Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Wind Power Curve Data Cleaning Algorithm via Image Thresholding*

Yahao Su[1,2], Fan Chen[1], Guoyuan Liang[1,**], Xinyu Wu[1]
[1]*Guangdong Provincial Key Lab of Robotics and Intelligent System*
CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems
Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences
Shenzhen, Guangdong Province, China
{*yh.su, fan.chen, gy.liang, xy.wu*}*@siat.ac.cn*

Yong Gan[2]
[2]*School of Mechanical and Electrical Engineering*
Guilin University of Electronic Technology
Guilin, Guangxi Province, China
ganyong@guet.edu.cn

***Abstract*— Wind turbine data from the Supervisory Control And Data Acquisition (SCADA) system is very important for wind turbine conditional monitoring, wind power prediction and wind turbine performance evaluation. However, the SCADA data usually contains lots of abnormal data. This paper presents an image-based algorithm for abnormal data cleaning of wind power curve (WPC) data via image thresholding. The basic idea is to build a gray-level representation of the original binary image of WPC which is able to preserve the normal part as much as possible. Therefore, the cleaning operation is then turned into a problem of image segmentation. The proposed algorithm includes the following steps: First, the scatter data is converted into a binary image. Then the median of four distances are computed from each pixel in the image to the nearest connected domain boundary along four directions, and a gray level image is generated to strengthen the normal part, in the meantime, weaken the abnormal part. For all possible threshold t, the optimal t_o which makes the smallest Hu moment based dissimilarity of the segmented normal part with a reference WPC template, is finally determined. The proposed algorithm is compared with some data-based algorithms as well as an image-based mathematical morphology operation (MMO) algorithm. Experiments carried out on WPC data of 17 wind turbines from a wind farm verified the effectiveness and accuracy of the proposed method.**

***Index Terms*— image thresholding; wind power curve; data cleaning; Hu moment;**

I. Introduction

Wind power is an important clean energy and renewable resource [1]. More and more wind turbines have been installed around the world in the past decade. Wind turbine data from SCADA system are widely used for wind power condition analysis [2]. WPC depicts the properties of wind power turbines, now becomes a popular research object recently [3]. SCADA data usually contains different types of abnormal data. Due to various wind speed and direction, wind power, in essence, has the characteristics of fluctuation, intermittence and randomness [4]. Therefore, wind power data cleaning is an important task for wind farm operations and managements. The uncleaned WPC, however, may distort the statistics of output wind power, which will interfere with the analysis of wind turbine operation status and characteristics [5].

So far, many kinds of data cleaning algorithms have been reported. Traditional algorithms are based on the scatter data, such as K-means clustering algorithm [6], Local Outlier Factor(LOF) algorithm [7], combined algorithm based on change point grouping and quartile(CA) algorithm [8]. Hu *et al.* [9] proposed a stepwise data cleaning algorithm through irregular space-division and nonlinear space-mapping. Kusiak *et al.* [10] proposed a curve model of k-nearest neighborhood to detect abnormal data. The main problem of those algorithms is that large amounts of stacked data can not be cleaned effectively.

In this paper, we propose an image-based algorithm which can turn the scatter data from continuous space into a digital space. In this case, we just need to consider a pixel instead of many scattered data points. This means less computation cost when calculating distance or density for data points. Besides, it is more intuitive to work directly with visual images. Moreover, it's possible to apply an effective image processing algorithm based on the characteristics of abnormal data pixels. Recently, an image-based MMO algorithm proposed by Long *et al.* [11] can clean the abnormal data more effectively than traditional data-based algorithms. Inspired by MMO, this paper presents a better solution for abnormal data cleaning by thresholding on a gray level feature image generated by four-directional median length measurements.

II. The background of the WPC data cleaning

The WPC curve represents the relationship between wind speed and power of wind turbine, which can reflect the performance of the wind power control system. The ability

*This work is partially supported by Joint Funds of the National Natural Science Foundation of China with Shenzhen City (No.U1813209), Shenzhen Peacock Technology Innovation Project (No.KQJSCX20170731164301774), Fundamental Researche Project of Shenzhen City (No.JCYJ20170818153048647) and the NSFC-Shenzhen Robotics Research Center Project (U1613219).
**Corresponding author: G. Liang(email:gy.liang@siat.ac.cn)

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

of a wind turbine to produce electric power from varying wind is a function of three main factors, the wind power availability, the power curve of the machine, and the ability of the machine to respond to wind fluctuations[12]. So there are many factors can cause abnormal data in the WPC data including unplanned maintenance, wind turbine failure, wind reduction, sensor failure, sensor noise, wind reduction commands communication failures and so on. The abnormal data caused by different factors show different features in the WPC data. As is shown in Fig.1. The abnormal data based on WPC can be summarized into the following three types: Type I is negative abnormal data, Type II is scattered abnormal data, and Type III is stacked abnormal data[11].

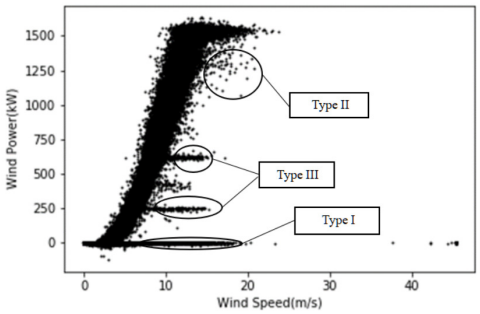

Fig. 1. Abnormal data in WPC.

The purpose of the WPC abnormal data cleaning algorithm is to remove these abnormal data. The reference WPC used in this paper is a reliable power curve model that is marked and cleaned by human expert. This model can be used as a reference for different wind turbines, but the parameters of WPC are different in different turbines. Wind power curve modeling is a challenging task in which the recorded wind power is far from the theoretical wind power[13]. There are also some differences in the curve. So an algorithm with high precision and robustness is needed to clean the data accurately and effectively.

III. Methodology and Algorithm

A. The overview of the proposed algorithm

To clean the abnormal data, we proposed an image-based algorithm which can turn the scatter data from continuous space into a digital space. In this case, we just need to consider a pixle instead of many scatter data points. This means less computation cost when calculating distance or density for data points. Then we build the gray-level representation of the original binary image of WPC which is able to perserve the normal part as much as possible. The gray value is the median of four distances from each pixel in the image to the nearest connected domain boundary along four directions, and a feature image is generated to strengthen the normal part, in the meantime, weaken the abnormal part. In this way, we effectively separate abnormal data from normal data. Therefore the cleaning operation is then turned into a problem of image segmentation. For all possible threshold t, the optimal t_o which makes the smallest Hu moment based dissimilarity of the segmented normal part with a reference WPC template, is finally determined. Finally, the WPC data is effectively cleaned according to the threshold segmentation result. The flow chart of the proposed algorithm is shown in Fig.2. Details of each step for the algorithm will be described in the following sections.

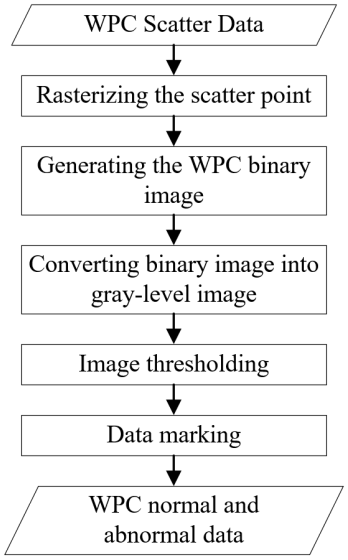

Fig. 2. Flow chart of the proposed algorithm.

B. The generation of the WPC image

1) The generation of binary image: The scatter points data need to be rasterized because it exists in continuous space. As is shown in Fig.3. The scatter points of the scatter plot can be divided into different boxes by rasterization. Each small box represents a pixel in the digital image. The value of the pixel can be obtained in many ways. In this paper, we set the pixel with scattered points to 255, the rest of the pixels are set to 0.

Based on the above principles, let (x, y) be the pixel of the WPC binary image $f(x, y)$, where $x = 1, \ldots, M, y = 1, \ldots, N$. The i-th wind point is (v_i, p_i) where v is the wind speed and p is the power. The scaling parameter between the pixel and the data point is (Δv,Δp), where Δv=0.2, Δp=7. $P_{\max}$, $P_{\min}$ represent the maximum and minimum power, $v_{\max}$, v_{min} represent the maximum and minimum wind speeds respectively. The(x_i, y_i) which can be calculated as Eq.(1). corresponding pixel is set to 255, the remaining pixels are set to 0.

$$\begin{aligned} y_i &= \lfloor (p_i - p_{\min}) / \Delta p \rfloor \\ x_i &= \lfloor (v_i - v_{\min}) / \Delta v \rfloor \end{aligned} \tag{1}$$

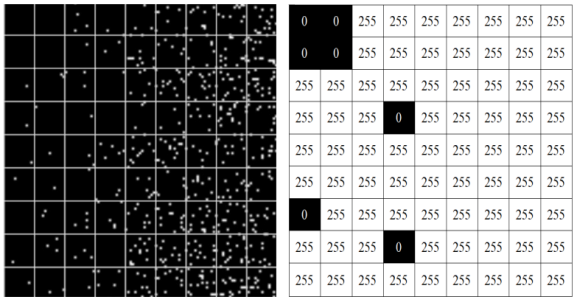

Fig. 3. Rasterize scatter points and convert the plot into the digital image. (a) Scattered point plot. (b) Binary WPC image.

The image size is $\mathrm{M} \times \mathrm{N}$ is calculated as Eq.(2). The binary image can be shown in Fig.4.

$$\begin{aligned} \mathrm{N} &= \lfloor (P_{\max} - P_{\min}) / \Delta p \rfloor \\ \mathrm{M} &= \lfloor (v_{\max} - v_{\min}) / \Delta v \rfloor \end{aligned} \tag{2}$$

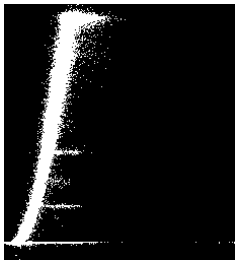

Fig. 4. Binary WPC image.

2) The generation of feature image: To distinguish the abnormal data from the normal data, we build a feature image in the second step. In this image, the pixels representing the normal data are brightened and the pixels representing the abnormal data are darkened.

This step is to calculate the distances of the valued pixels to the pixel of the connected domain boundary in four directions. In this paper, this distance is set as the number of pixels in that direction. The four directions are shown in Fig.5.

Then We can get a set of data: $X_1, \cdots, X_4$. Sort it in order from smallest to largest: $X_{(1)}, \cdots, X_{(4)}$.We can calculate the median of the data as Eq.3.

$$m_{0.5} = \frac{X_{(2)} + X_{(3)}}{2} \tag{3}$$

$m_{0.5}$ represents the gray value of the corresponding pixel in the image which can turn the binary image into the gray-level image.

To improve the uniformity of gray image, the gray-level image is firstly filtered by means and then calculated in this paper. We use the simplest of the mean filters. S_{xy} are the

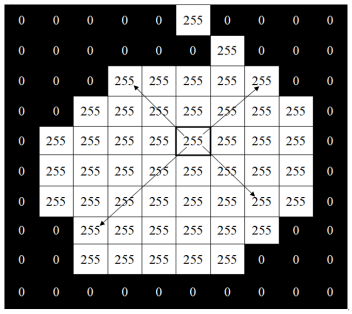

Fig. 5. The four directions when calculate the distances to the connected domain boundary, the distance is set as the number of pixels in that direction.

set of coordinates in a rectangular sub-image neighborhood of size $m \times n$ and (x, y) is the center point. The value of the center point (x, y) is replaced by the arithmetic mean in the region $g(s, t)$ defined by S_{xy}. The restored new image is $\hat{f}$. In other words,

$$\hat{f}(x, y) = \frac{1}{mn} \sum_{(s,t) \in S_{xy}} g(s, t) \tag{4}$$

Mean filters use the spatial filter of size $m \times n$ and smooth local variations in an image. Then we can improve the uniformity of the gray-level image as is shown in Fig.6.

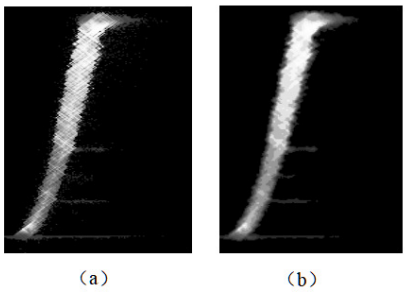

Fig. 6. (a) Original image. (b) Filtered image. The shown image has been equalized to show the distribution of the gray values.

C. Image thresholding

1) The foreground after thresholding: In the feature image, the foreground represents the normal data and the background represents the abnormal data. One obvious way to extract the foreground from the background is to select a threshold that separates these modes. Then, any point (x, y) in the image at which $f(x, y) > T$ is called a foreground point; otherwise, the point is called a background point. In other words, the segmented image, $g(x, y)$, is given by

$$g(x, y) = \begin{cases} 255 & \text{if } f(x, y) > T \\ 0 & \text{if } f(x, y) \leq T \end{cases} \tag{5}$$

In this paper, T is applied over an entire image, the process given in this equation is called global thresholding.

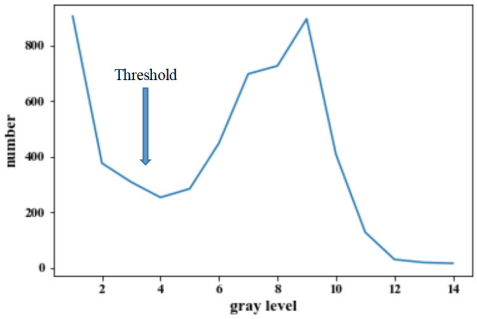

Fig. 7. The line graph of the filtered image, the horizontal coordinate represents the gray level, the vertical coordinate represents the number of pixels corresponding this gray level (under the threshold we can get the foreground and the background of the image).

Fig. 8. The foreground after image thresholding when threshold equals to 3, 4 and 5.

2) Hu moment dissimilarity measurement: Hu moment is a linear combination of normalized central moments. Hu moment can maintain the invariance in the image rotation, zoom, translation and other operations. So often use Hu moment to identify the characteristics of the image[14]. A set of seven invariant moments $\phi_1, \phi_2, \ldots, \phi_7$ can be derived from the second and third moments of the image. To calculate the image dissimilarity, the Hu moments are transferred as

$$m_i = \text{sign}\left(\phi_i\right) \cdot \log\left(\phi_i\right), i = 1 \ldots 7 \tag{6}$$

The ith Hu moments of image a and image b are transferred as $m_{a,i}$ and $m_{b,i}$. The dissimilarity $D(a,b)$ of the image a and image b is calculated as

$$D(a,b) = \sum_{i=1}^{7} \left| \frac{1}{m_{a,i}} - \frac{1}{m_{b,i}} \right| \tag{7}$$

When the threshold t changes, the corresponding dissimilarity between different segment contours and template WPC are calculated as $d(t)$.

3) The optimal thresholding and data marking: When the threshold is t, the smaller dissimilarity $\mathrm{d(t)}$ between the segmentation and the reference WPC, the better the segmentation result of current threshold t.

When threshold t is changing, $\mathrm{d(t)}$ is shown in table I. Where $\mathrm{d(t)}$ get the smallest value, and the most reasonable segmentation result is obtained as shown in Fig.9.

TABLE I

D(T) UNDER DIFFERENT THRESHOLDS

Threshold t	Dissimilarity d(t)
3	6.2173
4	9.1495
5	12.4696
6	33.6704
7	41.8459
8	30.8477
9	27.7144
10	24.5566
11	22.7469
12	23.7082
13	23.9665

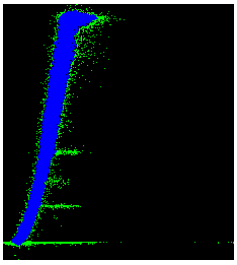

Fig. 9. Thresholding result, the blue pixels represent normal data, the green pixels represent abnormal data

D. Data marking

The last step is to extract the normal data from abnormal data in the scatter data. According to the image thresholding result, we can effectively distinguish between pixels that represent normal data and those that represent abnormal data. Suppose (x, y) is the pixel corresponding the normal data, with the scaling parameter (Δv,Δp) between the pixel and the data point, we can get the normal data points (v_i, p_i) in this pixel. Then we can extract the normal data representing normal pixels from the former rasterization step. The scatter plot of the normal scatter data and the abnormal scatter data is shown in Fig.10.

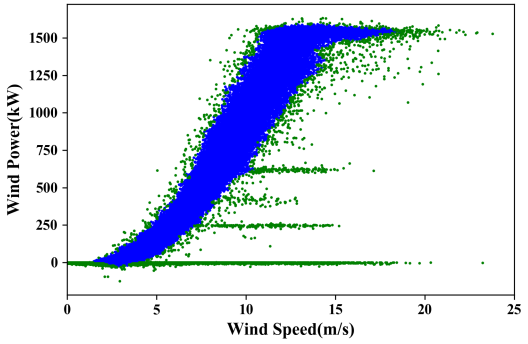

Fig. 10. Data cleaning result, the blue scatter points represent normal data, the green scatter points represent abnormal data.

IV. Experiments

A. The experimental results

To prove the effectiveness and generalization of this algorithm. The SCADA data of 17 wind turbines from real wind farm are being cleaned. Some of the result is shown in Fig.11.

The specification of wind turbines are shown in Table II. The cleaning rate of image thresholding algorithm is shown in Table III.

TABLE II

THE WIND TURBINE SPECIFICATION

Specification	Wind turbine data
Cut-in Speed(m/s)	3.0
Rated Speed(m/s)	13.0
Cut-out Speed	25.0
Rate Power(kw)	1500
Number of Wind Turbines	17

TABLE III

THE ABNORMAL DATA CLEANING RATE

WT name	MMO	LOF	CA	K-means	Thresholding
	Cleaning rate				
A1-007	12.04	9.99	7.47	13.48	12.74
A1-008	19.47	9.92	10.87	10.61	20.32
A2-019	11.40	9.93	10.39	17.73	48.31
A2-020	19.53	9.95	10.67	16.54	15.47
A2-021	20.00	9.94	10.15	12.28	39.22
A2-024	22.33	9.93	12.79	10.94	45.57
A2-026	19.13	9.93	11.32	10.92	41.00
A2-027	16.29	9.91	7.21	11.76	15.44
A2-028	18.85	9.89	13.32	11.49	43.71
A2-029	13.78	9.94	8.94	11.74	11.62
B2-045	8.93	9.94	9.38	13.63	12.90
B2-046	12.24	9.94	5.12	16.22	31.51
B3-062	16.49	9.94	6.81	10.17	18.41
B3-063	18.17	9.94	7.48	12.36	14.90
B3-064	14.25	9.95	7.10	11.40	25.11
B3-065	13.26	9.89	5.61	14.81	17.88
B3-066	16.78	9.96	5.12	14.12	14.52

As is shown in table III and Fig.11. The proposed algorithm can work well in 12 wind turbines. But in some wind turbines, the proposed algorithm has a bad performance. Because the scaling invariance of Hu moment, the prospect with smaller shape has lower dissimilarity.

B. Comparison with other algorithms

To prove the effectiveness of the proposed image thresholding algorithm, compare this algorithm with other algorithms such as image-based MMO algorithm, k-means clustering algorithm [6], LOF algorithm [7], CA algorithm [8]. The cleaning rata of the compared algorithm is cited from the paper proposed by long *et al.* [11] and shown in Table III. However, the cleaning rate does not mean the higher the better too. We need more accurate quantitative indicators in the future work.

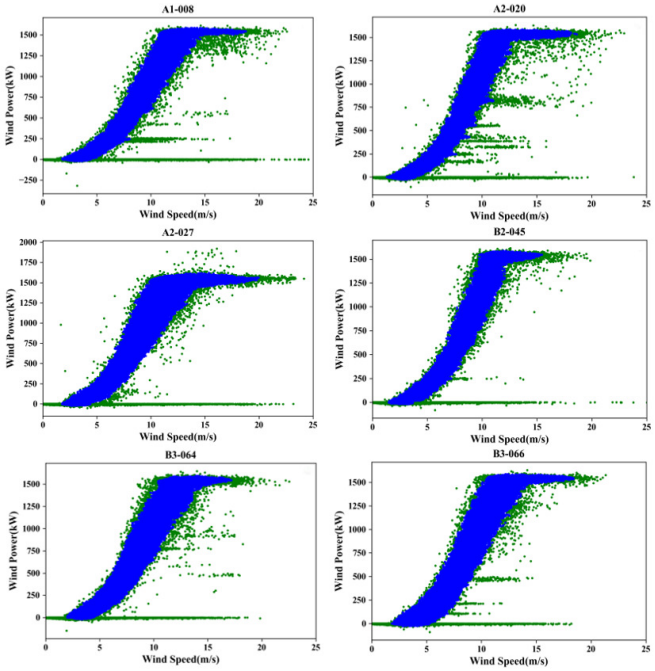

Fig. 11. The results of image thresholding algorithm.

Traditional data based cleaning algorithms like k-means clustering algorithm, LOF algorithm, and CA algorithm have some problems with the classification of stacked data and scattered data as is shown in Fig.12.

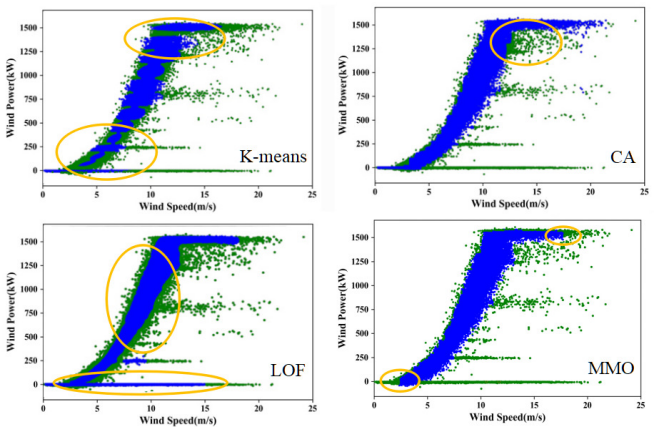

Fig. 12. The results of other algorithms.

The algorithm of this paper is mainly compared with the image-based MMO algorithm proposed by Long *et al.* [11]. The proposed algorithm and the MMO algorithm are implemented by Python and operated on the compute with CPU Intel(R) Core(TM) i7-8700 @3.2GHz, RAM 16GB. The MMO algorithm is much faster than traditional algorithms [11]. As is shown in Table IV, although the proposed algorithm is a little slower than MMO algorithm, it is basically at the same

TABLE IV

THE COMPUTATIONAL TIME(S) OF MMO ALGORITHM AND THE PROPOSED ALGORITHM.

WT name	MMO	The proposed algorithm				
		step1	step2	step3	Cleaning	Total
A1-007	1.04	0.14	0.67	0.13	0.79	1.72
A1-008	1.01	0.12	0.65	0.14	0.76	1.66
A2-019	1.01	0.13	0.59	0.09	0.83	1.55
A2-020	1.13	0.13	0.62	0.12	0.80	1.67
A2-021	1.03	0.13	0.56	0.13	0.79	1.59
A2-024	1.02	0.13	0.55	0.11	0.78	1.57
A2-026	0.99	0.13	0.41	0.07	0.78	1.39
A2-027	1.05	0.13	0.73	0.14	0.77	1.78
A2-028	0.97	0.13	0.45	0.07	0.78	1.43
A2-029	1.05	0.13	0.68	0.15	0.79	1.75
B2-045	1.03	0.13	0.62	0.12	0.82	1.68
B2-046	1.01	0.13	0.52	0.09	0.78	1.52
B3-062	0.92	0.12	0.49	0.07	0.73	1.41
B3-063	0.92	0.12	0.44	0.07	0.74	1.37
B3-064	0.92	0.12	0.50	0.07	0.73	1.42
B3-065	0.99	0.13	0.64	0.12	0.74	1.63
B3-066	0.99	0.13	0.64	0.12	0.74	1.63

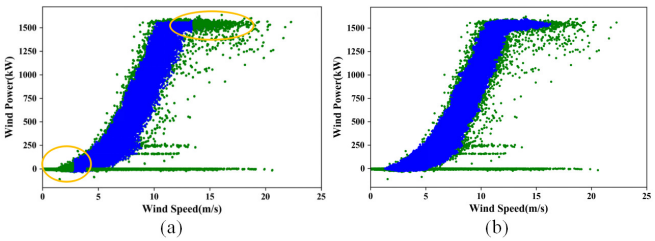

Fig. 13. (a) The result of MMO for B3-063. (b) The result of image thresholding for B3-063.

level. The step 1 is generating the binary WPC image, the step 2 is generating the gray-level image and the step 3 is image thresholding. The last step is cleaning the abnormal data in the scatter data according to the image thresholding result. The image thresholding algorithm is based on the gray-level WPC image can consider the position of the data in four directions which means more information is included. The main problem of the MMO algorithm is that all the highlighted parts of the image will be eliminated, and the normal data will be cleaned as is shown in Fig.13.

V. CONCLUSION

This paper presents an image-based algorithm for abnormal data cleaning of WPC data via image thresholding. The wind power curve data were converted to the binary image and then converted to the gray-level image. The proposed algorithm is mainly based on the this feature image generated by four-directional median length measurement. Then the Hu moment based dissimilarity of the segmented normal part with a reference WPC template is calculated to guide the image thresholding. The wind turbine SCADA data from 17 wind turbines are cleaned by the algorithm to prove the effectiveness and generalization of this algorithm. Compared with the data-based algorithm and other image-based algorithm, the proposed image-based algorithm has the best performance in the cleaning of abnormal data. But in some cases, because of the scaling invariance of Hu moment, the prospect with smaller shape has lower dissimilarity. We need other indicators to guide the image thresholding segmentation in future works.

VI. ACKNOWLEDGEMENT

The authors would like to thank Dr. Long Huan for providing data for experiments as well as her valuable comments on the algorithm described in this paper.

REFERENCES

[1] G. Shen, B. Xu, Y. Jin, S. Chen, W. Zhang, J. Guo, H. Liu, Y. Zhang, and X. Yang, "Monitoring wind farms occupying grasslands based on remote-sensing data from china's gf-2 hd satellite—a case study of jiuquan city, gansu province, china," *Resources, Conservation and Recycling*, vol. 121, pp. 128 – 136, 2017.

[2] P. Cambron, R. Lepvrier, C. Masson, A. Tahan, and F. Pelletier, "Power curve monitoring using weighted moving average control charts," pp. 126–135, 2016.

[3] M. Lydia, S. S. Kumar, A. I. Selvakumar, and G. E. P. Kumar, "A comprehensive review on wind turbine power curve modeling techniques," *Renewable & Sustainable Energy Reviews*, vol. 30, no. Complete, pp. 452–460, 2014.

[4] D. Yi, C. Singh, L. Goel, J. Ostergaard, and W. Peng, "Short-term and medium-term reliability evaluation for power systems with high penetration of wind power," *IEEE Transactions on Sustainable Energy*, vol. 5, no. 3, pp. 896–906, 2014.

[5] S. Swapna, P. Niranjan, B. Srinivas, and R. Swapna, "Data cleaning for data quality," in *International Conference on Computing for Sustainable Global Development*, 2016.

[6] M. Yesilbudak, "Partitional clustering-based outlier detection for power curve optimization of wind turbines," in *IEEE International Conference on Renewable Energy Research and Applications*, 2017.

[7] Z. Le, H. Wei, and M. Yong, "Raw wind data preprocessing: A data-mining approach," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 1, pp. 11–19, 2017.

[8] X. Shen, X. Fu, and C. Zhou, "A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 1, pp. 46–54, 2018.

[9] Y. Hu, Y. Xi, C. Pan, G. Li, and B. Chen, "Daily condition monitoring of grid-connected wind turbine via high-fidelity power curve and its comprehensive rating," *Renewable Energy*, vol. 146, pp. 2095 – 2111, 2020.

[10] A. Kusiak, A. Verma, and S. Member, "Monitoring wind farms with performance curves," *IEEE Transactions on Sustainable Energy*, vol. 4, no. 1, pp. 192–199, 2013.

[11] H. Long, L. Sang, Z. Wu, and W. Gu, "Image-based abnormal data detection and cleaning algorithm via wind power curve," *IEEE Transactions on Sustainable Energy*, pp. 1–1, April 2019.

[12] T. Üstüntaş and A. D. Şahin, "Wind turbine power curve estimation based on cluster center fuzzy logic modeling," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 96, no. 5, pp. 611–620, 2008.

[13] W. Yun, Q. Hu, D. Srinivasan, and W. Zheng, "Wind power curve modeling and wind power forecasting with inconsistent data," *IEEE Transactions on Sustainable Energy*, vol. PP, no. 99, pp. 1–1, 2018.

[14] M. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.