Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 20(2): 125-130, Jun. 2022 Regular paper

Filtering Correction Method and Performance Comparison for Time Series Data

Jongwoo Baek, Jiyoung Choi, and Hoekyung Jung*, *Member*, *KIICE*

Department of Computer Engineering, PaiChai University, Daejeon 35345, South Korea

Abstract

In modern society, as many data are used for research or commercial purposes, the value of data is gradually increasing. In related fields, research is being actively conducted to collect valuable data, but it is difficult to collect proper data because the value of collection is determined according to the performance of existing sensors. To solve this problem, a method to effectively reduce noise has been proposed, but there is a point in which performance is degraded due to damage caused by noise. In this paper, a device capable of collecting time series data was designed to correct such data noise, and a correction technique was performed by giving an error value based on the representatively collected ultrafine dust data, and then comparing before and after Compare performance. For the correction method, Kalman, LPF, Savitzky-Golay, and Moving Average filter were used. Savitzky-Golay filter and Moving Average Filter showed excellent correction rate as an experiment. Through this, the performance of the sensor can be supplemented and it is expected that data can be effectively collected.

Index Terms: Acquisition, Correction, Comparative analysis, Noise, Time series data

I. INTRODUCTION

In recent years, the seriousness of environmental pollution due to industrialization is on the rise, and safety accidents due to air pollution occur frequently. Among air pollution, fine dust represents a dust diameter of 10 μm or less, and ultrafine dust represents a dust diameter of 2.5 μm or less and is also referred to as PM10 and PM2.5. Fine dust enters the body and penetrates deeply into the alveoli, causing various respiratory diseases, and since it is invisible to the naked eye, most of them are not regarded as a major cause, causing pain later [1]. In order to prevent safety accidents, these time series data are monitored in real time through devices with good performance in each country, but other devices for research are devices that allow individuals to purchase sensors and measure air quality, including outdoors and indoors. In addition to the need to install data, it is difficult to obtain accurate measurement values because error values may occur depending on constraints such as sensor performance and environment [2].

As a related research, there are cases in which a filtering algorithm has been developed to correct the error value of time series data in which observations have a temporal order. Case A uses kNN algorithm and Fuzzy Clustering Mean algorithm to improve harmonic detection sensitivity and signal-to-noise ratio, and to reduce noise, a method that can reduce noise effects using a limited RBM-based classifier is proposed. The causes of noise in this study are largely divided into internal and external background noise of the classifier, and errors are included in the data collected by the detector, which degrades the performance of the NLJD (Nonlinear junction detector) system, and results from damage caused by noise. There was a problem

Received 21 July 2021, **Revised** 06 August 2021, **Accepted** 09 August 2021
***Corresponding Author** Hoekyung Jung (E-mail:hkjung@pcu.ac.kr, Tel:+82-42-520-5640)
Department of Computer Engineering, PaiChai University, Daejeon 35345, South Korea.

Open Access https://doi.org/10.6109/jicce.2022.20.2.125 print ISSN: 2234-8255 online ISSN: 2234-8883

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © The Korea Institute of Information and Communication Engineering

that the model performance was degraded due to difficulty in extraction [3]. Case B proposes a filtering technique that efficiently processes continuous data in a wireless network environment. Existing filtering is a technique that creates and applies a filter, but since this technique is not good in terms of energy efficiency, a technique that can significantly reduce data by creating a filter and applying the filter range to the entire sensor was proposed [4-6].

In this paper, an experiment was performed using data of representative ultrafine dust to correct the error value of time series data. For the data, different devices A and B with fine dust sensors were manufactured, and the values observed indoors for a certain period were used, and the observed data of device B was randomly given an error value, followed by a filtering correction technique. The resulting value was compared and analyzed with Device A. There are a total of 4 correction techniques, and the performance comparison is expressed through correlation coefficients.

II. RELATED WORKS

This chapter describes filtering techniques and data purification and shows the contents of correlation analysis for comparison.

A. Data Cleaning

In the data analysis, only necessary values of the observed data are extracted and analyzed after purification. Data cleansing is an important step for data analysis. Missing values must be removed but required data can be prevented from missing and analysis is possible. Missing values are called NA or Not Available, and there are three types of missing values: Completely Random Missing (MCAR), Random Missing (MAR), and Non-random Missing (MNAR) [7]. These missing values can cause a variety of problems and, firstly, incorrect substitutions can lead to bias within the observed data. Second, in the process of processing missing values, the opinion of the analyst may be reflected, and the direction of the analysis may be changed. Third, if all missing values are removed, an error may occur in the analysis and data loss may occur. This data purification technology is also used in news, SNS (Social Network Service), and blogs. Recently, data can be collected using the Retweet function or the fetch function, but there may be duplicate data among the collected data. Since such redundant data requires a large storage space and takes a lot of time to analyze, the reliability of the results can be lowered. As a related research, text data is expressed as a feature vector, and document vectors and Shingles, which can detect duplicate data by measuring the similarity between vectors, have been studied. Min-Hash, Simhash, Mod-p Shingles, etc. that can be detected have been studied [7].

B. Data filtering

Data filtering is a data processing technique that smoothes high-frequency fluctuations included in data and helps to remove trends periodically for a specific frequency of data. Typical filtering techniques are Gaussian Filter, Median Filter, Bilateral Filter, and Kalman Filter. Gaussian Filter replaces the current pixel value with the current pixel value using the weighted average of the current pixel value and surrounding values, and the Median Filter refers to the center value by replacing each pixel value with the median value of the neighboring value. In addition, the Bilateral Filter is known as a filter that has been developed to compensate for the parts of the Gaussian Filter whose edge changes and is a nonlinear filter that removes noise while preserving the edge. Bilateral Filter replaces each pixel value with a weighted average of surrounding pixel values. The last Kalman Filter was developed by Rudolf E. Kalman as a recursive filter that estimates linear dynamics based on measurements including noise. Also, Kalman Filter estimates the current joint distribution based on past measured values, and the definition of Kalman Filter is organized as follows [8].

$$\overline{x_k} = A\bar{x}_{k-1}$$
$$\overline{P_k} = AP_{k-1}A^T + Q. \quad (1)$$

x_k : state
z_k : Measures
A : State matrix
w_k : State matrix
v_k : Measurement noise
Q : The larger it is, the greater the influence of the measured value. Diagonal matrix.
R : The larger it is, the less affected by the measured value. Constant value.

Initial values are determined by r_0, P_0, and P_0 means a diagonal matrix of size $n \times n$ if the size of state x_k is $n \times 1$. Then, the estimated value and the error covariance are predicted through eq. (1) through the above values, and finally filtering is performed through calculation of Kalman gain and estimated value [9].

C. Correlation Analysis

Correlation analysis refers to a method of analyzing whether there is a nonlinear or linear relationship between two variables. At this time, the two variables may be a

relationship that is correlated with each other, and a relationship between the two variables is called a correlation. Unilaterally, the correlation between 0.1 and -0.1 is said to be an almost negligible linear relationship, and between 0.1 and 0.3, the two variables are said to have weak quantitative linear relationships, and between 0.3 and 0.7, the two variables have a distinct quantitative linear relationship. Can be seen. And if it is between 0.7 and 1.0, it is said that the two variables are a very strong quantitative linear relationship. If the values of the two variables are negative, it can be seen that the relationship is negative, not positive. Thus, there is a Pearson correlation coefficient, a method that is used unilaterally to know the relationship between the two variables. This indicates that, assuming that the values of the variables X and Y exist, the coefficient r is the degree to which X and Y change at the same time as the variables X and Y change. The value of r obtained through the Pearson correlation coefficient method has a value of +1 if the variables X and Y are completely identical, has a value of -1 if the variables X and Y are completely identical, and has a value of -1 if they are completely different[9].

III. SYSTEM DESIGN

This chapter deals with the overall system configuration and design for analysis. Content composition describes the system configuration and device design, refinement flow and algorithm, and represents the step-by-step process.

A. System configuration diagram

The system configuration for filtering correction and analysis is shown in Fig. 1.

Devices for data collection consist of a WI-Fi module and an ultra-fine dust sensor, and the device transmits ultra-fine dust data to the Java Development Kit (JDK) through WI-Fi. Here, the voluntary development kit grants permission to Cloud Build and integrates with App Engine in Google Cloud Platform. When the linking is performed, the device is activated and data can be checked through the Datastore. In addition, the collected ultrafine dust data is separately exported to a bucket and saved as a .CSV file through BigQuery. After data purification is performed to facilitate analysis of the .CSV file storing the last collected ultrafine dust data, data correction is performed, and the corrected ultrafine dust data is analyzed.

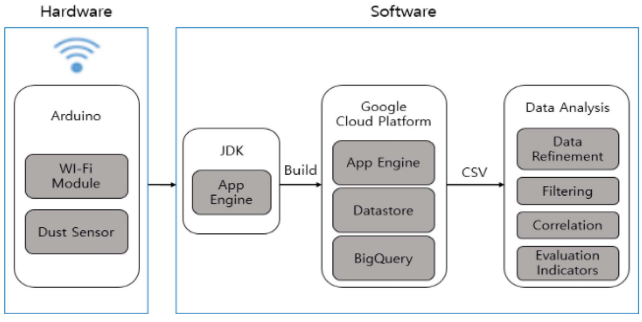

Fig. 1. Whole system configuration

B. Data Cleaning Design

Since the ultrafine dust data released through the device contains error data, smooth analysis is possible only after data purification. The following Fig. 2 shows the data cleaning flow chart.

The flow chart in Fig. 2 proceeds through the collected ultrafine dust data. Among the first raw data, all duplicate data due to errors must be removed. This is because it can be a hindrance factor in future correction techniques or correlation analysis. Second, the collected values must be converted into a form suitable for analysis. In addition, the data was checked for missing values, and all missing values were replaced with the previous values, and the data was refined by adding one variable representing the median value to be used for the last correction. In addition, Device B was refined by adding an arbitrary error value of 256 in units of 100 to compare the performance of the calibration. Therefore, Fig. 3 shows the refinement algorithm before adding the median value.

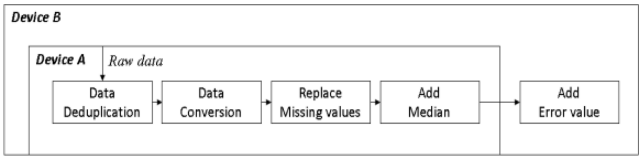

Fig. 2. Data cleaning flow chart.

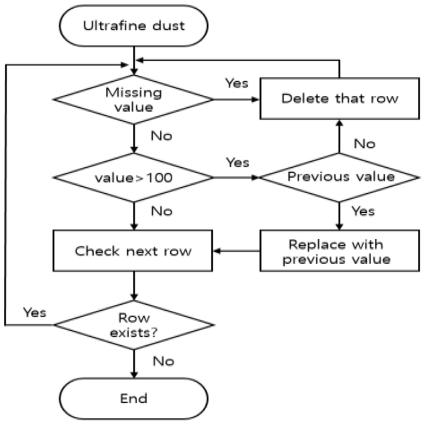

Fig. 3. Ultrafine dust purification algorithm.

Fig. 3 is an algorithm for data purification, and when the value of ultrafine dust is empty or exceeds 100, it is replaced with the previous value. At this time, if the previous value is empty or there is no value, the row is deleted and the next row value is refined. And check whether the refinement is done well for the next line. Refining repeats the above process until the end of the row, and if there are no rows, it ends.

IV. SYSTEM IMPLEMENT

The processor of the computer specifications used in this chapter is Intel i7-8700K, the system type is 64-bit operation and *64-based processor, Windows 10 Pro version and 16.0 GB RAM, but it is Colaboratory used in the implementation. The CPU using Ubuntu 18.04.5 LTS is the Xeon of the Intel series, the memory and drive used is 13333552 KB, 202 GB, and I used the 3.6.9 Python version.

A. Apply Filtering

In this section, data correction was performed using the ultra-fine dust data of device B. The correction technique was optimized by adjusting the Window Size and Polynomial Order in the filtering settings using each Kalman Filter, LPF Filter, Savitzky-Golay Filter, and Moving Average Filter. Fig. 4 shows the values with the Kalman filter applied. Zoom in and expand a portion (2,000) of the entire (10,000) data for precise application.

The Kalman Filter in Fig. 4 shows that all error values have been corrected to a moderate degree, and the remaining normal values have also been smoothed.

The values of the Moving Average Filter and LPF Filter are relatively stable compared to the Kalman Filter and have been corrected. Moving average filter shows that all error values are filtered inside 50, and LPT filter shows that the correction is not better than moving average filter, but the correction is better than that of the Kalman Filter.

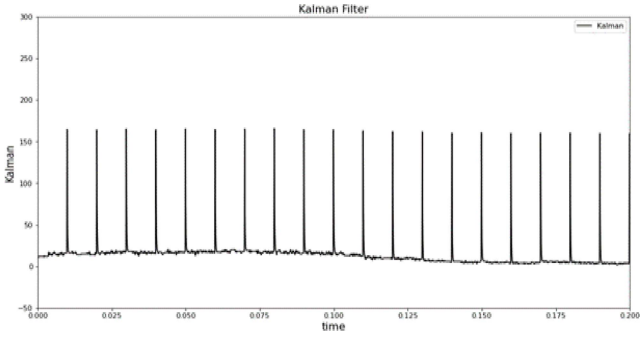

Fig. 4. Kalman filter.

The last Fig. 5 shows the Savitzky–Golay Filter correction applied value. It can be seen that the correction has been made to negative values little by little because of the excessive correction response to the error value, but it can be seen that the correction has generally been performed for the error values.

B. Comparing Instrument B and Calibration Data

In this section, the ultrafine dust data using the correction technique and the device B before correction were compared. The comparison target was the Savitzky-Golay and Moving Average Filter correction techniques that were stably corrected, and the device B before correction was displayed as a figure in consideration of the point that was displayed as a single line. Fig. 6 shows the compari-

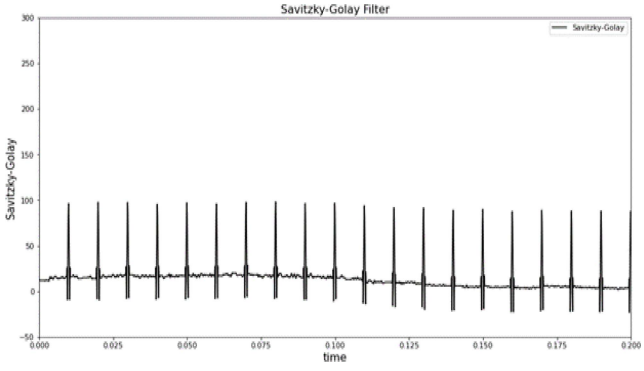

Fig. 5. Savitzky–Golay filter.

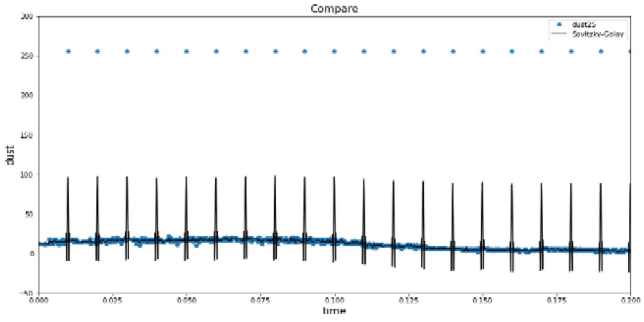

Fig. 6. Device b and the savitzky-golay filter comparison.

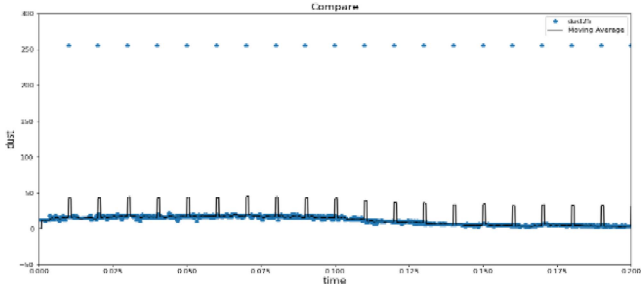

Fig. 7. Device b and the moving average filter comparison.

son of device B's data and Savitzky-Golay Filter, and Fig. 7 shows the comparison of Moving Average Filter.

C. Correlation Comparison Analysis

In this experiment, the data that had been calibrated to represent the correlation coefficient and the normal time series data device A were averaged and stored in units of 500. Table 1 shows the calibration data and the average of 5 instrument A.

Table 1 is data for extracting the correlation coefficient, and is composed of reference data, the ultrafine dust data of device A, and data to which the correction technique is applied. Based on this, Table 2 shows the correlations for the data.

The correlation coefficient in Table 2 shows a slightly different result from the graph shown with the naked eye in the 4.2 experiment. The correction technique for the actual data device A Kalman Filter, Moving Average Filter, Savitzky-Golay Filter, and LPF Filter all show a clear quantitative linear relationship, of which Savitzky-Golay Filter has the highest linear relationship with 0.9507. This means that the correction became smooth even for the actual ultrafine dust measurement values excluding the error value. Table 3 below shows the evaluation of MAE for the correction values.

Table 3 indicates an absolute value for error, and the closer the value is to 0, the better the evaluation value. As a result of MAE, Kalman, Moving Average, LPF, and Savitzky-Golay Filter all showed satisfactory values as their values were close to 0. Among them, Savitzky-Golay Filter showed the best evaluation value. MSE is the most commonly intuitive and is a method of calculating the mean by squared error values. Likewise for this evaluation, the closer the value is to 0, the better the evaluation value is. The results of MSE also showed good evaluation values in all correction techniques, and the same as MAE, Savitzky-Golay Filter was 1.19757, indicating that the evaluation value was the best. Unlike MSE, RMSE more intuitively represents the average of the error values themselves. In this evaluation, the smaller the value, the better the evaluation value, and the Savitzky-Golay Filter showed the highest evaluation value with similar values to the previous evaluation as a result of RMSE. The last Fig. 8 shows the average value of the ultrafine dust data of the Savitzky-Golay Filter, which showed the highest correction rate, and the reference device A, in one graph. The x-axis of Fig. 8 represents the time and the y-axis represents the observed value, and when looking at the graph, the Savitzky-Golay Filter correction method for the error value as an overall average shows a graph similar to that of the Reference, and shows excellent filtering correction by catching high error values.

Table 1. Instrument a and calibration data average

Ultrafine dust	Kalman	Moving average	LPF	Savitzky-Golay
18.312	18.22876	17.6578	18.13117	18.2853
19.542	19.54242	19.55667	19.54317	19.54211
11.532	11.54201	11.60844	11.54743	11.53192
6.938	6.940824	6.960444	6.942438	6.936844
7.188	7.185993	7.166667	7.184621	7.188275

Table 2. Correlation to data

	Ultrafine dust	Kalman	Moving average	LPF	Savitzky-Golay
Ultrafine dust	1	0.9503	0.9452	0.9496	0.9507
Kalman	0.9503	1	0.9995	0.9999	0.9999
Moving average	0.9452	0.9995	1	0.9996	0.9994
LPF	0.9496	0.9999	0.9996	1	0.9999
Savitzky-Golay	0.9507	0.9999	0.9994	0.9999	1

Table 3. MAE/MSE/RMSE evaluation index

Filtering Type	MAE	MSE	RMSE
Kalman Filter	0.8629712	1.2074798	1.0988538
Moving Average Filter	0.90861124	1.3268273	1.1518799
LPF Filter	0.8699039	1.2239242	1.1063111
Savitzky-Golay Filter	0.85859567	1.19757	1.0943354

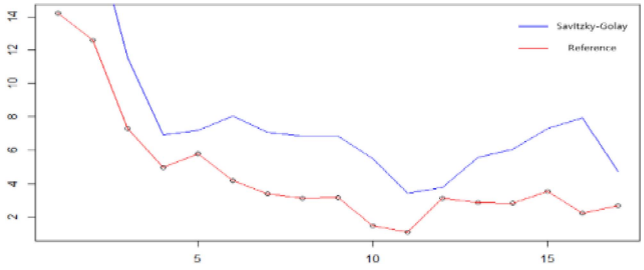

Fig. 8. Average of data from savitzky-golay filter and device a.

V. CONCLUSIONS

In modern society, real-time monitoring is provided nationwide to prevent safety accidents. However, there were limitations in the results and data provided, and sensors used for various studies or personal safety are also difficult to make devices themselves, and due to limitations in performance, error values occur for time-series data collecting measured values for a long time. There was a ship. Therefore, in this paper, in order to effectively correct the error value of the time series data, representatively, the experiment was conducted after collecting data of ultrafine dust. Kalman Filter, LPF Filter, Savitzky-Golay Filter, and Moving Average Filter were used as cor-

rection techniques, and the result of the correction was visualized by displaying a graph. As a result of comparing four types of correction techniques, the Moving Average Filter and Savitzky-Golay Filter showed excellent correction rates. As a result of the comparison, a high error value was caught, and the graph result of a curve similar to the reference was shown.

In future research, more devices should be installed to study the calibration method for multiple sensors. As a result, sensor performance for time series data can be supplemented, and it is expected that accurate data can be collected and contributed to various studies.

ACKNOWLEDGEMENTS

This work was supported by the PaiChai University research grantin 2022.

REFERENCES

[1] J. W. Lee, G. Wu, and H. Jung, "Deep learning document analysis system based on keyword frequency and section centrality analysis," *Journal of Information and Communication Convergence Engineering*, vol. 19, no. 1, pp. 48-53, Mar. 2021. DOI: 10.6109/jicce.2021.19.1.48.

[2] C. H. Hwang, H. Kim, and H. Jung, "Detection and correction method of erroneous data using quantile pattern and LSTM," *Journal of information and communication convergence engineering*, vol. 16, no. 4, pp. 242-247, Dec. 2018. DOI: 10.6109/jicce.2018.16.4.242.

[3] B. Lim, H. Lim, S. Hong, and H. Jung, "Air pollution monitoring based on Bonferroni Multi-analysis," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 24, no. 8, pp. 963-969, Aug. 2020. DOI: 10.6109/jkiice.2020.24.8.963.

[4] S. W. Byeon and S. -Y. Kim, "Has air pollution concentration increased over the past 17 years in Seoul, South Korea? : The gap between public perception and measurement data," *Journal of Korean Society for Atmospheric Environment*, vol. 36, no. 2, pp. 240-248, Apr. 2020. DOI: http://doi.org/10.5572/KOSAE.2020.36.2.240.

[5] C. -J. Lee, M. -S. Hong, "Spatiotemporal variations of fine particulates in and around the Korean Peninsula," *Journal of Korean Society for Atmospheric Environment*, vol. 35, no. 6, pp. 675-682, Dec. 2019. DOI: 10.5572/KOSAE.2019.35.6.675.

[6] S. Lee, S. Park, and H. Jung, "Web monitoring based encryption web traffic attack detection system," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 25, no. 3, pp. 449-455, Mar. 2021. DOI: 10.6109/jkiice.2021.25.3.449.

[7] A. Chowdhury, O. Frieder, D. Grossman, and M. C. McCabe, "Collection statistics for fast duplicate document detection," *ACM Transactions on Information Systems*, vol. 20, no. 2, pp. 171-191, Apr. 2002. DOI: 10.1145/506309.506311.

[8] J. -Y. Choi, "A study on a effective information compressor algorithm for the variable environment variation using the Kalman Filter," *Journal of the Korean Society Of Computer And Information*, vol. 23, no. 4, pp. 65-70. Apr. 2018. DOI: 10.9708/jksci.2018.23.04.065.

[9] J. Bang, D. Hwang, and H. Jung, "Product recommendation system based on user purchase priority," *Journal of Information and Communication Convergence Engineering*, vol. 18, no. 1, pp. 55-60, Mar. 2020. DOI: 10.6109/jicce.2020.18.1.55.

Jongwoo Baek

Jong-Woo Baek majored in Computer Science at Pai Chai University in 2018. From 2019 to the present, he has been working as a researcher at the AI·SW Centered University Project Group at Pai Chai University. He is currently a Doctorate course in Department of Computer Engineering of Paichai University. His current research interests are artificial intelligence, big data, deep learning, and AR.

Jiyoung Choi

She received MS in communication engineering from KAIST in 2000. From 1991 to 1999, she worked at ETRI, researched and developed network management for transmission networks, and then took charge of wireless communication and vehicle communication development in ventures. Since 2010, she has served as an industry-academic adjunct professor at Pai Chai University. She is currently researching a method of measuring evacuation routes for dangerous situations based on artificial intelligence and image recognition and the field of big data for this

Hoekyung Jung

Hoekyung Jung received his M.S. and Ph. D. degrees in 1987 and 1993, respectively, from the Department of Computer Engineering, Kwangwoon University, Korea. From 1994 to 1995, he worked for ETRI as a researcher. Since 1994, he has worked in the Department of Computer Engineering at PaiChai University, where he now works as a professor. His current research interests include machine learning, IoT, big data, and artificial intelligence.