A Coal Mine Underground Localization Algorithm Based on the Feature Vector

Guo Yinjing[1,*], Song Xianqi[1], Yang Lei[1] & Lv Wenhong[1]

Shandong University of Science and Technology
No. 579, Qianwan'gang Road, Qingdao Economic & Technical Development Zone,
Qingdao City, Shandong Province, 266510, PRC
*E-mail: gyjlwh@126.com

Abstract. To enhance the position estimation accuracy of an underground localization system for coal mine roadways, an algorithm based on the feature vector of received signals is presented in this paper. The algorithm includes three steps: the construction process of a feature vector database and a distance database, the vector matching process and the localization process. When a signal vector is received, it only needs to calculate the distance from the received vector to the center vector of each subset and then compare it with the data in the distance database. After multiple filtering and comparing the source of the strongest signal, the coordinates closest to the received vector are found. The experiment showed that the maximum error of this algorithm was 4 m and the average error was 1.62 m. Furthermore, within a localization error of 1 m, the X-axis localization accuracy was 98% while the Y-axis localization accuracy was 86%. Also, the algorithm took much less time compared to the KNN and WKNN algorithms, so the algorithm meets the requirements of coal mine safety systems and underground personnel localization systems.

Keywords: *underground localization algorithm; RSSI; feature vector; database construction; FVMA.*

1 Introduction

In the case of a coal mine accident, the accuracy of the localization system directly affects rescue efficiency. The RSSI (received signal strength indication) based localization algorithm is popular because of its low cost and high precision. It has become one of the most popular research directions in localization algorithms. The RSSI-based localization algorithm is divided into two main phases: offline data acquisition and online location matching [1-3]. However, the underground complex environment greatly affects the propagation of electromagnetic waves [4]. Experts and scholars have proposed a variety of localization methods to improve the stability and accuracy of RSSI. Hart & Cover proposed an algorithm called the K-neighbor method [5-6]. By comparing and calculating the distance between the received signal and the element of a fingerprint database, the nearest coordinate is selected from the

Received April 2nd, 2018, Revised January 8th, 2019, Accepted for publication February 2nd, 2019.

database as the coordinate of the locating point. In 2008, Lakmali *et al.* proposed a localization algorithm based on the K-neighbor method and evaluated it using a cellular network. They focused on the gathering of fingerprint data and the establishment of the fingerprint database [7]. For the instability of the RSSI signal, Chunyan & Wangjian have improved the KNN (k-nearest neighbor) algorithm using a database with geometric information. They used the point spoke strength of the nearest sample point to determine the control point network structure of the reference point where the mobile terminal is located. Based on this, their method dynamically selects the key parameter K of the KNN algorithm [8]. For the problem of low accuracy of the KNN algorithm, Kong & Chunlei proposed the WKNN localization algorithm to adjust the weights of K nearby points according to the fluctuations of the mean and variance of the RSSI values [9]. In addition, other scholars have proposed a weighted centroid algorithm [10], a K-means algorithm [11-12], a Bayesian algorithm [13] and an algorithm combined with Kalman filtering [14]. These algorithms improve positioning accuracy to varying degrees.

Some algorithms in the above references are very complicated and some cannot meet the requirements of high localization accuracy for more complicated environmental conditions in coal mine roadways. A new algorithm based on the feature vector of received signals, called the feature vector-matching algorithm (FVMA), is presented in this paper. In the process of constructing databases, first, several typical vectors are selected as seed vectors in the sample set. After several iterations, the sample set is divided into multiple subsets and the central vector of each subset is also calculated. Then, the distances between each vector and the center vectors of the subset, the previous subset and the next subset are calculated and stored in the distance database. In the vector matching process, the subset which the received vector belongs to can be found by calculating the distance between the received vector and the center vector of each subset. Then, according to the distance database, the two vectors closest to the distance of the received vector to the center vectors of the subset and the previous subset (or the next subset) are selected. Finally, by comparing the source of the strongest signal of these two vectors and the received vector, one single localization vector is determined and the localization process is completed. This study focused on the construction method of the databases and the localization method. Also, the accuracy and localization time of FVMA were tested by a simulation experiment.

2 Composition of the Underground Localization System

The roadway localization system based on the feature vector is composed of a localization server, a number of wireless access points (APs) and WiFi terminal nodes. The signals transmitted by each AP are determined by their amplitude and frequency. All transmitted signals have the same amplitude. AP points that are far apart can use the same transmission frequency. Since the amplitude value of the signal is more attenuated as the distance increases, the terminal node can receive vectors of amplitude and frequency from nearby APs. The terminal node sends the received vector to the localization server and the localization algorithm software embedded in the server can calculate the coordinate of the vector, thus realizing localization in the roadway.

3 Establishment of the Feature Vector Database and Distance Database

The foundation of the FVMA algorithm is the establishment of an underground localization feature vector database and a distance database. Considering the complexity of the underground environment, the heavy multipath effect and the complex roadway node topology, it is difficult to establish a suitable training set based on theoretical analysis. In this study, the training set was constructed by gathering the data vectors in the field.

3.1 Construction of the 'Training Set' for the Feature Vector Database

When a node moves in the coal mine roadway, feature vectors with coordinates are created; these vectors correspond to the points through which the node has passed. To ensure accuracy, every point in the coal mine roadway should be taken into account. Because each AP is fixed, the coordinate and the timescale signal frequency of the AP are predefined. For example, at a point designated as *k*, the moving terminal node receives certain timescale signals from nearby APs, denoted by $AP_1, AP_2, AP_3 ... AP_l$, and records the frequencies and amplitudes of the signals, after which a vector for point *k* is constructed as:

$$S_k = [(f_1, A_{k1}), (f_2, A_{k2}), (f_3, A_{k3}) ... (f_l, A_{kl}), (x_k, y_k)] \tag{1}$$

Here, A_{ki} is the signal amplitude value received at (x_k, y_k), and f_i is the frequency of the signal transmitted by AP_i.

While the terminal node traverses all points in the roadway of the coal mine, a 'training set' is obtained as follows:

$$\left\{\begin{array}{c} S_1 = [(f_1, A_{11}),(f_2, A_{12}),(f_3, A_{13})......(f_n, A_{1l}),(x_1, y_1)] \\ S_2 = [(f_1, A_{21}),(f_2, A_{22}),(f_3, A_{23})......(f_n, A_{2l}),(x_2, y_2)] \\ S_3 = [(f_1, A_{31}),(f_2, A_{32}),(f_3, A_{33})......(f_n, A_{3l}),(x_3, y_3)] \\ \vdots \\ S_n = [(f_1, A_{n1}),(f_2, A_{n2}),(f_3, A_{n3})......(f_n, A_{nl}),(x_n, y_n)] \end{array}\right\} \qquad (2)$$

3.2 Selection of the Vector Elements of the Localization Feature Database

To reduce the computational complexity of the localization algorithm, the sample feature vector set is divided into multiple subsets and a representative vector of the feature database is selected based on the feature vector sample set. The selection steps are as follows:

Step 1: Divide the ‘training set’ into p subsets. To reduce the error of the localization algorithm, one representative vector must be selected for each subset. As the selection procedure requires initialization, a random vector S_i is selected from the ‘training set’ as the representative vector of subset L_i. We obtain a seed vector database Da, described as follows:

$$Da = \left\{S_1, S_2, S_3, \cdots, S_p\right\} \quad (3)$$

Step 2: Initialize the variable values of the next steps: Let $D_{sum} = 0$, $D^{'}_{sum} = 0$.

Step 3: Calculate the distances $D_1, D_2, D_3, \cdots, D_p$ between a sample vector of every point on the roadway and every seed vector $S_1, S_2, S_3, \cdots, S_p$ based on the Euclidean distance formula focusing on $A_1, A_2, A_{3.} ... A_l$. Find the minimum distance $D_i = \min\left\{D_1, D_2, D_3, \cdots, D_p\right\}$, divide the sample vector into subset L_i, and let $D_{sum} = D_{sum} + D_i$.

Step 4: Repeat Step 3 until all vectors in the ‘training set’ have been allocated to the corresponding subset.

Step 5: Calculate the central vectors of the subsets. The central vector of subset L_k is:

$$S_k^{'} = [(f_1, A'_{k1}),(f_2, A'_{k2}),(f_3, A'_{k3})...(f_l, A'_{kl}),(x_k, y_k)] \qquad (4)$$

where $A'_{k1} = \frac{1}{q}\sum_{i=1}^{q} A_{i1}$, $A'_{k2} = \frac{1}{q}\sum_{i=1}^{q} A_{i2}$, .. $A'_{kl} = \frac{1}{q}\sum_{i=1}^{q} A_{il}$, and q is the number of sample vectors in subset L_k .

Step 6: Assign a sample vector S_k in subset L_k as the new seed vector, where S_k has the shortest distance to S_k in subset L_k. Then, a new seed vector database Da' is obtained in which all subsets have new seed vectors. Let $\varepsilon = \frac{\left|D_{sum} - D'_{sum}\right|}{D_{sum}}$.

Step 7: If $\varepsilon > \varepsilon_0$, let $D'_{sum} = D_{sum}$ and $Da = Da'$, and then repeat Step 3 to Step 7; otherwise, finish the selection procedure. Here, ε_0 is a stop threshold constant.

3.3 Construction of Distance Database

After performing all the steps described above, the localization feature database is created. Then, the distances from all the sample vectors of each subset to the center vectors of the subset, the previous subset and the next subset can be calculated. Thus, for each subset a distance database D_b is built. For example, the distance database of subset L is stored as shown in Table 1. The index of the last column is the index value of the point in the entire feature vector database.

Table 1 Distance database of subset L.

Sample Position	Distance to Self-subset	Distance to Previous Subset	Distance to Next Subset	Index
(x_{l1}, y_{l1})	$d_1(l,l)$	$d_1(l,l-1)$	$d_1(l,l+1)$	l_1
(x_{l2}, y_{l2})	$d_2(l,l)$	$d_2(l,l-1)$	$d_2(l,l+1)$	l_2
...	...	...	...	...
(x_{lq}, y_{lq})	$d_q(l,l)$	$d_q(l,l-1)$	$d_q(l,l+1)$	l_q

4 Localization Procedure Based on Feature Vector Matching

The WiFi terminal device at K will receive signals from nearby APs and construct a vector as follows:

$$S_k = [(f_1, A_{k1}), (f_2, A_{k2}), (f_3, A_{k3}), \ldots (f_n, A_{kn})] \tag{5}$$

One data packet containing all the information of vector S_k is sent to the server via the WiFi network. At the server side, S_k is recovered and then the localization steps are as follows:

Step 1: Initialization cutoff threshold constant $\eta_0 = 1, \eta_1 = 1, \eta_2 = 1$.

Step 2: Calculate the distance from S_k to the center vector of each subset based on the values of $D_i = \min\{D_1, D_2, D_3, ..., D_P\}$. The terminal device is considered to belong to subset i.

Step 3: Based on the distance database of subset i, some vectors that are close to S_k can be filtered out by the formula $|d(i,i) - D_i| < \eta_0$. Here, η_0 is a stop threshold constant. If the number found is 0, then increase the value of η_0 and repeat Step 3, otherwise proceed to the next step.

Step 4: Calculate the distance $D_{(k,i-1)}$ from S_k to S_{i-1}. The vectors obtained from Step 3 are filtered again by formula $|d(i,i-1) - D_{(k,i-1)}| < \eta_1$. Here, η_1 is a stop threshold constant. Similarly, if the filtered number is 0, then increase the value of η_1 and repeat Step 4, otherwise proceed to the next step.

Step 5: The vector of $|d(i,i) - D_i| < \eta_2$ in Step 4 is filtered. If there is such a vector, the smallest two are selected to go to Step 6, otherwise the results of this localization are discarded. Adjust the size of η_2 here to adjust the localization accuracy.

Step 6: Compare the source of the strongest signal of the two vectors obtained in Step 5 and S_k. Then, the only localization vector is determined and the index of the vector is known.

Step 7: According to the index obtained in Step 6, the coordinates of the localization point can be found in the feature vector database and the localization procedure of point K is completed.

5 Experiment and Discussion

Using the MATLAB software, an underground coal mine roadway with a length of 400 m and a width of 5 m was simulated. There were 21 WiFi APs distributed at equal distance along the roadway and 2000 points were selected as the training set. The location of the test points on the roadway was randomly generated. For convenience of statistics and observation, the random process produced only integer coordinate values. By comparing the true location and the location obtained by FVMA, the localization accuracy of the algorithm can be clearly seen. As the number of localization points increases, the error range and accuracy rate can be seen intuitively by comparing the difference between the vertical and horizontal coordinates.

5.1 Pretreatment of Experimental Data

In the RSSI-based localization algorithm a lognormal distribution model is mostly used. It is expressed as follows:

$$PL(d) = PL(d_0) + 10n * \lg\left(\frac{d}{d_0}\right) + v_k \tag{6}$$

where *PL(d)* is the path loss after the signal passes distance *d*, the unit is *dB*; v_k represents a Gaussian random variable with mean 0 and its standard deviation is usually 4~10. The range of *n* is generally 2~5; d_0 indicates the reference distance, which is generally 1 m. The signal strength value received by the final terminal node is:

$$RSSI = P + G - PL(d) \tag{7}$$

where *P* is the transmission power of the signal and *G* is the antenna gain of the AP node.

Usually the mobile terminal receives a negative value and the larger the distance, the smaller its absolute value. The relationship between signal strength and distance is shown in Figure 1. Of course, this empirical model does not accurately simulate signal attenuation in a real environment, but it can reflect the trend of signal attenuation.

The information used for localization mainly comes from the signal transmitted from the AP closest to the mobile terminal. In order to facilitate the calculation and observe directly, while expanding the force of nearby AP signals and reducing the effect of long-distance AP signals, the received signal strength was taken as an absolute value; then the countdown was taken and multiplied by 10000 to increase the distance. The degree of signal recognition was taken as:

$$RSSI^{'} = \frac{10000}{|RSSI|} \tag{8}$$

The relationship between signal strength and distance after preprocessing is shown in Figure 2.

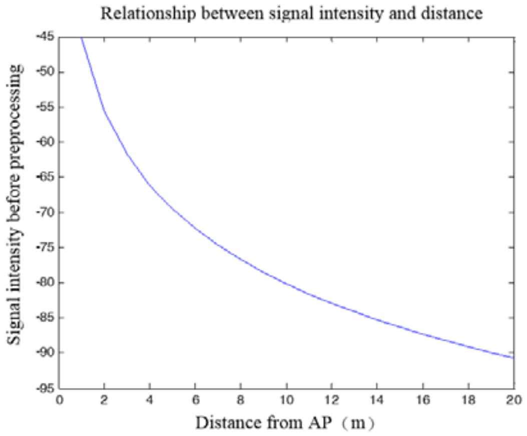

Figure 1 Trends in the relationship between signal strength and distance.

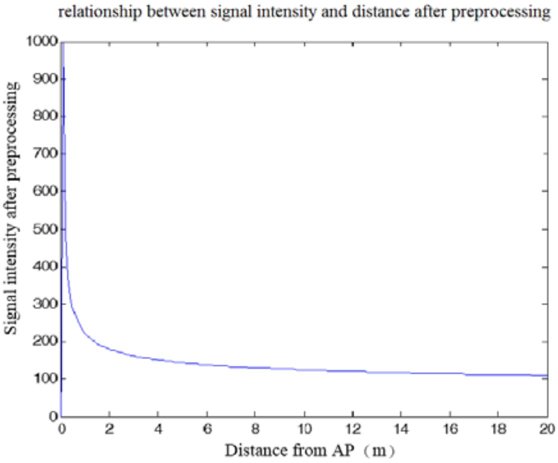

Figure 2 Trends in the relationship between signal intensity and distance after preprocessing.

5.2 Comparison of Real Coordinates and Computed Coordinates

A comparison was made between 20 random localization results. In the roadway with a length of 400 m and a width of 5 m, the real and computed coordinates of the mobile terminal were as shown in Table 2. As can be seen from the table, the average error of the localization results of these 20 random points was 0.4 m, and the maximum error was no more than 2 m.

Table 2 Comparison of 20 random localization results.

Point Number	Real Coordinates	Computed Coordinates	Point Number	Real Coordinates	Computed Coordinates
1	(41,1)	(41,1)	11	(267,3)	(266,3)
2	(149,1)	(149,1)	12	(338,4)	(338,4)
3	(56,4)	(56,4)	13	(176,3)	(175,4)
4	(173,3)	(173,2)	14	(279,4)	(279,4)
5	(287,3)	(285,3)	15	(379,1)	(379,1)
6	(108,2)	(108,2)	16	(79,4)	(79,4)
7	(352,3)	(352,4)	17	(101,5)	(101,4)
8	(276,1)	(276,1)	18	(412,2)	(412,2)
9	(347,2)	(348,2)	19	(85,1)	(85,1)
10	(150,1)	(150,1)	20	(322,2)	(322,1)

5.3 Analysis of Coordinates Error and Accuracy Rate

MATLAB randomly generated a hundred pairs of coordinates. The following experiment shows the difference between the real coordinates and the computed coordinates of these points, and the average accuracy rate is also given. Figure 3 shows the X-error of the 100 points in the 400-m long roadway. It can be clearly seen that most of the points have an error of 0 to 1 m and a few of them have an error of 1 to 2 m.

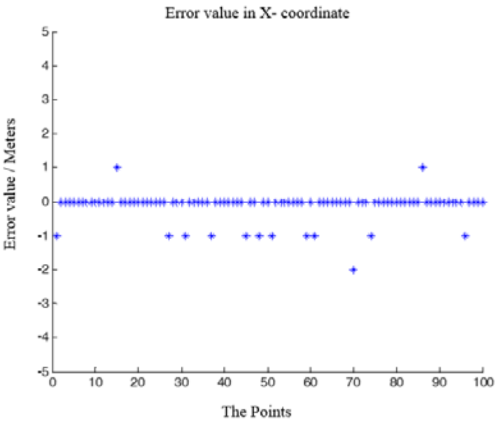

Figure 3 The error value in the direction of the length of the roadway (X-coordinate).

The experiment was conducted five times. The error values of the X-coordinates of these points were counted, as shown in Table 3. Times refers to the number of experiments, Range refers to the error range of the real and calculated coordinates, and Numbers refers to the number of points within each error range. Then, the accuracy rate shown in Table 4 below was calculated.

Table 3 Distribution of X-coordinate errors.

Times Numbers Range(abs）	1	2	3	4	5
0-1 m	99	98	98	99	97
1-2 m	1	2	1	1	2
2-3 m	0	0	1	0	1
3-4 m	0	0	0	0	0
4-5 m	0	0	0	0	0
More than 5 m	0	0	0	0	0

Table 4 Accuracy rate of X-coordinates.

Abs/M(Err or value)	0-1	1-2	2-3	3-4 M	4-5	More than 5
Proportion	98.2%	1.4%	0.4%	0	0	0

Figure 4 shows the Y-error in the width of the roadway. It can be seen from the graph that the error of most points is within 2 m.

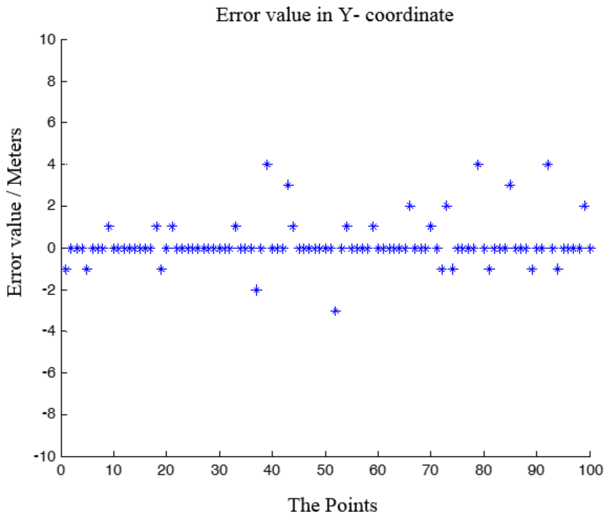

Figure 4 The error value in the direction of the width of the roadway (Y-coordinate).

This test was also carried out five times and the error statistics of the Y-coordinate of these tests are shown in Table 5. Meanwhile, Table 6 shows the distribution of the accuracy rate in different ranges.

Table 5 Distribution of Y-coordinate errors.

Times Numbers Range (abs）	**1**	**2**	**3**	**4**	**5**
0-1 m	88	85	86	91	84
1-2 m	5	9	8	5	9
2-3 m	5	5	5	1	2
3-4 m	2	1	1	3	5
4-5 m	0	0	0	0	0

Table 6 Accuracy rate of Y-coordinates.

Abs/M(Error value)	**0-1**	**1-2**	**2-3**	**3-4**	**4-5**
Proportion	86.8%	7.2%	3.6%	2.4%	0

5.4 Cumulative Distribution Function Comparison

In order to demonstrate the effectiveness of FVMA, it was compared with the KNN and WKNN algorithms. Figure 5 shows the error cumulative distribution function for the three algorithms, FVMA, KNN, and WKNN. The abscissa of Figure 5 is the error distance and the ordinate is the cumulative distribution probability.

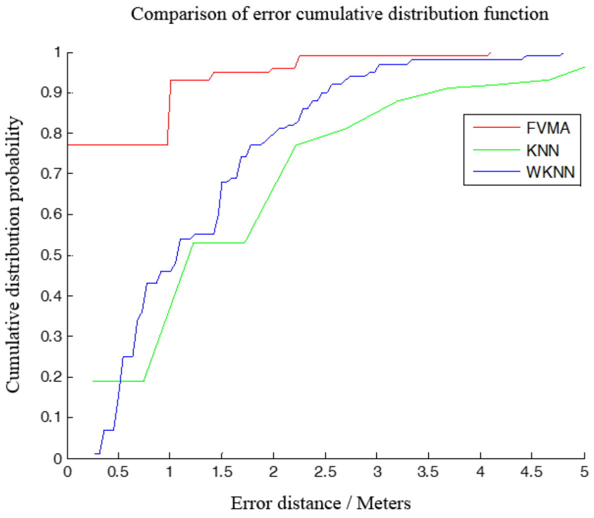

Figure 5 Cumulative error distribution function of three algorithms.

For the same error value, the larger the value of the cumulative distribution probability, the higher the credibility of the positioning result and the higher the accuracy. For the same cumulative distribution probability, the larger the value of the error distance, the less accurate the localization result. It can be seen from the figure that within the same error range, FVMA had the highest localization accuracy, followed by the WKNN algorithm, and finally the KNN algorithm.

5.5 Analysis of Localization Time

In addition to higher localization accuracy, the FVMA algorithm also has the advantages of low localization time and high real-time performance. In order to reflect the time-consumption advantage of FVMA, the time-consumption comparison results of the KNN, WKNN, and FVMA methods to locate 1000 points are shown in Figure 6.

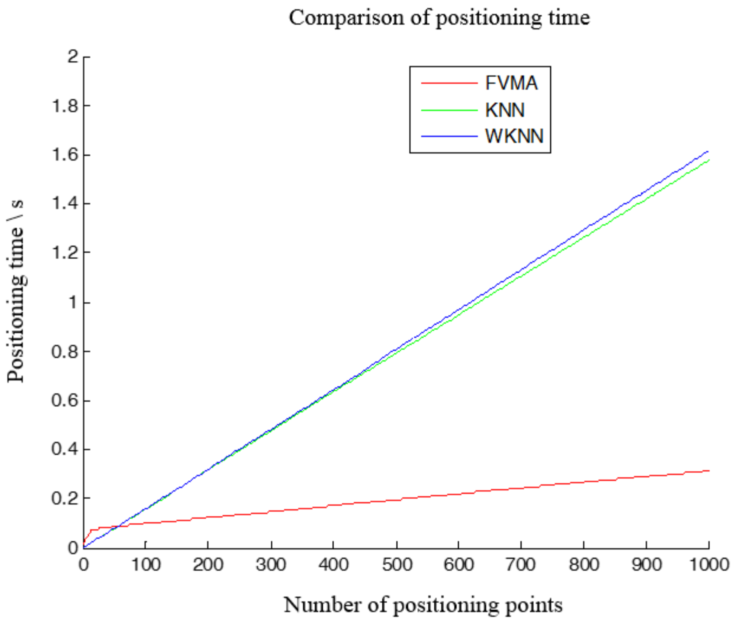

Figure 6 Localization time comparison of the three algorithms.

As can be seen from the figure, KNN and WKNN required similar time consumption, while that of WKNN was slightly larger than that of KNN, but the localization time of FVMA was much shorter than that of the former two. Of course, as the number of data in the database increases, the time-consumption gap between the different algorithms also increases.

6 Conclusion

The amplitude-frequency vector method proposed in this paper can effectively reduce the multipath effect in an underground roadway without increasing the hardware cost. The pre-calculation of the fingerprint database can reduce the time-consumption in localization and improves the real-time performance. After receiving the localization signal, the localization accuracy can be effectively improved by using the matching algorithm proposed in this paper. The localization accuracy can also be set by adjusting the size of η_2 to obtain the required localization result. In the simulation analysis of FVMA, the following conclusions can be drawn:

1. The vector-matching algorithm proposed in this paper can meet the requirements of localization accuracy in downhole incidents for disaster assistance. In the simulation experiment, FVMA showed good localization performance in the roadway of a coal mine and the localization error was less than 4 m. Although the experimental results showed that the error of the Y-coordinate was slightly larger than that of the X-coordinate, the result is acceptable because the width of the roadway is limited.
2. Using the feature vector, the FVMA integrates all the electromagnetic wave information of the APs near the node, thereby minimizing the adverse impact of the multipath effect. By selecting the vector elements for the localization feature database, the FVMA identifies the representative vectors that denote typical points distributed on the entire roadway area; as a result, the localization error caused by imprecise matching is limited. These factors ensure localization accuracy.
3. The preprocessing calculation of FVMA is the basis of the localization algorithm. The pretreatment calculation reduces the computing complexity and improves the real-time performance of the localization algorithm without increasing the computing burden of the mobile nodes. In the process of localization, most of the calculations are carried out by the server, thus improving the energy efficiency of the mobile terminal nodes.

Acknowledgements

The work was supported by the next founds: National Natural Science Foundation of China 61471224 and China National Coal Association MTKJ 2016-288.

References

[1] Taniura, Y. & Oguchi, K., *Indoor Location Recognition Method Using RSSI Values n System with Small Wireless Nodes*, 2017 40th International

Conference on Telecommunications and Signal Processing (TSP), pp. 52-55, 2017.

[2] Weixing, X. & Weining, Q., *Improved Wi-Fi RSSI Measurement for Indoor Localization*, IEEE Sensors Journal, **17**(7), pp. 2224-2230, 2017.

[3] Simon, Y. & Marzieh, D., *Wireless RSSI Fingerprinting Localization*, Signal Processing, **131**(1), pp. 235-244, 2017.

[4] Jin, L. & Yinjing, G., *Design of Underground Wireless Positioning System Based on Fingerprint Algorithm*, Journal of Shandong University of Science and Technology (Natural Science Edition), **6**(2), pp. 47-45, 2013.

[5] Jiang, S., Pang G., Wu, M. & Kuang, L., *An Improved K-Nearest-Neighbor Algorithm for Text Categorization*, Expert Systems with Applications, **29**(1), pp. 1503-1509, 2012.

[6] Shaowu, M., Huanguo, Z. & Chongchao, H., *Application of Improved K Shortest Path Algorithm in Communication Network*, Journal of Wuhan University (Science Edition), **5**(6), pp. 534-538, 2013.

[7] Lakmali, B.D.S. & Dias, D., *Database Correlation for GSM Location in Outdoor & Indoor Environments*, International Conference on Information and Automation for Sustainability, pp. 42-47, 2008.

[8] Chunyan, L. & Wangjian, *Constrained KNN Indoor Positioning Model Based on Set Clustering Fingerprint Database*, Journal of Wuhan University (Science Edition), **39**(11), pp. 1287-1292, 2014.

[9] Kong, C., Chunlei, S. & Jiabin, C., *Positioning Fingerprint Indoor Positioning Algorithm Based on Improved WKNN*, Navigation Positioning and Timing, **3**(4), pp. 58-64, July.2016.

[10] Lu, Y., *The Design of Underground Personnel Positioning System Based on Improved Weighted Centroid Algorithm*, China Mining Magazine. **26**(2), pp. 169-173, 2017.

[11] Zhong, Y., Wu, F., Zhang, J. & Dong, B., *Wifi Indoor Location Based on K-means*, Proceedings of International Conference on Audio, Language and Image Processing (ICALIP), pp. 663-667, 2016.

[12] Abdullah, O. & Abdel-Qader, I., *K-Means-Jensen-Shannon Divergence for a WLAN Indoor Positioning System*, IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2016.

[13] Zhou, F. & Lin, K., *RSSI Indoor Localization Through A Bayesian Strategy*, Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), pp. 1975-1979, 2017.

[14] Cheng, Z. & Jiazheng, Y., *Bluetooth Indoor Positioning Based on RSSI and Kalman Filter*, Wireless Personal Communications, **96**(3), pp. 4115-4130, 2017.