Available online at www.ijpe-online.com

International Journal of Performability Engineering

Vol. 13, No. 4, July 2017, pp. 469-478
DOI: 10.23940/ijpe.17.04.p14.469478

Classification of Potato External Quality based on SVM and PCA

Juntao Xiong[a], Linyue Tang[a], Zhiliang He[a], Jingzi He[a], Zhen Liu[a], Rui Lin[a], Jing Xiang[b,*]

[a]*College of mathematics and informatics, South china agricultural university, Guangzhou, 510640*
[b]*Hubei University for Nationalities, Enshi,445000*

Abstract

It is very important to classify and identify the quality of potato by computer vision. In order to realize the accurate and fast classification of potato, a classification and recognition method based on support vector machine and PCA is proposed. Study uses normal potato, green potato, germinated potato and damaged potato as the experiment sample. A total of 600 images were collected where 150 images of each sample was collected. The SVM multi classifier is designed to train the classifier based on the PCA principal component vector, and the key parameters of the classifier are optimized to improve the overall recognition rate of 96.6%. Separately, the normal potato recognition rate is 97.5%, the greened potatoes is 96.3%, the damaged potato is 95.0% and the germinated potato is 97.5%. The research results provide technical support for the intelligent grading of fruit and vegetable quality.

Keywords: potato, SVM, PCA, classification;

(Submitted on February 3, 2017; Revised on May 2, 2017; Accepted on June 8, 2017)

1. Introduction

Potato is one of the four major food crops in human society. It has great potential for processing because of the short growth cycle, strong adaptability and high yield. However, some of the external defects of the potato caused a serious impact on its quality. The traditional manual selection and classification method has the disad-vantages of low efficiency, high labor intensity and high false detection rate, which hinder the accurate and fast classification of potato in the actual production process.

Machine vision is one of the important techniques of non-destructive test for the quality of agricultural product [1,2,3,4]. With the development of machine vision tech-nology, domestic and overseas scholars have done a lot of research work on the de-tection and classification of the external quality of potatoes. Yu Xiaojuan [5] ex-tracted RGB components of potato images, propose a method of greened potatoes detection based on hue threshold division; Li Jinwei [6] proposed a fast gray intercep-tion segmentation threshold method to extract the dark part of potato surface and the ten color model for defects detection. Yang Dongfeng [7] proposed a method for greened surface detection of potatoes based on color character and neutral network. Zhou Zhu [8] proposed the volume method based on minimum circumscribed cylin-der to grade potatoes according to their size, the ratio of width and used length of the longest diameter circum-rectangle to grade potatoes according to their shape. Al-Mallahi [9] developed a machine vision system based on ultraviolet imaging to detect potato tubers on the potato harvester. Michael Barnes [10] introduced a method for detecting blemishes in potatoes using machine vision. Navid Razmjooy [11] pro-posed a hierarchical grading method based on machine vision using support vector machines combining with size sorting system. Hassankhani R [12] uses a compound of colour and physical properties of defects to classify potatoes.

* Corresponding author.
E-mail address: 110807515@qq.com.

In this paper, a method based on SVM (support vector machine) and PCA (principal component analysis) is proposed to classify and identify potatoes. The study realizes the effective detection and classification of normal potato, green potato, germinated potato and damaged potato. The overall classification recognition rate reached 96.6%, and the research results provide technical support for the application of machine vision technology in the intelligent grading of potato.

2. Materials and Samples

2.1. Experimental Materials

The varieties of the potatoes which as the experimental samples is Holland No. 15, produced in Guangdong China and purchased in Guangzhou Tianhe Changban Market. In order to detect the different status of the potato, the sample was bought at two different times. The first time, we purchased 300 normal potatoes. To make the potatoes sprouting and turning green, we sprayed a small amount of water mist on the surface of 150 samples after cleaning them, then deposited them into a black bag and placed in a dark storage cabinet. The remaining 150 were exposed to the sun and illuminated by incandescent lamp in overcast and night. After two weeks, we got the samples of germinated potatoes and green potatoes. At the same time, we purchased 150 normal potatoes and 150 damaged potatoes. Potatoes in different states are shown in Figure 1 the average bud length of germinated potato was about 3mm.

2.2. Image Acquisition System

The potato's shape is oval shaped or pear shaped and the center of mass is located in the mid-lower part. This paper mainly gathered static images and built an image acquisition system based on the above features. The image acquisition system is composed mainly of lighting system, industrial camera, digital image input interface and computer, as shown in figure 2.

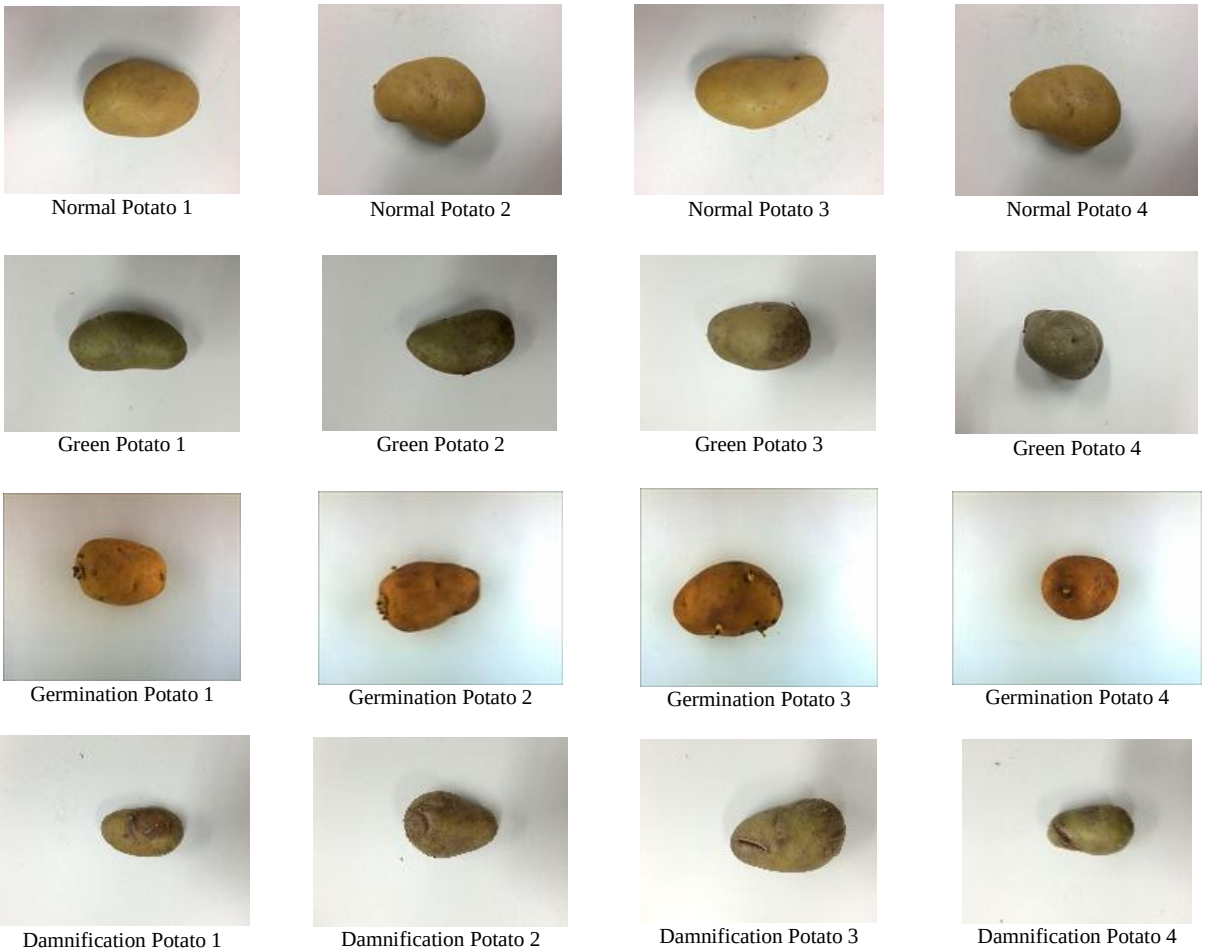


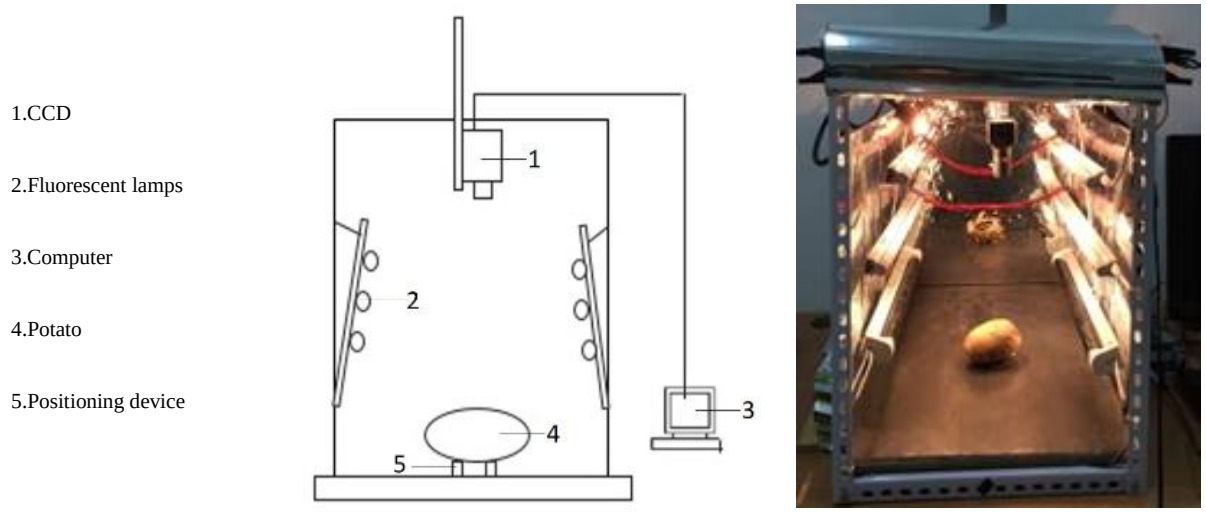
Figure 1. State of different quality potatoes

The lighting system is a square sealed light box, and the inner wall of the box body is a white non-reflecting paper, and the light source is composed of 6 white fluorescent lamps, which are distributed on both sides of the box body. The bottom of the lamp box is an image collecting area, and the background of it is white. In order to facilitate the collection of images in all directions, a white adjustable positioning device is arranged on the bottom and the industrial camera is installed at the upper part of the collecting area. In order to ensure the images are clear and without distortion, according to the preliminary experimental results, set the exposure time is 30ms, the object distance between camera and sample is 500mm and the CCD camera FOV is 20 degrees. The image acquisition process is completed in the sealed light box to avoid the external light source and other noise interference. In this study, the side projection of potato is the main research object.

3. Materials and Samples

3.1. Image Classification Algorithm

Image classification is based on the different characteristics of the image information of classification object, using the computer to carry on the quantitative analysis to the image, so as to separate the target area of different types, including image acquisition, image preprocessing, image feature selection and extraction and image classification modeling. In general, the difference between the different images or image regions should be as large as possible, and the internal properties of the same kind of image or image region should be guaranteed to be stable. The steps for classification using support vector machines are as follows.



(a) Diagram (b) Picture of real products

Figure 2 Image acquisition system

- Step 1: collect training sets and test sets for each class
- Step 2: preprocess the sample image, such as noise reduction, gray scale, etc.
- Step 3: extract image characteristic of training set and test set for image classification
- Step 4: use the classification characteristic to carry on the training and obtain the SVM classifier
- Step 5: use the trained classifier to classify the test set
- Step 6: verify whether the classification results meet the requirements
- Step 7: optimization parameters, retraining until the classification results to meet the requirements.

The classification process is shown in Figure 3.

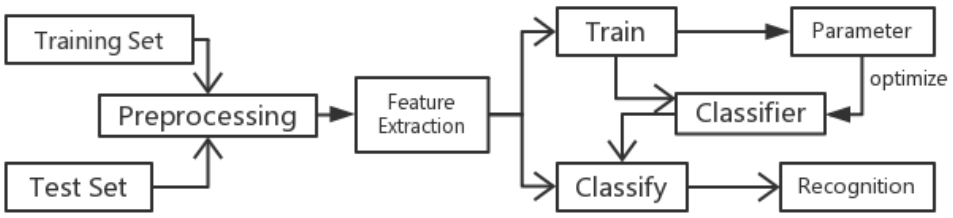


Figure 3. Diagram of image classification flow

3.2. Support Vector Machine

The main idea of SVM (Vector Machine Support) is to establish a hyper plane as a sample classification surface. For the two-dimensional linear separable case, it can not only separate the two kinds of samples, but also make the largest classification margin. In the high dimension space, the optimal classification line becomes the optimal classification hyper plane. The optimization function can be expressed as follows.

$$\begin{aligned} &\min \frac{1}{2}\|w\|^2 \\ &s.t. y_i[w^T X_i + b] - 1 \geq 0, i = 1, 2, ..., n \end{aligned} \tag{1}$$

Where x is the d-dimensional feature vector, $y \in \{+1,-1\}$ is category labels, w is the weight vector and b is a threshold of classification. The sample points, which make the constraint equal, are called support vector. By constructing the Lagrange function, the equation can be transformed into the following expression.

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{n}\alpha_i \tag{2}$$

Where α_i is the Lagrange multiplier. The corresponding discriminant function of function 2 is shown below.

$$D(x) = sgn(\sum_{i=1}^{n} \alpha_i^* y_i (x \cdot x_i) + b^*) \tag{3}$$

In the case of nonlinear separable, the support vector machine maps the input variables to the high dimensional feature space by kernel function, and the optimal classification hyper plane is constructed in the high dimension space. The common kernel [13,14] functions include radial basis function, polynomial function and Gauss kernel function, etc., as shown in Figure 4 is a schematic diagram of Gauss kernel function.

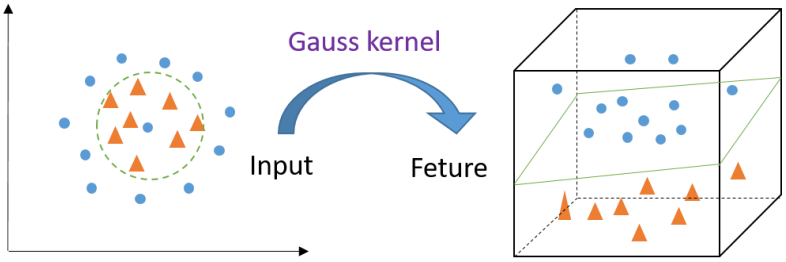


Figure 4. Diagram of Gauss kernel function

In addition, the slack variables, which indicate the amount of allowed deviation of the sample point from the margin edge, are introduced into the optimization function to reduce the influence of outliers on the classification results. At the same time, the optimization function is transformed into the following equation by adding constraints to the slack variables.

$$\begin{aligned} &\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi^{(i)} \\ &s.t. y^{(i)}(w^T x^{(i)} + w_0] \geq 1 - \xi^{(i)}, \forall i \\ &\xi^{(i)} \geq 0, \forall i \end{aligned} \tag{4}$$

Where $\xi^{(i)}$ is the slack variable, and C is the penalty parameter which is used to control the weight between the maximum margin and the deviation in the objective function.

3.3. Support Vector Machine

In the actual classification process, the classification accuracy of SVM is reduced when the samples of a certain class are mixed into another one, which will lead to that the hyper plane of classification is not optimal. To solve this problem, the paper proposes a method to improve the hyper plane by traversing the parallel planes of the classification plane in the margin. Specific steps are as follows.

Bulleted lists may be included and should look like this:

- Step 1: solve the hyper plane equation
- Step 2: use discriminant function D(x) for classification, record the current recognition rate as R0
- Step 3: set F(x) $= sgn(\sum_{i=1}^{n} \alpha_i^* y_i(x \cdot x_i) + b^* + d)$, n=0, d=1, d*=0, R*=R0, T=0.6;
- Step 4: use F(x) as the discriminant function for classification and record the corresponding recognition rate as R
- Step 5: compare the size of R and R*. If R*<R, set d*=d, R*=R; if R*⩾R, calculate the threshold P=exp ((R-R*) / (2-0.1n)). If P>T, set d*=d, R*=R, otherwise, don't update the value of d* and R*.
- Step 6: set d=d-0.1, n=n+1;
- Step 7: repeat step 4-6 until d= -1;
- Step 8: use d* to update F(x) and use the discriminant function for classification.

If d*=0, then there is no better hyper plane than the original classification.

4. Experiment and Analysis

4.1. Prepare Training Set and Test Set

The experiment has prepared normal potato, green potato, germination potato and damnification potato in the preparatory stage. Using image acquisition system to collect images of each potato, a total of 600 valid samples were obtained. From each kinds of potato images, selected 70 images that have representative and as far as possible including the shape of potatoes as training set, and 80 samples are used as the test set. Each kinds of potato has collected 150 samples.

The experimental samples may have both germination and green two defects at the same time. In the view of this situation, we mainly to detect germination defect. Because germinated potatoes and green potatoes are not edible, and germination is more likely to be detected than green.

4.2. PCA Feature Extraction

PCA (Principal Component Analysis) is one of the most widely used feature extraction methods. The main idea of PCA is to extract the main features of the original data, reduce the data redundancy, making the data to be treated in a low dimensional feature space, while maintaining most of the information in the original data, so as to solve the bottleneck problem of high dimension of data space [15]. The general steps are as follows.

- Step 1: calculate the mean of the data, subtract the mean with the original data to get the standard data
- Step 2: calculate the covariance matrix R
- Step 3: calculate the characteristic value and characteristic vector of R
- Step 4: sort the feature value from large to small, and the feature vectors corresponding to top m eigenvalues are transformed into the matrix A
- Step 5: transform the n-dimensional vector into a new m-dimensional vector by Y=ATX

In this paper, in addition to four main features, respectively as normal, green, sprouting and epidermal injury, the potato images also includes other feature such as color, texture, and shape. On the other hand, the pixel size of the image captured by the image acquisition system is 1200*950, so extracting features from the pixel level will increase the complexity of the calculation. Based on the above two factors, this paper does not use a single feature as the classification basis, and does not calculate all pixels. Instead, the PCA dimension reduction method is used to remove the correlation of pixels, and the linear combination of the extracted feature vectors is used as the classification feature.

The feature vectors are extracted from all the training samples and projected onto the two-dimensional space which can be expressed in the form of gray scale images. Figure 5 shows the top 20 principal component images (from left to right and then from top to bottom for pc1~pc5... Pc16~pc20).

4.3. Data Normalization

When there is a need to use two or more features, due to different physical meaning of features, the range of values is also very different and does not have the direct comparability. Therefore, before the comprehensive utilization of different features, we need to normalize the values of different features [16]. In this paper, we use the minimum eigenvalue l and the maximum eigenvalue u to normalize the eigenvalue *x* to [0, 1].

$$x' = \frac{x - l}{u - l} \tag{5}$$

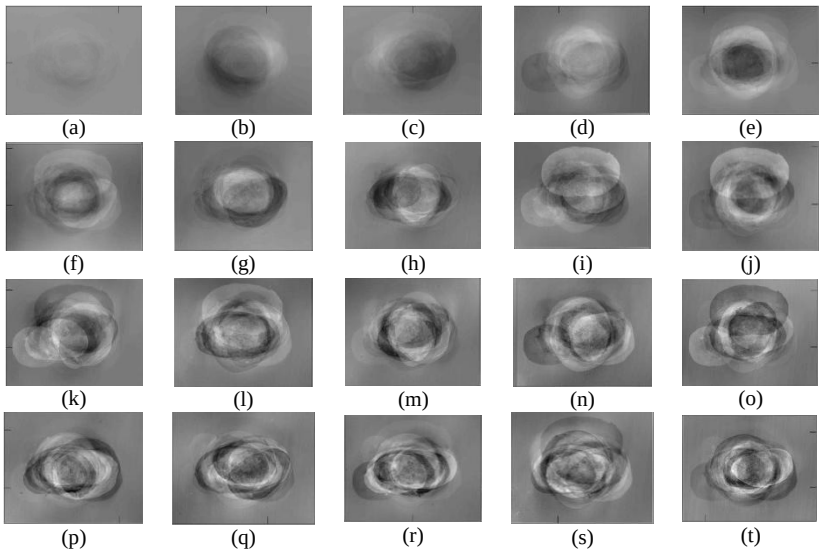


Figure 3. The top 20 principal components image

4.4. SVM Multi Classification

The basic SVM classifier is only applicable to the binary classification problem, for more than 3 kinds of classification, it is necessary to construct a suitable SVM multi classifier. At present, the main multi classification strategies have one-to-many maximum response strategy, one-to-one voting strategy and one-to-one elimination strategy. After testing, the effect of the one to one voting strategy is better in this paper, the concrete algorithm is as follows.

- Step 1: each of the two class samples constructs a classifier, set as (A,B), (A,C), (A,D), (B,C), (B,D), and (C,D)
- Step 2: set: V(A)= V(B)= V(C)= V(C)=0;
- Step 3: vote: the test samples will be fed into 6 classifiers, if the classification result of training set (A,B) is that the sample belongs to class A, then V (A) = V (A) +1, otherwise, the sample belongs to the B class, V (B) = V (B) +1... And so on, finally will get a set of results of V
- Step 4 decision: take Max (V(A),V(B),V(C),V(D)) and judge the sample belong to the class which the maximum value corresponding to.

The research is based on the Matlab software platform, using one-to-one voting strategy to train the classifier, and establish the classification system, and use the test set to test the classifier. Test results are shown in Figure 6. The algorithm flow chart is shown in Figure 7.

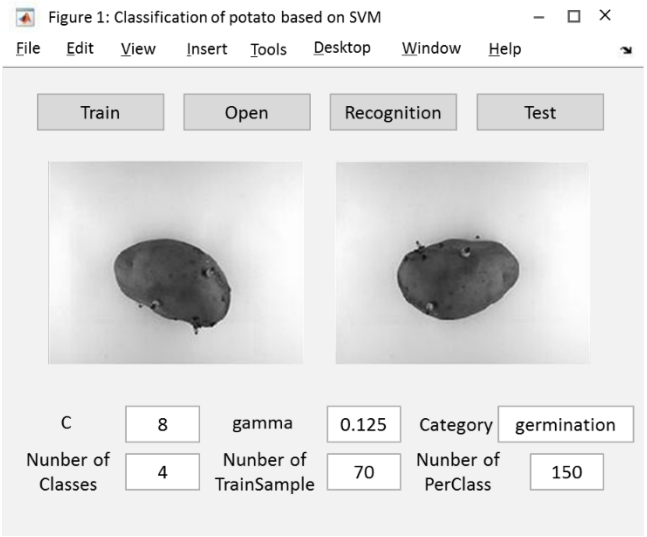


Figure 6. Test result of SVM classifier

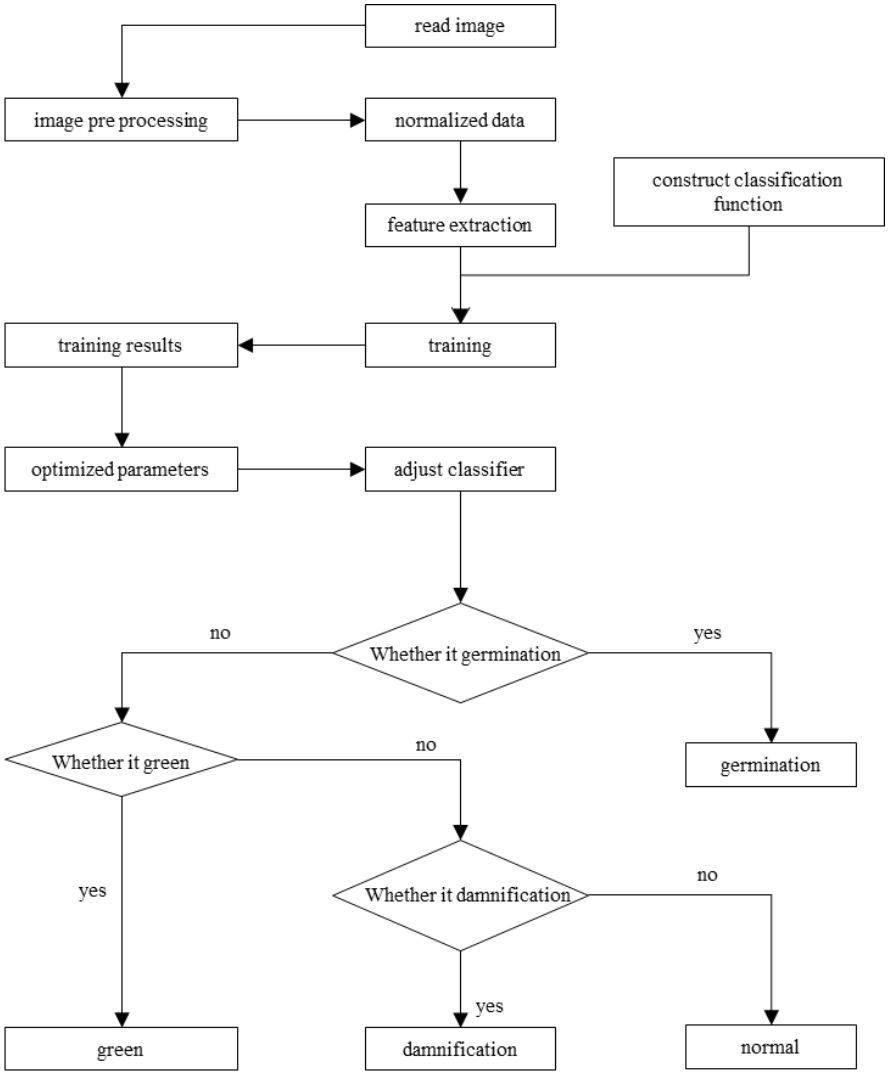


Figure 4. The flow chart of algorithm

4.5. Parameter Optimization and Analysis

The classification effect of the nonlinear classification problem has two key factors: the kernel function and the penalty parameter C. The kernel function used in this paper is the radial basis function (RBF), and its expression is as follows.

$$K(x, y) = \exp(-\gamma \|x - y\|^2) \quad (6)$$

In the expression, it is needed to determine the parameters of gamma, gamma and C to compute the impact of the recognition rate of the test set. The process of determining the optimal parameters is the process of the optimization problem. With the aid of the parameter optimization tool provided by Dr. Lin Zhiren in LibSVM [17], based on cross validation and grid search, the parameter optimization results were obtained as C=8 and gamma=0.125, and the process of finding the parameters as shown in Figure 8.

In order to further determine the superiority of the parameters, the values of the 4 pairs of parameters appearing in the process of seeking parameters are selected to test the test sets. The statistical results are shown in Table 1.

As can be seen from table 2, C=8 and gamma=0.125 are the best parameters for the recognition effect of the current data sets. Using the parameters to classify the test set, the results obtained are shown in Table 2.

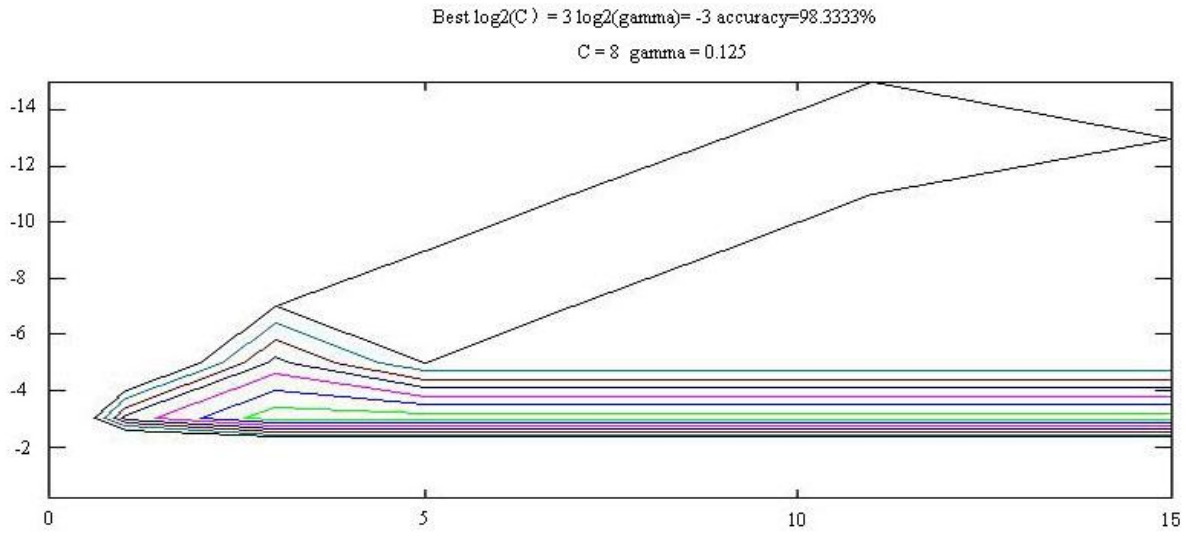


Figure 5. The result of parameter optimization

Table 1. An example of a table

C value	value	recognition rate
8	0.125	96.6%
8	0.0312	91.4%
512	0.0078	74.7%
2048	0.0078	68.3%

Table 2. The statistics of recognition results of potatoes

Sample state	Sample number	Identification number	Recognition rate	Processing time
Normal	80	78	97.5%	0.137s
Green	80	77	96.3%	0.142s
Germination	80	78	97.5%	0.146s
Damnification	80	76	95.0%	0.163s
Statistics	320	309	96.6%	0.147s

Compare the recognition rate with the not improved discriminant function results. The comparative results are shown in figure 9. It can be seen from the figure that after improving the discriminant function, the recognition rate of each category is improved, among which the amplitude of recognition rate of green potatoes and damaged potatoes is the largest, increased by 4.8% and 5% respectively.

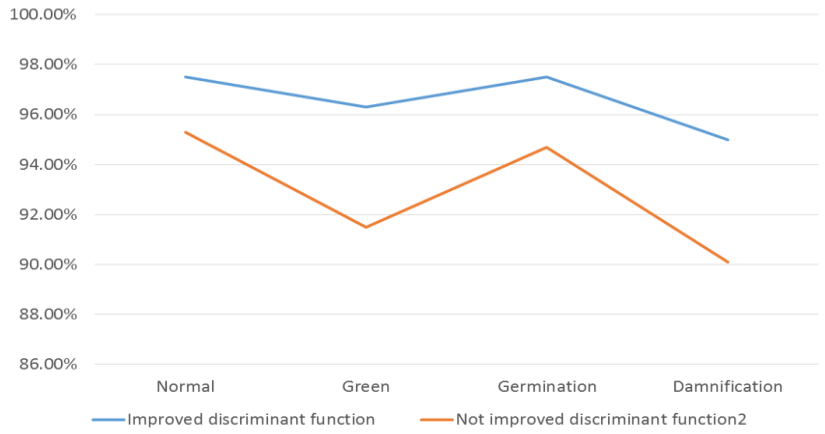


Figure 9. The comparative result of discriminant function

Before using SVM for classification, we also tried the PLA and KNN methods. For the green potato, the perceptron learning algorithm (PLA) was used to distinguish the normal potato and green peel potato, and the K-nearest neighbor (KNN) classification algorithm combined with edge detection method was used to recognize the germinated potato. Then the median filter and the area of the solution were used to determine the potato skin injury, finally realizes the classification of potato

quality. The potato quality test results with PLA and KNN as follows: the recognition accuracy of normal potato is 96.8%, the green potato is 89.7%, the damaged potato is 90.4%, and the germinated potato is 96%. The comparative results of two methods are shown in figure 10.

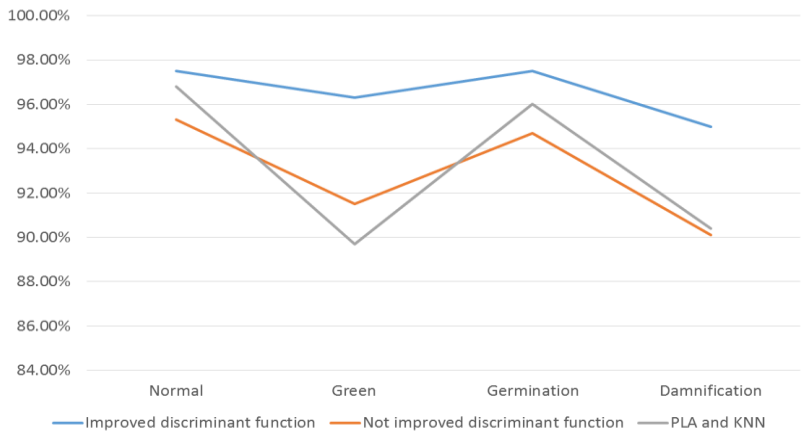


Figure 10. The comparative result of two methods

From the test results of research, it can be seen that improved SVM has a better recognition effect and greater generalization ability, especially for green potatoes and damaged potatoes. But there are still some shortcomings. Analyzing the error of the test results, the main reasons are as follows: the sprout of germinated potato is deviated from the camera angle, which is easy to miss; the damage of some damaged potatoes is not obvious, which is easy to be confused with the normal potato.

In view of the above problems, follow-up study can be from the following improvements: for the germinated potato, we can capture 3D images from multiple sides, or perform real-time detection by rotation; for the damaged potato, the contrast of the damaged parts can be increased by the enhancement algorithm in image preprocessing, which is advantageous to the classification of the damage characteristics.

5. Conclusion

The paper studies the application of combining PCA and SVM algorithm for image classification in the field of fruit and vegetable. In this paper, the potato is selected as the experimental object. The paper used principal component vectors as the classification feature, analyzed and designed the SVM multi classification system and improved the classification discriminant functions effectively identify normal potato, green potato, germinated potato and damaged potato. Using Lib-SVM parameter optimization tool to obtain the best classification parameters (C=8 and gamma=0.125), the recognition rate of the classification of the overall sample reached 96.6%.

Acknowledgements

This work is supported by: (1) the National Natural Science Foundation of China, NO: 31201135, (2) Pearl River S&T Nova Program of Guangzhou: 201506010081, (3) Natural Science Foundation of Guangdong Province of China (2015A030310258).

References

1. Al-Mallahi A., Kataoka T and Okamoto. Detection of potato tubers using an ultraviolet imaging-based machine vision system. *Biosystems Engineering*. (2010), 105(2): 257~265.
2. Barnes M and Duckett T. Visual detection of blemishes in potatoes using minimalist boosted classifiers. *Journal of Food Engineering*, (2010), 98(3): 339~346
3. Gao Xuwei. Feature Extraction Method Based on KPCA and Its Application. *Nanjing: Nanjing University of Aeronautics and Astronautics*, (2009).
4. Hassankhani R, Navid H and Seyedarabi H. Potato surface defect detection in machine vision system. *African Journal of Agricultural Research*, (2012), 7(5): 844~850.
5. Li Jiangbo, Rao Xiuqin and Ying Yibin. Detection of navel surface defects based on illumination-reflectance model. *Transactions of the Chinese Society of Agricultural Engineering*, (2011), 27(7): 338~342.

6. Li Jinwei, Liao Guiping and Jin Jing. Method of potato external defects detection based on fast gray intercept threshold segmentation algorithm and ten-color model. *Transactions of the Chinese Society of Agricultural Engineering*, (2010), 26(10): 236~242.
7. Liu He and Wang Maohua. Method for classification of apple surface defect based on digital image processing. *Transactions of the Chinese Society of Agricultural Engineering*, (2004), 20(6): 138~140..
8. Liu Xiaotong. Study on Data Normalization in BP Neural Network. *MECHANICAL ENGINEERING &AUTOMATION*, (2010), 03: 122~126.
9. Navid Razmjooy, Somayeh Mousavi and Soleymani. A real-time mathematical computer method for potato inspection using machine vision. *Computers & Mathematics with Applications*, (2012), 63: 268~279.
10. Song Hui, Xue Yun and Zhang Liang-jun. Research on Kernel Function Selection Simulation Based on SVM Classification. *JISUANJI YU XIANDAIHUA*, (2011), 08: 133~136
11. Song Yihuan, Rao Xiuqin and Ying Yibin, "Apple stem/calyx and defect discrimination using DT-CWT and LS-SVM," *Transactions of the Chinese Society of Agricultural Engineering*, (2012), 28(9): 114~118.
12. Yang Dongfeng and Chen Zhengguang. Greened Surface Detection of Potatoes Based on Color Character. *Journal of Heilongjiang Bayi Agricultural University*, (2011), 23(1): 83~87.
13. Yang Hai. Research and Application on SVM Kernel Parameters Optimization. Hangzhou: Zhejiang University, (2014).
14. Yu Xiaojuan, Liao Guiping and Li Jinwei. Greened potatoes detection based on hue threshold division. T*ransactions of the Chinese Society of Agricultural Engineering*, (2009), 25(Supp.2): 314~319.
15. Zhan Hui, Li Xiaoyu and Wang Wei. Determination of chestnuts grading based on machine vision. *Transactions of the Chinese Society of Agricultural Engineering*, (2010), 26(4): 327~331.
16. Zhou Zhu, Huang Yi and Li Xiaoyu. Automatic detecting and grading method of potatoes based on machine vision. *Transactions of the Chinese Society of Agricultural Engineering*, (2012), 28(7): 178~183.
17. Zhu Aihong, Zhao Shuai and Mao Mingliang. Research for Classification System of projects text based on LIB-SVM. *Microcomputer Information*, (2011), 27(4)