Effect of pre-processing on plant image classification

Aishwarya Sawant
Department of Computer Science
University of New South Wales
Sydney, Australia
aishwaryasawant9527@gmail.com

***Abstract*—In the process of plant classification, the challenge lies in the suppression of noise associated with the background and lighting. This paper performs a comparative study of the advantages of pre-processing on the overall performance of the classifiers. The local features for classification are extracted using SIFT, clustered using kMeans and are finally fed to the three classifiers: SVM, Random Forest and KNN. Results evaluated using the classification metrics show a significant improvement with the use of pre-processed images.**

Keywords—SVM, SIFT, KNN, Random Forest, Image pre-processing

I. Introduction and Background

The term plant phenotyping refers to quantitative description of plant's anatomical, physiological, and biochemical properties. This information is obtained on large scale by acquiring images of many plants simultaneously. Previously identifying and evaluating phenotypes using these characteristics was done manually but the combination of image-based and automated machine learning approaches has paved a path for more sophisticated and accurate methods for plant classification [1]. This finds its application in the agricultural industry for decision support to the breeders, to select best genotypes to suit their requirements. The quality of plant phenotyping images is important, therefore, for the purpose of this experiment we use the IPPN dataset (online at https://www.plant-phenotyping.org/datasets-download). It has a collection of raw and annotated top-view color images of Tobacco and Arabidopsis plant. The images have varied orientation, resolution, and background, which can help the classifier learn from a diverse range of data.

To obtain better classification results, there has been tremendous work towards segmenting plants from its complex background. Those with strict restrictions on the scene rely either on calibrated thresholds or on histogram-based methods such as Otsu's thresholding to identify data driven thresholds [2][3]. Shape, colour and texture features are common features involved in several applications, such as in Hiremath and Pujari [4]. These structures make the leaves of the plant determinant as they grow into specific patterns, rather than roots or stems [5]. The bag-of-visual-words method has also been used as a feature representation tool in image analysis with the idea of getting the frequency of each image feature [6][7]. All these methods can be used at different levels of image classification to perform a more effective plant classification.

II. Method

The proposed method for image classification of Tobacco and Arabidopsis plant includes three basic steps that is pre-processing, local feature extraction and a machine learning algorithm for classification. The details of each of these steps are discussed below.

A. Pre-processing

For this experiment we take 165 images of the Arabidopsis plant under the 'Ara2013-Canon' folder in the 'Plant' section of the dataset, and 62 images of the Tobacco plant under the 'Tobacco' folder. These RGB images have a complex background especially the tobacco plant. A close look at the images of the Arabidopsis plant tell us that the images are zoomed enough to avoid the inclusion of the undesired background. But this is not true for the Tobacco plant wherein the tray edges and other irrelevant background objects are captured. A pre-processing technique to segment the plant from its background will aid the next step of local feature extraction. To cater to different images, we use Otsu's thresholding to automatically choose a mean value which separates the background from the foreground [8]. Although this method works well for the Arabidopsis plant, but the Tobacco plant still has some background information which can hinder the performance of the classifier. An alternate approach is to perform colour thresholding based on the amount of saturation and luminance since our area of interest is the plant which is predominantly in the shades of green. For this we convert the RGB image into HSV and LAB colour space and obtain the saturation and luminance channel to perform a thresholding on them. The output image is then filtered for salt and pepper noise using median filter. A kernel size of 7 best suits the dataset to get rid of smaller bits from the pre-processing step. At the end both the images are used as a mask on the original image to retrieve just the plant area.

The processed images are stored in the respective folders but with a 70:30 split for training and test images. To perform the comparative study of the effect of this pre-processing, we split the original RGB images into training and test set as well.

B. Shape Feature Extraction

Image features are set of points in an image that are distinctively identified even with change in the rotation or illumination of the images. One of these feature points can be the shape features which are extracted using Scale-Invariant Feature transform (SIFT). This study uses the region-based shape feature extraction for plant classification.

The algorithm will detect key points and descriptors of each pre-processed RGB image, to obtain a collection of all the descriptors for all the training and test images. The training features obtained from SIFT are then clustered using kMeans clustering algorithm to generate a bag of visual words with key as the center of each cluster. This bag of visual words is then used to predict the class of test images based on the test features.

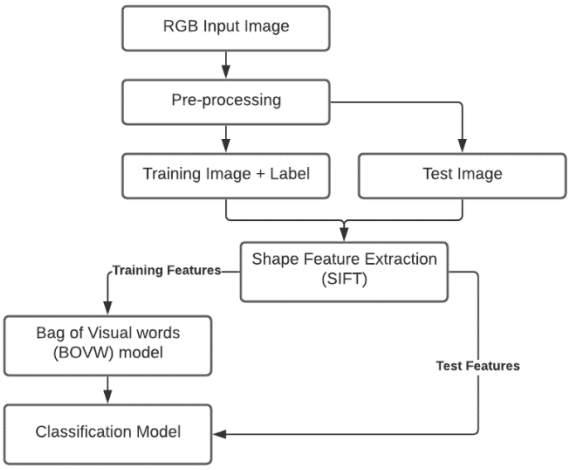

Fig. 1. Flow of the proposed method for plant image classification

C. Classification Model

Three different models are used to predict the class of the test images. If the model predicts the class as 0 it is the Arabidopsis plant and tobacco if class 1 is predicted. The local features thus obtained are scaled using fit and transform by computing the mean and standard deviation.

a) Support Vector Machine: The objective of this algorithm is to find a hyperplane which separates the data-points. The dimension of the hyperplane depends on the number of features chosen for the classification. In our case we have selected 1 shape feature, hence it would be a 1-dimensional space. SVM is an efficient algorithm for classification as it produces high accuracy with less computation power [13].

b) K-Nearest Neighbours: This algorithm captures the idea of similarity with the help of mathematical calculation of distance between the two data points. The correct value of K (nearest neighbours) plays an important role in determining the performace of the classification. K is chosen to be 5 in our case [12].

c) Random Forest: This algorithm is an ensemble of many decision trees, each of which predict a class for the test data, and the class with the most votes becomes our model's prediction [11]. The chance of making the correct prediction increases the number of estomator trees in our model. For the plant classification it is kept as 100.

III. Experiment

A. Experimental Setup

The pre-processing was done on all the 227 images of the two plants. The idea was to extract the plant from its complex background to get maximum out of the feature extraction method. For this experiment we converted the rgb image into gray-HSV image using an external library called 'plantcv'. As this color space conversion was not available in 'OpenCV' and was the turning point in obtaining the image in Fig. 2, plantcv library was used [15].

Initially, three features were selected, shape feature (hu-moments and SIFT) and colour feature (color histogram) for the pre-processed images [10]. These three features were concatenated together to get a global feature vector which could be used for classification. However, the computational time for this method was too high which could have affected the prediction algorithm. Alternatively, only the SIFT feature was used for the classification of the two plant images.

The ratio of split of train and test images was set to 70:30 after cross validation using a split of 60:40 and 80:20. An additional layer of pre-processing was added for the tobacco image to get a better foreground image, but this showed no significant improvement in the overall precision. This additional step included getting rid of the smaller pixel areas around the tobacco plant as shown in Fig, 2. Before the prediction of the test images we need to fit and transform the global features. This is done using the 'StandardScaler' function, which transforms data such that its distribution will have a mean value 0 and standard deviation of 1. This scaling would assure the normalization of the data for modeling.

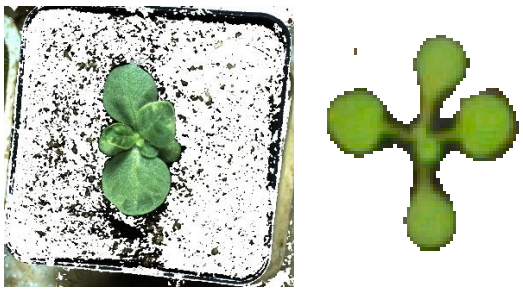

Fig. 2. Pre-processed image of Tobacco (left) and Arabidopsis plant (right)

B. Evaluation Methods and Metrics

The final classification models were evaluated based on their prediction on different ratio of split and the value of the parameters like the number of neighbors and number of estimators. The metrics used to measure the overall performance are discussed below:

a) Precision: This metric was used to measure the ability of the classifiier to predict a class that actually belong to that class. As we have a multi-class classification we set the average value to 'macro'.

b) Recall: In order to find the prediction of a class out of all the predictions from that class we use the metrics recall. If the precision-recall curve is in between the baseline and perfect classifier then the classifier can be taught of as a good classifer.

c) AUC: This is the measure of the area under the ROC curve (receiver operating characteristic curve). Typically it measures the probability of prediction of a correct class as against the wrong class. Range of AUC lies between 0.0 to 1.0.

The precision and recall were calculated using the inbuilt sklearn function 'precision_score' and 'recall_score'. For calculating the AUC we first obtained the prediction probability of the target. We select the probability of only the positive outcome and the calculate the AUC score using 'auc_roc_score' function of sklearn. The overall comparative

TABLE I.

Model	Metrics			
	Accuracy	***Precision***	***Recall***	***AUC***
Support Vector Machine	0.97	0.99	0.97	1.00
Random Forest	0.98	0.99	0.97	0.99
K-Nearest Neighbours	0.98	0.99	0.97	0.97

Fig. 3. Metrics comparison of three models (with pre-processing)

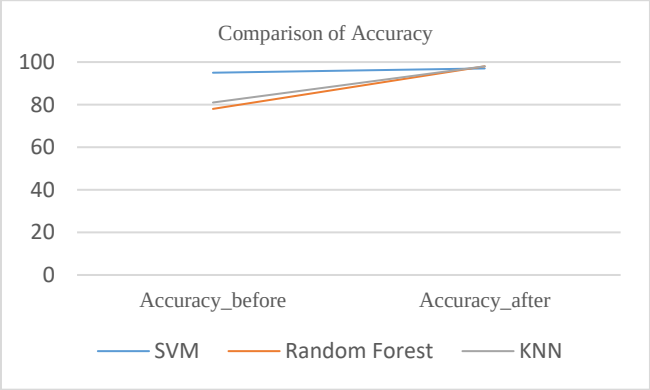

Fig. 4. Comparison of accuracy of three models

study shows the difference between the values of these metrics before and after pre-processing.

IV. Results and Discussion

The performance metrics for the three classification models with pre-processing is tabulated in Table I. Although precision and recall are the parameters which give us the ability of the model to predict correctly, AUC is the defining metric in this experiment. AUC is typically the area under the ROC curve which gives us a comparative graph between the false positive rate and true positive rate. It is evident that SVM has more ability to determine a positive class of the plant as compared to Random Forest and KNN. A small comparative study can be performed considering the accuracy improvement with the addition of pre-processing steps. We observe from Table I and II that the accuracy of SVM is slightly less as compared to KNN and Random Forest. The overall accuracy of all the three models has increased considerably with pre-processing. More than the accuracy the AUC of the three models is inclining more towards the fuller range indicating better prediction. It is evident from these results that SVM proves to be a better model for the classification of the Tobacco and Arabidopsis plant with the pre-processing involved in the initial steps.

TABLE II.

Model	Metrics			
	Accuracy	***Precision***	***Recall***	***AUC***
Support Vector Machine	0.95	0.97	0.92	0.99
Random Forest	0.78	0.73	0.76	0.86
K-Nearest Neighbours	0.81	0.83	0.67	0.97

Fig. 5. Metrics comparison of three models (without pre-processing)

Though this method uses just the local shape feature (SIFT) we can have a combination of global and local features to have get the best out of the SVM model. This topic can be expanded using the current pre-processing and supervised learning capabilities of SVM in combination with the global features like color histogram, hu-moments as well as other shape features.

References

[1] H. Scharr, M. Minervini, A.P. French, C. Klukas, D. Kramer, Xiaoming Liu, I. Luengo Muntion, J.-M. Pape, G. Polder, D. Vukadinovic, Xi Yin, and S.A. Tsaftaris. Leaf segmentation in plant phenotyping: A collation study. Machine Vision and Applications, pages 1-18, 2015.

[2] A. Hartmann, T. Czauderna, R. Ho mann, N. Stein, F. Schreiber, HTPheno: An image analysis pipeline for high-throughput plant phenotyping, BMC Bioinformatics 12 (1) (2011) 148+.BMC Bioinformatics 12 (1) (2011) 148+.

[3] J. De Vylder, F. Vandenbussche, Y. Hu,W. Philips, D. Van Der Straeten, Rosette Tracker: An Open Source Image Analysis Tool for Automatic Quantification of Genotype E ects, Plant Physiology 160 (3) (2012) 1149–1159.

[4] P. Hiremath, & J. Pujari, “Content based Image Retrieval based on Color, Texture and Shape Features Using Image and Its Complement”, International Journal of Computer Science and Security , vol. 1 (4), pp. 44-50, 2011.

[5] Z. Shanwen and Y. F., 2010 "Plant Leaf Classification Using Plant Leaves based on Rough Set." Proc. International Conference on Computer Application and System Modelling (ICCASM 2010), 2010, pp. 521-525.

[6] Zhu QQ, Zhong YF, Zhao B et al. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. Ieee Geosci Remote S. 2016;13(6):747–51.

[7] Wang JY, Li YP, Zhang Y, et al. Bag-of-features based medical image retrieval via multiple assignment and visual words weighting. Ieee T Med Imaging. 2011;30(11):1996–2011.

[8] Sezgin, Mehmet, and Bülent Sankur. "Survey over image thresholding techniques and quantitative performance evaluation." Journal of Electronic imaging 13.1 (2004): 146-166.

[9] Hussin, N. A. C., Jamil, N., Nordin, S., & Awang, K. (2013, December). Plant species identification by using scale invariant feature transform (sift) and grid based colour moment (gbcm). In 2013 IEEE conference on open systems (ICOS) (pp. 226-230). IEEE.

[10] Kabbai, L., Abdellaoui, M. & Douik, A. Image classification by combining local and global features. Vis Comput 35, 679–693 (2019). https://doi.org/10.1007/s00371-018-1503-0

[11] Tony Yiu (2019, July 12). Understanding Random Forest.

[12] Onel Harrison (2018, Sept 11). Machine Learning Basics with the K-Nearest Neighbors Algorithm.

[13] Rohith Gandhi (2018, Jun 8). Support Vector Machine — Introduction to Machine Learning Algorithms.

[14] Gurkan Demir, Bag of visual words, (2019), GitHub repository, https://github.com/gurkandemir/Bag-of-Visual-Words

[15] Donald Danforth, Plantcv, (2017), GitHub repository, https://github.com/danforthcenter/plantcv