ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

Early Detection of Plant Virus Infection Using Multispectral Imaging and Dual Spatial Information Fusion

Jianzhi Fan

School of, The University of Manchester, Manchester M13 9PL, United Kingdom

Abstract

Cassava brown steak diseases (CBSD) has caused serious reduction to the cassava productivity in Africa and is hard to detect it at the early stage of infection. Exist approaches requires large number of samples or complex experiment process, which cannot be implement in practice. This dissertation proposed a model that can detect infected plants at an early stage based on dual information fusion. Two main steps applied in this model is feature extraction and spatial optimization. For feature extraction, kernel principal component analysis (KPCA) is used to extract crucial information and transform original data to linear separable data, the data is then classified by support vector machines (SVM) to obtain a probability map. For spatial optimization, extended random walker is applied to generate another probability map. Then, the probability map created through feature extraction and spatial optimization are combined following a decision fusion algorithm. The model was applied on three groups of plants from three different trials, the result shows that this model has the potential to classify infected plants, but more trials are required to evaluate the reliability of this model in practice environment.

Keywords

Hyperspectral Classification; Plant Diseases; Feature Extraction; Decision Fusion.

1. Introduction

1.1 Background and Motivation

With the growth of the world's population, the demand for agricultural products has been increasing. Amid the COVID-19 pandemic, government quarantine policies have affected the production of agricultural products, and the spread of plant viruses has aggravated the food crisis. In the past, the main method of crop pest control was to spread pesticides evenly over different tillage process. However, the virus was initially existed in limited areas, and this approach resulted in unnecessary increase of cost and pesticide residues. PCR is a common method for plant virus control, but it requires equipment and time, and it often cannot reliably detect viruses in the early stages of infection. Therefore, a low-cost method is needed to detect virus-infected plants at an early stage.

With the development of imaging sensors, hyperspectral images can be obtained to extract the characteristics of objects in different spectra. These characteristics are useful for object classification. Some common classifier includes multinomial logistic regression (MLR), support vector machine (SVM) and random forest. Fusion of dual spatial information for hyperspectral image was successfully implemented for Indian Pine dataset. The reflectance of plant leaves under different spectra can be used to calculate vegetation indices and predict plant stress. Therefore, it is feasible to use computer to identify the hyperspectral images of leaves and detect the virus in the early stage of infection.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

1.2 Aims and Objectives

This dissertation aims to implement early detection of plant virus infection using fusion of dual spatial information for hyperspectral image. Detailed objectives are shown below:

1) Selection of appropriate dataset: Choose a set of multi-spectral image of plant for the experiment.

2) Dataset sampling: Do sampling to the dataset by crop the image into several patches with same size and random location. By sampling, the scale of data can be increased without further experiment.

3) Patch based vegetation indices calculation: Calculate the average reflectance for each patch and use reflectance in different spectral to obtain vegetation indices.

4) Spatial feature extraction: Find the correspondence of neighbor pixels and extract the spatial features to form a structural profile (SP). SVM should then be applied to obtain the class probability.

5) Spatial probability optimization: First obtain the initial class probability by directly apply SVM to original image. Extended random walker (ERW) should then be adopted to this initial probability to optimize it.

6) Decision fusion: Merge the class probabilities to obtain a final label.

7) Evaluation: Obtain the classification accuracy by calculating the ratio of correct classified leaves and total number of samples.

2. Literature Review

Cassava Brown Steak Disease (CBSD) is widespread in Africa, causing serious reduction in local food production. This disease is caused by cassava brown steak virus (CBSV), which is mainly transmitted by whiteflies [1][2]. Molecular Techniques are a common method for the detection of plant viruses and can obtain accurate results. According to Lopez, molecular techniques can detect bacteria from 10 to 10^6 colony forming units/mL [3]. The limitations of the molecular techniques are that they cannot detect cassava brown steak disease in an early stage, for they have strict requirement to sample quality and are time-consuming [4]. Therefore, spectroscopic and imaging techniques were raised to provide accurate and timely result to avoid the spreading of CBSD.

Spectroscopic and imaging techniques have potential in plant disease detection. In 1965, Gates et al. did research on the reflectance of various plants in different spectra [5]. Later, S. Jacquemoud proposed a model for understanding high spectral resolution data. Anatoly A. Guelson obtained the chlorophyll content of plants through reflectance. Thus, the physiological status of plants can be indirectly predicted by analysing its hyperspectral data [6] [7]. Classification based on hyperspectral data has better performance compared to traditional RGB colour model image, for it contains extra spectral information that cannot be directly obtained by human vision. According to Prasad, not all the spectrums have close relationship with the biophysical characteristics of crops. Based on their experimental results, 12 specific narrow bands ranging from 350nm to 1050nm were recommended to provide optimal crop information [8]. It is also suggested that the sharp change in reflectance between 680 nm and 750nm is especially suitable for early plant stress detection [9]. The ratio of reflectance under different spectra can also reflect the stress of plants. Carter calculated the correlation between the ratio and plant stress, and indicated that compared with a single spectrum, the ratio of reflectance under 695nm and 420nm or 760nm have better performance in indicating the change of plant stress. These ratios were named vegetation indices and were later successfully applied in the detection of pests and diseases in agriculture [10]. However, the symptom of CBSD cannot be detected directly from the image of different spectral, and it is hard to combine the information of all spectrums. For some machine learning algorithms are suitable in processing multi-dimensional data, they were applied in the classification of images with multiple spectra.

With the development of computer vision and deep learning, Hyperspectral images (HSIs) were used in classify related tasks in geology and agriculture [11]. Support vector machines (SVMs) is a widely used classifier, its performance was evaluated by Melgani, the result shows that binary SVMs is effective in the classification of hyperspectral dataset compared to pattern recognition approaches

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

[12]. By combining multinomial logistic regression (MLP) and subspace projection, images with noise and mixed pixels can be classified with high accuracy [13]. Random forest (RF) is an approach based on decision trees, which can provide results within acceptable processing time. The problem with these methods is that they are sensitive to the total number of samples. These methods often do not perform well when the number of samples available for training is limited. Due to the large number of samples required, this kind of methods are also computationallyexpensive, which is a time-consuming work for poor performance processors. Another problem is "Hughes" phenomenon, this refers to the fact that when the dimensionality of the dataset is high, the accuracy may decrease [14].

To solve the problems that are elaborated above, other classification methods were proposed. The first method is morphological profile (MP), it applies a set of morphological operations such as opening and closing to the original image with different size of structuring elements [15]. The resulted profile contains a large set of features, but these features may be mixed with redundant information. Therefore, an extended morphological profile (EMP) was applied to distinguish important features from redundancy in [16]. The extended method adds feature extraction as an additional step compared with the original one. Applying both method to an urban hyperspectral dataset, the result shows that with principal component analysis (PCA) as the approach for feature extraction, the classification accuracy of EMP is higher than MP. Kernel principal component analysis (KPCA) is one kind of nonlinear PCA, it has an advantage in processing data with multiple dimensions [14], therefore KPCA is used in this paper for feature extraction. The detail is discussed in the methods section. In addition to feature extraction, image segmentation is also a typical method for solving the problem aforementioned. In [17], the extended random walker (ERW) algorithm is used for the classification of hyperspectral images. As this algorithm should be applied based on an existed probability map, another classification should be implemented before to generate an initial probability for the belonged class of each pixel. In this paper, SVM is used to complete the preliminary classification. By applying ERW to the result of SVM, the spectral information between different spectrums, the spatial relationship of neighbouring pixels and the difference between samples can be combined. With this addition information, the result can be optimized compared to the experiment with only SVM applied, therefore an acceptable accuracy can be obtained with relatively small number of training samples [17]. Detail of ERW is given in the methods section.

For feature extraction with support vector machine is effective in the recognition of large-scale object, and extended random walker is suitable for the modelling of small-scale object, merging the results of these two algorithms can produce better classification performance [18]. In the study of this dissertation, the datasets used for training and testing were obtained by an active multispectral imaging (A-MSI) sensor system [19]. Two probability maps should be calculated by different approaches in this dissertation. For the first approach, feature extraction based on kernel principal component analysis is applied to the obtained hyperspectral image as a preliminary processing. In the classification of plant diseases, patch-based voting method shows a better performance compared to whole leaf [19]. In order to implement patch-based voting, the resulted hyperspectral image of all leaves should be cropped into a set of patches that does not contain the leaf vein. These patches are labelled and fed to SVM for classification to generate the first probability map. For the second approach, original hyperspectral image should be classified by SVM directly and generate an initial probability map. Then, this map should be optimized with ERW to obtain the refined class probability. With class probabilities calculated by two approaches, a weighted decision fusion rule is applied to merge these two results to obtain the final result.

3. Methods

The general flowchart of the process implemented in this dissertation is shown in Fig. 1. In the beginning, the initial hyperspectral image that was obtained by A-MSI system is processed by two approaches separately. For first approach, key step is feature extraction implemented based on kernel principal component analysis. For second approach, key step is probability map optimization based on extended random walker. Detail of each step is discussed below.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

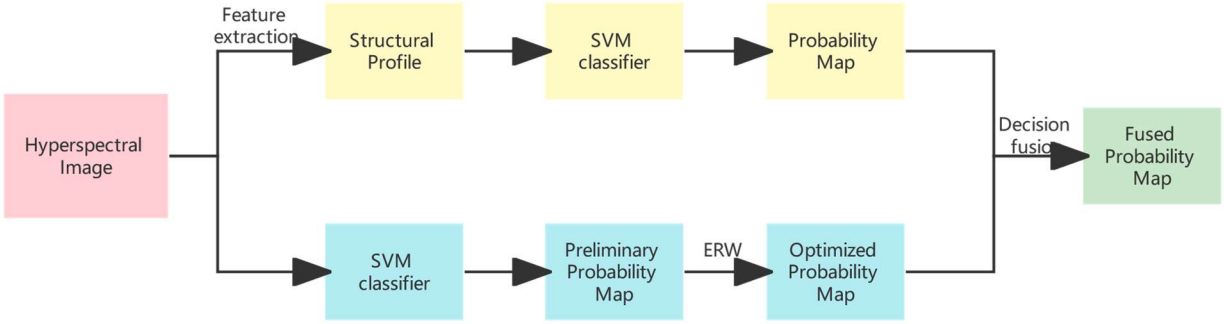

Figure 1. Flowchart of the proposed classification framework

3.1 Data Set

The dataset used in this experiment was generated by A-MSI, the dataset contains three trials. The leaves in each trial were processed by different operations to generate three treatment groups. The first group was inoculated by Uganda cassava brown steak virus (UCBSV) and is regarded as the infected group. The second group was not inoculated and raised in controlled environment to ensure they are not infected. However, in first group UCBSV may not be the only factor that affect the plant stress, the operation of inject can also have influence on the plant. To ensure the change of reflectance is caused by the virus, not the injection, it is necessary to set the third group. In the third group, each leaf was inoculated with an empty control E. coli plasmid that does not have substantial effect on the leaves, so the inject operation is the only parameter that affect the plant stress. These three groups of plants are then observed at specific day post inoculation (dpi). For each trial, the observation day is different. A-MSI system was applied in the observation to obtain reflectance of 15 wavebands, it used isotropic illumination and a combination of an integrating hemisphere to achieve a minimized specular reflectance [20]. The observation time for each trial is shown in table. 1, and the wavebands observed in this experiment is presented in table .2.

Table 1. Three Scheme comparing

Trial	Groups	No. of plant	Observation dpi
1	Infected	12	7, 28, 53, 88
	Mock	12	
	Uninfected	24	
2	Infected	18	14, 28, 54
	Mock	18	
	Uninfected	18	
3	Infected	18	7, 14, 28, 52
	Mock	18	
	Uninfected	18	

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

Table 2. Three Scheme comparing

Band no.	Centre band of wavelength
8	395 nm
9	415 nm
10	470 nm
11	528 nm
12	532 nm
13	550 nm
14	570 nm
0	585 nm
1	590 nm
2	610 nm
3	625 nm
4	640 nm
5	660 nm
6	700 nm
7	880 nm

3.2 Patch Cropping

Before the image processing of two branches, the original images of leaves are cropped into small patches. This step has two purposes, first is to enlarge the samples in dataset, so a larger number of samples can be used in the training of SVM to improve the performance of classifier. Second purpose is to specify the region of interest, as for each leaf, the reflectance of main leaf veins should avoid being taken into consideration since they do not tend to change significantly with the groups of plant. The location of the cropped patches is randomly selected. Depends on the size of the leaf image, the patch size in pixels varies from 16×16 to 48×48. An example 40×40 patch and its detail are shown in fig. 2. For each leaf, the number of cropped patches should be odd to ensure that there will not be equal vote for both groups in the voting step. In this experiment, 7 patches were cropped from each leaf.

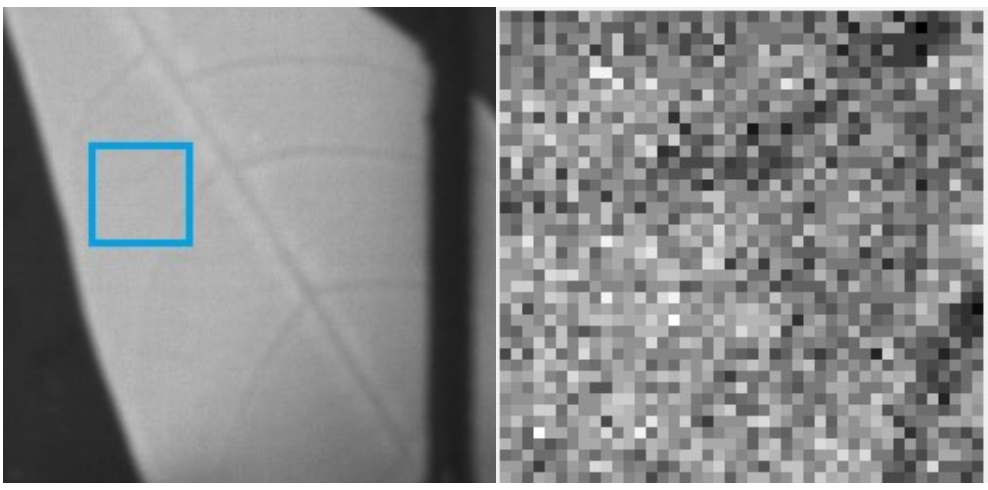

Figure 2. Flowchart of the proposed classification framework

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

3.3 Spatial Feature Extraction

Spatial feature extraction by KPCA is applied in the first branch of data processing. PCA can be used in linear dimensionality reduction, and kernel PCA can implement non-linear dimensionality reduction to dataset. Therefore, KPCA is applied to the linearly inseparable data in this experiment. The basic of KPCA is, for a matrix X in the input space, a nonlinear mapping is used to map all samples in X to a higher dimensional space called feature space. In feature space, these samples can be linearly separable, and dimensionality reduction by PCA can be applied in this space. Consider a two-dimensional data, a mapping operation ϕ can be used to map it to three-dimensional space as shown below:

$$\phi([x_1, x_2]^T) = \left[x_1^2, x_2^2, \sqrt{2}x_1x_2\right]^T \tag{1}$$

Where x_1, x_2 are the coordinates of data in two-dimensional space, $x_1^2, x_2^2, \sqrt{2}x_1x_2$ are corresponded three-dimensional coordinates after mapping. This mapping procedure can be presented as:

$$\phi(x): R^K \rightarrow R^D, D > K \tag{2}$$

Where K is the dimension of original space, D is the dimension of space after mapping. Assume original space is X, feature space after mapping is F:

$$X = [x_1, x_2, \dots, x_N] \tag{3}$$

$$F = \phi(X) \tag{4}$$

Where N is the number of samples in X. Dimensionality reduction can then be applied to the feature space F.

To apply kernel principal component analysis, four parameters need to be determined in advance. First parameter is the number of training samples N_s, it influences the required training time and accuracy. This number should be smaller than the quantity of total pixels in the input image, which in this case is within a range of 576 – 1764. Consider the memory of processor used in the experiment, 100 samples are selected for training. Second parameter is the number of dimensions required to be extracted, in this experiment is equal to the number of bands of hyperspectral image. Third parameter is related to the form of applied kernel function, in this experiment, Gaussian kernel function is used for it is suitable in transforming linearly inseparable data to linearly separable data. In the training step of KPCA, the scale parameter σ can be calculated by:

$$\sigma = \frac{\sum_{i=1}^{N_s^2} \sqrt{d_i}}{N_s^2} \tag{5}$$

Where d is the squared distances between training samples. Then, the kernel k for training data can be calculated by:

$$k_i = \exp(-\frac{d_i}{2\sigma^2}) \tag{6}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

The number of kernels equals to the square of training samples. Kernel matrix is centred in order to calculate the eigenvector. Then, the complete image is also kernelized and centred, by multiplying the resulted matrix with the eigenvector calculated by training samples, the kernel principal components can be obtained. These principal components are the first probability map of the image.

3.4 Probability Optimization with ERW

Then, probability optimization with ERW is applied to generate the second probability map. Before the optimization, SVM classifier is applied directly on the initial image to obtain a preliminary map. Then, ERW is applied to the map in the second branch of the experiment to implement optimization. It first transforms the original image into a weighted graph $G = (V, E)$. Where V is a set of pixels and E is a set of links that connect adjacent pixels. Then, a 8-connected lattice is build, each edge of the lattice has a weight denoted by ω_{ij}. The weight can be calculated by:

$$\omega_{ij} = \exp\left(-\beta\left(g_i - g_j\right)^2\right) \tag{7}$$

Where g_i and g_j are the intensity of pixel at point i and j. β is a free parameter, which in this experiment is determined as 710. The calculated weights show the intensity difference between pixel i and j, it should be normalized to obtain better performance. Solving random walker probabilities is same as solving combinatorial Dirichlet problem. The Dirichlet integral can be defined as:

$$D[u] = \frac{1}{2}\int_{\Omega} |\nabla \mathrm{u}|^2 d\Omega \tag{8}$$

Where u is the field and Ω is the region of this problem, and the harmonic function which satisfies $\nabla^2 u = 0$ should be found should be found to minimize the Dirichlet integral. As Laplace equation is the Euler-Lagrange equation for the Dirichlet integral, a combinational Laplacian matrix is defined as:

$$L_{ij}\begin{cases}\sum\omega_{ij} & if\ i = j \\ -\omega_{ij} & if\ i\ and\ j\ are\ adjacient\ pixels \\ 0 & otherwise\end{cases} \tag{9}$$

Then, the optimized probability can be obtained by solving the energy function shown below:

$$E^n(p_n) = E^n_{spatial}(p_n) + \gamma E^n_{aspatial}(p_n) \tag{10}$$

Where γ is a free parameter defined as 0.1^5, and $E^n_{spatial}(p_n)$ is an energy function of spatial terms represented by:

$$E^n_{spatial}(p_n) = p_n^T L p_n \tag{11}$$

Where n is the label of class, by minimizing this function, the probability of random walker starting from pixel i to labeled pixel n can be calculated. Similarly, $E^n_{aspatial}(p_n)$ is another energy function for aspatial term, represented by:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

$$E^n_{aspatial}(p_n) = \sum_{q=1,q\neq n}^{N} p_q^T \Lambda_q p_q + (p_n - 1)^T \Lambda_n (p_n - 1) \quad (12)$$

Where Λ_t is a diagonal matrix. The optimized probability can be obtained by choosing the maximum value of p_n, and generate the second probability map.

3.5 Decision Fusion

After the probability map of two branches are obtained, these two results can be merged through weighted decision fusion to improve the accuracy. Assume P_1^i is the probability that a pixel belongs to class i generated by the first branch, P_2^i is the probability generated by the second branch. A free parameter μ is multiplied as the weight. The final classification result P can then be calculated by:

$$P = \underset{i}{argmax}\{\mu P_1^i + (1 - \mu P_2^i)\} \quad (13)$$

3.6 Patch-based Voting

Patch based voting is to first apply classification to the patches cropped from each leaf, then determine the group of leaves based on the classified groups of their patches. According to the experiment presented by [19], classification with patch-based voting can provide higher accuracy compared with classification based on whole leaf in most cases.

The patches of a leaf may be classified to different groups, in this experiment the group that majority patches belong to is regarded as the group for this leaf. As explained in section 2.2, 7 patches were cropped for each leaf, which means if the number of patches belong to a group is larger or equal to 4, their corresponded leaf can be determined belonging to this group.

4. Results and Discussion

4.1 Infected Versus Uninfected

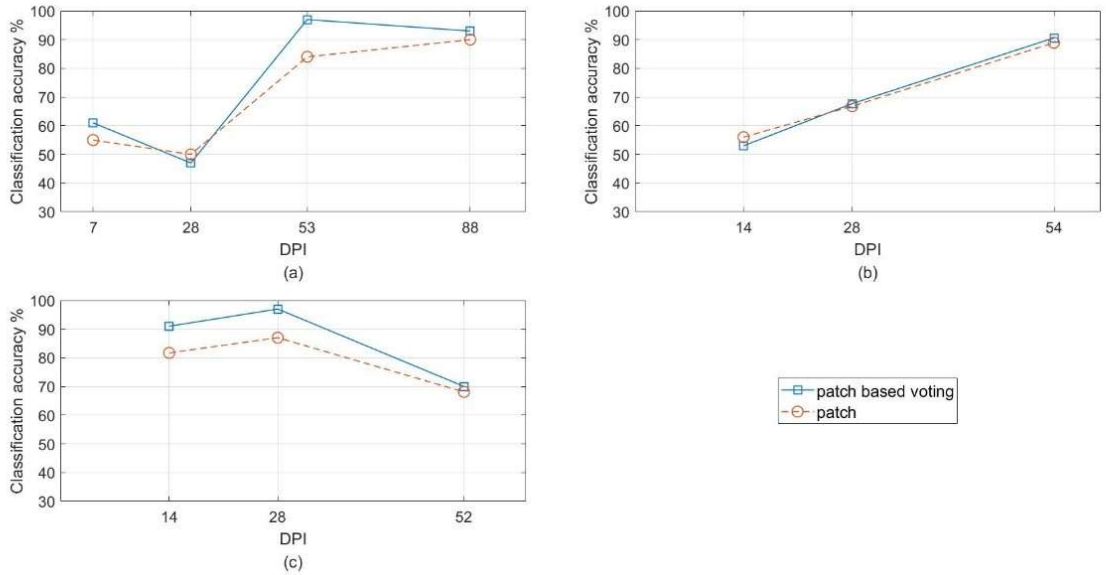

Figure 3. The classification accuracy (%) of infected versus uninfected leaf for three trials. (a) Classification accuracy of trial 1, sampled at 7,28, 53, 88 dpi. (b) Classification accuracy of trial 2, sampled at 14, 28, 54 dpi. (c) Classification accuracy of trial 3, sampled at 14, 28, 52 dpi.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

In the experiment, the model was first applied in the classification of infected and uninfected plants. Each plant was cropped into 7 patches and labelled manually, if one leaf belongs to the infected group, all patches cropped from it are labelled as "infected". Then, these patches are separated to test set and training set. For each trial, half of the data were used for training the model and the rest were used for testing. All patches in the testing set were labelled again by the classifier, by comparing the label given by classifier to the label given manually, the patch classification accuracy can be calculated by dividing the number of successfully matched labels to the overall number of patches. Patch-based voting was also applied to obtain the label for each leaf, and the patch-based voting classification accuracy was calculated similarly. The corresponded accuracy for each trial and dpi is shown in fig.3.

The classification accuracy for trial 1 is shown in fig. 3 (a). At 7 dpi, patch-based voting classification accuracy is 61%. Then, the accuracy dropped to below 50% at 28 dpi and rise to 97% at 53 dpi. From 53 dpi to 88 dpi, the accuracy has a small decrease from 97% to 93%.

In theory, the classification accuracy should rise with the increase of dpi, but there are two notable point in fig. 3 (a), which are the accuracy at 28 dpi and 88 dpi. At this two dpi, the classification accuracy was decreased compared to their previous dpi, and there are two possible explanations for this result. First explanation is that the accuracy was affected by the location of cropped patches, though for patch-based voting classification, the accuracy seems dropped significantly at 28 dpi and 88 dpi, the accuracy only dropped 5% for patch classification. In this scale, the selection of patches may cause some error, as at 28 dpi, it is possible that not all patches show symptoms of infected, for the sign of infection can only appear in particular area of the plants at early stage. As the testing set of trial 1 contains 210 patches, the number of mis detected patches was around 10, which can be regarded as a reasonable error caused by the selection of patches. For 88 dpi, the classification accuracy was in fact increased compared to previous dpi, but the accuracy was reduced by the voting step. Second explanation is that the symptom of infection may reduce slightly with the increase of dpi. As it is presented in [19], the symptom caused by CBSV at 73 dpi is more obvious than 81 dpi. This remission of diseases may be caused by the weather condition, though it is temporary, and the symptom worsening is inevitable on a long-term basis, it may reduce the classification accuracy at particular dpi. For trial 2 shown in fig. 3 (b), the classification accuracy of patch-based voting for different dpi increases smoothly from 53% at 14 dpi to 68% at 28 dpi, and at 54 dpi it has the highest accuracy of 91%. For trial 3 shown in fig. 3 (c), the classification accuracy of patch-based voting was 91% at 14 dpi, then it increased slightly to 97% at 28 dpi. However, the accuracy dropped significantly to 70% at 52 dpi. The possible explanation of this problem is similar to the situation in trial 1, but in this case the accuracy dropped nearly 20% and indicates that the number of misclassified patches has an increase of 40, which is not acceptable, so the decrease of infection symptom is a more convincing explanation to the problem occurred in trial 3.

4.2 Infected Versus Mock

As for infected plants, both CBSV and the operations for inoculating the virus could have effect on the plant, it is necessary to ensure that the classification accuracy is mostly affected by the virus. Therefore, the model was also applied on the sample sets of mock versus infected. In this experiment, mock is a group of plants that were treated the same way as the infected group, except that mock groups were inoculated with 'empty' control E. coli plasmid instead of CBSV. Therefore, by applying classifier on mock and infected group, it can be measured that how much did the virus affected the classification. In theory, if the classification is mostly based on the virus but not the inoculation operation, the infected group and mock group should not be difficult to classify and can have a classification accuracy close to infected versus uninfected as measured in the previous section. The practical classification accuracy of infected versus mock is shown below:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

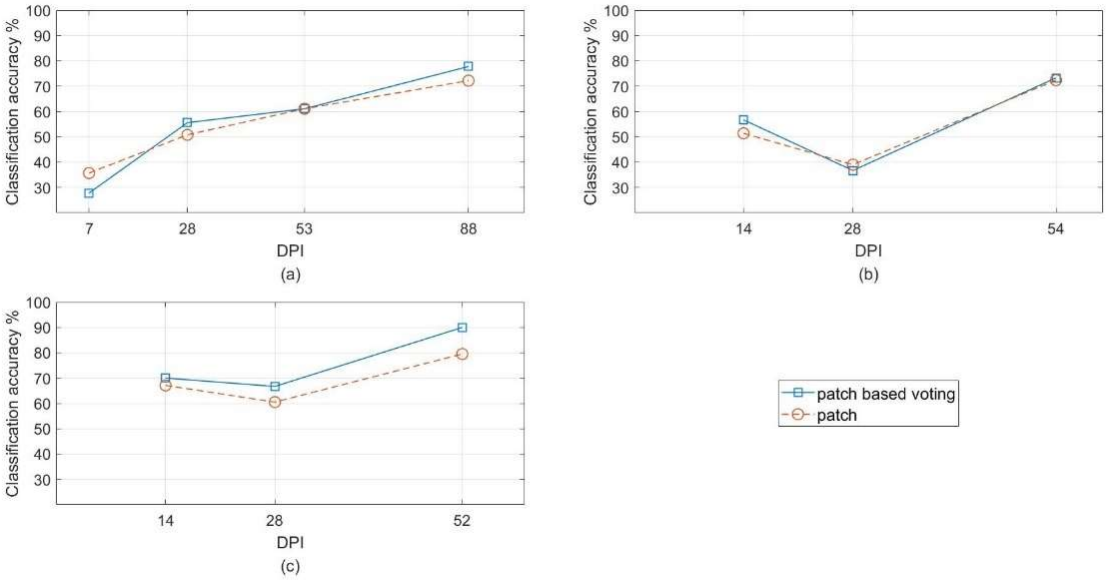

Figure 4. The classification accuracy (%) of infected versus mock leaf for three trials. (a) Classification accuracy of trial 1, sampled at 7,28, 53, 88 dpi. (b) Classification accuracy of trial 2, sampled at 14, 28, 54 dpi. (c) Classification accuracy of trial 3, sampled at 14, 28, 52 dpi.

Fig. 4 (a) shows the classification accuracy of infected versus mock. At 7 dpi, the accuracy is 27.8%, which is extremely low. With the increase of dpi, the accuracy first increases to 55.6% at 28 dpi, and then rises to 61.1% and 77.8% at 53 and 88 dpi respectively. To explain the low classification accuracy at 7 dpi, a mesh surface of the reflectance after KPCA is plotted as shown below:

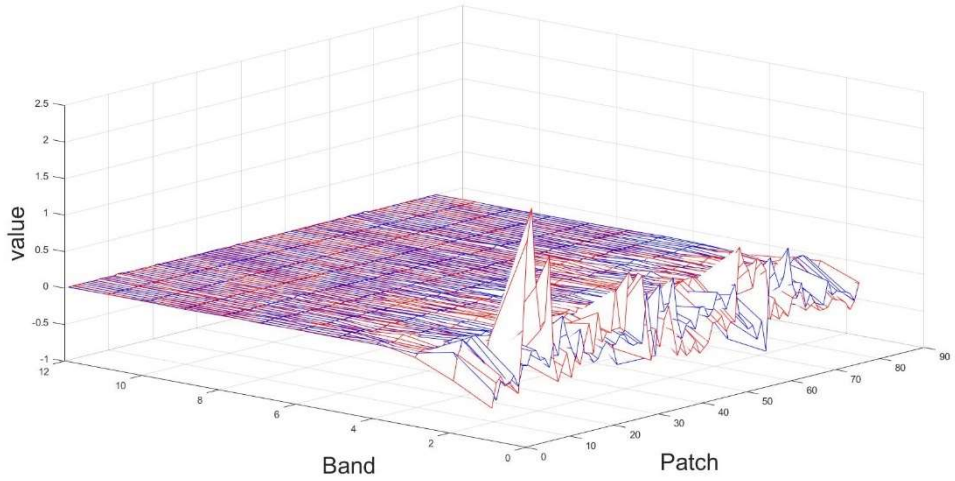

Figure 5. Mesh surface of reflectance after KPCA for trial 1 at 7dpi, red represents the reflectance of infected group, blue represents the mock group.

As it is shown in fig. 5, after KPCA, the reflectance value of each band varies between different patches. It can be seen that for the first half sets, infected patches tend to have a higher reflectance compared with mock patches, but for the second half sets, mock patches are more likely to have higher reflectance than infected patches. As half of the patches are used as the training set and the other half are testing set, it is possible that the trained model tends to recognise patches with higher reflectance in band 1 to be infected, which does not match the situation in the testing set and may

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

cause low classification accuracy. For the selection of training and testing set is fixed in this experiment, this problem can be attributed to the limited size of sample sets and is possible to be solved by reallocate the training set or increase the size of samples. For trial 2 shown in fig. 4 (b), classification accuracy at 14 dpi and 28 dpi are 56.7% and 36.7%. Then the accuracy increases to 73.3% at 54 dpi. The low accuracy at 28 dpi could be attribute to the similar reason in trial 1 7 dpi. For trial 3 shown in fig. 4 (c), classification accuracy was measured as 70% and 66.7% at 14 dpi and 28 dpi. It has a high accuracy at 52 dpi of 90%. The classification results shows that when dpi is low, it is difficult for the model to distinguish two groups, which indicates that the reflectance difference between infected group and mock group is not significant. Then, as dpi increases, there is higher possibility for the model to classify correctly between infected and mock patches. Therefore, it can be concluded that at the early stage of infection, CBSV does not has significant influence on the plant stress for trial 1 and trial 2. Only when dpi exceeded 50 did the virus caused distinct effect on the plants.

4.3 Uninfected Versus Mock

In the experiment, it is also necessary to quantify how much did the inoculation operation affected the plant. Therefore, the model was applied to classify uninfected group and mock group. In theory, if the injection does not have impact on the plants, the classification accuracy between mock and uninfected groups should be around 50%, which is close to random guessing. The practical result generated by the model is shown below:

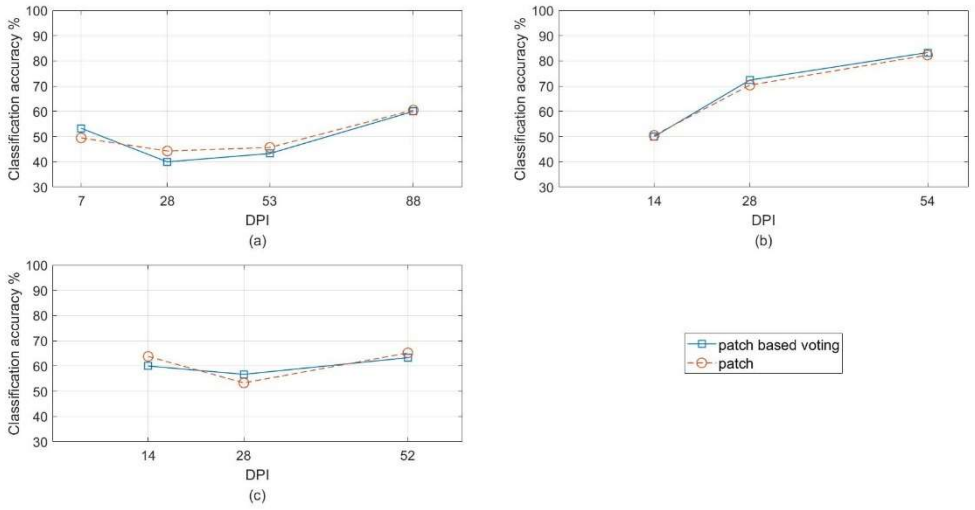

Figure 6. The classification accuracy (%) of uninfected versus mock leaf for three trials. (a) Classification accuracy of trial 1, sampled at 7,28, 53, 88 dpi. (b) Classification accuracy of trial 2, sampled at 14, 28, 54 dpi. (c) Classification accuracy of trial 3, sampled at 14, 28, 52 dpi.

Fig. 6 (a) shows that for trial 1, the accuracy varies between 40% to 60%, which is 50% $\pm$ 10%. It should also be noticed that for trial 1, the correlation between accuracy and dpi is low, which indicates that virus does not involve in the classification, as the effect of virus changes with time. For trial 2 shown in fig. 6 (b), classification accuracy is 50% at 14 dpi, but it then rises to 72.4% at 28 dpi and 83.3 at 54 dpi, which does not match the theoretical situation. As the accuracy changes significantly with the increase of dpi, the factor that caused this problem may not be the inoculation operation, for the accuracy is reasonable at the beginning. One possible explanation is that the environment of trial 3 was not perfectly controlled, the mock group may be infected or was influenced by other factors in the growing period which causes the difference between mock and uninfected group. For trial 3 shown in fig. 6 (c), the classification accuracy is fluctuated around 60%, and it does not change with dpi.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

Therefore, it can be estimated that the inoculation operation affected the mock group in trial 3, which caused slight difference between mock and uninfected plants.

Based on the three different experiments elaborated above, the result can be further discussed by taking all the results into consideration. As the classification accuracy of trial 1 in mock versus uninfected is around 50%, its accuracy measured for infected versus uninfected is more reliable compared with other 2 trials. According to the result of trial 1, the proposed model is not able to provide reliable classification at an early stage, but the accuracy can increase significantly with dpi. However, the symptoms of infected plants for different trial may also varies and influence the accuracy. As it is shown in fig. 4, the accuracy for trial 3 is much higher than trial 2, this may indicate that the symptoms appeared on the plants of trial 3 is more serious than trial 2 and could explain the high classification accuracy of trial 3 in fig. 3 (c). The impact of inoculation operation on each trial also varies. According to fig. 6, the operation caused more significant influence on trial 3, this may also contribute to the high classification accuracy in fig. 3 (c).

5. Conclusions and Future Work

5.1 Conclusions

In conclusion, this dissertation proposed a model for the classification of CBSV based on hyperspectral image. Firstly, a dataset obtained by A-MSI was selected to implement the experiment, it contains three trials, each trial has three groups and the total number of bands that can be detected is 15. Then, the images are sampled to enlarge the size of dataset by cropping leaves into patches. According to the result in [19], vegetation indices is abandoned in this experiment due to its poor performance, and patch-based voting method is applied instead. The sampled dataset is processed by two different algorithms. For the first algorithm, kernel principal component analysis is applied to the original dataset for the implementation of feature extraction, the crucial information is extracted and transformed to a linear separable set. The outcome of KPCA is classified by SVM to obtain the first probability map. For the second algorithm, SVM is directly applied to the initial dataset, and the outcome is further optimized by extended random walker to achieve another probability map. The probability map generated by these two algorithms is then fused to obtain the final result. The result shows that, this model can provide accurate classification when dpi exceeds 50. However, at early stage of infection, the classification accuracy varies between trials, which is not reliable. The result also shows that, patch-based voting can boost the accuracy in most cases. However, if the original accuracy is low, patch-based voting may weaken the performance of the model.

5.2 Future Work

In the future, the quality of cropped patches can be improved. The patches in this experiment are randomly selected and marked based on the group of corresponding leaves, which means some patches in the infected group may have no symptoms and mislead the model in the training process. Therefore, image processing algorithm can be applied to the initial dataset to assist the selection of patches. To avoid occasional case, it is possible to enlarge the size of samples or use random training set instead of a fixed set and take the average accuracy. In this experiment, the classification accuracy for three trials has significant difference, which indicated that the virus may cause different effect in different trials. Therefore, more trials are needed and tested to ensure that the result is reliable in all situations. As the result also shows that the effect of virus may not be the only factor involved in the classification, it is necessary to further improve the inoculation process to weaken the influence of factors other than the virus.

References

[1] J.P. Legg et al., “Comparing the regional epidemiology of the cassava mosaic and cassava brown streak virus pandemics in Africa,” Virus Research, Vol. 159, no. 2, pp. 161-170, 2011, doi: 10.1016/j.virusres.2011.04.018.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202210_8(10).0014

[2] M.N. Maruthi et al., “Transmission of Cassava brown streak virus by Bemisia tabaci (Gennadius),” Journal of Phytopathology, Vol. 153, no. 5, pp. 307-312, 2005, doi: 10.1111/j.1439-0434.2005.00974.x.

[3] M.M. López et al., “Innovative tools for detection of plant pathogenic viruses and bacteria,” International Microbiology, Vol. 6, no. 4, pp. 233-243, 2003, doi: 10.1007/s10123-003-0143-y.

[4] R. J. Hillocks and D. L. Jennings, “Cassava brown streak disease: A review of present knowledge and research needs,” International Journal of Pest Management, Vol. 49, no. 3, pp. 225-234, doi: 10.1080/ 0967087031000101061.

[5] D. M. Gates, H. J. Keegan, J. C. Schleter, and V. R. Weidner, "Spectral Properties of Plants," Appl. Opt, Vol. 4, no. 1, pp. 11-20, 1965, doi: 10.1364/AO.4.000011.

[6] A. A. Gitelson, Y. Gritz and M. N. Merzlyak, “Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves,” Journal of Plant Physiology, Vol. 160, no. 3, pp. 271-282, 2003, doi: 10.1078/0176-1617-00887.

[7] S. Jacquemoud and F. Baret, “PROSPECT: A model of leaf optical properties spectra”, Remote Sensing of Environment, Vol. 34, no. 2, pp. 75-91, 1990, doi: 10.1016/0034-4257(90)90100-Z.

[8] P. S. Thenkabail, R. B. Smith and E. D. Pauw, “Hyperspectral vegetation indices and their relationships with agricultural crop characteristics,” Remote sensing of Environment, Vol. 71, no. 2, pp. 158-182, 2000, doi: 10.1016/S0034-4257(99)00067-X.

[9] D. N. H. Horler, M. Dockray and J. Barber, “The red edge of plant leaf reflectance,” International Journal of Remote Sensing, Vol. 4, no. 2, pp. 273-288, 1983, doi: 10.1080/01431168308948546.

[10] A. K. Mahlein et al., “Development of spectral indices for detecting and identifying plant diseases,” Remote Sensing of Environment, Vol. 128, pp. 21-30, 2013, doi: 10.1016/j.rse.2012.09.019.

[11] A. F. H. Goetz, B. Curtiss and D. A. Shiley, “Rapid gangue mineral concentration measurement over conveyors by NIR reflectance spectroscopy,” Minerals Engineering, Vol. 22, no. 5, pp. 490-499, 2009, doi: 10.1016/j.mineng.2008.12.013.

[12] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," in IEEE Transactions on Geoscience and Remote Sensing, vol. 42, no. 8, pp. 1778-1790, Aug. 2004, doi: 10.1109/TGRS.2004.831865.

[13] J. Li, J. M. Bioucas-Dias and A. Plaza, "Spectral–Spatial Hyperspectral Image Segmentation Using Subspace Multinomial Logistic Regression and Markov Random Fields," in IEEE Transactions on Geoscience and Remote Sensing, vol. 50, no. 3, pp. 809-823, March 2012, doi: 10.1109/TGRS.2011. 2162649.

[14] G. Hughes, “On the mean accuracy of statistical pattern recognizers,” IEEE Transactions on Information Theory, vol. 14, no. 1, pp. 55–63, Jan. 1968, doi: 10.1109/TIT.1968.1054102.

[15] M. Pesaresi and J. A. Benediktsson, “A new approach for the morphological segmentation of high-resolution satellite imagery,” IEEE Trans. Geosci. Remote Sens., vol. 39, no. 2, pp. 309–320, Mar. 2001, doi: 10.1109/36.905239.

[16] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, “Classification of hyperspectral data from urban areas based on extended morphological profiles,” IEEE Trans. Geosci. Remote Sens., vol. 43, no. 3, pp. 480–491, Mar. 2005, doi: 10.1109/TGRS.2004.842478.

[17] X. Kang, S. Li, L. Fang, M. Li and J. A. Benediktsson, "Extended Random Walker-Based Classification of Hyperspectral Images," in IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 1, pp. 144-153, Jan. 2015, doi: 10.1109/TGRS.2014.2319373.

[18] P. Duan, P. Ghamisi, X. Kang, B. Rasti, S. Li and R. Gloaguen, "Fusion of Dual Spatial Information for Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 9, pp. 7726-7738, Sept. 2021, doi: 10.1109/TGRS.2020.3031928.

[19] Y. Peng et al., “Early detection of plant virus infection using multispectral imaging and spatial–spectral machine learning”. Scientific Reports, vol. 12, no. 1, pp. 1-14, Feb. 2022, doi: 10.1038/s41598-022-06372-8.