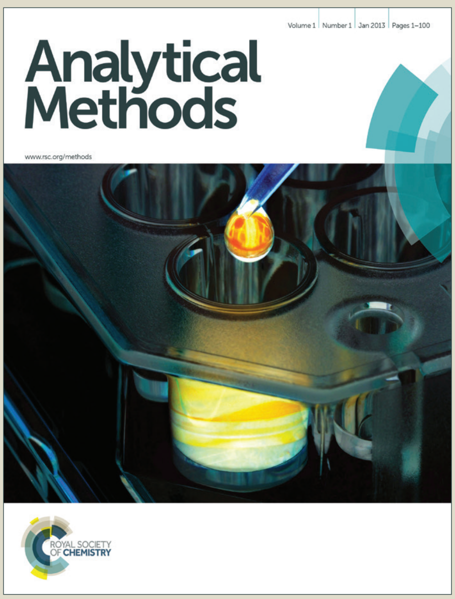
Analytical Methods

Accepted Manuscript



Non-destructive gender identification of silkworm cocoon using x-ray imaging technology coupled with multivariate data analysis

Running title: ***Gender Identification of Silkworm Cocoon with X-ray Imaging***

Jian-rong Cai*, Lei-ming Yuan, Bin Liu and Li Sun
School of Food & Biological Engineering, Jiangsu University, Xuefu Road 301, Zhenjiang City, Jiangsu Province, 212013, China.
*Corresponding author: Jian-rong Cai, e-mail: jrcai@ujs.edu.cn.

Abstract

A rapid, reliable and non-destructive method for gender discrimination of silkworm cocoon is of great importance to the mulberry silkworm industry for producing high quality silk. The objective of this study was to determine the feasibility of soft X-ray imaging with multivariate data analysis to classify the gender of silkworm cocoon. X-ray images of silkworm cocoon were obtained and pre-processed, and then region of interest (ROI) of chrysalis was segmented. Totally 11 morphological characters of chrysalis were extracted and compressed by principal component analysis (PCA) to visualize the cluster trends. In developing the discrimination classifiers, four kinds of algorithms including K-nearest neighbors (KNN), linear discriminant analysis (LDA), back propagation artificial neural network (BP-ANN) and support vector machine (SVM) were comparatively employed, and the performance of those models were optimized by cross validation. The results indicated that the correct identification rates of these classifiers were all high and in the range from 92.57% obtained via SVM to 93.68% obtained via KNN. On the consideration of the running time, LDA classifier was highlighted for its 93.31% accuracy. The results indicated this X-ray imaging technique with multivariate analysis could successfully be used in the screening of the gender of silkworm cocoon in the mulberry silkworm industry.

Key words: *X-ray imaging technique; silkworm cocoon; pattern recognition; image process; gender identification*

1. Introduction

Silkworm silk is one of the most important materials utilized extensively in the textile industry. Undoubtedly, the silkworm silk with high quality draws consumers' purchasing tendency and gains a bigger market share. The mulberry silkworm & silk industry is a traditional industry in China, and the exportation of silk is going to increase every year. There are gender differences in mulberry silkworm industry, and the silk of male silkworm has advantages over the female silkworm, and its price is higher. According to experience, the silkworm silk produced by male silkworm basically has several advantages such as more even fineness, higher grade of raw silk, and better elasticity. Thus silkworm cocoons are usually separated according to their genders manually. This can greatly improve the quality of silk products without increasing any inputs.

At the present stage, the mostly used method is that cutting the cocoon by manual operation and identifying the pupa gender by its physical features. This traditional method seems to be less effective when large amount of silkworms need to be classified at the breeding season or before the reeling process because of consuming mass labor. Moreover it is likely to damage the pupa and judge falsely, which causes the male and female pupa mixed. The mix up of pupas brings about parental pupa copulations, which will weaken the quality of silkworm eggs. Thus, it is of great importance to separate the female from the male, and to exert the advantages of male silkworm cocoon. In this situation, a rapid, reliable and nondestructive method, which can reduce manpower, accelerate identification and be limited damage to pupas, is urgently required for discriminating the gender of silkworm cocoon.

Near infrared spectroscopy (NIRS) is an all-purpose analysis method, and its excellent analysis has been proved in agriculture products [1, 2]. The attempt of reflectance spectra of pupas with chemo-metrics has been used to inspect gender of silkworm and obtained more than 90% identification rates for gender discrimination [3]. However, some pre-processes were needed before NIRS approaching, such as cocoon cutting, fixed point of spectral reading, et al. What’s more, the accuracy of NIR determination was affected by temperature, humidity, the distance between detector and acquiring point.

Other promising non-destructive approaches can be investigated via the differences of morphological characters according to the gender of silkworm cocoon. With the image processing technique, shape representation and description techniques can be developed to distinguish the gender of silkworm cocoon. Thus the projective information of silkworm chrysalis is possible to study the discrimination. With the illumination of a UVA light source, female silkworm pupa can be distinguished from the male via fluorescent imaging technique [4]. But this technique is limited to some species because of the improvement of silkworms through dye-modified diet and cross breeding [5]. Magnetic resonance imaging (MRI) technology was employed to capture the weighted images and was considered as a potential technique to classify the silkworm genders [6], but with key disadvantages on the speed and the equipment cost. In order to settle above problems, a few low-cost light emitting diodes (LED) were proposed and demonstrated on silkworm pupa [7]. Each LED emitted different wavelength spectra on pupa, and special characterizers only found on female chrysalis [8]. This optical penetration-based silkworm pupa gender identification was studied and 88% discrimination accuracy was obtained [9]. Obviously, the identification rates of those attempts based on the computer vision were not better than the result of NIRS approach. But the exercisable ability of computer vision in identifying the gender of silkworm pupas let it highlight from other approaches, such as NIRS [3], weighting [10], MRI [6], DNA analysis [11] et al.. However, a key problem that exists in the application of computer vision is that the cocoon must be cut before capturing the projective information, which will tremendously weaken the practical utilization.

X-ray, also called roentgen ray, is a type of electromagnetic radiation with the wavelength 0.01~ 10 nm. The energy of photon in X-ray system usually is higher than 100 keV, which results a strong penetrability to the target objects. The wavelength in range 1~10 nm is often called as soft X-ray, whose applications were arisen in food inspection. The principle of soft X-ray inspection is on the basis of density of product and the contaminant. X-ray is one of the most effective tool to inspect the agriculture products, such as the inside defects of pecan [12], fungal infection in wheat [13], the defrosted fruit [14].

Due to the differences in the density of cocoon and chrysalis, X-ray image of silkworm cocoon can exhibit the discrepant information: penetration intensity in cocoon area is higher than that in chrysalis area. These discrepant representations can be described by image process [15] and are used as the input of supervised classifiers, which are calibrated by pattern recognition methods. Therefore, in this study an appropriate X-ray imaging technique coupled with image process, pattern recognition methods was proposed. The objective of this study was to determine the potential of soft X-ray imaging with multivariate data analysis to classify the gender of silkworm cocoon in the mulberry silkworm industry.

2. Materials and methods

2.1 Samples Collection

Three hybrid varieties of silkworm samples (*Bombyx mori L.*) were collected for this experiment, and all samples were provided by Zhenjiang Sericulture Research

Institute, Chinese Academy of Agricultural Sciences. The hybrid varieties were *'JingSong'×'HaoYue', 'Shuangkang'×'Yongkang'* and *'Suzhen'×'Chunguang'* respectively. Couples of batches of samples were gathered from two seasons 2009.06 and 2010.06. Each batch of about 80~100 samples was carried to the laboratory for at least 12 hours storage before X-ray imaging operation. Once finishing the X-ray scan, the samples were delivered to Sericulture Research Institute and the gender was checked by the traditional approach. The technologist cut the cocoon and identified the gender by the inherent morphological features of pupa. In the routine of identifying the gender of silkworm pupas, some unqualified samples were got rid of, such as rotten or dead chrysalis. **Fig.1** shows the photograph information of silkworm cocoon and silkworm pupa.

2.2 X-ray Imaging Acquirement

In this experiment, a soft X-ray digital imaging system (**Fig.2**) was utilized, which consisted of a frame grabber, X-ray controller, detector, tube, stepper motor, conveyor belt, protective cover and computer, along with appropriated software (X-view4.2). The wave, which generated by the X-ray tube (model: *T80-1-60*, voltage: 20kV~160kV, current: 0.1mA~1mA, *Beijing mechanical and electrical technology company, China*), penetrated through the detected sample on the conveyor belt and was captured by X-ray detector (model: *0.4f2-307*, line scanning 1*768, pixel size: 0.4mm, scanning speed: 0.1~0.8m/s, *DT company, Finland*). Then the incoming wave of X-ray detector was first converted to visible light and accurately measured. After signal processing by X-ray detector, the digital data signal was sent to the frame grabber. During samples moving on conveyor belt, sequential signals were acquired by this linear scan frame grabber. At last an image (1024*768 pixel size) featured with gray information was manually obtained and then displayed on computer screen.

Before grabbing the radiographs of silkworm cocoon, some parameters of the X-ray system were optimized by preliminary works, and those parameters were: a) the speed of conveyer belt was set to 0.15m/s; b) the integration time of X-ray detector was set to 2.22ms; c) the voltage of X-ray tube was set to 40kV; d) the current was set to 0.6mA. The X-ray detector was calibrated as the literature [16] described to eliminate the differences in six modules of X-ray detector. The cocoon and pupa could be both penetrated through by X-ray energy. What's more, the differentiation in penetration through cocoon and pupa was different due to their density properties. Thus image part of pupa region was darkened and highlighted from cocoon or belt. After setting the imaging parameters and calibrating X-ray detector, several samples were put simultaneously without any touch or overlap to each other on the conveyer belt and grabbed radiographs by this soft X-ray digital imaging system. Images were acquired with silkworm cocoons in a random orientation and were read in 12-bit resolution.

2.3 Image processing and features extraction

A full depiction of the key steps involved in image processing algorithms for identifying the gender of silkworm cocoons is shown in a flow chart presented in **Fig.3**. The algorithm had various steps of process. The first step was for image pre-processing and segmentation. The second step was for geometrical feature extraction, such as area, perimeter, parameters of ellipse and shape. The third step was to build a robust classification model for gender discrimination. The first two steps were conducted in a portable computer with OpenCV1.0 (opening computer vision library, version 1.0) and VC++6.0, and the third step was investigated in Matlab (version 2008a, the Mathworks, USA) to obtain an appropriate classifier.

2.3.1 Image pre-processing and segmentation

Fig.4 shows the binary image of the silkworm cocoon after being processed, and some differences can be seen between pupas. **Fig.5a** shows the original X-ray image, where the region of the pupa is darker than others. Before image segmentation, grey contrast enhancement and de-noising filter were employed to enhance the region of pupa part. And after segmentation, morphological operations were needed to filter out the unwanted regions (i.e. noises).

Firstly, in the aspect of grey contrast enhancement, the grey value of pixel was stretched as **Eq.1** described. The value lower than 150 in *f(x,y)* was set to 0 and that higher than 250 was set to 255, while the value in the interval [150, 250] was stretched to [0, 255].

$$f'(i,j)=\begin{cases}0, & 0 \le f(i,j)<150\\ \dfrac{255-0}{250-150}\times[f(i,j)-150], & 150\le f(i,j)\le 250\\ 255, & 250< f(i,j)\le 255\end{cases} \qquad \text{Eq.1}$$

Where *f(x,y)* is the grey value in original image **Fig.5a**, and *f'(x,y)* is the stretched grey value **Fig.5b**.

Then, de-noising filter was tried to remove the noise, which might be solid particle, cocoon or other sundries. A Gaussian filter of 3×3 window size was used to remove those noises and the result was better than average, median, and wiener filters on the enhanced image. A Gaussian filter is a line filtering technique which allows the edges to be preserved while filtering out the unwanted noise. It replaces the original pixel with the Gaussian of its 8-neighbouring pixel values. **Fig.5c** shows the de-noising image, and the edge of pupa region can be clear from the background (i.e. the conveyor belt and the cocoon).

Next, a global threshold segmentation method as **Eq.2** shown was utilized to separate the silkworm pupa from the background. The grey value of pixels which was lower than 250, was set to 0, and grey value of the rest pixels was set to 1. The result of the threshold image was a discrete binary image, which as **Fig.5d** shown, mixed with some small regions, which needed to remove.

$$g(i,j)=\begin{cases}0, & f'(i,j)<250\\1, & f'(i,j)\geq 250\end{cases} \qquad \text{Eq.2}$$

Where $f'(x,y)$ is the filtered image **Fig.5**c, and $g(x,y)$ is the binary image **Fig.5d** (**Fig.4** is a result of the above process).

Finally, a morphological operation was utilized in order to reduce or eliminate the unwanted regions (undesired speckles outside of the pupa). Opening operation of 3×3 window size was highlighted over erosion or dilation on the performance of the segmented output as **Fig.5e** shown, small regions of noises were removed and only the region of pupa was remained.

2.3.2 Features extraction

A simple trick was needed to check whether the outputs were the region of interest (ROI) for silkworm chrysalis. Firstly outputs were sorted by their columns or rows and each region was read one by one according to its sequence. Then two parameters (i.e. *Area* and *Perimeter*) were extracted from this output region to validate whether it was the ROI for pupae by the discriminated criteria: a) area >1000; b) perimeter >120. If one of region's parameters could not meet the designed criteria, the output would be ignored and returned to a next region. Only the extracted parameters of output regions met the both two conditions, the output should be acknowledged as ROI for pupa region. An object can be represented in features of its external characteristics, and there are many features can be used to describe an object. Thus, usually the objects of one class can be distinguished from the others by their shapes, which are physical dimensional measurements that characterize the appearance of an object. According to the geometrical differences between male and female pupa, several invariant features about size and shape of the ROI were extracted orderly as depicted in the literature [15].

The descriptors of ellipse about major axis (Ra), minor axis (Rb), and the Ratio (Ra/Rb), and the descriptors of shape about eccentricity, roundness, rectangularity, and complexity were calculated. Additionally, the rested shape features *S1* and *S2*, which used to describe the concave-convex of region, were counted by **Eq.3**.

$$\begin{cases} S1 = Area / S_c \\ S2 = Area / S_r \end{cases} \qquad \text{Eq.3}$$

Where *Area* is the area of the ROI, and S_c is the area of external minimum convex polygon, and S_r is the area of minimum enclosing rectangle.

2.4 Multivariate Data Analysis

The Matlab software was used for data treatment with ChemAC toolbox and libsvm toolbox on portable computer (ASUS, Inter core i2, CPU6670@2.2GHz).

2.4.1 Principle component analysis

Mass original data can be transformed to a few new variables, usually called PC (principal component, PC) by principal component analysis (PCA), which is a data reconstruction and dimensional reduction method. Each PC is the linear combination from the input variables, and as a vector is orthogonal simultaneously to the rest. PCs are arranged in the order of eigenvalues from large to small. Usually, the first few PCs can account for the greatest proportion of the original information. Also, PCA is an unsupervised pattern recognition method which is used for visualizing samples trends in a limited dimensional space. Thus, the plot of some top PCs can be used to handily examine the distance between different classes. Prior to perform PCA, the feature matrix is normalized to the range of 0~1, which can be eliminated the numerical difference of the features (e.g. usually *Area* >1000, while *Rb* <40). In this work, PCA was performed on all variables, and some of the top PCs were employed as the input of classifiers.

2.4.2 Pattern recognition methods

Supervised classifiers of linear discriminant analysis (LDA), K nearest neighbors (KNN), back propagation artificial neural network (BP-ANN) and support vector machine (SVM), were employed in this research in order to discriminate the gender of silkworm cocoon.

According to the principle of supervised pattern recognition, all collected samples were divided into two subsets: calibration set and prediction set. The samples in calibration set used to develop the classifiers, while the samples in prediction set used to test the performance of the developed classifiers. Correct identification rate (CR) is commonly used to appraise the performance of the classifiers on correctly predicting the sample's category of the prediction set. The higher the CR is, the better the classifier has the ability of predicting unknown samples.

Besides, the average prediction time of samples by the classifiers was taken into account. The prediction time can be accounted for the discrimination model's complexity to some extent. With the satisfied prediction, the less the time is consumed, the simpler the classifier is better, which will add the applicability of the model to the practice and can be suitable for fast determination.

3 Results and Discussion

3.1 Samples Statistics

As **table 1** shown, totally about 1071 samples were collected for this experiment. These samples came from two seasons and varied in three hybrid varieties. The gender ratio of all samples was close to 1:1, which was similar to Pan's study [3]. Samples from *'JingSong'*×*'HaoYue'* variety were gathered sub-totally 294 (27.45% of all) in 2009.06 and 603 (56.3% of all) in 2010.06, respectively. Another two varieties, *'Shuangkang'*×*'Yongkang'* 86 samples and *'Suzhen'*×*'Chunguang'* 88 samples, were both brought in 2010.06, and had 8.03 and 8.22 percentage of all repetitively. Diversity of samples in this research can expand the generalization of the developed classifiers. **Table 2** shows the statistical result of morphological features of silkworm. Those features were extracted from X-ray image after a series of image process operations. The *Area* of ROI (i.e. region pixel of pupa in X-ray image) has the biggest coefficient of variable among all descriptors. This illustrated the projective area of pupa was in a broad range of parameter *Area*. Largish distribution differences of other features (i.e. *Perimeter, Ra, Rb, Ra/Rb, S2*) also indicated diversities of morphological features in silkworm pupas.

There was a dataset by matrix sizes of 1071 rows×12 columns (i.e. 11 features and 1 gender in column). Dataset was divided into calibration set and prediction set, respectively. The ratio between the number of calibration set and prediction set was set to 3:1. As a result, totally 802 samples were included in calibration set, and 269 samples were included in prediction set. At the aspect of dividing, each sample was randomly divided into calibration or prediction set according to its arrival batch or variety. Detail of the dividing result is shown in **table 3**.

3.2 PCA performance

One object has many descriptors about its shape, region, contour, etc., and those descriptors have their inner relationships. Consequently, these morphological characterizers of pupa contain overlapped information (i.e. redundant information) for one sample [17]. In this experiment, all features (i.e. 11 variables) extracted from X-ray images were used for the PCA. Albeit PCA is not a supervised classifier tool, its properties could provide morphological information trends as a result of the gender difference of silkworm. In order to visualize the cluster trends of these samples, a 2D score scatter plot was figured with employing the top two PCs (i.e. PC1 and PC2) of the calibration set samples as shown in **Fig.6**. The labels of "Female" and "Male" were attached to samples according to their actual genders. PC1 explained 39.59% variance (i.e. got 39.59% information from 11 raw variables), and PC2 also explained 32.9% variance. A total of top two PCs captured accumulatively 72.49% variability of original information. In this 2D scatter plot, most samples clustered well along the axis plane. It was apparent that there was a separation between the gender clusters which indicated possibly to be separated, as well as a transition region (i.e clusters overlapped) which indicated a slight similarity of morphological features between genders. This overlapped part could be attributed partly to the projection imaging of randomly placing pupas. The goal of this work was to classify the gender of silkworm by means of X-ray imaging approach with the help of the supervised pattern recognition methods. Therefore, it was crucial to choose a suitable pattern recognition method for developing a classifier.

3.3 Results of pattern recognition methods

The geometrical exploration of above 2-D plot by PCA only gives the data cluster trend. Therefore, it is a key step to choose the suitable classification tool. In this work, four linear or nonlinear classification tools (i.e. LDA, KNN, BP-ANN, and SVM) were attempted to develop the discrimination models (i.e. classifiers). These classifiers were optimized by cross validation. The group sets of cross validation came from the five-equal segments in calibration set when developing these discrimination models.

3.3.1 LDA result

LDA is the most commonly used algorithm among the supervised pattern recognition methods. LDA is considered to find the linear combination of inputs and the resulting combination that may separate two or more categories of events or objects. It maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability. It has been successfully utilized in cocoa beans [18], vinegar acetic fermentation [19], and tea varieties [20]. It is crucial that the appropriate number of PCs will enhance the performance of LDA classifier. Frequently, LDA model is optimized to select the appropriate number of PCs by means of cross-validation in developing this discrimination model. The optimal quantity of PCs was specialized according to the highest identification rates by cross-validation. Then the back identification rate was made for calibration and prediction set as **Fig.7** shown according to the consecutive PCs. As shown in **Fig.7**, increasing by the quantity of PCs, the CR for prediction set firstly trended from the low 79.93% to the high 93.31% and then remained unchanged at 92.94%. Obviously, the best LDA classifier was obtained when the top two PCs were utilized, and the correct identification rates were 93.52% and 93.31% in the calibration and prediction sets, respectively. **Table 3** shows the details of LDA classifier with top two PCs for calibration set and prediction set.

3.3.2 KNN result

KNN has been developed to be a powerful method of classification and has proved to be utilized successfully in many applications such as tea grade[21], pork storage time[22], et al.. It is a machine learning technique on the basis of linear supervised pattern recognition. An unknown sample in prediction set is classified depending on the majority of its K nearest neighbors (i.e. samples) in calibration set. Thereby parameter K greatly influences on the identification rate of KNN classifier. Generally the prediction accuracy for a given set of K values are estimated by cross-validation, and the value of parameter K that gives the maximum discrimination accuracy (i.e. lowest discrimination error) is selected by cross-validation. In this study, the quantity of PCs (PCs=2, 3,⋯, 11) and K values (K=1, 3, 5, 7, 9) were optimized simultaneously by cross-validation. According to the quantity of PCs and the K values, the developing KNN model was tested by the prediction set, and **Fig. 8** shows the correct discrimination rates. The same as the number of PCs affected the performance of KNN model, it was found that the value of parameter K influenced the discrimination accuracy as well. The best KNN classifier was achieved when K=3 and PCs=2, and the discrimination accuracy was 93.68% for prediction set.

3.3.3 BP-ANN result

In consideration of which the linear model may fail to provide a complete solution to the classification problem, some non-linear approaches such as artificial neural network (ANN) and support vector machine (SVM) were used in this study. As a classification algorithm ANN is well known, and some literatures have reported its applications in solutions to classification problem of agricultural products [23]. There are many types of ANN, and each type may perform well in a certain instance. Here a type of ANN called back propagation artificial neural network (BP-ANN) was adopted. In the routine of developing BP-ANN classifier, many parameters exerted to some extent to certainly influence on the performance of the conclusive model. These parameters were: the number of neurons in the hidden layer, scale functions, momentum, initial weights and learning rate. In this study, the classical BP-ANN with three layers topology (i.e. Input Layer, Hidden Layer, and Output Layer) was constructed to calibrate a discrimination model [23]. The PCs (n=1,2,⋯,11) were used as the Input Layer of BP-ANN model, and the gender (i.e. 0 = female, and 1 = male) was used as the Output Layer of BP-ANN model. In this work those parameters were optimized by cross validation on the basis of the calibration samples. In developing BP-ANN model, each model was calibrated at least 15 times. Finally, the parameter number of neurons in the Hidden layer was set to 5, and the parameter scale function was set to 'purelin', and the parameter learning rate was set to 0.1. Additionally, the result which was optimized by cross validation showed that some parameters were not the key and had hardly any impacts on the final classifier, such as initial weights and momentum. Besides, it was crucial to optimize the quantity of PCs in calibrating BP-ANN classifier. **Fig. 9** shows the correct identification rates by cross validation on the basis of the calibration samples according to the quantity of PCs. The optimal BP-ANN classifier was achieved when top four PCs (totally explained 94.74% variance) were included. The eventual BP-ANN classifier was structured with 4-5-1 topology. The back identification rate was 94.44% for the calibration set, and the correct identification rate was 93.31% for the prediction set.

3.3.4 SVM result

Another non-linear classification tool, support vector machine (SVM) was employed in this study and attempted to obtain a better result. SVM was firstly proposed by Vapnik and Chervone [24], and further developed by Cortes and Vapnik [25]. It works by creating a hyper-plane between two different categories for classification. The aim of SVM specializes to find a hyper-plane that can be used to distinguish these data points between two categories. If two classes can not be enough separated by a linear boundary in the low dimension space, it is possible to transform this linear boundary to a hyper-plane which should allow linear separation in the higher dimension space. This transformation is implemented by a kernel function [26], and usually three kernel functions are: polynomial, radial basis function (RBF), and sigmoid. As far as their computation, the structure of RBF is the fastest and simplest among those kernel functions. It is often a good choice to select RBF kernel function and resulting good performance of SVM. Therefore, only RBF kernel function was employed to undertake this transformation in this study.

When using RBF kernel function, there are two parameters are introduced, and they are penalty parameter C and kernel parameter γ respectively [27]. In this experiment, to obtain a good performance of SVM classifier, these two parameters were generally optimized by applying "grid search" to search the best combination (C, γ). Basically pairs of (C, γ) were conducted and the resulting one with the best performance of SVM classifier by cross validation was collected. The grid $\gamma=2^{-10}$, 2^{-9}, …, 2^{9}, 2^{10}, respectively, and $C=2^{-10}$, 2^{-9},…, 2^{9}, 2^{10}, respectively, were attempted referring to Chen' description[21]. Finally, this combination could be got from the contour map (**Fig.10a),** which showed a contour map for indicating various performances with different color. The optimal pair (C, γ) was achieved with $\log_2(C)=2$, $\log_2(\gamma)=1.4142$, (i.e. $C=4$, $\gamma=2.665$), which was marked by a blue asterisk in the map. Certainly it was also crucial to optimize the number of PCs for developing SVM model by cross validation in calibration set. **Fig.10b** shows the correct identification rates by cross validation for both calibration and prediction set according to the quantity of PCs. The optimum SVM classifier was achieved when including top three PCs (totally explained 86.19% variance) and the back identification rate was 95.51% in the calibration set, and the predictive discrimination rate was 92.57% in the prediction set.

3.3.5 Discussion of the Results

In order to obtain a good performance in the classification of the gender of silkworm cocoon via X-ray imaging technique, pattern recognition methods and parameters optimization were conducted systematically in this work. **Table 4** shows the discrimination results of LDA, KNN, BP-ANN, and SVM classifiers. As seen from this table, the linear models (i.e. LDA and KNN) achieved a little bit better results, as well as it employed a fewer PCs than the nonlinear classifiers (i.e. BP-ANN and SVM). The correct identification rates of the four classifiers were all higher than 92%, and the result of KNN model was the best performance with 93.68% correct identification rate. At aspect of the averaged running time, LDA model cost the least time 3.3 milliseconds (ms) in predicting per sample, while the best KNN discrimination model needed 21.5 ms to finish the identification process. Considering that the practical utilization placed emphasis on the speed of determination, LDA model would be highlighted for its fast identification speed and the satisfying identification rate from the rests. Also as investigated from this table, we found that the correct identification rate for female samples was slightly higher than that for male samples except for SVM model. This would decrease the error discrimination of male silkworm, and thus benefit for the mulberry silkworm industry because of the male silkworm usually producing higher quality of silk than the female.

Table 3 shows the details about the identification of LDA classifier for calibration set and prediction set. The correct identification rates for calibration or prediction samples collected in 2009.6 both were much lower than that collected in 2010.6, what illustrated that there were morphological differences of pupas between seasons and the season factor affected much the robustness of the LDA classifier. But there was a just slight difference of the correct identification rates between three varieties from a same season. This indicated many similarities between varieties of silkworm, and the developed classifier based on one or more certain varieties could be used to predict the samples from other varieties. This correct identification rate was better than other study [9]. From above results, we could draw the conclusion that X-ray imaging technique coupled with pattern recognition methods had a great potential to identify the gender of silkworm cocoon. The following two aspects gave detailed discussions to this potential.

First of all, as seen from the operating principle of X-ray imaging system, the region of silkworm can be projected by X-ray energy penetration. Due to the low density of cocoon, the region of pupa is highlighted from the cocoon, and can be captured directly without cutting cocoon. With the help of image processing techniques, shape representations can be digitized and extracted effectively. Those representations are generally on behalf of the morphological characterizers of pupa. Thus, the X-ray imaging system has the ability to capture the physical information of silkworm directly without cutting cocoon, which is a great advantage over other approaches [3, 8, 9]. However the representations' differences between the genders in morphological characterizers can not be easily identified by our naked eyes or experiences but with help of multivariate data analysis they can be differentiated effectively.

Secondly, as seen from the principles of multivariate data analysis methods, they act great roles in sexual identification of silkworm cocoon. PCA is used to eliminate the overlapped information and extract mainly useful information (i.e. PCs), and only top two PCs are involved in the final classifier. **Fig.11** shows the loadings of top two PCs against the 11 characteristic variables. A 2D scatter plot of variable-loadings for top two PCs is a good way to detect important variables since they represent the largest variations in the original information. Generally important variables have large absolute loadings value and are far way from origin point. The longer the distance between properties and coordinate origin point, the more the morphological property is important. As investigated from it, most of properties have contributed to the top two PCs, and each contributes much except one parameter *complexity*. However, limited information can be provided by PCA to estimate which property is important in developing a classifier, and this needs further explorations. Supervised pattern recognition methods have a strong capability of self-learning and self-adjusting, and can handle complex relationships between classes. Linear and nonlinear classifiers are employed comparatively to develop a robust classifier by optimizing parameters. Seen from the **table 4**, linear classifiers have included lower number of PCs, and got better performance than that of nonlinear classifiers. This is different to other studies where usually non-linear classifiers got better performances [22]. In some certain case the linear classifiers have a better capability of identifying the genders than nonlinear classifiers, as well as use a fewer number of PCs avoiding bringing noises to the classifiers. Moreover, it is of great importance to balance the model's performance and the running time when turning to the practical application. LDA classifier was considered as a desired discrimination model for its high correct identification rate and the least consuming time for predicting samples. Thus, with the help of LDA discrimination method, that the X-ray imaging system is proposed to identify the gender of silkworm made it possible.

4. Conclusion

The overall results sufficiently demonstrated that the X-ray imaging technique with an appropriate classification tool could be successfully used to discriminate the gender of silkworm cocoon. After x-ray image processing, features extraction and

PCA reconstruction, four different linear and nonlinear classification tools (LDA, KNN, BP-ANN and SVM) were attempted comparatively to develop the discrimination model in this research. Among these discrimination models, LDA model achieved the optimum discrimination 93.31% with the least running time of 3.3ms. Compared with other discriminated methods such as NIRS, computer vision based on visual light, and weighting, it can be concluded that X-ray imaging technique coupled with classification tool shows its superiority in solution to classify the gender of silkworm cocoon in the mulberry silkworm industry.

Acknowledgments

This work has been financially supported by the National Key Technology R&D Program of China (Grant No. 2012BAD29B04-4) and the National Natural and Science Foundation of China (30771243). Thanks to Zhenjiang Sericulture Research Institute for their help in providing silkworm samples and identification of the gender service. We are also grateful to many of our colleagues for stimulating discussion in this field.

Compliance with Ethics Requirements

Jian-rong Cai, Lei-ming Yuan, Bin Liu, and Li Sun declares that there is no conflict of interest. And this article does not contain any studies with human or animal subjects.

Reference

1. L.-M. Yuan, L. Sun, H. Lin, E. Han, H.-L. Liu and J.-R. Cai, *Spectrosc Spect Anal*, 2013, **33**, 2387-2391.
2. B. M. Nicolai, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K. I. Theron and J. Lammertyn, *Postharvest Biol Tec*, 2007, **46**, 99-118.
3. S.-Y. Pan, M. Tao, A.-Q. Sun and T.-M. Jin, *Chinese Acta Entomologica Sinica*, 1996, **39**, 360-365.
4. H. Xiaolong, X. Renyu, C. Guangli, Z. Xing, Z. Yilin, Y. Xiaohua, Z. Yuqing and G. Chengliang, *Molecular biology reports*, 2012, **39**, 1395-1409.
5. N. C. Tansil, Y. Li, C. P. Teng, S. Zhang, K. Y. Win, X. Chen, X. Y. Liu and M. Y. Han, *Advanced Materials*, 2011, **23**, 1463-1466.

6. C. Liu, Z. H. Ren, H. Z. Wang, P. Q. Yang and X. L. Zhang, Bmei 2008: Proceedings of the International Conference on Biomedical Engineering and Informatics, Vol 2, 2008.

7. S. Sumriddetchkajorn and C. Kamtongdee, *Appl Optics*, 2012, **51**, 408-412.

8. S. Sumriddetchkajorn, C. Kamtongdee and C. Sa-Ngiamsak, in *Sensing Technologies for Biomaterial, Food, and Agriculture 2013*, 2013, vol. 8881.

9. C. Kamtongdee, S. Sumriddetchkajorn and C. Sa-ngiamsak, *Comput Electron Agr*, 2013, **95**, 31-37.

10. M.-R. Babaei, H. Barari and K. Kara, *Pakistan Journal of Biological Sciences*, 2009, **12**, 443-446.

11. T. Fujii and T. Shimada, *Seminars in Cell & Developmental Biology*, 2007, **18**, 379-388.

12. N. Kotwaliwale, P. R. Weckler, G. H. Brusewitz, G. A. Kranzler and N. O. Maness, *Postharvest Biol Tec*, 2007, **45**, 372-380.

13. D. S. Narvankar, C. B. Singh, D. S. Jayas and N. D. G. White, *Biosyst Eng*, 2009, **103**, 49-56.

14. M. S. Nielsen, L. B. Christensen and R. Feidenhans'l, *Food Control*, 2014, **39**, 222-226.

15. C.-J. Du and D.-W. Sun, *Trends in Food Science & Technology*, 2004, **15**, 230-249.

16. G.-P. Qu, Q. Ling, L.-Y. Guo, L.-H. Zhao and P.-J. Cheng, *Chinese Nuclear Electronics & Detection Technology*, 2003, **23**, 414-416.

17. Q. Ouyang, J. Zhao and Q. Chen, *Food Res Int*, 2013, **51**, 633-640.

18. E. Teye, X. Huang, H. Dai and Q. Chen, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2013, **114**, 183-189.

19. Q. Chen, A. Liu, J. Zhao, Q. Ouyang, Z. Sun and L. Huang, *Sensors and Actuators B: Chemical*, 2013, **183**, 608-616.

20. Q. Chen, J. Zhao and J. Cai, *T Asabe*, 2008, **51**, 623-628.

21. Q. Chen, J. Zhao, Z. Chen, H. Lin and D.-A. Zhao, *Sensors and Actuators B: Chemical*, 2011, **159**, 294-300.

22. Q. Chen, J. Cai, X. Wan and J. Zhao, *LWT - Food Science and Technology*, 2011, **44**, 2053-2058.

23. C.-J. Du and D.-W. Sun, *Journal of Food Engineering*, 2006, **72**, 39-55.

24. V. N. Vapnik and A. Y. Chervonenkis, *Theory of Probability & Its Applications*, 1971, **16**, 264-280.

25. C. Cortes and V. Vapnik, *Machine learning*, 1995, **20**, 273-297.

26. U. Thissen, M. Pepers, B. Üstün, W. J. Melssen and L. M. C. Buydens, *Chemometr Intell Lab*, 2004, **73**, 169-179.

27. S. S. Keerthi and C.-J. Lin, *Neural Computation*, 2003, **15**, 1667-1689.

FIGURE CAPTIONS

Figure 1 Picture of silkworm cocoon and silkworm pupa

Figure 2 Schematic of X-ray digital imaging system

Figure 3 Workflow of image processing and gender discrimination

Figure 4 The X-ray binary image of silkworm cocoons

Figure 5 The result of Image processes. (a) Original image; (b) Grey enhancement; (c) Gaussian filter; (d) Global threshold segmentation; (e) Opening operation output.

Figure 6 2-D scatter plot of the gender of silkworm cocoon. PC: principal component.

Figure 7 Correct identification rates of LDA model for calibration and prediction set according to different PCs.

Figure 8 Correct identification rate of KNN classifier for prediction set according to different PCs and *K* values

Figure 9 Correct identification rate of BP-ANN model for calibration and prediction set according to different PCs.

Figure10 (a) Contour plot of the optimal parameter pairs (C, γ) by cross validation based on the calibration samples; (b) Correct Identification rates of SVM classifier with different PCs for calibration and prediction set.

Figure 11 The loadings of top two PCs against the 11 morphological variables

Table 1 Statistical result of the sets of silkworm samples.

Experiment Time	Hybrid Varieties	Female Number	Male Number	Total (percentage)
2009.06	*'JingSong'×'HaoYue'*	150	144	294 (27.45%)
	'JingSong'×'HaoYue'	299	304	603 (56.3%)
2010.06	*'Shuangkang'×'Yongkang'*	37	49	86 (8.03%)
	'Suzhen'×'Chunguang'	45	43	88 (8.22%)
Total (percentage)		531 (49.58%)	540 (50.42%)	1071 (100%)

566 **Table 2** Statistical results of the morphological features. SD: standard deviation; CV:
567 coefficient of variable.

Features	Range	Mean	SD	CV(%)
Area	1051~2672	1897.1	243.2	12.82
Perimeter	139~225	186.5	12.205	6.54
Ra	56.7~91.9	76.43	5.013	6.56
Rb	23.4~38.7	31.7	2.324	7.33
Ra/Rb	1.88~2.80	2.42	0.129	5.33
eccentricity	0.85~0.93	0.91	0.011	1.21
roundness	0.61~0.76	0.68	0.023	3.38
rectangularity	0.73~0.78	0.78	0.003	0.38
complexity	16.6~20.6	18.4	0.623	3.39
S1	0.93~0.99	0.97	0.005	0.52
S2	0.51~0.86	0.74	0.057	7.70

568
569

Table 3 Statistical results of LDA classifier for predicting calibration set and prediction set. NFM: the number of actual female silkworms; NM: the number of actual male silkworms; FM: female silkworm; M: male silkworm; CR: correct identification rate.

Dataset	Season	Variety	Discriminated Gender NFM FM	Discriminated Gender NFM M	Discriminated Gender NM FM	Discriminated Gender NM M	CR for FM (%)	CR for M (%)	Total CR (%)
Calibration (NFM= 400; NM=402)	2009-06	*'JingSong'×'HaoYue'*	103	8	14	95	88.03	92.23	90.00
	2010-06	*'JingSong'×'HaoYue'*	206	19	8	219	96.26	92.02	94.03
		'Shuangkang'×'Yongkang'	30	0	1	33	96.77	100.0	98.44
		'Suzhen'×'Chunguang'	34	0	2	30	94.44	100.0	96.97
	Count		373	27	25	377	93.72	92.32	93.52
Prediction (NFM= 131; NM=138)	2009-06	*'JingSong'×'HaoYue'*	35	4	6	29	85.37	87.88	86.49
	2010-06	*'JingSong'×'HaoYue'*	69	5	2	75	97.18	93.75	95.36
		'Shuangkang'×'Yongkang'	6	1	0	15	100.0	93.75	95.45
		'Suzhen'×'Chunguang'	11	0	0	11	100.0	100.0	100.0
	Count		121	10	8	130	93.80	92.86	93.31

Table 4 Statistical results of classifiers for predicting the gender of silkworm cocoon. LDA: linear discriminant analysis; KNN: *K* nearest neighbors, BP-ANN; back propagation artificial neural network; SVM: support vector machine; NFM: the number of actual female silkworms; NM: the number of actual male silkworms; FM: female silkworm; M: male silkworm. CR: correct identification rate.

Classifiers	Number of PCs	Discriminated Gender				CR for FM (%)	CR for M (%)	Total CR (%)	Running time (ms)
		NFM (131)		NM (138)					
		FM	M	FM	M				
LDA	2	121	10	8	130	93.80	92.86	93.31	3.3
KNN	2	120	11	6	132	95.24	92.31	93.68	21.5
BP-ANN	4	120	11	7	131	94.45	92.25	93.31	9.3
SVM	3	121	10	10	128	92.37	92.75	92.57	14.2

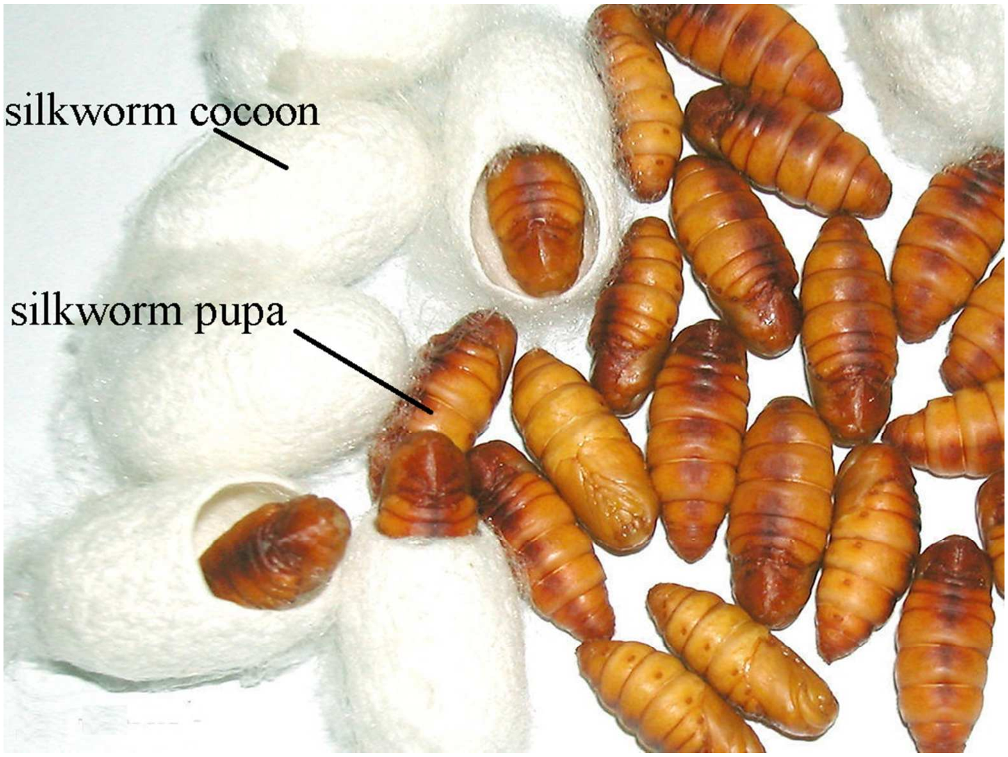


Figure1 Picture of silkworm cocoon and silkworm pupa
361x270mm (72 x 72 DPI)

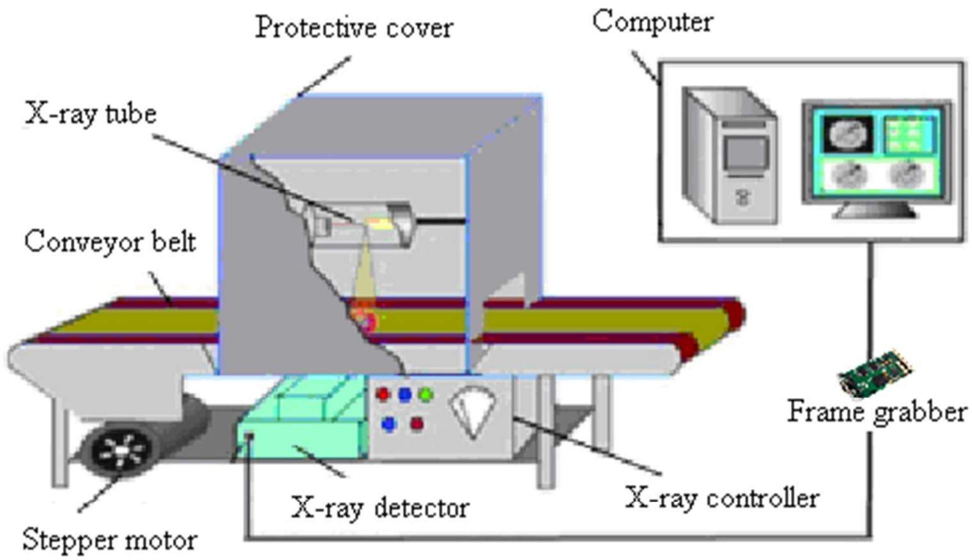


Figure2 Schematic of X-ray digital imaging system
132x82mm (96 x 96 DPI)

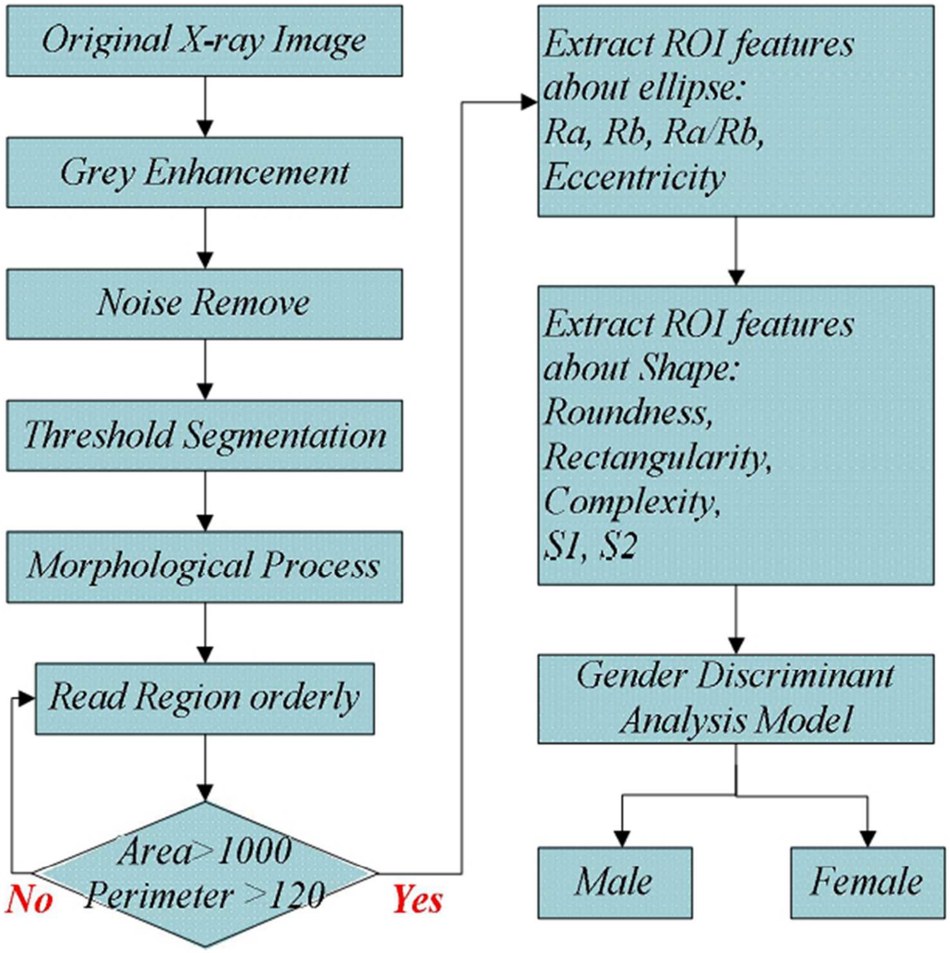


Figure3 Workflow of image processing and gender discrimination
170x167mm (96 x 96 DPI)

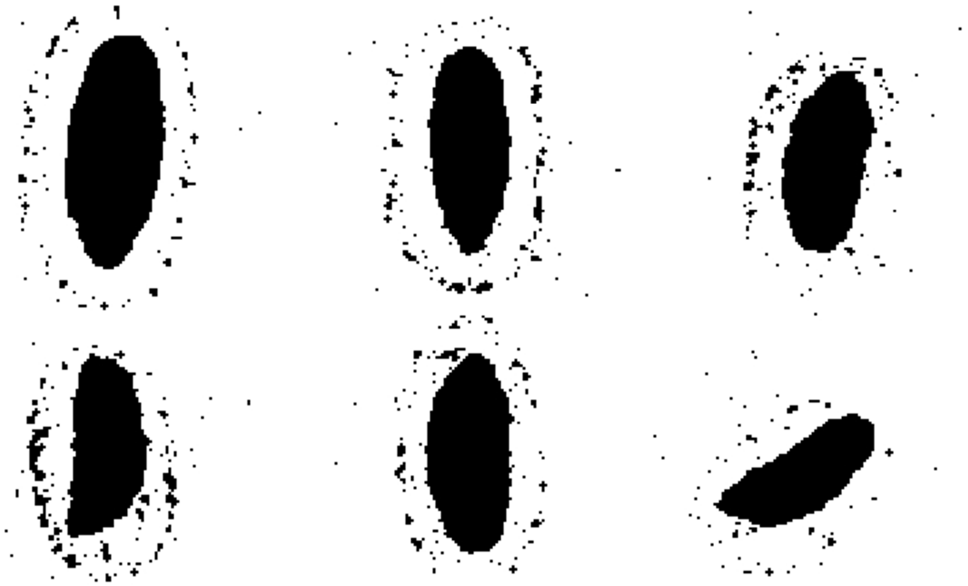


Figure4 The X-ray binary image of silkworm cocoons
62x39mm (300 x 300 DPI)

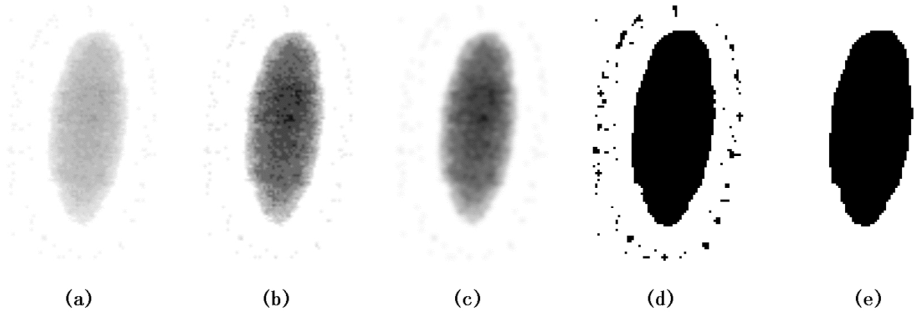


Figure5 The result of Image processes. (a) Original image; (b) Grey enhancement; (c) Gaussion filter; (d) Global threshold segmentation; (e) Opening operation output.
148x46mm (300 x 300 DPI)

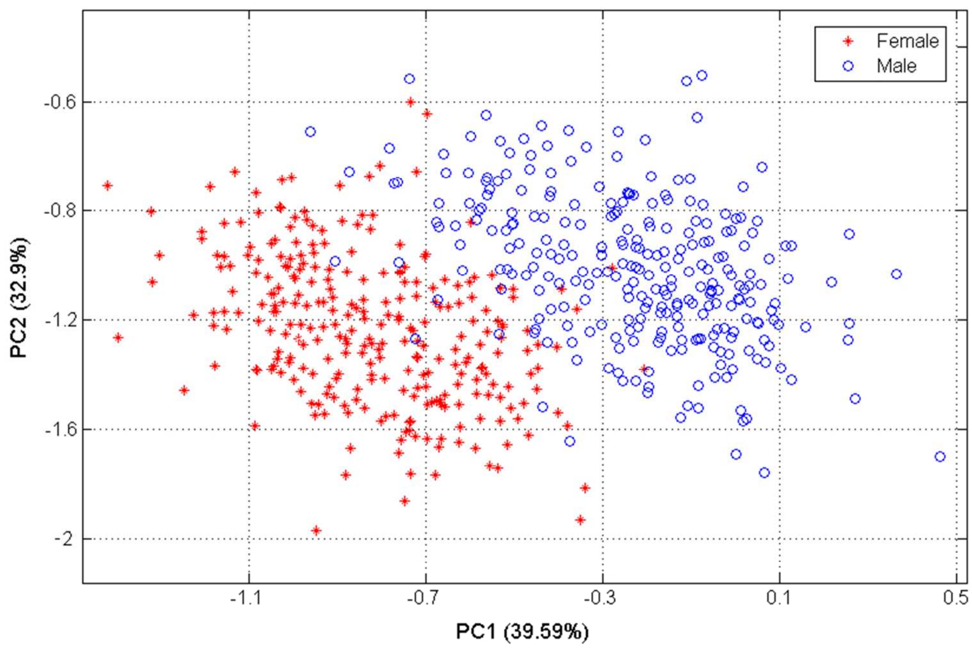


Figure6 2-D scatter plot of the gender of silkworm cocoon. PC: principal component.
224x150mm (96 x 96 DPI)

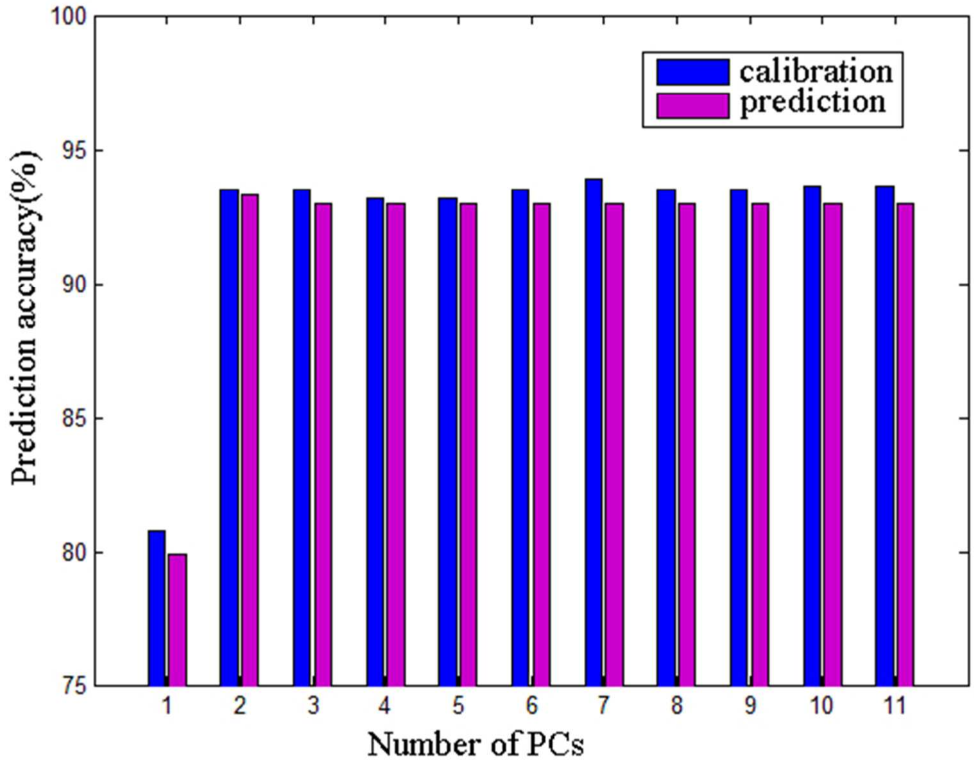


Figure7 Correct identification rates of LDA model for calibration and prediction set according to different PCs.
147x114mm (96 x 96 DPI)

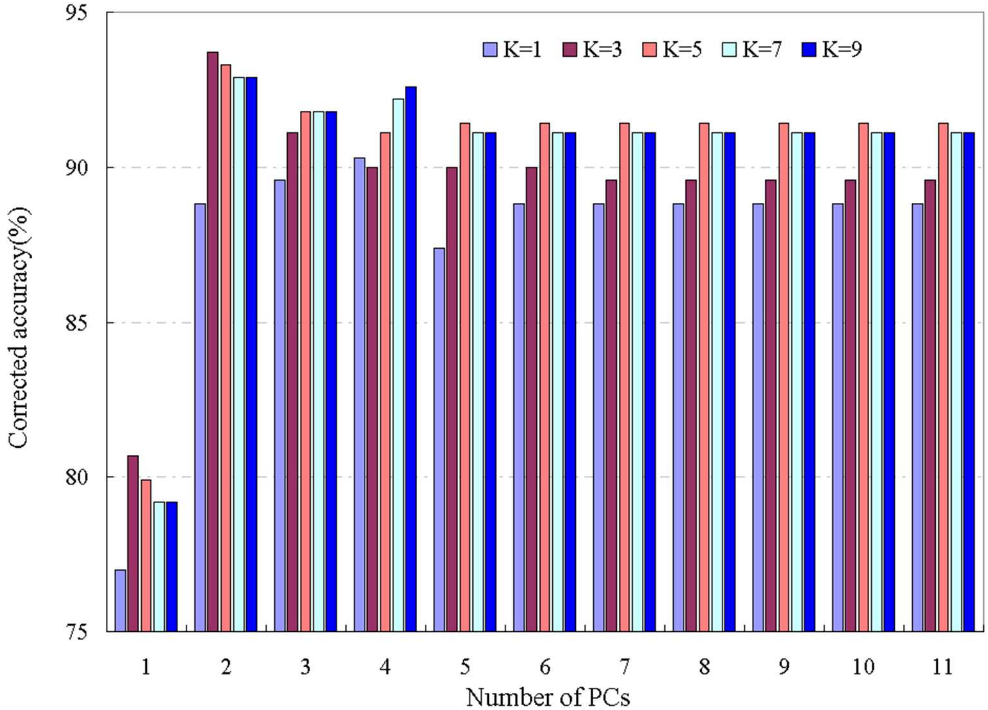


Figure8 Correct identification rate of KNN classifier for prediction set according to different PCs and K values
251x186mm (96 x 96 DPI)

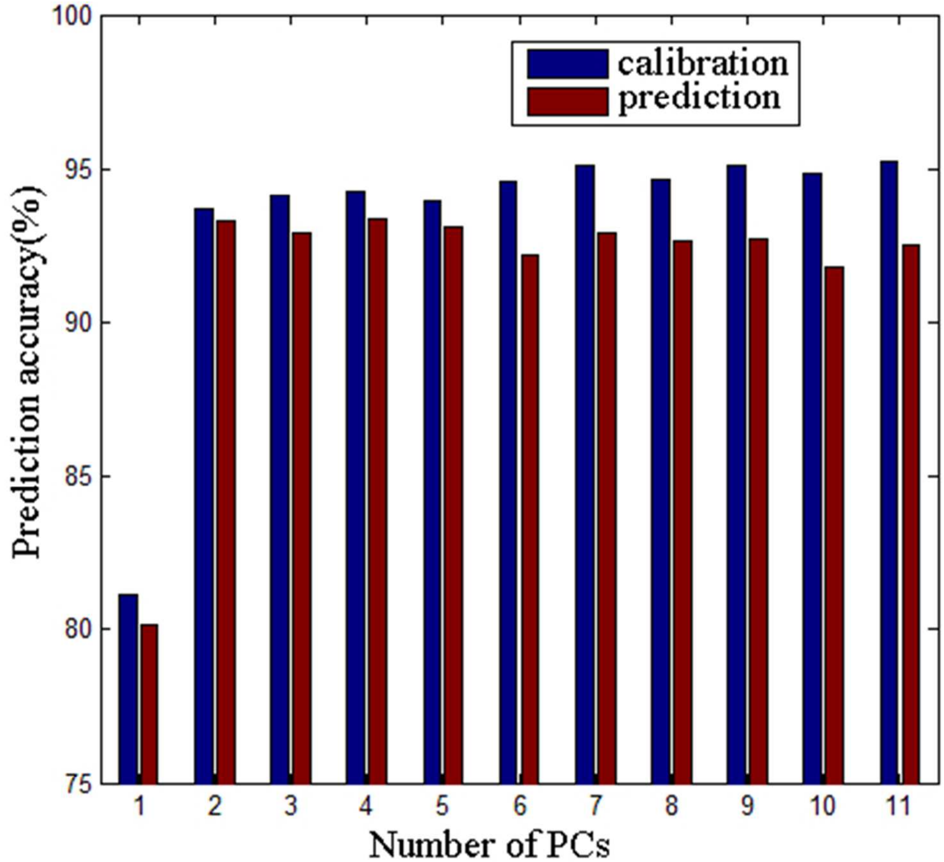


Figure9 Correct identification rate of BP-ANN model for calibration and prediction set according to different PCs.
132x115mm (96 x 96 DPI)

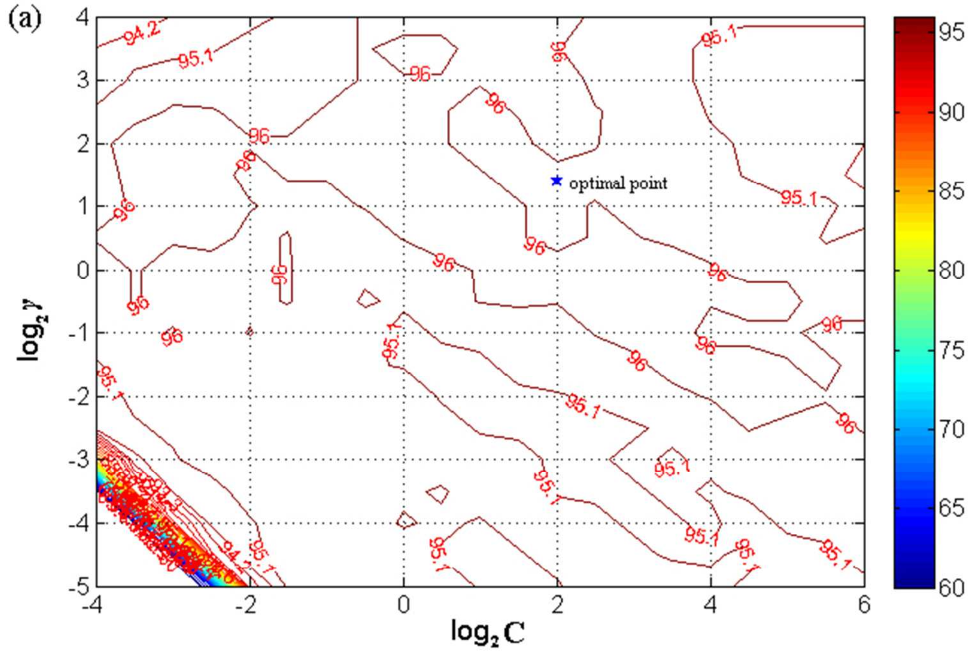


Figure10 (a) Contour plot of the optimal parameter pairs (C, γ) by cross validation based on the calibration samples; (b) Correct Identification rates of SVM classifier with different PCs for calibration and prediction set.
177x119mm (96 x 96 DPI)

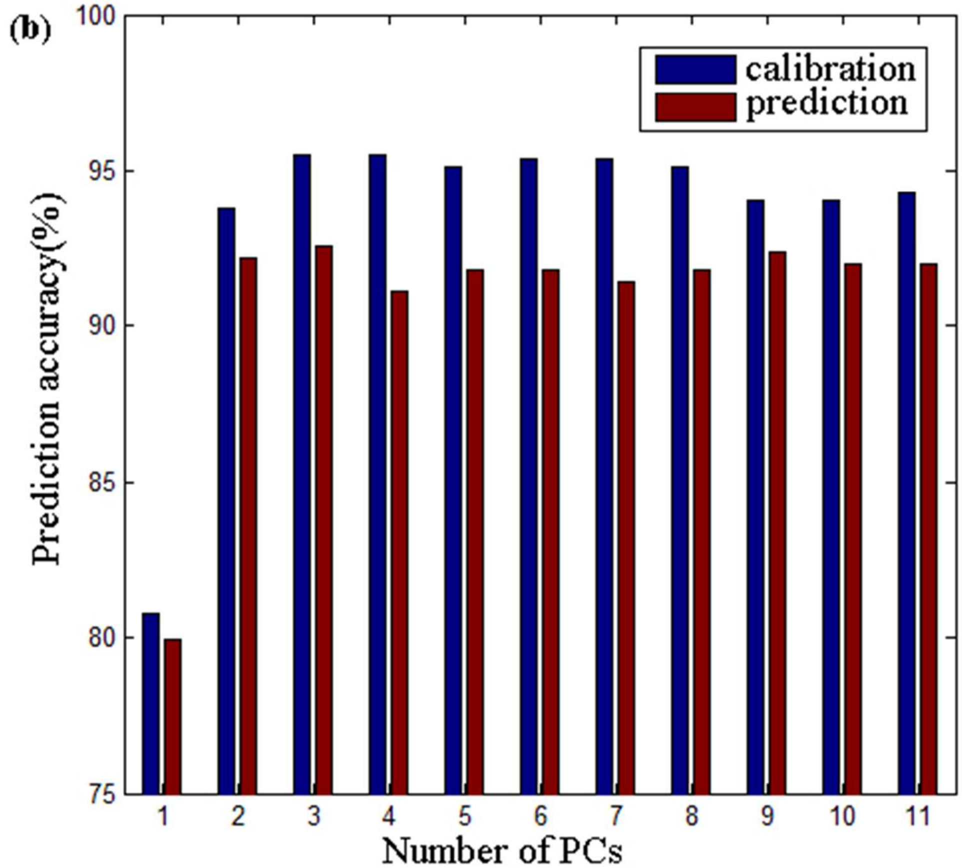


Figure10 (a) Contour plot of the optimal parameter pairs (C, γ) by cross validation based on the calibration samples; (b) Correct Identification rates of SVM classifier with different PCs for calibration and prediction set.
133x117mm (96 x 96 DPI)

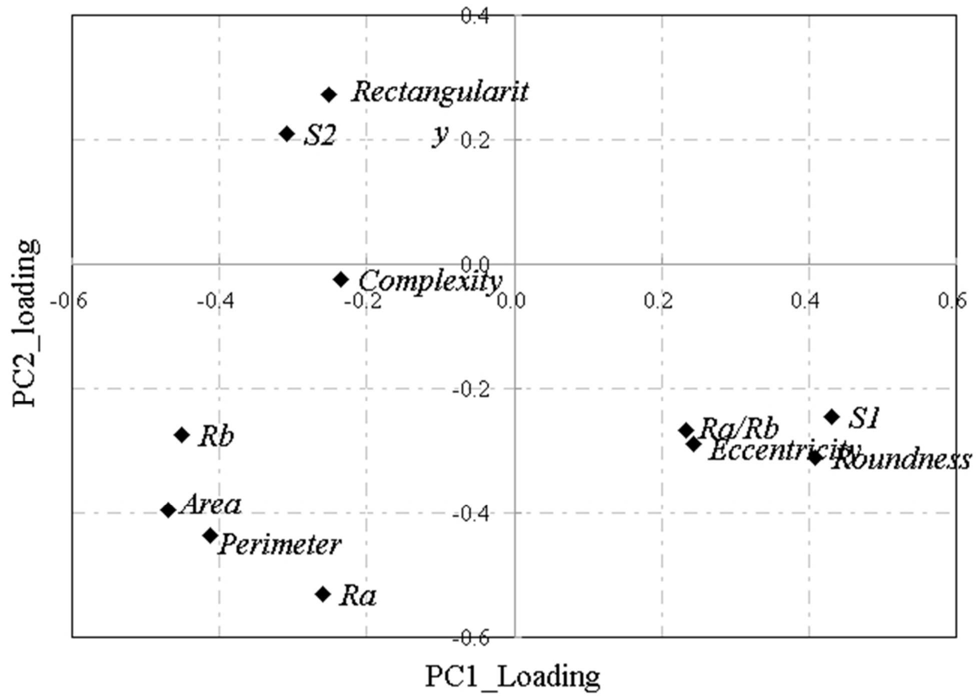


Figure 11 The loadings of top two PCs against the 11 morphological variables
176x127mm (96 x 96 DPI)