmus, and M. Haddar, Eds., pp. 291–308, Cham: Springer International Publishing, 2014.

[28] A. Djebala, N. Ouelaa, C. Benchaabane, and D. F. Laefer, "Application of the wavelet multi-resolution analysis and Hilbert transform for the prediction of gear tooth defects," *Meccanica*, vol. 47, no. 7, pp. 1601–1612, 2012.

[29] X. Zhang and S. Fan, *Identification and modification of closed defect of the location function for N-lobed noncircular gears*, Springer Singapore, Singapore, 2016.

[30] X. Tian, L. Jiao, Y. Duan, and X. Zhang, "Video denoising via spatially adaptive coefficient shrinkage and threshold adjustment in Surfacelet transform domain," *Signal, Image and Video Processing*, vol. 8, no. 5, pp. 901–912, 2014.

[31] H. D. J. O. Domínguez and V. G. Jiménez, "Evaluation of denoising methods in the spatial domain for medical ultrasound imaging applications," in *Current Trends on Knowledge-Based Systems. Intelligent Systems Reference Library, vol 120*, G. Alor-Hernández and R. Valencia-García, Eds., pp. 263–281, Springer, Cham, 2017.

[32] J. Yang, S. Li, Z. Wang, and G. Yang, "Real-time tiny part defect detection system in manufacturing using deep learning," *IEEE Access*, vol. 7, pp. 89278–89291, 2019.

Hindawi
Journal of Sensors
Volume 2019, Article ID 5370838, 12 pages
https://doi.org/10.1155/2019/5370838

Research Article

Presenting a New Wireless Strain Method for Structural Monitoring: Experimental Validation

Amedeo Gregori,[1] Emidio Di Giampaolo,[2] Alessandro Di Carlofelice,[2] and Chiara Castoro[1]

[1] *Department of Civil, Building and Environmental Engineering, University of L'Aquila, Via Giovanni Gronchi 18, 67100 L'Aquila, Italy*
[2] *Department of Industrial and Information Engineering and Economics, University of L'Aquila, Via Giovanni Gronchi 18, 67100 L'Aquila, Italy*

Correspondence should be addressed to Chiara Castoro; chiara.castoro@libero.it

Received 14 March 2019; Revised 7 July 2019; Accepted 23 August 2019; Published 3 November 2019

Guest Editor: Ye Lu

The structural health monitoring (SHM) of large and complex infrastructures as well as laboratory tests of new structures and materials resorts to strain gauge measurements to check mechanical stress. A wireless measurement of the strain gauge response is desirable in many practical applications to avoid the cost and the difficulty of wiring, particularly in large structures requiring several sensors and in complex objects where the measurement points are difficult to access. In this paper, a wireless strain gauge which is a hybrid between an RFID tag and a usual thin-film resistive strain gauge is experimented. Installation and maintenance problems of the wireless sensor networks are overcome allowing a high level of measurement accuracy, comparable to that of wired strain sensors, together with a long measurement distance. A large set of measurements has been performed using reference specimens and readings in order to validate the sensor and to develop a calibration procedure that makes the sensor suitable for a large number of different applications in civil engineering.

1. Introduction

The structural monitoring of large and complex infrastructures requires the use of a number of specific sensors distributed on a large area or volume to form a monitoring system composed of independent or interconnected sensor nodes like a smart skin. Because of the complexity of that monitoring system, some kind of intelligence is required to make easy the managing of the system and the handling of the large amount of measured data. In particular, the communication between the sensor nodes and a managing unit that collects and stores the information gathered from the sensors spread over the structure is a critical issue that preferably has a solution with wireless systems. In fact, wireless sensors permit a fast and easy installation even in points difficult to access while their cost is lower than that of a wired system.

Wireless sensor networks (WSNs) are a favorite candidate for structural health monitoring (SHM) of large structures. As an interdiscipline consisting of sensor, communication, and wireless technology, WSNs were initially applied in the military and then extended to environmental monitoring, agriculture, medical treatment, and civil engineering [1–13]. Various types of sensors have been developed during the past decades (such as strain gauge and optical fiber sensors) while SHM technology is becoming important in a wide range of technical fields, most of the time in combination with structural use of special composites and high-performance materials [14]. Actually, it is more and more understood as the SHM system can improve safety and reliability of structures by autonomously monitoring the conditions or detecting critical damage. In [9], authors state that in comparison with traditional wired sensor networks, wireless systems for SHM have numerous advantages in terms of better flexibility, software or hardware expandability, cost effectiveness, and fault tolerance. Interesting applications of wireless sensors made in aerospace, civil, and environmental engineering are discussed in [15–18] comparing the practical needs concerning space requirements and cost increments with those of traditional wired techniques. In [19], a dipole antenna-based wireless sensor for the damage detection of a composite rotor blade is investigated. In [20], the use of WSNs has been presented as a useful tool even in forest fire surveillance, allowing real-time acquisition, evaluation, and analysis of environmental information including temperature, humidity, sound, vibrations, and smoke as well as pictures of buildings and forest. Some advances in research, development, and implementation of smart sensor networks and health monitoring systems for civil infrastructures are presented in [21, 22]. In particular, cases of study of WSNs and their integrated systems and implementations in offshore platform structures, hydraulic engineering structures, large span bridges, and large space structures have been reported. Other researches on wireless sensing technology applied to SHM for buildings and civil engineering structures are discussed in [23].

During the entire last decade, implementation of wireless transmitters for continued structural damage monitoring became a promising research field to be often related to new patented inventions and devices. In [24], the authors present a new methodology for operating a monitoring system that provides near-real-time structural condition assessment for extreme events and long-term deterioration information, using MEMS-type accelerometers. This proposed structural monitoring system comprises modular, battery-powered data acquisition devices which transmit structural information to a central data collection and analysis device over a wireless data link. Data acquisition devices comprise mechanical vibration sensors, data acquisition circuitry, wireless transmitter, and battery. For sophisticated analysis after a natural hazard or extreme event, the authors suggested that powerful computers may be interfaced with the central device.

While WSNs have been extensively investigated in recent years, many practical challenges are still to be faced when employing such a technology for many SHM applications, including civil and mechanical infrastructures [25]. In fact, a WSN must remain in operation over multiple decades with maintenance costs low enough to justify its integration into a given structural maintenance strategy. These technical barriers include ensuring reliable power sources for sensor nodes, reducing installation and maintenance costs, and automating the collection and analysis of data acquired by a WSN.

As a possible solution to overcome the mentioned challenges, in [26] is discussed the use of sensor nodes that collect measurements from a structure in a completely passive manner without any electrical power. In [27], a wireless sensor network with temperature-compensated measuring technology for long-term structural health monitoring of buildings and infrastructures is presented. A brief summary and comparison among benefits and disadvantages related to active and passive wireless sensors are given in [28] together with the presentation of a passive wireless structural health monitoring sensor made with a flexible planar dipole antenna. In fact, it is understood that chipless passive wireless sensors can give real-time structural information for SHM without space and battery constraints in harsh environmental conditions [10–13]. Chipless passive wireless strain and damage detection sensors based on a frequency selective surface are presented in [29].

In this paper, a new kind of sensor node for mechanical stress detection which is obtained as a hybridization of an RFID (Radio Frequency Identification) tag and a resistive strain gauge is exploited.

It is a semipassive wireless strain sensor tag, which uses a piezoresistive thin-film strain gauge (like a wired sensor), but it can be passively interrogated as an RFID tag. Like an RFID system, there is an interrogation unit (i.e., a commercial RFID reader) that radiates an electromagnetic wave that impinges on the antenna of the sensor tag waking it up. The strain gauge varies its resistance in accordance with the applied strain and drives an oscillating circuit that modulates the electromagnetic wave backscattered by the sensor tag antenna. The modulating frequency is dependent on the applied strain and can be easily measured by means of a spectrum analyzer or a frequency meter once the modulated backscattered wave is received back by the interrogation unit.

The prototype of the proposed sensor tag makes use of a battery to power the oscillating circuit that is maintained turned off for all the time except for a short time interval during the measurement interrogation. For this reason, the life span of the battery can be very long. A detailed description of the sensor tag is reported in [30] and in Materials and Methods.

Since each sensor tag is autonomous and independent from other nearby deployed sensors, the resulting network has the simple star topology where the central unit is an RFID reader that interrogates one-by-one all the deployed sensor tags. In this paper, the application of this sensor tag to realistic cases concerning the field of civil and mechanical engineering is shown. The assessment of the sensor tag is demonstrated by means of measurements of the Young's modulus of different materials and by a comparison with the results obtained using a calibrated wired system. A calibration procedure of the new sensor tag has been developed, and a detailed measurement campaign using many instrumented specimens is also reported.

With this study, we have assessed the possibility to transfer the measured information from the sensor tag to a central unit analogically avoiding the analog to digital conversion at the sensor tag level but performing it at the level of the central unit. We transmit the measured data analogically as a backscattered frequency modulated continuous signal that can be sampled at the receiving unit using an appropriate device having a high data rate so that very fast phenomena can be detected easily, with high accuracy, using cheap sensor tags. The feasibility of this new sensor tag has been investigated; a measurement campaign shows its effectiveness and proves the advantages of this new sensor tag in particular in measuring vibrations and dynamic phenomena. The managing of the system results is simple because it does not suffer of restriction on the data rate, since the strain information travels over analogical signals.

2. Materials and Methods

The overall SHM system consists of an interrogating unit (i.e., a reader) placed in a convenient position (e.g., near the ground in Figure 1) and several sensor tags deployed on the structure to be monitored (e.g., in positions difficult to access). Tags are fixed to the structure; they are designed to remain operative for several years, while the interrogation unit is intended to be portable and placed at the measurement location only when needed (obviously, it can be also used for a permanent monitoring of the structure). All sensor tags are quiescent (i.e., inactive) for all the time except for the short time interval when interrogated. During interrogation, they measure the strain affecting the portion of the structure where each of them is stuck on and send back to the interrogation unit a signal encoding the strain measured value. The optimal operation of the system is achieved when the interrogation unit has a line of sight with each sensor tag, but the system is able to work even in nonline of sight conditions. As sketched in Figure 1, the distance between the interrogation unit and the sensor tags can be of several meters. Like in logistic applications, the number of tags that can be handled by a single interrogation unit is large (e.g., a commercial reader can interrogate several tags per second), but that number can be larger or smaller in agreement with the repetition time of interrogations. In case of static objects (i.e., the change of the status of the monitored structure is slow compared to the measurement time), the interrogation rate can be low (the number of tags interrogated in the time unit is not an issue), and as a consequence, the number of tags that can be handled by an interrogation unit is limited only by the maximum interrogation distance, i.e., the distance over which the tag is not getting enough power to be woken up.

The developed sensor consists of three main circuital blocks: an RFID block, a supply block, and a sensing block as shown in Figure 2(a).

The RFID block consists of an antenna, a commercial RFID microchip (i.e., an NXP GMiL+), and a pin diode with its feeding network. The antenna is a dipole-like antenna operating at 868 MHz; the pin diode is connected to the antenna terminals by means of a feeding network, and it is used to modulate (on-off) the backscattered signal. The RFID microchip, powered by an external battery, allows the reading/writing distance of the sensor tag up to 30 m (nominal) and allows the remote control of the voltage level of a logic pin by means of an appropriate writing of the configuration word in its memory. This voltage level is used to switch the supply and sensing blocks on and off.

The sensing block is essentially a resistance-to-frequency converter circuit whose output is a squared wave signal that is used to drive the pin diode connected to the antenna terminals. Under the squared wave signal, the input impedance of the pin diode switches between two values (low and high impedances) performing an amplitude modulation of the electromagnetic wave that is backscattered by the antenna (Figure 2(a)). Since the frequency of the squared wave signal is proportional to the strain gauge stretch, the backscattered signal, amplitude modulated by the squared wave signal, carries the information concerning the strain measured by the strain gauges.

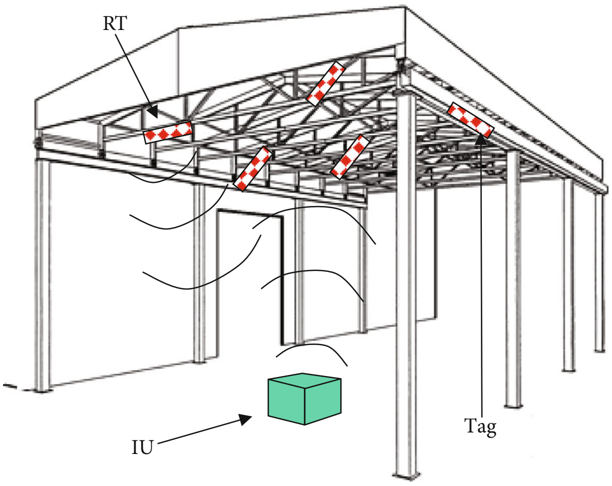


Figure 1: Scheme of the system measurement set-up. Tags are represented as dotted rectangles; the interrogation unit (IU) is shown as a box on the ground while RT means responding tag.

Details of the resistance-to-frequency converter are shown in Figure 2(b). It is composed of a full Wheatstone bridge strain gauge circuit (i.e., four piezoresistive thin-film strain gauges like in a wired sensor) and an operational amplifier that amplifies the small voltage changes across the bridge.

Actually, the Wheatstone bridge is a well-known circuit consisting of four resistive arms with resistances R_1, R_2, R_3, and R_4 and an excitation voltage, V_e, applied across the bridge. The output voltage V_o is zero when $R_1/R_2 = R_4/R_3$ and the bridge is said to be balanced. Any change in resistance in any arm of the bridge will result in a nonzero output voltage.

A possible configuration for strain measurements is the so-called quarter-bridge. It has a strain gauge at one arm of the bridge (active arm), e.g., R_4, which is the only changing resistance of the bridge. $R_4 = R_0 + \Delta R$, where R_0 is the unstressed resistance (or nominal resistance) of the strain gauge and ΔR models the strain-induced change in resistance due to an applied strain ε.

It is $\Delta R = R_0 \cdot \mathrm{GF} \cdot \varepsilon$, where GF is the Gauge Factor, i.e., the sensitivity of the strain gauge to strain which for metallic strain gauges is typically around 2.

If we assume that $R_1 = R_2$ and $R_3 = R_0$, the output voltage is

$$V_o = -\frac{\mathrm{GF} \cdot \varepsilon}{4}\left[\frac{1}{1 + \mathrm{GF} \cdot \varepsilon/2}\right] V_e. \quad (1)$$

In practice, considering that the strain measurements rarely involve quantities larger than a few millistrain, the output voltage (Equation (1)) cannot be larger than a few thousandths of the excitation voltage. For these reasons, wired

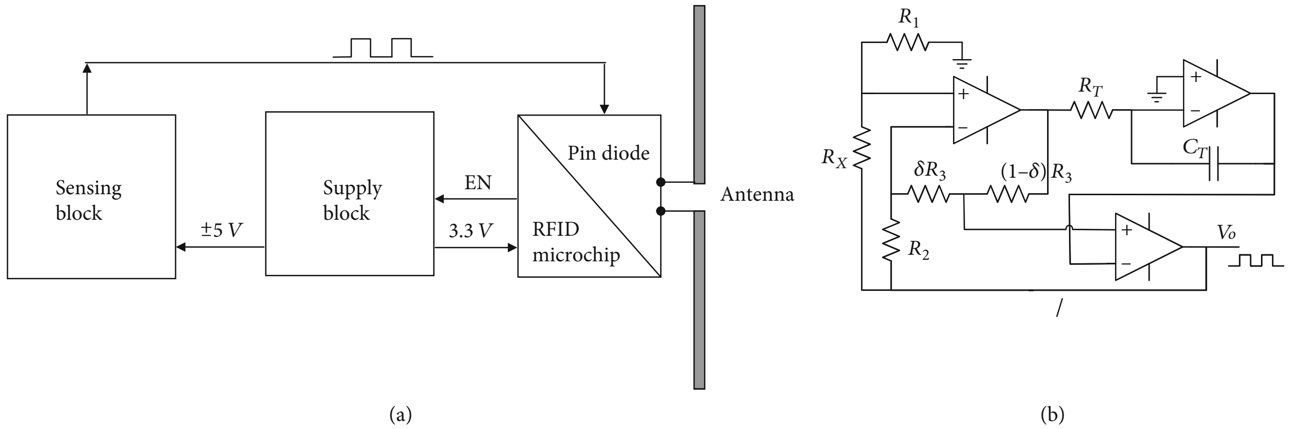


Figure 2: (a) Scheme of the sensor tag; three main circuital blocks: an RFID block, a supply block, and a sensing block. The sensor tag is quiescent and can be activated on demand. (b) Schematic of the resistance-to-frequency converter circuit.

systems resort to an amplification of the output voltage to bring the signal to a level where it can be conveniently handled for indication or recording. In the proposed wireless system, instead, the resistance-to-frequency technique is used in order to codify the strain measurement into a modulation frequency while preserving the high sensitivity of the Wheatstone bridge. A deeper description of the resistance-to-frequency converter is provided in [30].

The unbalance voltage due to the resistance change is then integrated, and its polarity is fed back to the bridge as the bias voltage to sustain the oscillation. The oscillation frequency changes quite linearly with the stretching of the strain gauges. The output of this circuit drives a pin diode circuit for modulating the backscattered signal.

The interrogating unit consists of a commercial RFID reader (the same kind used in logistics) and a spectrum analyzer connected to an antenna and to a personal computer (Figure 3); alternatively, we have used a Universal Software Radio Peripheral (National Instruments NI USRP 2920) particularly suited for dynamic measurements. At the start of measurements, the interrogation unit sends, by means of the RFID reader, an electromagnetic wave which delivers both the energy to wake up the tags and a query command to boost the tags to reveal themselves. Awaken tags modulate the electromagnetic wave that scatters back from their antennas with a random numeric code revealing their readiness for communication. Then, the RFID reader performs the inventory of all the responding tags which identify themselves sending back their ID. Once the inventory is completed, the interrogation unit starts with the measurements of the strain gauge status of the inventoried tags.

The measurement procedure is repeated identically for each one of the responding tags and consists of three steps. In the first step, the interrogation unit opens a specific reading/writing session with the ith responding tag identified by means of its ID and enables that tag (with appropriate writing of the configuration word located in the memory of the tag) to switch on the supply circuit block which supplies energy to the sensing block. After that, during the second step, the interrogation unit sends a continuous wave (CW) signal and waits for the signal backscattered from the tag which delivers the strain gauge status information. In fact, the strain gauge varies its resistance in accordance with the applied strain and drives an oscillating circuit (inside the sensing block) which modulates the electromagnetic wave backscattered by the tag antenna. The modulating frequency is linearly dependent on the applied strain, so the backscattered signal has a different frequency modulation in accordance with the strain gauge status. Once the modulated backscattered signal is received by the interrogation unit, the modulating frequency is easily measured by means of a spectrum analyzer and the strain status is retrieved. Once the strain measurement is completed, the interrogation unit becomes the step three. A new reading/writing session is open to disable (by writing the appropriate memory location of the tag) the supply circuit block, and consequently, the sensing block is switched off. The modulation of the backscattered signal finishes, and a CW signal arrives to the interrogating unit which closes the communications with that tag and starts to interrogate another inventoried tag repeating the above-described three steps. Therefore, the sensing circuit is kept turned off all the time except for the short time interval during the measurement interrogation (step two).

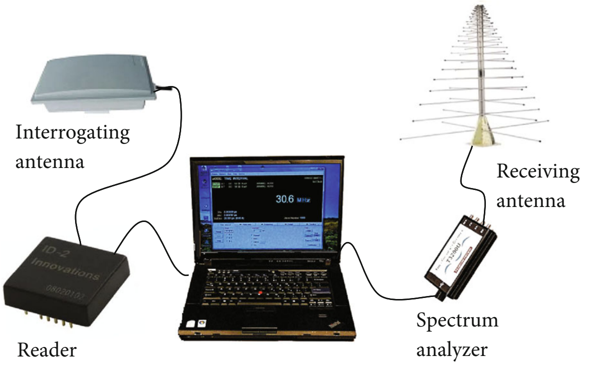


Figure 3: The interrogation unit consists of a commercial RFID reader and a spectrum analyzer connected to an antenna and to a personal computer.

3. Set-Up of the Experimental Tests

To evaluate the effectiveness of the proposed sensor tag, a campaign of measurements has been performed. Tasks of this experimental activity were the calibration of the sensor and the comparison of the wireless measured data with those achieved with a consolidate method based on wired sensors. Specific tests have been also performed to estimate the maximum interrogation distance for the wireless sensor. Inside a laboratory room, with several obstacles making this environment noisy, the sensor tag was proved to allow for an interrogation distance even larger than 20 meters.

To prove the effectiveness of the proposed new wireless strain measurement technique, the elastic properties of three materials have been investigated using wireless sensor tags in place of the wired strain gauges. In particular, the experimental procedure described in this section refers to the elastic modulus assessment of brass, aluminium, and steel samples, respectively. These estimations have been first performed in accordance with a common wired strain gauge technology. Subsequently, the estimation of the elastic modulus of the three materials has been repeated on the base of the wireless strain measurements. The two procedures are discussed in Results and Discussion, together with a comparison of the experimental results.

The experimental tests were performed in accordance with the wired strain measurement technique, and those concerning the new proposed method have been all carried out operating a specific apparatus. In particular, the experimental apparatus included a manual compensator (working with the principle of the Wheatstone bridge) and a vertical metal frame (Figure 4) that mainly consists of a rigid steel-made structure built for a laboratory test and standing in the vertical plane and designed to apply controlled forces up to 10 kN in tension. In particular, a manually driven wheel is mounted on the top of the testing frame. The wheel controls a lever mechanism by which a variable tensile force is introduced in the specimen causing it to extend. A load cell mounted in series with the specimens (and electrically connected to the compensator) allows the measurement of the tensile force applied during the tensile test.

Operating the just mentioned apparatus in a proper way, the elastic modulus E (Young's modulus), the shear modulus G, and/or the Poisson's ratio ν of any elastic material can be assessed. So far, only specimens of material previously instrumented with wired strain gauges have been considered for tests.

Three different solid metal pieces made of brass, aluminium, and steel, respectively, have been selected and machined to obtain a couple of identical specimens from each of them. The two specimens of each couple have been assumed to be homogeneous and to univocally represent the same material. To perform tensile tests, the specimens had to be previously worked and instrumented. Once machined, specimens have been characterized from a dog's bone shape, with the two hands kept bigger to secure the specimen to the testing apparatus and to favorite a regular strain diffusion along the central, cylindrical portion of the specimen (stalk) when applying tensile forces. The cylindrical portion of the specimens measures 8 mm in diameter and about 100 mm in length (Figure 5).

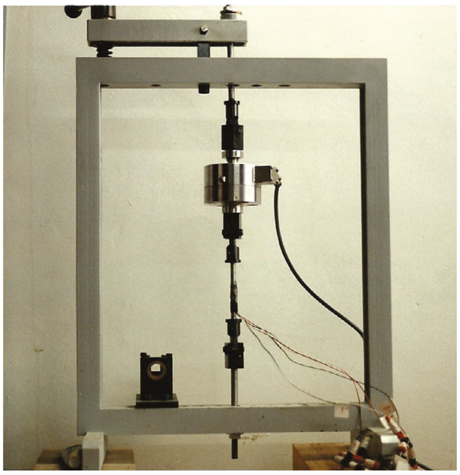


Figure 4: Steel frame for the laboratory tests.

During the test, the applied tensile forces have been measured and controlled by means of a load cell mounted in series with the specimen and electrically connected to the compensator. The load cell is characterized by a maximum load capacity of 20 kN, a maximum output signal of 4000 mV/V, and a scale factor $k = 2$.

The number and the magnitude of the load increments applied during the tensile test have been set in accordance with the different strengths and stiffness of the various metal specimens. In particular, load increments have been limited in the range of 0.75 ÷ 1 kN (equivalent signal variation: 150 ÷ 200 mV/V) and have been repeated several times upward and downward, respectively. A complete load sequence includes a number of steps upward to a maximum load value and an equal number of load decrements downward to a zero load value, so to obtain a closed loop. The maximum applied load has been fixed in advance and kept low enough to avoid the specimens to yield: about half of the yield threshold was reached as maximum stress.

According to the elastic theory, tensile strains resulted in the specimen while loads are being applied. During the experiments, tensile strains have been detected and measured by means of several strain gauges previously mounted on the specimen and electrically connected to the compensator. In particular, each specimen has been instrumented with two single grid strain gauges glued on the opposite side of the specimen, the grids being aligned with the longitudinal axis of the specimen. Common strain gauges 3/120LY4x type, characterized from $R = 120\,\Omega$ and a scale factor $k = 2$, were adopted.

The number of strain gauges used in each single test varies from two to four according to the half or to the whole bridge configuration, respectively. The reason for multiple gauges mounted on the same specimen first comes from the undesired possibility that applied tensile force could act eccentrically with respect to the ideal longitudinal axis of the specimen. In fact, applied loads acting with casual eccentricities introduce undesired bending actions so that a not

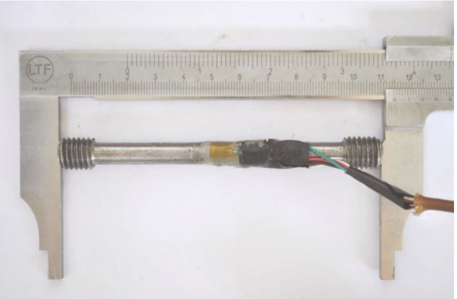


Figure 5: One of the instrumented steel specimens.

uniform strain distribution across the transversal section of the specimen may result. An average signal computed from strain gauges mounted on opposite sides of the same specimen allows to compensate for this undesired effect and accounts for the ideal, uniform tensile strain to be considered in the tensioned specimen. On the other hand, when more arms of the Wheatstone bridge have been made active, the sensitivity of the bridge increased. Moreover, due to possible temperature change during the test, the need for a second signal compensation rises as well. For this reason, signals from gauges mounted on the loaded specimen are combined with those recorded, at the same time, from gauges mounted on a second, identical specimen kept unloaded. In this research work, both the mechanical and thermal compensations of the strain measurements have been performed at once, recording signals from gauges in a whole bridge configuration.

4. Results and Discussion

The elastic modulus of the three materials has been determined by means of tensile tests performed in accordance with the wired method first, then using wireless strain gauges.

4.1. Experimental Data from the Wired Method. Experimental curves obtained from separate tensile tests carried out on the brass specimen, on the aluminium specimen, and on the steel specimen, respectively, have been reported in the graph of Figure 6. The values of the applied stress σ have been reported on the vertical axis and measured strains ε in the horizontal one.

As expectable, the experimental data clearly proved the elastic behavior of the specimens. According to Hooke's law ($\sigma = E \cdot \varepsilon$), the estimation of the elastic modulus E of each different material was then obtained as a slope (angular coefficient) of the linear regressions built on the experimental data. Both the ascending and descending branches of each experimental curve have been considered in these calculations. In fact, it is worth to note that a curve actually consists of a very narrow loop in which the ascending and the descending branches, respectively, do not precisely overlap each other. This fact (undesired) is usually assumed to depend from hysteretic phenomena (unavoidable) taking place during the loading/unloading cycles. In our research, the loop amplitude was made sure to remain always negligible and not to affect the test repeatability. This explains why loops are not clearly visible from curves plotted in Figure 6.

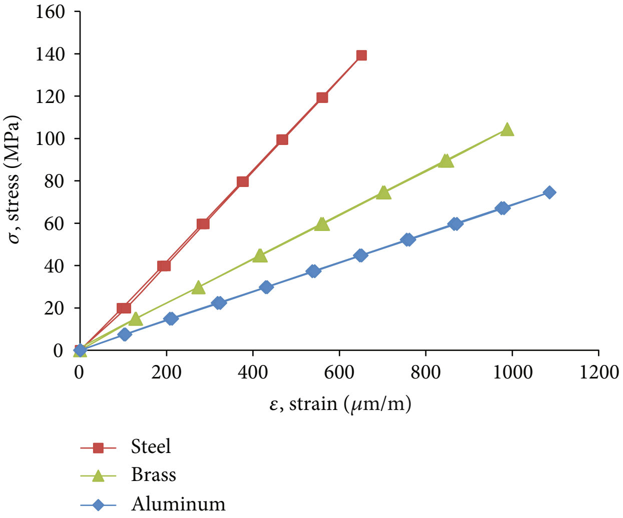


Figure 6: Experimental stress-strain curves obtained from tensile tests on steel, brass, and aluminium specimens using the wired method.

On the other hand, linear regressions built on experimental data are found to be very accurate and characterized by R^2 values almost equal to 1. In this experimental work, the mean value of the three measurements repeated on each metal specimen has been assumed as the reference value for the elastic modulus E_{ref} of the considered material. Single experimental results and E_{ref} values have been reported in Table 1.

4.2. Experimental Data from the Wireless Method: Static Measurements. To prove that the wireless sensor tag is a reliable measuring tool, the elastic modulus assessment previously carried out on the base of the wired method has been attempted in the wireless mode. As before, tensile loads (and stresses) continued to be measured by means of the same load cell and compensator.

Data plotted in Figure 7 show the results of a tensile test performed on the brass specimen and recorded in accordance with the new wireless method. Values of the applied stress σ have been reported on the vertical axis, and the recorded bridge frequency changes on the horizontal one. To be consistent in the experimental procedure, stress increments and load sequence have been repeated as in previous tests carried out to estimate reference elastic modulus E_{ref} of the brass specimen in accordance with the wired method.

Experimental data in Figure 7 highlight that a relationship exists among the applied load increments and the frequency changes (i.e., the modulation frequency of the backscattered signal) measured from the interrogation unit. Actually, a linear regression built on the experimental data was found to be very satisfactory, with the coefficient of determination R^2 approaching unit ($R^2 = 0.9931$). On the other hand, considering experimental data more into details, it is noted that the experimental curve yet consists of a very narrow loop (as for wired gauge strain measurements) in

Table 1: Estimation of the reference elastic modulus E_{ref} of the materials and R^2 of the single linear regressions. Wired method.

Material	Stress step (MPa)	Max. stress (MPa)	Cycle 1		Cycle 2		Cycle 3		E_{ref} (MPa)
			E_1 (MPa)	$R_1{}^2$	E_2 (MPa)	$R_2{}^2$	E_3 (MPa)	$R_3{}^2$	
Steel	20	140	219200	0.9999	223700	0.9999	223100	1.000	**222000**
Brass	15	105	106100	0.9995	106600	1.000	105400	0.9996	**106000**
Aluminium	7.5	75	68300	1.0000	69200	1.000	68900	1.000	**68800**

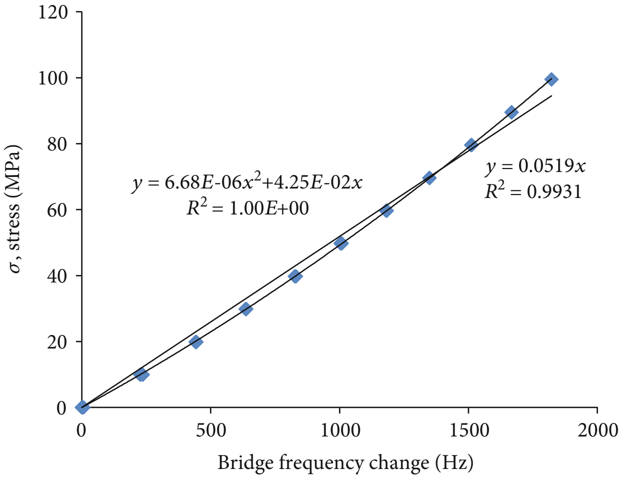


Figure 7: Wireless sensor tag frequency variation recorded during the tensile test on the brass specimen.

which the ascending and the descending branches tend to a unique monotonic curve perfectly fitted by a second order polynomial regression ($R^2 = 1.00$) (Figure 7). This was confirmed also for different materials.

About the weak nonlinearity of the experimental curve, it is worth to note that it mainly depends on the actual value of the hardware parameters, i.e., the tolerances of the electronic components (in particular, resistances and capacitances) that make the response of the sensor tag not perfectly linear as explained in [30]. Fortunately, the level of nonlinearity can be made small by means of an accurate choice of the electronic components so that the resulting nonlinearity can be ignored as shown by the R^2 value calculated for the linear regression plotted in Figure 7, approaching unit.

On the other hand, observing the four curves of Figure 8, referring to tensile tests carried out on a unique aluminium specimen, it can be noted that they significantly differ from each other for the entity of the frequency increments recorded in consequence of the same strain variation produced in the specimen. This is because the frequency response of the sensor tag changes in the four experiments; in particular, the frequency domain and the slope of the response are different. The change of the frequency response is achieved by means of different choices of the values of the electronic components of the oscillating circuit [30]. Therefore, the sensor tag can be tuned to have different frequency responses, and each response is characterized by a different reference frequency value f_0 (i.e., the frequency in unstressed condition). For these reasons, the four series of measurements shown in Figure 8 have been operated into different domains and have different slopes.

Linear regressions plotted in Figure 8 are characterized from different slopes $\Delta\sigma/\Delta f$ and clearly show the highest bridge frequency changes when using the sensor tag characterized by the lowest initial frequency $f_0 = 8648\,\text{Hz}$ (red square dots curve). Consequently, the sensor tag characterized by higher initial frequencies results to be less and less sensitive to the same stress (strain) increments applied to the specimen.

4.3. Calibration. According to the wired method, a unique, direct proportion exists among the strain variations, the gauge resistance changes, and the consequent current change in the Wheatstone bridge of the compensator. Similarly, the sensor tag gives a change of the modulating frequency of the backscattered signal in accordance with the real strain increment occurring in the gauge, but it depends on the reference frequency in unstressed condition, and it suffers from possible nonlinearity. In order to investigate these characteristics of the sensor tag, a large campaign of new tensile tests has been carried out making the initial frequency f_0 variable in a wide range of values. As explained in [30], this is possible operating a suitable tuning of the electrical resistance of the circuit. Tensile tests are extended to specimens made of steel, brass, and aluminium, searching for a general validation of the proposed method. As already done in Figure 7, the linear regressions are calculated on the experimental data and the angular coefficient $E_{\text{app}} = \Delta\sigma/\Delta f$ of each of these regressions is regarded as an apparent value of elastic modulus for the tested specimens. Results of these calculations, made with respect to a significant number of different f_0 values and extended to all the considered materials, have been reported in Figure 9. The ratio $k = E_{\text{app}}/E_{\text{ref}}$ among the calculated values of apparent elastic modulus E_{app} and the reference value of the Young elastic modulus E_{ref} (previously determined for various materials using the wired method) have been plotted with respect to the considered initial frequency f_0. In particular, a unique linear relationship has been highlighted among experimental k values and f_0 ones, independently from the considered materials. That relationship can also be written as

$$k = \frac{E_{\text{app}}}{E_{\text{ref}}} = \frac{\Delta\sigma/\Delta f}{\Delta\sigma/\Delta\varepsilon} = \frac{\Delta\varepsilon}{\Delta f} \qquad (2)$$

so that the variable k represents the ratio among the real strain increment to be measured ($\Delta\varepsilon$) and the measured

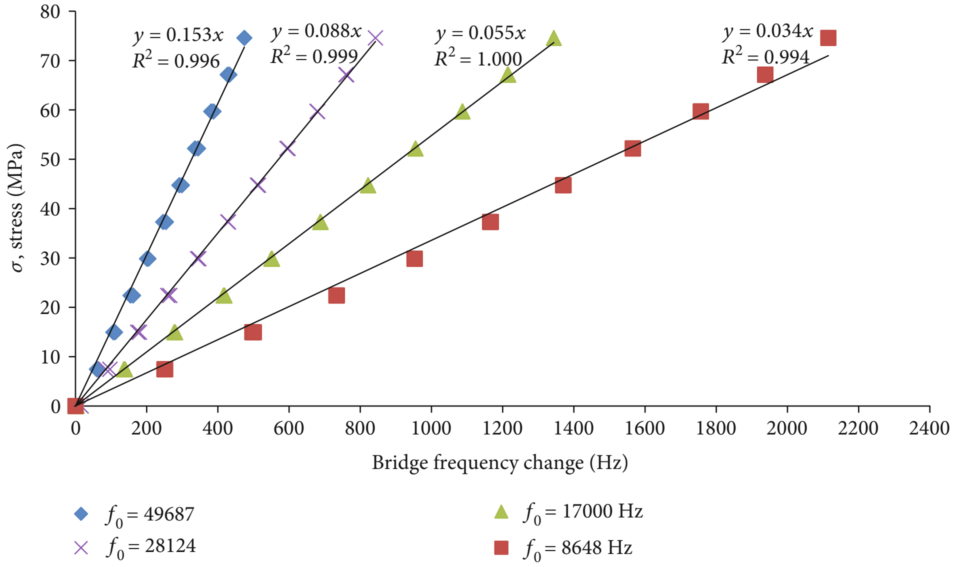


FIGURE 8: Effect of the tuning of the sensor tag. The frequency f_0 in unstressed condition depends on specific circuital parameters.

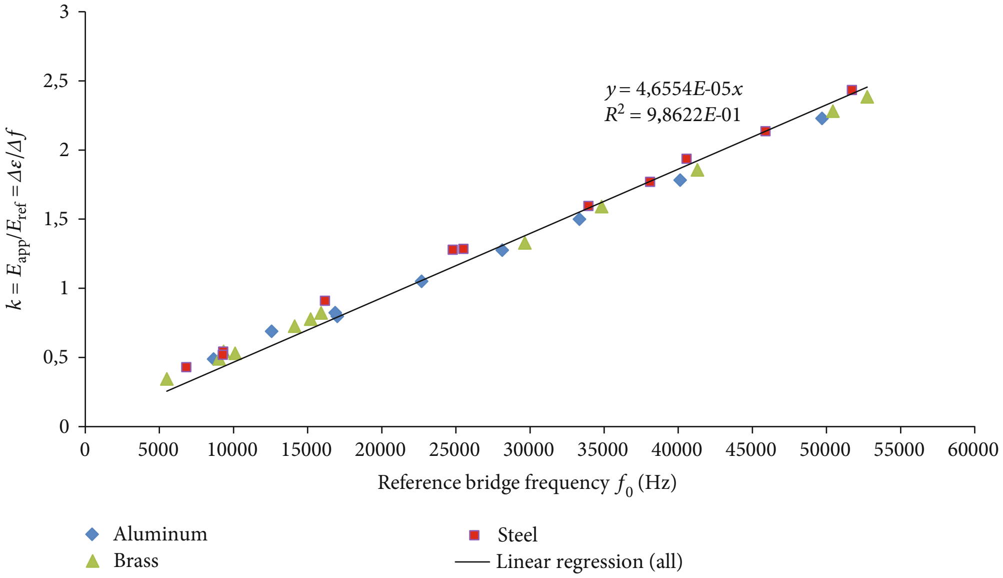


FIGURE 9: Experimental values for $k = E_{\text{app}}/E_{\text{ref}}$ as a function of the initial frequency f_0. Values of k also express the ratio among the real strain increment in the gauge ($\Delta\varepsilon$) and the measured frequency variation Δf.

frequency change Δf. This ratio depends on the considered frequency f_0. Then, linear regression calculated in Figure 9 provides the calibrating law to translate the measured frequency changes into the actual strain variation experienced by the specimen.

If nonlinearity of the experimental $\Delta\sigma$-Δf curves is taken into account, at each step i of a single tensile test, stress increments $\Delta\sigma_i = \sigma_{i+1}$-$\sigma_i$ and, consequently, single real strain increments $\Delta\varepsilon_i = \varepsilon_{i+1}$-$\varepsilon_i$ may be directly related to the correspondent frequency variations $\Delta f_i = f_{i+1}$-f_i. This allows to compute several local values of the apparent elastic modulus, here named secant values $E_{\text{sec},i} = E_{\text{sec}}(f_i) = (\sigma_{i+1}$-$\sigma_I)/(f_{i+1}$-$f_i)$ which depends from the actual frequency f_i at the step i. Consequently, a new definition for the ratio k (Equation (2)) can be given with respect to the $E_{\text{sec},i}$ values in place of the E_{app} ones, highlighting the continued dependence of ratio $k = E_{\text{sec}}/E_{\text{ref}}$ from the current f_i value in place of the initial value f_0. In these terms, ratio k can be regarded as a continue function of the actual bridge frequency f_i in accordance with the following expression:

$$k = k(f_i) = \frac{E_{\text{sec}}(f_i)}{E_{\text{ref}}} = \frac{(\sigma_{i+1} - \sigma_i)/(f_{i+1} - f_i)}{(\sigma_{i+1} - \sigma_i)/(\varepsilon_{i+1} - \varepsilon_i)} = \frac{\Delta\varepsilon_i}{\Delta f_i}. \quad (3)$$

Equation (3) highlights the existing relation between the real strain increment to be measured $\Delta\varepsilon_i$ and the frequency change Δf_i measured in place of it. The ratio k between these two variables is modelled as a continuous function of the instantaneous frequency f while strain increments (and consequent frequency changes) are taking place.

A plot of $k = k(f_i)$ calculated for each tensile test is shown in Figure 10.

In particular, experimental $\Delta\sigma$ vs. Δf curves recorded from tensile tests were first considered as sequences of individual steps i, whose stress increments $\Delta\sigma_i$ could be converted into the real strain variations $\Delta\varepsilon_i$ in accordance with Hooke's law $\Delta\varepsilon_i = \Delta\sigma_i/E_{\text{ref}}$ for each specific material. Subsequently, the values of the ratio $k = \Delta\varepsilon_i/\Delta f_i$ among the calculated real strain variations $\Delta\varepsilon_i$ and the related frequency changes Δf_i recorded at the step i have been computed and plotted with respect to the instantaneous bridge frequency value f_i considered at the beginning of each considered step i.

Data in Figure 10 refer to the entire bunch of experimental values obtained from tensile tests performed on all the three different couples of metal specimen. These data clearly show the real strain increment and the correspondent frequency change to be strictly related to each other. In particular, several regression laws have been calculated on the experimental data, including an exponential form, a second order polynomial, and a linear equation, respectively. The first two regression laws showed to fit experimental data not significantly better than the linear regression one, this latter being characterized from an R^2 value very close to unity.

Assuming $k = k(f_i)$ as a continue function of f_i, the differential $k = \Delta\varepsilon/\Delta f$ can be considered and the real strain variation $\Delta\varepsilon$ is calculated in accordance with the following integral expression:

$$\Delta\varepsilon = \int k(f)\,df. \tag{4}$$

If the linear regression law $k = 2.382E\text{-}05 \cdot f$ (indicated in Figure 10) is substituted in Equation (4), the strain increment $\Delta\varepsilon$ occurring in the gauge while the measured frequency changes from value f_1 to value f_2 is simply calculated as

$$\Delta\varepsilon = 1.1912 \cdot 10^{-5}\left(f_2^2 - f_1^2\right). \tag{5}$$

Once again, it is worth noticing that Equation (5) is independent from the material of the specimen. Also, the validity of Equation (5) extends to a very wide range of frequency values (from just a few Hz up to 60 kHz). On the other hand, measurements performed with a sensor tag having lower initial frequency result to be more sensitive than that having higher initial frequency. In fact, in the former case, the ratio $k = \Delta\varepsilon/\Delta f$ is less than unity. In this condition, the frequency variation of 1 Hz corresponds to a strain variation lower than 1 ε, representing a great improvement with respect to strain measurements performed with traditional techniques. In fact, electrical current variations in gauges are generally detected with such a limited resolution that only a precision of several ε may correspond, this representing a limit for the traditional method of measuring strains with wired gauges. In comparison, the proposed new method may enhance the strain measurement resolution even below 0, 2 ε if the initial frequency is set lower than about 10 kHz.

4.4. Experimental Data from the Wireless Method: Dynamic Measurements. A steel cantilever about 25 cm long has been anchored at one end to the structure of a steel staircase as shown in Figure 11 so that it can vibrate after an initial displacement of its free end. The cantilever is instrumented with the proposed sensor that changes the modulation frequency of the backscattered signal dynamically according to the vibration movement. In fact, the strain gauge alternates extension and compression in agreement with the natural frequency of the beam while the backscattered signal changes its modulating frequency according to the extension and compression of the strain gauge. The analog modulated signal coming back from the sensor is received, sampled, and recorded by a Universal Software Radio Peripheral (National Instruments NI USRP 2920). A subsequent demodulation and filtering permit to determine the vibration frequency and the damping of the vibration. Figure 12 shows the modulation frequency of the received signal with respect to time in the case of repeated bending (with a pulse load) of the free end of the cantilever with a time interval of about 5 seconds. After each bending, the load is quickly removed permitting the cantilever to have free vibration. Figure 12 shows four responses of the cantilever under repeated loading and free vibration. Figure 13 shows an enlargement of the third response. Until about 13.8 s the cantilever is static, the small ripple concerns hardware and numerical noise. From 13.8 s until about 14.1 s, the modulation frequency increases because of the progressive loading of the free end of the cantilever. After 14.1 s, the load is quickly removed and the cantilever is under free vibration; the modulation frequency becomes oscillating since it follows the alternating extension and compression of the strain gauge glued on the cantilever. The frequency of that oscillation corresponds to the natural frequency of the cantilever as shown in Figure 14 that is the spectrum of the signal of Figure 13. The continuous component, in Figure 14, depends on the part of the signal corresponding to the static condition of the cantilever, the peak at 33.14 Hz is the natural frequency of vibration of the cantilever, and the small peak at 50 Hz is the hardware noise (from supplying network). Because of the damping of vibration, the oscillation of Figure 13 (i.e., the modulation frequency) decreases along the time. The response of the sensor has been compared to that measured by a laser detector (also shown in Figure 12) obtaining excellent agreement (shown in Figure 15).

5. Conclusions

A new method for measuring strains has been proposed. It consists in a wireless technique based on the traditional strain gauges and integrated into an electrical circuit together with an RFID tag. While traditionally strain measurement is usually performed by reading the current variation occurring in the circuit as a consequence of the resistance change in the strain gauge, in the proposed new method, an interrogating antenna is used to detect the modulation frequency of an

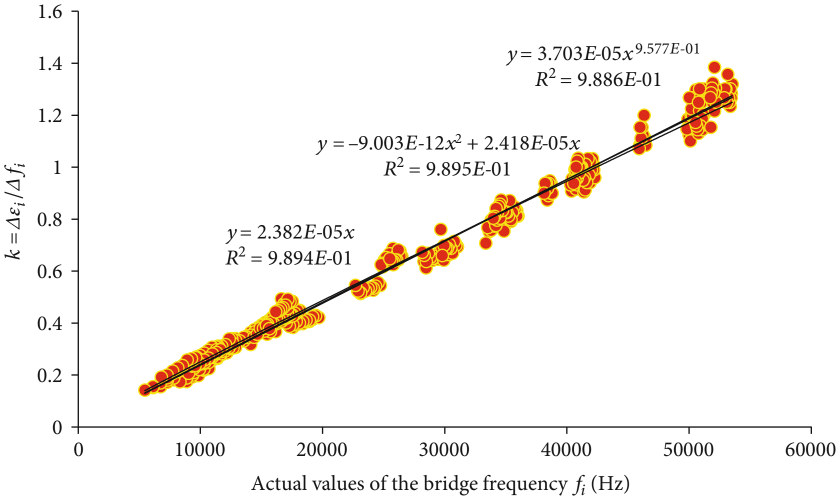


Figure 10: Different regressions carried out on experimental values of the ratio $k = \Delta\varepsilon_i/\Delta f_i$ considered as a function of the current bridge frequency f_i.

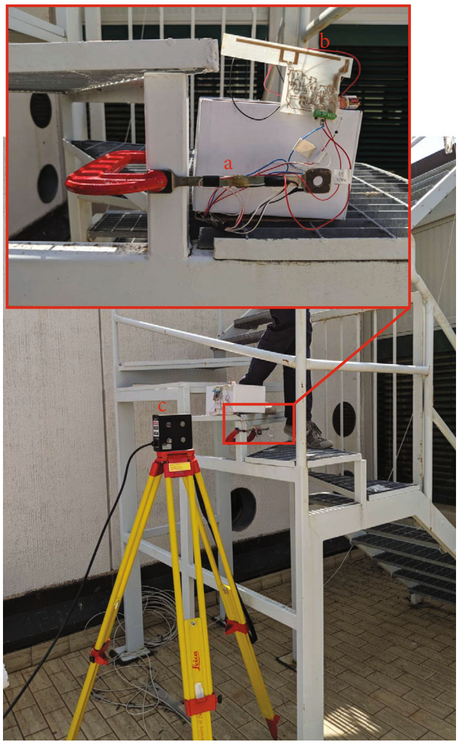


Figure 11: Dynamic measurement set-up: (a) instrumented cantilever anchored at one end to the structure of a steel staircase; (b) wireless sensor tag; (c) laser detector.

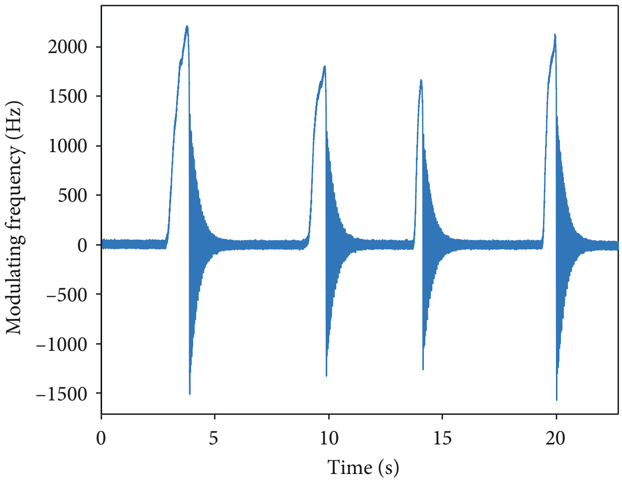


Figure 12: Response of the oscillating cantilever under repeated pulse loading. The enlargement of free vibration is shown in Figure 13.

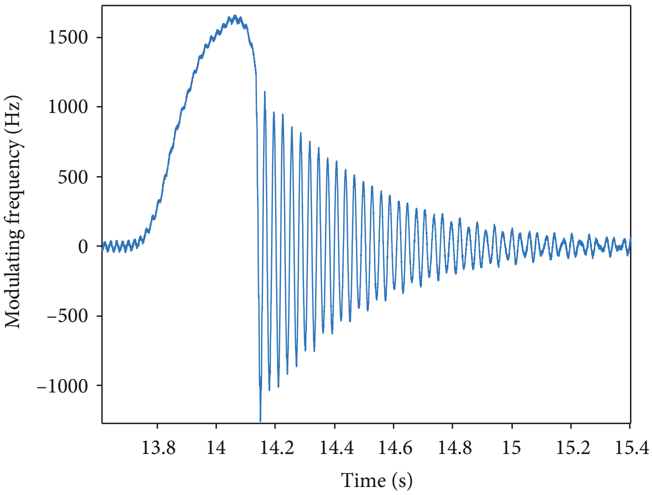


Figure 13: Enlargement of the third response of Figure 12. It shows the loading and free vibration of cantilever.

electromagnetic signal that varies in accordance with the strain to be measured.

Several tensile tests have been carried out on metal specimens made of different materials, including steel, brass, and aluminium. Feasibility of this new method has been proved determining a unique clear relationship among the strain values and the measured frequency changes.

The tag sensor is able to measure dynamic phenomena as assessed by experiments with vibrations of tens of Hz. The

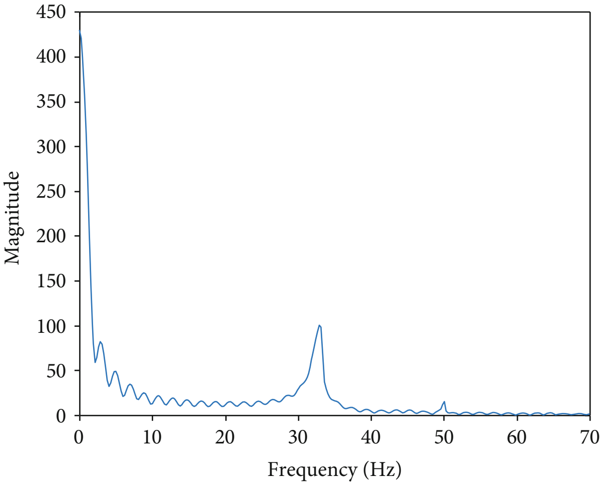


FIGURE 14: Spectrum of the signal of Figure 13: the continuous component (frequency zero) depends on the part of the signal corresponding to the static condition of the cantilever, the peak at 33.14 Hz is the natural frequency of vibration of cantilever, and the small peak at 50 Hz is the hardware noise (supplying network of the central unit).

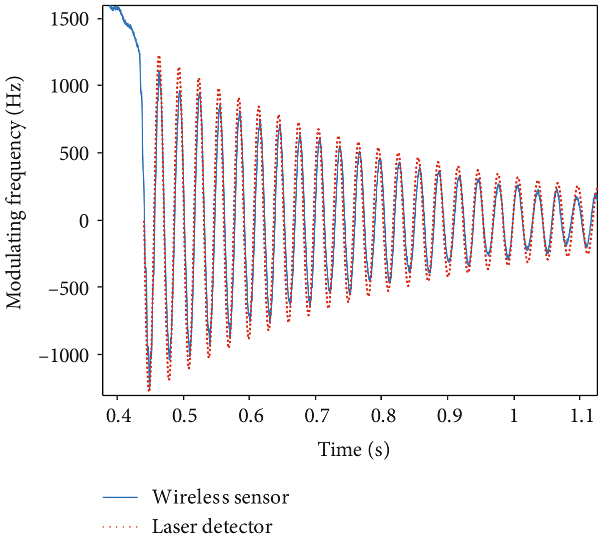


FIGURE 15: Comparison between the response of the wireless sensor and measurement with a laser detector (after a suitable scaling and translation).

detectable frequency of vibration is not limited by the sampling rate because the sampling is performed in the central unit (not in the tag sensor) that can be equipped with a suitable sampling device. The limit in measuring oscillating phenomena is given by the bandwidth of the strain gauge that is in the order of KHz, while the bandwidth of electronic components is much larger in the order of MHz. So, the tag sensor should be able to measure oscillations up to a few kHz.

A calibration of the measuring system has been proposed, showing that it remains valid for a large range of working frequencies and for large strain intervals as well. In comparison with a traditional strain measuring procedure, the accuracy of the proposed new technique has been proved to be potentially higher. Effectiveness of the proposed wireless method has been proved up to a maximum interrogating distance of 20 meters in a laboratory room and outdoor, making this new strain measuring technique suitable for structural monitoring.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This article was supported by the Department of Civil, Building and Environmental Engineering of the Università degli Studi dell'Aquila and the Department of Industrial and Information Engineering and Economics of the University of L'Aquila, Via Giovanni Gronchi 18, 67100, L'Aquila, Italy.

References

[1] B. F. Spencer, "Opportunities and challenges for smart sensing technology," in *Proceedings of the First International Conference on Structural Health Monitoring and Intelligent Infrastructure*, pp. 65–71, Japan. A A Balkema Publisher, 2003.

[2] E. G. Straser, A. S. Kiremidjian, T. H. Meng, and L. Redlefsen, "A modular, wireless network platform for monitoring structures," in *SPIE Vol. 3243, Proceedings of the 16th International Modal Analysis Conference*, pp. 450–456, 1998.

[3] J. P. Lynch, A. Sundararajan, K. H. Law, A. S. Kiremidjian, T. Kenny, and E. Carryer, "Embedment of structural monitoring algorithms in a wireless sensing unit," *Structural Engineering and Mechanics*, vol. 15, no. 3, pp. 285–297, 2003.

[4] K. Mitchell, S. Sana, V. S. Balakrishnan, V. S. Rao, and H. J. Pottinger, "Micro sensors for health monitoring of smart structures," in *Proceedings Volume 3673, Smart Structures and Materials 1999: Smart Electronics and MEMS*, pp. 351–358, Newport Beach, CA, USA, July 1999.

[5] Y. Yu, H. W. Li, and J. P. Ou, "Wireless acceleration sensor used for civil engineering structure monitoring and its integration technique," *Proceedings of the 3rd International Symposium on Instrumentation Science and Technology*, 2004, pp. 741–748, Harbin Institute of Technology Publisher, 1, China, 2004.

[6] Y. Yu, H. W. Li, and J. P. Ou, "*Design and development of wireless acceleration sensor applied to civil engineering structure*," *Journal of Transcluction Technology*, 2004.

[7] G. Park, T. Rosing, M. D. Todd, C. R. Farrar, and W. Hodgkiss, "Energy harvesting for structural health monitoring sensor networks," *Journal of Infrastructure Systems*, vol. 14, no. 1, pp. 64–79, 2008.

[8] M. Jeong, J.-G. Bae, and B.-H. Koh, "A feasibility study of damage tracking through the diffusive communication of wireless sensors," *International Journal of Precision Engineering and Manufacturing*, vol. 11, no. 1, pp. 23–29, 2010.

[9] L. Liu and F. G. Yuan, "Wireless sensors with dual-controller architecture for active diagnosis in structural health monitoring," *Smart Materials and Structures*, vol. 17, no. 2, article 025016, 2008.

[10] R. Melik, E. Unal, N. Kosku Perkgoz, C. Puttlitz, and H. V. Demir, "Flexible metamaterials for wireless strain sensing," *Applied Physics Letters*, vol. 95, no. 18, article 181105, 2009.

[11] R. Melik, N. K. Perkgoz, E. Unal, C. Puttlitz, and H. V. Demir, "Bio-implantable passive on-chip RF-MEMS strain sensing resonators for orthopaedic applications," *Journal of Micromechanics and Microengineering*, vol. 18, no. 11, article 115017, 2008.

[12] I. Mohammad and H. Huang, "An antenna sensor for crack detection and monitoring," *Advances in Structural Engineering*, vol. 14, no. 1, pp. 47–53, 2011.

[13] J.-T. Lin, K. W. Walsh, D. Jackson et al., "Development of capacitive pure bending strain sensor for wireless spinal fusion monitoring," *Sensors and Actuators A*, vol. 138, no. 2, pp. 276–287, 2007.

[14] F. Gasco, P. Feraboli, J. Braun, J. Smith, P. Stickler, and L. DeOto, "Wireless strain measurement for structural testing and health monitoring of carbon fiber composites," *Composites Part A: Applied Science and Manufacturing*, vol. 42, no. 9, pp. 1263–1274, 2011.

[15] P. C. Chang, A. Flatau, and S. C. Liu, "Review paper: health monitoring of civil infrastructure," *Structural Health Monitoring*, vol. 2, no. 3, pp. 257–267, 2003.

[16] A. Kesavan, S. John, and I. Herszberg, "Strain-based structural health monitoring of complex composite structures," *Structural Health Monitoring*, vol. 7, no. 3, pp. 203–213, 2008.

[17] U. Tata, S. Deshmukh, J. C. Chiao, R. Carter, and H. Huang, "Bio-inspired sensor skins for structural health monitoring," *Smart Materials and Structures*, vol. 18, no. 10, article 104026, 2009.

[18] A. Ibrahim and D. R. S. Cumming, "Passive single chip wireless microwave pressure sensor," *Sensors and Actuators A: Physical*, vol. 165, no. 2, pp. 200–206, 2011.

[19] R. Matsuzaki, M. Melnykowycz, and A. Todoroki, "Antenna/sensor multifunctional composites for the wireless detection of damage," *Composites Science and Technology*, vol. 69, no. 15-16, pp. 2507–2513, 2009.

[20] Y.-s. H. Son and J.-G. Kim, "A design and implementation of forest-fires surveillance system based on wireless sensor networks for South Korea Mountains," *IJCSNS International Journal of Computer Science and Network Security*, vol. 6, no. 9B, 2006.

[21] O. Jinping, "Research and practice of intelligent sensing technologies in civil structural health monitoring in the mainland of China," in *Proceedings Volume 6176, Nondestructive Evaluation and Health Monitoring of Aerospace Materials, Composites, and Civil Infrastructure V*, San Diego, CA, USA, March 2006.

[22] Y. Yu and J. Ou, "Wireless sensing experiments for structural vibration monitoring of offshore platform," *Frontiers of Electrical and Electronic Engineering in China*, vol. 3, no. 3, pp. 333–337, 2008.

[23] S. J. Dyke, J. M. Caicedo, and E. A. Johnson, "Monitoring of a benchmark structure for damage identification," in *Proceedings of the Engineering Mechanics Speciality Conference*, pp. 21–24, Austin, TX, USA, May 2000.

[24] Stanford University, USA, "Wireless structural damage monitoring," *Smart Materials Bulletin*, vol. 2002, no. 1, p. 16, 2002.

[25] J. P. Lynch and K. Loh, "A summary review of wireless sensors and sensor networks for structural health monitoring," *The Shock and Vibration Digest*, vol. 38, no. 2, pp. 91–128, 2006.

[26] D. D. L. Mascarenas, E. B. Flynn, M. D. Todd et al., "Development of capacitance-based and impedance-based wireless sensors and sensor nodes for structural health monitoring applications," *Journal of Sound and Vibration*, vol. 329, no. 12, pp. 2410–2420, 2010.

[27] E. Köppe and M. Bartholmai, "Wireless sensor network with temperature compensated measuring technology for long-term structural health monitoring of buildings and infrastructures," *Procedia Engineering*, vol. 25, pp. 1032–1036, 2011.

[28] S.-D. Jang and J. Kim, "Passive wireless structural health monitoring sensor made with a flexible planar dipole antenna," *Smart Materials and Structures*, vol. 21, no. 2, article 027001, 2012.

[29] S.-D. Jang, B.-W. Kang, and J. Kim, "Frequency selective surface based passive wireless sensor for structural health monitoring," *Smart Materials and Structures*, vol. 22, article 025002, no. 2, 2013.

[30] E. DiGiampaolo, A. DiCarlofelice, and A. Gregori, "An RFID-enabled wireless strain gauge sensor for static and dynamic structural monitoring," *IEEE Sensors Journal*, vol. 17, no. 2, pp. 286–294, 2017.

Hindawi
Journal of Sensors
Volume 2019, Article ID 4581672, 14 pages
https://doi.org/10.1155/2019/4581672

Research Article

A Deep Learning Model for Concrete Dam Deformation Prediction Based on RS-LSTM

Xudong Qu,[1] **Jie Yang**,[1] **and Meng Chang**[2]

[1] *Institute of Water Resources and Hydro-Electric Engineering, Xi'an University of Technology, Xi'an 710048, China*
[2] *Department of Development, Sino Hydro Engineering Bureau 15 Co., Ltd, Xi'an 710016, China*

Correspondence should be addressed to Xudong Qu; qxd@stu.xaut.edu.cn

Received 23 June 2019; Accepted 14 September 2019; Published 31 October 2019

Guest Editor: Samir Mustapha

Deformation is a comprehensive reflection of the structural state of a concrete dam, and research on prediction models for concrete dam deformation provides the basis for safety monitoring and early warning strategies. This paper focuses on practical problems such as multicollinearity among factors; the subjectivity of factor selection; robustness, externality, generalization, and integrity deficiencies; and the unsoundness of evaluation systems for prediction models. Based on rough set (RS) theory and a long short-term memory (LSTM) network, single-point and multipoint concrete dam deformation prediction models for health monitoring based on RS-LSTM are studied. Moreover, a new prediction model evaluation system is proposed, and the model accuracy, robustness, externality, and generalization are defined as quantitative evaluation indexes. An engineering project shows that the concrete dam deformation prediction models based on RS-LSTM can quantitatively obtain the representative factors that affect dam deformation and the importance of each factor relative to the effect. The accuracy evaluation index (AVI), robustness evaluation index (RVI), externality evaluation index (EVI), and generalization evaluation index (GVI) of the model are superior to the evaluation indexes of existing shallow neural network models and statistical models according to the new evaluation system, which can estimate the comprehensive performance of prediction models. The prediction model for concrete dam deformation based on RS-LSTM optimizes the factors that influence the model, quantitatively determines the importance of each factor, and provides high-performance, synchronous, and dynamic predictions for concrete dam behaviours; therefore, the model has strong engineering practicality.

1. Introduction

Due to unique advantages in design, construction, and operational management, concrete dams account for a large proportion of all dams and have become the preferred dam type for the construction of high dams. However, most of the concrete dam projects are located in harsh alpine valleys. Thus, the dams are subjected to various dynamic, static, and special cyclic loads during service, and the design, construction, and operational management must be tailored to these conditions. Therefore, service safety behaviour involves a nonlinear dynamic process that includes material and structure interactions and multiple factors [1]. As a comprehensive variable that reflects the safety state of concrete dams, deformation can be used as an important index of structural behaviours and trends. Therefore, strengthening the prediction models for deformation, conducting safety monitoring, and establishing early warning systems are important ways to ensure long-term service safety of concrete dams [2].

In recent years, the successful application of dam engineering theory, finite element theory, and artificial intelligence (AI) technology has greatly promoted the development of concrete dam deformation prediction models. The most commonly used methods [3] for influential factor selection in concrete dam deformation prediction models include prior knowledge, linear correlation coefficient, stepwise regression, principal component analysis (PCA), and grey correlation analysis methods. However, in actual applications, the prior knowledge method relies too much on experience and has large errors. Notably, the water pressure, temperature, and dam age are generally selected as influential factors in hydrostatic seasonal temporal (HST) models considering simplified physical models of dams and dam foundations, the burial conditions of monitoring equipment, prototype monitoring data, engineering mechanical analysis, and deductive investigation. The limitation of the PCA method is that only linear relations between variables are considered. If the dependence is nonlinear, the misinterpretation of results may occur. The grey correlation analysis method can only sort factors according to their relevance, and there is no clear criterion for selecting influential factors. Moreover, multiple collinearity can exist among the factors selected by conventional methods, which may reduce the accuracy of the model and adversely affect the prediction results [4]. Meanwhile, prediction models do not consider the influence of nonquantitative factors such as the seepage flow, crack opening degree, and lifting pressure; the dam construction materials; the construction quality; and the geological conditions. Additionally, model interpretation is important for evaluating the performance of prediction models, especially the model accuracy. The HST model has been traditionally used to identify the response of a dam to a considered action, such as a hydrostatic load, or to variations in factors such as temperature and time [5]. However, such analyses are only valid if the predictor variables are independent, which is not generally true [6]. In contrast, intelligent models (such as neural network, multilayer perceptron, and support vector machine models) have not been applied to interpret dam behaviour. Traditional models are frequently termed "black box" models, in reference to their lack of interpretability. Therefore, in the selection process of the factors that influence concrete dam deformation prediction models, imperfect selection criteria and neglecting important factors can seriously affect the prediction performance of the model. Single-point statistical models, deterministic models, and hybrid models [7–10] have evolved into multipoint intelligent models [11–16]. Based on the traditional statistical model, Gu et al. treated deformation at multiple measurement points and the spatial coordinates of these points as variables and established a spatiotemporal distributed prediction model of the deformation field of a concrete dam. Li et al. investigated the spatial and temporal expression of the factors that affected the deformation of an RCC dam and established a spatiotemporal deformation prediction model for RCC dams based on measured data. The prediction results agreed to the actual dam deformation data. Li et al. used the strong functional nonlinear mapping ability of a back propagation (BP) neural network to replace the complex factor subset in the traditional spatial deformation field model with water level, temperature, time, and measurement point variables as the input of the neural network. A BP network prediction model was established for dam deformation at multiple points. Chen et al. proposed a spectral decomposition method to decompose the monitoring data collected at multiple measurement points into several mutually independent latent variables for noise reduction and monitoring data processing. A least square support vector machine prediction model was established between the environmental data and latent variables, and the horizontal displacement of Mianhuatan Dam was successfully predicted. Many scholars have addressed these issues. The successful application of new methods has expanded the theoretical knowledge of dam deformation prediction and model establishment and provided important guiding significance for engineering practice. However, due to the complexity of concrete dam engineering, the structural volatility of dams, and the uncertainty of working conditions, there are still some shortcomings in existing prediction models. It is difficult for some models to process massive amounts of monitoring data in real time with extensive mining data mechanisms for high-performance prediction targets, such as those in practical applications. It is important to appropriately evaluate the prediction performance of a model from all angles because the practical value of the models can be guaranteed, different models can be compared, and different warning thresholds can be defined. There are various indexes [17] that can be used to assess how well a model matches the observed data, among which the most commonly used are the mean squared error (MSE), root mean squared error (RMSE), coefficient of determination (R^2), mean absolute error (MAE), mean absolute percentage error (MAPE), and average relative variance (ARV). The result of any of these indexes is frequently equivalent to a given prediction task. Specifically, an accurate model will have small MSE, RMSE, MAE, and MAPE values and high R^2 and ARV values. However, these accuracy indexes have differences that can be relevant but are often not considered [18]. Commonly, robustness and generalization ability are neglected in the model assessment, and quantitative evaluation indexes are not always used in practical applications. Therefore, it is necessary to explore methods for factor selection, establish high-performance, dynamic, synchronous prediction models, and design a scientific and comprehensive evaluation system which are urgent for concrete dam deformation prediction.

Attribute reduction is one of the core concepts of RS theory, which addresses incompleteness, redundancy, and ambiguity in data in the field of machine learning. This approach avoids the use of complex discernibility matrices and uses attribute importance as heuristic information to obtain inductive sets and importance analysis results; excellent results can be obtained in factor selection for prediction models based on RS theory [19–21]. Moreover, long short-term memory (LSTM) based on the memory architecture in deep learning (DL) can overcome the memory shortage and vanishing gradient issues of recurrent neural networks (RNNs). Besides, this method is characterized by controllable memory and rapid convergence. LSTM has achieved good practical application results in the dynamic and deep processing of massive, long-term, dependent data series [22–25]. To overcome the shortcomings of existing concrete dam deformation prediction models, RS theory and an LSTM network are applied to a concrete dam deformation prediction model in virtue of Tensor Flow. Finally, a concrete dam deformation prediction model based on RS-LSTM is established, and a new predictive model evaluation system is proposed.

2. Materials and Methods

2.1. Rough Set Theory. RS theory was proposed by Polish scholar Pawlak in the 1980s. The core objectives are the mining and refining of essential information under the premise of maintaining equivalence relations. The main tasks in this approach are attribute reduction, correlation analysis, and importance evaluation for uncertain information systems.

2.1.1. Information System. To describe the samples that encompass the necessary information in RS theory, a quaternary information system S is established, and it can be expressed as follows:

$$S = \{U, R, V, f\}, \tag{1}$$

where U is a nonempty finite set of all samples; R is a set of attributes, including a set of conditional attributes C and a set of decision attributes D; V is the attribute value set; and f is the information function, also known as the decision table.

2.1.2. Attribute Reduction. For arbitrary $P \subseteq R$ and $P \neq \emptyset$, the indistinguishable relationship between P and U is defined as follows.

$$\mathrm{IND}(P) = \left\{(x, y) \in U^2 \,|\, \forall \alpha \in P, \alpha(x) = \alpha(y)\right\}. \tag{2}$$

For an arbitrary set of objects $X \subseteq U$ and attributes $B \subseteq C$ in a given information system S, the approximation of X is defined as $B\underline{X} = \{x | [x]_B \subseteq X\}$; the approximate definition of X is defined as $\bar{B}X = \{x | [x]_B \cup X \neq \emptyset\}$; and the boundary area of x is defined as $BN_B(X) = \bar{B}X - \underline{B}X$. In this case, $[x]_B$ represents the set of indistinguishable relations for the division of U by B.

If $BN_B(X)$ is not empty, then X is called a rough set of B. The positive region of B relative to D is as follows.

$$\mathrm{POS}_B(D) = \left\{\frac{\underline{B}X | X \in U}{\mathrm{IND}(D)}\right\}. \tag{3}$$

When $\mathrm{SIM} = \mathrm{POS}_C(D) - \mathrm{POS}_{C-\{a\}}(D) = 0$, where $a \in C$, a can be omitted. Additionally, when each element in C is not omissible from D, it can be concluded that C is independent of D. When $C' = C - C^*$, where C' is independent of D and all the elements in C^* can be omitted, then C' is called the relative reduction of D.

2.1.3. Importance Evaluation. In attribute reduction, the importance of the attribute can be defined by the degree of interdependence between the attribute sets B and D. The degree of interdependence between P and R is defined as follows:

$$\gamma_B(D) = \frac{|\mathrm{POS}_B(D)|}{|U|}, \tag{4}$$

where $|\cdot|$ represents the cardinality value of a set.

The importance of the conditional attribute a to the decision attribute D based on the attribute dependency degree is defined as follows.

$$\mathrm{Sig}(\alpha, B, D) = \gamma_B(D) - \gamma_{B-\{\alpha\}}(D). \tag{5}$$

2.2. LSTM Network Based on a Memory Architecture. LSTM is obtained by improving the hidden layer of the RNN structure. LSTM based on a memory architecture can overcome memory shortage and vanishing gradient problems. The LSTM model structure is shown in Figure 1. The key advantages of LSTM are twofold. Notably, the hidden layer includes a hidden state and a cell state, and a threshold mechanism is established in the RNN. These factors strengthen the ability of the model to learn current information, extract the information and rules associated with the data, and simultaneously transmit information to reduce memory use. The threshold mechanism uses input gates, forget gates, and output gates to selectively memorize the feedback parameters of the feedback error function as the gradient decreases, achieving rapid gradient convergence [26].

2.2.1. Input Gate Updates. The input gate controls the information $x^{(t)}$ transmitted from the input of the network at moment t and hidden state at the final moment $h^{(t-1)}$ to the cell state $C^{(t)}$. The function of the input gate is to filter new information. The structure of an input gate is shown in Figure 2.

Figure 2 shows that the input gate consists of two parts. The first part selects the sigmoid activation function, for which the output is $i^{(t)}$, and the second part selects the tanh activation function, for which the output is $a^{(t)}$. The two partial outputs are multiplied to update the cell state. The renewal process can be mathematically expressed as follows:

$$\begin{aligned} i^{(t)} &= \sigma\left(W_i h^{(t-1)} + U_i x^{(t)} + b_i\right), \\ a^{(t)} &= \tanh\left(W_a h^{(t-1)} + U_a x^{(t)} + b_a\right), \end{aligned} \tag{6}$$

where W_i, U_i, b_i, W_a, U_a, and b_a are the weights and biases of the input gate and σ is the sigmoid activation function.

2.2.2. Forget Gate Updates. The forget gate controls the information transmitted from the cell state $C^{(t-1)}$ at moment $t-1$ to the cell state $C^{(t)}$ at moment t, and the information that should be discarded is identified. The structure of the forget gate is shown in Figure 3.

Figure 3 shows that the hidden state $h^{(t-1)}$ at moment $t-1$ and the input $x^{(t)}$ at moment t activate the sigmoid function, and the output $f(t)$ is in the range of [0, 1]. This value represents the probability of forgetting the information associated with the cell state at a previous moment. The renewal process can be mathematically expressed as follows:

$$f^{(t)} = \sigma\left(W_f h^{(t-1)} + U_f x^{(t)} + b_f\right), \tag{7}$$

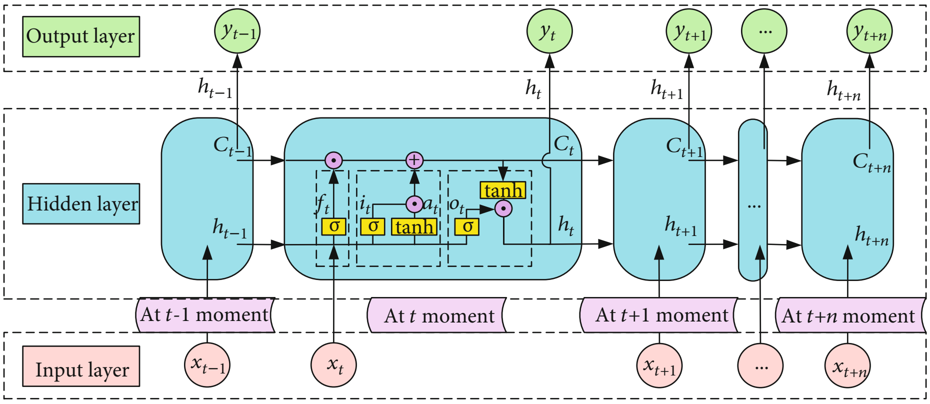


FIGURE 1: Long short-term memory network model structure.

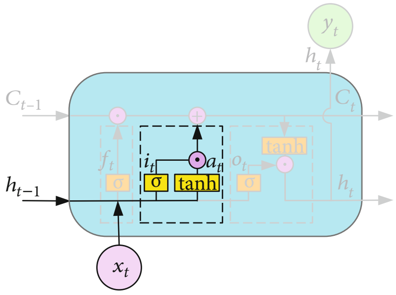


FIGURE 2: Input gate structure.

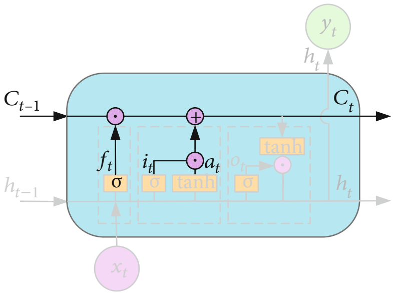


FIGURE 4: Cell state structure.

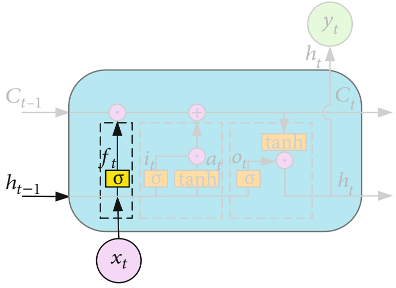


FIGURE 3: Forget gate structure.

where W_f, U_f, and b_f are the weights and biases of the forget gate.

2.2.3. Cell State Updates. The cell state controls the information $a^{(t)}$ transmitted from the result of the input gate $f(t)$ and the result of the forget gate $i^{(t)}$ to the cell state $C^{(t)}$. The structure of a cell state is shown in Figure 4.

Figure 4 shows that the cell state updating result $C^{(t)}$ is mainly determined by the cell state $C^{(t-1)}$ at moment $t-1$ and the results of the input and forget gates ($f(t)$, $i^{(t)}$, and $a^{(t)}$) at moment t. The renewal process can be mathematically expressed as follows:

$$C^{(t)} = C^{(t-1)} \odot f^{(t)} + i^{(t)} \odot a^{(t)}, \tag{8}$$

where $\odot$ is the Hadamard product.

2.2.4. Output Gate Updates. The output gate controls the information transmitted from the hidden state $h^{(t-1)}$ at moment $t-1$, the cell state $C^{(t)}$ at moment t, and the input $x^{(t)}$ at moment t. The function of the output gate is to determine the final retained information. The structure of an output gate is shown in Figure 5.

Figure 5 shows that the hidden state $h^{(t)}$ at moment t contains two parts. The first part $o^{(t)}$ is determined by the hidden state $h^{(t-1)}$ at moment $t-1$, the input $x^{(t)}$ at moment t, and the sigmoid activation function. The other part is determined by the cell state $C^{(t)}$ at moment t and the tanh activation function. The renewal process can be mathematically expressed as follows:

$$\begin{aligned} o^{(t)} &= \sigma\left(W_o h^{(t-1)} + U_o x^{(t)} + b_o\right), \\ h^{(t)} &= o^{(t)} \odot \tanh\left(C^{(t)}\right), \end{aligned} \tag{9}$$

where W_f, U_f, and b_f are the weights and biases of the output gate.

2.2.5. Output Layer Updates. The output of the model is determined by the hidden state $h^{(t)}$ at moment t and the sigmoid activation function. The renewal process can be mathematically expressed as follows:

$$y\wedge^{(t)} = \sigma\left(Vh^{(t)} + c\right), \tag{10}$$

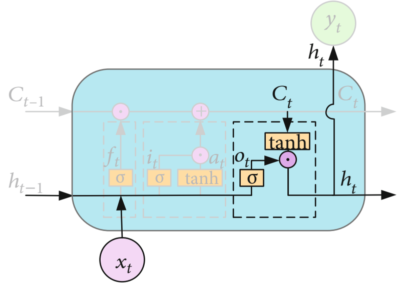


Figure 5: Output gate structure.

where V and c are the weight and bias of the output layer, respectively.

2.2.6. Model Parameter Updating. To obtain the optimal solution, this paper iteratively updates all the parameters in the LSTM model based on the gradient descent algorithm and error BP algorithm.

The objective loss function $L(t)$ is defined to minimize the sum of squared residuals between the predictions $y\wedge^{(t)}$ of the output layer and the target outputs $y^{(t)}$. $L(t)$ is divided into two parts: the loss $l(t)$ at moment t and the subsequent loss $l(t+1)$ moments later.

$$L(t) = \begin{cases} l(t) + l(t+1), & t < \tau, \\ l(t), & t = \tau. \end{cases} \tag{11}$$

The gradients of the hidden state $h^{(t)}$ and cell state $C^{(t)}$ are defined as $\delta_h^{(t)}$ and $\delta_C^{(t)}$, respectively, and the gradient at position τ can be expressed as follows.

$$\delta_h^{(\tau)} = \frac{\partial L(\tau)}{\partial h^{(\tau)}} = \frac{\partial L(\tau)}{\partial O(\tau)} \frac{\partial O(\tau)}{\partial h^{(\tau)}} = V^T \left(y\wedge^{(\tau)} - y^{(\tau)} \right), \tag{12}$$

$$\begin{aligned} \delta_C^{(\tau)} = \frac{\partial L(\tau)}{\partial C^{(\tau)}} = \frac{\partial L(\tau)}{\partial h(\tau)} \frac{\partial h(\tau)}{\partial C(\tau)} = \partial h(\tau) \odot o^{(\tau)} \odot \\ \cdot \left(1 - \tanh^2 \left(C^{(\tau)} \right) \right). \end{aligned} \tag{13}$$

The output gradient error at a given moment is determined in two parts, respectively, because $\delta_h^{(t)}$ and $\delta_C^{(t)}$ are obtained for $l(t)$ and $l(t+1)$. Thus, according to equations (12) and (13), the gradients of the hidden state $h^{(t)}$ and cell state $C^{(t)}$ can be expressed as follows.

$$\delta_h^{(t)} = \frac{\partial L}{\partial h^{(t)}} = V^T \left(y\wedge^{(t)} - y^{(t)} \right) + \delta_h^{(t+1)} \partial h^{(t+1)} / \partial h^{(t)}, \tag{14}$$

$$\delta_C^{(t)} = \frac{\partial L}{\partial C^{(t)}} = \delta_C^{(t+1)} f^{(t+1)} + \delta_h^{(t)} \odot o^{(t)} \odot \left(1 - \tanh^2 \left(C^{(t)} \right) \right). \tag{15}$$

According to equations (14) and (15), the following formula can be obtained.

$$\begin{aligned} \frac{\partial L}{\partial W_f} &= \sum_{t=1}^{\tau} \frac{\partial L}{\partial C^{(t)}} \frac{\partial C^{(t)}}{\partial f^{(t)}} \frac{\partial f^{(t)}}{\partial W_f} \\ &= \sum_{t=1}^{\tau} \left[\delta_C^{(t)} \odot C^{(t-1)} \odot f^{(t)} \odot \left(1 - f^{(t)} \right) \right] \left(h^{(t-1)} \right)^T. \end{aligned} \tag{16}$$

The other parameters in the model are derived similarly. The updating step size and learning rate of the model are defined as λ and α, respectively. The parameters in the LSTM model are iteratively updated using the gradient BP algorithm. The corresponding formula can be expressed as follows:

$$\beta_{t+1} = \beta_t - \alpha \frac{\partial L}{\partial \beta}, \tag{17}$$

where β represents the parameters in the LSTM model and $\partial L / \partial \beta$ represents the gradients of the parameters.

In summary, the updating process of the parameters in the LSTM network model based on a memory architecture can be expressed as follows. First, a parameter initialization process is implemented. Second, the iterative process is repeated by the gradient descent algorithm and the error BP algorithm until the target loss function converges. Finally, the parametric optimal solution of the LSTM model is obtained. Moreover, the Dropout algorithm is adopted in the training process of the LSTM model to avoid the overfitting phenomenon [27] and improve network performance by preventing feature detectors from working together.

2.3. Concrete Dam Deformation Prediction Model Based on RS-LSTM. With the advantages of RS, the mapping relationship between the factors that influence dam operating behaviour and the corresponding effects is established under the premise of retaining the key information. Additionally, the redundant information is eliminated, the expression space of the influential factors is simplified, and the importance of each factor is evaluated. Moreover, because the LSTM model overcomes the memory shortage and gradient dissipation issues of traditional RNNs and is characterized by controllable memory and fast gradient convergence, the model yields high-performance dynamic predictions based on long-term data series. Therefore, by combining the advantages of RS theory and the LSTM network, this paper establishes a concrete dam deformation prediction model based on RS-LSTM, and the prediction model is optimized considering the relevant influential factors and interactive mechanisms between these factors and concrete dam deformation in a quantitative manner. The process of establishing a concrete dam deformation prediction model based on RS-LSTM is shown in Figure 6. The specific modeling steps are as follows.

2.3.1. Data Acquisition. Statistical methods are used to perform gross error processing for concrete dam monitoring data. Such methods provide a reliable data foundation for the establishment of prediction models. Attribute reduction in

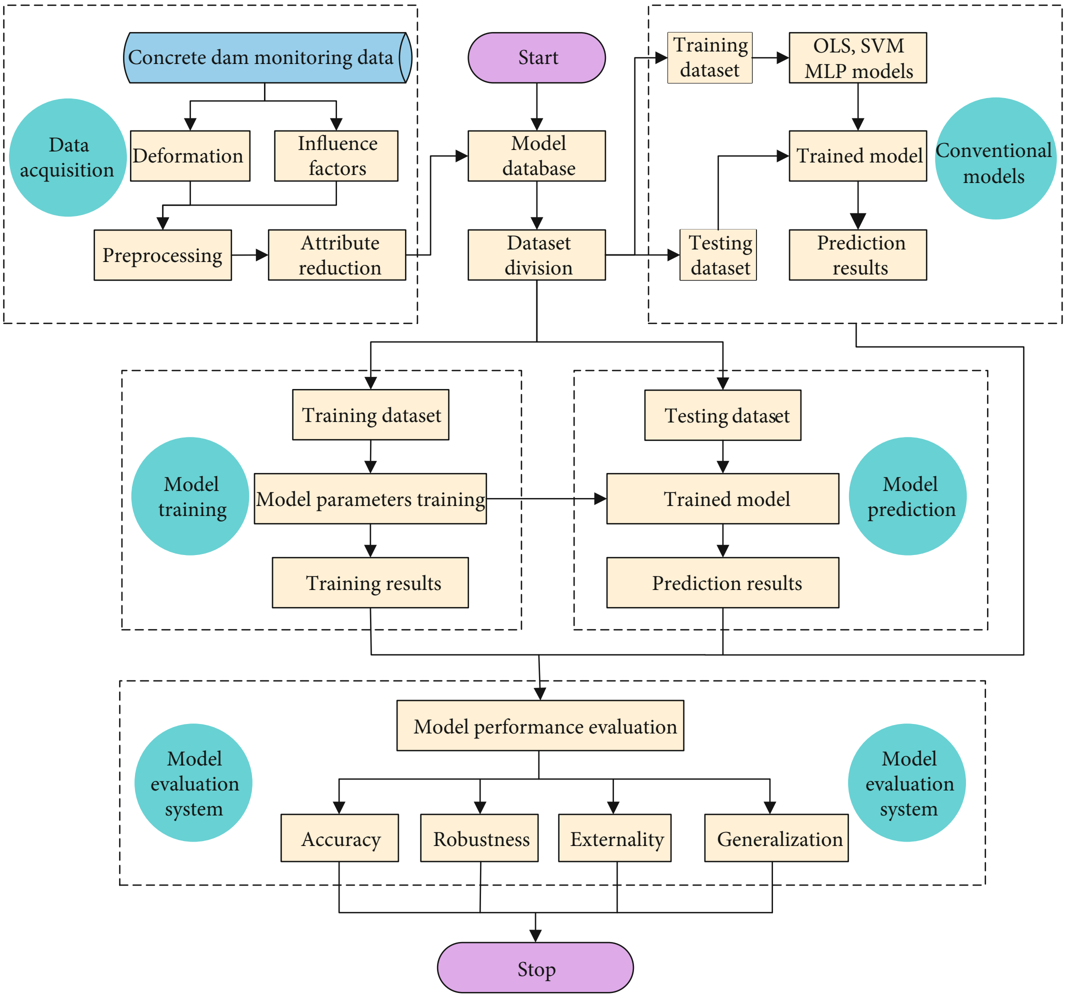


Figure 6: Process of establishing the concrete dam deformation prediction model based on RS-LSTM.

RS theory is conducted based on a complex multivariate dataset composed of water depth, temperature, seepage flow, fracture aperture, and uplift pressure information to accurately obtain the representative factors that affect the deformation behaviours of concrete dams. Deformation monitoring data and the representative influential factors corresponding to certain measurement points are selected as the model dataset. The representative factor dataset is standardized using an independent standardization formula, and the model dataset is divided into a training set and testing set by a cross-validation method.

2.3.2. Model Training. The preprocessed and standardized training set samples are used as model inputs. Error back propagation based on the gradient descent algorithm drives the model loss function to converge, and the optimal model parameters are obtained. The Dropout algorithm is used to overcome the problem of overfitting in training, and finally, a prediction model with optimal parameters is obtained.

2.3.3. Model Prediction. The testing set samples are input into the trained prediction model to obtain the corresponding deformation prediction results.

2.3.4. Model Performance Evaluation. According to the established evaluation system, the results of the concrete dam deformation prediction model based on LSTM, a classical least squares (OLS) model, a support vector machine (SVM) model, and a multilayer perceptron (MLP) model with 2 hidden layers are compared based on accuracy, robustness, externality, and generalization.

2.4. Evaluation System for the Concrete Dam Deformation Prediction Model. A concrete dam deformation prediction model plays an important role in operational behaviour monitoring, real-time abnormality detection, and decision-making, and its performance directly affects condition assessments and early warning strategies. In actual application processes, a single accuracy evaluation index may have certain limitations, and it is often impossible to evaluate the robustness, externality, and generalization of a model. Therefore, a complete evaluation system for concrete dam deformation prediction models must be established for practical applications. Therefore, this paper evaluates model performance from the aspects of accuracy, robustness, externality, and generalization, and quantitative evaluation indexes are used to comprehensively evaluate the performance of the concrete dam deformation prediction model based on statistical theory.

2.4.1. Accuracy. The accuracy of the concrete dam deformation prediction model refers to the degree of agreement between the predicted and true values. This evaluation index is the most widely used in model assessment. In actual engineering, the MAPE, MSE, and RMSE are usually selected to evaluate the accuracy of a model. Considering the nonstationarity of deformation monitoring data and the overlap among evaluation indexes, the RMSE_P and MAPE_P are selected to establish the accuracy evaluation index (*AEI*) of the concrete dam deformation prediction model. The corresponding formulas are defined as follows.

$$\begin{aligned}\mathrm{RMSE}_P &= \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t - y\wedge_t\right)^2},\\ \mathrm{MAPE}_P &= \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|.\end{aligned} \tag{18}$$

2.4.2. Robustness. The robustness of a concrete dam deformation prediction model refers to its resistance to the inherent errors in training data. Model training and prediction are performed by establishing normal training samples and training samples with a certain degree of random error. The ability of a model to learn the true nonlinear mapping relationships when there is a small gross error in the training set is tested. The absolute difference between the RMSE_O of the training model prediction results with no gross error and the RMSE_E of the training model prediction results with gross error is selected as the robustness evaluation index (REI) for the concrete dam deformation prediction model. The corresponding formula is defined as follows.

$$\mathrm{REI} = \left|\mathrm{RMSE}_O - \mathrm{RMSE}_E\right|. \tag{19}$$

2.4.3. Externality. The externality of the concrete dam deformation prediction model refers to its adaptability to accurately process samples outside the training set with the same mapping relationship. A high-performance model based on its externality ability can learn the mapping relationships hidden in data through training set. Even if some samples are outside the training set, a model with a satisfactory externality can achieve accurate predictions. The samples outside the training set are fused with the testing samples, and the prediction performance of the model based on a training set with the same mapping relationship is tested. The accuracy index of the model under this condition, the RMSE_P, is selected as the externality evaluation index (EEI). The corresponding formula is defined as follows.

$$\mathrm{EEI} = \mathrm{RMSE}_P = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t - y\wedge_t\right)^2}. \tag{20}$$

2.4.4. Generalization. The generalization of a concrete dam deformation prediction model refers to its adaptability to process samples with the same mapping relationship. A poor generalization ability can lead to overfitting. In such cases, the model error for the training set is very low, but the error is very large for the testing set. The model is optimized by adding training samples, performing regularization processing, and applying the Dropout algorithm to improve its generalization performance. Selecting the ratio of RMSE_T in the training process to RMSE_P in the prediction process is selected as the generalization evaluation index (GEI). The corresponding formula is defined as follows.

$$\mathrm{GEI} = \frac{\mathrm{RMSE}_P}{\mathrm{RMSE}_T}. \tag{21}$$

In each evaluation index formula above, $y(t)$ represents a measured value; $\hat{y}(t)$ represents a predicted value; n represents the number of predicted samples; the subscript T represents the training process; the subscript P represents the prediction process; the subscript O represents samples with no gross error; and the subscript E represents samples with gross error.

2.5. Simulation Environment and Engineering Project. Concrete dam deformation prediction models based on OLS, SVM, MLP, and LSTM are established in accordance with the horizontal displacement of concrete gravity dams, and the evaluation system is used to evaluate the accuracy, robustness, externality, and generalization of each model. Additionally, a comparative analysis is performed. The simulation environment includes the Windows 10 operating system, an Intel Core i5 CPU, 8 GB of memory, the Python programming language version 3.7.2rcl, and the TensorFlow deep learning framework version 1.12.0.

2.5.1. Engineering Situation. Zhouning Hydropower Station is a diversion-type power station on the Muyang River in Fujian Province that performs step exploitation. The total installed capacity is 250 MW, the total storage capacity of the reservoir is 47 million m^3, and the designed flood level is 633.00 m. The power station consists of a barrage, a sluice building, a water conveyance system, an underground powerhouse, and a ground switch station. The barrage is an RCC gravity dam with a foundation plane elevation of 562.00 m, a maximum dam height of 72.40 m, and a dam crest length of 206.00 m. The body of Zhouning Dam is divided into nine dam sections, of which Nos. 1-4 and Nos. 7-9 are nonoverflow sections and Nos. 5-6 are overflow sections.

The deformation monitoring data collected by Zhouning Hydropower Station include horizontal and vertical dam displacement data. The horizontal displacement monitoring of the dam crest is performed by the extension wire alignment method. The fixed end of the extension wire with a total length of 200.75 m is arranged at Sta. R01+107.025 and the guide end is placed at Sta. L0+93.50. In total, 11 monitoring points are arranged along the dam, of which nine datum points are located at the top of each dam section and two checkpoints are set at the left and right ends of the extension wire to check the displacement of each end. The extension wire system was automated in April 2005 with an observation frequency of 1 time per day. The layout of the extension wire

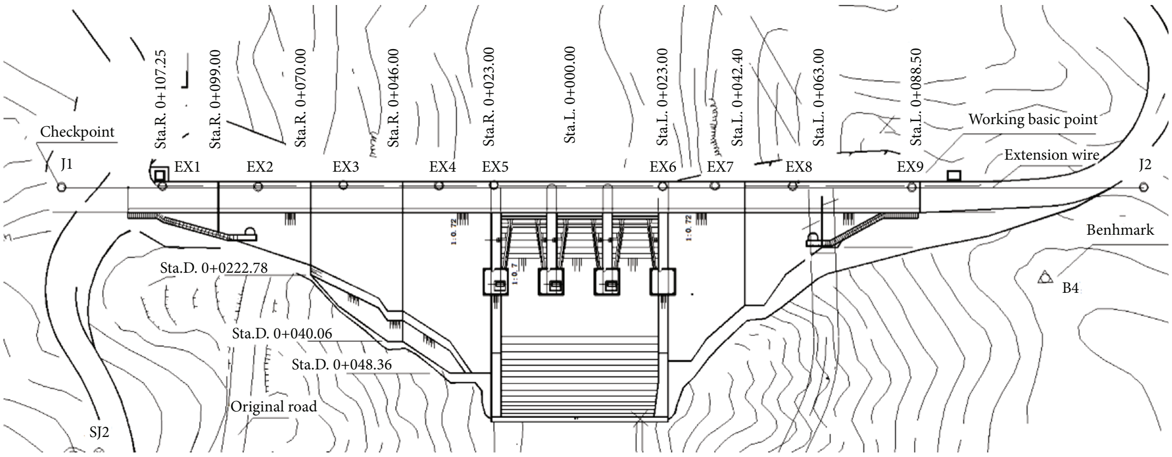


Figure 7: Layout of the extension wire measuring points for horizontal displacement.

measurement points for horizontal displacement is shown in Figure 7.

2.5.2. Selection and Optimization of Influential Factors in the Prediction Model. According to theoretical knowledge, monitoring data, expert experience, etc., the initial selection of empirical influence factors was as follows:

$$\left\{(H-H_0)^1,(H-H_0)^2,(H-H_0)^3,(T_5-T),(T_{20}-T),\right.$$
$$\left.(T_{60}-T),(T_{90}-T),\theta,\ln(1+\theta),J,Q,U\right\}, \tag{22}$$

where H is the water depth on a day when observations are collected, H_0 is the water depth on the base day; T_i is the mean reservoir region temperature i days ago, and T is the annual mean temperature. Additionally, $\theta=(t-t_0)/100$, where t is the observation date and t_0 is the date of the base day. J is the average fracture aperture at measurement points, Q is the seepage flow, and U is the average uplift pressure at measurement points.

The initial empirical influential factors are selected as the conditional attributes X, and the horizontal displacements obtained by the dam crest extension wire (to the left bank is positive and to the right bank is negative) at point EX1 are set as the decision attributes Y in the single-point prediction model. Additionally, the horizontal displacements of EX1, EX2, EX4, EX5, EX6, and EX7 are selected as the decision attributes Y_M in the multipoint prediction model (because the extension wire at EX3 contacted a stainless-steel rod, the monitoring data at the point are not reliable). Overall, 864 monitoring samples of horizontal displacement and influential factors were selected as the sample set U. The attribute range V was determined based on the K-means clustering algorithm with adaptive discretization, and the number of clusters was experimentally determined to be 7. To eliminate irrelevant or weakly informative input variables and keep only the representative factors that affect concrete dam deformation, the RS theory is used to conduct an attribute reduction and importance evaluation and obtain an initial information table $S=\{U,X\cup Y,V,f\}$. The attribute reduction and importance evaluation results for the single-point and multipoint prediction models are shown in Table 1.

According to attribute reduction and importance evaluation results, the influential factors of the single-point prediction model are $\{H\text{-}H_0,(H\text{-}H_0)^2,(H\text{-}H_0)^3,(T_5\text{-}T),\ (T_{20}\text{-}T),\theta,\ln(1+\theta)\}$, and the importance evaluation values for each component of horizontal displacement at EX1 are 0.12, 0.08, 0.13, 0.42, 0.20, 0.02, and 0.03, respectively. The influential factors of the multipoint prediction model are $\{H\text{-}H_0,(H\text{-}H_0)^2,\ (H\text{-}H_0)^3,(T_5\text{-}T),(T_{20}\text{-}T),\theta,\ln(1+\theta),J,Q,U\}$, and the importance evaluation values of each component of the horizontal displacement at EX1, EX2, EX4, EX5, EX6, and EX7 are 0.10, 0.07, 0.06, 0.33, 0.19, 0.00, 0.00, 0.02, 0.04, 0.06, 0.08, and 0.05. Therefore, it can be concluded that the horizontal displacement of the extension wire is greatly affected by temperature changes and water level fluctuations. Specifically, the temperature component accounts for 60% of the horizontal displacement, and the lag period of the water level is approximately 20 days.

2.5.3. Sample Selection for Prediction Models. According to attribute reduction and importance evaluation results, the influential factor monitoring data of the single-point and multipoint prediction models are selected to obtain samples as independent variables, and the horizontal displacement at points EX1-EX7 (except EX3) is selected to obtain samples as dependent variables. The dataset is established between June 2, 2016, and October 22, 2018, and has a total of 864 samples of data. The dataset of 700 samples selected from June 2, 2016, to May 10, 2018, is used as the training set, and the dataset of 164 samples selected from May 11, 2018, to October 22, 2018, is adopted as the testing set. Investigations of the concrete dam deformation prediction model based on the OLS, SVM, MLP, and LSTM methods are performed using the dataset with 864 samples of data. Variations in the water depth and horizontal displacement are shown in Figures 8 and 9.

2.5.4. Model Parameter Setting. The performance of SVM, MLP, and LSTM models depends greatly on the setting of some parameters. According to experience and experiment

Table 1: Attribute reduction results of the single-point and multipoint prediction models.

Experience impact factors	Component name	Single-point model *SIM*	Reduction	Importance evaluation Sig (a, X, Y)	Multipoint model *SIM*	Reduction	Importance evaluation Sig (a, X, Y)
H-H_0	Water pressure	-5	No	0.12	-4	No	0.10
$(H$-$H_0)^2$		-2	No	0.08	-2	No	0.07
$(H$-$H_0)^3$		-2	No	0.13	-4	No	0.06
$(T_5$-$T)$	Temperature	-5	No	0.42	-7	No	0.33
$(T_{20}$-$T)$		-4	No	0.20	-2	No	0.19
$(T_{60}$-$T)$		0	Yes	0.00	0	Yes	0.00
$(T_{90}$-$T)$		0	Yes	0.00	0	Yes	0.00
θ	Aging	-2	No	0.02	-2	No	0.02
$\ln(1+\theta)$		-1	No	0.03	-3	No	0.04
J	Fracture	-1	Yes	0.00	-2	No	0.06
Q	Seepage	-2	Yes	0.00	-4	No	0.08
U	Uplift pressure	-3	Yes	0.00	-3	No	0.05

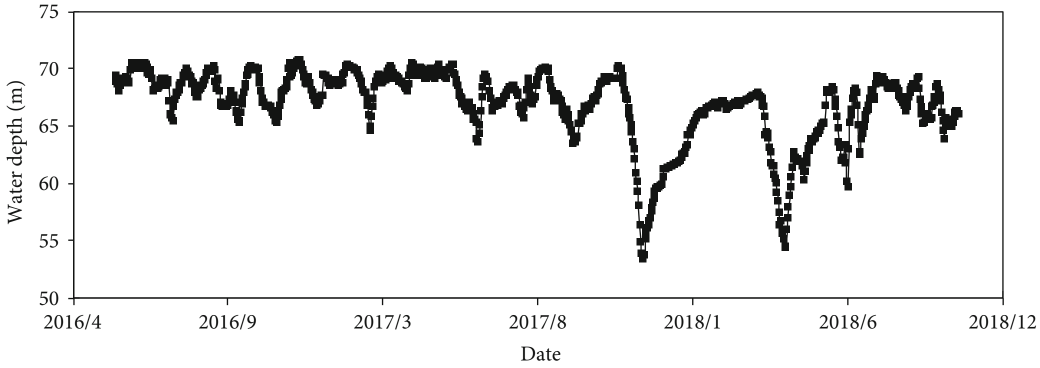


Figure 8: Variations in the water depth.

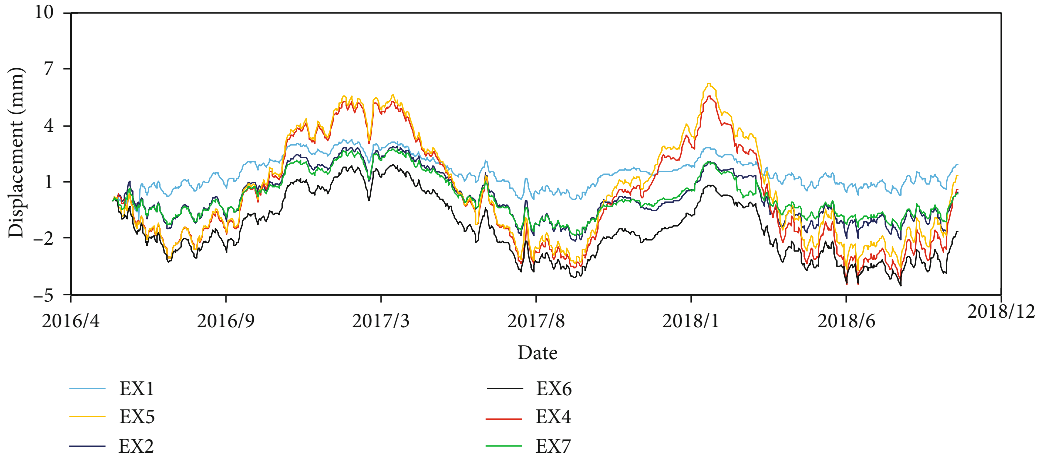


Figure 9: Variations in horizontal displacement.

results, parameters of the adapted algorithms, namely, regularization parameters, kernel function parameters, network parameters, learning rates, and so on, are given before the simulation.

Parameters in the SVM model: the kernel function is determined as a radial basis function (RBF) according to experience. Parameter range of the SVM model is determined based on experience, penalty parameter $C \in [-256,$

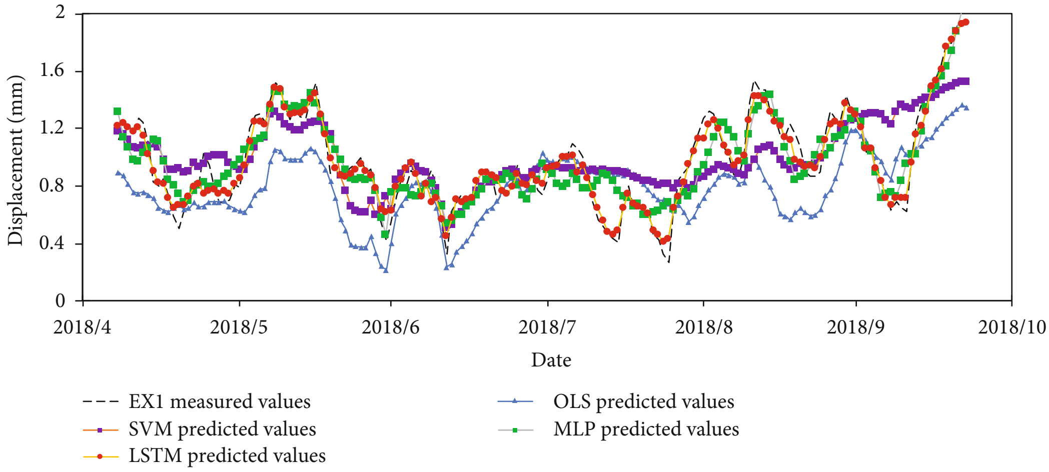


Figure 10: Measured and model-predicted values.

256], kernel parameter $\gamma \in [-256, 256]$. Parametric tuning is implemented with Grid Search, C is set to 8, and γ is set to 0.72 according to the experimental relationship between the objective functions and parameters.

Parameters in the MLP model: according to experiment results, the network is composed of input layer, hidden layers, and output layer with the three-layer topology of 7-15-15-1 for single-point prediction model and 10-15-15-1 for multipoint prediction model, and the learning rate is also set to 0.08. The network variable parameter weights and biases are initialized randomly and calculated by gradient descent algorithms, and the activation function is the ReLU function.

Parameters in the LSTM model: hyperparameter range of the LSTM model is determined based on experience, batch size $\in [0, 1000]$, timestep $\in [20, 300]$, hidden layers $\in$ [20,200], and the initial value of learning rate is set to 0.1. Taking the minimum RMSE value as the objective function, and according to the experimental relationship between the objective function and parameters, parametric tuning is implemented with Grid Search. Finally, batch size, timestep, hidden layers, and learning rate are set to 12, 46, 42, and 0.12, respectively. The network variable parameters including weights and biases are generated using the glorot_uniform initializer and calculated by gradient descent algorithms.

3. Results

To verify the superiority of the concrete dam deformation prediction model based on LSTM compared to other models in terms of accuracy, robustness, externality, and generalization, a comparative analysis of the OLS, SVM, MLP, and LSTM models is conducted based on the prediction results with preprocessed training and testing samples.

3.1. Single-Point Prediction Model. Single-point prediction models for concrete dam deformation based on the OLS, SVM, MLP, and LSTM algorithms are established to facilitate comparative analysis.

3.1.1. Model Prediction Analysis. Concrete dam deformation prediction models based on the OLS, SVM, MLP, and LSTM algorithms were established based on the preprocessed standardized environmental dataset and the unnormalized deformation dataset. Based on the objective functions, the training samples are used to train the models, and the optimal model parameters are obtained. Finally, a concrete dam deformation prediction and performance analysis are performed. The measured and predicted values of concrete dam deformation based on the OLS, SVM, MLP, and LSTM models are shown in Figure 10.

Figure 10 shows that the predicted values of the OLS model largely deviate from the measured values, but the overall trend is similar to that for the measured values. The deviation between the predicted values of the SVM, MLP, and LSTM models and the measured values is small, but the late-stage prediction trend of the SVM model deviates significantly from the measured values. The LSTM model not only exhibits the highest degree of agreement between the predicted and measured values but also yields the same trend as that for the measured values. Therefore, the prediction performance of the concrete dam deformation prediction model based on LSTM is significantly better than that based on the OLS, SVM, and MLP models.

3.1.2. Model Performance Evaluation. A prediction model for concrete dam deformation is an important tool for quantitatively evaluating the safety status of dams, revealing abnormalities in the service status and ensuring engineering safety. A high-performance deformation prediction model should meet the relevant accuracy, robustness, externality, and generalization requirements to implement effective early warning strategies and feedback control for engineering safety. To verify the reliability of the concrete dam deformation prediction model based on LSTM, concrete dam deformation prediction models based on OLS, SVM, MLP, and LSTM were compared and analyzed according to the evaluation system established in this paper. The horizontal displacement residuals of each prediction model are shown in Figure 11. Evaluation index values of all prediction models are shown in Table 2.

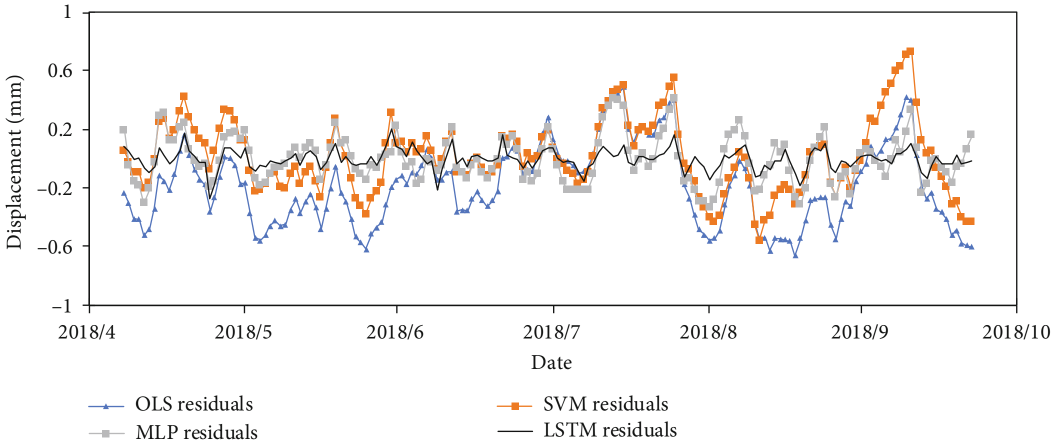


Figure 11: The horizontal displacement residuals of each prediction model.

Table 2: Evaluation index values of all prediction models.

	Prediction model			
	OLS	SVM	MLP	LSTM
AVI				
RMSE_P	0.32833	0.2466	0.1604	0.0690
MAPE_P	29.7077	24.7789	16.3695	6.3213
REI	0.3253	0.3672	0.2865	0.1760
EEI	0.4032	0.2739	0.1892	0.1023
GEI	0.8266	0.8788	0.9336	0.9610

(1) Accuracy. Figure 11 and AVI value of each model in Table 2 show that the horizontal displacement residuals of the concrete dam deformation prediction model based on LSTM are the smallest, RMSE_P is lower than 0.1, MAPE_P is lower than 10, and all these are in the low range compared with those of the models based on OLS, SVM, and MLP. Therefore, the concrete dam deformation prediction model based on LSTM displays better accuracy than the other models, and the prediction results better agree with the real data.

(2) Robustness. The REI value of each model in Table 2 shows that the concrete dam prediction models based on OLS, SVM, MLP, and LSTM are all affected by the gross error, resulting in different degrees of prediction accuracy. The REI value of the concrete dam deformation prediction model based on LSTM is the smallest among the REI values of all the models; thus, the gross error associated with the data sample has little impact on the prediction results of the proposed model, which displays the strongest robustness.

(3) Externality. The EEI value of each model in Table 2 shows that the accuracy of the models decreases after adding samples outside the training set to the model testing samples. Nevertheless, the concrete dam deformation prediction model based on LSTM exhibits the smallest EEI, representing the strongest externality and the most powerful learning ability.

(4) Generalization. The GEI value of each model in Table 2 shows that the generalization of the concrete dam deformation prediction models based on OLS and SVM is poor. These models likely experience overfitting during training, resulting in an increase in the error for the testing set and poor performance. The concrete dam deformation prediction models based on MLP and LSTM display good generalization performance, and the LSTM model yields the best performance.

In summary, the successful application of machine learning technology has greatly promoted the development of concrete dam deformation prediction model compared with using traditional statistical methods. The concrete dam deformation prediction models based on SVM, MLP, and LSTM all displayed high accuracy, but the performance of each model in terms of robustness, externality, and generalization varies. The concrete deformation prediction model based on LSTM displays the highest accuracy, robustness, externality, and generalization by comparison with the performance of the other models. Therefore, the application of LSTM to concrete deformation prediction models further promotes the development of concrete dam prediction model.

3.2. Multipoint Synchronized Prediction Model for Concrete Dam Deformation.

According to the theoretical, mathematical, and mechanical principles of concrete dams, the concrete dam deformation is affected not only by loads such as water pressure and temperature loads but also by adjacent local factors. The sudden displacement of some dam parts will influence the surrounding areas, and a single-point prediction model for concrete dam deformation does not consider the relationships among points; therefore, it is difficult to grasp the displacement field under a given load. It is necessary to establish a multipoint synchronized prediction model for concrete dam deformation that can effectively improve the prediction performance compared to that of traditional models and the accuracy of mechanical parameter inversion and feedback analysis. Additionally, such a method could improve the

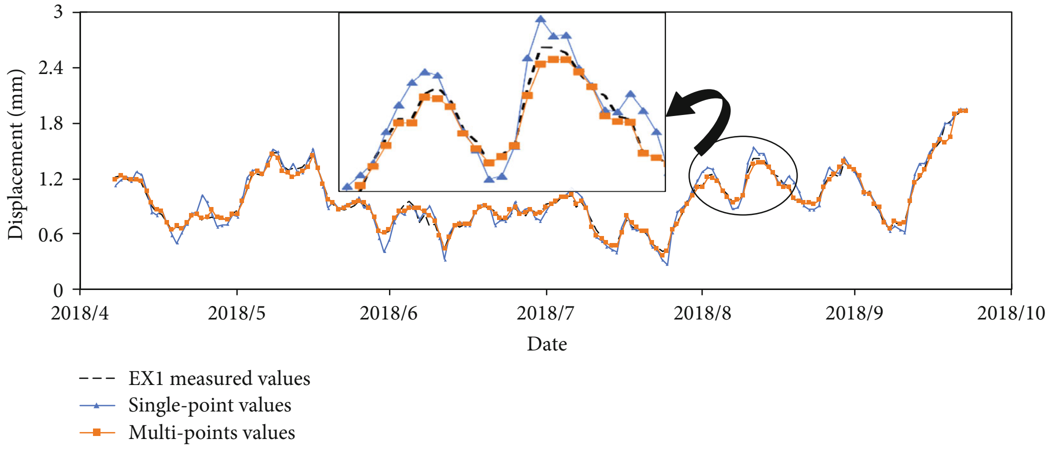


Figure 12: Actual values and predicted values with the single-point and multipoint models.

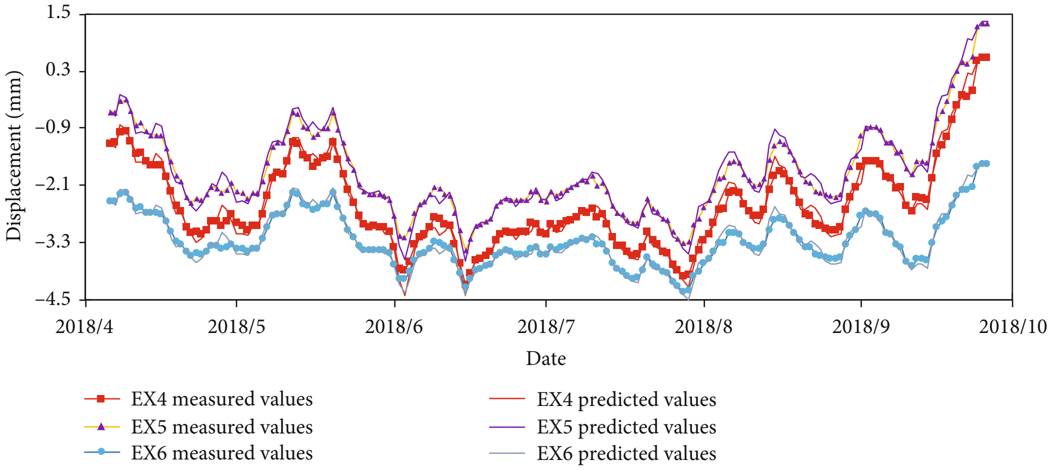


Figure 13: Measured values and values predicted with the multipoint deformation prediction model.

safety monitoring level of concrete dams. This paper establishes a multipoint synchronized prediction model for concrete dam deformation based on the data collected at multiple points and the advantages of LSTM for multiple inputs and outputs.

3.2.1. Model Prediction Analysis. The factors that influence the multipoint synchronized prediction model for concrete dam deformation are determined by attribute reduction. All the data are normalized and used as samples of the independent variables in the LSTM model, and the deformation monitoring data from points EX1-EX7 (except EX3) are selected as samples of the dependent variable (no normalization processing). The training data and testing set are divided in the same way as in the single-point model. The output layer is a multidimensional fully connected layer, the model learning rate is 0.18, and other parameters are the same as those in the single-point model. The six-point synchronized prediction model for concrete dam deformation with optimal parameters is obtained by training, and the deformation values are predicted based on the testing samples. The actual measured values and the predicted values of the single-point and multipoint models are shown in Figure 12 (taking the measured values at EX1 as an example). The measured values and values predicted with the multipoint synchronized deformation model are shown in Figure 13 (taking the measured values at EX4-EX6 as examples).

3.2.2. Model Performance Evaluation. Since the prediction model is based on the deformation values at multiple points, the error of the multipoint model includes the error at all points. According to error theory, RMSE of the multipoint model is the weighted average of RMSE at each point, and it can be expressed as follows:

$$S = \sqrt{\frac{ns_1^2 + ns_2^2 + \cdots + ns_k^2}{kn}} = \sqrt{\sum_{i=1}^{k} \frac{s_i^2}{k}}, \tag{23}$$

where S represents the RMSE of the multipoint model, s_i represents each point, and k represents the number of testing samples.

The RMSE values of the multipoint synchronized prediction model and single-point prediction model are compared and analyzed, as shown in Table 3.

Table 3: RMSE values of the multipoint prediction model and single-point prediction model.

Prediction model	Single-point model of EX1	Multipoint model of EX1	Average of all single-point models	Multipoint value
S	0.0690	0.0575	0.0739	0.0592

Figures 12 and 13 and Table 3 show that the performance of both models is good, and the error is within the acceptable precision range. The RMSE value of the multipoint prediction model is smaller than that of the single-point prediction model, and the predicted values of the multipoint model are closer to the measured values. Additionally, the weighted average RMSE of the single-point prediction model is larger than the RMSE of the multipoint model, which indicates that the prediction accuracy at each point in the multipoint model is high. Therefore, the multipoint synchronized prediction model for concrete dam deformation based on LSTM exhibits good performance, and the analysis results are locally meaningful and spatially representative at large scales.

4. Conclusions and Discussion

RS theory and an LSTM network are introduced for concrete dam safety monitoring in the TensorFlow framework, and single-point and multipoint concrete dam deformation prediction models based on LSTM are established. Moreover, a new evaluation system and quantitative evaluation indexes for the concrete dam deformation prediction model are proposed. The following conclusions were obtained from application examples.

(1) RS theory is applied to optimize the selection and evaluate the importance of the factors that influence concrete dam deformation based on the internal relationships among the monitoring dataset. This approach overcomes the deficiencies of intelligent prediction models related to the quantitative interpretation and ensures the objectivity of prediction model analysis

(2) According to statistical theory, an evaluation system is proposed, and accuracy, robustness, externality, and generalization evaluation indexes are given as performance inspection criteria to comprehensively evaluate the performance of concrete dam deformation prediction models in practical engineering

(3) The single-point prediction model for concrete dam deformation based on LSTM displays high prediction accuracy and strong robustness, externality, and generalization. Moreover, the multipoint synchronized prediction model for concrete dam deformation based on LSTM is locally pertinent and spatially representative at large scales. Thus, the multipoint approach can be effectively used in the deformation prediction of concrete dams at large scales

The continuous improvements in concrete dam technology have resulted in high requirements for prediction model performance, and establishing high-performance spatiotemporal prediction models will be important as concrete dam safety monitoring continues to progress. Therefore, the combination of AI, deep learning theory, online dynamic learning, and space-time deformation prediction models should be promoted to establish an ideal concrete dam monitoring system and achieve the goal of "intelligent monitoring."

Data Availability

"Dataset of Zhouning dam.xlsx" used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 51809212), the Key Projects of Natural Science Basic Research Program of Shaanxi Province (No. 2018JZ5010), and the Water Science Plan Project of Shaanxi Province (No. 2018SLKJ-5).

Supplementary Materials

The supplementary material submitted is the dataset used to build the prediction model based on RS-LSTM in the article. The dataset named "Dataset of Zhouning dam.xlsx" includes values of Zhouning dam's horizontal displacement at points EX1-EX7 (except EX3) and its impact factors from June 2, 2016, to October 22, 2018, which are selected to obtain samples. The dataset of 700 samples selected from June 2, 2016, to May 10, 2018, is used as the training set, and the dataset of 164 samples selected from May 11, 2018, to October 22, 2018, is adopted as the testing set. *(Supplementary Materials)*

References

[1] B. F. Zhu, "On the expected life span of concrete dams and the possibility of endlessly long life of solid concrete dams," *Journal of Hydraulic Engineering*, vol. 43, no. 1, pp. 1–9, 2012.

[2] C. S. Gu, H. Z. Su, and S. W. Wang, "Advances in calculation models and monitoring methods for long-term deformation behaviour of concrete dams," *Journal of Hydroelectric Engineering*, vol. 35, no. 5, pp. 1–14, 2016.

[3] X. Q. Li, D. J. Zheng, and Y. P. Ju, "Input factor optimization study of dam seepage statistical model based on copula entropy theory," *Journal of Hohai University (Natural Sciences)*, vol. 44, no. 4, pp. 370–376, 2016.

[4] J. Yang, D. X. Hu, and Z. R. Wu, "Multiple co-linearity and uncertainty of factors in dam safety monitoring model," *Journal of Hydraulic Engineering*, vol. 35, no. 12, pp. 99–105, 2004.

[5] F. Kang, L. S. Zhao, and Y. Wang, “Structural health monitoring of concrete dams using long-term air temperature for thermal effect simulation,” *Engineering Structures*, vol. 180, pp. 642–653, 2019.

[6] F. Q. Li, *Research on Data Analysis Method of Dam Safety Monitoring*, Zhejiang University, 2012.

[7] Y. Zhao, X. H. Hua, and M. Li, “Application of stepwise regression model to dam radial displacement monitoring,” *Journal of Geomatics*, vol. 21, no. 1, pp. 1–5, 2012.

[8] L. Pei, Z. Y. Wu, M. Cui, Q. Zhang, and J. K. Chen, “Research and application on the displacement hybrid-model of high earth dam,” *Journal of Sichuan University*, vol. S1, pp. 7–12, 2012.

[9] B. Li, J. Li, K. Jiang, and L. H. Pi, “Spatial-temporal model of monitoring the displacement of roller compacted concrete dam,” *Journal of Yangtze River Scientific Research Institute*, vol. 30, no. 1, pp. 90–92, 2013.

[10] D. Y. Li, Y. C. Zhou, and X. Q. Gan, “Research on deterministic displacement monitoring model for multiple survey points of concrete arch dam,” *Journal of Hydraulic Engineering*, vol. 42, no. 8, pp. 981–985, 2011.

[11] C. S. Gu and Z. R. Wu, *Safety Monitoring of Dams and Dam Foundations -Theory, Methods and Application*, Hohai University Press, Nanjing, China, 1st edition, 2006.

[12] F. Kang, J. J. Li, and J. H. Dai, “Prediction of long-term temperature effect in structural health monitoring of concrete dams using support vector machines with Jaya optimizer and salp swarm algorithms,” *Advances in Engineering Software*, vol. 131, pp. 60–76, 2019.

[13] D. Y. Li and Y. C. Zhou, “Application of BP network to multiple-spot model of dam deformation monitoring,” *Journal of Yangtze River Scientific Research Institute*, vol. 25, no. 6, pp. 52–55, 2005.

[14] R. X. Chen and L. Cheng, “The monitoring model of multiple monitoring points with latent variables based on LS-SVM,” *Water Resources and Power*, vol. 10, pp. 80–82, 2012.

[15] F. Kang, J. Liu, J. Li, and S. Li, “Concrete dam deformation prediction model for health monitoring based on extreme learning machine,” *Structural Control and Health Monitoring*, vol. 24, no. 10, article e1997, 2017.

[16] B. Dai, C. S. Gu, E. F. Zhao, and X. Qin, “Statistical model optimized random forest regression model for concrete dam deformation monitoring,” *Structural Control and Health Monitoring*, vol. 25, no. 6, article e2170, 2018.

[17] F. Salazar, R. Moran, M. Á. Toledo, and E. Oñate, “Data-based models for the prediction of dam behaviour: a review and some methodological considerations,” *Archives of Computational Methods in Engineering*, vol. 24, no. 1, pp. 1–21, 2017.

[18] D.-A. Tibaduiza, M.-A. Torres-Arredondo, L. E. Mujica, J. Rodellar, and C.-P. Fritzen, “A study of two unsupervised data driven statistical methodologies for detecting and classifying damages in structural health monitoring,” *Mechanical Systems & Signal Processing*, vol. 41, no. 1-2, pp. 467–484, 2013.

[19] Y. Yang, D. Chen, and H. Wang, “Active sample selection based incremental algorithm for attribute reduction with rough sets,” *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 825–838, 2016.

[20] Y. Jing, T. Li, H. Fujita, B. Wang, and N. Cheng, “An incremental attribute reduction method for dynamic data mining,” *Information Sciences*, vol. 465, pp. 202–218, 2018.

[21] C. Yuan, X. Qin, L. Yang, G. Gao, and S. Deng, “A novel function mining algorithm based on attribute reduction and improved gene expression programming,” vol. 7, IEEE Access, 2019.

[22] B. B. Yang, K. L. Yin, and J. Du, “Dynamic prediction model of landslide displacement based on time series and long and short time memory networks,” *Chinese Journal of Rock Mechanics and Engineering*, vol. 37, no. 10, pp. 2334–2343, 2018.

[23] W. T. Tong, L. X. Li, X. L. Zhou, A. Hamilton, and K. Zhang, “Deep learning $\mathrm{PM}_{2.5}$ concentrations with bidirectional LSTM RNN,” *Air Quality, Atmosphere & Health*, vol. 12, no. 4, pp. 411–423, 2019.

[24] T. T. Huang and L. Yu, “Application of SDAE-LSTM model in financial time series prediction,” *Computer Engineering and Applications*, vol. 55, no. 1, pp. 142–148, 2019.

[25] Y. X. Liu, Q. X. Fan, Y. Z. Shang, Q. M. Fan, and Z. W. Liu, “Short-term water level prediction method for hydropower stations based on LSTM neural network,” *Advances in Science and Technology of Water Resources*, vol. 39, no. 2, pp. 56–60+78, 2019.

[26] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: continual prediction with LSTM,” *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[27] Y. Ying and M. Pontil, “Online gradient descent learning algorithms,” *Foundations of Computational Mathematics*, vol. 8, no. 5, pp. 561–596, 2008.

Hindawi
Journal of Sensors
Volume 2019, Article ID 5494901, 12 pages
https://doi.org/10.1155/2019/5494901

Research Article

Optimal Uncalibrated RSS Indoor Positioning and Optimal Reference Node Placement Using Cramér-Rao Lower Bound

Xavier Tolza,[1,2] Pascal Acco ,[1] Jean-Yves Fourniols,[1] Georges Soto-Romero,[1,3] Christophe Escriba,[1] and Manuel Bracq[2]

[1] *LAAS-CNRS, Université de Toulouse, CNRS, INSA, Toulouse, France*
[2] *FFLY4U, 3 Avenue Didier Daurat, Toulouse, France*
[3] *EA4660, C3S Health-Sport Department, Sports University, Besançon, France*

Correspondence should be addressed to Pascal Acco; pacco@laas.fr

Received 11 February 2019; Accepted 28 April 2019; Published 4 June 2019

Academic Editor: Qiang Wu

In this paper we propose a global positioning algorithm of multiple assets based on Received Signal Strength (RSS) measurements that takes into account the gain uncertainties of each hardware transceiver involved in the system, as well as the uncertainties on the Log-Distance Path Loss (LDPL) parameters. Such a statistical model is established and its Maximum Likelihood Estimator (MLE) is given with the analytic expression of the Cramér-Rao Lower Bound (CRLB). Typical values of those uncertainties are given considering whether calibration is done in production, *in situ*, or if hardware is used uncalibrated, in order to know what is the expected accuracy in function of the calibration setup. Results are tested by numerical simulations and confronted to real measurements in different room configurations, showing that the theoretical bound can be reached by the proposed MLE algorithm.

1. Introduction

Assets positioning raised a great interest in the last decade, especially with the Internet of Things (IoT) business. In this context, we expect indoor and outdoor positioning to be achieved with the same hardware, a low energy constraint for an autonomy of a few years. Economical aspect can even constrain each reference node, named Access Point (AP), used to locate the target assets, to run on battery.

Outdoor positioning is mainly achieved using Global Navigation Satellite System (GNNS) with an accuracy of a few meters that is enough for this kind of applications. Even if GNNS receiver energy consumption has been greatly improved in the past decade [1–5], this still mainly limits the asset autonomy to a few years. There is more energy efficient but less accurate work in progress systems where the measurement effort and position estimation are reported from the asset to the infrastructure using mainly a Low Power Wide Area Network (LPWAN) connection [6–10]. Sub-GHz bands allow using the same transceiver for both indoor and outdoor positioning.

Indoor positioning using GNNS requires additional infrastructure because of the poor signal level; the monolithic solutions [11–14] can be very precise but the need of a great number of external antenna does not meet most economical requirements.

This paper focuses on indoor solutions that meet the constraints of low energy (for assets and APs) with minimal infrastructure and setup to reach a precision of a few meters. We particularly focus on the static assets positioning use-case, which implies that some techniques such as pedestrian dead reckoning are not applicable for the solutions we are studying. This area of indoor positioning using LPWANs, WIFI, or Bluetooth Low Energy (BLE) raised a great interest in the last decade and many solutions have been proposed, based on Received Signal Strength Indication (RSSI), Time of Arrival (ToA), and Angle of Arrival (AoA). Reviews on recent advances and capacities can be found [15–17], and theoretical bounds like Cramér-Rao Lower Bound (CRLB) and algorithms have been given and reviewed for most techniques [18–20].

The common framework widely used proceeds in two steps instead of direct estimation to reduce complexity even if this is suboptimal in general [19]:

(i) A given number of measurements related to distance are collected in a short time slot; outliers and noise are filtered and an approximate distance measurement is inferred from those, which are not coherent one with the other due to multipath and measurement noise.

(ii) Those distances are then aggregated using algorithms, which can belong to the following domains:

   (a) Geometry (finding the intersection of distance circles).

   (b) Machine learning (neural networks, Smallest M-vertex Polygon (SMP), or Support Vector Machine (SVM)).

   (c) Cellular algorithms (closest neighbour, weighted neighbors, etc. . . .).

   (d) Statistical estimation algorithms (mainly maximum likelihood).

ToA based solutions can be very accurate (within a meter or centimeter), such as Ultra Wide Band (UWB) or collaborative positioning, but they require expensive hardware and moreover synchronization signals are involved which increases the power consumption. On the other hand such a precision is not needed for asset positioning.

This paper then considers less precise but lower cost solutions like Received Signal Strength (RSS) positioning using low energy protocols like BLE. In this case the accuracy is strongly limited by the fast and slow fading effect arising in dense multipath environments. Fast fading can be mitigated using several measurements in time and their first step preprocessing (averaging and removing outliers from the average, i.e., values that are far from the mean value, or using median value) before running the second step positioning algorithm (which is known as time diversity [21]).

Contrariwise, slow fading effects need spatial diversity or frequency diversity to be reduced. For a given multipath configuration, frequency diversity changes fading effects on the RSS allowing simple algorithms like averaging or maximum selection to improve the measurement at the preprocessing stage of the positioning [22, 23]; this solution is promising because it is only at the energy cost of a few emissions using an agile emitter. Theoretical bounds and finding algorithms should be more developed in the future to tackle the slow fading effect.

Fingerprinting methods include the multipath effect for each specific room in the propagation model to compensate the static part of slow fading effects. Depending on the AP placement, this method can drastically improve the accuracy of positioning compared to physical models based solutions. The main drawback is the learning process energy consumption and hardware cost that should be deployed at the setup for static environments, and moreover the continuous learning process needed to correct slow changes in the environment.

An energy efficient solution to reduce slow fading is to use a sufficient number of APs in sparse places of the room to give redundant measurements, with different multipath biases. This spatial diversity may increase the accuracy of the positioning algorithm when it is able to take advantage of redundant measurements. There is then a trade-off between the hardware cost of a great number of APs and the desired accuracy of positioning. Part of the hardware cost resides in its installation process which is greatly reduced when those receivers are battery powered (so they do not need to be connected to a power source), when no calibration of the hardware gain is needed, and moreover when precise coordinates of AP placement in the room are not needed.

Calibration of APs gains can be done in the fabrication process or at setup time which takes into account the antenna coupling with the environment, or calibration can be run jointly to positioning as an unsupervised learning process. Many works deal with the calibration of radio map with pure machine learning [24–27], where others introduce path-loss model into the learning process [17, 24, 28]. It is difficult to find how much radio map learning can improve the positioning performances and what will be quantitatively the positioning error after the learning phase.

To help the design of a positioning system, the authors of the previous work proposed theoretical bounds of positioning error with uncalibrated propagation model that are reachable in an analytic way [29]. This bound has already been derived for RSS, ToA, and Difference Time of Arrival (DToA) techniques but not in the specific case of uncalibrated Aps [19]. In the case of RSS phenomena like measurements quantification are taken into account but lead to numerical expression of the CRLB [30]. The CRLB is generally used in statistics to bound the error variance of a given estimation problem using the FIM derived from the likelihood function of an estimate. This bound cannot be used to find an estimation algorithm but gives a limit to the performance of any estimator that can be proposed to solve this specific estimation problem. Moreover this bound can sometimes be optimistic when there is no existing algorithm capable of reaching the performances of the bound.

We first consider in this paper the whole estimation of multiple assets positions, as it may improve performances compared to the independent estimation of each single target in presence of common APs miss-calibrated parameters. Then full equations are given and compared to previous paper single asset positioning [29]. The Maximum Likelihood Estimator (MLE) is also confronted to real experimentation and measurement. Comparing error performances to CRLB gives the efficiency of each solution so one can judge the loss of precision in regard to the algorithm complexity and energy efficiency. Moreover the analytic expression of the bound helps to design, without any experimentation or simulation, the best APs coverage density and calibration efforts to be run with a given precision objective.

In the next section, we first give the stochastic propagation model and state the positioning problem as an uncertain static parameter estimation. Then analytic CRLB and the

MLE reaching this bound are derived. Section 3 aims to verify previous results using simulations: the analytic bound is compared with variance and RMSE of numerically computed estimates.

2. Optimal RSS Positioning Algorithm with Uncertain Calibration Parameters

We consider the positioning problem of I target assets which sends N signals to J APs placed in a given room. The positioning algorithm takes part of a propagation model which contains uncertain parameters model for each asset and APs.

It must be noted that this estimation problem is not a joint estimation of the model parameters and the position at the same time (like it is often done when dealing with fingerprinting) but rather a position estimator which takes into account the residual uncertainties of model parameters from the calibration process to estimate a position.

The propagation model and estimation problem are stated in the next section; the likelihood and maximum likelihood estimator is derived in Section 2.2. The analytic CRLB is given in Section 2.3, and finally the impact of asset gain, APs gain, and reference model gain are shown using this analytic formula at the end of Section 2.3.

2.1. Problem Statement and Uncertain Model. We consider the position estimation problem to be a static parameter estimation: knowing the RSSI channel model, the uncertainties on the parameters model, we want to find $\widehat{\theta}$, the best estimate of position θ, inside a given room from some measurements $r(\theta)$ by maximizing the likelihood function $\mathscr{L}(\theta \mid r(\theta))$:

$$\widehat{\theta} = \arg\max_{\theta} \left(\mathscr{L}\left(\theta \mid r\left(\theta\right)\right)\right) \tag{1}$$

The use of probabilistic algorithms requires a propagation model; the Log-Distance Path Loss (LDPL) model is a common choice for indoor (NLOS) propagation [31], which defines the received power in relation with the distance as it follows:

$$P_R\left(d\right) = G + a_0 - 10\gamma \log_{10}\left(d\right) + R + \mathscr{V}\left(\sigma\right) \tag{2}$$

where a_0 is the loss at a reference distance of 1 m, γ is commonly known as path loss exponent, path loss factor, or path loss gradient, R (G, resp.) is the loss due to receiver's antenna (transmitter, resp.), and $\mathscr{V}(\sigma)$ is the measuring noise modeled as a random Gaussian variable.

Fast fading phenomenon is removed by the preprocessing step, usually by removing outliers using median or Kalman filtering. NLOS situation is taken into account by the log-normal probability of shadowing $\mathscr{V}(\sigma)$ and does not worth to be modeled separately from measurement noise as for time delay techniques.

Path Loss parameters are known with relative precision depending on the calibration setup when it exists; its value differs from the theoretical value of 2 (valid in the case of isotropic propagation in vacuum). Density of obstacles and antenna directivity change this value which can be different from an AP to another. However, it is possible to find average values for those parameters working fine in most case and improving them with the collected RSSI over time [27, 31] and measure a mean value and a variance for G and R.

If we do not calibrate G (R, a_0, and γ, resp.), then we can consider it as a random variable normally distributed around an average value $\overline{G}$ with a variance σ_G. Similarly, we can model all those parameters by a mean value and a variance expressing the residual uncertainty of the measure. We want to provide some typical values for three different calibration modes: production calibration where the static gains are measured prior to installation, *in situ* calibration where the gains are measured with the device fixed at its final position (more complexes and being costly for industrials, but it takes into account the possible change in gain because antenna coupling with the material the device is attached on), and finally no calibration where we only know the general Probability Density Function (PDF) of the gains from mean and variance measurements. We conducted experiments in three different rooms (confined, semiconfined, and open environment) to get typical values for our hardware: we measured RSSIs at known distance from multiple APs with several target assets and measured the mean and standard deviation of each parameter of the model using a Root Minimum Square (RMS) optimization. We also measured the static gain in an anechoic chamber using an USRP B200 from Ettus Research, measuring as well the variance of our measurement; results are compiled in Table 1. It must be noted that mean gains ($\overline{G}$ and $\overline{R}$) are set to zero for *in situ* calibration because their mean value is reported to be $\overline{a_0}$ as we do not have a reference measuring tool such as in production calibration.

We propose studying the influence of those variances over the accuracy and precision of the likelihood estimate using CRLB, for instance, to know which accuracy we could expect by skipping or not the calibration of the receivers or the APs.

Let us consider that I assets send N signals to J APs in the same room. The APs have known positions and are placed at coordinates $\theta_j^{ap} = {}^T[x_j \;\; y_j \;\; z_j]$ if we are interested in 3D coordinates or $\theta_j^{ap} = {}^T[x_j \;\; y_j]$ when estimating 2D coordinates with $j \in [0, J-1]$, and the assets have unknown coordinates named $\theta_i^t = [x_i \;\; y_i \;\; z_i]$ or $\theta_i^t = {}^T[x_i \;\; y_i]$, $i \in [0, I-1]$, respectively, for 3D and 2D estimation.

The path loss model (2) applied to the $n \in \mathscr{N} = \{n,\ 0 \leq n < N\}$ signal sent from the ith asset and received at the jth AP gives a strength measurement (in dB) r_{ijn} and its expectation $\overline{r_{ijn}}$ is expressed by

$$\begin{aligned} \overline{r_{ijn}} &= \overline{g_j} + \overline{a_0} + \overline{r_i} - \overline{\gamma_j}\Delta_{ij}\left(\theta_i^t, \theta_j^{ap}\right) \\ r_{ijn} &= \overline{r_{ijn}} + \mathscr{V}_{g_j} + \mathscr{V}_{a_0} + \mathscr{V}_{r_i} + \Delta_{ij}\mathscr{V}_{\gamma_j} + \mathscr{V}_m^{ijn} \end{aligned} \tag{3}$$

with $\Delta_{ij}(\theta_i^t, \theta_j^{ap}) = 5\log_{10} g(d_{ij}(\theta_i^t, \theta_j^{ap})^2)$ being the log–distances and $d_{ij}(\theta_i^t, \theta_j^{ap})^2 = {}^T(\theta_i^t - \theta_j^{ap})(\theta_i^t - \theta_j^{ap})$ being the distance between the target i and the AP number j.

It is shown in the Appendix that the vector r of all $I \times J \times N$ measurement can be modeled as the statistical expectation

Table 1: Typical uncertain parameters models when production calibration is run, or when calibration is done at the setup process *in situ*, or for the uncalibrated scenario. The $\overline{x}$ notation is the value obtained from calibration where $\mathcal{N}(\nu,\ \sigma^2)$ is the normal law of mean ν and variance σ^2.

Parameters	Production calibration	*In situ* calibration	Not calibrated
$G(dB)$: AP gain mismatch	$\mathcal{N}\left(\overline{G},\ 0.1^2\right)$	$\mathcal{N}(0,\ 1)$	$\mathcal{N}\left(0,\ 2^2\right)$
$R(dB)$: asset gain mismatch	$\mathcal{N}\left(\overline{R},\ 0.1^2\right)$	Non Applicable	$\mathcal{N}\left(0,\ 2^2\right)$
γ: room path loss exponent	Nonrelevant	$\mathcal{N}\left(1.4,\ 0.1^2\right)$	$\mathcal{N}\left(1.4,\ 0.4^2\right)$
$a_0(dB)$: reference gain	$\mathcal{N}\left(52,\ 2.3^2\right)$		
$\nu(dB)$: noise measurement and shadow probability	$\mathcal{N}\left(0,\ 6^2\right)$		

of the measurement vector $\overline{r}$ given all the target positions θ^t added to multivariate Gaussian vectors $\mathcal{W}_\gamma(\theta^t)$ and $\mathcal{W}_L$. The final expression (A.7) is recalled here:

$$\overline{r}\left(\theta^t\right) = M_\Gamma\left(\theta^t\right)\cdot\overline{\Gamma} \tag{4}$$

$$r\left(\theta^t\right) = \overline{r}\left(\theta^t\right) + \mathcal{W} \tag{5}$$

with

$$\mathcal{W} \sim \mathcal{N}\left(\mathbb{O}_{IJN}, \Sigma_{\mathcal{W}}\left(\theta^t\right) = \Sigma_{\mathcal{W}_\gamma}\left(\theta^t\right) + \Sigma_{\mathcal{W}_L}\right) \tag{6}$$

$$\begin{aligned}\Sigma_{\mathcal{W}}\left(\theta^t\right) = &\left(\sigma_\gamma{}^2 \underset{IJ}{\operatorname{diag}}\left(\Delta_{ij}\left(\theta_i^t,\theta_j^{ap}\right)^2\right) + \sigma_{a_0}{}^2 \mathbb{1}_{IJ}^{\square}\right.\\ &\left. + \sigma_G{}^2 \mathbb{1}_I^{\square} \otimes \mathbb{I}_J + \sigma_R{}^2 \mathbb{I}_I \otimes \mathbb{1}_J^{\square}\right) \otimes \mathbb{1}_N^{\square}\end{aligned} \tag{7}$$

$$M_\Gamma\left(\theta^t\right) = \left[\Delta_{diag}\left(\theta^t\right)\ \ \mathbb{1}_{IJ}\ \ \mathbb{1}_I \otimes \mathbb{I}_J\ \ \mathbb{I}_I \otimes \mathbb{1}_J\right] \otimes \mathbb{1}_N \tag{8}$$

Thus the measurements are affected linearly by a multivariate covariant random uncertain vector, so the likelihood and MLE can be obtained.

2.2. Log-Likelihood Expression. The positioning problem is to find all the assets coordinates components θ_k^t of the vector θ^t with $k \in \mathcal{K} = \{k, 0 \le k < K\}$, $K = 3I$, for 3D positioning and $K = 2I$ for 2D where $\theta^t = [\theta_0 \cdots \theta_K]$. The likelihood of a measurement vector r is simply the probability density of the multivariate random vector $\mathcal{W}$ to equal $r - \overline{r}$. Then the probability density function of such a vector follows a multivariate normal law and can be expressed in matrix form [32], which gives the following expressions for likelihood ($\mathcal{L}$) and log–likelihood ($\mathcal{L}\mathcal{L}$):

$$\begin{aligned}&\mathcal{L}\left(r \mid \theta^t\right)\\ &\quad = \frac{1}{\sqrt{2\pi}^{IJN}\sqrt{\left|\Sigma_{\mathcal{W}}\left(\theta^t\right)\right|}} e^{-(1/2)\,{}^{T}(r-\overline{r}(\theta^t))\Sigma_{\mathcal{W}}(\theta^t)^{-1}(r-\overline{r}(\theta^t))}\\ &\quad - 2\mathcal{L}\mathcal{L}\left(r \mid \theta^t\right)\\ &\quad = IJN\ln(2\pi) + \ln\left|\Sigma_{\mathcal{W}}\left(\theta^t\right)\right|\\ &\quad + {}^{T}\left(r-\overline{r}\left(\theta^t\right)\right)\Sigma_{\mathcal{W}}\left(\theta^t\right)^{-1}\left(r-\overline{r}\left(\theta^t\right)\right)\end{aligned} \tag{9}$$

Then the maximum likelihood to be solved is

$$\widehat{\theta^t} = \underset{\theta^t\in\mathbb{R}^K}{\arg\max}\,\mathcal{L}\left(r \mid \theta^t\right) = \underset{\theta^t\in\mathbb{R}^K}{\arg\max}\,\mathcal{L}\mathcal{L}\left(r \mid \theta^t\right) \tag{10}$$

Analytic solution to this optimization problem seems complicated as long as uncertainties on γ_j (involved in the covariance matrix $\Sigma_{\mathcal{W}}(\theta^t)$) are to be taken into account: this involves a covariance matrix which depends on the unknown optimization. This is the scope of future work because the part of $\Sigma_{\mathcal{W}}$ that depends on θ^t is the diagonal matrix $\Sigma_{\mathcal{W}_\gamma}$ in an additive way and further calculations may be solved analytically in the future.

In this paper we consider that all path-loss exponents are certain, or sufficiently calibrated, with known values $\overline{\gamma}$. Then only miss-calibrated gains on assets and APs are considered uncertain as follows. Then $\Sigma_{\mathcal{W}}$ does no longer depend on γ, which simplifies (9) and (15) to

$$\begin{aligned}\mathcal{L}\mathcal{L}\left(r \mid \theta^t\right) = &-\frac{IJN}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\Sigma_{\mathcal{W}_L}\right|\\ &-\frac{1}{2}\,{}^{T}\left(r-\overline{r}\left(\theta^t\right)\right)\Sigma_{\mathcal{W}_L}{}^{-1}\left(r-\overline{r}\left(\theta^t\right)\right)\\ \frac{\partial\mathcal{L}\mathcal{L}\left(r \mid \theta^t\right)}{\partial\theta_k^t} = &-{}^{T}\left.\frac{\partial\overline{r}}{\partial\theta_k^t}\right|_{\theta^t}\Sigma_{\mathcal{W}_L}{}^{-1}\left(r-\overline{r}\left(\theta^t\right)\right)\end{aligned} \tag{11}$$

The positioning problem becomes the following nonlinear least square formulation:

$$\begin{aligned}\widehat{\theta^t} &= \underset{\theta^t\in\mathbb{R}^K}{\arg\max}\,\mathcal{L}\mathcal{L}\left(r \mid \theta^t\right)\\ &= \underset{\theta^t\in\mathbb{R}^K}{\arg\min}\left[{}^{T}\left(r-\overline{r}\left(\theta^t\right)\right)\Sigma_{\mathcal{W}_L}{}^{-1}\left(r-\overline{r}\left(\theta^t\right)\right)\right]\end{aligned} \tag{12}$$

which is efficiently solved in an iterative way [33]:

$$\begin{aligned}\widehat{\theta^t}(k+1) = \widehat{\theta^t}(k) &+ \left({}^{T}\left.\frac{\partial\overline{r}}{\partial\theta^t}\right|_{\widehat{\theta^t}(k)}\Sigma_{\mathcal{W}_L}{}^{-1}\left.\frac{\partial\overline{r}}{\partial\theta^t}\right|_{\widehat{\theta^t}(k)}\right)^{-1}\\ &\cdot {}^{T}\left.\frac{\partial\overline{r}}{\partial\theta^t}\right|_{\widehat{\theta^t}(k)}\Sigma_{\mathcal{W}_L}{}^{-1}\left(r-\overline{r}\left(\widehat{\theta^t}(k)\right)\right)\end{aligned} \tag{13}$$

2.3. Analytic Cramer-Rao Lower Bound. To compute the theoretical bound, the FIM matrix should be derived using

$$\mathcal{I}\left(\theta^t\right) = \mathrm{E}\left[{}^{T}\left(\frac{\partial\mathcal{L}\mathcal{L}\left(r \mid \theta^t\right)}{\partial\theta^t}\right)\frac{\partial\mathcal{L}\mathcal{L}\left(r \mid \theta^t\right)}{\partial\theta^t}\right] \tag{14}$$

Once again, the log-likelihood derivative with respect to θ^t does not permit analytic expression of the bound because of $\Sigma_{\mathscr{W}}$ depending on γ:

$$\frac{\partial\mathscr{L}\mathscr{L}(r\mid\theta^t)}{\partial\theta^t} = \left(\frac{\partial\mathscr{L}\mathscr{L}(r\mid\theta^t)}{\partial\theta^t_k}\right)_{k\in\mathscr{K}}$$
$$\text{with } \frac{\partial\mathscr{L}\mathscr{L}(r\mid\theta^t)}{\partial\theta^t_k} = -\frac{\partial^{T}(r-\bar{r})}{\partial\theta^t_k}\bigg|_{\theta^t}\Sigma_{\mathscr{W}}^{-1}(\theta^t)\left(r-\bar{r}(\theta^t)\right) - \frac{1}{2}\mathrm{Tr}\left(\Sigma_{\mathscr{W}}^{-1}\right)\cdot\frac{\partial\Sigma_{\mathscr{W}_\gamma}}{\partial\theta^t_k}\bigg|_{\theta^t} + \frac{1}{2}\,{}^{T}(r-\bar{r})\Sigma_{\mathscr{W}}^{-1}(\theta^t)\,\frac{\partial\Sigma_{\mathscr{W}_\gamma}}{\partial\theta^t_k}\bigg|_{\theta^t}\Sigma_{\mathscr{W}}^{-1}(\theta^t)\,(r-\bar{r}) \tag{15}$$

In this case the FIM can only be obtained in a numerical way. Taking the assumption of a well-calibrated path-loss exponents (14) and (11) gives

$$\begin{aligned}\mathscr{I}(\theta^t) &= \mathrm{E}\left[{}^{T}\frac{\partial\bar{r}}{\partial\theta^t}\bigg|_{\theta^t}\Sigma_{\mathscr{W}_L}^{-1}\left(r-\bar{r}(\theta^t)\right)\right.\\ &\quad\left.\cdot\,{}^{T}\left(r-\bar{r}(\theta^t)\right)\Sigma_{\mathscr{W}_L}^{-1}\frac{\partial\bar{r}}{\partial\theta^t}\bigg|_{\theta^t}\bigg|\right] = {}^{T}\frac{\partial\bar{r}}{\partial\theta^t}\bigg|_{\theta^t}\\ &\quad\cdot\Sigma_{\mathscr{W}_L}^{-1}E\left[\mathscr{W}_L\,{}^{T}\mathscr{W}_L\right]\Sigma_{\mathscr{W}_L}^{-1}\frac{\partial\bar{r}}{\partial\theta^t}\bigg|_{\theta^t} = {}^{T}\frac{\partial\bar{r}}{\partial\theta^t}\bigg|_{\theta^t}\\ &\quad\cdot\Sigma_{\mathscr{W}_L}^{-1}\frac{\partial\bar{r}}{\partial\theta^t}\bigg|_{\theta^t}\end{aligned}\tag{16}$$

Then the CRLB can be used to obtain the inequality

$$\mathrm{Var}\left(\widehat{\theta^t}\right) \geq \mathscr{I}^{-1} = {}^{T}\frac{\partial\bar{r}}{\partial\theta^t}\Sigma_{\mathscr{W}_L}^{-1}\frac{\partial\bar{r}}{\partial\theta^t}^{-1} \tag{17}$$

Then as $\partial\bar{r}/\partial\theta^t$ is overdetermined compared to the dimension of $\mathscr{W}_L$ we can use the Moore-Penrose Pseudo Inverse (MPPI) [34]:

$$\mathscr{I}^{-1} = \frac{\partial\bar{r}}{\partial\theta^t}^{+}\left(\Sigma_{\mathscr{W}_L}^{-1}\right)^{+}\,{}^{T}\frac{\partial\bar{r}}{\partial\theta^t}^{+} = \frac{\partial\bar{r}}{\partial\theta^t}^{+}\Sigma_{\mathscr{W}_L}\,{}^{T}\frac{\partial\bar{r}}{\partial\theta^t}^{+} \tag{18}$$

where A^+ and $(A)^+$ stand for the MPPI of A.

As the measurements sensibility is independent from an asset to another and the expectation of measurements is independent from a measure to another, the derivative $\partial\bar{r}/\partial\theta^t$ is a block diagonal repetition of the form $\mathrm{diag}_{i\in\mathscr{I}}(\partial\overline{r_{in}}/\partial\theta^t_i\otimes\mathbb{1}_N)$ where $\overline{r_{in}} = (\overline{r_{ij}})_{j\in\mathscr{J}}$ is the vector of all expected measurement for the i$^{\text{th}}$ asset. Then pseudoinverse is expressed as a vertical vector as

$$\begin{aligned}\frac{\partial\bar{r}}{\partial\theta^t}^{+} &= \operatorname*{diag}_{i\in\mathscr{I}}\left(\left(N\,{}^{T}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}\right)^{-1}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}\right)_{i\in\mathscr{I}}\otimes\mathbb{1}_N\\ &= \frac{1}{N}\operatorname*{diag}_{i\in\mathscr{I}}\left(\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\right)\otimes\mathbb{1}_N\end{aligned}\tag{19}$$

Then using the covariance matrix structure (A.5) of $\mathscr{W}_L$ with expression (18) the inverse FIM is simplified in

$$\begin{aligned}\mathscr{I}^{-1} &= \frac{1}{N^2}\left(\operatorname*{diag}_{i\in\mathscr{I}}\left(\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\right)\otimes\mathbb{1}_N\right)\cdot\Big(\Big(\sigma_{a_0}^{\;2}\mathbb{1}^{\square}_{IJ}\\ &\quad+\sigma_G^{\;2}\mathbb{1}^{\square}_{I}\otimes\mathbb{I}_J+\sigma_R^{\;2}\mathbb{I}_I\otimes\mathbb{1}^{\square}_{J}\Big)\otimes\mathbb{1}^{\square}_{N}+\sigma_m^{\;2}\mathbb{I}_{IJN}\Big)\\ &\quad\cdot\left(\operatorname*{diag}_{i\in\mathscr{I}}\left({}^{T}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\right)\otimes{}^{T}\mathbb{1}_N\right)\\ &= \operatorname*{diag}_{i\in\mathscr{I}}\left(\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\left(\left(\sigma_G^{\;2}+\frac{\sigma_m^2}{N}\right)\mathbb{I}_J\right.\right.\\ &\quad\left.\left.+\left(\sigma_{a_0}^{\;2}+\sigma_R^{\;2}\right)\mathbb{1}^{\square}_{J}\right){}^{T}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\right) = \operatorname*{diag}_{i\in\mathscr{I}}\left(\left(\sigma_G^{\;2}\right.\right.\\ &\quad\left.+\frac{\sigma_m^2}{N}\right)\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\,{}^{T}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}+\left(\sigma_{a_0}^{\;2}+\sigma_R^{\;2}\right)\\ &\quad\left.\cdot\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\mathbb{1}^{\square}_{J}\,{}^{T}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\right)\end{aligned}\tag{20}$$

The inverse FIM shows that each asset positioning accuracy is independent one from the other, which shows that estimating positions of I multiple assets is equivalent to making I different estimations of one asset. Hence, information measured by each asset does not improve positioning of the others, and results from previous work [29] can also be used for multiple assets without loss of optimality. Then the CRLB of the i$^{\text{th}}$ target positioning can bound the RMSE of its estimate error ϵ_i with

$$\begin{aligned}\mathrm{E}\left[{}^{T}\epsilon_i\epsilon_i\right] &\geq \mathrm{Tr}\left(\mathrm{Var}\left(\epsilon_i\right)\right) \geq \mathrm{Tr}\left(\mathscr{I}^{-1}\right)\\ &\geq \left(\sigma_G^{\;2}+\frac{\sigma_m^2}{N}\right)\mathrm{Tr}\left(\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\,{}^{T}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\right)\\ &\quad+\left(\sigma_{a_0}^{\;2}+\sigma_R^{\;2}\right)\mathrm{Tr}\left(\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\mathbb{1}^{\square}_{J}\,{}^{T}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\right)\\ &= \left(\sigma_G^{\;2}+\frac{\sigma_m^2}{N}\right)\mathrm{Tr}\left(\left({}^{T}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}\right)^{-1}\right)\\ &\quad+\left(\sigma_{a_0}^{\;2}+\sigma_R^{\;2}\right)\mathrm{Tr}\left(\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\mathbb{1}^{\square}_{J}\,{}^{T}\frac{\partial\overline{r_{in}}}{\partial\theta^t_i}^{+}\right)\end{aligned}\tag{21}$$

This formula shows that the increase of assets to be positioned does not improve accuracy (as the positioning algorithm does not calibrate the gain mismatch, this result seems fine). The number of measurements N only helps to mitigate the fast fading and noises measurement σ_m^2 and does not compensate the calibration of the APs gains. Hence, we see that we can stop measuring values when we reach $\sigma_G{}^2 \gg \sigma_m^2/N$. Two terms depending on the geometrical configuration represent the Geometrical Dilution of Precision (GDoP) of the positioning. This means that the optimal AP placement depends on the quality of assets, APs gain calibrations, and fading probability.

3. Simulations and Experiments

In this section we compare analytic expression with simulations. The first subsection validates the analytic form of the FIM with simulations, showing analytic versus simulated covariance matrices. Then Section 3.2 compares results from real measurements against CRLB expectations for various results from a numerical simulation. Those results were made for several different configurations of APs.

3.1. Numerical Verification of Cramer-Rao Lower Bound. For the numerical verification of previous equations, we used a simulator generating RSSIs with an additional pseudorandom noise. From those values we computed a position estimate using the MLE, and we measured the covariance matrix of those estimates. We then plotted the estimates coordinates, the 2-sigma ellipsis of the numerical verification, and the CRLB covariance matrix. An example of the results is shown in Figure 1 with a room of11 by 6 meters, one AP at each corner, $\sigma_G = \sigma_R = \sqrt{2}$ and $\sigma = \sqrt{3}$, and two target assets at positions $x, y = [5 \ \ 5]$ and $x, y = [2 \ \ 2]$. In this figure, we can see that the simulated covariance is bigger than the CRLB which is consistent with the theory. On regions closer to APs (position of transmitter 1 in Figure 1), the likelihood function is highly nonlinear due to the nonlinearity of the LDPL model close to zero; hence the estimated positions will no longer be spread linearly but rather in circle around the AP combined with the fact that we have further all APs and the GDOP is not good in this configuration; some noisy measurements might be estimated outside the room, which is distorting the covariance ellipsis: the CRLB ellipsis less fits the simulated values; however estimates covariance is always above the CRLB, so the results are still theoretically correct. We can see that filtering those outliers leads to a closest match between CRLB and simulated values. On the rightmost part of the figure you can see the RMSE of the Euclidean distance ($(1/N)\sum_i^N \sqrt{(\widehat{x}_i - x)^2 + (\widehat{y}_i - y)^2}$ where $\widehat{x}$ and $\widehat{y}$ are the estimates of the real coordinates x and y) versus the trace of the CRLB (the criterion in equation (21)).

3.2. Comparison with Real Data. To make sure that the CRLB matches real case scenarios, we conducted experiments in different room configurations, with several APs configurations at various known positions and computed the Cumulated Density Function (CDF) of the distance error between estimated and real positions. This subsection depicts the setup and the results.

3.2.1. Presentation of Hardware and Experimental Setup. After a numerical verification of the equations, we confronted the CRLB results with real measurements. Experiments took place in 11 by 6 meters office with two devices from firm FFLY4U: FFLYdot and Myria, respectively, for APs and tracked devices (see Figure 2 for a schematic of the setup and pictures of the devices). Both devices use a Nordic NRF52 for Bluetooth communication and RSSI ranging. It must be noted that for practical issues in this scenario the APs were set as BLE advertisers (iBeacon) and the RSSIs were measured by the tracked asset. This is done without any loss of generality and its impact on the algorithm results in a change of gains between receivers and transmitters (R and G values are swapped). This inverted scenario is necessary because the APs only have a broadcast capability. The APs were broadcasting with a period of 25 milliseconds and a total of 140 measurements were collected for each AP. The tracked device was placed on a regular grid of 135 points spaced with a distance of 0.6 meter, using existing marks on the ground, whose size was fine-measured using a laser rangefinder. To be able to easily change the APs coordinates for position optimization tests, each AP was mounted on a mobile wooden pillar.

3.2.2. Cumulative Density Function of the Error for Simulated Data, Real Data, and CRLB in Various AP Configurations. To compare those measurements, we also computed a CDF using our RSSI simulator: for all 135 positions of measurements, we simulated 140 measurements, from which we computed the estimated position. For the CRLB, we generated 1000 positions estimates from the covariance matrix at each measuring point. It must be noted that the number of estimates is significantly higher than the CRLB because we want to generate a theoretical smooth curve and the random generated position is just an easier way to compute the theoretical curve from the analytic expression, taking into account all the different values of the CRLB at the different positions of the room. We made measurements with 7 and 4 APs placed at easiest mount points in the room (close to existing pillars, e.g., not on windows); the resulting CDF curves of those simulations and measurements are shown in Figure 3, showing that the CRLB is as expected slightly better than reality but not too much conservative or optimistic. Hence, the CRLB could be used as a good indicator of the performance of the AP topology, giving a metric of the error in meters. Moreover, simulated values also produced a more slightly better result as they do not include effect of multipath propagation and fading.

3.3. Impact of Calibration on the Expected Accuracy. Calibration is a step costly in time, which involves measuring in postproduction the static gain of all devices (APs and tracked assets), which might also vary in time for devices on battery. However, it could be skipped by measuring the average and variance of gains and model parameters and using those values in the model. As it could save time and

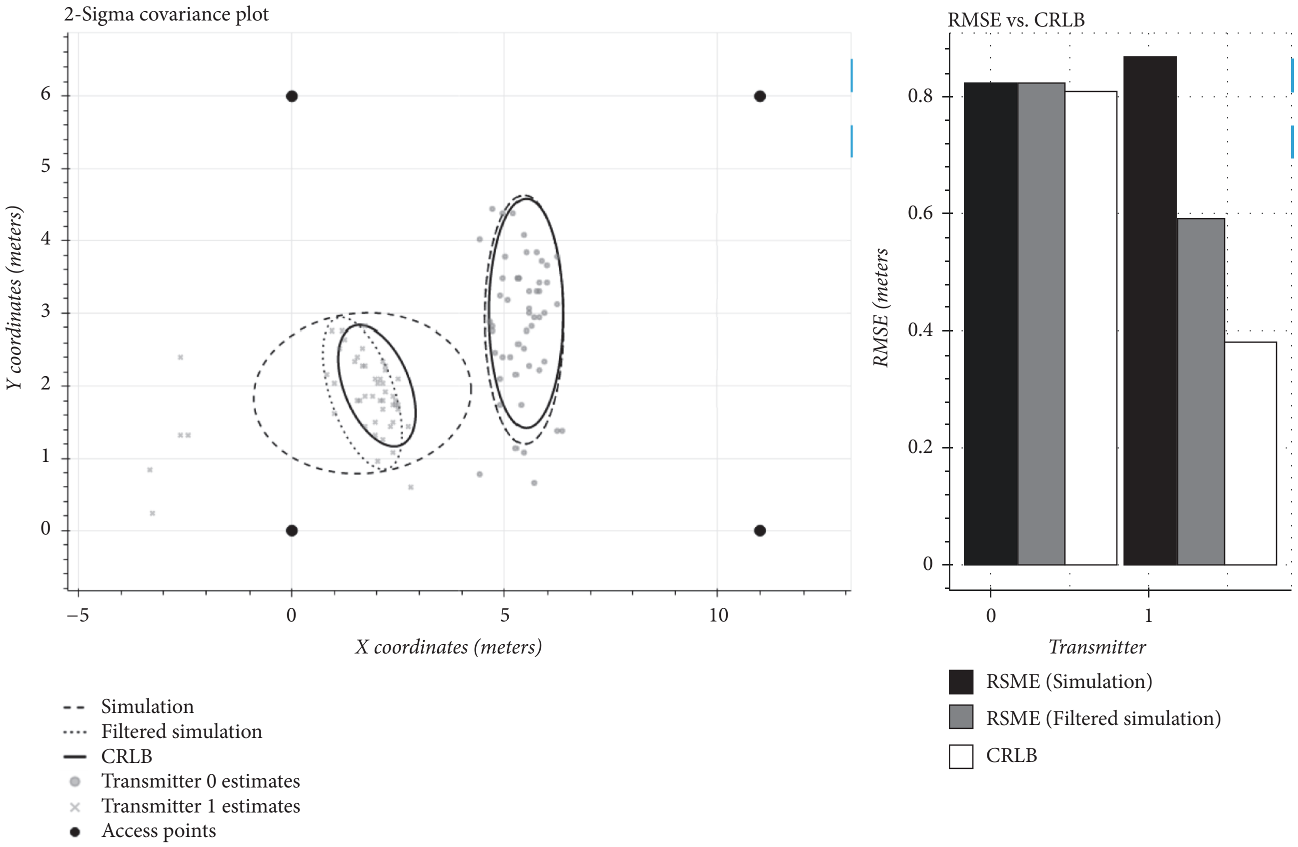


Figure 1: Numerical verification of the CRLB equation. The left figure shows 2-sigma covariance of simulated position estimates using MLE estimator vs. the covariance from our CRLB expression. Right figure shows the RMSE from simulation data and the CRLB trace.

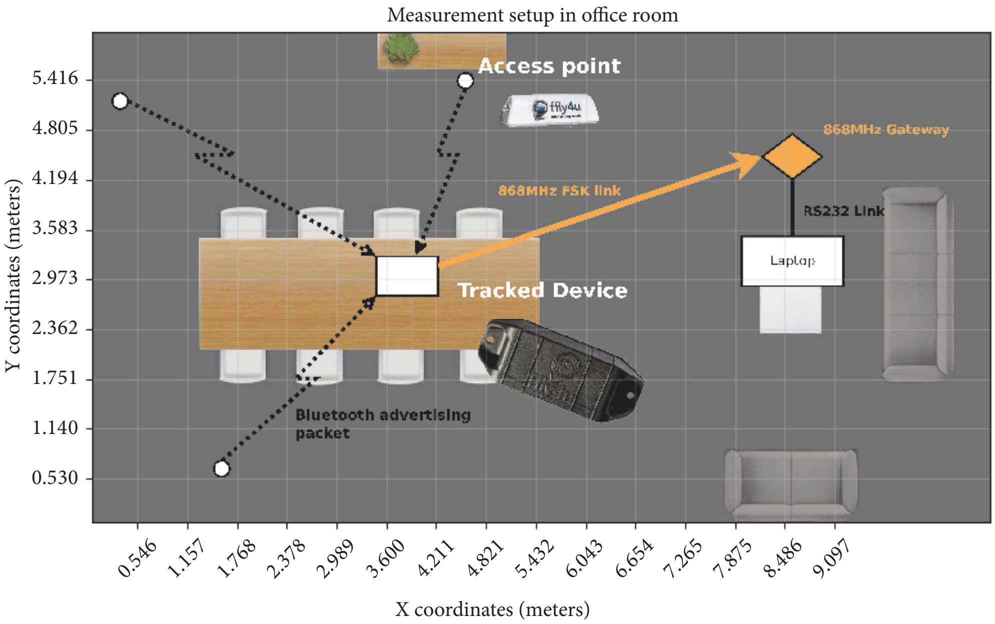


Figure 2: Measurement setup in one of the rooms. The grid shows all the measurements points.

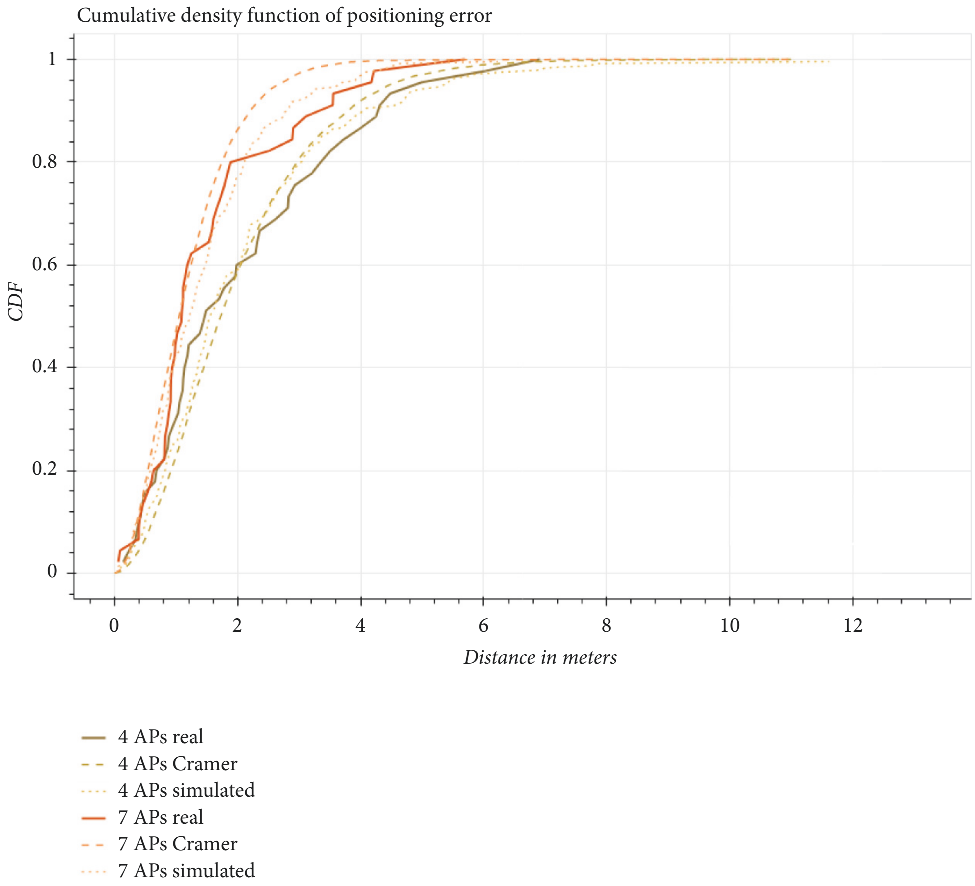


Figure 3: Experimental values matched against CRLB.

money for industrial deployments, we want to study its impact on the expected accuracy. Now, we showed that our CRLB is consistent with real measurements; we will use it as a criterion to see if calibration is required depending on the expected accuracy. We simulated the CRLB in the case of a squared room with an AP at each corner in function of the size of the room (Δ, which symbolizes the meshing density), our relative coordinates $[\delta_x, \delta_y] = [x, y]/\Delta$, and the number of measures N. We compared the result in the case where no calibration has been made ($\sigma_G = \sigma_R = 2$), where APs are calibrated in production ($\sigma_R = 2, \sigma_G \approx 0$), and where both APs and tracked devices are calibrated in production ($\sigma_R = \sigma_G \approx 0$). Results are shown in Table 2.

What we can see on those results is that if we do not calibrate the APs gains, the lower error achievable is equivalent to $\Delta/2$; that means the positioning is equivalent to a cellular algorithm. If we want a higher precision than cellular setup, we necessarily have to calibrate the APs.

4. Conclusion

The positioning problem of multiple assets with uncertain receivers gain and uncertain propagation model has been addressed. Typical values of uncertainties have been given from the observation of multiple setups and calibration realized in different industrial environments. Based on this model, the global MLE algorithm is given and formulated, having an iterative nonlinear least square solver. The CRLB is expressed in the general form and analytically for the specific case of well-calibrated path-loss exponents. Simulations show that the bound is reached by the MLE and moreover real measurements and estimations show that this result is not too much conservative; that is, the bound is a good estimation of the expected accuracy. Using this analytic CRLB, we discuss the impact on the accuracy of access points gain calibration, assets gain calibration, precision of measurements, number of measurements repetitions, and number of joint assets to be positioned.

It first shows that in contrary to joint calibration and positioning algorithms the number of assets to be estimated has no effect on the accuracy of each estimation. One can then reduce the solver complexity by running independent estimations without loss of optimality.

Secondly, it shows that the positioning of RMSE is split into two terms involving APs calibration, measurement precision and repetition, weighted by a parameter connected to the geometric position of the APs inside the room on the one hand. And on the other hand the asset calibration gain and LDPL reference gain are weighted by a different geometrical factor to impact the accuracy. Then it shows that depending on different calibration quality or efforts different geometrical terms can be involved and then different optimal

Table 2: CRLB RMSE. Coordinates are δ_x and δ_y: fractions of Δ the inter-APs distance (the meshing density).

Δ	(δ_x,δ_y)	I	Not calibrated	Calibrated APs	Calibrated APs & targets
5	(0.5, 0.5)	1	2.13	0.48	0.48
5	(0.5, 0.5)	20	2.08	0.11	0.11
5	(0.1, 0.5)	1	3.14	1.95	1.16
5	(0.1, 0.5)	20	3.09	1.87	1.01
10	(0.5, 0.5)	1	4.26	0.96	0.96
10	(0.5, 0.5)	20	4.15	0.21	0.21
10	(0.1, 0.5)	1	6.27	3.89	2.31
10	(0.1, 0.5)	20	6.17	3.73	2.03
20	(0.5, 0.5)	1	8.51	1.92	1.92
20	(0.5, 0.5)	20	8.30	0.43	0.43
20	(0.1, 0.5)	1	12.54	7.79	4.62
20	(0.1, 0.5)	20	12.35	7.47	4.05

APs configurations can arise. Using a numerical application, we also showed that with typical hardware the error was close to those of cellular positioning if we do not calibrate APs.

More generally, this can be used for dimension and to optimize a setup to reach a desired accuracy. It showed an example of how to infer the number of APs to reach accuracy in a given room. Future work could infer the number of measurements required for a given accuracy in function of the calibration error of the APs or even to find the optimal AP disposition in a given room for a given calibration.

Appendix

Full Measurement Vector Construction

Equation (3) givesthe RSS measurement obtained for signal number $n \in \mathcal{N}$ received by the AP number $j \in \mathcal{J} = \{j, 0 \le j < J\}$ sent from asset number $i \in \mathcal{I} = \{i, 0 \le i < I\}$. In this section we use matrix algebra with the Kronecker product, noted $\otimes$, to give the full IJN measurements vector r and its expectation value $\overline{r}$.

We first stack in the vector $\overline{r_{ij}}$ the N expectations of the strength measurement transmitted between a couple $(i, j) \in \mathcal{I} \times \mathcal{J}$ of target and AP. This expectation is the same for all measurements; only the measurement and shadowing effect noise $\mathcal{V}_m^{ijn} \sim \mathcal{N}(0, \sigma_m^2)$ change from one measurement to another. We then define the Gaussian random vector variable $\mathcal{V}_m^{ij}$ as ${}^T\mathcal{V}_m^{ij} = {}^T[\mathcal{V}_m^{ij0} \;\ldots\; \mathcal{V}_m^{ij(N-1)}] = \mathcal{N}(\mathbb{O}_N, \sigma_m^2 \mathbb{I}_N)$, where $\mathbb{O}_N$ is the null vector of size N, and $\mathbb{I}_N$ is the identity matrix of size N, and express those measurements as

$$\begin{aligned}
\overline{r_{ij}}\left(\theta_i^t\right) &= \left(\overline{g_j} + \overline{a_0} + \overline{r_i} + \overline{\gamma_j}\Delta_{ij}\left(\theta_i^t, \theta_j^{ap}\right)\right) \otimes \mathbb{1}_N \\
r_{ij}\left(\theta_i^t\right) &= \overline{r_{ij}}\left(\theta_i^t\right) \\
&\quad + \left(\mathcal{V}_{g_j} + \mathcal{V}_{a_0} + \mathcal{V}_{r_i} + \Delta_{ij}\left(\theta_i^t, \theta_j^{ap}\right)\mathcal{V}_{\gamma_j}\right) \\
&\quad \otimes \mathbb{1}_N + \mathcal{V}_m^{ij} \\
&= \overline{r_{ij}}\left(\theta_i^t\right) + \left[\mathbb{1}_N \;\; \mathbb{I}_N\right] \cdot \begin{bmatrix} \mathcal{V}_{a_0} \\ \mathcal{V}_m^{ij} \end{bmatrix} \\
&\quad + \left(\mathcal{V}_{g_j} + \mathcal{V}_{r_i} + \Delta_{ij}\left(\theta_i^t, \theta_j^{ap}\right)\mathcal{V}_{\gamma_j}\right) \otimes \mathbb{1}_N
\end{aligned} \tag{A.1}$$

For stacking all APs on the vector we need to define the mismatch gain vector G of those J receivers as ${}^T\overline{G} = {}^T[\overline{g_0} \;\ldots\; \overline{g_{J-1}}]$ and the corresponding gain uncertainties vector ${}^T\mathcal{V}_G = {}^T[\mathcal{V}_G^0 \;\ldots\; \mathcal{V}_G^{J-1}] \sim \mathcal{N}(\mathbb{O}_J, {\sigma_G}^2 \mathbb{I}_J)$. Similarly for the path-loss exponent we get $\overline{\gamma} = {}^T[\overline{\gamma}_0 \;\ldots\; \overline{\gamma}_{J-1}]$ and ${}^T\mathcal{V}_\gamma = {}^T[\mathcal{V}_\gamma^0 \;\ldots\; \mathcal{V}_\gamma^{J-1}] \sim \mathcal{N}(\mathbb{O}_J, {\sigma_\gamma}^2 \mathbb{I}_J)$. The log-square-distance vector between the ith asset and all the APs is noted ${}^T\Delta_i(\theta_i^t) = {}^T[\Delta_{ij}(\theta_i^t, \theta_0^{ap}) \;\ldots\; \Delta_{ij}(\theta_i^t, \theta_{J-1}^{ap})]$. Then the vector r_i of the JN measurements and its expectation $\overline{r_i}$ concerning an asset number $i \in \mathcal{I}$ are written (we consider that the Kronecker operator $\otimes$ has priority on the matrix product):

$$\begin{aligned}
\overline{r_i}\left(\theta_i^t\right) &= \left[\left(\operatorname{diag}\left(\Delta_i\left(\theta_i^t\right)\right)\overline{\gamma} + \overline{G}\right) + \left(\overline{a_0} + \overline{r_i}\right) \otimes \mathbb{1}_J\right] \\
&\quad \otimes \mathbb{1}_N \\
&= \left[\operatorname{diag}\left(\Delta_i\left(\theta_i^t\right)\right) \otimes \mathbb{1}_N \;\; \mathbb{1}_J \otimes \mathbb{1}_N \;\; \mathbb{I}_{JN}\right] \cdot \begin{bmatrix} \overline{\gamma} \\ \overline{a_0} \\ \overline{G} \end{bmatrix} \\
&\quad + \overline{r_i} \otimes \mathbb{1}_{JN}
\end{aligned} \tag{A.2}$$

$$\begin{aligned}
r_i\left(\theta_i^t\right) &= \overline{r_i}\left(\theta_i^t\right) \\
&\quad + \left[\operatorname{diag}\left(\Delta_i\left(\theta_i^t\right)\right) \otimes \mathbb{1}_N \;\; \mathbb{1}_J \otimes \mathbb{1}_N \;\; \mathbb{I}_{JN} \;\; \mathbb{I}_{JN}\right] \\
&\quad \cdot \begin{bmatrix} \mathcal{V}_\gamma \\ \mathcal{V}_{a_0} \\ \mathcal{V}_G \\ \mathcal{V}_m^i \end{bmatrix} + \mathcal{V}_{r_i} \otimes \mathbb{1}_{JN}
\end{aligned}$$

where $\mathscr{V}_m^i \sim \mathcal{N}(\mathbb{O}_{JN}, {\sigma_m}^2 \mathbb{I}_{JN})$ and $\mathbb{1}_k$ ($\mathbb{O}_k$, resp.) is the vertical vector of k ones (zeros, resp.).

Finally, we note $\overline{R}$ being the vector of the I expected assets gain mismatch defined as ${}^T\overline{R} = {}^T[\overline{r_0} \ \ldots \ \overline{r_{I-1}}]$ and its random multivariate vector ${}^T\mathscr{V}_R = {}^T[\mathscr{V}_R^0 \ \ldots \ \mathscr{V}_R^{I-1}] \sim \mathcal{N}(\mathbb{O}_I, {\sigma_R}^2 \mathbb{I}_I)$.

For the full log-square-distance vector defined as ${}^T\Delta(\theta^t) = {}^T[\Delta_i(\theta_0^t) \ \ldots \ \Delta_i({}^T\theta_{I-1}^t)]$, we added a rectangular form ${}^T\Delta_{diag}(\theta^t) = {}^T[\operatorname{diag}(\Delta_i(\theta_0^t)) \ \ldots \ \operatorname{diag}(\Delta_i(\theta_{I-1}^t))]$ to allow matrix expression. All the targets related measurements are stacked to obtain the full NJI measurements vector r and its expectation $\overline{r}$ being then expressed as

$$\begin{aligned} \overline{r}\left(\theta^t\right) &= \left[\operatorname{diag}\left(\Delta\left(\theta^t\right)\right) \cdot \mathbb{1}_I \otimes \overline{\gamma} + \overline{R} \otimes \mathbb{1}_J + \mathbb{1}_I \right. \\ &\quad \left. \otimes \left(\overline{G} + \overline{a_0} \otimes \mathbb{1}_J\right)\right] \otimes \mathbb{1}_N \\ &= \left[\Delta_{diag}\left(\theta^t\right) \ \ \mathbb{1}_{IJ} \ \ \mathbb{1}_I \otimes \mathbb{I}_J \ \ \mathbb{I}_I \otimes \mathbb{1}_J\right] \otimes \mathbb{1}_N \cdot \begin{bmatrix} \overline{\gamma} \\ \overline{a_0} \\ \overline{G} \\ \overline{R} \end{bmatrix} \\ r\left(\theta^t\right) &= \overline{r}\left(\theta^t\right) \\ &\quad + \left[\Delta_{diag}\left(\theta^t\right) \otimes \mathbb{1}_N \ \ \mathbb{1}_{IJN} \ \ \mathbb{1}_I \otimes \mathbb{I}_J \otimes \mathbb{1}_N \ \ \mathbb{I}_I \otimes \mathbb{1}_{JN} \ \ \mathbb{I}_{IJN}\right] \\ &\quad \cdot \begin{bmatrix} \mathscr{V}_\gamma \\ \mathscr{V}_{a_0} \\ \mathscr{V}_G \\ \mathscr{V}_R \\ \mathscr{V}_m \end{bmatrix} \end{aligned} \tag{A.3}$$

where $\mathscr{V}_m \sim \mathcal{N}(\mathbb{O}_{IJN}, {\sigma_m}^2 \mathbb{I}_{IJN})$.

The path-loss exponents γ_j are the only parameters involved with the targets positions θ^t in this equation; thus we can define the vector $\overline{\Gamma_L}$ of expected uncertain parameters linearly involved in the equation as ${}^T\overline{\Gamma_L} = {}^T[\overline{a_0} \ \ {}^T\overline{G} \ \ {}^T\overline{R}]$ and its mismatch noise vector as ${}^T\mathscr{V}_L = {}^T[\mathscr{V}_{a_0} \ \ {}^T\mathscr{V}_G \ \ {}^T\mathscr{V}_R \ \ {}^T\mathscr{V}_m] \sim \mathcal{N}(\mathbb{O}, \Sigma_{\Gamma_L})$ which is a zero mean independent random variables vector whose diagonal covariance matrix is $\Sigma_{\Gamma_L} = \operatorname{diag}({\sigma_{a_0}}^2, {\sigma_G}^2 \mathbb{I}_J, {\sigma_R}^2 \mathbb{I}_I, {\sigma_m}^2 \mathbb{I}_{IJN})$.

Then (A.3) becomes

$$\begin{aligned} \overline{r}\left(\theta^t\right) &= M_\gamma\left(\theta^t\right) \cdot \overline{\gamma} + M_{\Gamma_L} \cdot \overline{\Gamma_L} \\ r\left(\theta^t\right) &= \overline{r}\left(\theta^t\right) + M_\gamma\left(\theta^t\right) \cdot \mathscr{V}_\gamma \\ &\quad + \left[M_{\Gamma_L} \mid \mathbb{I}_{IJN}\right] \cdot \mathscr{V}_L \end{aligned} \tag{A.4}$$

$$\text{with } M_\gamma\left(\theta^t\right) = \Delta_{diag}\left(\theta^t\right) \otimes \mathbb{1}_N$$

$$M_{\Gamma_L} = \left[\mathbb{1}_{IJ} \ \ \mathbb{1}_I \otimes \mathbb{I}_J \ \ \mathbb{I}_I \otimes \mathbb{1}_J\right] \otimes \mathbb{1}_N$$

Then the independent random Gaussian vector $\mathscr{V}_\gamma$ is linearly mixed by $M_\gamma(\theta^t)$ to give the multivariate Gaussian vector $\mathscr{W}_\gamma \sim \mathcal{N}(\mathbb{O}, \Sigma_{\mathscr{W}_\gamma}(\theta^t))$. The covariance is given by

$$\begin{aligned} \Sigma_{\mathscr{W}_\gamma}\left(\theta^t\right) &= {\sigma_\gamma}^2 M_\gamma \cdot {}^T M_\gamma \\ &= {\sigma_\gamma}^2 \operatorname*{diag}_{IJ}\left(\Delta_{ij}\left(\theta_i^t, \theta_j^{ap}\right)^2\right) \otimes \mathbb{1}_N^\square \end{aligned} \tag{A.5}$$

where $\operatorname{diag}_{(i,j)\in I\times J}(A_{ij})$ is the block diagonal matrix defined as $((A_{ij})_{j,j})_{i,i}$.

The independent measurement noise vector $\mathscr{V}_m$ is added to the independent Gaussian vector Γ_L linearly mixed with M_{Γ_L} to give the multivariate Gaussian vector $\mathscr{W}_L \sim \mathcal{N}(\mathbb{O}, \Sigma_{\mathscr{W}_L})$:

$$\begin{aligned} \Sigma_{\mathscr{W}_L} &= \left[M_{\Gamma_L} \mid \mathbb{I}_{IJN}\right] \cdot \Sigma_{\mathscr{V}_L} \cdot {}^T\left[M_{\Gamma_L} \mid \mathbb{I}_{IJN}\right] \left({\sigma_{a_0}}^2 \mathbb{1}_{IJ}^\square \right. \\ &\quad \left. + {\sigma_G}^2 \mathbb{1}_I^\square \otimes \mathbb{I}_J + {\sigma_R}^2 \mathbb{I}_I \otimes \mathbb{1}_J^\square\right) \otimes \mathbb{1}_N^\square + {\sigma_m}^2 \mathbb{I}_{IJN} \end{aligned} \tag{A.6}$$

where $\mathbb{1}_k^\square$ is the square $k \times k$ matrix full of ones.

Finally measurements are modeled by

$$\begin{aligned} \overline{r}\left(\theta^t\right) &= M_\gamma\left(\theta^t\right) \cdot \overline{\gamma} + M_{\Gamma_L} \cdot \overline{\Gamma_L} \\ r\left(\theta^t\right) &= \overline{r}\left(\theta^t\right) + \mathscr{W}_\gamma\left(\theta^t\right) + \mathscr{W}_L \end{aligned} \tag{A.7}$$

The two multivariate random vectors $\mathscr{W}_\gamma$ and $\mathscr{W}_L$ are independent so they can be added to form a single multivariate vector which simplify expression to

$$\begin{aligned} &\overline{r}\left(\theta^t\right) = M_\Gamma\left(\theta^t\right) \cdot \overline{\Gamma} \\ &r\left(\theta^t\right) = \overline{r}\left(\theta^t\right) + \mathscr{W} \\ &\text{with } \mathscr{W} \sim \mathcal{N}\left(\mathbb{O}_{IJN}, \Sigma_{\mathscr{W}}\left(\theta^t\right) = \Sigma_{\mathscr{W}_\gamma}\left(\theta^t\right) + \Sigma_{\mathscr{W}_L}\right) \\ &\Sigma_{\mathscr{W}}\left(\theta^t\right) = \left({\sigma_\gamma}^2 \operatorname*{diag}_{IJ}\left(\Delta_{ij}\left(\theta_i^t, \theta_j^{ap}\right)^2\right) + {\sigma_{a_0}}^2 \mathbb{1}_{IJ}^\square \right. \\ &\qquad \left. + {\sigma_G}^2 \mathbb{1}_I^\square \otimes \mathbb{I}_J + {\sigma_R}^2 \mathbb{I}_I \otimes \mathbb{1}_J^\square\right) \otimes \mathbb{1}_N^\square \\ &M_\Gamma\left(\theta^t\right) = \left[\Delta_{diag}\left(\theta^t\right) \ \ \mathbb{1}_{IJ} \ \ \mathbb{1}_I \otimes \mathbb{I}_J \ \ \mathbb{I}_I \otimes \mathbb{1}_J\right] \otimes \mathbb{1}_N \end{aligned} \tag{A.8}$$

Abbreviations

AoA:	Angle of Arrival
AP:	Access Point
BLE:	Bluetooth Low Energy
CDF:	Cumulated Density Function
CRLB:	Cramér-Rao Lower Bound
DToA:	Difference Time of Arrival
FIM:	Fisher Information Matrix
GDoP:	Geometrical Dilution of Precision
GNNS:	Global Navigation Satellite System
IoT:	Internet of Things
LDPL:	Log-Distance Path Loss
LPWAN:	Low Power Wide Area Network
MLE:	Maximum Likelihood Estimator
MPPI:	Moore-Penrose Pseudoinverse
NLoS:	Nonline of Sight

PDF:	Probability Density Function
RMS:	Root Minimum Square
RMSE:	Root Minimum Square Error
RSS:	Received Signal Strength.
RSSI:	Received Signal Strength Indication
SMP:	Smallest M-vertex Polygon
SVM:	Support Vector Machine
ToA:	Time of Arrival
UWB:	Ultra Wide Band.

Data Availability

Position estimation algorithm and Cramér-Rao bound ellipses can be computed following the instructions and using the code provided in https://github.com/XavierTolza/RssiCRLB-plots.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] K. S. Raju, Y. Pratap, V. Patel et al., "Implementation of multichannel GPS receiver baseband modules," in *Advances in Computer Science, Engineering and Applications*, vol. 1, pp. 817–824, Springer, Berlin, Germany, 2012.

[2] T. H. Meng, "Low-power GPS receiver design," in *Proceedings of the IEEE Workshop on Signal Processing Systems, SIPS '98*, pp. 1–10, IEEE, 1998.

[3] B. H. Ong, M. J. Wu, W. L. Lien, C. F. Kuo, S. L. Chew, and C. H. Lu, "Low power CMOS GPS/GALILEO RF front-end receiver," in *Proceedings of the International Symposium on Radio-Frequency Integration Technology, RFIT '09*, pp. 16–19, IEEE, 2009.

[4] B. Z. Tang, S. Longfield, S. A. Bhave, and R. Manohar, "A low power asynchronous GPS baseband processor," in *Proceedings of the International Symposium on Asynchronous Circuits and Systems*, pp. 33–40, 2012.

[5] F. M. Jumaah, S. J. Hashim, R. M. Sidek, and F. Z. Rokhani, "Low power GPS baseband receiver design," in *Proceedings of the 4th Annual International Conference on Energy Aware Computing Systems and Applications, ICEAC '13*, pp. 65–68, 2013.

[6] Sigfox, "Sigfox Geolocation, the simplest and cheapest IoT location service," 2017.

[7] N. Podevijn, D. Plets, J. Trogh et al., "TDoA-based outdoor positioning with tracking algorithm in a public LoRa network," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 1864209, 9 pages, 2018.

[8] M. Jol, "LoRa Geolocation," 2016.

[9] B. C. Fargas and M. N. Petersen, "GPS-free geolocation using LoRa in low-power WANs," in *Proceedings of the Global Internet of Things Summit, GIoTS '17*, 2017.

[10] J. N. Nine and S. Boudaud, "LPWAN as enabler for widespread geolocation solutions," in *Proceedings of the Embedded World*, 2017.

[11] C. Kee, D. Yun, H. Jun, B. Parkinson, S. Pullen, and T. Centimeter-accuracy indoor navigation using GPS-like pseudolites, "GPS World," 2001.

[12] N. Jardak and N. Samama, "Indoor positioning based on GPS-repeaters: Performance enhancement using an open code loop architecture," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 45, no. 1, pp. 347–359, 2009.

[13] R. Xu, W. Chen, Y. Xu, and S. Ji, "A new indoor positioning system architecture using GPS signals," *Sensors*, vol. 15, no. 5, pp. 10074–10087, 2015.

[14] S. Nirjon, J. Liu, G. DeJean, B. Priyantha, Y. Jin, and T. Hart, "Coin-Gps," in *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys '14*, pp. 301–314, 2014.

[15] J. M. Fresno, G. Robles, J. M. Martínez-Tarifa, and B. G. Stewart, "Survey on the performance of source localization algorithms," *Sensors*, vol. 17, no. 11, article 2666, 2017.

[16] R. F. Brena, J. P. García-Vázquez, C. E. Galván-Tejada, D. Muñoz-Rodriguez, C. Vargas-Rosales, and J. Fangmeyer, "Evolution of indoor positioning technologies: a survey," *Journal of Sensors*, vol. 2017, Article ID 2630413, 21 pages, 2017.

[17] A. Yassin, Y. Nasser, M. Awad et al., "Recent advances in indoor localization: a survey on theoretical approaches and applications," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 1327–1346, 2017.

[18] M. R. Gholami, E. G. Ström, H. Wymeersch, and M. Rydström, "On geometric upper bounds for positioning algorithms in wireless sensor networks," *Signal Processing*, vol. 111, pp. 179–193, 2015.

[19] S. Gezici, "A survey on wireless position estimation," *Wireless Personal Communications*, vol. 44, no. 3, pp. 263–282, 2008.

[20] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 37, no. 6, pp. 1067–1080, 2007.

[21] R. Yamasaki, A. Ogino, T. Tamaki, T. Uta, N. Matsuzawa, and T. Kalo, "TDOA location system for IEEE 802.11b WLAN," in *Proceedings of the Wireless Communications and Networking Conference*, vol. 4, pp. 2338–2343, IEEE, March 2005.

[22] V. C. Paterna, A. C. Augé, J. P. Aspas, and M. A. P. Bullones, "A bluetooth low energy indoor positioning system with channel diversity, weighted trilateration and kalman filtering," *Sensors*, vol. 17, no. 12, article 2927, 2017.

[23] T. E. Abrudan, M. Paula, J. Barros, J. P. S. Cunha, and N. M. G. B. de Carvalho, "Indoor location estimation and tracking in wireless sensor networks using a dual frequency approach," in *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN '11)*, pp. 1–4, Guimarães, Portugal, 2011.

[24] M. Ali, S. Hur, and Y. Park, "LOCALI: calibration-free systematic localization approach for indoor positioning," *Sensors*, vol. 17, no. 6, article 1213, 2017.

[25] K. Majeed, S. Sorour, T. Y. Al-Naffouri, and S. Valaee, "Indoor localization and radio map estimation using unsupervised manifold alignment with geometry perturbation," *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2794–2808, 2016.

[26] C.-C. Huang, W.-C. Chan, and M. Hung-Nguyen, "Unsupervised radio map learning for indoor localization," in *Proceedings of the 4th IEEE International Conference on Consumer Electronics - Taiwan, ICCE-TW '17*, pp. 79-80, June 2017.

[27] H. Wang, "Bayesian radio map learning for robust indoor positioning," in *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation, IPIN '11*, 2011.

[28] J. Wang, Q. Gao, H. Wang, H. Chen, and M. Jin, "Differential radio map-based robust indoor localization," *EURASIP Journal on Wireless Communications and Networking*, vol. 2011, no. 1, article 17, 2011.

[29] X. Tolza, P. Acco, and J.-Y. Fourniols, "Impact of calibration on indoor positionning precision," in *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN '18)*, pp. 1–3, Nantes, France, 2018.

[30] H. Shi, X. Li, Y. Shang, and D. Ma, "Cramer-Rao bound analysis of quantized RSSI based localization in wireless sensor networks," in *Proceedings of the 11th International Conference on Parallel and Distributed Systems (ICPADS '05)*, vol. 2, pp. 32–36, IEEE, 2005.

[31] S. Mazuelas, A. Bahillo, R. M. Lorenzo et al., "Robust indoor positioning provided by real-time rssi values in unmodified WLAN networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 5, pp. 821–831, 2009.

[32] N. L. Johnson and T. W. Anderson, "Introduction to multivariate statistical analysis," *Journal of the Royal Statistical Society*, vol. 121, no. 4, article 482, 1958.

[33] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, vol. 1 of *Society for Industrial and Applied Mathematics*, 1995.

[34] R. Penrose and J. A. Todd, "A generalized inverse for matrices," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, no. 3, pp. 406–413, 1955.

Hindawi
Journal of Sensors
Volume 2019, Article ID 3409525, 15 pages
https://doi.org/10.1155/2019/3409525

Research Article

Traffic Sensing Methodology Combining Influence Line Theory and Computer Vision Techniques for Girder Bridges

Xudong Jian,[1] Ye Xia,[2] Jose A. Lozano-Galant,[3] and Limin Sun[1]

[1] *State Key Laboratory for Disaster Reduction in Civil Engineering, Tongji University, Shanghai 200092, China*
[2] *Department of Bridge Engineering, Tongji University, Shanghai 200092, China*
[3] *Department of Civil Engineering, University of Castilla-La Mancha, Ciudad Real 13071, Spain*

Correspondence should be addressed to Ye Xia; yxia@tongji.edu.cn

Received 22 January 2019; Revised 8 March 2019; Accepted 7 May 2019; Published 27 May 2019

Guest Editor: Samir Mustapha

Collecting the information of traffic load, especially heavy trucks, is crucial for bridge statistical analysis, safety evaluation, and maintenance strategies. This paper presents a traffic sensing methodology that combines a deep learning based computer vision technique with the influence line theory. Theoretical background and derivations are introduced from both aspects of structural analysis and computer vision techniques. In addition, to evaluate the effectiveness and accuracy of the proposed traffic sensing method through field tests, a systematic analysis is performed on a continuous box-girder bridge. The obtained results show that the proposed method can automatically identify the vehicle load and speed with promising efficiency and accuracy and most importantly cost-effectiveness. All these features make the proposed methodology a desirable bridge weigh-in-motion system, especially for bridges already equipped with structural health monitoring system.

1. Introduction

Modern bridges are mainly constructed for traffic purposes. Accordingly, collecting the information of traffic including vehicle weight, velocity, quantity, type, and spatiotemporal distribution, is crucial for bridge design refinement, safety evaluation, and maintenance strategies [1–3]. To this end, a number of studies on traffic information identification have been conducted. Among these methods, the bridge-weigh-in-motion (BWIM) technique is highlighted [4, 5].

The concepts behind BWIM techniques were initially proposed by Moses [6], who used an instrumented bridge as the weighing scale to estimate vehicle weights. Compared with other weigh-in-motion (WIM) techniques, such as the pavement-based WIM systems [7, 8], BWIM techniques are cost-efficient, durable, and unbiased as they are not impacted by repeated axle loads and do not require interrupting the traffic to cut the pavement. All these advantages have made BWIM a preferable tool to weigh vehicles, especially heavy trucks, attracting many follow-up research and engineering applications. Up to date, this research topic has progressed significantly in aspects as diverse as the identification results, such as time-history moving load identification [9–11], or the types of sensors, such as portable accelerometers [12].

One of the most simple and practical BWIM techniques verified by field tests is the gross vehicle weight (GVW). This identification method is based on the static influence line/surface theory, which is already applied by Moses in his earliest research [13]. However, key problems arise in obtaining accurate results when multiple vehicles cross the bridge deck simultaneously or move transversely [14]. In this scenario, combining supplemental vehicles position information and the influence surface instead of influence line might help to mitigate the problem. To position the vehicles on the bridge, traffic sensors such as radar, road tubes, and embedded axle detectors are recommended by Snyder et al. [15]. Lamentably, these sensors are too costly for its massive installation in actual structures. Alternatively, Xiao et al. [16] and Yamaguchi et al. [17] innovatively utilized the longitudinal ribs strains of an orthotropic steel bridge to detect the transverse position of vehicle axles. Unfortunately, concrete bridges without ribs are insensitive to single axle loads, making this method ineffective for this kind of structures. Yu et al. [18] proposed a novel BWIM algorithm that was able to identify the lateral position of a single vehicle on a bridge by using only seven strain gauges installed transversely at the bottom of the beams. That paper, however, admitted that identifying the presence of multiple-vehicle is still one of the main challenges faced by BWIM technology.

To address the multiple-vehicle presence challenge, using visual information is an innovative and feasible solution on the basis that a large number of bridges have been equipped with surveillance cameras for traffic monitoring of late years. In fact, the rich visual information recorded by the surveillance cameras enables obtaining the exact position of the vehicles on the bridge deck with nothing but a common webcam. Chen et al. [19] proposed an identification approach for the spatiotemporal distribution of traffic loads on bridges using the information from the pavement-based WIM and background subtraction technique. This approach relies on high quality video image, which limits its range of application. Another disadvantage of this method is the fact that it is nonsemantic, which means deep information contained in the video image, such as type and axle number of vehicles, is difficult to obtain. Similar problems also exist in studies aiming to detect vehicle axles using traditional computer vision techniques [1, 20].

In recent years, deep learning methods have dramatically improved the state of the art in visual object detection and recognition with amazing efficiency and robustness [21]. Inspired by the tremendous advance of computer vision techniques, this paper presents a traffic information identification methodology in combination with influence line theory and deep learning based computer vision techniques.

This paper is organized as follows. Firstly, the theoretical background of both aspects of structural analysis and computer vision techniques is presented. Next, field tests on a box-girder bridge were conducted to evaluate the proposed methodology in various aspects. Finally, both the advantages and the potential engineering applications of the methodology are discussed.

2. Structural Analysis

2.1. Bridge Response Analysis. One of the most concerning traffic information is vehicle weight. To estimate the vehicle weight, BWIM technology traditionally uses the bridge strains. Therefore, bridge structural strain analysis is essential in the process of vehicle weight identification.

Most BWIM systems are applied on girder bridges with small or medium span due to its structural simplicity. Compared with long-span bridges, middle-small span girder bridges perform linear elasticity under normal operation, making them ideal weighing scales to estimate vehicle weights. Moreover, load effects on such bridges are relatively simple, which can be expressed as follows:

$$\varepsilon_{bridge} = \varepsilon_{environment} + \varepsilon_{vehicle} \tag{1}$$

$$\varepsilon_{vehicle} = \varepsilon_{dynamic} + \varepsilon_{static} \tag{2}$$

where ε_{bridge} is the directly measured bridge strain; $\varepsilon_{environment}$ is the bridge strain caused by environmental factors, such as temperature, wind, slight earth pulse, and creep of concrete; $\varepsilon_{vehicle}$ is the bridge strain induced by vehicles, which includes dynamic $\varepsilon_{dynamic}$ and static ε_{static} components.

According to the influence theory, the static component ε_{static} is to be extracted from ε_{bridge} by filtering $\varepsilon_{environment}$ and $\varepsilon_{dynamic}$ for the purpose of traffic load identification. In this paper, the filtering process is divided into two steps: (i) the ε_{bridge} time-history curve is robustly smoothed to get $\varepsilon_{environment}$ and subtract $\varepsilon_{environment}$ from ε_{bridge} to get $\varepsilon_{vehicle}$, and (ii) the $\varepsilon_{vehicle}$ time-history curve is smoothed to get the desired ε_{static}. The whole procedure is shown in Figure 1 for intuitive illustration.

To achieve the filtering process in time domain, a local regression algorithm named locally weighted scatterplot smoothing (LOWESS) is used. Chief attractions of this algorithm are the accuracy and convenience. It is not required to specify a global function of any form to fit a model to the data, only to fit segments of the data so that satisfactory local accuracy is achieved. According to Cleveland and Devlin [22], the basic principle of the LOWESS is expressed as follows.

First of all, the LOWESS belongs to the regression analysis, which aims to fit the mathematical relationship between two sequences x_i and y_i. In this paper, x_i is considered as the time sequence t_i, while y_i is the bridge strain data sequence ε_i.

The LOWESS adopts the polynomial regression model, expression of which is [23]

$$\varepsilon_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \cdots + \beta_d t_i^d + \varepsilon_i = \sum_{j=0}^{d} \beta_j t_i^j + \theta_i \quad (i = 1, 2, \ldots, n) \tag{3}$$

where β_j is the coefficient of the polynomial regression model, d is the order of the polynomial, θ_i is the random error, and n is the length of local sequence segment. For LOWESS, taking $d = 2$ should almost always provide adequate smooth and computational efficiency.

To get appropriate coefficient $\widehat{\beta}_j(t_i)$ of the polynomial, the LOWESS chooses weighted least squares estimate method, which means $\widehat{\beta}_j(t_i)$ are the values that minimize the following function:

$$E = \sum_{k=1}^{n} w_k(t_i) \left(\varepsilon_k - \beta_0 - \beta_1 t_k - \cdots - \beta_d t_k^d\right)^2 \tag{4}$$

where $w_k(t_i)$ are weights defined for all $t_k (k = 1, \ldots, n)$. The tri-cube weight function is adopted to provide adequate smooth results.

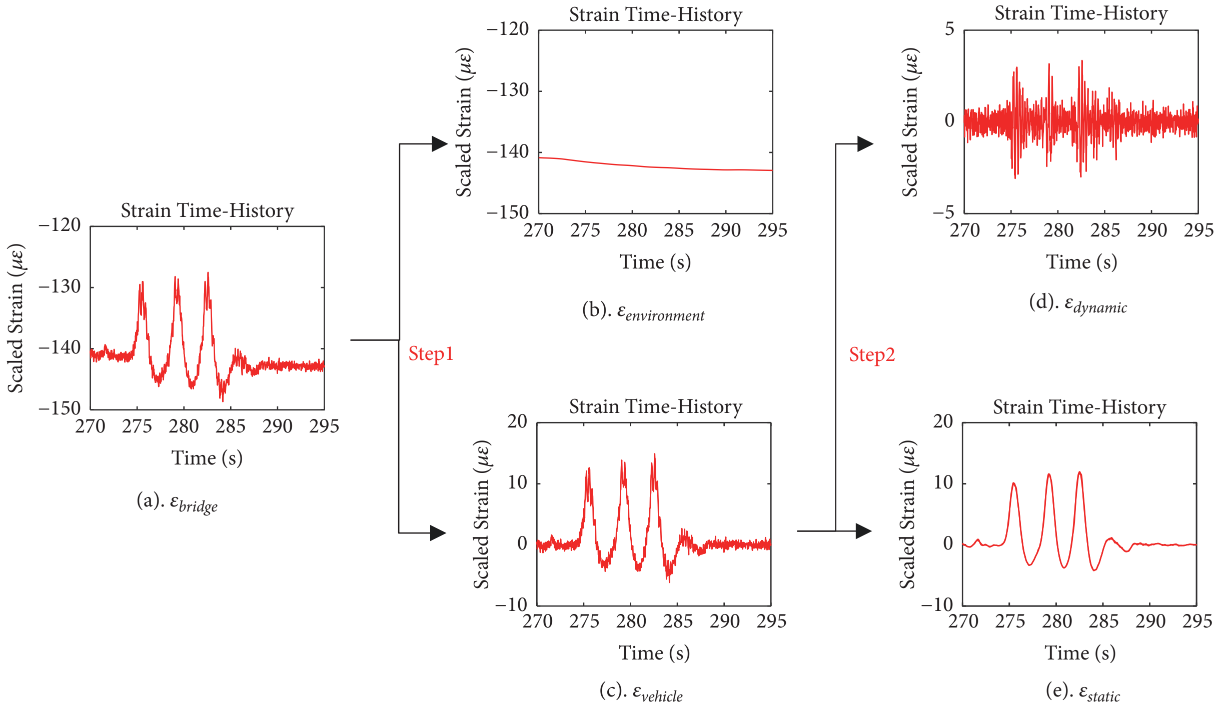
Thus $\widehat{\beta}_j(t_i)$ can be obtained by

$$\frac{\partial E}{\partial \beta} = 0 \tag{5}$$

Finally, smoothing results are

$$\widehat{\varepsilon}_i == \sum_{j=0}^{d} \widehat{\beta}_j(t_i) t_i^j \quad (i = 1, 2, \ldots, n) \tag{6}$$

where $\widehat{\varepsilon}_i$ is the smoothed strain sequence.



(a). ε_{bridge}

(b). $\varepsilon_{environment}$

(c). $\varepsilon_{vehicle}$

(d). $\varepsilon_{dynamic}$

(e). ε_{static}

Figure 1: Procedure of the vehicle induced static strain extraction.

After preselecting d, the order of the polynomial, and $w_k(t_i)$, the weight function, the only parameter left to be defined, is the length of local sequence segment, n. This parameter can be chosen on the basis of the data properties. In this paper, n is selected as 50 when smoothing the $\varepsilon_{vehicle}$ time-history curve. Because the longest vibration period of girder bridges is less than 1s, and the sampling frequency of strain sensors in this research is fixed at 50 Hz, which means that 50 data points are recorded per minute. Choosing the length of the data sequence segment as 50 for locally smoothing is enough to filter the $\varepsilon_{dynamic}$ from the $\varepsilon_{vehicle}$ hence. Similarly, $n = 500$ can be assumed for smoothing the ε_{bridge} time-history curve when a vehicle crosses a bridge with small or medium span. In these cases, the frequency is usually within 10s.

The LOWESS algorithm is capable of smoothing the $\varepsilon_{vehicle}$ time-history curve to get the desired ε_{static}. However, using LOWESS to smoothing ε_{bridge} might not be satisfactory enough. Compared with $\varepsilon_{dynamic}$, strain variation caused by vehicle weight is much more significant. Thus, the apparent peaks will distort the smoothed results as shown in Figure 2.

To prevent seriously deviant data from distorting the smoothed results, robust locally weighted regression (RLOWESS) algorithm was proposed on the basis of LOWESS [24]. Based on the size of the residual $e_k = \hat{\varepsilon}_k - \varepsilon_k$, a different set of weights, $\delta_k \cdot w_k(t_i)$, is defined for each (t_i, ε_i) as

$$\delta_k = \begin{cases} \left[1 - \left(\dfrac{e_k}{6s}\right)^2\right]^2, & \text{for } \left|\dfrac{e_k}{6s}\right| < 1 \\ 0, & \text{for } \left|\dfrac{e_k}{6s}\right| \geq 1 \end{cases} \qquad (7)$$

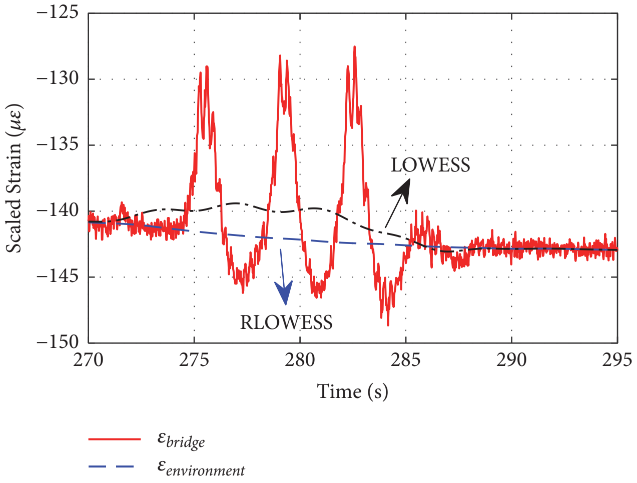


Figure 2: Normal strain signal collected from the web of a concrete box girder bridge.

where δ_k is the robust factor of weights, $e_k = \hat{\varepsilon}_k - \varepsilon_k$ is the smoothing residual, and s is the median of the $|e_k|$. By introducing δ_k, large residuals result in small weights and small residuals result in large weights. In this way, distortion produced by seriously deviant data points can be effectively mitigated as shown in Figure 2.

2.2. Influence Line Calibration. Bridge influence lines can be used to weigh vehicles and they are vital tools for BWIM analysis. In fact, obtaining an adequate accuracy of the influence line is critical for the BWIM system to achieve

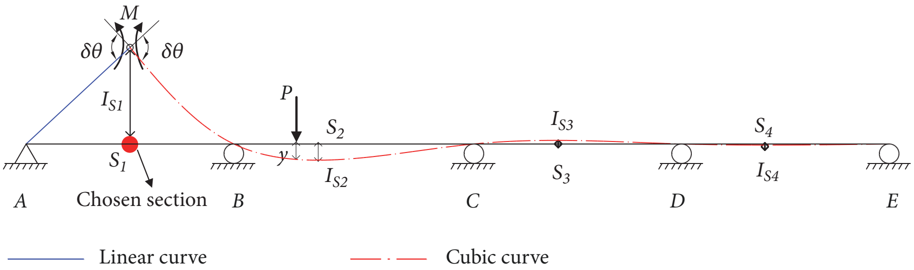


Figure 3: Diagram of the kinematic method aiming to obtain the influence line of section S_1.

convincing results. According to previous studies on BWIM, there are two methods to obtain the influence line of a bridge. One is the theoretical simulation method [6, 25, 26], and the other is the calibration method carried out in field tests [8, 27].

Apparently, numerical simulation is unable to fully reproduce the mechanical behavior of a real bridge. To fill this gap, a method fitting strain influence line with measured strain data from field calibration tests is presented in this paper. The method includes the following two steps.

Step 1 (theoretically derive the shape of the influence line). In the first step, the theoretical shape of the strain influence line of the analyzed girder bridge is obtained. According to the structural mechanics, one of the most common methods to obtain the influence line of a chosen beam section is the kinematic one [28].

Normal strain of the chosen bridge cross-section is used to weigh vehicles in this work. As it is well known, the Euler-Bernoulli beam theory states that the normal strain of the chosen cross-section of a beam under vertical loads is proportional to its bending moment. According to Timoshenko and Gere [29], the proportional relationship can be expressed as

$$\varepsilon = \frac{My}{EI} \tag{8}$$

where ε is the normal strain of a point on the chosen beam cross-section, M is the bending moment at that cross-section, y is the distance between the point and the neutral axis of the cross-section, E is the elastic modulus of the beam material, and I is the moment of inertia of the cross-section.

Equation (8) indicates that, for a fixed point on the chosen beam cross-section, the normal strain ε of that point is proportional to the bending moment M at the cross-section. Thus the shape of strain influence line of a fixed point is similar to that of the bending moment influence line at the chosen cross-section where the fixed point is located. In other words, it can be said that the two influence lines are scaled.

To illustrate the kinematic method a four-span continuous beam presented in Figure 3 with nodes from A to E is considered. This method assumes that an element of the beam at the chosen cross-section, like S_1 section in Figure 3, is replaced with an ideal hinge. It allows relative rotation between the two portions of the beam and a system with one degree of freedom is obtained in this manner. If a load P is applied at any point on the movable system, for equilibrium, a pair of two equal and opposite bending moments M is needed at the hinge. Meanwhile, virtual displacement of the movable system will be produced by the loads. For the left movable portion AS_1, the displacement curve is linear, and, for the right structure portion S_1-E, the displacement curve is a cubic [29], as shown in Figure 3.

According to the principle of virtual work, the sum of corresponding virtual work of load P and the couple M equates zero, that is

$$\begin{gathered} M \cdot \delta\theta - P \cdot y = 0 \longrightarrow \\ M = P \cdot \frac{y}{\delta\theta} \end{gathered} \tag{9}$$

where $\delta\theta$ is the total angular displacement between the two parts of the beam and y is the vertical displacement of the point where load P is applied. Thus $y/\delta\theta$ refers to the influence coefficients for bending moment at the chosen section S_1, and the diagram of structural displacement has the shape of the influence line.

Step 2 (calibrating the derived influence line with field tests data). In the second step, numerical values of the strain influence line are calibrated from field test data. Figure 3 illustrates how the influence line can be numerically fitted after introducing the real measurements (I) obtained in a field test at points S_1, S_2, S_3, and S_4 (that is to say I_{S1}, I_{S2}, I_{S3}, and I_{S4}).

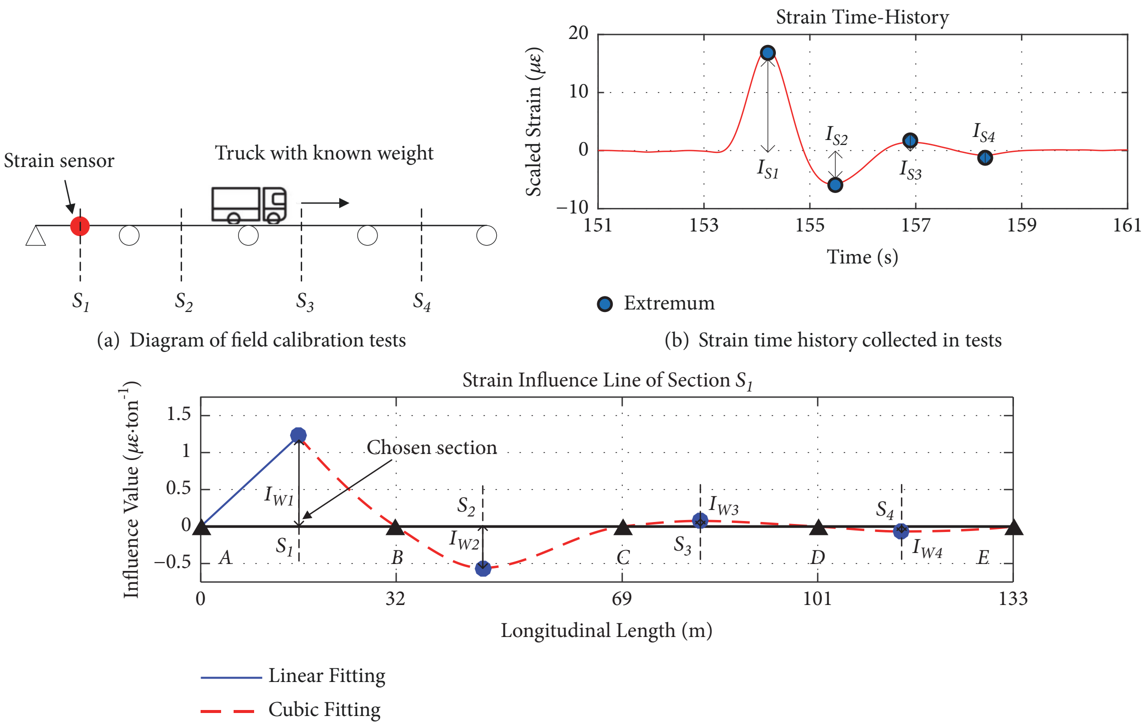
The calibration approach begins with arranging a calibration truck with known weight to cross the instrumented bridge for several times, as Figure 4(a) shows.

Since bridges are usually long relative to the spacing of vehicle axles, gross vehicle weight is more important than individual axle load [30]. Besides, for the linear elastic structures, the mechanics principle of superposition works. Considering this, the vehicle load can be simplified as a concentrated load P, which is written as

$$P = W \cdot \mathrm{g} \tag{10}$$

where W is the vehicle weight, and g is the gravitational acceleration.

According to the influence line theory, there will be an extreme on the strain time history curve recorded by a fixed strain sensor when a moving load passes a bridge span. For a four-span continuous girder bridge passed by the



(a) Diagram of field calibration tests

(b) Strain time history collected in tests

(c) Calibrated influence line of the chosen section

Figure 4: Procedure of the influence line calibration.

calibration truck, the ε_{static} time-history curve of a fixed point on bridge has four extremes, as shown in Figure 4(b). The deflections I_{s1}, I_{s2}, I_{s3}, and I_{s4} occur when the calibration truck passes cross-section S_1, S_2, S_3, and S_4. Then it is feasible to numerically fit the strain influence line of a desired point on the chosen section with nine points, A~E and $S_1 \sim S_4$ in Figure 3, of which the coordinates are determined.

Finally, the strain influence line is normalized to obtain the direct relationship between vehicle weight and bridge for BWIM application convenience. The normalization equation is as follows:

$$I_W = \frac{I_S}{W} \tag{11}$$

where I_W is the static strain caused by per unit vehicle weight, and I_S is the obtained value of the strain influence line. An example of calibrated strain influence line is shown in Figure 4(c). In this figure, four static strain values per unit vehicle weight (I_{w1}, I_{w2}, I_{w3}, and I_{w4}) are considered. As discussed previously, the polynomial order for fitting purposes depends on the order of the displacement curve.

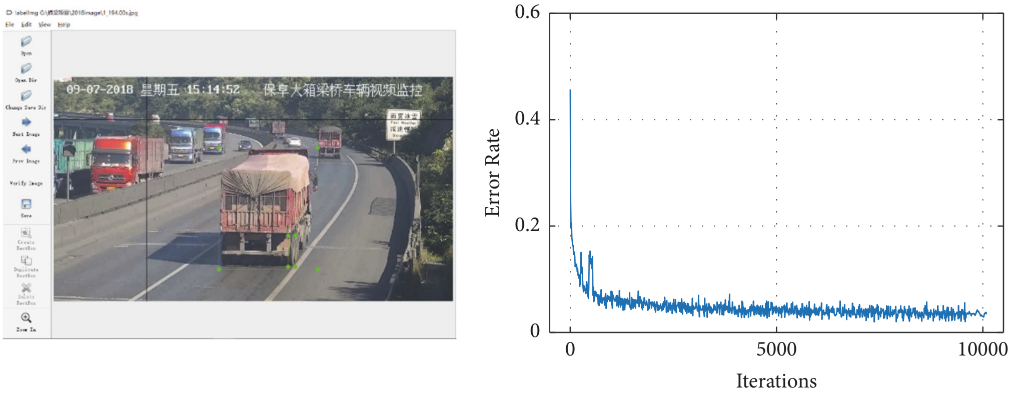
This calibration procedure of influence lines has a number of advantages, such as low calculation, operation simplicity, and no need to close traffic, enabling convenient recalibration if the mechanical performances of the instrumented bridge change [31]. However, it is noteworthy that the usage of influence line, instead of influence surface [32], is a simplification for real bridges, because vehicles may move transversely on bridge. But this simplification is still acceptable under the assumption that heavy trucks this research focuses on seldom change the traffic lane when crossing the bridge.

3. Computer Vision Technique

3.1. Deep Learning Approach. Convolution neural network (CNN) is one of the most notable deep learning approaches employed for object detection, classification, and segmentation tasks [33]. Here, learning means that CNN automatically learns useful features from the training data and distinguishes the target object and the others based on these features. Actually, that is how humans recognize objects. The CNN is therefore classified as artificial intelligence (AI) method. The learning ability is a qualitative leap over traditional manual feature extraction methods and can thus drastically reduce the workload of operation. Besides, the intelligent character also improves the robustness and generalization capacity because of the invariance to complex background, geometric distortion, and illumination.

Due to such advantages, new CNN based computer vision algorithms with better performance have been unceasingly proposed, and most of them are open source. This research applies the most advanced algorithm named YOLO V3 to fulfill the vehicle recognition tasks for its multi-scale and deeper feature extraction capacity as well as the fastest recognition speed up to date [34]. The application procedure has the following three steps.

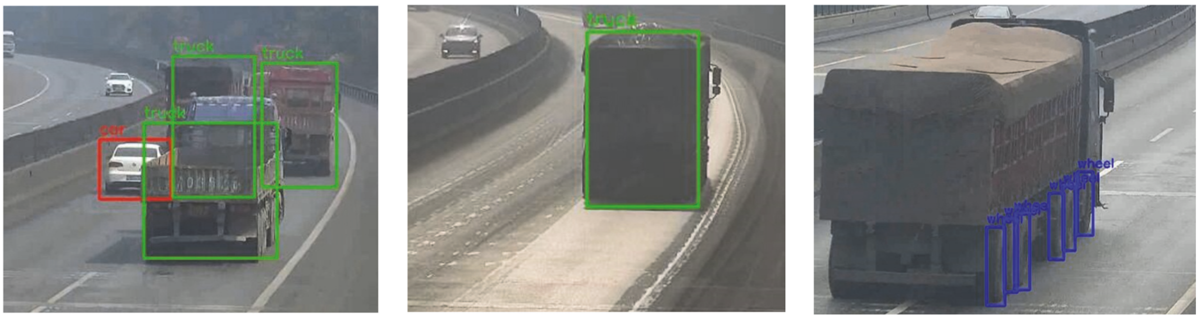
Step 1 (preparing training data sets). As stated previously, CNN based computer vision algorithms will not work without effective training. Thus training data sets need to be prepared for the YOLO V3 algorithm at first. The preparatory work includes singling out segment of videos in which vehicles exist and manually labelling cars, trucks, and wheels



(a) Manual labeling

(b) Error curve in the training process

Figure 5: Preparation procedure of YOLO V3 algorithm.



(a) Multi-vehicle overlap

(b) Illumination change

(c) Axle segmentation

Figure 6: Vehicle recognition results for different scenarios.

in every video frame, as shown in Figure 5(a). In this paper, a total of 1000 video images with the same camera visual angle are selected as the training set and different types of objects including cars, trucks, and wheels are labelled in each image.

Step 2 (training CNN of YOLO V3). CNN is essentially a set of weight coefficients capable of recognize objects using the pixel data of an image. Errors are inevitable in the process of recognition, and training CNN intends to obtain the optimal weight coefficients that minimizes the errors. To that end, the gradient descent method [33] is used in this optimization problem. This technique states

$$\frac{\partial e}{\partial w_i} \leq r \tag{12}$$

where e is the recognition error, w_i are the CNN weights, and r is the convergence threshold.

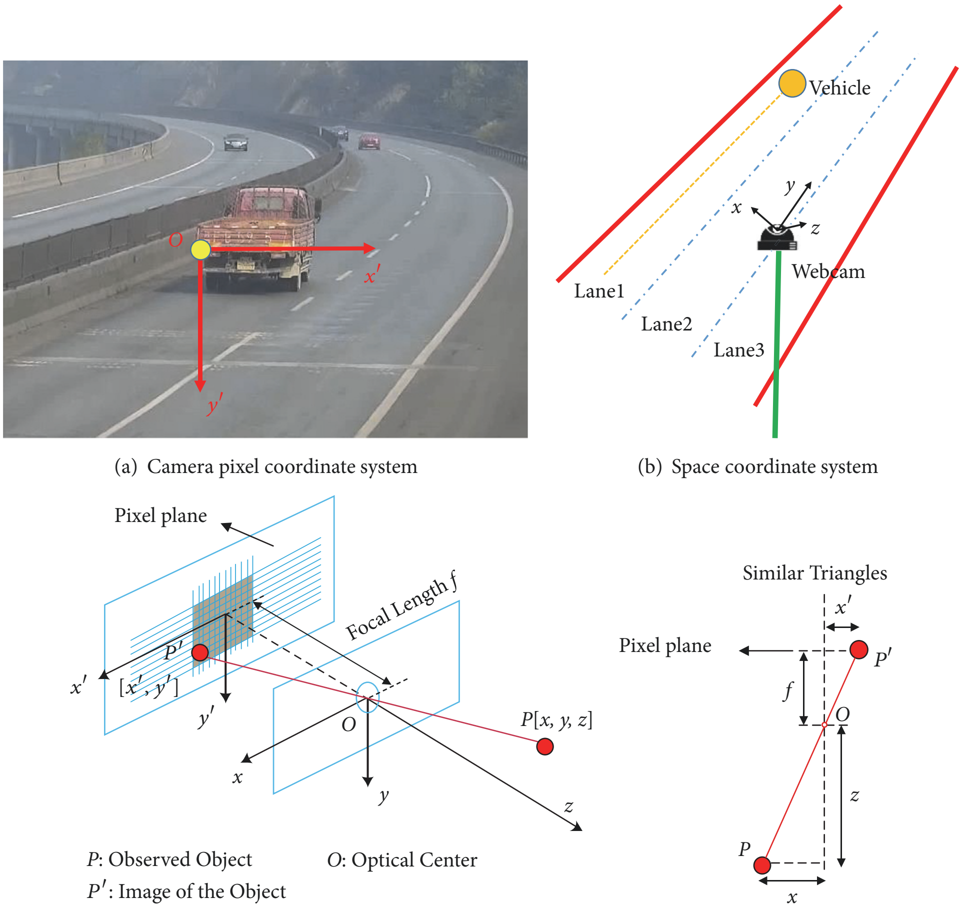
Numerical iteration is required to achieve (12), and the iteration process is shown in Figure 5(b). In this research, iteration times are set to 10000 in the six-hour training process accelerated by a NVIDIA 1080Ti GPU.

Step 3 (applying YOLO V3). After a well-trained CNN is obtained, the YOLO V3 is used to recognize vehicles in real time. Recognition results in this research were quite satisfactory in the different scenarios shown in Figure 6. It is remarkable that the closely spaced wheels are successfully recognized as shown in Figure 6(c), proving the splendid recognition capability of the YOLO V3 algorithm. The recognized pixel coordinates of the detection box are collected for further vehicle positioning tasks.

Vehicle overlap and insufficient illumination might produce inevitable recognition errors. Diversifying neural network training sets and utilizing infrared camera at dark night will help to improve the recognition accuracy.

3.2. Coordinate Transformation. After the successful recognition by YOLO V3 algorithm, precise position of vehicles has to be determined. To address this problem, coordinate systems are established as illustrated in Figure 7 [35] and a coordinate transformation method is proposed in this work. There are two coordinate systems in the process of coordinate transformation, namely, camera pixel coordinate in the video image as shown in Figure 7(a) and space coordinate in the real world as shown in Figure 7(b). The relations between them are shown in Figures 7(c) and 7(d), respectively, where the parameters with same marks are equal. Based on the relations, the spatial coordinate of a point $P'(x', y')$ on pixel plane can be transferred into $P(x, y, z)$ as follows:

$$\begin{aligned} x &= x' \cdot t \\ y &= y' \cdot t \\ z &= f \cdot t \end{aligned} \tag{13}$$



(a) Camera pixel coordinate system

(b) Space coordinate system

(c) Camera imaging spatial model

(d) Camera imaging planar model

Figure 7: Coordinate systems.

where f is the focal length of the camera and t is the similarity coefficient between the two similar triangles.

In the space coordinate, the bridge deck can be considered as a spatial plane represented by the following equation:

$$Ax + By + Cz + D = 0 \tag{14}$$

where x, y, z are the spatial coordinates of the observed object such as a truck, and A, B, C, D are unknown parameters determining the bridge deck plane equation. If the bridge slope is negligible, which is the usual case, x and y directly determine position of vehicles on the bridge deck. Then vehicle coordinates on the deck are attainable after obtaining A, B, C and D.

Instinctively, both location and orientation of the webcam are needed to obtain parameters A, B, C and D. However, a number of field conditions, such as heavy traffic flow, make it difficult to obtain this information. To solve this problem, this paper proposes a method to obtain A, B, C, D directly from the video image without knowing the camera location and/or its orientation. The proposed method only needs two lines of equal space length in the image. For example, lines $P_1'P_2'$ and $P_3'P_4'$ in Figure 8 have the equal length of 3.75m. The coordinates of their endpoints can be measured directly from the image.

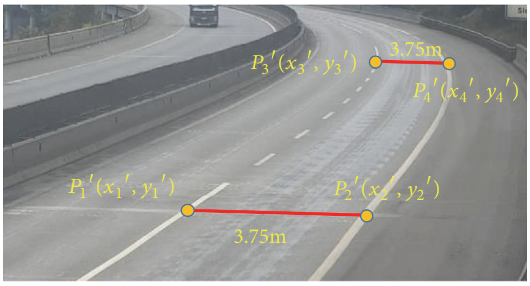
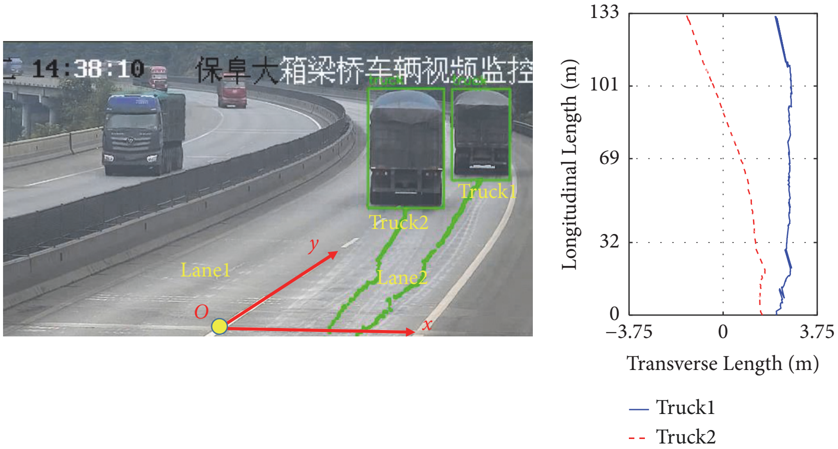


Figure 8: Diagram for A, B, C, D calculation from two reference lines P_1'-P_2' and P_3'-P_4'.

According to Figure 8, the relations between both lines can be written as follows:

$$\begin{gathered}\Delta x_1 = x_1' - x_2', \\ \Delta y_1 = y_1' - y_2' \\ \Delta x_2 = x_3' - x_4', \\ \Delta y_2 = y_3' - y_4' \\ \sqrt{\Delta x_1^2 + \Delta y_1^2} \cdot t_1 = \sqrt{\Delta x_2^2 + \Delta y_2^2} \cdot t_2 = L\end{gathered} \tag{15}$$



(a) Trajectory in image

(b) Trajectory on bridge deck

FIGURE 9: Recognized vehicle trajectory.

where (x_1', y_1'), (x_2', y_2'), (x_3', y_3'), and (x_4', y_4') are the coordinates of the endpoints P_1', P_2', P_3', and P_4', t_1 and t_2 are similarity coefficients of the lines, and L is the line length. With (15), parameter t of the two lines can be separately calculated, then spatial coordinates of the four endpoints are obtained with (13). Substituting coordinates of the four endpoints of the two equal length lines into (14), the four unknowns A, B, C and D can be directly obtained.

The main advantage of this method is its simplicity, while the trade-off is the loss of accuracy assuming that parameter t is equal for endpoints P_1' and P_2' as well as P_3' and P_4', which is true when the selected lines are far enough from the camera and the line length is short. Another noticeable error source comes from the camera imaging distortion, which is complicated and will not be discussed in this paper. Figure 9 depicts vehicle trajectory tracked by the aforementioned method.

4. Field Tests

4.1. Test Setup. In order to verify the applicability of the proposed traffic information identification methodology in real structures, field tests were conducted on a 32 m+37m+32m+32m continuous concrete box girder bridge of Baoding-Duping Highway, China. There are three traffic lanes on the bridge in total, and each of them is 3.75m wide. Among these traffic lines, lane3 is an emergency lane where vehicles are prohibited to drive under normal conditions. The bridge is slightly curved with a bending radius of 2600 m and a central angle of 2.93°; thus the curvature effects are negligible in the analysis. A structural health monitoring system comprising a pavement-based WIM system, six resistance-type strain sensors, and a webcam is installed on this bridge. All the discussed information is shown in Figure 10.

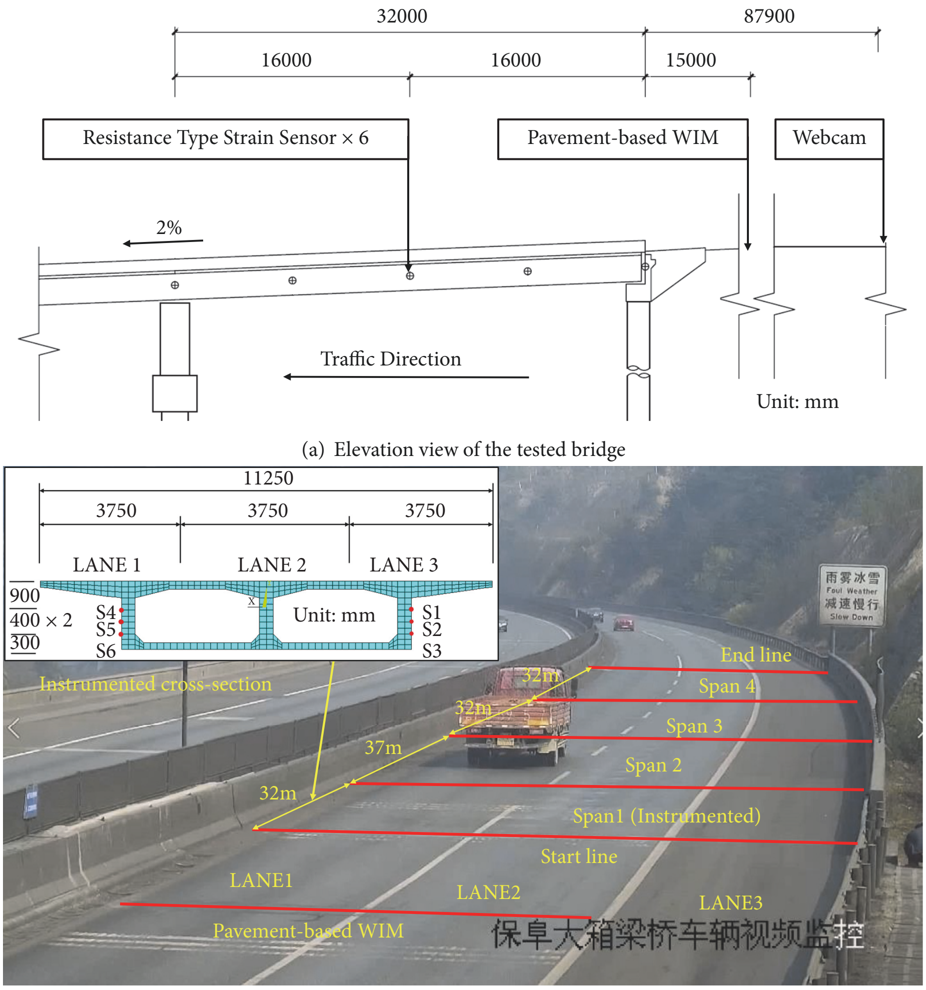
In the field tests, the normal strain data were collected by the six resistance-type strain sensors mounted on the mid-span section of the first span and stored in an online server. Video recorded by the webcam is also available on line, providing a basis for long-term online application. Vehicle weight and velocity recognized by pavement-based WIM system are used as contrast to evaluate the accuracy of this proposed methodology.

4.2. Calibration Tests. First of all, field calibration tests as on the lanes presented in Figure 10(b) were implemented following the method mentioned in Section 2.2 up front. To do so, an ordinary truck weighing 14.86t, as shown in Figure 11, was arranged to drive on the tested bridge for four times. Detailed test conditions are shown in Table 1. Lane3 was ignored for the calibration because it is an emergency lane where vehicles are prohibited to drive under normal conditions.

Figure 12 shows the influence line calibration results of the “S6” strain sensor in Figure 10(b). Differences between test1 and test2, as well as test3 and test4, are slight, which verifies the feasibility and reliability of the proposed influence line calibration method. It is also observed that the influence value of traffic lane1 is larger than that of lane2, as strain sensor “S6” is located closer to lane1.

4.3. Strain Data Processing. Next, strain data collected by six resistance-type strain sensors, shown in Figure 10, is processed with the LOWESS algorithm mentioned in Section 2.1 of this paper. Taking a segment of the processed strain time history shown in Figure 13(a) as an example, an obvious linear relationship between data peak values of strain sensors mounted on the same box web is observed in Figure 13(b). The linear relationship confirms the plane section assumption of Euler-Bernoulli beam theory mentioned in (8) above and thus validates the effectiveness of the strain data processing method.

4.4. GVW Calculation. The gross vehicle weight (GVW) can be calculated by combining the calibrated strain influence line, the processed bridge strains, and the vehicle position. Basically, there are only three elementary vehicle distribution



(a) Elevation view of the tested bridge

(b) Diagram of the tested bridge and the instrumented section

Figure 10: Bridge for field tests.

Table 1: Conditions of calibration tests.

	Test1	Test2	Test3	Test4
Weight	14.86t	14.86t	14.86t	14.86t
Velocity	60km/h	80km/h	60km/h	80km/h
Lane	Lane1	Lane1	Lane2	Lane2

scenarios presented in Figure 14. They are single vehicle in Figure 14(a), one-by-one vehicles on the same lane in Figure 14(b), and side-by-side vehicles on different lanes in Figure 14(c).

For the first single vehicle scenario, it is simple to calculate weight of the vehicle through the following equation:

$$W = \frac{S^{peak}}{I^{peak}} \tag{16}$$

where S^{peak} is the peak value of vehicle induced static strain signal, I^{peak} is the peak value of the calibrated strain influence line, and W is the GVW of the vehicle.

For the second one-by-one vehicles scenario, GVW of the first front vehicle can still be calculated through (16). Then GVW of the rear vehicles can be calculated after subtracting effects of the front vehicle whose GVW is already known. This process is written as

$$W_{rear} = \frac{S_{rear}^{peak} - I_{front} \cdot W_{front}}{I_{rear}^{peak}} \tag{17}$$

where S_{rear}^{peak} is the peak value of the rear vehicle induced static strain signal, I_{front} is the strain influence value related to the position of the front vehicle, which can be obtained with the aid of the aforementioned computer vision technique, W_{front} is the GVW of the front vehicle calculated through (16), I_{rear}^{peak} is the peak value of the calibrated strain influence line of

Table 2: Statistics of the relative errors compared with pavement-based WIM.

Sensor	Mean of errors (%)	Standard deviation of errors (%)
S1	35.2	37.4
S2	-3.6	18.7
S3	-2.8	20.8
S4	38.6	25.2
S5	-1.8	12.4
S6	-5.2	11.6

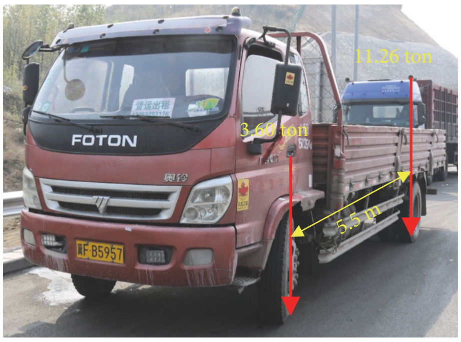


Figure 11: Calibration truck.

traffic lane where the rear vehicle drives, and W_{rear} is the GVW of the rear vehicle.

It is important to highlight that using (16) to calculate the weight of the front vehicle in the one-by-one vehicle queue is not applicable to circumstances when the rear vehicle enters the bridge before the front vehicle passes the instrumented bridge cross-section, because S^{peak} in (16) involves the effects of the rear vehicle under such circumstances. Fortunately, this problem does not exist in this research; for a sizeable safety margin, no less than 30 m, between front and rear vehicles, is demanded when driving on highways in China. The distance between the instrumented cross-section and the start point of the bridge, however, is only 16 m.

Challenge arises when two vehicles driving side by side, however. In this scenario, one strain signal peak corresponds to two indistinguishable vehicles, which makes the above GVW calculation methods ineffective.

Finally, a segment of 15 minutes' strain signal and video when there are no side-by-side trucks is analyzed. Cars are ignored and weights of a total of 61 trucks are calculated. Statistics of the relative errors compared with the results recognized by the pavement-based WIM system are listed in Table 2. Plots of the GVW results of the six sensors S1~S6 are also shown in Figure 15, in which each point corresponds to a vehicle. In this figure, the further away the point is from the baseline, the larger the error is.

According to the GVW calculation results, though errors of several vehicles are unpleasantly significant, accuracy of the rest is still acceptable, except results based on strain sensors named S1 and S4. Close distance to the section neutral axis of S1 and S4 explains their significant errors. Because, under the plane section assumption, the closer the strain sensor is to the neutral axis, the smaller its strain value, making the relative error larger in contrast. To avoid this problem, BWIM sensors should be installed far from the section neutral axis for higher accuracy.

4.5. Vehicle Velocity Calculation. Theoretically, instantaneous velocities of vehicles can be calculated through (18) with the recognized vehicle position in each video frame and fixed time interval between frames.

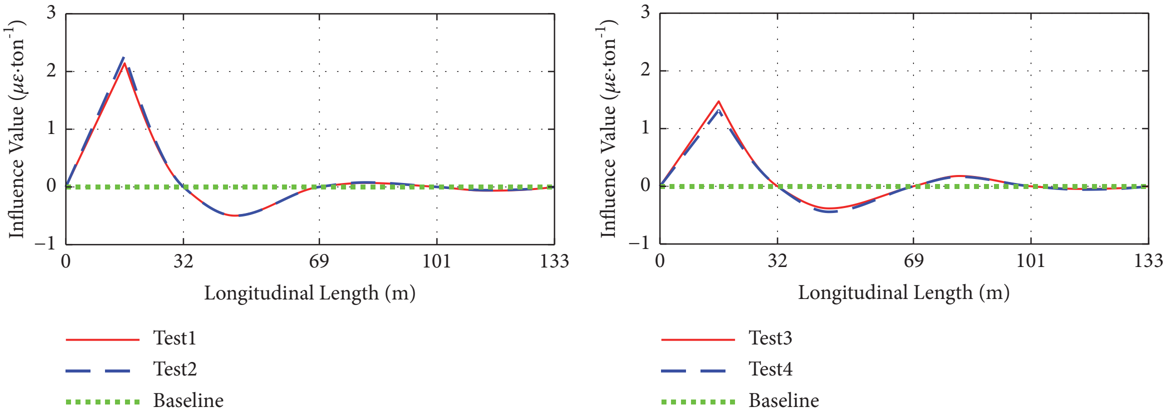
$$v = \frac{\Delta S}{\Delta t} \tag{18}$$

where v is the vehicle velocity and ΔS is the vehicle displacement within a period of time Δt.

However, calculated instantaneous velocities of vehicles appear to fluctuate drastically. Average vehicle velocity in three seconds, which means $\Delta t = 3$s, is calculated instead of instantaneously, and the calculation results are quite accurate as shown in Figure 16. The mean value of errors is -0.8%, the standard deviation of errors is 9.2%, and the maximum value of errors is 23.1%.

4.6. Vehicle Type and Axle Recognition. Identification of closely spaced axles, including tandem axles, is a key factor to ensure accurate classification of the passing vehicles. Real-time traffic characterization on a bridge is beneficial for asset managers and bridge owners because it provides statistical data about the configurations of the passing vehicles. The nothing-on-road (NOR) technique is generally utilized to obtain the information about the axles with sensors located underneath the bridge girder [36, 37].

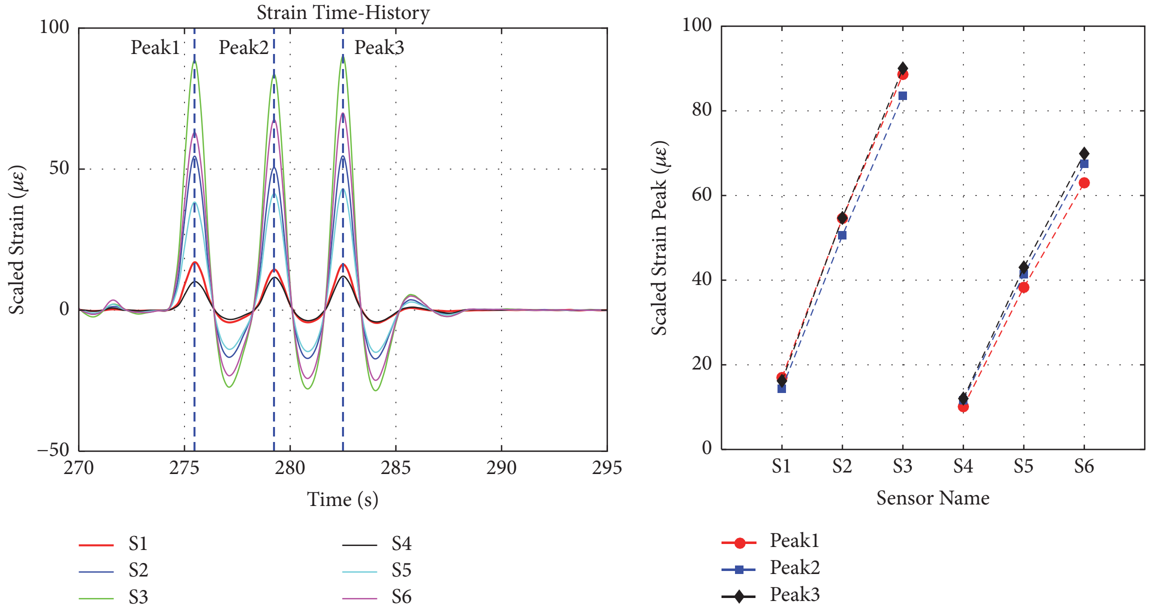
As a supplement, this paper obtains information about vehicle type and number of axles with visual information provided by only a webcam. Figures 6(a) and 6(c) show that the well-trained YOLO V3 algorithm is capable of directly recognizing vehicle types and the number of axles similarly to humans. The vehicle type recognition accuracy is 100% and the axle recognition results of 61 trucks (including 6 trucks with 2 axles) and 50 cars in the field tests are shown in Figure 17, which is still quite satisfactory compared with the pavement-based WIM. Errors are inevitable because of vehicle overlap, limited visual angle of the webcam, and illumination conditions. For instance, if cars are obscured by trucks with large size or the illumination is rather dim, wheels of cars will thus not be recognized. For instance, if cars are obscured by trucks with larger size or the illumination is rather dim, cars wheels will thus not be recognized. That is the reason why the computer vision made mistakes. This problem



(a) Results of test1 and test2

(b) Results of test3 and test4

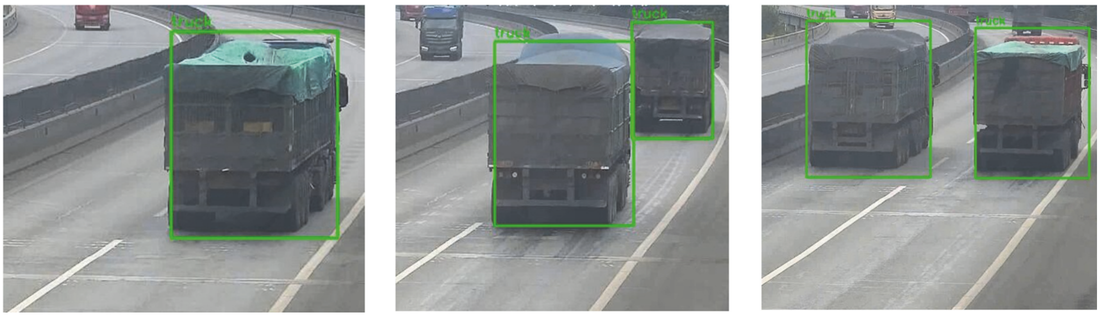
Figure 12: Influence line calibration results.



(a) Processed strain time-history curve

(b) Values of peaks

Figure 13: Strain data processing results.

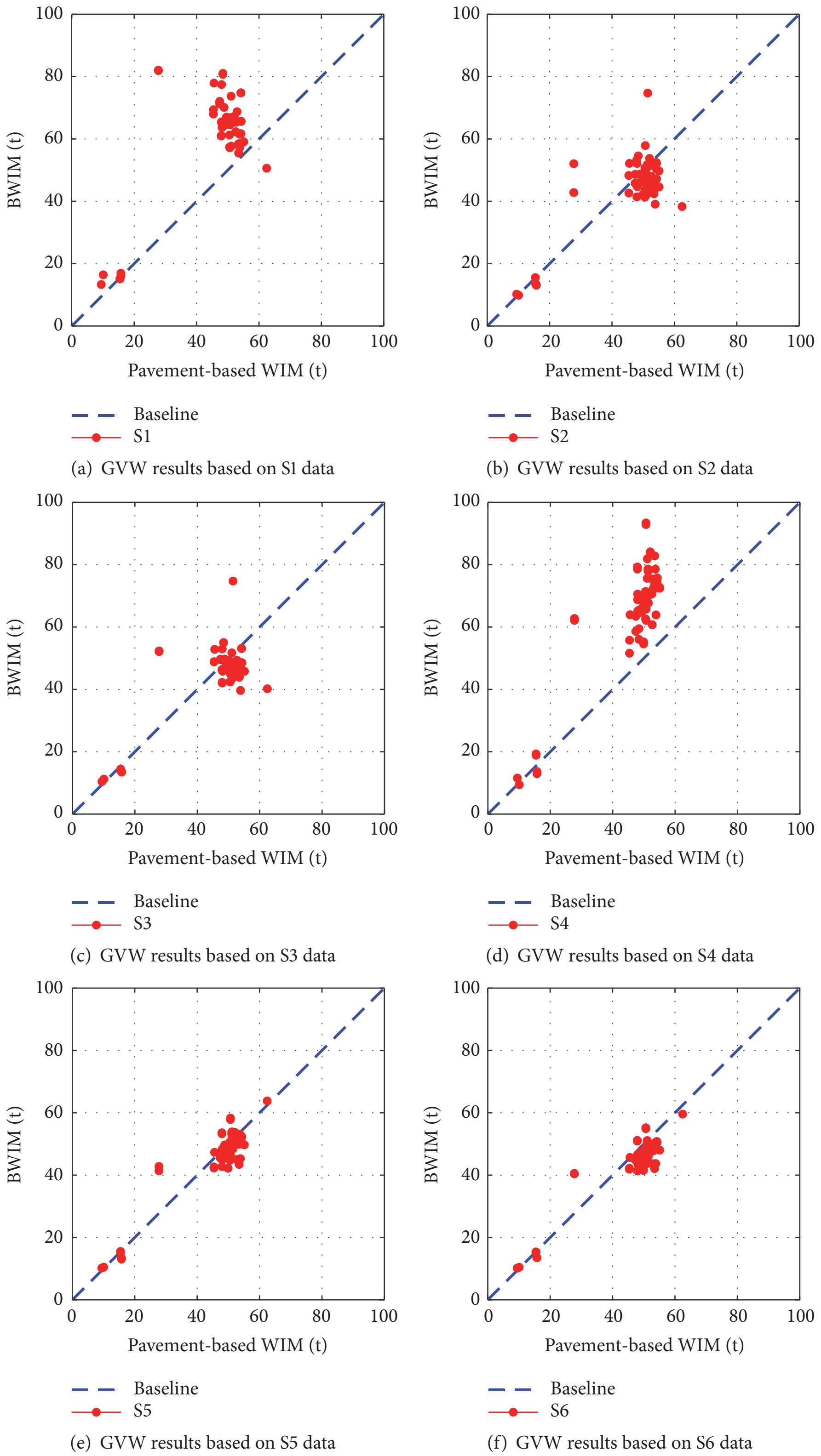


(a) Single vehicle

(b) One-by-one vehicles

(c) Side-by-side vehicles

Figure 14: Scenarios of vehicle distribution on bridge.



(a) GVW results based on S1 data

(b) GVW results based on S2 data

(c) GVW results based on S3 data

(d) GVW results based on S4 data

(e) GVW results based on S5 data

(f) GVW results based on S6 data

Figure 15: GVW calculation results for the six strain sensors S1~S6.

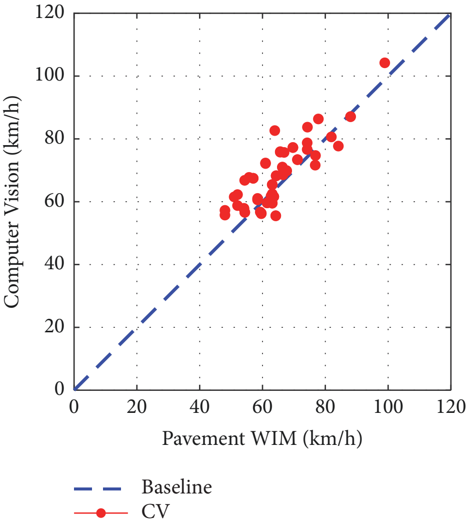


FIGURE 16: Vehicle velocity results.

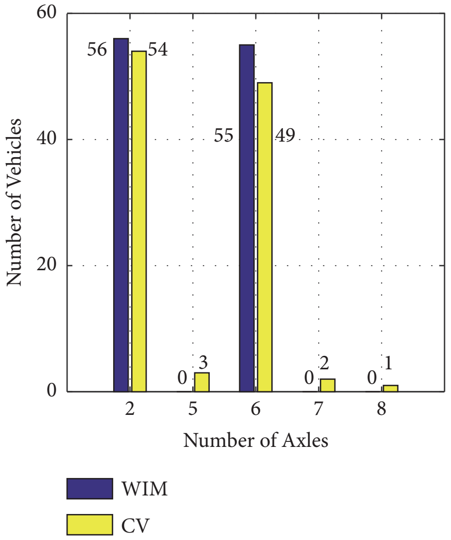


FIGURE 17: Axle recognition results.

did not appear in the pavement-based WIM as illustrated in Figure 17.

To prevent these errors, the visual angle of the webcam can be adjusted to observe the vehicle wheels more clearly. Another way to improve the quality of the vision is using an infrared camera to prevent dim illumination.

4.7. Error Analysis. Although the recognition accuracy is acceptable, error analysis is imperative for further improvement. To the author's knowledge, the following two reasons may account for calculation errors illustrated in Figure 18: (i) vehicle deviation from the traffic lane and (ii) vehicle positioning errors of the computer vision technique.

It is noted that vehicles on a bridge do not drive on the traffic lane strictly in some cases, but, in this research, only influence lines of traffic lanes are utilized. This assumption leads to significant errors when the vehicle deviates from the traffic lane severely as presented in Figure 18(a). To reduce this kind of error, the influence line can be substituted by the influence surface.

Another error source is inaccurate vehicle detection as shown in Figure 18(b), where multiple vehicles overlap in the image and leads to positioning errors. Particular labelling aiming at this phenomenon and diversifying the training sets for the deep neural network will help to mitigate the problem.

5. Conclusions

A traffic sensing methodology has been proposed in this paper in combination with influence line theory and computer vision technique. Field tests were conducted to evaluate the proposed methodology in various aspects. The main conclusions of this work might be listed as follows:

(1) The identification of vehicle positions, especially on transverse direction when passing a bridge, is quite critical to solve multiple-vehicle problem for BWIM systems. This paper introduces, for the first time, deep learning based computer vision technique to obtain the exact position of vehicles on bridges and successfully solves one-by-one vehicles scenario of multiple-vehicle problems for BWIM research with an average weighing error within 5%.

(2) The time series smoothing algorithm, LOWESS, is an effective tool to extract static component from directly measured bridge responses. Then, influence line or influence surface of a real bridge can be easily calibrated for BWIM purpose.

(3) Verified by field tests, the deep learning based computer vision technique is highly stable and efficient to recognize vehicles on bridges in real time manner. Therefore, it is proven to be a promising technique for traffic sensing.

(4) The proposed traffic sensing methodology is capable of identifying vehicle weight, velocity, type, axle number, and time-spatial distribution on small and medium span girder bridges in a cost-effective way, especially for those bridges already equipped with structure health monitoring systems and surveillance cameras.

Data Availability

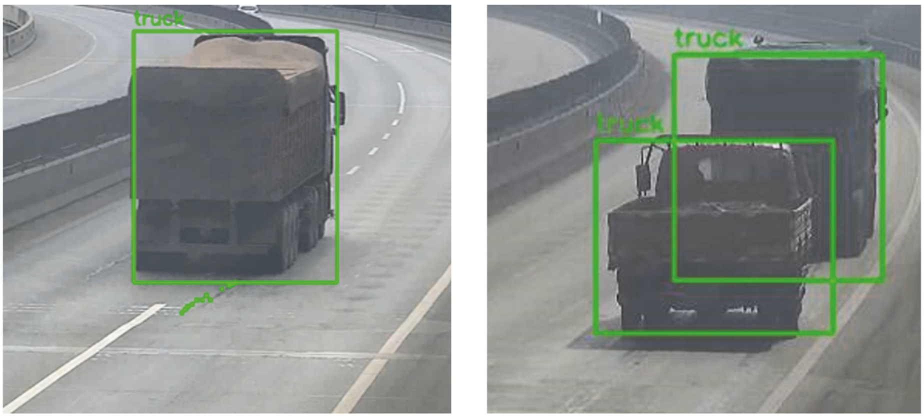
The video and testing data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper is supported by the National Key R&D Program of China (2017YFC1500605), Science and Technology Commission of Shanghai Municipality (17DZ1204300 and



(a) Vehicle deviation from traffic lane

(b) Inaccurate vehicle detection

Figure 18: Major error sources.

18DZ1201203), and the Fundamental Research Funds for the Central Universities.

References

[1] R. Zaurin and F. N. Catbas, "Integration of computer imaging and sensor data for structural health monitoring of bridges," *Smart Materials and Structures*, vol. 19, no. 1, Article ID 015019, 2009.

[2] B. Wei, C. Zuo, X. He, L. Jiang, and T. Wang, "Effects of vertical ground motions on seismic vulnerabilities of a continuous track-bridge system of high-speed railway," *Soil Dynamics and Earthquake Engineering*, vol. 115, pp. 281–290, 2018.

[3] B. Wei, T. Yang, L. Jiang, and X. He, "Effects of uncertain characteristic periods of ground motions on seismic vulnerabilities of a continuous track–bridge system of high-speed railway," *Bulletin of Earthquake Engineering*, vol. 16, no. 9, pp. 3739–3769, 2018.

[4] Y. Yu, C. S. Cai, and L. Deng, "State-of-the-art review on bridge weigh-in-motion technology," *Advances in Structural Engineering*, vol. 19, no. 9, pp. 1514–1530, 2016.

[5] M. Lydon, S. E. Taylor, D. Robinson, A. Mufti, and E. J. O. Brien, "Recent developments in bridge weigh in motion (B-WIM)," *Journal of Civil Structural Health Monitoring*, vol. 6, no. 1, pp. 69–81, 2016.

[6] F. Moses, "Weigh-in-motion system using instrumented bridges," *Journal of Transportation Engineering*, vol. 105, no. 3, 1979.

[7] J. Zhang, Y. Lu, Z. Lu, C. Liu, G. Sun, and Z. Li, "A new smart traffic monitoring method using embedded cement-based piezoelectric sensors," *Smart Materials and Structures*, vol. 24, no. 2, Article ID 025023, 2015.

[8] H. Zhao, N. Uddin, X. Shao, P. Zhu, and C. Tan, "Field-calibrated influence lines for improved axle weight identification with a bridge weigh-in-motion system," *Structure and Infrastructure Engineering*, vol. 11, no. 6, pp. 721–743, 2015.

[9] C. D. Pan, L. Yu, and H. L. Liu, "Identification of moving vehicle forces on bridge structures via moving average Tikhonov regularization," *Smart Materials and Structures*, vol. 26, no. 8, Article ID 085041, 2017.

[10] Z. Chen, T. H. T. Chan, and A. Nguyen, "Moving force identification based on modified preconditioned conjugate gradient method," *Journal of Sound and Vibration*, vol. 423, pp. 100–117, 2018.

[11] C.-D. Pan, L. Yu, H.-L. Liu, Z.-P. Chen, and W.-F. Luo, "Moving force identification based on redundant concatenated dictionary and weighted l1-norm regularization," *Mechanical Systems and Signal Processing*, vol. 98, pp. 32–49, 2018.

[12] H. Sekiya, K. Kubota, and C. Miki, "Simplified portable bridge weigh-in-motion system using accelerometers," *Journal of Bridge Engineering*, vol. 23, no. 1, Article ID 04017124, 2017.

[13] X. Q. Zhu and S. S. Law, "Recent developments in inverse problems of vehiclebridge interaction dynamics," *Journal of Civil Structural Health Monitoring*, vol. 6, no. 1, pp. 107–128, 2016.

[14] F. Schmidt, B. Jacob, C. Servant, and Y. Marchadour, "Experimentation of a bridge WIM system in France and applications for bridge monitoring and overload detection," in *Proceedings of the International Conference on Weigh-In-Motion ICWIM6*, p. 8p, 2012.

[15] R. E. Snyder and F. Moses, "Application of in-motion weighing using instrumented bridges," *Transportation Research Record*, vol. 1048, pp. 83–88, 1985.

[16] Z.-G. Xiao, K. Yamada, J. Inoue, and K. Yamaguchi, "Measurement of truck axle weights by instrumenting longitudinal ribs of orthotropic bridge," *Journal of Bridge Engineering*, vol. 11, no. 5, pp. 526–532, 2006.

[17] E. Yamaguchi, S.-I. Kawamura, K. Matuso, Y. Matsuki, and Y. Naito, "Bridge-weigh-in-motion by two-span continuous bridge with skew and heavy-truck flow in Fukuoka area, Japan," *Advances in Structural Engineering*, vol. 12, no. 1, pp. 115–125, 2009.

[18] Y. Yu, C. S. Cai, and L. Deng, "Nothing-on-road bridge weigh-in-motion considering the transverse position of the vehicle," *Structure and Infrastructure Engineering*, vol. 14, no. 8, pp. 1108–1122, 2018.

[19] Z. Chen, H. Li, Y. Bao, N. Li, and Y. Jin, "Identification of spatio-temporal distribution of vehicle loads on long-span bridges using computer vision technology," *Structural Control and Health Monitoring*, vol. 23, no. 3, pp. 517–534, 2016.

[20] J. F. Frenze, *A Video-Based Method for the Detection of Truck Axles*, National Institute for Advanced Transportation Technology, University of Idaho, 2002.

[21] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988.

[23] D. A. Freedman, *Statistical Models: Theory and Practice*, Cambridge University Press, 2009.

[24] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

[25] A. Znidaric and W. Baumgartner, "Bridge weigh-in-motion systems-an overview," in *Proceedings of the Second European Conference on Weigh-In-Motion of Road Vehicles*, Lisbon, Portugal, 1998.

[26] P. McNulty and E. J. O'Brien, "Testing of bridge weigh-in-motion system in a sub-arctic climate," *Journal of Testing and Evaluation* , vol. 31, no. 6, pp. 497–506, 2003.

[27] E. J. OBrien, M. J. Quilligan, and R. Karoumi, "Calculating an influence line from direct measurements," *Proceedings of the Institution of Civil Engineers: Bridge Engineering*, vol. 159, no. BE1, pp. 31–34, 2006.

[28] S. P. Timoshenko and D. H. Young, *Theory of Structures*, McGraw-Hill, 1965.

[29] S. P. Timoshenko and J. M. Gere, *Mechanics of Materials*, van Nordstrand Reinhold Company, New York, NY, USA.

[30] F. Moses, "Instrumentation for weighing truck-in-motion for highway bridge loads," Final Report FHWA-OH-83-001, 1983.

[31] Y. L. Ding, H. W. Zhao, and A. Q. Li, "Temperature effects on strain influence lines and dynamic load factors in a steel-truss arch railway bridge using adaptive FIR filtering," *Journal of Performance of Constructed Facilities*, vol. 31, no. 4, Article ID 04017024, 2017.

[32] M. Quilligan, *Bridge weigh-in motion: development of a 2-D multi-vehicle algorithm [Doctoral thesis]*, Byggvetenskap, 2003.

[33] J. Schmidhuber, "Deep learning in neural networks: an overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[34] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[35] G. Xu and Z. Zhang, *Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach*, Springer, 1996.

[36] H. Kalhori, M. M. Alamdari, X. Zhu, B. Samali, and S. Mustapha, "Non-intrusive schemes for speed and axle identification in bridge-weigh-in-motion systems," *Measurement Science and Technology*, vol. 28, no. 2, Article ID 025102, 2017.

[37] H. Kalhori, M. M. Alamdari, X. Zhu, and B. Samali, "Nothing-on-road axle detection strategies in bridge-weigh-in-motion for a cable-stayed bridge: case study," *Journal of Bridge Engineering*, vol. 23, no. 8, Article ID 05018006, 2018.