Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 14, No. 1, 2025

Study on Oil Production Conditions Recognition Based on Standard Error Algorithm

Xinhao Yan[1, *]

[1] School of Mechanical Engineering, Xi'an Shiyou University, Xi'an 710000, China
[*] Corresponding author: Xinhao Yan (Email: 13137882873@163.com)

Abstract: The dynamometer card contains plenty of information features of equipment status, technology, and conditions in oil production. It is of great significance to identify such important information accurately, rapidly, and conveniently for safe and reliable operation of the pumping unit, improving oil production technology, and real-time monitoring of oil production performance. A new method was presented to recognize oil production equipment conditions based on a standard error algorithm and also develop a mathematical model of a standard error matching algorithm. An argument about matching algorithms correlating to the degree of difference between graphics is first proposed. The reliability of the standard error matching algorithm and classic matching algorithm are compared in detail. An intensive study shows that the standard error matching algorithm highly correlates to the degree of difference between graphics, and the reliability of the standard error algorithm is better than that of the classical matching algorithms. A large number of dynamometer card recognitions have shown that standard error-matching algorithm has a fairly high reliability. Furthermore, this algorithm has terrific recognition accuracy especially to those dynamometer cards with minute differences.

Keywords: Standard error; oil production conditions; dynamometer cards; recognition technique.

1. Introduction

At present, indicator diagram diagnosis and identification technology has been developed rapidly in China and has become the most important practical tool for oil well condition diagnosis. China currently has hundreds of thousands of pumping wells. In recent years, the oilfield has been committed to the research of oil well condition diagnosis technology. Because the indicator diagram contains multidimensional information characteristics of oil production conditions, the establishment of a sample indicator diagram database of oil well conditions is the basis of oil well conditions diagnosis technology. Through the real-time acquisition of the pumping unit indicator diagram, and the comparison and identification with the sample database, the working condition of the oil production system is determined, to realize the real-time monitoring and production guidance of the oil well working condition.

The key to oil well condition diagnosis is the accurate identification of indicator diagrams. At present, the main methods used in indicator diagram recognition are the grid method, gray matrix method, vector method, artificial neural network method, etc. The core of the above method is feature extraction, which is to extract the relevant feature information from the two-dimensional array of indicator diagrams of different scenes and eliminate the irrelevant information. The second is to establish a classifier, which is a kind of matching algorithm to identify the working conditions. The key to indicator diagram recognition is how to extract its most representative features and what kind of classifier to use for type recognition.

Both the grid method and gray matrix extract the information of the indicator diagram boundary through the characteristic matrix, which is either true (1) or false (0) [1, 2]. Due to the fluctuation of the contour line of the indicator diagram, there is a possibility that different curves will have great differences when passing through the same grid point. As shown in Fig. 1 (a) and (b), both A and B curves pass through the same grid, and the value is 1. However, the local difference between a curve and a B curve is very large, so the extracted feature matrix cannot accurately reflect the local features of the indicator diagram. Even if the classifier algorithm is very accurate, it is still inaccurate. The artificial neural network method [3-5] is also based on the grid method and gray matrix to extract feature information, so it still has the defects of the grid method and gray matrix. These three methods belong to feature recognition, they are only applicable to the classification and recognition of graphics, and cannot realize the accurate recognition of graphics. The vector method belongs to pattern recognition, its characteristic matrix is the quotient of matrix elements and matrix modulus, and its matching value is the cosine value of the vector angle. Both the grid method and gray matrix extract the information of the indicator diagram boundary through the characteristic matrix, which is either true (1) or false (0) [1, 2]. Due to the fluctuation of the contour line of the indicator diagram, there is a possibility that different curves will have great differences when passing through the same grid point. As shown in Fig. 1 (a) and (b), both A and B curves pass through the same grid, and the value is 1. However, the local difference between the A curve and the B curve is very large, so the extracted feature matrix cannot accurately reflect the local features of the indicator diagram. Even if the classifier algorithm is very accurate, it is still inaccurate. The artificial neural network method [3-5] is also based on the grid method and gray matrix to extract feature information, so it still has the defects of the grid method and gray matrix. These three methods belong to feature recognition, they are only applicable to the classification and recognition of graphics, and cannot realize the accurate recognition of graphics. The vector method belongs to pattern recognition, its characteristic matrix is the quotient of matrix elements and matrix modulus, and its matching value is the cosine value of the vector angle.

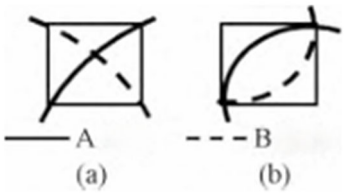

Figure 1. Difference diagram of dynamometer card contour through the grid

With the continuous accumulation of the historical data of the indicator diagram of the oilfield, the types of oil production conditions that have been identified are more and more accurate and refined. The fine classification of indicator diagrams of the same working condition type can reach several or even dozens. The accuracy of the final condition-type diagnosis results depends heavily on the quality of the contour extraction of the indicator diagram, Or the low correlation between the matching algorithm and the degree of graphic difference is low. Therefore, they cannot meet the requirements of fine identification of modern indicator diagrams.

This paper proposes a matching algorithm based on the standard error (SE) algorithm to realize the matching between the indicator diagram to be identified and the sample indicator diagram. The recognition algorithm based on Se theory can also be applied to the accurate recognition of two-dimensional graphics. At present, there are many algorithms to measure the similarity of graphics. The main algorithms are cross-correlation algorithm [6] (COR), mean absolute difference algorithm (MAD) [6,7], mean square deviation algorithm (MSD) [6,8,9], camper distance algorithm [10], Hausdorff algorithm [8,11], normalized product correlation algorithm (NPROD) [12], which are the most widely used and classic graphics recognition algorithms. The reliability of these algorithms is studied in detail.

2. Working Condition Identification of Indicator Diagram Based on Standard Error Algorithm

2.1. Isometric displacement and dimensionless calibration of the indicator diagram

The indicator diagram data is a discrete two-dimensional array [S_i, W_i] (i=1, 2,...), S_i is the displacement, and W_i is the load. Since the displacement value points of the indicator diagram are different, it is necessary to carry out S_i coordinate equivalent calibration for the indicator diagram to be identified and the sample indicator diagram. Our value can be obtained by the Lagrange algorithm.

The geometric contour shape of a figure is the relevant element to be extracted, while the size of the actual [S_i, W_i] value is irrelevant to the figure recognition. Therefore, the graph must be dimensionless calibrated.

$$\begin{cases} Y_{\mathrm{d}} = \dfrac{(W_i - W_{min})}{(W_{max} - W_{min})} \\ X_{\mathrm{d}} = \dfrac{(s_i - s_{min})}{(s_{max} - s_{min})} \end{cases} \tag{1}$$

In equation (1), $W_{\max}$ is the maximum load on the dynamometer diagram, N. $W_{\min}$ is the minimum load on the dynamometer diagram, N. $S_{\max}$ is the maximum displacement, m. $S_{\min}$ is the minimum displacement, m。

2.2. Classifier Algorithm——Standard Error

In a dimensionless coordinate system, the similarity between two geometric shapes is characterized by the degree of convergence of the same displacement point. The indicator diagram consists of a dataset, with the calibration sample indicator diagram dataset being the true value and the identification indicator diagram dataset being the measured value. In this way, the similarity between the indicator diagram to be identified and the sample indicator diagram can be defined as the degree to which the measured value approaches the true value, that is, the deviation between the measured value and the true value, which is called the standard error (SE). SE is very sensitive to large or small errors in a set of measurement values, so, SE can effectively reflect the degree to which the identified indicator diagram dataset approaches the sample indicator diagram dataset.

The SE of the indicator diagram dataset A (y_A, x_A) is to be identified and the sample indicator diagram dataset B (y_B, x_B) is

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

SE matching value range: $0 \le d \le 1$. Two geometric shapes that are completely similar have d=0; The closer the geometric shape, the smaller the SE value of the shape. 0 similarity is an equivalent scale used to measure the degree of similarity between two geometric shapes. The similarity of completely similar shapes is 1, and the more similar the shapes are, the higher their similarity. Define the similarity between two graphics as

$$R = 1 - d \tag{3}$$

By performing similarity recognition between the comparison image and the reference image, the degree of similarity of the graphics can be accurately determined based on the similarity level.

3. Research on the Correlation Between Similarity Matching Algorithm and Graphic Difference Degree

At present, the main similarity-matching algorithm models based on graphics include Vector, COR, MAD, MSD, Comberra, Hausdorff, and NProd algorithms. These algorithms are classic algorithms for determining whether two shapes are similar. However, there is currently no report in domestic and foreign literature on whether these algorithms can accurately determine the similarity of graphics, or which algorithm can more accurately determine the degree of similarity of graphics. The argument that there is a correlation between the similarity-matching calculation model and the degree of geometric difference has been proposed for the first time. The size of the matching value should be able to reflect the degree of difference between two geometric shapes, that is, the greater the difference between two geometric shapes, the greater the difference in their matching values should also be, and they should have monotonicity. Therefore, the correlation between the similarity matching algorithm and the

degree of graphical difference between the identified indicator diagram A and the sample indicator diagram dataset B based on dimensionless calibration can be expressed as follows: the change in matching value Δ d has the same trend as the degree of graphical difference Δ R.

$$\Delta d(A,B) \propto \Delta R(A,B) \tag{4}$$

Figure 2 shows a sample dynamometer diagram B. Applying fluctuations locally to this sample diagram can obtain the dynamometer diagram A to be recognized, as shown in Figure 3. By gradually increasing the fluctuation points of Dataset A, the difference between Dataset A and Dataset B is increased. Apply Vector, COR, MAD, MSD, Comberra, Hausdorff, NProd, and SE algorithms to study the changing trends of matching values.

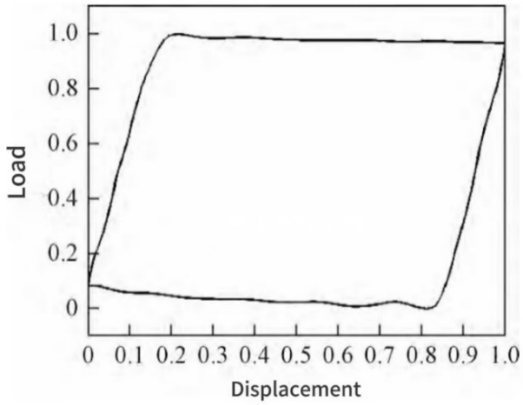

Figure 2. Sample dynamometer card B

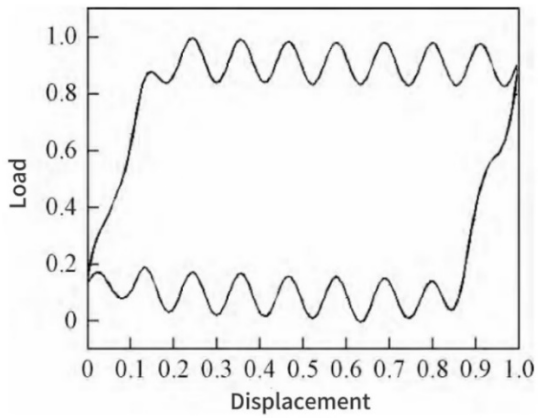

Figure 3. Waiting for recognizing dynamometer card A

3.1. Vector Algorithm

The indicator diagram to be identified and the sample indicator diagram are represented by 2n dimensional vectors, and their matching algorithm is the dot product of the two, which is the cosine value of the vector angle.

$$R_n(A,B) = A^{\mathrm{T}} \cdot B \tag{5}$$

3.2. COR Algorithm

Cross-correlation algorithm.

$$R_n(A,B) = \frac{\frac{1}{N}\sum_{i=1}^{N}[y_A(x)y_B(x)-\overline{y_{AB}}(x)][r_B(x)-\overline{y_B}(x)]}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}[r_{AB}(x)-\overline{y_{AB}}(x)]^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}[V_B(x)-\overline{y_B}(x)]^2}} \tag{6}$$

3.3. MAD Algorithm

Mean absolute difference algorithm.

$$R_n(A,B) = 1 - \frac{1}{N}\sum_{i=1}^{N}|y_A(x) - y_B(x)| \tag{7}$$

3.4. MSD Algorithm

Mean square deviation algorithm.

$$R_n(A,B) = 1 - \frac{1}{N}\sum_{i=1}^{N}[y_A(x) - y_B(x)]^2 \tag{8}$$

3.5. Camberra Algorithm

The Camberra distance between the identified indicator vector and the sample indicator vector can be used as a distance measure between the two feature vectors.

$$R_n(A,B) = 1 - \frac{\sum_{i=1}^{N}|y_A(x) - y_B(x)|}{\sum_{i=1}^{N}|y_A(x) + y_B(x)|} \tag{9}$$

3.6. SE Algorithm

Standard error algorithm.

$$R_n(A,B) = 1 - \sqrt{\frac{1}{N}\sum_{i=1}^{N}[y_A(x) - y_B(x)]^2} \tag{10}$$

3.7. Hausdorff Algorithm

The Hausdorff distance is defined by a general mathematical formula, which describes a measure of the similarity between two sets of points. It is a definition of the distance between two sets of points.

$$\begin{aligned} &R_n(y_A,y_B) = 1 - max[d_{AB}(y_A,y_B), d_{BA}(y_B,y_A)] \\ &d_{AB} = \mathrm{maxmin}_{y_A\in Ay_B\in B}\|y_A - y_B\| \\ &d_{BA} = \mathrm{maximin}_{y_B\in By_A\in A}\|y_B - y_A\| \end{aligned} \tag{11}$$

3.8. NProd Algorithm

Normalized product correlation algorithm.

$$R_n(A,B) = \frac{\sum_{i=1}^{N}[y_A(x)-\overline{y_A}(x)][y_B(x)-\overline{y_B}(x)]}{\sqrt{\sum_{i=1}^{N}[y_A(x)-\overline{y_A}(x)]^2}\sqrt{\sum_{i=1}^{N}[y_B(x)-\overline{y_B}(x)]^2}} \tag{12}$$

In equations (5) to (12), $\overline{y_A}, \overline{y_B}$ The mean values of datasets A and B respectively; N=1, 2, 3..., where N is the number of fluctuating points; $\|\cdot\|$is the distance norm of datasets A and B (e.g., L2 or Euclidean distance).

According to the above algorithm, a similarity curve can be drawn. The relationship curve between similarity and fluctuation points of the above algorithm is shown in Figure 4. Table 1 shows the similarity values of the above algorithm.

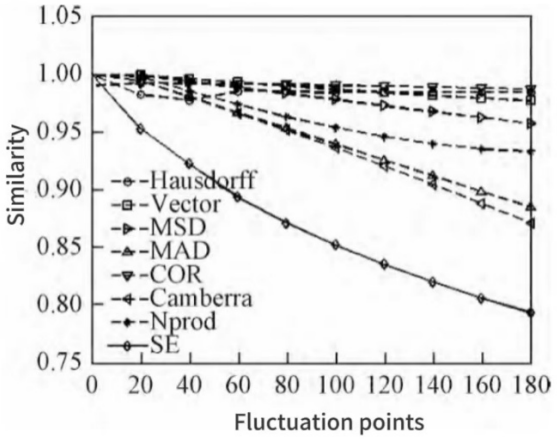

Figure 4. Relation curves of similarity and fluctuation points indifferent algorithms

Table 1. Similarity value of different algorithms

Algorithm	Similarity									
Vector	1.000	0.999	0.996	0.993	0.990	0.987	0.984	0.981	0.979	0.977
COR	0.990	0.992	0.992	0.991	0.991	0.990	0.990	0.989	0.988	0.987
MAD	1.000	0.994	0.981	0.967	0.953	0.939	0.926	0.912	0.898	0.885
MSD	1.000	0.999	0.994	0.988	0.983	0.978	0.973	0.967	0.962	0.957
Camberra	1.000	0.994	0.981	0.966	0.951	0.936	0.921	0.905	0.888	0.871
SE	1.000	0.953	0.922	0.894	0.871	0.852	0.836	0.820	0.807	0.794
Hausdorff	1.000	0.982	0.977	0.985	0.985	0.984	0.984	0.984	0.984	0.984
Nprod	1.000	0.998	0.985	0.974	0.963	0.954	0.946	0.940	0.935	0.933

As shown in Figure 4 and Table 1, without applying point fluctuations, datasets A and B are completely similar, The similarity of Vector, MAD, MSD, Camberra, SE, Hausdorff, and NProd algorithms is all 1, while the similarity of COR algorithm is an uncertain value. As the number of fluctuation points in dataset A gradually increases, it indicates that the degree of difference between the identified indicator diagram A and the sample indicator diagram B will gradually increase, that is, the degree of dissimilarity between A and B will increase. Research has shown that.

(1) As the degree of difference between datasets A and B gradually increases, The similarity changes of Vector, COR, MSD, and NProd are very small, and the difference in similarity cannot reflect the degree of difference between the identified indicator diagram A and the sample indicator diagram B. This indicates that these four algorithms have low sensitivity to local fluctuations in the graph and low reliability. When the fluctuation of the indicator diagram A to be identified is significant, it clearly indicates that the indicator diagram A to be identified is not similar to the sample indicator diagram B. However, the matching values of Vector, COR, MSD, and NProd are still equal to 0.977, 0.987, 0.957, and 0.933, respectively. In actual discrimination, this may lead to incorrect identification.

(2) As the number of fluctuation points increases, The Hausdorff similarity value no longer changes, as shown in Figure 4 and Table 1. This is because the degree of difference between the two datasets A and B is determined by the Hausdorff distance of the least similar point. Once the Hausdorff distance of the newly added wave point is not greater than the distance of that point, its similarity will no longer change. Therefore, The Hausdorff algorithm is not suitable for discriminating overall similarity between two datasets A and B.

(3) As the degree of difference between datasets A and B gradually increases, The similarity changes between MAD and Camberra basically reflect the degree of difference between datasets A and B.

(4) The similarity change of SE can well reflect the degree of difference between datasets A and B. This algorithm is highly sensitive to local fluctuations in graphics. Figure 4 and Table 1 indicate that The reliability of the SE algorithm is the highest among all matching algorithms. Figure 5 shows the displacement load dynamometer diagram of the oil well, the sample dynamometer diagram NWC is the dynamometer diagram under standard operating conditions of the oil well, and the to-be-identified dynamometer diagram AWC1-7 is the dynamometer diagram under non-standard operating conditions of the oil well. The difference between AWC1-7 and NWC gradually increases. Apply Vector, MSD, COR, MAD, Camberra, Hausdorff, NProd, and SE algorithms to perform similarity recognition on NWC and AWC1-7. The similarity relationship curve between NWC and AWC1-7 of the above algorithm is shown in Figure 6. Table 2 shows the

similarity between NWC and AWC1-7.

As shown in Figure 6 and Table 2, when the difference between the sample dynamometer and the dynamometer to be identified gradually increases, we can draw the following conclusions: (1) The similarity changes of Vector, MSD, and Hausdorff are small, which is consistent with the conclusion of point fluctuation, indicating that the reliability of the algorithm is not high. (2) The similarity between COR and NProd varies greatly, which is inconsistent with the results of fluctuating application points, indicating that the reliability of these two algorithms is unstable, and the similarity between COR shows non-monotonic changes. (3) The conclusion regarding the similarity changes and application point fluctuations between MAD and Camberra is consistent. (4) The similarity of SE changes the most and is highly consistent with the fluctuation of the application point. Therefore, The SE algorithm has high reliability. Figure 6 and Table 2 indicate that The reliability of the SE algorithm is the highest among all algorithms.

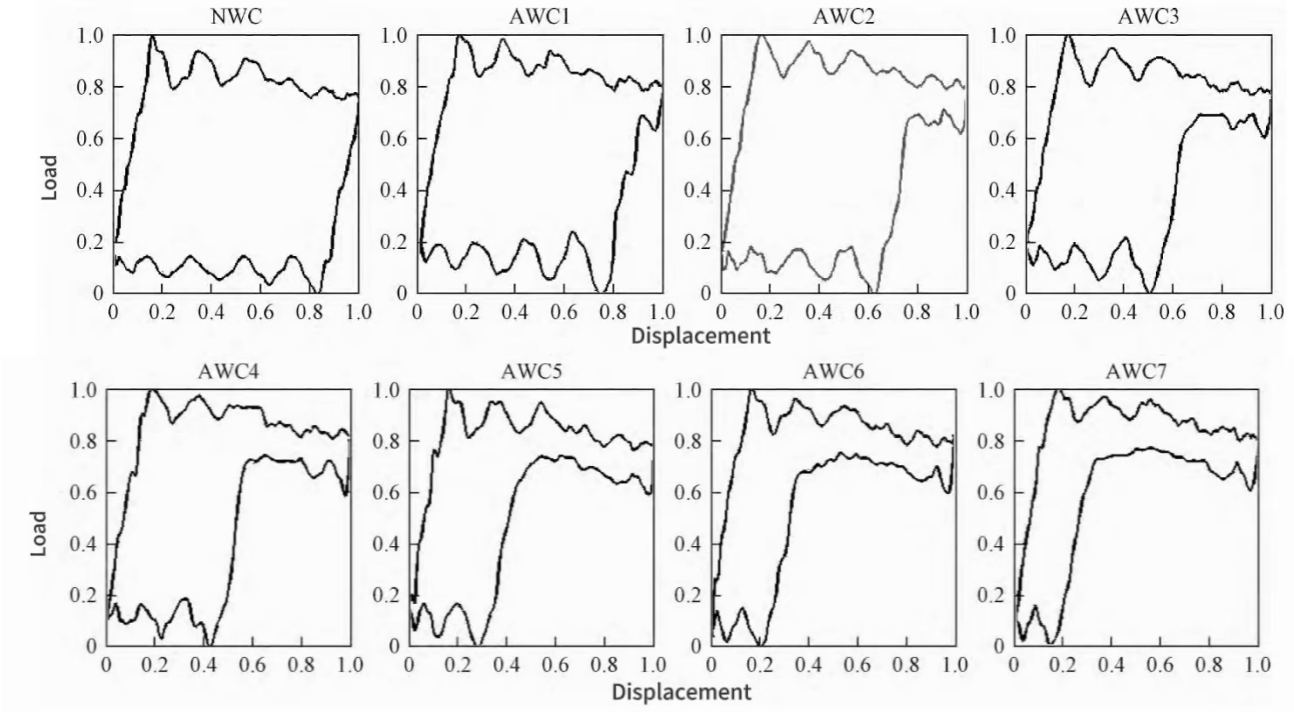

Figure 5. Displacement-load dynamometer card of oil well

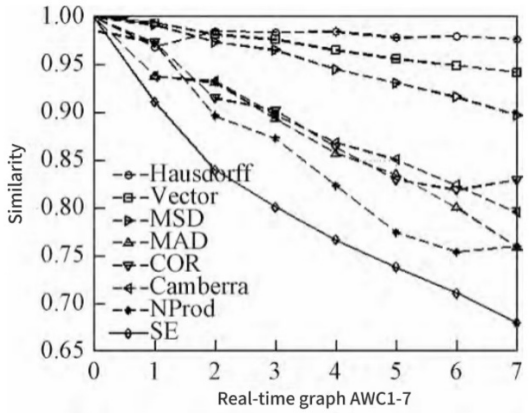

Figure 6. The similarity curves of NWC and AWC1-7

4. Application Examples

Nearly 10000 dynamometer diagrams were collected from more than 1000 wells in the Chuankou Oil Production Plant of Yanchang Oilfield, and nearly 300 suspected faulty pump dynamometer diagrams were selected. Table 3 shows some of the main fault types selected. The SE algorithm was used to perform similarity recognition between the indicator diagram to be recognized and the established sample indicator diagram database. The recognition results are shown in Table 3, and the recognition success rate reached 100%. This algorithm is particularly suitable for identifying indicator diagrams with slight differences. Table 3 shows the results of identifying the standard error of oil production conditions.

Table 2. The similarity values of NWC and AWC1-7

Algorithm	Similarity							
	NWC	AWC1	AWC2	AWC3	AWC4	AWC5	AWC6	AWC7
Vector	1.0000	0.9948	0.9819	0.9769	0.9655	0.9563	0.9495	0.9425
COR	0.9861	0.9735	0.9153	0.9023	0.8640	0.8298	0.8189	0.8299
MAD	1.0000	0.9385	0.9318	0.8937	0.8573	0.8357	0.8007	0.7590
MSD	1.0000	0.9922	0.9744	0.9656	0.9457	0.9314	0.9164	0.8975
Camberra	1.0000	0.9384	0.9331	0.8977	0.8685	0.8510	0.8243	0.7957
SE	1.0000	0.9115	0.8399	0.8010	0.7669	0.7381	0.7109	0.6798
Hausdorff	1.0000	0.9688	0.9855	0.9846	0.9848	0.9786	0.9805	0.9772
Nprod	1.0000	0.9731	0.8969	0.8732	0.8237	0.7744	0.7542	0.7604

Table 3. Results of production condition recognition by SE algorithm (%)

Failure Mode	Similarity											
	1	2	3	4	5	6	7	8	9	10	11	12
1. Fixed Valve Leakage	99.9	76.3	79.1	77.6	80.5	79.3	80.4	82.8	77.1	76.8	82.1	53.1
2. Moving Valve Leakage	82.9	99.7	73.1	91.0	75.1	86.6	69.9	80.1	82.1	73.3	83.0	57.2
3. Tubing Breakage	78.6	72.6	97.8	72.4	82.5	72.7	59.3	79.6	71.8	76.9	74.4	59.2
4. Sucker Rod Upper Breakage	77.7	91.0	72.9	98.6	71.8	86.3	71.9	76.3	84.0	72.3	79.8	59.0
5. Tubing Bending	80.2	74.9	82.6	71.7	99.2	75.1	59.0	84.6	71.6	78.7	76.9	56.5
6. Sand Lock	79.1	80.8	74.3	85.2	75.5	96.5	74.2	81.5	85.8	78.1	86.2	61.7
7. Gas Interference	61.9	70.1	59.9	72.2	59.3	74.4	98.4	63.3	77.9	67.6	69.1	66.1
8. Gearbox Wear	83	78	81	77	85	82	64	98	79	86	83	57.9
9. Belt Slippage	76.7	81.8	72.1	83.9	71.7	85.8	78.2	77.0	98.8	76.3	80.0	65.6
10. Gas Lock	73.8	71.2	74.7	70.9	75.4	75.6	68.6	79.7	75.4	95.4	79.7	61.5
11. Subsurface Pump Collision	82.1	82.9	74.9	79.8	77.1	85.8	68.8	83.4	80.0	81.7	99.1	62.6
12. Stuck Pump	52.9	57.1	59.1	58.9	56.2	60.8	65.5	56.5	65.2	61.9	62.2	99.2

5. Conclusions

For the first time, the SE algorithm was proposed to achieve accurate recognition of oil well conditions, and the algorithm was compared and analyzed with existing similarity recognition matching algorithms. The argument that similarity matching algorithms have a great correlation with the degree of graphic differences was also put forward for the first time, and the degree of correlation determines the superiority or inferiority of the algorithm. Algorithms with high correlation have high sensitivity to recognizing graphics with slight differences, indicating high accuracy in recognition. On the contrary, algorithms with low correlation have low sensitivity to recognizing graphics with small differences, that is, the accuracy of recognition is low. Figures 4, 6, Table 1, and Table 2 fully demonstrate this argument. The research results indicate that there is a high degree of correlation between the SE algorithm and the graphics, and the reliability of this algorithm is the highest among all algorithms. The application of this algorithm reduces the probability of misjudging similar working conditions and improves the accuracy and precision of recognition.

References

[1] Ding Yi, Li Xunming. The research and application of the eigenvalue extraction and selection in oilfield fault diagnosis. Electronic Design Engineering, 2014; 22(17): 148—150.

[2] Wang Xiufang, Guan Chuang, Wang Zi. Study on entropy-based grey correlation fault diagnosis of oil pumping well indicator diagram. Control and Instruments in Chemical Industry, 2013; 40 (11):1370-1373.

[3] Zhang Qiang, Xu Shaohua, Li Panchi. Pumping unit fault diagnosis based on quantum shuffled frog leaping algorithm and process neural networks. China Mechanical Engineering, 2014; 25 (12):1609-—1614.

[4] Wen Bilong, Wang Zhiqun, Jin Zongze, et al. Diagnosis of pumping unit with combing indicator diagram with fuzzy neural networks. Computer Systems & Applications, 2016; 25(1): 121—125.

[5] Li Chunsheng, Su Xiaowei, Wei Jun, et al. Research on diagrams identification of pumping unit based on support vector machine. Computer Systems & Applications, 2014; 24(8): 215-218.

[6] Lu Wentao, Wang Honglun, Liu Chang, et al. Design and simulation of terrain matching aided navigation system for UAVs. Electronics Optics & Control, 2014; 21(5):63—67.

[7] Fan Chengxiao, Zhang Ying, Hu Zhiqian. Improving project research. of digital map application arithmetic in terrain matching assist navigation. Science of Surveying and Mapping, 2010; 35(4): 89-90.

[8] Li Yingnan, Wang Yong, Yan Zhiyu. High-speed terrain matching algorithm and its realization on FPGA. Journal of Henan Institute of Engineering, 2012; 24(4): 57—62.

[9] Liu Hong, Gao Yongqi, Shen Jian. Underwater terrain matching techniques based on combination of PMF and

TERCOM Algorithms. Torpedo Technology, 2012; 20(6): 437—442.

[10] Ye Bin, Hu Xiulin, Zhang Yunyu, et al. 3D terrain matching algorithm and performance analysis based on 3D zernike moments. Journal of Astronautics, 2007;28(5):1241—1244.

[11] Ma Danshan, Wang Minghai, Nie Feng, et al. Terrain matching algorithm based on contour line and Hausdorff distance. Journal of Projectiles, Rockets, Missiles and Guidance, 2009; 29 (6):81-84.

[12] Tong Guang, Zha Yue. Simulation of confidence algorithm of underwater terrain matching. Journal of Chinese Inertial Technology, 2011 ; 19(5):549—552.