Tea leaf maturity levels based on ycbcr color space and clustering centroid

Bagaskara Aji Wicaksono¹, Ledya Novamizanti², Nur Ibrahim³
¹,²,³School of Electrical Engineering, Telkom University, Indonesia

E-mail : ¹bagaskaraaji@student.telkomuniversity.ac.id,
²ledyaldn@telkomuniversity.ac.id, ³nuribrahim@telkomuniversity.ac.id

Abstract. Technology develops very rapidly in all areas. Smart Farming 4.0 is a farming management concept that uses modern technology to increase quantity and quality. The picking of tea leaves during this time the farmer is only based on the quotes from the planting block. If the block is already arriving, then the block is taken in a thorough plucking. However, the picking time can be erratic due to weather factors. The design of the tea leaf maturity level identification system based on the digital image processing of tea leaves. The leaf image of the tea is then processed on a system that begins with segmentation preprocessing, the image that has been uniform then carried out the extraction of images transformed into the color features of YCbCr. After obtaining the luma and chroma values, the classification using Centroid is based on statistical characteristics. Then the extraction and classifying data is a system database that will then be used during the testing process. The total data of Peko tea leaves (P + 2) is used as much as 90 training images and 90 test images. The maturity classification of tea leaves uses a Centroid Clustering method with centroid 10 based on YCbCr color space and a minimum statistical feature, maximum, and variances get an accuracy value of 80% and computation time of 2.80 seconds.

1. Introduction
Technology develops very rapidly in all areas of the 4.0 era. Smart Farming is the concept of quick capture in agricultural business [1]. Smart Farming has the real potential to produce a more productive and sustainable agricultural production, based on a more precise and resource-efficient approach [1]. Tea as a popular drink favored by Asians, in particular, Chinese, green tea has received much global attention for its promotion of human health [2]. Indonesia is one of the countries that exported the highest value of tea and ranks 7th on the country's list of the main tea producer [3].

In previous research studies have been conducted on maturity using YCbCr color space. On [4], Identification of starfruits using Cr color layer on YCbCr with accuracy 96%. System performance using centroid clustering is high [5]. This research is done to identify the maturity level of tea leaves using digital imagery that shown in figure 1. The digital image of the tea leaves is observed in green color in the RGB color space (Red, Green, Blue) which is then transformed into the color space YCbCr where Y is a luma component while Cb and Cr are the blue and red difference chroma components. The grouping is then performed using the Centroid Clustering method based on statistical features in the YCbCr color space.
2. Basic Theory

2.1. Digital Imagery
A digital image can be defined as a discrete 2D space that has the origin of an analog image \((x, y)\) into a 2D continuous image through a sampling process or commonly called digitization. Image sizes are usually measured in the number of dots or are often called pixels, where a pixel has certain coordinates on digital imagery. Each pixel has a value representing the information contained in that pixel. Figure 2 shows three types of digital imagery are true color, grayscale, and binary imagery [6].

2.1.1. True color image. The image does not use a color map and an image is represented by three colour component intensities such as red, green, and blue. RGB image uses 8-bit monochrome standard and has 24 bit/pixel where 8 bit for each color (red, green and blue) [6].

2.1.2. Grayscale. The grayscale image is a monochrome image or one-color image. It contains brightness information only and no color information. Then grayscale data matrix values represent intensities. The typical image contains 8 bit/pixel allows the image to represent (0-255) different brightness (gray) levels [6], [7].

2.1.3. Binary imagery. The simplest type of image and has two values, black and white or ‘0’ and ‘1’ [6]. The binary image is referred to as a 1 bit/pixel image because it takes only one binary digit to represent each pixel [6]. Converting from another color space into a binary color space defined by equation (2).

\[
m(x,y) = \begin{cases} 
1 & \text{if } f(x,y) \geq T \\
0 & \text{if } f(x,y) < T 
\end{cases}
\] 

(1)

with \(f(x,y)\) is a color space component input and \(T\) is the threshold value [8].

Figure 1. Peko Tea (P + 2) gambung tea (a) mature (b) half-mature (c) immature.

Figure 2. Image (a) true color (b) grayscale (c) binary.
2.2. Color Space YCbCr

YCbCr color space separates luminance information from chrominance information. Where Y represents the Luma value, Cb and Cr represent the blue and red chroma values [9]. The Luma component presents the intensity of light on the image, while the blue and red chroma present the components of the blue chroma difference and the Red chroma component. Chroma (chrominance) is its use to convey color information [9].

2.3. Statistical Features

Statistical features are a collection of values based on their histogramme characteristics [10]. Histograms indicate the probability of emergence of Y, Cb, or Cr values on an image. The statistical features used include minimum, maximum, variance, and standard deviation.

- **Minimum** is the lowest value of the histogram of an image.
- **Maximum** is the highest value of the histogram of an image.
- **Variance** is the value variations of an image's histogram. The variance calculated by the formula

\[
\sigma^2 = \frac{\sum_{n=0}^{N} (p(fn) - \mu)^2}{N}
\]  

(2)

where \( p(fn) \) is the histogram value (Probabilities occurrence) of Y, Cb, or Cr. \( \mu \) represents the average value of Y, Cb or Cr calculated using the equation (3) [9].

\[
\mu = \frac{\sum_{n=0}^{N} p(fn)}{N}
\]  

(3)

- **Standard Deviation** is the value of the histogram on an image. The standard deviation calculated by the formula

\[
\sigma = \sqrt{\frac{\sum_{n=0}^{N} (fn - \mu)^2 \times p(fn)}{N}}
\]  

(4)

where \( p(fn) \) is the histogram value (Probabilities occurrence) of Y, Cb, or Cr [10]. Parameter \( \mu \) represents the average value of Y, Cb or Cr calculated using the equation (6).

2.4. Centroid Clustering

Clustering is to group data objects into different classes or clusters and is one of the most tasks in data analysis, such as pattern discovery, pattern recognition, data summary and image processing [11]. A centroid is defined as the point whose coordinates are obtained by computing the average of each of the coordinates of the points of the jobs assigned to the cluster [6]. Formally, the centroid clustering algorithm follows the following steps [6]:

1. Choose a number of the desired cluster, \( k \).
2. Choose \( k \) starting points to be used as initial estimates of the cluster centroids. These are the initial starting values.
3. Examine each point in the data set and assign it to the cluster whose centroid is nearest to it.
4. When each point is assigned to a cluster, recalculate the new \( k \) centroids.
5. Repeat steps 3 and 4 until no point changes its cluster assignment, or until a maximum number of passes through the data set is performed.

Each cluster has a central/ middle point called Centroid [12]. At the beginning of the centroid is determined randomly but after a centroid, the running algorithm can be searched using the formula (8) with \( C \) being a centroid on the cluster, \( X_i \) is the i-dot point/object, and N is the number of objects [11].
\[ C = \sum_{i=1}^{n} \frac{X_i}{n} \]  

(5)

Above is the formula to find the average value. There are other slightly different formulas but have the same essence, with \( \mu_k \) is the centroid of the k cluster; \( X_q \) is q-object of the K-cluster, and Q is the amount of data from the k –cluster [12].

\[ \mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} X_q \]  

(6)

Grouping the data based on the closest distance to the centroid is calculated using the Euclidean Distance with \( x_i \) and \( y_i \) is the x and y value of the matrix i [12].

\[ d(xy) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \]  

(7)

3. Proposed Method

Identification of tea leaf maturity is generally seen from the harvest age. But on this system, it uses a digital image, especially green color on the tea leaves. A tea leaf image is then carried out extraction. Image extraction results, which are converted RGB values which are then converted into YCbCr color space to get a certain value. Once acquired these values are done classifying using the Centroid Clustering method. Then carried out analysis of the performance and accuracy of the system. Figure 3 shows the diagram of the registration system.

3.1. Image Acquisition

Image acquisition is the beginning of the tea leaf maturity identification system, which is the collection of tea leaves data for the training and test of tea leaf maturity system. Tea leaves data obtained from several blocks on PT. Perkebunan Nusantara VIII located in Sinumbra, Bandung, West Java, as many as 60 images with the planting age of different tea leaves. The image acquisition uses the Canon 600D.
DSLR (Digital Single Lens Reflex) camera with an 18-55mm lens. Image measuring 1 Megabyte – 2 megabytes with the format. JPEG or. JPG.

3.2. Preprocessing
After taking the test data, enter the preprocessing that equate all test data before input into the system. Pre-processing on the tea leaf identification system divided into three stages: segmentation, cropping, and transformation from RGB to YCbCr. The preprocessing and feature extraction stages show in Figure 4.

Figure 4. Preprocessing and feature extraction diagram.

Segmentation is the process to remove image background and just take tea leaf image only. In cropping, the segmentation image is done cutting according to the upper border, bottom, right, and left of the tea leaves. So that the image is taken to be processed only the image of the tea leaves. RGB to YCbCr is the process of transform green color on the leaves to get the value of RGB which then the value is done converting into YCbCr color space to get the Luma value.

3.3 Features Extraction
The value of YCbCr obtained in the exposition of color features is taken by each of its modifiers. To get a feature of imagery, the coefficient is used for the extraction of statistical features by using equations (5) – (7). Then the resulting feature of the imagery is used as a database for testing.
- Minimum shows the lowest Y, Cb, and Cr values of an image.
- Maximum shows the highest Y, Cb, and Cr values of an image.
- Variances shows variations of values that can be calculated using equations (4) by calculating the average histogram value first using formula (5).
- Standard deviation shows the size of the Y, Cb, and Cr values of an image that can be calculated using equations (6).

3.4 Classification

The principle of Centroid Clustering with the average attention of each object is then combined based on the main general spacing using equation (9) and (10). The grouping of test data is divided by observation blocks that have different picking ages. Figure 5 is the stages of the classification process.

![Diagram of Recognition System and Centroid Clustering](image)

**Figure 5.** The diagram of (a) Recognition system (b) Centroid clustering
Specifies a value of K, which is identical to the number of clusters. Because there is no other data or calculation result, then the middle point is chosen freely. Once known the midpoint, then do the closest distance calculations of each point with each centroid using the equation (10) to define the class. Once known temporary classes are based on closest distances with the centroid. Then specify a new centroid point, using the average value with equation formula (8) – (9). Thus, it generates a new point C1, C2, and C2. Since the centroid value changes, it is required to recalculate the closest distance of each point to each centroid using the equation formula (10) to define each class.

4. Result and Discussion

4.1. Effect of Light Intensity on System Performance
In this scenario, testing is done with the difference in light intensity on the image of tea leaves. The intensity of light is measured using the Lux meter from the range of 10,000 – 30,000 and 90,000 – 110,000. In the span of two intensity, the Lux has the best possible accuracy and computing time for the designed system. In this scenario, it is given a temporary parameter i.e. the amount of data 50% practice 50% test, four statistical minimum features, maximum, variance, standard deviation, as well as a three-point centroid amount. Result of light influence testing against accuracy value and computational time shown in Table 1.

| Lux              | Accuracy | Time of Computation (s) |
|------------------|----------|-------------------------|
| 10.000 - 30.000  | 65.66%   | 2.34                    |
| 90.000 - 110.000 | 46.67%   | 2.39                    |

Based on Table 1, the intensity of light can affect the accuracy value and computing time. Best results earned at Lux value 10,000 – 30,000 with an accuracy value of 65.66%. The Lux value, which is too high, affects the Luma value of YCbCr, making it difficult to distinguish between the mature, half-cooked, or immature image of the tea leaves.

4.2. Effect of YCbCr Color Layers on System Performance
In this scenario, a layer analysis is performed on YCbCr color space against the accuracy and timing of system computing. Based on the results of the Lux scenario, the light intensity value used ranges from 10,000 – 30,000 range. In this scenario used the number of centroids while the number of three points and the amount of data 90 practice 90 test. From the three layers of Y, Cb, and Cr colors are performed in combination to find the best accuracy results.

| Layers | Accuracy | Time of Computation (s) |
|--------|----------|-------------------------|
| Y      | 63.33%   | 2.50                    |
| Cb     | 64.44%   | 2.07                    |
| Cr     | 58.89%   | 2.34                    |
| YCb    | 61.11%   | 2.87                    |
| YCr    | 60.00%   | 3.21                    |
| CbCr   | 64.44%   | 2.09                    |
| YCbCr  | 65.66%   | 2.34                    |

Based on Table 2, combination of the three layers of color space YCbCr has an influence on the value of system accuracy. The best color layer combination is YCbCr with an accuracy value of 65.66%. Because the value on the Luma and chroma color layers on the tea leaf image is influenced by lighting.
4.3. Effect of Statistical Features on System Performance

In this scenario, the analysis that is performed is the use of statistical features against the accuracy value and compute time. Based on the results of the Lux scenario, the light intensity value used ranges from 10,000 – 30,000 range. In this scenario used the number of centroids while the number of three points and the amount of data 50% practice 50% test. Some statistical features to be used include maximum value, minimum, variance, and standard deviation. Out of these four traits are combinations of features to find the best results. The test results are being tested in Table 3.

| Statistical Features | Accuracy | Time of Computation (s) |
|----------------------|----------|-------------------------|
| Min                  | 55.56%   | 2.28                    |
| Max                  | 51.11%   | 2.40                    |
| Var                  | 61.11%   | 2.76                    |
| Std                  | 60.00%   | 2.57                    |
| Min, Max             | 48.89%   | 2.98                    |
| Min, Var             | 65.56%   | 2.39                    |
| Min, Std             | 78.89%   | 3.17                    |
| Max, Var             | 61.11%   | 2.19                    |
| Max, Std             | 51.11%   | 2.18                    |
| Var, Std             | 65.56%   | 3.00                    |
| Min, Max, Var        | 73.33%   | 2.27                    |
| Min, Max, Std        | 51.11%   | 2.52                    |
| Min, Var, Std        | 64.44%   | 2.42                    |
| Max, Var, Std        | 66.67%   | 2.31                    |
| Min, Max, Var, Std   | 70.00%   | 2.30                    |

Based on Table 3, the highest accuracy value of 73.33% when using a combination of the minimum, maximum, and variance values. Because the third has a value that is varied so it has a considerable accuracy value.

4.4. The Effect of Total Centroids on System Performance

In this scenario, the analysis of the number of centroid changes in the classification against the accuracy and timing of the system computation. Based on previous scenarios the parameter used is the value of lux ranges from the range of 10,000 – 30,000, the statistical characteristics used are minimum, maximum, variance, and standard deviation. Also given parameters while the amount of data 50% practice 50% test. The test results of a centroid count can be seen in Figure 6.
Based on Figure 6, the maximum training accuracy result of 100% with a total centroid 10 dots. The addition of the number of dots does not guarantee that the accuracy of the system will increase, have a fixed accuracy possibility or can also be the accuracy value decreases depending on the spread of data and the closest distance with the centroid.

4.5. Testing Results
Based on previous training, it can be concluded that the tea leaf maturity level identification test can use the light intensity parameters 10,000 – 30,000 lux, YCbCr color layer, three minimum statistical features, maximum, variance, and total centroids as much as 10 dots. The final result of system tests can be seen in Table 4.

| Table 4. Testing Results. | Accuracy | Time of Computation (s) |
|---------------------------|----------|-------------------------|
|                           | Train    | Test        | Train | Test |
| 100%                      | 4,12     | 2,10        |       |       |

Based on Table 4, the accuracy value of 80% and the computing time of system testing for each tea leaf image of 2.01 seconds.

4.6. Testing Amount of Data
In this test, tests were conducted based on a comparison of the amount of training data and tea leaf test data. The comparison of training data and test data is calculated based on an overall percentage. Of these comparisons are sought-after that have the best accuracy and computing time. In this scenario, it is given temporary parameters of lux value 10,000 – 30,000, three minimum statistical characteristics, maximum, variances, as well as a 10-point centroid amount. Test results based on the percentage of data amount comparison.

Figure 7 can be seen that the amount of data 50%-50% has the best test accuracy value of 80%. Minimal training data can lead to a major failure rate in testing.

5. Conclusion
It has been designed the tea leaf Maturity level identification system based on the color features of YCbCr and classified using Centroid Clustering, generating an accuracy value of 80% with a compute time of 2.80 seconds. Determination of maturity level of tea leaves using green color on leaf image
which is then transformed into YCbCr color space. The color layer of the YCbCr color space generates maximum accuracy when the Y, Cb, and Cr layers. Because on observation there are influences of light intensity and color. The statistical features that generate maximum accuracy are the minimum, maximum, and variance characteristics. Because it has a varying values to represent each class. The number of centroid with a maximum accuracy of 10 dots.

6. References
[1] R. Gorli and Y. G. 2018 Future of Smart Farming with Internet of Things J. Inf. Technol. Its Appl., vol. 2, no. 1, pp. 27–38.
[2] W.-Y. Huang, Y.-R. Lin, R.-F. Ho, H.-Y. Liu, and Y.-S. Lin 2013 Effects of Water Solutions on Extracting Green Tea Leaves Sci. World J., vol. 2013, no. 111, pp. 1–6.
[3] D. Workman 2019 Tea Exports by Country World’s Top Exports [Online]. Available: http://www.worldstopexports.com/tea-exports-by-country/. [Accessed: 16-Jul-2019].
[4] R. Amirulah, M. M. Mokji, and Z. Ibrahim 2010 Starfruit color maturity classification using Cr as feature,” Proc. 6th Int. Conf. Signal Image Technol. Internet Based Syst. SITIS 2010, pp. 93–97.
[5] M. Sakthi and A. Selvadoss T 2011 An Effective Determination of Initial Centroids in K-Means Clustering Using Kernel PCA vol. 2, no. 3, pp. 955–959.
[6] K. Padmavathi and K. Thangadurai 2016 Implementation of RGB and grayscale images in plant leaves disease detection - Comparative study Indian J. Sci. Technol., vol. 9, no. 6, pp. 4–9.
[7] R. D. Atmaja, L. Novamizanti, J. Halomoan, M.A. Murti 2017 An image processing algorithm to find edges in grayscale image, Far East Journal of Electronics and Communications, vol: 17, issue : 4.
[8] Kalavathi P 2014 A Thresholding Method for Color Image Binarization SSRG Int. J. Comput. Sci. Eng., vol. 1, no. September, pp. 31–40.
[9] V. V K and S. C. Sajjan 2016 Fire Detection using YCbCr Color Model vol. 2, no. 2, pp. 698–701.
[10] M. F. Al-Saleh and A. E. Yousif 2017 Properties of the Standard Deviation that are Rarely Mentioned in Classrooms Austrian J. Stat., vol. 38, no. 3, p. 193.
[11] F. E. Only, M. E. Celebi, H. A. Kingravi, and P. A. Vela 2013 Edited by Foxit Reader Copyright ( C ) by Foxit Software Company , 2005-2007 Expert Systems with Applications A comparative study of efficient initialization methods for the k-means clustering algorithm Edited by Foxit Reader Copyright ( C ) by Foxit Soft,” Expert Syst. Appl., vol. 40, no. 1, pp. 200–210.
[12] R. Primartha 2018 Belajar Machine Learning Teori dan Praktik. Bandung: Informatika.