Identification of Tea Red Leaf Spot and Tea Red Scab Based on Hybrid Feature Optimization

Shulin Meng¹, Shuguang Wang²*, Tao Zhou¹ and Jiankun Shen¹

¹School of Electronics and Information Engineering, Anhui University, Hefei, Anhui, 230601, China
²Hefei Institute for Public Safety Research, Tsinghua University, Hefei, Anhui, 230601, China

*Corresponding author’s e-mail: wangshuguang@gsaft.com

Abstract. Tea leaf diseases seriously affect the quality and the yield of tea. In order to determine whether the tea leaves are infected by diseases or any types of infection, technical support is essential for taking appropriate measures of disease control. Images of normal tea leaves, tea leaves infected with Tea Red Leaf Spot, and leaves infected with Tea Red Scab disease were studied. An identification algorithm for both of the tea leaf diseases based on hybrid feature optimization was proposed. First, the image features were extracted using the Histogram of Oriented Gradient and the Inception v3 model. Then, hybrid feature optimization processing was performed on two types of extracted features. Finally, the Gradient Boosting Decision Tree algorithm was used as the classifier for the identification of tea leaf diseases. Experiments demonstrate that the hybrid feature optimization algorithm reduces the image feature from 36,068 to less than 150 dimensions while maintaining a high identification accuracy, which greatly reduces the complexity of the identification algorithm. At the same time, the identification accuracy of tea leaf diseases based on hybrid feature optimization algorithm were higher than 95%.

1. Introduction

Tea is frequently affected by different diseases, especially Tea Red Leaf Spot disease and Tea Red Scab disease. They affect the quality and the yield of tea, resulting in direct economic losses to tea farmers. The accurate identification of tea diseases can facilitate prevention and treatment measures. This not only promotes the industrialization of agricultural industry, but also improves the level of the agricultural industry, with a practical significance and social value.

In order to solve the problems of manual identification of tea leaves, Md. Selim Hossain[1] extracted 11 features of tea pests and disease images, and used SVM classifier to identification three kinds of tea leaves images, which are tea brown leaf spot leaf, tea algal spot leaf and normal leaf leaves. On the one hand, the [1] does not consider the problem of numerical matching and information redundancy between different features; On the other hand, it uses a large number of features defined by human, the identification algorithm migrates to other data set with uncertainty. Few scholars and experts have paid attention to the identification of tea diseases. There are few methods or algorithms to identify tea diseases. Imaging that identification of tea leaves[2-6]and other plant diseases or pests identification[7-10], which is similar to tea leaf diseases identification task, that could be solved by the methods and theories of machine learning. Generally speaking, there are three stages for this kind of identification program. The first stage is data acquisition (open data sets are also available) and data
preprocessing. Two commonly used data include spectral data[8] and image data. Usually, the amount of data obtained is small, which is not suitable for direct training of large and deep neural networks. The second stage is feature extraction[11] and optimization, extraction features such as Histogram of Oriented Gradient (HOG)[11], Local Binary Pattern (LBP)[5], Kernel Principal Component Analysis (KPCA)[12] and color. Good features and appropriate feature optimization algorithm has an important impact on identification accuracy. Finally, Support Vector Machine (SVM) [4] or Back Propagation (BP) [13] algorithm is used to complete the classification task.

In this paper, for the identification of tea leaf’s diseases, a hybrid feature optimization method is proposed. Firstly, using hog algorithm to extract hog features and by Inception v3 model to extract features. Then, the extracted features are optimized by principal component analysis (PCA), standardization and tree-based feature selection (TBFS) method. Finally, the Gradient Boosting Decision Tree (GBDT) algorithm is used to identification of the tea leaf’s diseases.

2. Materials and methods

2.1 Implementation environment and data
The experiment was performed using a laptop with Inter(R) Core (TM) i5-4210M CPU @2.60GHz; the system was Ubuntu16.04, and the programming language was Python3.6.

A total of 100 tea leaf images, were used, including normal tea leaf images (27), images of leaves infected with Tea Red Leaf Spot disease (58), and images of leaves infected with Tea Red Scab disease (15). A total of 80 tea images were randomly selected as training samples, and the remaining 20 images were used as test samples. As shown in Figure 1, images from left to right are of normal leaves, tea leaves infected with Tea Red Leaf Spot disease, and tea leaves infected with Tea Red Scab disease.

![Figure 1. Different types of tea leaf images](image)

2.2 Identification method for tea leaf diseases
The tea leaf disease identification process based on hybrid feature optimization is shown in Figure 2. After the tea leaves image is preprocessed, the features are extracted, including the HOG feature and the feature extracted by the Inception v3 model. Subsequently, the extracted features are optimized with PCA and standardization, respectively. Then, the TBFS algorithm is applied to select good features and appropriate weights. Finally, the optimized features are sent to the GBDT.

![Figure 2. Tea leaf disease identification flow chart based on hybrid feature optimization](image)

Before extracting the HOG feature, image preprocessing is necessary, including size normalization and gray image transformation. Before the tea leaf image can be imported into the Inception v3 model, the purpose of image preprocessing is to normalize the image size by converting the input image to 299*299. The feature extracted by the Inception v3 model is 2,048 dimensions, and the feature of the
2.3 Extraction features with the Inception v3 model

Inception v3[14] is a 46-layer convolutional neural network model consisting of 11 mixed layers, also known as inception blocks. A framework for extracting image features using the Inception v3 model is shown in Figure 3.

![Figure 3. Framework for extracting image features using the Inception v3 model](image)

The parameters and structures between the mixed layers are different. In general, the mixed layer processes the input matrix using filters of different sizes. The mixed layer has two characteristics: one is that the network learns the filter’s parameter by itself to avoid the problem of manually selecting the size of the filter, and the other is that the convolution layer with small kernel size is used to replace the layer of the large size.

The Inception v3 model uses the cross entropy loss function, which is shown in formula (1).

$$H(q, p) = - \sum_{k=1}^{K} \log(p(k)) q(k)$$

$$H(q, p) = - \sum \log p(k) q(k) = (1 - \varepsilon)H(q, p) + \varepsilon H(u, p)$$

(1)

Here, and is a pair of loss functions used to replace the single loss function, is a distribution independent of the training sample, measures the similarity between the predicted distribution and the real distribution, is the weight coefficient.

There aren’t enough tea leaf disease image training a Inception v3 network, thus using Parameter-based Approaches of Transfer Learning to Extraction Image Features. Specifically, an Inception v3 model is trained with a data set similar to the tea leaf data set, and the trained model is used to extract the image features of tea leaves. The Inception v3 model defaults to input 3 channels of 299 * 299 RGB images. Therefore, the size of the input image is normalized to 299 * 299 in the image preprocessing section. After 11 mixed layers, the image features of tea leaves of 2,048 dimensions are obtained from the maximum pooling layer of the size 8 * 8.

2.4 Optimization of the tea leaves’ hybrid features

Data fusion methods include data layer, a feature layer, and a decision layer. Aiming at the leaf image of tea, the feature layer fusion method was adopted in this paper. The method is mainly divided into three categories: direct combination method (DCM), weighted multi-feature fusion method, and kernel function method. The TBFS is a method to calculate the importance of features and eliminate irrelevant features. In this paper, the TBFS algorithm is used as a weighted multi-feature fusion algorithm. In addition to the TBFS, the ReliefF algorithm is a common feature selection method.

The optimization of tea leaves’ multi-feature consists of three steps: PCA dimensionality reduction, standardization, and multi-feature fusion.

Input data mapping is performed from the high-dimensional space to the low-dimensional space by the PCA algorithm. While most of the feature information of the original data is preserved, the redundant data and a small amount of low-information data are left. On the one hand, the PCA can
effectively reduce data dimension, compress the data, and improve the speed of the algorithm. On the other hand, it can also remove the noise to a certain extent and enhance the noise immunity of the data. In this paper, the PCA algorithm was used, and the highest feature dimension was 100. Therefore, the HOG feature was reduced by the PCA algorithm from 34,020 to 100 dimensions, and the features extracted using the Inception v3 model were reduced from 2,048 dimensions to 100 dimensions.

3. Results

Data set 1 (D1) contained two categories of tagging data. Normal tea leaves are labeled "1", and diseased leaves are labeled "0". As training samples, 80 images were randomly selected, and the remaining 20 images were used as test samples. The D1 was used to determine whether tea leaves were infected. Data set 2 (D2) contained three categories of tagging data: leaves with Tea Red Leaf Spot were labeled "0", normal tea leaves were labeled "1", and leaves with Tea Red Scab disease were labeled "2". Again, 80 leaf images were randomly selected as training samples, and the remaining 20 as test samples. The D2 can be used to distinguish the disease category.

Figure 4. Weight calculation for features extracted from D1 and D2 using TBFS algorithms

Figure 4 represents a line graph of the feature index and the feature coefficient drawn by the TBFS after the input of the feature. If the weight coefficient is 0, it will be removed manually.

| Identification accuracy | D1 | D2 |
|-------------------------|----|----|
| [HOG]                   | 70%| 70%| 60%| 45%| 65%| 40% |
| [Inception v3]          | 95%| 85%| 85%| 80%| 75%| 70% |
| [DCM-A]                 | 95%| 70%| 85%| 85%| 65%| 70% |
| [DCM-B]                 | 90%| 95%| 80%| 90%| 90%| 70% |
| [ReliefF]               | 95%| 85%| 80%| 90%| 65%| 70% |
| [TBFS]                  | 100%| 100%| 85%| 95%| 90%| 80% |

* The [DCM-A] represents processing the two types of features using the DCM, which are image [HOG] features and [Inception v3] features, after extracting the image [HOG] feature and the [Inception v3] feature. [DCM-B], [ReliefF], and [TBFS] represent the feature sets of tea leaf images. The [DCM-B] uses PCA dimension reduction and feature standardization; two types of features are extracted, followed by DCM processing. ReliefF represents the ReliefF method for feature selection and computation of [DCM-B]. The difference between [TBFS] and [ReliefF] is that [TBFS] uses the TBFS method as feature set, obtained by the weighted multi-feature fusion algorithm.

Compared with [DCM-A], [DCM-B] uses the PCA algorithm to reduce the dimensions and to standardize the feature vectors. Moreover, [DCM-B] is 200-dimensional, far below the 36,068-dimensional [DCM-A]. Based on the experimental results, the [TBFS], whether on D1 or D2, achieved the best identification accuracy on any classifier, reaching 95%, which is higher than that of the DCM and the ReliefF algorithm. The [DCM-B] has the same identification accuracy as the [TBFS] under certain conditions, such as the GBDT classifier on D1, the DT classifier and the SVM classifier used
on D2. The [ReliefF] accuracy in different data sets or different classifiers was lower than that of the [DCM-B] and the [TBFS], except when the DT classifier was used on D1.

In conclusion, the use of the Inception v3 model to extract tea leaf image features has a high identification accuracy. In addition, the hybrid feature optimization methods greatly reduce the input feature vector dimension (about 0.4% of all feature dimensions), and the experiment shows that it is highly discriminating. Moreover, the hybrid feature optimization algorithm greatly reduces the amount of computation and computation time of the model and provides a strong support for the identification model to be deployed in mobile devices. Finally, the use of GBDT with excellent performance further improves the accuracy of the tea leaf disease image identification algorithm; the accuracy of the proposed tea leaf disease identification algorithm based on hybrid feature optimization was higher than 95%.

4. Conclusions
By comparing the identification accuracies of different features, different classifiers, and different feature layer fusion methods, we could show that the proposed algorithm has an excellent performance. In the future, this image dataset will be expanded, both in terms of image category and quantity. The identification algorithm can distinguish more types of tea leaf disease images and has a higher robustness.

Acknowledgments
This work was supported in part by the National Natural Science Foundation of China under Grant Nr. 61672032 and by a 2016 Doctoral Research Initiation Fund (J01003220).

References
[1] Hossain, M. S.; Mou, R. M.; Hasan, M. M. (2018) Recognition and detection of tea leaf’s diseases using support vector machine. In: 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA). Malaysia. pp. 150-154.
[2] Laddi, A.; Sharma, S.; Kumar, A. (2013) Classification of tea grains based upon image texture feature analysis under different illumination conditions. Journal of Food Engineering, 115(2): 226-231.
[3] Li, X.; Nie, P.; Qiu, Z. J.(2011) Using wavelet transform and multi-class least square support vector machine in multi-spectral imaging classification of Chinese famous tea. Expert Systems with Applications, 38(9):11149-11159.
[4] Kundu, P. K.; Kundu, M. (2016) Classification of tea samples using SVM as machine learning component of E-tongue. In: International Conference on Intelligent Control Power and Instrumentation (ICICPI-2016). Kolkata. pp. 56-60.
[5] Tang, Z.; Su, Y.; Meng, J. E. (2015) A local binary pattern based texture descriptors for classification of tea leaves. Neurocomputing, 168(C):1011-1023.
[6] Palaciosmorillo, A.; Alczar, N.; De, Pablos. F.(2013) Differentiation of tea varieties using UV-Vis spectra and pattern recognition techniques. Spectrochimica Acta Part A Molecular & Biomolecular Spectroscopy, 103(4):79-83.
[7] Arivazhagan, S.; Shebihah, R. N.; Ananthi, S.(2013) Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features. Agricultural Engineering International: The CIGR e-journal, 15(1): 211-217.
[8] Wu, X.; Zhang, W.; Qiu, Z. (2016) A novel method for detection of pieris rapae larvae on cabbage leaves using nir hyperspectral imaging. Applied Engineering in Agriculture, 32(4):311-316.
[9] Anter, A. M.; Hassenian, A. E.; Oliva, D. (2019) An improved fast fuzzy c-means using crow search optimization algorithm for crop identification in agricultural. Expert Systems with Applications, 118:340-354.
[10] Muthukannan, K., & Latha, P. (2018). A ga_fnn algorithm applied for classification in diseased plant leaf system. Multimedia Tools and Applications, 77: 24387-24403
[11] Chen, X.; Wang, S.; Zhang, B. (2018) Multi-feature fusion tree trunk detection and orchard mobile robot localization using camera/ultrasonic sensors. Computers and Electronics in Agriculture, 147:91-108.

[12] Datta, A.; Ghosh, S.; Ghosh, A. (2017) Unsupervised band extraction for hyperspectral images using clustering and kernel principal component analysis. International Journal of Remote Sensing, 38(3):850-873.

[13] Dimililer, K; Zarrouk, S. (2017)ICSPI: intelligent classification system of pest insects based on image processing and neural arbitration. Applied Engineering in Agriculture, 33(4): 453-460

[14] Szegedy, C.; Vanhoucke, V.; Ioffe, S. (2016). Rethinking the Inception Architecture for Computer Vision. Computer Vision and Pattern identification. Las Vegas, Nevada. pp. 2818-2826.