

## Sunflower Diseases Recognition Algorithm Based on Wavelet Domain Feature Dimension Reduction

Zhu Zhongyang, Xiao Zhiyun\*, Guo Yi

College of Electric Power, Inner Mongolia University of Technology, Hohhot, China

\*xiaozhiyun@imut.edu.cn

**Keywords:** Wavelet analysis; Feature Vector Dimension Reduction; Probabilistic neural network

**Abstract.** A new sunflower diseases recognition algorithm has been presented, which is based on image processing and pattern recognition. Firstly, sunflower diseases were collected for segmenting disease spot. The color histogram based on RGB space and color features based on HSI space of leaf for disease are extracted, based on gray level co-occurrence matrix texture features, these features are arranged for one dimensional vector. Dimension of characteristic vector is reduced by wavelet analysis, and the original vector is replaced by discrete approximation information of characteristic vector. Finally, the dimension reduction series and identify diseases were trained and automatically determined by probabilistic neural network. Experimental results show that the system not only can accurately identify these three diseases, sunflower powdery mildew, sunflower black rot and sunflower downy mildew, but also can make the feature vector have low characteristics dimension, in the mean time, ensure the recognition accuracy.

### Introduction

Crops often under the influence of diseases and pests in the growth process, not only caused huge economic losses but also affect the quality of the crops. Accurate and rapid identification of diseases and pests in agriculture has becoming a pressing demand. As the main commercial crop, sunflower has become a global second categories of oil-bearing crops. The plant diseases and insect pests recognition system of sunflower leaf is mainly studied based on image processing and pattern recognition in this paper, and the main recognized objects are these three diseases, sunflower powdery mildew, sunflower black rot and sunflower downy mildew.

Domestic and overseas experts had done a lot of work on identification of crops diseases and insect pests, and the research contents mainly include the disease spot of segmentation, feature extraction and pattern recognition<sup>[1]</sup>. The threshold method was first used on the segmentation of crop disease and disease spot such as mushroom sick leaves in 1989 by Xinxiong S. et al.<sup>[2]</sup> Due to the application of RGB histogram was not representative, the segmentation results were not accurate and ideal. On this basis, Chen Jiajuan<sup>[3]</sup> completed the segmentation of cotton diseases and insect pest disease spot through the local threshold method to extract RGB histogram shape. He Dandan<sup>[4]</sup> succeeded to segment the rice leaves through the combination of median filtering algorithm, Canny templates edge detection algorithm and minimum bounding rectangle algorithm. Diao ZhiHua<sup>[5]</sup> studied the plant diseases and insect pests of wheat leaf, and they obtained high accuracy by the application of Hu's invariant moments features. Zhuhong et al.<sup>[6]</sup> studied the wheat nitrogen deficiency, and they improved the accuracy through extracting the color features from RGB and HIS color space. Patil JK et al.<sup>[7]</sup> conducted a study on corn pests, by extracting the HSV histograms and texture extraction by GLCM information, and they also got a good result. For image recognition, the classical method contains neural network, fuzzy clustering and support vector product method. Zhang Jianhua et al.<sup>[8]</sup> studied the cotton diseases and insect pests classification, and they apply the radial support vector product to pest identification such as cotton spider mites, cotton aphid and so on. Their identification accuracy reached more than 88%. Guan Zexin et al.<sup>[9]</sup> had carried on the classification of plant diseases and insect pests images of rice by extracting features such as color, texture and shape, and at the same time the recognition is studied by Bayes classifier and the accuracy rate reached 97.2%. However, due to excessive selected features, heavy workload and low efficiency need to be

improved.

This paper mainly studies a new sunflower diseases recognition algorithm, which is based on image processing and pattern recognition. Firstly, sunflower diseases were collected for segmenting disease spot. The color histogram based on RGB space and color features based on HSI space of leaf for disease are extracted, based on gray level co-occurrence matrix texture features, these features are arranged for one dimensional vector. Dimension of characteristic vector is reduced by wavelet analysis, and the original vector is replaced by discrete approximation information of characteristic vector. In the end, the dimension reduction series and identify diseases were trained and automatically determined by probabilistic neural network.

## Lesion Segmentation

Transfer these samples from RGB space to Lab space, the non-green part is segmented by using one-dimensional Otsu method in the channel. And the 8-bit grayscale image in the channel is segmented by two-dimensional OTSU method<sup>[10]</sup>. After segmentation, the results as samples were used to set up the double Gaussian model and produce the Markov distance image. Then the Markov distance image is segmented by Otsu method, comparison results are shown in figure 1.

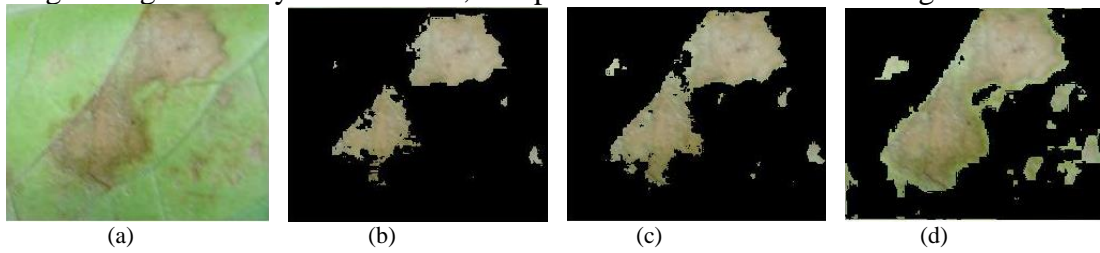


Fig.1 The contrast between the results

(a)The original image;(b)the segmentation results which using one-dimensional OTSU method; (c)The segmentation result which using OTSU threshold method of Curve; (d)The segmentation result which using double Gaussian model

## Feature Extraction

### Color features

This paper choose the color histogram based on the RGB space and color features based on the HSI space to extract color feature. And only choose the color histogram in the R channel and the B channel as characteristics. The three kinds of color histogram of the disease, sunflower powdery mildew, sunflower black rot and sunflower downy mildew, are shown in figure 2.

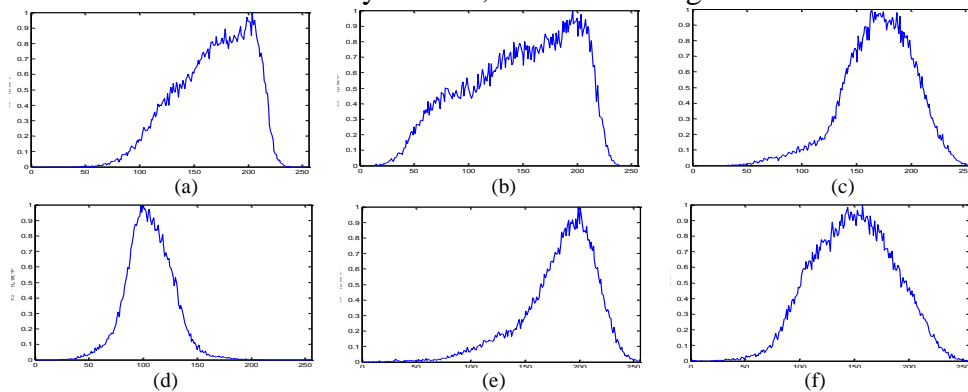


Fig.2 Three kinds of diseases of color histogram

(a)R channel color histogram of sunflower powdery mildew; (b)B channel color histogram of sunflower powdery mildew; (c)R channel color histogram of sunflower black rot; (d)B channel color histogram of sunflower black rot; (e)R channel color histogram of sunflower downy mildew; (f)B channel color histogram of sunflower downy mildew

In this paper, color moments of the HIS color space are also used to color feature as a supplement. We choose six color moments from the H channel and the S channel as color features.

The first order, second order and third order moment for the H channel and the S channel as follows:

$$M_1 = \frac{1}{N} \sum_{i=1}^N H(p_i) \quad M_2 = \left[ \frac{1}{N} \sum_{i=1}^N (H(p_i) - M_1)^2 \right]^{1/2} \quad M_3 = \left[ \frac{1}{N} \sum_{i=1}^N (H(p_i) - M_1)^3 \right]^{1/3} \quad M_4 = \frac{1}{N} \sum_{i=1}^N S(p_i) \quad M_5 = \left[ \frac{1}{N} \sum_{i=1}^N (S(p_i) - M_4)^2 \right]^{1/2} \quad M_6 = \left[ \frac{1}{N} \sum_{i=1}^N (S(p_i) - M_4)^3 \right]^{1/3}$$

## Texture Feature

Application of gray level co-occurrence matrix to describe the sunflower leaf disease texture features. The spatial relations  $\sigma$  such as horizontal, vertical, 45 degrees, which are often used in pattern recognition. When the spatial relations  $\sigma$  once identified, then the gray level co-occurrence matrix  $P_\sigma$  will be generated. In this article the 256 level of gray level compression to 16 levels, which greatly increase the computing speed. Then normalized processing and extract the texture feature of the gray level co-occurrence matrix, the texture feature of the sunflower leaves' disease can be gain.

The angular second moment calculation formula is: 
$$f_1 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P_\sigma^2(i, j) \quad (1)$$

The contrast second moment calculation formula is: 
$$f_2 = \sum_{i=0}^{L-1} n^2 \left[ \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P_\sigma^2(i, j) \right], n = |i - j| \quad (2)$$

The correlation second moment calculation formula is: 
$$f_3 = \frac{\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} ij P_\sigma^2(i, j) \mu_1 \mu_2}{\sigma_1^2 \sigma_2^2} \quad (3)$$

$$\mu_1 = \sum_{i=0}^{L-1} i \sum_{j=0}^{L-1} P_\sigma(i, j) \quad \mu_2 = \sum_{i=0}^{L-1} j \sum_{j=0}^{L-1} P_\sigma(i, j) \quad \sigma_1^2 = \sum_{i=0}^{L-1} (i - \mu_1)^2 \sum_{j=0}^{L-1} P_\sigma(i, j) \quad \sigma_2^2 = \sum_{i=0}^{L-1} (j - \mu_2)^2 \sum_{j=0}^{L-1} P_\sigma(i, j)$$

The entropy second moment calculation formula is: 
$$f_4 = - \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} i \sum_{j=0}^{L-1} P_\sigma^2(i, j) \lg P_\sigma^2(i, j) \quad (4)$$

This paper respectively choose 6 samples in the three diseases, to extract a total of 16 characters about angular second moment, contrast, correlation and entropy, respectively in horizontal, vertical, plus or minus 45 degrees.

In order to reduce the dimensions of the characteristic value and retain accuracy. In the texture feature, this paper take energy and correlation with  $\sigma$  for  $0^\circ$ ,  $45^\circ$ ,  $135^\circ$ , and the correlation of  $90^\circ$  degrees. A total of 7 eigenvalues. So this paper extracted the 525 characteristic value ( $256 + 256 + 6 + 7 = 525$ ), these characteristic value is enough to describe the color and texture of sunflower. However 525 eigenvalues for pattern recognition is too large, there is still a redundancy in these characteristics.

## Feature Dimension Reduction and Recognition

The number of eigenvalues and the effectiveness of the characteristic value determines the efficiency and accuracy of pattern recognition. The Mallat pyramidal decomposition algorithm in wavelet analysis has been adopted in this paper, and the original vector is replaced by discrete approximation information of characteristic vector. Using probabilistic neural network to train and recognition each resolution in wavelet analysis. Calculation accuracy of each resolution, when resolution drops and accuracy also drops, take the resolution of the upper story of eigenvalues.

### Feature Dimension Reduction

Multi-resolution analysis theory is an application of discrete wavelet transform (DWT). It is a layered framework theory to explanation signal or image information. Assumptions, the expression of discrete wavelet base is: 
$$h_{m,n}(t) = a_0^{-\frac{m}{2}} h(a_0^{-m} t - nb_0), (m, n \in \mathbb{Z}, a_0 > 1, b_0 \neq 0). \quad (5)$$

Discrete wavelet transform for any signal  $f(t)$  can be used under type said: 
$$DW_{m,n} = \int_{-\infty}^{+\infty} f(t) h_{m,n}(t) dt$$

Mallat pyramidal decomposition algorithm is applied in this article to dimension reduction of 512 features, this article firstly combine color histogram of the R channel with the B channel as a one dimensional function.

$A_S f(\phi) = \psi_s(\phi)$ ,  $\phi \in [1, 512]$ ,  $f(n) = \psi_s(\phi_n)$ ,  $n \in [1, 512]$ ,  $n \in \mathbb{Z}$ ,  $\psi_s(\phi_n)$  is a discrete function. The  $f(n)$  can be regarded as the discrete approximation of  $f(\phi)$  under  $2^j$  resolution.  $A_{2^j}^d f(n) = f(n) \in V_{2^j}$ ,  $j \in \mathbb{Z}$ ,  $A_{2^j}^d$  is  $f(\phi)$  in discrete approximation operators under  $2^j$  resolution,  $V_{2^j}$  is the vector space of  $A_{2^j}^d$ .  $f(n)$  can be decomposed into approximation information and detail information under  $2^{j-1}$  resolution by Mallat algorithm.

Algorithm is as follows:  $f(n) = A_{2^j}^d f(n) = A_{2^{j-1}}^d f(n) + D_{2^{j-1}}^d f(n) \quad (6)$ ,  $A_{2^{j-1}}^d f = HA_{2^j}^d f \quad (7)$ ,  $D_{2^{j-1}}^d f = GA_{2^j}^d f \quad (8)$

Formula (7) and (8) respectively represent the discrete approximation and the discrete detail under  $2^{j-1}$  resolution.  $H$  and  $G$  respectively represent the low-pass filter and its mirror image high-pass filter. Through the above algorithms, the approximate information and detail information can be obtained by filtering  $H$  and  $G$ .  $f^{(n)} = \psi_s(\varphi_n)$ ,  $n \in [1, 512]$ ,  $n \in Z$ , can be regarded as  $f^{(\varphi)}$  the discrete approximation of  $A_1^d f$  under 1 resolution. Figure 3 is the Mallat decomposition process under the  $\downarrow 2$  representative sampling. Figure 4-1 shows that the approximate information and details under the  $2^{j-1}$  resolution can be respectively obtain by the approximate information under 3 resolution through low-pass filter and high-pass filter and then after sampling.

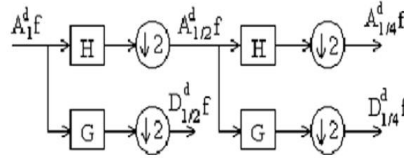


Fig.3 Mallat pyramidal decomposition algorithm process

The application of Mallat algorithm to do four decomposition in  $f^{(n)} = \psi_s(\varphi_n)$ ,  $n \in [1, 512]$ ,  $n \in Z$ , which can get the discrete approximation information with the resolution for  $1/2, 1/4, 1/8, 1/16$ . In this paper, the low pass filter ( $H$ ) we choose was the firming orthogonal filter with length of 14. In order to make the algorithm more anti-noise performance, this article will first do four decomposition on characteristic curve of  $\psi_s(\varphi_n)$ , and then adjust  $D_{1/2}^d f$ ,  $D_{1/4}^d f$ ,  $D_{1/8}^d f$ ,  $D_{1/16}^d f$  details to zero and reconstruction to get the denoising characteristic curve of  $\psi_s$ .

Due to a total of 512 characteristic value with high dimension. In this paper, using multi-resolution analysis improved the combination curve characteristics of the RB channel. And to reduce the number of eigenvalues with the application of low resolution to improved characteristic curve of  $\psi_s$ . In this section, apply Mallat pyramidal decomposition algorithm to decompose the combination curve characteristics of the RB channel for different resolution curve as shown in figure 4.

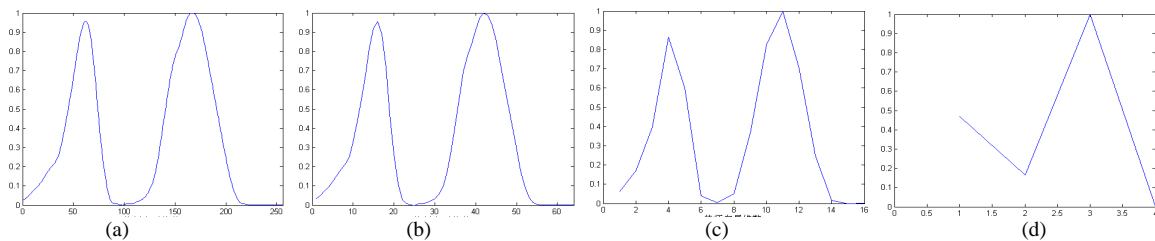


Fig.4 Mallat Sunflower downy mildew RB combination characteristic when resolution is 1/2 to 1/128

(a),(b),(c),(d)Sunflower downy mildew RB combination characteristic when resolution respectively is 1/2,1/8,1/32,1/128

Figure 4 shows that the resolution falling at the same time with the resolution of the waveform. And with the number of eigenvalues is declining, the speed of pattern recognition improved. In this paper, the probabilistic neural network automatically used for determine the resolution and identification.

### Resolution of the Automatic Selection and Sunflower Disease Spot Identification

In order to estimate the probability density in the Bayes rule, samples  $P_i$  can be applied to training network. The basic principle of probabilistic neural network is the summation layer linked together with the model units. And the Parzan window function method was used to know the sample probability, the conditional probability is:

$$p(X|f_k) = \frac{1}{(2\pi)^{m/2} \sigma^m n} \sum_{P_i \in f_k} \exp\left[-\frac{X^T W - 1}{\sigma^2}\right] \quad (9)$$

In type (9) :  $m$  is the vector dimension,  $x$  is the vector of unknown sample,  $X^T$  is matrix transpose,  $W$  is the training sample vector,  $f_k$  is the classification model,  $\sigma$  is the smoothing factor,  $n$  is the number of classification model.

The basic structure of probabilistic neural network as shown in figure 5.

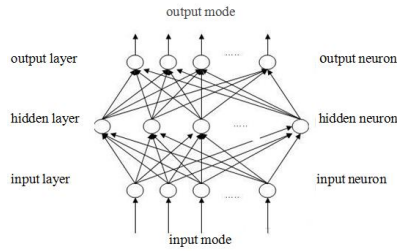


Fig.5 Probabilistic neural network structure

Selecting RB combination curve characteristics, color moment characteristics based on the HIS space and texture features based on gray level co-occurrence matrix in probabilistic neural network input layer. Identify three kinds of plant diseases and insect pests, each type of diseases has 30 pictures and 20 pictures for test.

This paper, specific steps of choice the optimal resolution are as follows.

Step1. The application of RB combination curve characteristics with a resolution of 1 in training and 60 images for the accuracy test.

Step2. Using RB combination curve characteristics with a resolution of 1/2 in training and testing accuracy with 60 images, compared with accuracy of the last step, continue to reduce resolution if correct as same as the previous step, or a slightly smaller.

Step3. Repeat step 2 until the current of the time difference is larger, choose the previous step as the optimal resolution.

Tab.1 recognition correct rate

resolution	Characteri- stic value dimension	powdery mildew recognition accuracy	black rot's recognition accuracy	downy mildew recognition accuracy
1	512	80%	75%	80%
1/2	256	80%	75%	80%
1/4	128	80%	75%	80%
1/8	64	80%	75%	80%
1/16	32	80%	75%	80%
1/32	16	80%	75%	80%
1/64	8	80%	75%	80%
1/128	4	70%	70%	75%

Table 1 shows that the recognition accuracy of sunflower powdery mildew and sunflower downy mildew are 80% when resolution change from 1 to 1/64, the sunflower black rot's recognition accuracy is 75%. And the identification accuracy reduce with the lower resolution. When the resolution is higher, with the high dimension characteristic value, characteristic information also rich and redundancy, so a very high accuracy can be achieved. Through training each layer of the resolution and do the test identification with the combination of RB curve characteristics, it can be seen the description of the sunflower leaves department disease spot information is more and more incomplete with the characteristic value dimension is reduced at the same time. This leads to with the increase of decomposition, and accuracy has fallen sharply. According to table 1 data, this paper choose the RB combination curve characteristics with the resolution is 1/64, when the number of eigenvalues is eight. So in this paper, application of wavelet analysis to the sunflower leaves department disease spot under the RGB space, the color histogram features of RB space go down from 512 to 8, and does not affect the recognition accuracy. At this point, and based on six colors torque characteristics of the HSI space and seven characteristics choosing from energy, inertia, entropy and contrast based on gray level co-occurrence matrix, in this paper, a total of 21 features (8+6+7=21) describe the sunflower leaves department disease spot.

## Conclusion

This article mainly studied the sunflower leaves department disease spot recognition algorithm based on wavelet domain feature dimension reduction, mainly includes:

Feature extraction methods: adopt color and texture features. In the color feature extraction, the application of color histogram based on the RGB space and color moment based on the HSI space. In the extraction of texture feature, using features such as energy, inertia, entropy and contrast based on gray level co-occurrence matrix. Through experimental study finally eliminated some useless features.

Feature dimension reduction and recognition features: the application of Mallat pyramidal decomposition algorithm in wavelet analysis to reduce the dimensions of the eigenvalues, and using discrete approximation details replace the original features. Training and identification each layer resolution characteristic value in the application of probabilistic neural network. By identifying the accuracy to determine the resolution, and then determine the number of eigenvalues. While decreasing the number of eigenvalues still not affect the identification accuracy, and solve the contradictions between characteristic value quantity and the identification accuracy.

In addition this paper only studies the sunflower leaf spot disease, sunflower diseases can be produced in a variety of locations, such as fruit and stem, etc. So the follow-up algorithms should also consider other parts of the disease spot of segmentation and recognition algorithm.

## Acknowledgment

This research has been supported by Natural Science Foundation of Inner Mongolia (Grant No. 2015MS0617), and Inner Mongolia Talent Development Fund. We would like to thank to all two reviewers for useful suggestions.

## References

- [1] Wang Jingjing,Zhang Wu Liu Lianzhong, et al. Crop diseases and pests of image recognition technology research review . Computer engineering and science, 2014, 4 (7) : 1363-1370
- [2] Xinxiong S. In accordance with the image extracting plant growth information. The 6th Branch Research Data of the Agricultural Society of Mechanical Engineers, 1989, 1(12): 1-9
- [3] Chen Jiajuan, Zhao Xuedu. Using computer vision for automatic determination of the degree of cotton insect pests. Journal of agricultural engineering, 2001, (2) : 157-160
- [4] He Dandan,Xu Mei. Fusion technology based on rice leaf edge detection algorithm research. Journal of northeast agricultural university,the lancet,2012 (5) : 65-68
- [5] Diao Zhihua,Zheng An-ping,Wu Yuanyuan. Shape feature extraction of wheat leaf disease based on invariant moment theory. IFIP Advances in Information and Communication Technology, 2012, (369):168-173
- [6] Zhu Hong,Zheng Limin,Yin Jianling etc. Based on wheat canopy image color feature extraction in nitrogen nutrition level evaluation. Journal of agricultural engineering, 2010, 26 (2) : 16-20
- [7] Patil J K,Raj K. Feature extraction of diseased leaf images. Journal of Signal and Image Processing, 2012, 3(1): 60-63
- [8] Zhang Jianhua, Ji Ronghua, Yuan Xue, etc. Cotton insect pests recognition based on radial basis of support vector machine . Journal of agricultural machinery, 2011, (8) : 178-183
- [9] Guan Zexin,Tang Jian,Yang Baojun,et al. Rice disease recognition method based on image study . Chinese rice science, 2010, 24 (5) : 497-502
- [10] Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on System Man and Cybernetic, 1979, 9(1): 62-6