Feature Selection Using Artificial Bee Colony for Fruit Classification

Mauj Hauder Abd Alkreem¹ and Prof. Dr. Abdulamir Abdullah Karim²

¹Department of Computer Science, University of Technology, Baghdad, Iraq
²Department of Computer Science, University of Technology, Baghdad, Iraq

Emails: cs.19.69@grad.uotechnology.edu.iq¹
110004@uotechnology.edu.iq²

Abstract. Extraction of features plays a very significant part in the creation of images. Different image pre-processing methods such as banalization, thresholding, resizing, normalization are added to the captured picture before receiving functionality. After that, techniques of extraction of features are applied to obtain features that will be useful in the classification and recognition of images. This paper aims to propose a fruit classification system based on an Intelligence algorithm, the proposed framework derives extracted features using merge harries corner detection with GLCM feature extraction. After that, the important features are selected using one of the methods of artificial intelligence, where the bee colony algorithm was used. After obtaining the selected features, two powerful classifiers are merged, which are the decision tree and naïve Bayes to classify the input fruit image. This work uses the Fruit 360 dataset. Where 70 percent of the data set was used during the training phase and 30 percent used during the test phase. It was observed that the proposed approach yielded good results and the accuracy was 96%.

1. Introduction

Fruit classification presents significant challenges due to interclass similarities and irregular interclass characteristics. data acquisition and feature representation approach are also crucial due to the huge diversity of the fruit field. The initial feature set consists of a large number of features that include some redundant features also. The high number of features may reduce the accuracy of the system and affect the classifier accuracy. In the real-time scenario, it would affect the computational efficiency of the overall system. So it is necessary to remove the redundant or unwanted features [1]. Feature selection is the process of selecting a suitable subset in feature space. The random search provides a random subset in the search space. So that many bio-inspired algorithms are developed to choose the optimal subset [2]. To select or minimize the feature set, a metaheuristic approach is followed in the proposed method. The optimal feature set will give effective features to the classifiers. The artificial bee colony method (ABC) is used in several domains in terms of wrapper and forward strategies. It is used in the optimization problem. ABC optimization method is used by very few researchers for feature selection. ABC algorithm highlights the brainy behavior [2] of bees that are scheduled by their tasks: hired, observers, and guides. found the hunting fashion of bees, interior, and exterior information. The information about food location is expressed in terms
of the waggle dance. The ultimate model of ABC shows the Intelligent behavior of bees concerning the following parameters: food location, engaged bees, and unengaged bees [3]. ABC algorithm consists of three groups of bees: engaged bees, unengaged bees, and scout bees. In this method number of bees in the engaged and unengaged are the same. The unengaged bees make the decision of food selection in the dance area. The unengaged bee is named as an engaged bee when it drives to the food sources. If the engaged bee has the food, then it returns as a scout bee and it will find a new food source by random search.

2. Literature Review
In this paragraph, a brief description of different schemes which have been proposed in fruit classification as shown below:

**Robin, B. [4]** two decision trees are compared to classify fruits and vegetables. Decision trees differ in their zoning structure to discover which form of structure is most efficient. Although the trees used similar data, the variation in the structure created completely different classifications. The classification accuracy for both decision trees in the final test was similar. Fortunately, the tree with non-binary cleavages depending on the entity type (color and shape) was found to be most effective because it was much faster and showed greater consistency than its equivalent based on the element.

**Mandeep Kaur and Reecha Sharma [5]**, presented the features of the image which were used to reveal the quality of the luminaries using the method of combining the features of the color and shape and size of the vegetables on the basis of a combination of the industrial grid. The system starts first taking pictures of the vegetables and then moves those images to the level of processing to be extracted features, followed by used the neural network, the discovery of the quality of the imagination by relying on the features extracted during the training and provide the result compared to those features. The experiment result gives 88% accuracy.

**Zhang, Y et. al. [6]** proposed a method for fruit classification. Eighteen different types of fruits are classified which includes berries, grapes, and pineapples. Color histogram, shape and texture features are extracted from the input image dataset. Feed forwarded neural network was used to classify the different types of berries, grapes and pineapples. The different fruit varieties are classified with a classification accuracy of 89%.

**Zhang, Y et. al. [7]**. The improved method of Zhang et al. [6] is proposed in Zhang et al. [7]. It includes biogeography based optimization and wavelet-based feature selection. So the accuracy of this system was improved to 90%.

**Sabzi, S. et. al. [8]** proposed a method for classifying orange varieties. Three varieties of oranges are classified namely bam, Thomson, and pay band. Three hundred color orange images are used for the classification. A hundred images are used under each category. A large number of features are extracted from each category of orange images. Three different metaheuristic algorithms are used to find the reduced feature set which are harmony search, artificial bee colony, particle swarm optimization. After getting the optimal feature set, the classifiers are used to classify three different types of oranges. The classification accuracy of the system is 96.7%.

**Jana, s .et.al.[9]** A shape-based approach to fruit recognition has been suggested. It requires a pre-processing step that normalizes the fruit image in terms of variations in translation, rotation, scaling, including the use of features which do not alter due to varying lengths, growth stages, and fruit surface appearances. The approach proposed was applied to 210 photos of seven groups of fruit. The overall accuracy of the identification ranges from 88 to 95%.
3. The Proposed Method
The system of fruit classification used several techniques, but in general, there are four basic steps: pre-processing, feature extraction, feature selection, and finally, classification as shown in Figure (1). The system aims to develop a method to classify the fruit based on an optimized feature set. Texture features are used in this system. In this paper, Artificial Bee Colony optimization (ABC) is used to identify the optimal feature set. To nullify the redundant features.

![The Framework of the Proposed Fruit Classification](image)

3.1 Image Acquisition
Acquisition of the image is always where vision systems begin to complete their set task. Once obtained, several different processing methods can be used to carry out several tasks concerning the image. The reason why the image acquisition is always the first step in the workflow sequence is that if there is no image, the processing is impossible. There are multiple methods to acquire images including but not limited to, the use of cameras or scanners. The image which has been acquired needs to retain all features. Where fruit360 dataset was used, which is characterized by several of them. Size of multiple fruit set: 45 images (more than one fruit (or fruit category) per image) Number of categories: 81 (fruit) Image size: 100 x 100 pixels, behind the fruits a white sheet is placed as a background and all kinds of image fruits RGB dataset is available at GitHub and Kaggle. Figure (2) shown the sample of the fruit 360 dataset.
3.2 Pre-processing stage using the HSV model
The first step in this work is to apply the pre-processing image to improve the fruit images. This technique is used to improve the image to show the image detail well because sometimes the images may be taken in conditions that are inappropriate in terms of light, noise or the size of the image is very large and does not produce good results. The color RGB model is converted into an HSV model where the parameters of H, S, and V are not fixed but flexible in the arrangement used to obtain an acceptable HSV image.

3.3 Harries and GLCM Feature Extraction
The most important stage in the identification system is feature extraction from samples such as fruit images, in this method to improve the Co-Occurrence Matrix method, the Harris corner detector has been used. There are three important parameters: the quality limit, the minimum distance between two feature points, and the maximum number of feature points that must be calculated. During all experiments, we used a distance of at least 5 pixels. The quality threshold is set to 0.001 to produce many features and respectively apply all angles by calculating the probability of a neighboring relation of two pixels at a specified distance (d) and angle (θ). After inserting the fruit image, Harris’ angle detector takes into account the difference of the corner points concerning the direction directly, instead of using the change spots for all 45-degree angles, after calculating all the fine edges and extracting the features for all, it enters another stage, which is the GLCM stage that is done from During which he calculated 7 features. The features taken are contrast, correlation, energy, and homogeneity, entropy, mean, stander deviation.

3.4 Feature Selection using ABC algorithm
In feature selection, an artificial bee colony algorithm is used, which is inspired by the intelligent behavior of the bees from the searches for food and the news of the rest of the workers about food from vibrating dances that were used to obtain the strong feature to facilitate the classification process.
An artificial group of bees in the ABC algorithm consists of three different groups: employed bees, onlooker bees, and scout bees. In the ABC algorithm, the number of bees employed in the colony also equals the number of onlooker bees. Additionally, the number of employed bees or onlooker bees equals the number of solutions in the population. An onlooker bee is a bee that waits in the dance area to make the food source selection decision. An onlooker bee is named employed bee once it goes to a food source. An employed bee that has consumed the food source turns into a scout bee, and it must perform a random search to discover new resources. Food supply position—which represents the solution to the optimization problem—and the amount of nectar in the food source depends on the quality of the associated solution. This value is calculated in (1).

\[
\text{fit}_i = \frac{1}{1 + f_i}
\]

SN in the algorithm indicates the size of the population where SN=1000 in the proposed method. At first, the ABC algorithm produces a distributed initial population \( P \) (\( C = 0 \)) of SN solutions (food source positions) randomly, where SN means the size of the population. Each \( z_i \) solution is a D-dimensional vector for \( i = 1, 2, 3, \ldots, \text{SN} \).
Here, \( D \) is the numbers of cluster products and input size for each dataset. After start-up, an investigation is repeated on employed bees=500, onlooker bees=500, and scout bees=100, the processes are repeated until the number of population of positions (\( C = 1, 2, \ldots, MCN \)) is completed.

3.5 Classification of fruit

The goal of fruit classification is to build or find a model to predict the category of data based on some predictor variables. Where two decision tree algorithms with naive Bayes were merged. A decision tree is a decision support tool as it uses a tree-like model of decisions and their potential consequences, including outcomes of encountered events, resource costs, and benefits. That is, it represents one of the ways to build an algorithm that only contains conditional control data, and therefore this algorithm is considered unstable, which means that a slight change in the data can lead to a significant change in the structure of the optimal decision tree. Therefore, it was combined with the naive bases method to give the probability. The error, though slight, calculates all the possibilities for the system, and thus the resulting decision is very accurate. The merge process is illustrated through a set of steps:

1. Transform data collection into a table of frequencies
2. Make a decision node attribute to that and split the dataset into smaller subsets.
3. Begins construction of tree by repeating this process recursively for each child until one condition matches:
   3.1: All tuples are of the same type as the attributes.
   3.2: There are no more attributes left over.
   3.3: No more instances do occur.
4. Build a table of probability by defining the probabilities
5. Use now the Naive Bayes equation to measure the later likelihood for each class. Prediction outcomes are the class with the highest posterior likelihood.

4. Experimental Results

In this paragraph, the results obtained are illustrated to find the best features of the fruit image and its classification.

4.1 Pre-processing stage:

This is the important stage during which the original image is converted to a HSV color model image as shown in Figures (3).

| Original Image | HSV image |
|----------------|----------------|
| ![Original Image](image1.png) | ![HSV image](image2.png) |
| ![Original Image](image3.png) | ![HSV image](image4.png) |
| ![Original Image](image5.png) | ![HSV image](image6.png) |

Figure (3) Pre-processing Image
4.2 Feature Extraction using Harries and GLCM Feature Extraction
After the improvement process of the GLCM Feature Extraction method by combining it with Harris, this method gave good results because Harris gives interest point.

| Harries and GLCM Feature Extraction |
|-------------------------------------|
| harries                             |
| mean                               |
| std                                |
| contrast                            |
| dissimilarity                      |
| homogeneity                        |
| ASM                                |
| energy                             |
| max                                |
| entropy                            |

4.3 Classify Using Decision Tree and Naïve Bayes
After the process of the feature extracting, and after applying the ABC algorithm to select the appropiable feature, the output of the ABC algorithm is feed to the decision tree and naïve Bayes classification to classify the fruit image, and hence confusion matrix has been constructed. based on the result of the confusion matrix the Sensitivity, Specificity, Accuracy measure has been calculated. Table (1) show the Evaluation performance of fruit classification.

1) Accuracy
Accuracy (ACC) is determined as a number for all predictions of correct (TP + TN) divided by the total number of data sets (P + N). The best accuracy equal to 1.0, while the worst equals 0.0. It can likewise be determined by 1 - error (ERR) as shown in equation (2).
\[ \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} \]  
\[ \text{SN} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{P}} \]  
\[ \text{SN} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{TN}}{\text{N}} \]

2) **Sensitivity (Recall or True Positive Rate)**

Determined the number of predictions of true positive (TP) divided by a total number of the positives (P) this method called Sensitivity (SN) or likewise Recall or the True Positive Rate (TPR)(REC). The sensitivity equal to 1.0 is best, whereas the worst equal 0.0 as shown in equation (3).

3) **Specificity (True Negative Rate)**

Determined the number of predictions of True Negative (TN) divided by the total number of negatives (N) this method called Specificity or True Negative Rate (TNR). The specificity equal to 1.0 is best, whereas the worst equal 0.0 as shown in equation (4).

**Table (1) Evaluation performance of Fruit Classification**

| Class    | Samples | FP | TN | TP | FN | Recall | Precision | Accuracy |
|----------|---------|----|----|----|----|--------|-----------|----------|
| Class 1  | 90      | 2  | 6  | 84 | 2  | 0.98   | 0.98      | 0.96     |
| Class 2  | 90      | 3  | 5  | 78 | 1  | 0.99   | 0.96      | 0.95     |
| Class 3  | 90      | 1  | 4  | 82 | 2  | 0.98   | 0.98      | 0.97     |
| Class 4  | 90      | 4  | 7  | 80 | 3  | 0.96   | 0.95      | 0.98     |
| Class 5  | 90      | 5  | 9  | 76 | 2  | 0.97   | 0.94      | 0.93     |
| **Average** |        |    |    |    |    | 0.98   | 0.96      | 0.96     |

5. Conclusion

In this work, the experiments conducted on fruit 360 dataset where, the pre-processing techniques are used and they are based on, converting RGB color model to HSV color model, it has been found that using HSV color model gives more appropriate results. The distinctive characteristics were obtained through the combination of Harris corner detection and the Gray Level Co-occurrence Matrix, where the results were invested from GLCM (gray system), and the results were improved by merging, so the results were more powerful to give strong feature concentrated at the edges and inside of the fruit, as well as a stronger selection of feature through the Bee Colony system. The obtained features are then used to classify fruit by using a decision tree and naive Bayes classifier. The merge of these two classification methods. provides a high-quality classification mainly because of its ability to reduce the expected error.
References

[1] Karaboga, D., Basturk, B., Ozturk, C., "Artificial bee colony (ABC) optimization algorithm for training feed-forward neural networks", Modeling Decisions for Artificial Intelligence, volume 4617/2007 of LNCS: 318-319, Springer, Berlin 2007.

[2] Diana Andrushia, A., & Trephena Patricia, A. (2018). Artificial Bee Colony Based Feature Selection for Automatic Skin Disease Identification of Mango Fruit. Nature Inspired Optimization Techniques for Image Processing Applications, 215.

[3] Schiezaro, M., Pedrini, H.: Data feature selection based on artificial bee colony algorithm. EURASIP J. Image Video Process. 47 (2013).

[4] Robin.B,, “ A Comparison of Decision Trees for Ingredient Classification”, Springer Science, 2012.

[5] Mandep Kaur1, Reecha Sharma2. " ANN based Technique for Vegetable Quality Detection". IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN: 2278-2834, p-ISSN: 2278-8735.Volume 10, Issue 5, Ver. I (Sep - Oct .2015), PP 62-70 www.iosrjournals.org.

[6] Zhang, Y., Wang, S., Ji, G., Phillips, P.: Fruit classification using computer vision and feedforward neural network. J. Food Eng. 143, 167–177 (2014)

[7] Zhang, Y., Phillips, P., Wang, S., Ji, G., Yang, J., Wu, J.: Fruit classification by biogeography-based optimization and feedforward neural network. Exp. Syst. 33(3), 239–253, (2016)

[8] Sabzi, S., Abbaspour Gilandeh, Y., Garcia Mateos, F.: A new approach for visual identification of orange varieties using neural networks and metaheuristic algorithms. Inf. Process. Agric. (2017)...

[9] Jana, S., Basak, S., & Parekh, R. (2017). Automatic fruit recognition from natural images using color and texture features. 2017 Devices for Integrated Circuit (DevIC).