A Method of Gas Source Localization from Sensor Network using Machine Learning

Worachat Arunothaikrit[1] Ayumi Ohnishi[1] Shuhei Tsuchida[1] Tsutomu Terada[1]
Masahiko Tsukamoto[1]

Abstract: Gas source localization (GSL) is one of the most important tasks to find the origin of the gas source to avoid potential danger. GSL in natural conditions is a big challenge because it has many complex conditions, especially when the wind blows in an unpredictable direction. Mobile robots use a lot of energy to work and still have short working time. That makes using an immobile sensor is better, just placing it in the right place will lead to a longer lifetime and use less energy. However, finding the location of a gas leak is difficult both on relevant and irrelevant factors. Here we show that the location and distance of the gas emission source between the gas source to the station gas sensor array in the indoor environment. We found that the machine learning algorithm is applicable to localize our experiment gas source using standard performance metrics for the regression problem in machine learning: Mean absolute error (MAE) metric. Currently, an estimated position of the source with a deviation of 3.90 cm (93.1% using R-Squared) by using Random Forests Regression (RF regression). Our results show how the stationary sensor network tends to work in finding GSLs in an indoor environment using machine learning to find the distance between the gas and the sensor in natural wind conditions. We expect our experiment to be the starting point for bringing GSL to more complex forms, for example finding distance in multiple wind direction conditions and using it on a daily basis.

1. Introduction

Gas leak is very dangerous whether it is indoors, outdoors, factories or mines. We can easily know where the gas leak is if it is in a specific area where gas is used, for example in the kitchen. Then, if the gas pipe is broken in the construction area or in an area where there are so many gas pipes we that cannot find the source of the gas leak easily. An easy way to find the gas leak is to use a gas alarm or gas monitor to find a gas leak. These devices can warn and measure gas only, but they cannot specify the location of the gas leak. That makes finding the source of the gas leak difficult, slow, and dangerous.

Gas source localization is one of the most important tasks to find the origin of the gas source to avoid potential danger. When the gas is taken by the wind to the receptors, they are distributed, resulting in lower concentrations of substances from the source, making it difficult to find the source of those substances. Natural environments make gas diffusion random.

Currently, there are two main methods for finding source of the gas. The first is the methods using mobile robot types and the other is the methods using stationary sensor types. The mobile use of robots is a convenient way because the robot will travel to find clues to gas, but it will waste energy, short operation time and it will look cluttered if mobile robots are constantly running or flying in the area. The stationary sensor does not need to be moved since we can put it where there is a risk of gas leakage. That makes it not using a lot of energy and does not require frequent care. But immobile is also a disadvantage because the sensor will only receive gas depending on the wind.

Most of GSL research uses land-based wheeled robot [4] or Air-based: Micro-Drone [5]. They use mobile robots to track spatial gas distribution and find the location of the gas source. Many algorithms and patterns of moving for mobile robots were applied, such as a pseudo-gradient-based plume tracking algorithm and a particle filter-based algorithm. Stationary sensor research also has a similar method to find a gas source location. Both of them use gas concentration and wind velocity to estimate gas source location. This research not just interested in the sensor blocks in the area that we want to find because the sensor limitations will require a lot of sensors to create the territory, as seen in Bilgera et al.[1]. However, this research did not focuses on specifying the distance between the sensor and the gas. Qiuming Li el al.[3], They provide an another GSL algorithm using stationary sensor network that combine the weight centroid method and particle filter to address the location of gas source under unknown gas source strength and wind.

In this paper, we analyzed the use of machine learning algorithms for gas source localization (GSL) in indoor environments using data collected from the sensor array. Machine learning was used to predict the distance from the gas source and gas sensor array, the problem caused by indefin-

[1] Graduated School of Engineering, Kobe University

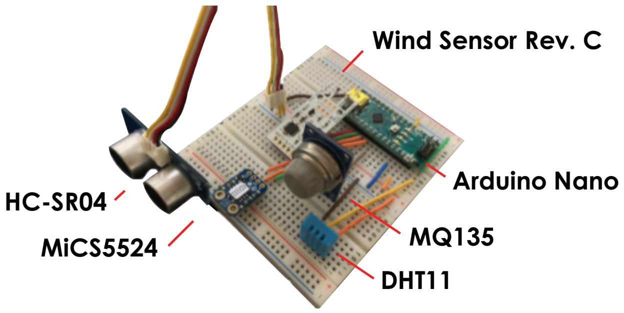

Fig. 1 Gas sensor array.

able wind airflow and gas distribution in an indoor environment. To train the machine learning model for GSL, we use data taken from a gas sensor array with 1 meter apart from the source origin in an indoor area.

2. Related researches

There are not many words on gas source localization. There are a variety of techniques used to find a location, whether it is to use sensors to find or to use mathematical simulations to find. Each of them has different advantages and disadvantages. Research has surveyed several techniques from the basics to use in this research to find gas source localization using sensors and machine learning.

2.1 Gas sensor

Gas sensors are a kind of small device that can transfer gas concentration information into electrical signals. We can easily classify the gas depending on its operation. The technology used for gas sensors consists of 4 main types: metal oxide semiconductors (MOS), electrochemical detection (EC), photoionization (PID), and infrared (IR). Most of these sensors are used with security, research, and the environment. The most common place that we can find this kind of sensor is building smoke alarms, environment monitoring, air quality measurements, or researches. In research, we can often find work in electronic-node and gas source localization problem. Gas sensors are designed to be cheap, commercially available, high sensitivity, low power devices, and new sensors technology are designed to enable onboard processing and internet-connected(cloud computing).

2.2 E-nose

The electronic nose (eNose, artificial olfaction) is electronic sensing that created to be similar to human olfaction for classification, comparison, and other applications with sensors array and pattern recognition.

2.3 Metal-oxide(MOX) gas sensor

The main concept of the MOX gas sensor is when a semiconductor gets heat, oxygen is absorbed on the particle surface by capturing free electrons so that no electric current flows through it. When reducing gas appears, the electrons released into the tin dioxide because reducing gases carry oxygen on the surface make oxygen decrease that allows electric currents to flow freely. It makes us monitor the value of the gas. The semiconductor is typically tin dioxide. MOX gas sensors perform different actions depending on the type of gas and the type of sensor.

The disadvantage of using a MOX gas sensor is that it takes time to heat before it is fully utilized. The distance between the gas and the sensor is limited, which depends on the type and size of the gas that affects the sensitivity of the sensor.

3. Preliminary Experiment

3.1 Experimental setting

To train the machine learning model for GSL, training and testing data were collected by wired sensor array with consist of air quality (AQ) sensor (MQ135, Hanwei Electronics Group Corporation), Carbon monoxide (CO), Alcohol and Volatile organic compounds (VOC) gas sensor (MiCS5524, Amphenol Corporation), temperature and humidity sensor (DHT11, Guangzhou Aosong Electronics Co., Ltd.), wind sensor (Wind Sensor Rev. C, Modern Device) and ultrasonic sensor (HC-SR04, ElecFreaks Technology Ltd.) as shown in Figure 1. The experiments were conducted in an indoor environment that is free of obstacles and allowed wind just from the window. The maximum distance between gas and sensor array is 1 meter, facing each other. Gas source generated randomly in a semicircle as see in Figure 2. Alcohol 70% is used as a gas source in the experiment. The average temperature is around 20-21 degrees Celsius. For the wind condition, we use only the wind that enters through the window, not from the wind generator as can be seen in Figure 3. Data is collected randomly from the point of gas source localization within 1-meter distance and collected at a sampling period of $\Delta t = 1$ second, resulting in data samples from gas sources localization.

Data was also collected by an automatic fan for comparison as well. All the setup for data collection in this experiment same condition as natural wind experimental. The only difference 2 things are using a fan to create the wind and data was collected on a different day. Since collecting data with a fan, it is necessary to close all windows before collecting data to prevent natural airflow and after every data collection, have to wait 30 minutes before collecting a new set to get rid of padding gas residue.

3.2 System construction

Choosing a suitable sensor depends on the usage according to the situation. In this research, the stationary sensor is used to store data, we do not have to consider the weight and duration of the battery reserve as the mobile sensor. However, the size of the sensor is also important to consider because we want the sensor to be small and not clutter. The stationary sensor cannot move on its own that makes wind is the main factor in the gas movement. Therefore, we use a wind sensor (Rev. C) to collect the wind speed. Another important factor of GSL using a stationary sensor is the gas sensor. We choose Alcohol as a compound because it can be

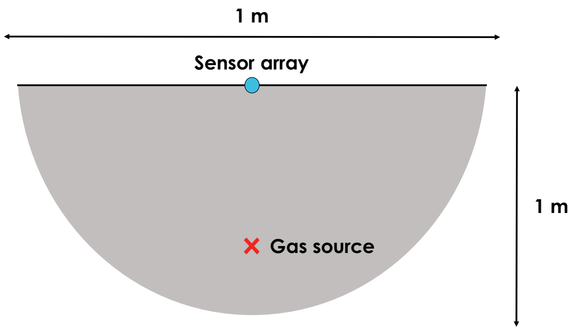

Fig. 2 Schematic of experimental setup

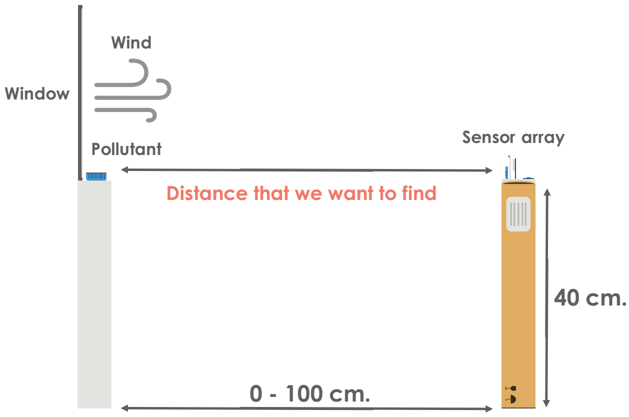

Fig. 3 Experimental setup.

easily found in an indoor and can be easily ignited. Therefore, we choose MQ135 and MiCS5524, which are AQI and Alcohol sensitive sensors, respectively. The gas detector can also measure combustible gases, e.g., Ammonia (NH3), Nitric oxide (NOx), Benzene, smoke, Carbon dioxide (CO2), etc. The purpose of this research is to find the distance of the gas source. We then use the ultrasonic sensor (HC-SR04) to measure the distance form the training set that use in a machine learning model.

3.3 Methodology

There are a lot of parameters that we get from the sensor network and many things are not necessary, so we cut out some parameters to choose only the important parameters. The heatmap in figure 4 and 5 shows that the relationship of some features have correlate to each other, but not correlate with the distance that is the output of the model. Therefore, we removed those features so that the model was not confused. Those features are shown in table 3.

Multiple traditional machine learning and deep learning algorithms in the regression task was applied to compare the result of this problem. They are Elastic net, k-nearest neighbor regression (KNN), Support Vector Regression, Random forests, Decision Tree Regression. K Nearest Neighbors regression (KNN) is the algorithm that we have tested that gives the best results to our data, which KNN is a simple machine learning algorithm that can be used in both classification and regression problems. KNN is a lazy learning algorithm witch generalization of the data and delayed until a query is made.

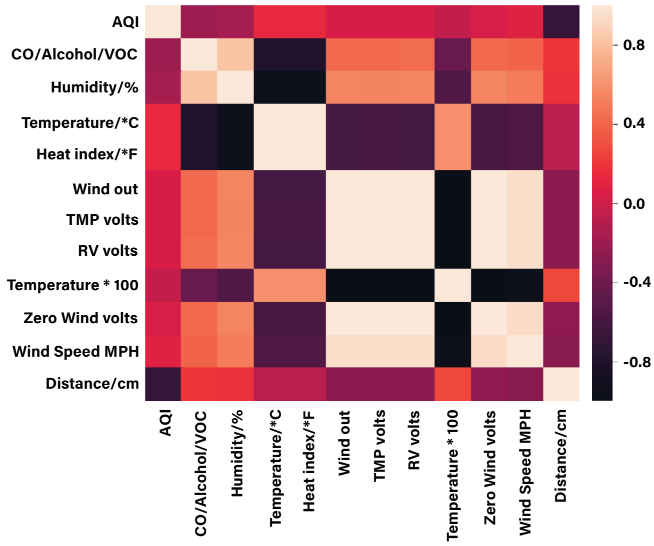

Fig. 4 The correlation heatmap for sensor network parameters.

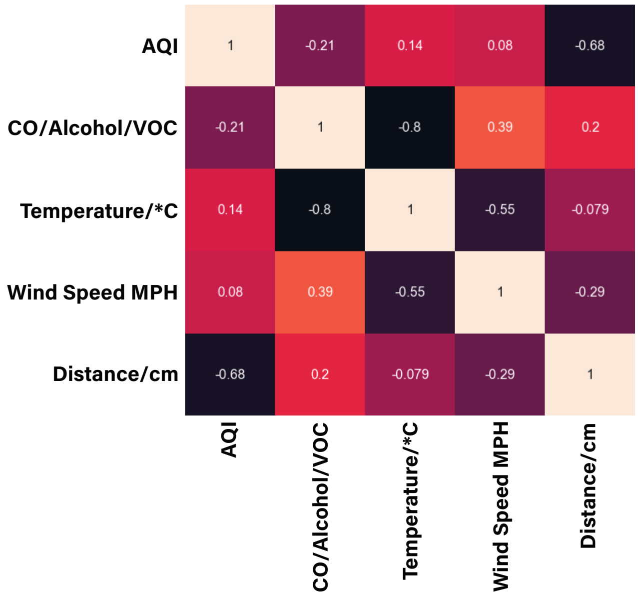

Fig. 5 The correlation heatmap only important features.

The target of the research is to find distance so KNN regression is an algorithm that used in this research. Minkowski metric was used to calculate the distance in KNN. The formula is often written as:

$$d(x, y) = \sqrt{(\sum_{i=1}^{n} |x_i - y_i|^p)}$$

where d is a distance and p is the power parameter of Minkowski metric.

In this research, $p = 2$ which is a Euclidean distance, that use to calculate the distance between two points in a plane. The Ball-tree is an algorithm that we used to compute the nearest neighbors. Ball-tree or metric tree is a multi-dimensional space with use divide-and-conquer approach. It partition data into herpespheres. All of the training data are partitioned into the nested set inside spheres as a ball.[6].

Decision Tree [8] is also used to compare in this research.

Table 1 Decision tree parameters

Parameters	Values
Criterion	Mean absolute error
Max features	12
Min samples leaf	3
Min samples split	5

Table 2 Random forests parameters

Parameters	Values
Criterion	Mean absolute error
Number of estimators	100
Minimum samples split	2
Minimum samples leaf	5
Maximum features	4
Maximum depth	7

Table 3 Parameters

Parameters	Units
AQI	parts-per-million (ppm, 10^{-6})
Carbon monoxide/Alcohol/ Volatile organic compound	parts-per-million (ppm, 10^{-6})
Temperature	Celsius (°C)
Speed of wind	Miles per hour (MPH)
Distance	Centimeter (cm)

Decision Tree is use in both regression and classification problems in the form of the tree structure. It separate the dataset into a smaller subset and structure it as decision nodes and leaf nodes.

Random forests[7] or random decision forests are an ensemble technique performing both classification and regression tasks. The idea is similar to the decision tree, but random forest uses multiple decision tree and bagging technique (Bootstrap Aggregation) that use to improve the accuracy of machine learning algorithm. For use bagging in random forest, it modified tree learning algorithm by split each learning candidate in the learning process (Feature bagging).

Before using machine learning for training the model, we preprocess the data by cutting outliers, handling empty value and normalizing data by using MinMax Scaling. The data is scaled to a fixed range of 0-1. The experiment is divided into 2 types for comparison by First, we calculate the result without moving average, the parameter used in the model is in Table 1.

During the data collection, we observed that wind and gas sensor response delay, so that make the values received from the sensors are not simultaneous. That lets us try to experiment with moving averages to improve performance to find GSL distances. The simple moving average (SMA) use to analyze data by creating a series of average data over the specified period.

$$SMA = \frac{G_1 + G_2 + \cdots + G_n}{n}$$

where G_n is a gas source value at period n, n is the number of time periods

4. Evaluation

The results showed that the machine learning algorithm is applicable to localize our experiment gas source using standard performance metrics for the regression problem in machine learning: Mean absolute error (MAE) metric. The test dataset was split by 25% from the dataset. The final output from the model is the distance between the source gas and the sensor.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

where y_i is the prediction value, x_i is the true value and n is the number.

Table 4 Performance metrics of the Machine Learning when the natural wind is an input

Models	Results (MAE)	R^2
Random forests (RF)	5.02	88.14%
Decision Tree Regression (DT)	5.60	76.36%
K-Nearest Neighbors (k-NN)	5.33	85.40%
Support Vector Regression (SVR)	9.62	70.80%

Table 5 Performance metrics of the Machine Learning when the natural wind is an input and number of time periods equal to 3.

Models	Results (MAE)	R^2
Random forests (RF)	5.01	89.99%
Decision Tree Regression (DT)	5.01	81.07%
K-Nearest Neighbors (k-NN)	5.86	84.35%
Support Vector Regression (SVR)	8.59	77.24%

Table 6 Performance metrics of the Machine Learning when the natural wind is an input and number of time periods equal to 10.

Models	Results (MAE)	R^2
Random forests (RF)	3.90	93.10%
Decision Tree Regression (DT)	4.13	85.48%
K-Nearest Neighbors (k-NN)	4.67	87.35%
Support Vector Regression (SVR)	7.48	81.23%

MAE is one of many metrics in machine learning for the regression problem with summarizing and assessing the quality of the model. It is the mean of the absolute values of the prediction errors. Each prediction error is the difference between true or actual values and predict values of a model with respect to a test set. The result of natural wind evaluating of the model is shown in Table 4. Table 4 contains the comparison of the performance between each machine learning algorithm using input data from natural wind condition. The bar graph in Figure 6 shows the comparison of the evaluation result using the MAE algorithm from natural wind dataset to find the distance of GSL.

R-Squared (R^2) is also used to measure the performance of the model. R^2 is also known as the coefficient of determination that evaluates how close the data are in the fitted regression line. The result of the R^2 describes the percentage of the variability in the output variable is considered for by the regression on the input variable.

The result of Table 4, 5, and 6 shows that Random forests performed the best result compared with other algorithms in natural wind conditions. In a generate wind condition, Decision Tree Regression does the best in the MAE metric. We noticed that both conditions of the natural wind and the wind generated from the fan have 3 algorithms that are good for finding GSL: K-Nearest Neighbors (K-NN), Decision Tree Regression (DT), and Random forests (RF) which have similar MAE depending on condition. The differences between the two conditions are due to the unpredictable

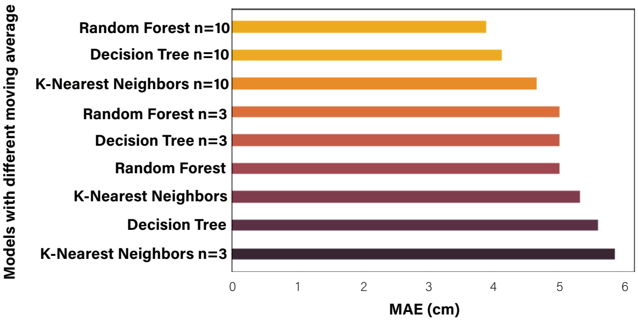

Fig. 6 The graph compares the efficiency of each moving average in a machine learning algorithm based on natural wind data sets. The model was evaluated by the mean absolute error (MAE) metric where a unit of each row in the graph is in centimeters. The smaller result is better.

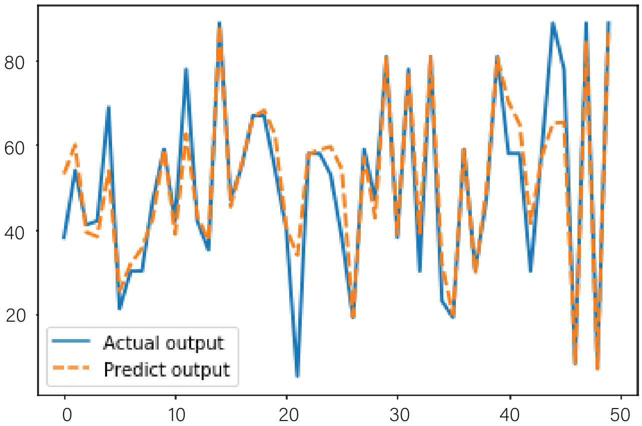

Fig. 7 Graph comparing the result of the actual output and the predicted output from the model using a random forest regression algorithm without moving average

Table 7 Performance metrics of the Machine Learning when the automatic fan is an input.

Models	Results (MAE)	R^2
Decision Tree Regression (DT)	1.93	95.80%
Random forests (RF)	2.25	94.42%
K-Nearest Neighbors (k-NN)	2.65	94.76%
Support Vector Regression (SVR)	7.88	83.76%

complexity of the natural wind and the wind direction. The results obtained from the experiment are somewhat different than expected because many sensors, especially gas sensors, are very sensitive, but the distance is very short, making the area of the experiment quite narrow. This experiment may not work in untrained weather conditions or temperatures.

We changed the number of time periods in moving average analysis to compare the result between n = 3 and n = 10. The result in Table 5, 6 shown shows how the wind delay affects the performance of gas source localization. The first thing we recognize is that the relationship between how long the wind takes the gas to the sensor is a good indication that the wind delay time can increase the model's performance, as we originally suspected.

To make it easier to understand the results, line graphs in Figure 7, 8, and 9 are used to show the comparison of predictions in each model between real data and predictive data. We can see that more timestamp that we use in moving average can predict better results for the source localization

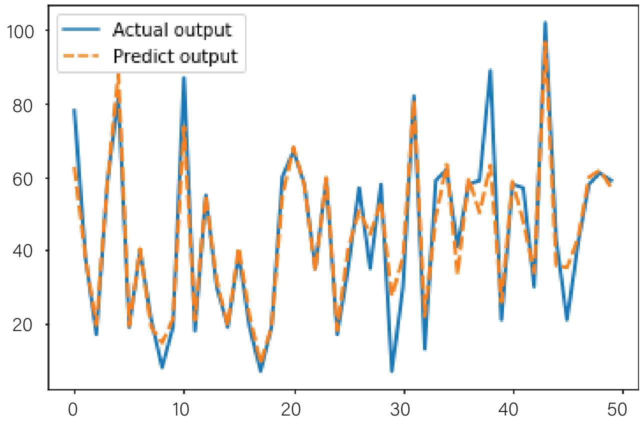

Fig. 8 Graph comparing the result of the actual output and the predicted output from the model using a random forest regression algorithm with moving average is 3.

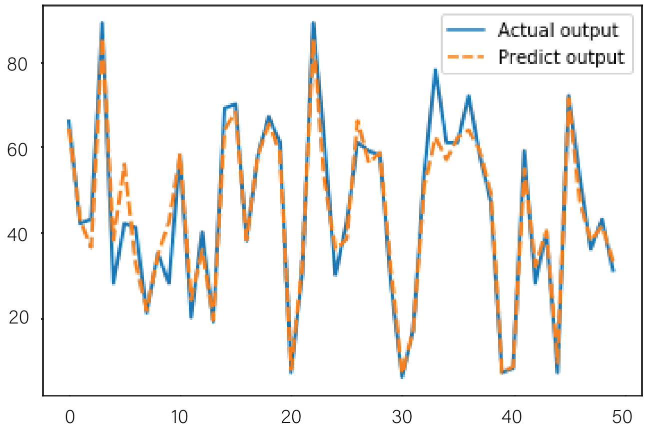

Fig. 9 Graph comparing the result of the actual output and the predicted output from the model using a random forest regression algorithm with moving average is 10.

because the wind slows down the value received from the sensor. These data are collected in the same condition as training data and have not been trained. We can summarize that as long as the data collected on the same weather conditions can be predicted using this method and it also shows that the model is not overfitted.

The data collected by using automatic fans also evaluated using the same machine learning algorithm as using natural wind.

The result of the model using wing data from an automatic fan is shown in Table 7. By using an automatic fan, data collected in this way is not complicated as natural wind because the wind generated from the fan is the wind in one direction that makes results is less than using natural wind in MAE metric. It shows that the complexity of the wind greatly affects the predictions of the model.

5. Conclusions

In this research, the machine learning model used to estimate the location of a gas source in an indoor environment with a windy condition using machine learning. The data were acquired in the indoor area for practicality. We evaluated several machine learning algorithms for comparison. We also compare the evaluation of natural wind condition with the wind that made from a fan. Many limitations occurred in the research. The limitation of the gas sensor was affected by this research because of the gas sensor can receive gas value in a limited distance, this makes the distance from the gas sensor and the gas source was close to each other. Also, The limitations of the wind are a problem as many wind directions make gas distributed without directions, so the sensor does not receive enough gas to calculate further. For future research, we plan to find the localization of the gas source by adapting the trilateration algorithm[2] with a wireless sensor network (WSN) to finding an effective and accurate positioning method to locate the source of the gas source. Another challenging thing is finding the location of multiple gas source and multiple gas type. We will also collect data in an outdoor environment condition with multiple wind directions. Use it as a wearable device, attached to the car or electric pole is one thing that is interesting as we can collect data from many places to alert the gas leak to the relevant parties to solve the problem. However, due to the large size of the sensors and the low receiving distance limitation, the sensor must still be developed first.

Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR18A3, Japan.

References

[1] Application of Convolutional Long Short-Term Memory Neural Networks to Signals Collected from a Sensor Network for Autonomous Gas Source Localization in Outdoor Environments: Sensors: C. Bilgera, A. Yamamoto, M. Sawano, H. Matsukura and H. Ishida, Sensors(online) available from ⟨https://doi.org/10.3390/s18124484⟩ (accessed 2018-12-18).

[2] Trilateration based localization algorithm for wireless sensor network: International Journal of Science and Modern Engineering (IJISME): O.S Oguejiofor, A. Aniedu, H.C Ejiofor, A.U Okolibe, IJISME(online) available from ⟨https://doi.org/10.1155/2015/874532⟩ (accessed 2018-9).

[3] A gas source localization algorithm based on wireless sensor network: International Journal of Distributed Sensor Networks (IJDSN): Q. Li, Z. Liu, J. Wang and X. Xiao, IJDSN(online) available from ⟨10.1109/WCICA.2014.7053119⟩ (accessed 2015-11-04).

[4] Gas Source Localization Using an Olfactory Mobile Robot Equipped With Wind Direction Sensor: International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM): H. Widyantara, M. Rivai and D. Purwanto, CENIM(online) available from ⟨10.1109/CENIM.2018.8711381⟩ (accessed 2018-11).

[5] Gas source localization with a micro-drone using bio-inspired and particle filter-based algorithms: International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM): P. P. Neumann, V. H. Bennetts, A. J. Lilienthal, M. Bartholmai and J. H. Schiller, CENIM(online) available from ⟨10.1080/01691864.2013.779052⟩ (accessed 2013-4-02).

[6] Ball*-tree: Efficient spatial indexing for constrained nearest-neighbor search in metric spaces: arXiv: M. Dolatshah, A. Hadian and B. Minaei-Bidgoli, CENIM(online) available from ⟨ https://arxiv.org/abs/1511.00628⟩ (accessed 2015-11-02).

[7] Random decision forests: Proceedings of 3rd International Conference on Document Analysis and Recognition: T. K. Ho, IEEE(online) available from ⟨10.1109/ICDAR.1995.598994⟩ (accessed 1995-8-16).

[8] Induction of decision trees: Machine Learning: J. R. Quinlan, Springer(online) available from ⟨https://link.springer.com/article/10.1007/BF00116251⟩ (accessed 1986-3).