Classification of Apple Types Using Principal Component Analysis and K-Nearest Neighbor

Moh. Arie Hasan 1

1 Universitas Nusa Mandiri, 14002250@nusamandiri.ac.id, www.nusamandiri.ac.id, Jakarta, Indonesia

Paper Information

Submit : May 20, 2023
Revision : May 28, 2023
Accept : June 10, 2023

Abstract

Apple is a fruit that is quite popular in Indonesia and is widely consumed by people. This fruit has various types of shapes and colors. Types of apples can be distinguished by their color, size, and shape, but it is still difficult for ordinary people to type apples that are more similar in color and size, such as the examples of Braeburn and Crimson Snow apples. This gave rise to the idea of researching image processing to classify the types of apples. This is to help determine the differences between the two types of apples. The classification process of apples is done by testing the image of an apple based on existing training data.

The research method consisted of preprocessing image segmentation with morphological operations and feature extraction into Principal Component Analysis (PCA). The classification algorithm used is a K-Nearest Neighbor (KNN).

Using adequate training data will further improve the classification of types of apples. The final results of this study amounted to 91.67%.

Keywords

Apple
Principal Component Analysis
Image Processing
K-Nearest Neighbor

1. Introduction

Apples are a type of fruit with various colors on its skin, and some are red, green, or yellow. The scientific name of the apple is Malus Domestica. Apples themselves belong to a genus called the genus malus. Where the Malus genus itself has a diversity center, namely in Eastern Turkey, it has been said that this apple is the first crop in the world of agriculture. The word apple is taken from an old English language, aeppl. From the Iron Age to the 1st century AD in Roman times, it has been found that apples have six types of apples. Because it has a shape almost similar to other fruits, many ordinary people find it difficult to classify the types of apples [1].

In image processing, computer graphics and computer vision can be considered "translating" an input image into a suitable output image [2]. An essential part of image processing is color. In addition to being seen by visualization, images also have essential information in presenting image quality. The external feature of color and the firmness of the interior features are the most critical factors that the consumer (wholesaler or retailer) observes to determine the quality of the fruit [3].

Based on these problems, we need to classify two types of apples, Braeburn apples, and Crimson Snow, using image processing. The process of classifying types of apples uses PCA and KNN methods. This study aimed to classify the types of Braeburn and Crimson Snow apples from the existing images. Furthermore, the image is
processed using the Matlab application to obtain the classification results of Braeburn and Crimson Snow apples.

2. Related Work

Many researchers have been working to solve the problem of not knowing how to differentiate apples over the past few years. Research conducted by Bhargava and Bansal in 2021 [4] finds evaluating the quality of fresh and rotten apples important because of their impact on human health and the agricultural sector. The proposed computer-based technique involves segmentation, feature extraction, and classification, resulting in an accurate classification with SVM achieving 98.42% accuracy. Future research should test this technique in various real-life environments and explore using deep learning-based features to reduce computational costs.

Guo et al. 2020 [5] study focused on using electronic noses to analyze apple gas information before and after inoculation with Penicillium expansa, a type of rot-causing fungus. Various classification models were created using PCA, PCA-DA, LDA, PLS-DA, and KNN, with PCA-DA showing the best prediction accuracy (100% for the training set and 97.22% for the prediction set). This study also uses selection methods such as SI, GA, and CARS to efficiently extract the relevant feature variables to build a PLS model to predict the area of rotten apples. Among these methods, the CARS-optimized PLS model produces the best prediction results (Rc = 0.953, RMSEC = 1.28, Rp = 0.972, and RMSEP = 1.01) for the damaged area.

This study by Zou et al. [6] addresses the challenges associated with the Apple quality assessment process, such as the time-consuming detection period and the inability to assess internal quality. A new electronic nose detection system (e-nose) based on the K nearest neighbor support vector machine (KNN-SVM) was developed, and the nasal cavity structure of the e-nose was optimized using computational fluid dynamics (CFD) simulations. The KNN-SVM classifier is proposed to overcome the limitations of traditional SVM. The KNN-SVM classifier is constructed using native 18-dimensional data and dimension-reduced PCA and LDA data. LDA data achieved a classification accuracy of 97.78%. Comparing the proposed classifier with other popular classifiers, it performs better with accuracy and speed.

3. Research Methods

The data used in this study consisted of 60 images of apples consisting of 30 training data for Braeburn apples and 30 training data for crimson snow apples. The test data consisted of 6 images of Braeburn apples and six of crimson snow apples [7]. An example of the image of these two types of fruit can be seen in Figure 1.

![Figure 1. Braeburn Apple (a) and Crimson Snow Apple (b)](source)

The process of classifying the type of apple image can be seen in Figure 2.

![Figure 2. Design of Classification of Types of Apples](source)
The figure shows that the method used to classify the apple image type starts from the apple image's input and then segments the image to obtain the segmentation result. Furthermore, the feature extraction process becomes red, green, blue, hue, saturation, and area values. The result of feature extraction that has been obtained is converted into a principal component. The next step is to classify with the KNN algorithm to determine the two types of apple images.

**RGB Image Input**

The RGB color space is widely used and is usually the default color space for storing and representing digital images. We can get other color spaces from RGB or non-linear transformations. RGB color space is used by computers, graphics cards, and monitors or LCDs [8]. This process aims to display the managed Apple image's RGB (Red, Green, Blue) color space. RGB formula:

\[
\begin{align*}
 r &= \frac{R}{R+G+B} \quad \text{(1)} \\
 g &= \frac{G}{R+G+B} \quad \text{(2)} \\
 b &= \frac{B}{R+G+B} \quad \text{(3)}
\end{align*}
\]

**Image Segmentation**

Image segmentation divides an image into homogeneous areas based on specific similarity criteria between a pixel's gray level and its neighboring pixels' gray level. Then the results of this segmentation process will be used for further processing. The Otsu method is a method for segmenting digital images using automatic threshold values, namely converting a grey digital image to black and white based on comparing the threshold value with the pixel color value of the digital image. To get the threshold value, some calculations must be done. The first step that must be done is to make a histogram. From the histogram, it can be seen the number of pixels for each grey level. The grey level of the image is expressed as i to L. The ith level starts from 1, which is pixel 0. For L, the maximum level is 256, with a pixel worth 255[9]. The threshold value to be sought from a grayscale image is expressed by k. The value of k ranges from 0 to L-1, with a value of L = 256. So the probability of each pixel at the level the equation expresses me:

\[
P_i = \frac{n_i}{N} \quad \text{(4)}
\]

The formula for the cumulative sum of \( p_i \), where \( L = 0, 1, 2, ..., L-1 \):

\[
\omega(k) = \sum_{i=0}^{k} p_i \quad \text{(5)}
\]

The formula for the cumulative mean of (k), where \( L = 0, 1, 2, ..., L-1 \):

\[
\mu(k) = \sum_{i=0}^{k} i \cdot p_i \quad \text{(6)}
\]

The formula for calculating global mean intensity (k) \( \mu_T \) :

\[
\mu_T(k) = \sum_{i=0}^{L-1} i \cdot p_i \quad \text{(7)}
\]

The equation for the between-class variance:

\[
\sigma_B^2(k) = \frac{[\mu_T\omega(k)-\mu(k)]^2}{[\omega(k)[1-\omega(k)]]} \quad \text{(8)}
\]

The result of the calculation between class variance is sought for the maximum value. The most significant value is used as a threshold or threshold value (k) with the equation.

\[
\sigma_B^2(k*) = \max_{1≤x≤L}\sigma_B^2(k) \quad \text{(9)}
\]

Between classes, variance aims to find the threshold value of a grayscale image. The threshold value is a reference value to convert a grayscale to a binary image. Each image has a different threshold value [9].

**Hue Saturation Value (HSV)**

The input image in RGB color space is converted to HSV color space using transformations. An HSV image is a collection of three images: hue, saturation, and value [10].
HSV is closely related to the RGB system in describing colors humans can see. HSV functions to reduce the intensity of light from outside and can detect particular objects. Here's the RGB to HSV formula:

\[ C_{\text{max}} = \max(R', G', B') \]
\[ C_{\text{min}} = \min(R', G', B') \]
\[ \Delta = C_{\text{max}} - C_{\text{min}} \] ............................ (10)

Cmax functions to determine the most significant constant value for RGB values, while Cmin determines the smallest value for RGB values.

Principal Component Analysis (PCA)

Main Component Analysis is used to project or convert an extensive data set into a form of data presentation with a smaller size. PCA transformation of a ample data space will produce several orthonormal base vectors in the form of a collection of eigenvectors from a specific covariance matrix that can optimally represent the data distribution. [11]

In this case, the covariance method is used with the following algorithm:

Collects data in the form of a grey-level matrix \( \mathbf{X} \) of size \( M \times N \). Suppose that it is a vector \( \mathbf{x} \):

\[ \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_N \end{bmatrix} \]

Calculating the average:

\[ \mathbf{x} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{x}_i \] ............................ (11)

Calculating the difference in average:

\[ \Phi_i = \mathbf{x}_i - \mathbf{x} \] ............................ (12)

Determine the covariance matrix of the matrix \( \mathbf{X} = [\Phi_1 \Phi_2 \ldots \Phi_M] \) (matrix \( N \times M \)), Calculate covariance:

\[ \mathbf{C} = \frac{1}{M} \sum_{n=1}^{M} \Phi_n \Phi_n^T = \mathbf{X} \mathbf{X}^T \] ............................ (13)

Determine the characteristic value and characteristic vector of the covariance matrix \( \mathbf{C} \):

\[ \lambda_1 > \lambda_2 > \ldots > \lambda_N \] ............................ (14)

Sort the characteristic vector \( \mathbf{u} \) and the characteristic value \( \lambda \) in a diagonal matrix in descending order according to each characteristic vector's most considerable cumulative probability value to obtain the dominant characteristic values. [12].

K-Nearest Neighbor (KNN)

The following method uses the K-Nearest Neighbor (KNN) algorithm to classify new objects based on attributes and training samples, where the results of the new test samples are classified based on the majority of the categories on the KNN. In the classification process, no model is used to match and only based on memory [13].

The working principle of the KNN is to find the closest distance between the data to be evaluated and its closest neighbors in the training data. The training data is projected into a multi-dimensional space, where each dimension represents a data feature. This room is divided into sections based on the classification of training data. A point in this space is marked class \( c \) if class \( c \) is the classification mainly found in the k nearest neighbors of that point. Near or far neighbors are usually calculated based on the Euclidean distance with the following formula:

\[ d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2} \] ............................ (16)

Where \( x_1 = \) sample data, \( x_2 = \) test data, \( i = \) data variable, \( \text{dist} = \) distance, and \( p = \) data dimension. In the learning phase, this algorithm only stores feature vectors and classifications from the learning data. The same features are calculated for the test data in the classification phase. The distance of this new vector to all of the learning data vectors is calculated, and the nearest k number is taken. The point that has just been classified is predicted to be included in the largest classification of these points. The
best k value for this algorithm depends on the data. Generally, a high k value reduces the effect of noise on the classification but makes the boundaries between each classification more blurred [14].

**Data Distribution Plotting**

Plotting of data distribution is done to test and view the graph of the distribution of image data that is processed based on hue and saturation values to see the results of the accuracy of testing the apple-type image processing. The data distribution plot will be displayed in the distribution of training data in each class, the distribution of training data for each class, the boundary lines, and the distribution of test data in each class.

4. Results and Discussion

The initial process of classifying the types of apples is inputting the apple image from the training data in this study. Training data can be seen in Figure 3

![Figure 3. Apple Fruit Training Data Image](image3.jpg)

Furthermore, the data that has been inputted is performed for image segmentation using the morphological operation method to improve the segmentation results. It converts a grayscale image to determine the foreground and background areas with a binary image value, as shown in Figure 4.

![Figure 4. Binary Image](image4.jpg)

Then the color space transformation from RGB image to HSV image (Hue, Saturation, Value) is used as a reference for recognizing the color of an object in a digital image and reducing the light intensity from the outside which can be seen in Figure 5.

![Figure 5. Image of Segmentation Results](image5.jpg)

Based on the image segmentation results obtained, feature extraction is done to get the Apple image's RGB, hue, saturation, value, and area values. After that, the reduction was made using the PCA algorithm to get the image classification results of Braeburn apples and Crimson Snow.

The test results can be seen from the processed image data plotting graph. The plotting of the distribution of training data in each class is shown in Figure 6.

![Figure 6. Distribution of Training Data](image6.jpg)
Based on the training data obtained, testing is carried out using test data. The following displays training data distribution and test data based on boundary lines using the PCA and KNN algorithms seen in Figure 7.

![Figure 7. Distribution of Test Data and Training Data](image)

Furthermore, tests were carried out using the GUI application using the Matlab application to get the classification accuracy of types of apples. The application design consists of several functions, namely image input, image segmentation process, feature extraction, and determining classification results. The following shows the Matlab GUI application that has been created.

![Figure 8. Display of the Matlab GUI Application Image Input](image)

The following is the appearance of the Matlab GUI application of the segmented image, consisting of binary and segmented images.

![Figure 9. Display of Matlab GUI Application Image segmentation results](image)

The following shows the results of image feature extraction consisting of Red, Green, Blue, Hue, Saturation, Value, and Area features in the Matlab GUI Application:

![Figure 10. Display of the Matlab GUI Application of Feature Extraction Results](image)

Based on the feature extraction results obtained, a process is carried out to determine the classification using the PCA and KNN algorithms. The following is a display of the Matlab GUI classification results for Braeburn apples:

![Figure 11. Display of the Matlab GUI Application Results of Braeburn Apples Classification](image)
Figure 12. Display of the Matlab GUI Application Classification Results of Crimson Snow Apples

The test results were carried out on the test data of 12 apples using the Matlab GUI application, which can be seen in Table 1.

Table 1. Results of Apple Image Processing

<table>
<thead>
<tr>
<th>No.</th>
<th>Test Image</th>
<th>Original Class</th>
<th>Output Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Apel1.jpg</td>
<td>Braeburn Apples</td>
<td>Braeburn Apples</td>
</tr>
<tr>
<td>2</td>
<td>Apel2.jpg</td>
<td>Braeburn Apples</td>
<td>Braeburn Apples</td>
</tr>
<tr>
<td>3</td>
<td>Apel3.jpg</td>
<td>Braeburn Apples</td>
<td>Braeburn Apples</td>
</tr>
<tr>
<td>4</td>
<td>Apel4.jpg</td>
<td>Braeburn Apples</td>
<td>Braeburn Apples</td>
</tr>
<tr>
<td>5</td>
<td>Apel5.jpg</td>
<td>Braeburn Apples</td>
<td>Braeburn Apples</td>
</tr>
<tr>
<td>6</td>
<td>Apel6.jpg</td>
<td>Crimson Snow Apple</td>
<td>Braeburn Apples</td>
</tr>
<tr>
<td>7</td>
<td>Apel7.jpg</td>
<td>Crimson Snow Apple</td>
<td>Crimson Snow Apple</td>
</tr>
<tr>
<td>8</td>
<td>Apel8.jpg</td>
<td>Crimson Snow Apple</td>
<td>Crimson Snow Apple</td>
</tr>
<tr>
<td>9</td>
<td>Apel9.jpg</td>
<td>Crimson Snow Apple</td>
<td>Crimson Snow Apple</td>
</tr>
<tr>
<td>10</td>
<td>Apel10.jpg</td>
<td>Crimson Snow Apple</td>
<td>Crimson Snow Apple</td>
</tr>
<tr>
<td>11</td>
<td>Apel11.jpg</td>
<td>Crimson Snow Apple</td>
<td>Crimson Snow Apple</td>
</tr>
<tr>
<td>12</td>
<td>Apel12.jpg</td>
<td>Crimson Snow Apple</td>
<td>Crimson Snow Apple</td>
</tr>
</tbody>
</table>

Based on Table 1 above, 11 apples were classified according to their type, but 1 Crimson Snow Apple is classified as Braeburn Apple. The final result of testing accuracy for apples is 91.67%, obtained from 11 (number of correct) / 12 (number of test data) * 100%. The resulting accuracy value indicates that the PCA and KNN algorithms can be applied to classify fruit types.

5. Conclusion

This study obtained how ordinary people can quickly determine the type of apple fruit only from image processing. From the results of the image processing classification process for Braeburn and Crimson Snow apples, the accuracy is 91.67%. Using the Principal Component Analysis algorithm and K-Nearest Neighbour with the parameter k = 3, it can be used in the classification process for the type of apple. Image quality is very influential in the classification results and the amount of training data used to obtain classification results. The more training data used, the better the result of the classification accuracy of the type of apple fruit. The classification of types of apples still uses two types: Braeburn and crimson snow. Developing further research using more than two types of apples is recommended.

6. Referensi


