The Analysis of Features Importance in Electrical Infrared Images Faults Diagnosis

Qi Zhao
School of Mathematical Sciences,
Fudan University
Shanghai, China
hellozhaoqi@gmail.com

Lei Su
State Grid Shanghai Electric Power
Research Institute
Shanghai, China

Zhaoyang Shi
School of Mathematical Sciences,
Fudan University
Shanghai, China

Ping Ling
State Grid Shanghai Electric Power
Research Institute
Shanghai, China

Nannan Yan
State Grid Shanghai Maintenance
Company, Shanghai
Shanghai, China

Chunjie Gu
State Grid Shanghai Maintenance
Company
Shanghai, China

Zhixiong Shi
State Grid Shanghai Municipal
Electric Power Company
Shanghai, China

ABSTRACT

This paper proposes an effort to study on input features and classifier about fault detection and diagnosis of infrared images within electrical installations and employ the results in fault detection robots with the purpose to improve the accuracy and be more convenient saving the waste of money and workforce. The features are extracted from infrared images of electrical equipments and classified by using random forest algorithm. In the experiments, the classification performances of various input features are evaluated. The commonly used indicators to describe the classification performance, including sensitivity(TPR), specificity(TNR), accuracy(ACC) and area under curve (AUC) are employed to identify the most suitable input feature as well as the best configuration of classifiers. The results of the experimental demonstrate that the combination of features including Skewness, Max, Kurtosis, 0.95 percentile, Gradient Direction Histogram, Max-Min, 0.75 percentile and other features can result in the best effect for infrared images fault diagnosis. In addition, the Random Forest performs better than the support vector machine(SVM) using radial basis kernel function (RBF) or gaussian kernel function. At the end of the paper, we further discuss how those features effect the fault diagnosis with infrared images. And as for employing the results in fault detection robots, it shows a considerably good effect in terms of accuracy and convenience compared with some traditional methods. It endows those robots with the abilities to make a real time monitoring and detect the faults.

IoPARTS '18, June 10, 2018, Munich, Germany

ACM ISBN 978-1-4503-5843-9/18/06.
https://doi.org/10.1145/3215525.3215539

CCS CONCEPTS

• **Mathematics of computing → Trees**; • **Computing methodologies → Bagging**; **Feature selection**; • **Computer systems organization →** *Real-time systems*; • **General and reference →** *Experimentation*;

KEYWORDS

infrared image; faults diagnosis; features importance; random forest

ACM Reference Format:

Qi Zhao, Lei Su, Zhaoyang Shi, Ping Ling, Nannan Yan, Chunjie Gu, and Zhixiong Shi. 2018. The Analysis of Features Importance in Electrical Infrared Images Faults Diagnosis. In *Proceedings of IoPARTS '18: International Workshop on Internet of People, Assistive Robots and ThingS (IoPARTS '18).* ACM, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.1145/3215525.3215539

1 INTRODUCTION

In the field of power grid, it is of great significance to improve the reliability of the systems, monitor the safety of each device and set up an early warning plan. With the development of data science and machine learning algorithm, fault detection robots with those mentioned technologies can offer a lot of help thus saving the waste of money and workforce [3].

As for fault detection and diagnosis with infrared images, there exist many methods including those using support vector machine(SVM) [1, 2], nonsubsampled contourlet transform(NSCT) [1], Hu moments features and texture features [6], Otsu thresholding and valley emphasis method [11]. However, due to the restraint of the camera and the image data, there is still much work to do to improve the accuracy of the results. Therefore, to find features with good effect and build an evaluating system for them is what this paper puts emphasis on.

The image data are from the FLIR camera which has the capacity to take visible light pictures and infrared pictures simultaneously,

Figure 1: Fault detection robots

Table 1: Example of artificial fault detection rule [2]

Faults	Changes of temperature	Meanings
1	less than 0.5 $°C$	early stage of overheating, needing to be controlled
2	5-30 $°C$	developed overheating, needing to be fixed
3	more than 30$°C$	strong overheating, needing to be fixed instantly

although there are papers dealing with extra ultraviolet pictures [5]. Infrared pictures can provide the information of temperature of devices and the corresponding environment [8, 14]. Visible light pictures can provide the location of each device and some information about the physical damages.

Researches show that temperature is one of the most reliable factors [2]. Temperature of those damaged device often deviates from the normal values (as an example in **Table 1**). Therefore, to process the data about the temperature of devices and the corresponding environment, this paper uses many indicators like skewness, kurtosis, the maximum and minimum value and the quantile of the data.

However, with the advent of the era of big data, machine learning is widely used in all areas of the power grid. Using machine learning technologies, we can more precisely detect subtle changes in temperature, and more easily dig out the non-linear correlation between the faults and their temperature, making fault detection robots more accurate and more capable than experts to some degree.

The existing papers are mainly focused on the image enhancement [7]. But besides improving the quality of image information, in order to have better accuracy of fault diagnosis, information about the target devices is more important with some proper features extracted. This paper has adopted many methods to not only filter the noise of images, highlight the information of target devices but also extract some features to improve the accuracy of the fault diagnosis with the results (AUC=0.9178; ACC=0.9787; TPR=0.7619; TNR=0.8844). The mainly process of this paper includes five points: fundamental information and related methods of fault detection and diagnosis of power grid, data collecting and processing, algorithms and implementation, results and evaluating systems, discussion and conclusion. And in the end of the conclusion, we especially make some comparison between fault detection robots using our methods and some traditional methods to detect faults with infrared images. There are two highlight merits of this paper: effective methods of electrical thermal faults diagnosis and good evaluating systems to analysis the importance of features.

2 METHOD

2.1 Features Selection

To better describe the effective information on infrared images, we have extracted 60 statistics and Gradient's features from the infrared images. The details about features and reasons for selecting them are listed in the following **Table 2**.

2.2 Random Forest

The random forest is a machine learning algorithm based on the decision tree. The random forest algorithm is robust to situations when there exist missing data and unbalanced data. Additionally, the speed of the calculation is relatively fast. More remarkably, the advantages of random forests over multidimensional features can also be used for feature selection. Therefore, the random forest algorithm has been widely used in various classification, prediction, feature selection and outlier detection problems [9, 10, 12].

2.3 Analysis of Features Importance Based on Random Forest

The feature selection method based on the Gini index of node impurity and the classification accuracy rate based on OOB (out-of-bag) data are the most two commonly used feature selection methods in random forests [9]. For a given node t, its corresponding data set is D and the Gini index is defined as

$$G(D) = 1 - \sum_{k=1}^{Q} p^2(k|D) \tag{1}$$

Where $p(k|D)$,k=1,...,Q is the proportion of k-type samples on D. For the two-class classification problem, $Q = 2$. $G(D)$shows the probability of randomly selecting two samples from data set D whose class labels are inconsistent. Therefore, the smaller the $G(D)$ is, the higher the purity will be. When we measure the Gini index for each discrete attribute $j = 1, \dots, m$

$$G_j(D) = \sum_{i=1}^{V} \frac{|D^v|}{|D|} G(D^v) \tag{2}$$

Where $|D^v|$ represents the number of samples of the j^{th} possible value of the j^{th} attribute. For continuous attributes, we can do the same calculation after discretizing it. For each attribute j, the average $\overline{G_j}$ of K-trees in the Gini index represents the importance.

$$\Delta(G_j) = 1 - \overline{G_j} \tag{3}$$

Table 2: Statistical characteristics

Index	Features	Description of extraction features
1-2	Size of infrared image	The resolution of images taken by different infrared devices may be different.
3-5	Max,Min and Max-Min of infrared image	The maximum temperature is often the important factor to determine whether the equipment is malfunctioning or not, while ignoring the impact of ambient temperature on the maximum.
6-10	95%,75%,50%,25%,5% percentile of infrared image	Compared to the maximum and minimum values, the quantile is more robust and can remove noise caused by environmental factors such as lighting.
11-15	Mean,Standard Deviation, Entropy ,Skewness, Kurtosis	These are common statistics of distribution and can effectively describe the distribution of the overall temperature.
16-35	Gradient Magnitude Histogram	From the perspective of the gradient, the temperature changes on the device can be clearly characterized. We did histogram statistics on Magnitude and direction and chose Prewitt as a gradient operator. Magnitude's range is from 0-100, divided into 20 intervals.
36-60	Gradient Direction Histogram	The direction is -180 degrees to 180 degrees, and every 15 degrees is an interval.

2.4 Feature-Based Random Forest Algorithm

We adopted a random forest algorithm based on feature selection. The characteristics of the variable importance of the random forest are used to rank the features in the descending order, and the former p% is taken as the feature subset. Randomly dividing the training set and the test set, 80% as the training set and 20% as the test set, and after calculating the average accuracy, we get the mean and standard deviation of the AUC values after multiple times of averaging. After the algorithm was redesigned and improved, the whole algorithm is in the following algorithm.

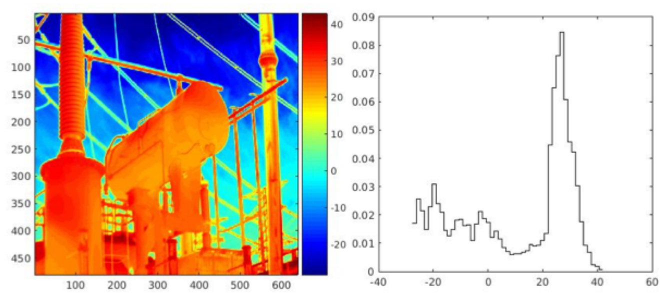


Figure 2: Infrared image and the temperature distribution of faulty insulator

3 RESULTS

3.1 Data Description

We use information from infrared images gathered by power grids in one area of East China in the past year. Totally, there were 19,633 photos, of which 270 were abnormal infrared photos and 19,363 were normal photos.

3.2 feature Selection

Form Feature-Based Random Forest Algorithm step5, we obtain values about the importance of variables, from which we can see Skewness, Max, Kurtosis, 95% percentile, Gradient Direction Histogram, Max-Min, 75% percentile and other features play an important role in the classification.

3.3 Classification Result

Due to the imbalance of data, we can't focus on the accuracy rate only, so we use three categories of AUC values: Accuracy (ACC), Sensitivity (TPR) and Specificity (TNR) to measure the results of the classification. At the same time, we also compare the classification results using support vector machines (SVM). The following data are the results of repeating 500 times. The feature used by the SVM is the one whose importance is greater than zero. And the threshold for calculating the ACC, TPR, and TNR values is set to $270/19,633 \approx 0.01375$, $n_{tree} = 500$, $m_{try} = 7$.

From the **Table 3**, we firstly use max and 95% percentile temperature which are what the inspection robots are using in current fault diagnosis. Then, we compare the random forest with SVM, which is widely used in the classification. We can see that the AUC and TNR value of the random forest are significantly better than those of SVM, indicating that the Random forest performs better about fault detection and it shows some difficulties to classify unusual

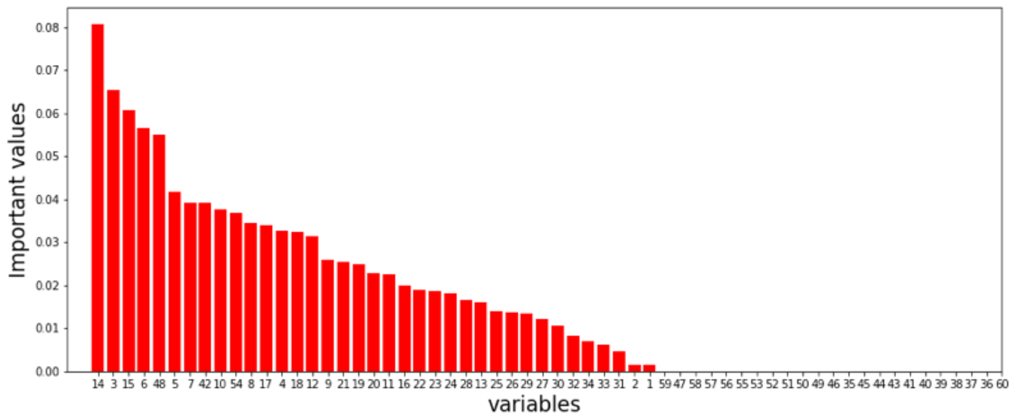


Figure 3: Variables important values

Algorithm 1 Feature-Based Random Forest

Require: Raw data set $S = \{x_1, x_2, ..., x_n\}$
Ensure: Importance of Classification Accuracy, Average and Standard Deviation of AUC Values and Their Corresponding Feature Sets Rank

step1:Initialize, set the number of repeat training N_1 and N_2
for $i = 1$ to N_1 **do**
  Step2: In the raw data set S, randomly select K training subsets with the bootstrap placed back, and construct K independent categorical regression trees. The complement of each training subset is denoted as out-of-bag. (OOB)
  Step3: For m features, m_{try} features are randomly selected in each node of each tree to generate feature subsets, and the Gini index of each feature is calculated. The feature with the smallest Gini index in the feature subset is selected as a split feature.
  Step4: The maximum growth of each tree, not pruning.
  Step5: Average the importance of each feature, arranged in descending order, get Rank, and take the former $p\%$ as the feature subset.
end for
for $i = 1$ to N_1 **do**
  Step6: Randomly disassemble the data set S and divide 80% training set and 20% test set. Repeat Step 2-4.
  Step7: The generated K-trees are composed of random forests, and the classification result is determined by the vote of the tree classifier.
  Step8: Calculate the importance of each feature and the accuracy of the model in the test set and the AUC value.
  Step9: Calculate the mean and standard deviation of accuracy and AUC values.
end for
Where $m_{try} = floor(log_2(m) + 1)$

pictures in normal pictures, which is of more importance in robots' fault detection.

4 CONCLUSIONS

We also tested the SVM using all the features of the classification to make a comparison and found the effect of the classification did not improve, indicating that these features are difficult to be detected in the SVM, thus revealing the superiority of the random forest for unbalanced data sets. Further, we analyzed the effect of Top N features on the classification performance. From the above

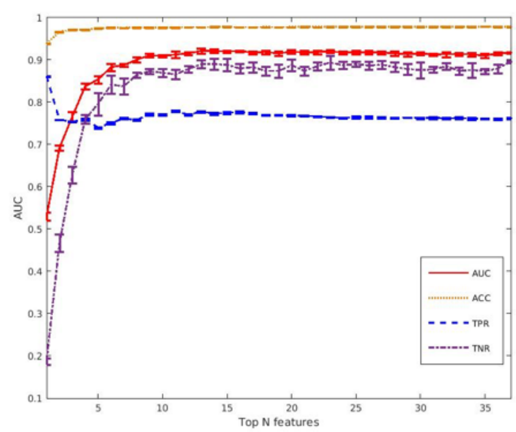


Figure 4: Top N features on classification performance

figure, we find that the first 7 features have a great influence on the classification performance. And it shows that those features like Skewness, Max, Kurtosis, 95% and 75% percentile play a considerably great role. Statistic indicators like Skewness and Kurtosis represent the trend of deviation and the peakedness of the data which mean that the temperature, especially the maximum of it, does serve as an important warning signal of damaged devices.

And results show that fault detection robots using our methods with those features can offer a better outcome compared with some conventional methods like using only the temperature threshold in

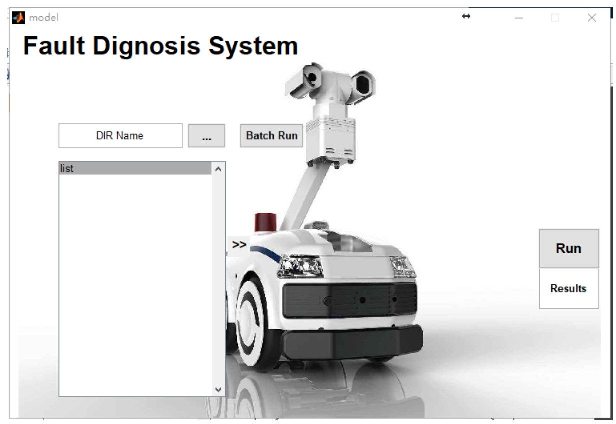
Table 3: Prediction performance between different classifiers

Classifier		AUC	ACC	TPR	TNR
SVM(default)	Mean	0.5256	0.9778	0.9908	0.0452
	std	0.0046	0.0013	0.0014	0.0066
SVM(kernel:'Gaussian')	Mean	0.7410	0.9861	0.9958	0.2933
	std	0.0067	0.0002	0.0008	0.0121
SVM(kernel:'RBF')	Mean	0.7277	0.9856	0.9953	0.2956
	std	0.0070	0.0002	0.0002	0.1030
Random Forest	Mean	**0.9178**	0.9787	0.7619	**0.8844**
	std	0.0038	0.0006	0.0015	0.0121

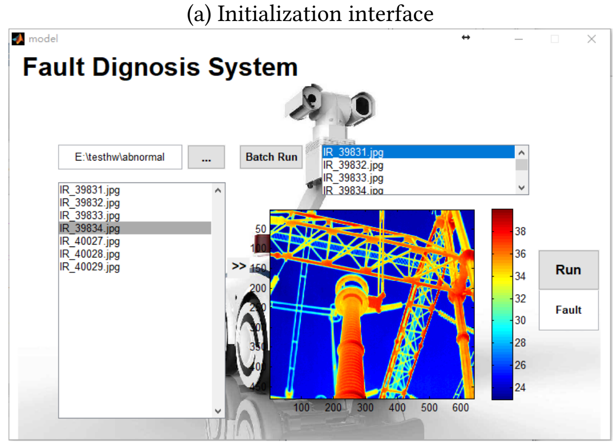
terms of accuracy and convenience.

As for improving the quality of image data, some existing papers have mentioned digital image segmentation using functions like edge identification, gradient administration, Laplacian, Hough light simple and adaptive threshold [4]. Those functions can lead to a successful segmentation. Therefore, the data from the image will be more accurate to locate the target devices.

As for improving the effect of classification, there are many alternative methods about features extracting like principal direction extracting [13], which can have a great advantage over the extraction of features thus improving the effect of classification.



(a) Initialization interface



(b) Display results

Figure A1: real-time fault diagnosis system

A REAL-TIME FAULT DIAGNOSIS SYSTEM

In this appendix,we use MATLAB GUI to real-time display the diagnosis results.We can diagnose the image one by one under the button of **Run** and batch diagnose the fault images under the button of **Batch Run**.

ACKNOWLEDGMENTS

We thank all the volunteers, publications support, staff, and authors who wrote and provided helpful comments on previous versions of this document. We gratefully acknowledge the grant from State Grid Corporation of Science and Technology Project, China (52090016002M).

REFERENCES

[1] Tangbo Bai, Laibin Zhang, Lixiang Duan, and Jinjiang Wang. 2016. NSCT-based infrared image enhancement method for rotating machinery fault diagnosis. *IEEE Transactions on Instrumentation and Measurement* 65, 10 (2016), 2293–2301.
[2] Y Laib dit Leksir, M Mansour, and A Moussaoui. 2017. Localization of thermal anomalies in electrical equipment using Infrared Thermography and support vector machine. *Infrared Physics & Technology* (2017).
[3] Li Li Hongwei Li, Binhai Wang. 2010. Research on the infrared and visible Power-Equipment image fusion for inspection robots. *Applied Robotics for the Power Industry* (2010).
[4] Deepak Kumar, Amit Kumar, and MA Ansari. 2018. A Flexible Scheme to Fault Detection for Electrical Assets Using Infrared Thermography. In *Advances in Electronics, Communication and Computing*. Springer, 537–548.
[5] Yu Li, Fengyuan Yu, Qian Cai, Kun Yuan, Renzhuo Wan, Xiaoying Li, Meiyu Qian, Pengfeng Liu, Junwen Guo, Juan Yu, et al. Image fusion of fault detection in power system based on deep learning. *Cluster Computing* (????), 1–9.
[6] Hang Liu, Ting Xie, Jian Ran, and Shan Gao. 2017. An Efficient Algorithm for Server Thermal Fault Diagnosis Based on Infrared Image. In *Journal of Physics: Conference Series*, Vol. 910. IOP Publishing, 012031.
[7] Tao Liu, Wei Zhang, and Shaoze Yan. 2015. A novel image enhancement algorithm based on stationary wavelet transform for infrared thermography to the debonding defect in solid rocket motors. *Mechanical Systems and Signal Processing* 62 (2015), 366–380.
[8] Jae-Do Park. 2015. Ground fault detection and location for ungrounded DC traction power systems. *IEEE Transactions on Vehicular Technology* 64, 12 (2015), 5667–5676.
[9] David M Reif, Alison A Motsinger, Brett A McKinney, James E Crowe, and Jason H Moore. 2006. Feature selection using a random forests classifier for the integrated analysis of multiple data types. In *Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB'06. 2006 IEEE Symposium on*. IEEE, 1–8.
[10] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC*

bioinformatics 9, 1 (2008), 307.

[11] M Suguna, S Mohamed Mansoor Roomi, and I Sanofer. 2016. Fault localisation of electrical equipments using thermal imaging technique. In *Emerging Technological Trends (ICETT), International Conference on*. IEEE, 1–3.

[12] Antanas Verikas, Adas Gelzinis, and Marija Bacauskiene. 2011. Mining data with random forests: A survey and results of new tests. *Pattern recognition* 44, 2 (2011), 330–349.

[13] Liu Zhaohui, Fu Weiping, Yin Zihui, Liu Yunpeng, Wang Jiangwei, and Pei Shaotong. 2017. Insulator identification method based on infrared image. In *Smart Grid and Smart Cities (ICSGSC), 2017 IEEE International Conference on*. IEEE, 137–141.

[14] Zhang Zhixia, Liu Xiao, and Piao Zailin. 2014. Fault line detection in neutral point ineffectively grounding power system based on phase-locked loop. *IET Generation, Transmission & Distribution* 8, 2 (2014), 273–280.