A Study on Similarity Matching Algorithms and Reliability of Graphics Based on Standard Error

Tao REN*, Xiaoqing KANG, Wen SUN

School of Mechanical Engineering, Xi'an Shiyou University, Xi'an, Shaanxi, 710065, China

Abstract **— Similarity recognition of graphic objects is widely used in engineering and terrain contour matching. The accuracy of similarity recognition is determined by a matching algorithm. The reliability of the matching algorithm determines the accuracy of similarity recognition. In this paper we study the reliability of the widely used classical matching algorithm. An argument about matching algorithm correlating to the degree of difference between graphic objects is first proposed, followed by a detailed analytical discussion. We then present a new graphic recognition method based on the standard error algorithm and formulate its mathematical model. The standard error algorithm and the classic matching algorithm are compared in detail. Our research results show that the standard error matching algorithm and the degree of difference between graphic objects have a high correlation, and the matching algorithm has a very high reliability. The reliability of standard error algorithm is better than that of the classical matching algorithm.**

Keywords - standard error; graphic similarity; matching algorithm; reliability

I. INTRODUCTION

Similarity recognition of two-dimensional graphics is widely used in the field of engineering and terrain contour matching (TERCOM) [1]. At present, there are many algorithms to evaluate the degree of similarity between standard graphic (SG) and contrast graphic (CG), such as Grid and gray matrix (GGM) [2], Artificial neural network (ANN) [3], Vector, cross correlation algorithm (COR)[4-6], mean absolute difference algorithm (MAD) [4,6,7], mean square difference algorithm (MSD)[4,5,6,8,910], Camberra distance algorithm[11], Hausdorff algorithm[8,9,10,11,12], normalized product correlation (Nprod) [13]. They are all the most classical and most widely used algorithms at home and abroad. Grid and gray matrix algorithm and artificial neural network algorithm are all based on the grid characteristic matrix to extract the boundary information of graphic. Because there may be great difference for the path of a curve to pass through the same grid, characteristic matrix extracted from them can not reflect local feature of the graphic exactly, even if the algorithm is very exact, the methods are still a characteristic recognition and only applicable for graphic classification recognition rather than exact recognition. Vector, COR, MAD, MSD, Camberra, Hausdorff and Nprod algorithms belong to graphic recognition. At present, there is little research on the reliability of graphic recognition algorithm. Matching algorithm correlating to the degree of difference between graphics is first proposed, and makes a detailed discussion and analysis. This paper presents a new similarity matching algorithm based on standard error to realize recognition of standard graphic and contrast graphic.

II. SIMILARITY MATCHING ALGORITHM OF GRAPHIC BASED ON STANDARD ERROR

A. *Dimensionless Calibration of Graphic*

The graphic data are discrete two-dimensional array[Xi，Yi] (i=1，2，….). Since x coordinates value of graphics are not the same, the graphics should be calibrated by equal x coordinate value. The y coordinate value can be obtained by Lagrange algorithm.

$$y(x)=\sum_{k=1}^{n} y_k\left(\prod_{\substack{j=1\\ j\neq k}}^{n}\frac{x-x_j}{x_k-x_j}\right) \tag{1}$$

The shape of the geometric contour of a graphic is the relevant element to be extracted. While the actual [Xi, Yi] value are irrelevant elements for recognizing, so, dimensionless calibration should be performed to graphics.

$$\begin{cases} Y_d=(y_i-y_{\min})/(y_{\max}-y_{\min}) \\ X_d=(x_i-x_{\min})/(x_{\max}-x_{\min}) \end{cases} \tag{2}$$

Where ymax is maximum y, N, ymin is minimum y, N, xmax is maximum x, m, xmin is minimum x, m.

B. ***Similarity Matching Algorithm Based on Standard Error***

In dimensionless coordination system, the degree of similarity or proximity of standard graphic and contrast graphic is represented by proximity of same coordinate point. A graphic is formed by data set, calibrate standard graphic data set as truth value, data set of contrast graphic as test value. Thus, similarity between standard graphic and contrast graphic can be defined as proximity of test value to truth value, namely, deviation between test value and truth value, called standard error (SE). SE reflects sensitively to those extremely big or small errors in a set of test value, so SE can well reflect the accuracy of data set of contrast graphic approaching to data set of standard graphic.

SE of data set A(y_A, x_A) from contrast graphic and data set B(y_B, x_B) from standard graphic is

$$d(A,B)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_A(i)-y_B(i))^2} \tag{3}$$

SE value $0 \le d \le 1$, SE of two graphics that have completely similar geometry is 0. The closer the geometry come, the smaller the SE value is.

Similarity is an equivalent scale to evaluate the similarity degree of two graphics. Similarity of complete similar graphics is 1, the more similar the two graphics are, the bigger the similarity value is. Similarity of two graphics can be defined as follow:

$$R = 1 - d \tag{4}$$

When making similarity recognition of standard graphic and contrast graphic, according to the similarity value, the similarity degree of two graphics can be estimated accurately.

III. STUDY ON CORRELATION BETWEEN THE SIMILARITY MATCHING ALGORITHM AND THE GRAPHICS DIFFERENCE DEGREE

The present algorithm models of similarity matching based on graphic mainly are: Vector algorithm, COR algorithm, MAD algorithm, MSD algorithm, Camberra algorithm, Hausdorff algorithm and Nprod algorithm. They all are classical algorithms to determine whether the contrast graphic is similar to standard graphic. However, whether these algorithms can make a precise determination of similarity between two graphics, or which algorithm of them can make a better determination of the similarity of graphics, there is no report in Chinese and foreign literatures. An argument about matching algorithm correlating to the degree of difference between graphics is first proposed: difference of the matching value should represent the degree of the difference between two graphics, that is, the more different the two graphics are, the more different the matching values are, and also monotonic. Therefore, the correlation of similar matching algorithm and the degree of difference between graphics of contrast graphic data set A and standard graphic data set B based on dimensionless calibration can be represented as: the variation of matching value Δd and the degree of difference between two graphics ΔR have same variation trend.

$$\Delta d(A,B) \propto \Delta R(A,B) \tag{5}$$

Figure1 shows a standard graphic B. Contrast graphic A can be obtained by adding fluctuation to graphic B, shown in Figure 2. By gradually adding more fluctuation points to data set A, it became more and more different from data set B. The variation trend of matching value is studied by using Vector algorithm, COR algorithm, MAD algorithm, MSD algorithm, Comberra algorithm, Hausdorff algorithm, NProd algorithm and SE algorithm.

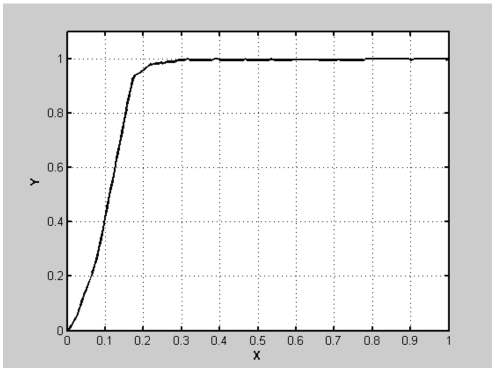

Figure 1. Standard graphic B

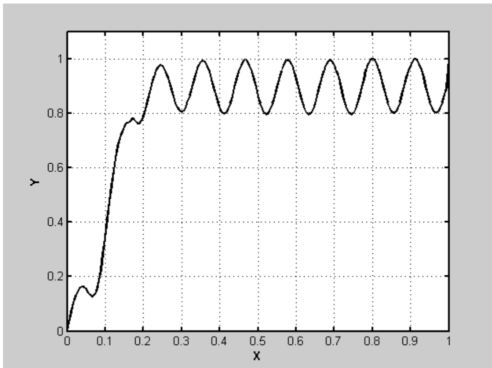

Figure 2. Contrast graphic A

(1) Vector algorithm

Standard graphic and contrast graphic are represented by 2n dimension vectors; matching algorithm is their dot product, that is, cosine value of the included angle between vectors.

$$R_n(A,B) = A^T \bullet B \tag{6}$$

(2) COR algorithm

Cross correlation (COR) algorithm:

$$R(A,B)=\frac{\frac{1}{N}\sum_{i=1}^{N}(y_A(i)y_B(i)-\overline{y_{AB}})(y_B(i)-\overline{y_B})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_A(i)y_B(i)-\overline{y_{AB}})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_B(i)-\overline{y_B})^2}} \tag{7}$$

(3) MAD algorithm

Mean absolute difference (MAD) algorithm:

$$R(A,B)=1-\frac{1}{N}\sum_{i=1}^{N}|y_A(i)-y_B(i)| \tag{8}$$

(4) MSD algorithm

Mean square difference (MSD) algorithm:

$$R(A,B)=1-\frac{1}{N}\sum_{i=1}^{N}(y_A(i)-y_B(i))^2 \tag{9}$$

(5) Camberra algorithm

The Camberra distance between vectors of standard graphic and contrast graphic can be used as distance measure between two characteristic vectors.

$$R(A,B)=1-\frac{\sum_{i=1}^{N}\left|y_A(i)-y_B(i)\right|}{\sum_{i=1}^{N}\left|y_A(i)+y_B(i)\right|} \tag{10}$$

(6) SE algorithm

Standard error (SE) algorithm:

$$R(A,B)=1-\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_A(i)-y_B(i))^2} \tag{11}$$

(7) Hausdorff algorithm

Hausdorff distance is defined by general mathematical equation. It is a definition of the distance between two point sets and a measure to describe their similarity as well.

$$\begin{aligned}R(A,B)&=1-\max[d_{AB}(y_A,y_B),d_{BA}(y_B,y_A)]\\ d_{AB}&=\max_{y_A\in A}\min_{y_B\in B}\left\|y_A-y_B\right\|\\ d_{BA}&=\max_{y_B\in B}\min_{y_A\in A}\left\|y_B-y_A\right\|\end{aligned} \tag{12}$$

(8) Nprod algorithm

Normalized Product Correlation (Nprod) algorithm:

$$R(A,B)=\frac{\sum_{i=1}^{N}(y_A(i)-\overline{y_A})(y_B(i)-\overline{y_B})}{\sqrt{\sum_{i=1}^{N}(y_A(i)-\overline{y_A})^2}\sqrt{\sum_{i=1}^{N}(y_B(i)-\overline{y_B})^2}} \tag{13}$$

Where, R is the similarity between A and B, $\overline{y_A}$, $\overline{y_B}$ are the mean value of A and B respectively, $\overline{y_{AB}}$ is the mean value of A and B products, $\left\|y_A-y_B\right\|$, $\left\|y_B-y_A\right\|$ are distance norm of set A and B, set B and A respectively.

According to the above algorithm, we can draw the curve of similarity and similarity variation. Figure 3 and 4 show relation curves of similarity and fluctuation points, curves of similarity variation rate based on the above algorithms respectively. Table 1shows the similarity values and similarity variation rate of these algorithms.

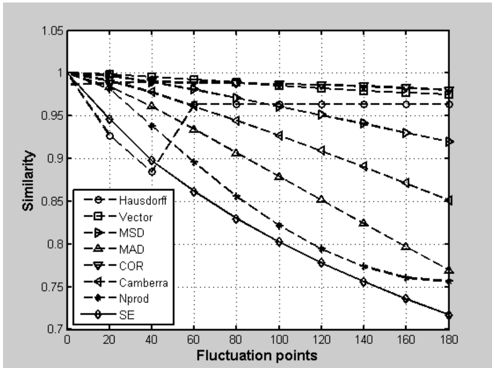

Figure 3. Relation curves of similarity and fluctuation points in different algorithms

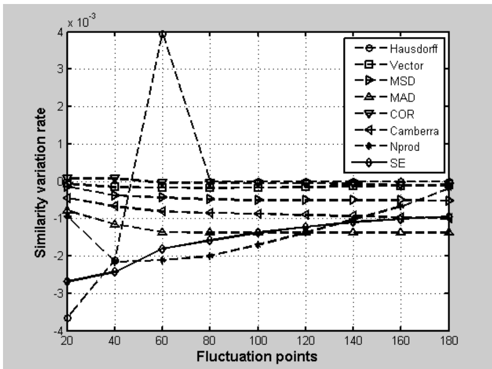

Figure 4. Curves of similarity variation rate in different algorithms

TABLE 1. SIMILARITY VALUES AND ITS VARIATION RATE OF DIFFERENT ALGORITHMS (SIMILARITY VALUES / SIMILARITY VARIATION RATE ($\times 10^{-3}$))

Algorithm	Fluctuation points									
	0	20	40	60	80	100	120	140	160	180
Vector	1.00	0.99/ -0.06	0.99/ -0.15	0.99/ -0.16	0.98/ -0.17	0.98/ -0.17	0.98/ -0.15	0.97/ -0.13	0.97/ -0.11	0.97/ -0.09
COR	0.98	0.98/ 0.07	0.98/ 0.07	0.98/ -0.04	0.98/ -0.04	0.98/ -0.04	0.98/ -0.05	0.98/ -0.06	0.98/ -0.09	0.98/ -0.12
MAD	1.00	0.98/ -0.79	0.96/ -1.15	0.93/ -1.36	0.90/ -1.37	0.87/ -1.37	0.85/ -1.37	0.82/ -1.37	0.79/ -1.37	0.76/ -1.37
MSD	1.00	0.99/ -0.14	0.98/ -0.37	0.98/ -0.43	0.97/ -0.48	0.96/ -0.50	0.95/ -0.51	0.94/ -0.51	0.93/ -0.51	0.91/ -0.51
Camberra	1.00	0.99/ -0.44	0.97/ -0.66	0.96/ -0.81	0.94/ -0.84	0.92/ -0.87	0.90/ -0.90	0.89/ -0.93	0.87/ -0.96	0.85/ -1.00
SE	**1.00**	**0.94/ -2.68**	**0.89/ -2.43**	**0.86/ -1.80**	**0.82/ -1.58**	**0.80/ -1.37**	**0.77/ -1.21**	**0.75/ -1.09**	**0.73/ -1.01**	**0.71/ -0.94**
Hausdorff	1.00	0.92/ -3.66	0.88/ -2.11	0.96/ 3.94	0.96/ 0.00	0.96/ 0.00	0.96/ 0.00	0.96/ 0.00	0.96/ 0.00	0.96/ 0.00
Nprod	1.00	0.98/ -0.93	0.93/ -2.16	0.89/ -2.11	0.85/ -2.00	0.8/ -1.70	0.79/ -1.37	0.77/ -1.01	0.76/ -0.66	0.75/ -0.19

(1) When the difference between A and B is gradually increasing, the similarity of Vector, COR and MSD vary little. The difference of similarity value cannot show the degree of difference between data set A and B. Thus, the three kinds of algorithms are not sensitive to the local fluctuation of graphic, and the reliability of these algorithms is low. It is apparent that contrast graphic A and standard graphic B are not similar when there are big fluctuations in A, but the similarity values of the three algorithms are still 0.97, 0.98 and 0.91, which may result in wrong recognition in practical application.

(2) The similarity value of Hausdorff algorithm varies greatly when a small number of fluctuation points are added, indicating a high sensitivity to local fluctuation. But with increasing of the number of fluctuation points, the similarity value doesn't vary any more, shown in Figure 4. This is because the degree of difference between data set A and B is determined by Hausdorff distance of the most mismatching points. Once the Hausdorff distance of the new added point is not more than the original one, the similarity value will not vary any more. Therefore, Hausdorff algorithm is not applicable to recognize the entire similarity between two data sets A and B.

(3) When the difference between A and B is gradually increasing, the similarity of MAD, Camberra and Nprod vary greatly. The difference of similarity value can show the degree of difference between data set A and B. It shows that the reliability of these three algorithms is high.

(4) SE similarity variation can effectively reflect the degree of difference of data sets A and B. This algorithm is very sensitive to the local fluctuation of graphic. Figure 3, 4 and table 1 show that the reliability of SE algorithm is the highest in all matching algorithms.

IV. THE APPLICATION OF SIMILARITY MATCHING ALGORITHM IN GRAPHIC RECOGNITION

Figure 5 shows the displacement-load graphics in the oil field. The standard graphic (SG) is the displacement-load graphic of oil well in normal working conditions. Contrast graphics (CG), as shown in Figure 5, are the displacement-load graphic of oil well in abnormal working conditions. The current reservoir permeability can be recognized by the similarity of standard graphic and contrast graphic.

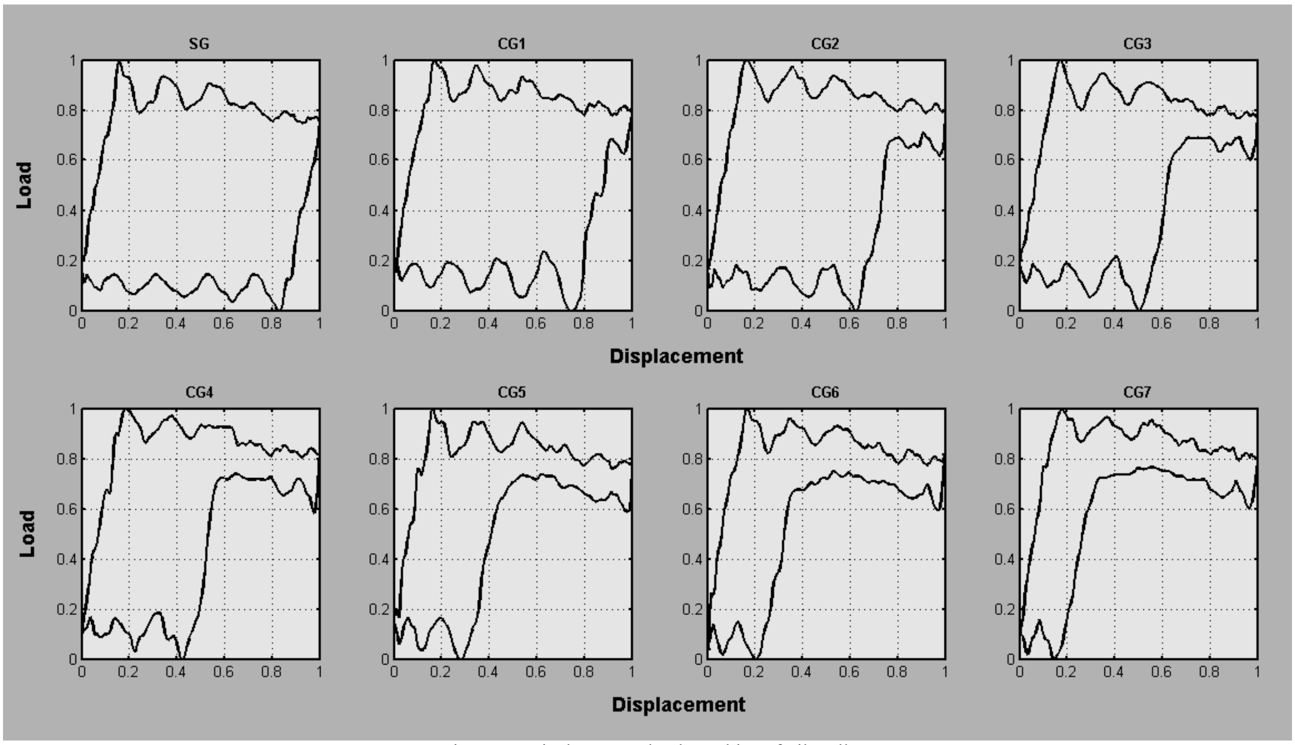

Figure 5. Displacement-load graphics of oil well.

Vector, MSD, COR, MAD, Camberra, Hausdorff, Nprod and SE algorithms are applied to recognize the similarity of standard graphic and contrast graphic. Figure 6 shows the similarity curves of standard graphic and contrast graphic based on the above algorithms. Table 2 shows the similarity values of standard graphic and contrast graphic based on these algorithms.

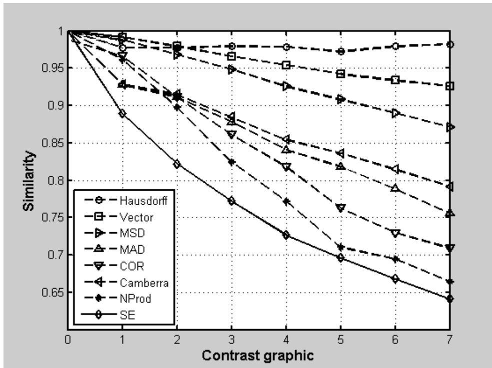

Figure 6. The similarity curves of standard graphic and contrast graphic

TABLE 2 THE SIMILARITY VALUES OF STANDARD GRAPHIC AND CONTRAST GRAPHIC

Algorithm	SG	CG1	CG2	CG3	CG4	CG5	CG6	CG7
Vector	1.00	0.991	0.979	0.965	0.954	0.942	0.933	0.926
COR	0.988	0.966	0.911	0.860	0.817	0.762	0.729	0.709
MAD	1.00	0.927	0.911	0.878	0.840	0.818	0.788	0.755
MSD	1.00	0.987	0.968	0.948	0.925	0.907	0.889	0.871
Camberra	1.00	0.927	0.914	0.884	0.854	0.835	0.814	0.791
SE	**1.00**	**0.889**	**0.821**	**0.772**	**0.726**	**0.696**	**0.668**	**0.641**
Hausdorff	1.00	0.977	0.977	0.979	0.978	0.972	0.979	0.982
Nprod	1.00	0.961	0.897	0.824	0.771	0.710	0.694	0.664

As shown in Figure 6 and table 2, when the difference between standard graphic and contrast graphic is gradually increasing, we can get the following conclusions.

(1) The similarity of Vector and MSD vary little, this is consistent with the results of adding fluctuation points, which shows the reliability of these two algorithms is low.

(2) The similarity of COR varies greatly, and the similarity of Hausdorff varies little, they are inconsistent with the results of adding fluctuation points. It shows that the reliability of these two algorithms is not stable.

(3) The similarity of MSD and MAD are basically consistent with the results of adding fluctuation points.

(4) The similarity of MAD, Nprod and SE vary greatly, and they are consistent with the results of adding fluctuation points. Thus, these three algorithms have high reliability. Figure 6 and table 2 show that the reliability of SE algorithm is the highest in all algorithms.

V. CONCLUSION

This paper first proposes a SE algorithm to realize the accurate similarity recognition of graphic, and the reliability of this algorithm and the existing similarity matching algorithms are compared and analyzed. An argument about matching algorithm correlating to the degree of difference between graphics is first proposed, and the degree of correlation determines quality of an algorithm. Algorithm of high correlation degree has high sensitivity to the recognition of small difference graphics, i.e. high recognition accuracy. Contrariwise, Algorithm of low correlation degree has less sensitivity to the recognition of small difference graphics, i.e. low recognition accuracy. Figure3, 4, 6 table1 and 2 amply demonstrate this argument. Research results show that SE algorithm has high correlation to the degree of difference between graphics. Application of this algorithm will greatly improve the accuracy and precision of the recognition.

ACKNOWLEDGMENTS

The authors would like to thank the Shaanxi Science and Technology Innovation Scheme (2015KTZDGY06-02), Key Problems of Industrial Science and technology of Shaanxi Province (2015GY110), (2016GY-185) and the Local Service Scheme of Shaanxi Province Education Department (15JF027) for financial support.

REFERENCES

[1] J. H. Zhao, K. Zhang, H. M. and Zhang, et al, “Study on underwater terrain matching positioning based on the integration of TERCOM and MCTM,” OCEANS'10 IEEE Sydney, OCEANSSYD 2010. Sydney, Australia. 2010.

[2] X. F. Wang, C. Guan and Z. Wang, “Study on entropy-based grey correlation fault diagnosis of oil pumping well indicator diagram,” Control and Instrument in Chemical Industry, vol. 40, pp. 1370-1373, 2013.

[3] W. Wu, G. d. Chen and Y. He, “Fault diagnosis system for pump work indicating diagram based on neural network and gray-level matrix,” Journal of Xi’an Shiyou University (Natural Science Edition), vol. 22, pp. 119-121, 2007.

[4] W. T. Lu, H. L. Wang and C. Liu, et al, “Design and simulation of terrain matching aided navigation system for UAVs,” Electronics Optics & Control, vol. 21, pp. 63-67, 2014.

[5] N. N. Li, X. Y. An and G. J. Tang, et al, “Synthetic terrain-matching algorithm used in integrated navigation of cruise missile,” Flight Dynamics, vol. 26, pp. 60-63, 2008.

[6] K. Wu and L. Zhao, “Research on the TAN fitting for aviation,” Piezoelectrics & Acoustooptics, vol. 32, pp. 754-757, 2010.

[7] C. X. Fan, Y. Zhang and Z. Q. Hu, “Improving project research of digital map application arithmetic in terrain matching assist navigation,” Science of Surveying and Mapping, vol. 35, pp. 89-90, 2010.

[8] Y. N. Li, Y. Wang and Z. Y. Yan, “High-speed terrain matching algorithm and its realization on FPGA,” Journal of Henan Institute of Engineering, vol. 24, pp. 57-62, 2012.

[9] H. Liu, Y. Q. Gao and J. Shen. “Underwater terrain matching techniques based on combination of PMF and TERCOM algorithms,” Torpedo Technology, vol. 20, pp. 437-442, 2012.

[10] H. Cheng, J. W. Tian and J. Ma, et al, “Selection of suitable 3D terrain matching filed based on dual-neighbor pattern and proximate distance,” Journal of Astronautics, vol. 28, pp. 230-232, 2008.

[11] B. Ye, X. L. Hu and Y. Y. Zhang, et al, “3D terrain matching algorithm and performance analysis based on 3D zernike moments,” Journal of Astronautics, vol. 28, pp. 1241-1244, 2007.

[12] D. S. Ma, M. H. Wang and F. Nie, et al, “Terrain matching algorithm based on contour line and hausdorff distance,” Journal of Projectiles, Rockets, Missiles and Guidance, vol. 29, pp. 81-84, 2009

[13] G. Tong and Y. Cha, “Simulation of Confidence Algorithm of Underwater Terrain Matching,” Journal of Chinese Inertial Technology, vol. 19, pp. 549-552, 2011.