Rapid Identification of Rice Varieties by Grain Shape and Yield-Related Features Combined with Multi-class SVM

Chenglong Huang[1,2], Lingbo Liu[3], Wanneng Yang[1,2,4], Lizhong Xiong[4], and Lingfeng Duan[1,2(✉)]

[1] College of Engineering, Huazhong Agricultural University, Wuhan 430070, People's Republic of China
hcl@mail.hzau.edu.cn, ywn@mail.hzau.edu.cn, duanlingfeng@mail.hzau.edu.cn

[2] Agricultural Bioinformatics Key Laboratory of Hubei Province, Huazhong Agricultural University, Wuhan 430070, People's Republic of China

[3] Britton Chance Center for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics-Huazhong University of Science and Technology, 1037 Luoyu Rd., Wuhan 430074, People's Republic of China
firbo007@gmail.com

[4] National Key Laboratory of Crop Genetic Improvement and National Center of Plant Gene Research, Huazhong Agricultural University, Wuhan 430070, People's Republic of China
lizhongx@mail.hzau.edu.cn

Abstract. Rice is the major food of approximately half world population and thousands of rice varieties are planted in the world. The identification of rice varieties is of great significance, especially to the breeders. In this study, a feasible method for rapid identification of rice varieties was developed. For each rice variety, rice grains per plant were imaged and analyzed to acquire grain shape features and a weighing device was used to obtain the yield-related parameters. Then, a Support Vector Machine (SVM) classifier was employed to discriminate the rice varieties by these features. The average accuracy for the grain traits extraction is 98.41 %, and the average accuracy for the SVM classifier is 79.74 % by using cross validation. The results demonstrated that this method could yield an accurate identification of rice varieties and could be integrated into new knowledge in developing computer vision systems used in automated rice-evaluated system.

Keywords: Computer vision · Rice varieties identification · Grain shape · Rice yield · Multi-class SVM

1 Introduction

Rice is one of the most significant cereals in the world, especially for china (Zhu et al. 2011). Thousands of rice varieties could be produced daily by modern breeding technique (Bagge and Lubberstedt, 2008). And a large number of rice

D. Li and Z. Li (Eds.): CCTA 2015, Part I, IFIP AICT 478, pp. 390–398, 2016.
DOI: 10.1007/978-3-319-48357-3_38

germplasm-resources need to be exploited by breeders for the rice improvement (Xing and Zhang, 2010). However, the characterization for the various rice varieties are technically challenging due to the slight difference (Tanabata et al. 2012). Rice variety is also regarded as one of the most important factors related to cooking and processing quality, which was resulted by the variations in size, shape, and constitution (Zhang, 2007). Therefore rice variety identification is of great significance.

Since the identification of rice varieties is so important for rice-related research. A lot of work had been reported about it. Namaporn Attaviroj tried to identify the rough and pure rice varieties using fourier-transform NIR (Attaviroj et al., 2011). Liu Hongyun had tried to indentify rice varieties by tolerance and sensitivity to copper (Liu et al. 2007). Liu Feng tried to identify rice vinegar variety using visible and near infrared spectroscopy (Liu et al. 2011). However, the above study only focused on a few of special rice varieties, and the identification for the massive ordinary rice varieties were still an urgent problem.

Machine vision was a practical technology and had recently been widely applied in the agriculture. Dual-camera rice panicle length measuring system was proposed by Dr. Huang (Huang et al. 2013). A machine-vision-facility was developed for rice traits evaluation (Duan et al. 2011). A hyperspectral imaging system was designed for biomass prediction (Feng et al. 2013). Yang et al. applied x-ray computed tomography for rice tiller measurement (Yang et al. 2011). Duan et al. had counted filled/unfilled spikelets using Bi-modal imaging (Duan et al. 2011). Support vector machine (SVM), first proposed in 1995 by Cortes and Vapkin, has a lot of advantages, such as nonlinear, small-sample, and high dimensional pattern recognition and can be easily extended to other machine learning problems. However, since it is originally used for binary classification (Cortes and Vapnik 1995; Vapnik, 1999), it requires extra algorithm support to meet practical needs.

This research aimed to propose a feasible method for rapid identification of rice varieties. In this study, the features of grain shape and yield-related traits were extracted by image analysis. And the specific Muti-SVM classifier was developed to discriminate the rice variety.

2 Materials and Methods

The Rice varieties used in this study are selected from the Chinese core-germplasm resources. 79 rice varieties were tested and each variety had four samples. Three quarter of the rice samples were taken as training set, meanwhile the other were testing set. The rice grains were threshed from the panicles manually. And the filled spikelets were selected out by wind separator.

The technical method for rice variety identification is described as Fig. 1. Firstly, the rice grain were imaged and analyzed for shape and weight parameters. Then features of the training set were applied to build the SVM model. With the SVM model, the testing set was applied to evaluate the rice variety identification accuracy.

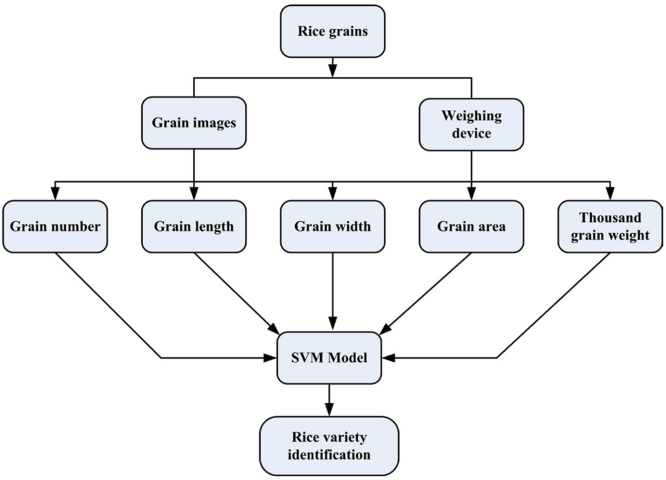

Fig. 1. The technical method for rice variety identification

2.1 Rice Feature Extraction

The features of each rice sample were obtained as shown in Fig. 2. The rice grains per sample were spread on the scanner manually (Fig. 2a). And the image was acquired and transferred to the computer. Then grain image was analyzed for grain number (GN), grain width (GW), grain length (GL), grain area (GA) (Figs. 2b and c). And the grain weight was obtained by the electro-weighing device (Fig. 2d).

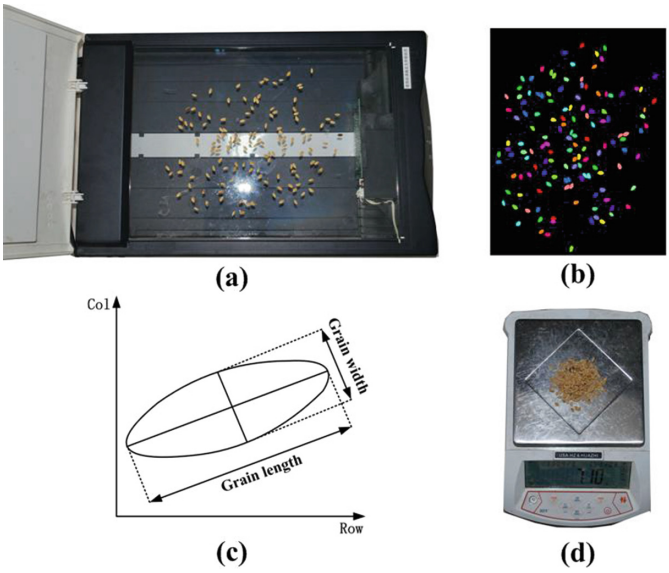

Fig. 2. Rice grain features acquisition

2.2 Multi-SVM Classifier

Since we focus on the classification of rice varieties, we should know the biological classification criteria of rice. Generally, different type of rice will have different genotypes, which usually leads to different phenotypes. Obviously, one of the important problems is how to divide the training rice samples into different subsets by binary tree based SVM-BTA algorithms. When it comes to plants classification, we wish the two subsets will have at least one totally different genotype. Knowing the fact that disparate varieties all have different genotypes, one possible way is using the K-Means clustering method, which will resign each data to the nearest cluster repeatedly, just like combining analogous genes. The problem is to determine the evaluation function.

In order to reduce the algorithm running time of the partitioning process, it's necessary to improve the KMeans clustering. Since we always want to divide the input set into two clusters, a pretreatment of the data using average threshold algorithm will work. In the next section, we introduce one partition function for evaluation, and then propose MBT-SVM based on the K-Means clustering with optimal partition function.

2.2.1 Partition Function

Suppose the problem's center $c_{problem}$ as the all data mean value in the i.th column of the input of a non-leaf node. The following partition function can be adopted to split the node (Huang et al. 2013):

$$PF(I_1 \cup I_2) = \sum_{j=1}^{2} \sum_{i=1}^{l} \frac{d(c_i, c_{problem})}{\sigma_i} \tag{1}$$

Where $d(c_i, c_{problem})$ is the Euclidean distance and σ_i is the variance of column i.

The larger PF is, the better it works. So we have also determined our termination criterion, just to traversal all the possible combinations to find the largest PF, or to reassign the object one by one until the value PF won't become larger in a whole round. An initial partition is needed to reduce the algorithm running time. The judging function using average threshold algorithm was described as followed:

$$J(x) = \begin{cases} 1, & \sum_{i=1}^{v} x_i \geq \frac{1}{l} \sum_{j=1}^{l} \sum_{i=1}^{v} x_{ij} \\ -1, & else \end{cases} \tag{2}$$

Where v is the number of features. And x_{li} is the i.th feature of the j.th sample in the training set.

The judging function compares the average value of all the features of a sample with the average value obtained by the whole training set. Since different features of the rice will have different magnitude, we have to standardize the input data. The standardization includes data integrity check, linear unification and repacking.

2.3 Kernel Function

The efficiency and accuracy of SVM is determined by the kernel type and parameters, as well as the parameter c. To determine the best type of kernel function, we can try the three basic kernel functions and pick the one with the best accuracy according to cross-valid. In general, the Gaussian kernel with a single parameter γ is a good choice. The c and γ is usually calculated by a grid searching method, in which the cross validation is applied, then we will pick the one with the highest accuracy, such as $c = \{2^{-5}, 2^{-4}, 2^{-3}, \ldots 2^{8}\}$; $\gamma = \{2^{-10}, 2^{-9}, 2^{-8}, \ldots 2^{3}\}$. The final model, will then training set was applied by the chosen type of kernel function and with the optimized parameters Duan et al. 2011. As is shown in Fig. 3, an inappropriate combination of kernel type and parameters will cause under-fitting or over-fitting problems.

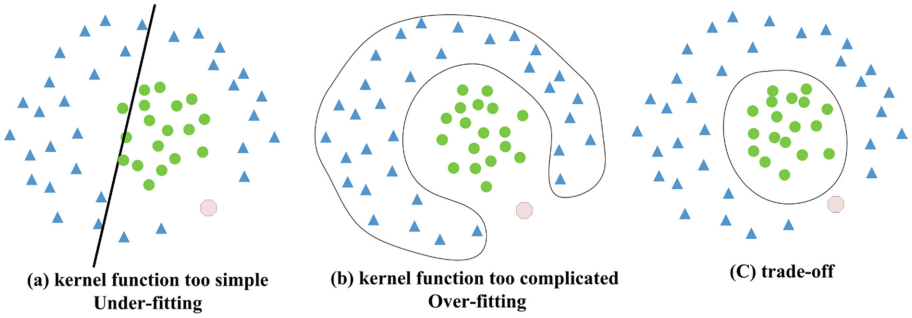

Fig. 3. Classification models generated by different kernel type and parameters

3 Results and Discussion

3.1 Grain Traits Extraction Accuracy

Totally, 79 copys of rice grains were measured automatically and manually, and the parameters of the GN, GL, GW, GA, grain weight were all obtained for each copy. The measurement accuracy for each traits were shown in Fig. 4, the MAPE were calculated according tho the Eq. 3.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_{ai} - x_{mi}|}{x_{mi}} \tag{3}$$

The measurement results showed that the average MAPE for grain number measurements was 1.33 %; the average MAPE for grain length measurement was 1.25 %; the average MAPE for grain width measurement was 2.20 %. The results demonstrated that the automatical measurements performed a good relationship with manual measurements.

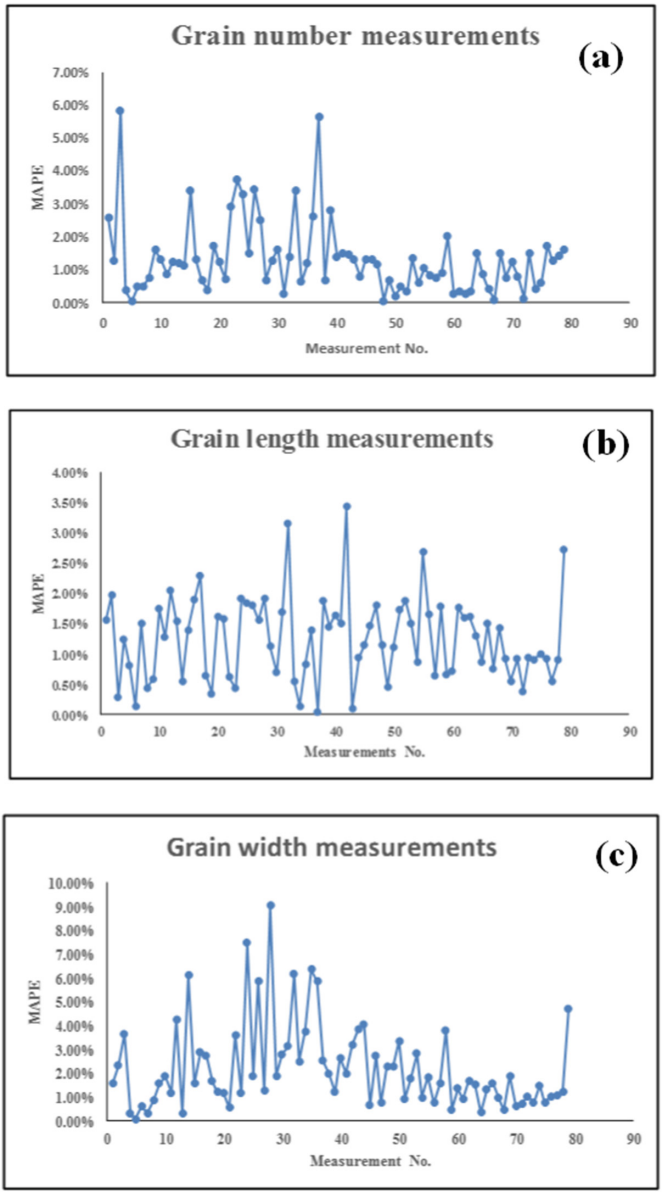

Fig. 4. System measurements accuracy evaluation. (a) Grain number, (b) Grain length, (c) Grain width

3.2 Rice Varieties Identification

The rice feature measurements were shown above. From the results, it was proved that the difference between outer-varieties and inner-varieties were both slight. In the SVM identification, for each tree node, an RBF-kernel SVM was first adopted after partitioning the input data by a clustering method and the linear function was the last one to try. A grid search based on cross-validation was employed for parameter optimization. The result for the whole experiment is shown in Table 1. The training time includes partitioning, tree construction and SVM training. It represented the CPU running time (milli second), for a quad core computer, the real-time approximately equals to a quarter of the CPU time.

The standard data from UCI and the rice sample data were all tested and the classifying results were shown in Table 1. It was proved that the algorithm had high classification accuracy when processing these standard data. Since the testing samples are randomly selected, every kind of data set was tested only three times and the worst result was recorded in order to avoid anthropogenic interference (like continually running the algorithm until it gets a good output). From the data associated to rice in Table 1, it was seen that building a classification binary tree for a set of data with many classes needed a relatively long time.

Rice-s79-1 is a basic pre-experiment focusing on verification of the algorithm compatibility. There was not a linear kernel function and the number of attributions was fixed to 13. Rice-s79-2 was a grid-search experiment aiming at finding a suitable number of attributions for the next experiment. Since the algorithm needed to traverse all the possible combinations of the attribution, and there were four kernel functions to be examined for each combination, it's naturally to have a very large training time (nearly ten hours of real-time). After tracking the misclassified samples, we find that about half of the errors occurred in the tree nodes with a relatively high cross-valid accuracy other than the low ones. Clearly, the model is over-fitted. It is necessary to improve the optimization function. We needed to pick the SVM classifier with an appropriate cross-valid accuracy instead of the ones with the highest. Rice-s79-3 was the result of the formal experiment. As was mentioned above, we repeat the experiment three times and record the result with the worst accuracy rate. And the average accuracy was about 79.74 %.

Table 1. The class identification results by Muti-SVM

Data	Class	Dimensions	Size	Training set	Testing set	Accuracy	Training time (CPU time ms)
iris	3	4	150	120	30	100 %	0.875
wine	3	13	178	133	45	97.78 %	28.03
seeds	3	7	210	100	110	96.67 %	161.03
vehicle	4	18	846	746	100	86 %	213.07
glass	6	9	214	166	48	72.91 %	321.04
rice s79-1	79	13	316	237	79	64.56 %	1226.6
rice s79-2	79	13	316	237	79	25.32 %	124457
rice s79-3	79	13	316	237	79	79.74 %	55108

4 Conclusions

In this paper, a support vector machine working in the multi-space-mapped mode (MBT) was proposed for a rice multi-class classification task. The result showed that this study performed high accuracy for the grain traits extraction and also proved a good performance for the rice varieties classification. In future work, we will further analyze the data of tree nodes from experiments to develop more effective algorithms. The range of the parameters will change according to the size of the input training set which will greatly reduce the computation time, which will reduce the time complexity of the algorithm. Therefore we could apply this method for larger rice sample sets for varieties recognition. The results also demonstrated that this method would provide new knowledge for automated rice-vision-evaluated system.

Acknowledgment. This work was supported by the Fundamental Research Funds for the Central Universities (2662014BQ036, 2662015QC006, and 2662015QC016), the Natural Science Foundation of Hubei Province (2015CFB529), the National High Technology Research and Development Program of China (2013AA102403), the National Natural Science Foundation of China (30921091, 31200274).

References

Bagge, M., Lubberstedt, T.: Functional markers in wheat: technical and economic aspects. Mol. Breed. **22**(3), 319–328 (2008)

Zhu, J., Zhou, Y., Liu, Y., Wang, Z., Tang, Z., Yi, C., Tang, S., Gu, M., Liang, G.: Fine mapping of a major QTL controlling panicle number in rice. Mol. Breed. **27**, 171–180 (2011)

Xing, Y., Zhang, Q.: Genetic and molecular bases of rice yield. Annu. Rev. Plant Biol. **61**, 11.1–11.22 (2010)

Tanabata, T., Shibaya, T., Hori, K., Ebana, K., Yano, M.: SmartGrain: high-throughput phenotyping software for measuring seed shape through image analysis. Plant Physiol. **160**, 1871–1880 (2012)

Liu, H., Zhang, H., Wang, G., Shen, Z.: Identification of rice varieties with high tolerance or sensitive to copper. J. Plant Nutr. **31**(1), 121–136 (2007)

Liu, F., Yusuf, B., Zhong, J., et al.: Variety identification of rice vinegars using visible and near infrared spectroscopy and multivariate calibrations. Int. J. Food Prop. **14**(6), 1264–1276 (2011)

Zhang, Q.-F.: Strategies for developing green super rice. PNAS **104**(42), 16402–16409 (2007)

Attaviroj, N., Kasemsumran, S., Noomhorm, A.: Rapid variety identification of pure rough rice by fourier-transform near-infrared spectroscopy. Cereal Chem. **88**(5), 490–496 (2011)

Huang, C., et al.: Rice panicle length measuring system based on dual-camera imaging. Comput. Electron. Agric. **98**, 158–165 (2013)

Duan, L., et al.: A novel machine-vision-based facility for the automatic evaluation of yield-related traits in rice. Plant Meth. **7**, 44 (2011a)

Feng, H., et al.: A hyperspectral imaging system for an accurate prediction of the above-ground biomass of individual rice plants. Rev. Sci. Instrum. **84**, 095107 (2013)

Duan, L., et al.: Fast discrimination and counting of filled/unfilled rice spikelets based on bio-modal imaging. Comput. Electron. Agric. **75**, 196–203 (2011b)

Yang, W., Xu, X., Duan, L., Luo, Q., Chen, S., Zeng, S., Liu, Q.: High-throughput measurement of rice tillers using a conveyor equipped with x-ray computed tomography. Rev. Sci. Instrum. **82**, 025102–1–025102-7 (2011)

Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)

Vapnik, V.: An overview of statistical learning theory. IEEE Trans. Neural Netw. **10**(5), 988–999 (1999)

Kumar, M., Gopal, M.: Reduced one-against-all method for multiclass SVM classification. Exp. Syst. Appl. **38**, 14238–14248 (2011)

Hsu, C., Chang, C., Lin, C.: A Practical Guide to Support Vector Classification. Department of Computer Science and Information Engineering. National Taiwan University (2003)