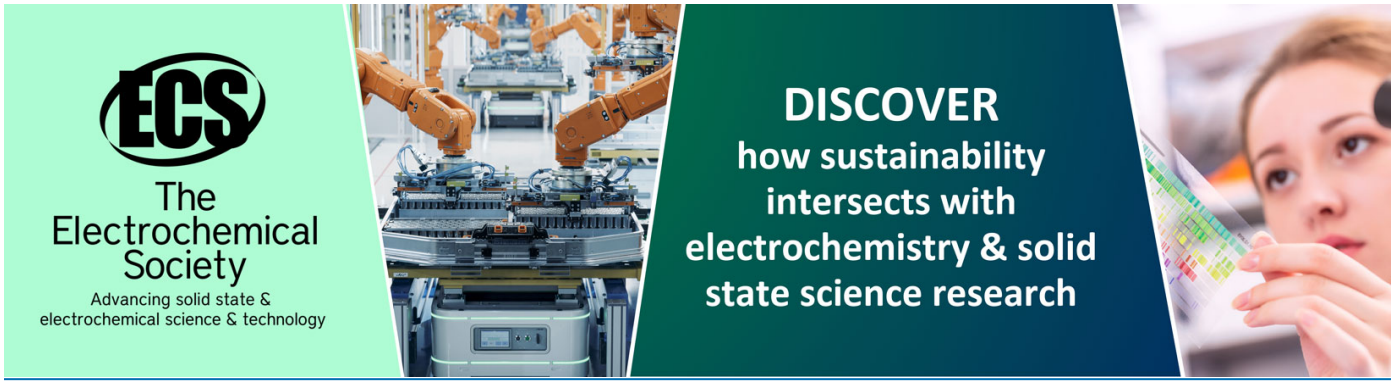
The quality of coffee bean classification system based on color by using k-nearest neighbor method



The quality of coffee bean classification system based on color by using k-nearest neighbor method

Nelly Oktavia Adiwijaya[1], Hammam Iqomatuddin Romadhon[1], Januar Adi Putra[1], Dewangga Putra Kuswanto[1]
[1]Informatics Department, Faculty of Computer Science, University of Jember, Indonesia

E-mail: nelly.oa@unej.ac.id, ramadan455@gmail.com, januaradi.putra@unej.ac.id, dewanggaputra570@gmail.com

Abstract. Sorting coffee bean nowadays is still done manually, although there is already a support machine for separation through size, but to determine the quality of the seeds remain manual using human power. This coffee bean sorting is in the spotlight to research whether it can be more effective if there is a tool that can directly find out the quality of coffee. This system will make it easier for workers in the field. In addition to saving time, costs will decrease and also the work of workers will be reduced. This paper present the implementation of machine learning method to classify the coffee bean quality. The dataset use 90 coffee bean for three classes and 30 for each class. From the experimental result, the highest accuracy obtain 83%.

1. Introduction

Coffee is one of the commodities in the world that is cultivated by more than 50 countries. With regard to agricultural commodities, coffee is the fourth largest foreign exchange earner for Indonesia after palm, rubber and cocoa [5]. The quality of coffee beans is closely related to the value of defects that coffee beans have by paying attention to how the processing is given to the coffee beans. The good character of coffee beans can be seen from the physical shape of the coffee beans, such as the color, shape and uniformity of size. Handling coffee beans to have good physical uniformity, done using the standards that have been set, namely based on coffee SNI number 01-2907-2008. Coffee bean sorting is currently still fairly manual, although there is already a support machine for separation through size, but to determine the quality of the beans remain manual using human power [8].

Indonesia's Coffee and Cocoa Research Center has engineered a conveyor table type sorting machine that is suitable and affordable by small entrepreneurs, both technologically and price-only [7]. The coffee beans will spread over a conveyor rubber belt, and workers sort out good physical-grade coffee beans from the deformed coffee beans manually. This coffee bean sorting is in the spotlight to research whether it can be more effective if there is a tool that can directly find out the quality of coffee. This system will make it easier for workers in the field. In addition to saving time, costs will decrease and also the work of workers will be reduced. The expected system will use optics to read the value of the coffee beans [3, 9].

Coffee quality measurement has been carried out based on coffee taste testing using support vector

machine. The accuracy obtained by the study was 48.33% [10]. Classification of coffee bean types is also carried out by extracting textures and then classifying them using the K-Nearest Neighbor method. The results obtained are 80% to 85% accuracy with K = 5 and K = 7 [2]. In this study data processing using the K-Nearest Neighbor method because this method is able to calculate the distance between two image objects to get the closest distance. The closest distance is used to classify the quality values in coffee according to the value k. The processing results will conclude the classification of quality in coffee beans based on color features, namely the average value of red, green, and blue in the training data. The benefit of this application is that it can help in quality classification by providing a perception of the same size and can save on production costs. The research aims to design and create applications that can classify quality using matlab and desktop-based programming languages.

2. Research Methods

2.1 System Development Techniques

The system development stage is carried out after the data analysis has been completed and used as material to build the system according to existing needs. Software development on this study using the waterfall model. According to [6] waterfall models take a systematic and sequential software development approach starting at the level of system progress until analysis, design, code, testing, and maintenance. As shown in Figure 1.

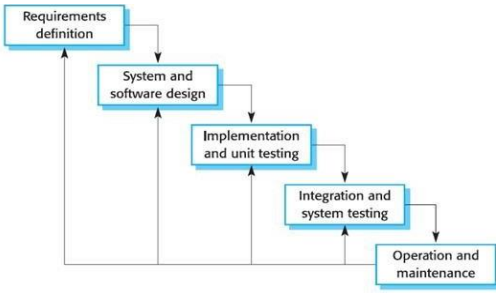


Figure 1. Waterfall method stages.

2.2 Data Analysis

Data analysis is the stage that is done after data collection or observation [4]. The data that has been collected is processed by taking pictures from the Camera Oppo A5 2020 Smartphone 12 MP. The dataset contains 90 images of coffee bean and 30 of each class. The label for classification is three classes. A data processing flow diagram can be seen in Figure 2.

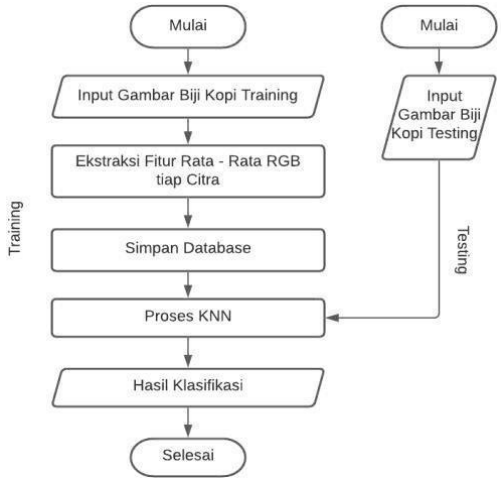


Figure 2. Data training processing flow diagram.

2.3 RGB Color Space

Rgb color space is commonly applied to CRT monitors and most computer graphics systems. This color space uses three basic components namely red (R), green (G), blue (B). Each pixel is formed by these three components. Rgb models are usually presented in the form of three-dimensional cubes, with red, green, and blue in the corner of the axis [12]. The RGB model on a three-dimensional cube can be seen in Figure 3.

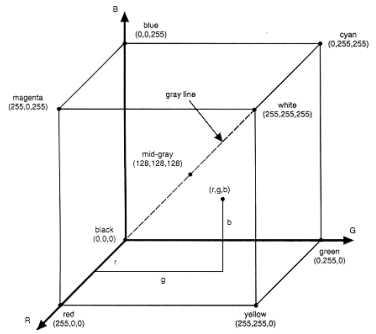
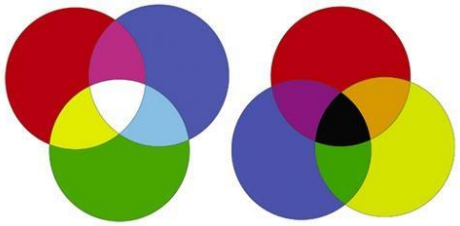


Figure 3. RGB model on three-dimensional cubes.

The image above is an image that shows the geometric shape of the RGB color model to specify colors using the Cartesian coordinate system. The grayscale spectrum is a color formed from a combination of three main colors of the same number, being on the line connecting the black and white dots. The formula for finding rgb values can be seen in equation 1:

$$r = \frac{R}{(255)}, g = \frac{G}{(255)}, b = \frac{B}{(255)} \quad (1)$$

Colors are represented in an additional beam to form new colors, and are associated to form mixed rays. The image below shows the mixture by adding the main colors red, green, and blue to form secondary colors of yellow (red+green), cyan (blue+green), magenta (red+blue) and white (red+green+blue). Additive color models and subtrative color models can be seen in Figure 4.



(a) (b)

Figure 4. (a) Additive color models, and (b) substantive color models.

2.4 K-Nearest Neighbor Method

K-Nearest Neighbor is a tough classification method for data training that has a lot of noise and this method has a very high accuracy rate and is effective if the data training used is large [11]. This simple algorithm works according to the shortest distance from test data to training data to determine K-NN. The training data consists of n attributes and a k value to determine its closest distance. A high k value reduces the error rate, but makes the boundaries between each classification even more blurred. One way to measure the distance of proximity between new data and old data (training data) is Euclidean distance. Euclidean distance is the formula for finding distance d from the root of the difference between the data on the record to i and to j according to the formula stated in equation 2 [4]:

$$d\left(X_i, X_j\right) = \sqrt{\sum_r^n \quad \left(a_r\left(X_i\right)\right) - \left(a_r\left(X_j\right)\right)^2} \tag{2}$$

Information:

$d(X_i, X_j)$	: Distance	a_r	: r data
(x_i)	: record to – i		: 1,2,3,.....n
(x_j)	: record to – j		

To measure the distance from attributes that have great value can be done by normalization. Normalization can be done with min-max normalization or Z-score standardization [1]. If the training data consists of mixed attributes between numeric and category, it is better to use min - max normalization. To normalize you can use the Z-score formula in equation 3:

$$V' = \frac{(v-A)}{\sigma A} \tag{3}$$

Information:

V'	: Z-score normalization results
v	: the value to be normalized
A	: the average value of attribute A
σA	: standard deviation attribute A

3. Results and Discussion

3.1 Data Set

At the classification stage of the application requires a data set that is used as data training. Data training is used as learning materials at the classification stage. The amount of data sets that exist affects the quality of data training carried out, the more data sets it will produce a higher accuracy value. Training data is obtained by extracting the histogram value of the digital image of coffee beans. The digital image of coffee was taken using the Oppo A5 2020 smartphone device that has a 12 MP camera quality with a shooting distance of approximately 15 cm, white background and has the same lightning level. The use of a white background is intended so as not to cause and affect noise at the time of pixel value extraction. The number of coffee beans used for shooting is 30 coffee beans per quality. Digital image capture is done in a mini studio box with LED lights and coffee placed under led lights, it is done to maximize the light against coffee beans to clearly provide pixel edge limits.

Figure 5. Example of digital image training data A.

Coffee bean shooting is done 90 times from 3 classes, where each class consists of 30 digital images. For training data used 24 images for each class. As for data testing is used each - 6 images each class. The results of digital image training data with a total of 72 images are then extracted to get the rgb average value. The expansion of RGB values is done per pixel as many pixels as in one digital image which is then used as data training is the average value of RGB per pixel. Rgb histogram values are used as training and classification data. Table 1 shows the dataset used as training data.

Table 1. RGB data set.

No	Average Red	Average Green	Average Blue	Type Quality
1	182,130	181,949	178,359	A

No	Average Red	Average Green	Average Blue	Type Quality
2	182,010	182,404	177,909	A
3	182,542	181,878	178,658	A
4	182,309	182,096	179,091	A
5	184,208	183,882	180,502	A
6	181,909	182,199	177,687	A
7	184,372	183,954	180,473	A
8	184,557	184,295	180,758	A
9	182,787	181,935	179,378	A
10	184,706	184,371	180,744	A
11	180,354	180,032	176,520	A
12	184,746	183,645	180,017	A
13	182,125	181,687	177,197	A
14	182,561	182,171	178,824	A
15	184,294	183,795	180,133	A
16	184,372	183,954	180,473	A
17	182,787	181,935	179,378	A
18	184,706	184,371	180,744	A
19	181,129	180,283	177,926	AA
20	183,023	182,076	179,784	AA
21	183,631	182,612	180,089	AA
22	181,400	181,030	177,953	AA
23	183,492	182,480	180,037	AA
24	181,926	180,886	178,436	AA
25	181,296	180,336	177,925	AA
26	183,474	182,479	179,731	AA
27	181,744	180,878	178,424	AA
28	180,587	179,665	177,132	AA
29	181,260	180,181	177,442	AA
30	183,423	182,223	179,387	AA
31	183,615	182,593	180,115	AA
32	184,951	183,984	181,646	AA
33	183,189	182,165	179,452	AA
34	181,400	181,030	177,953	AA
35	183,474	182,479	179,731	AA
36	184,951	183,984	181,646	AA
37	180,998	180,229	177,987	B
38	181,453	180,627	178,322	B
39	181,755	180,822	178,406	B
40	181,627	180,832	178,280	B
41	179,823	178,986	176,647	B
42	182,001	181,170	178,707	B
43	180,990	180,264	177,901	B
44	182,765	181,903	179,466	B
45	182,363	181,479	179,154	B
46	180,622	179,924	176,492	B
47	182,624	181,932	178,499	B
48	182,890	182,017	179,542	B
49	180,280	179,953	176,637	B
50	182,747	181,933	179,524	B

No	Average Red	Average Green	Average Blue	Type Quality
51	181,135	180,314	177,906	B
52	181,755	180,822	178,406	B
53	182,890	182,017	179,542	B
54	182,765	181,903	179,466	B

3.2 Application Accuracy Testing

Here are the results of matching application calculations with calculations or manual formulas using dataset testing amounting to 6 data in each class, so that the total data testing is 18 data. The test is done by selecting a lot of data to be trained with k = 3, k = 5 and k = 7. The data is sorted randomly but still places each class flat so that there is no data inequality. Test results with k = 3 are shown in Table 2, test results with k = 5 can be seen in Table 3, and Table 4 displays test results with k = 7.

Table 2. Test Results with k = 3.

Test Results 18 Data	
True Classifications	15
Incorrect Classifications	3
%Successful	83
%Wrong	17

Table 3.Test Results with k = 5.

Test Results 18 Data	
True Classifications	15
Incorrect Classifications	3
%Successful	83
%Wrong	17

Table 4. Test Results with k = 7.

Test Results 18 Data	
True Classifications	15
Incorrect Classifications	3
%Successful	83
%Wrong	17

Based on these results, the accuracy value obtained is the same which is 83% for the testing scheme of 18 data testing using k = 3, k = 5, and k = 7.

4. Conclusion

The development of a classification application for the classification of coffee bean quality classification based on color using the K-Nearest Neighbor method is based on the waterfall method. At the analysis stage, the application development requirements are obtained from observations and interviews to coffee farmers. In the next stage, the application development is carried out based on the results of the analysis.

The application of classification of coffee bean quality level by color using the K-Nearest Neighbor method is able to implement digital image processing through the capture of coffee bean imagery through digital image data which is then taken rgb average value.

The application of classification of coffee bean quality level by color using the K-Nearest Neighbor method is able to implement digital image processing through the capture of coffee bean imagery through digital image data which is then taken the average rgb value and determines the quality of coffee beans based on the parameters of the average RGB value of the coffee bean image. The RGB average is calculated distance using the Euclidean Distance formula. Classification is carried out when the distance value between the training data and the test data has been sorted based on the closest distance and concluded in accordance with the specified K value. In this study the K-Nearest Neighbor method had an accuracy rate of 83% on test data 18 to k = 3, k = 5 and k = 7. However, the application of coffee bean quality level classification requires further development on devices that have better camera quality to get better image results.

References

[1] Agusta Y 2007 K-means, Penerapan, Permasalahan dan Metode Terkait *Jurnal Sistem dan Informatika* **3** pp. 47-60
[2] Alessandra O 2017 *Klasifikasi Jenis Biji Kopi Dengan Ekstraksi Tekstur Berbasis Histogram Pengolahan Citra Digital Menggunakan Algoritma K-Nearest Neighbour*
[3] Cairns D 2009 *Intisari Kimia Farmasi Edisi Kedua* Translator: Puspita Rini
[4] Han J and Kamber M 2006 Data Mining: Concept and Techniques Second Edition *Elsevier Inc*
[5] Investments Indonesia 2017 *Indonesia Investments: Kopi* Retrieved 20 November 2017 from https://www.indonesia-investments.com/id/bisnis/komoditas/kopi/ item186?searchstring=kopi
[6] Pressman, R.S. 2012. *Rekayasa Perangkat Lunak Pendekatan Praktisi*. Yogyakarta: Andi
[7] Ridwansyah 2003 *Pengolahan Kopi* Fakultas Pertanian Universitas Sumatera Utara
[8] Saputra E 2008 *Kopi* Harmoni Yogyakarta
[9] Widyotomo S Mulato S & Suharyanto E 2006 *Optimasi Mesin Sortasi Biji Kopi Tipe Meja Konveyor untuk Meningkatkan Kinerja Sortasi Manual* Pelita Perkebunan **22**(1) 57-75
[10] Windrawati a. n. 2020 *Klasifikasi Varietas Kopi Arabika menggunakan Motode Support Vector Machine (SVM)* Yogyakarta
[11] Yofianto E 2010 *K-Nearest Neighbor (KNN). http://kuliahinformatika.wordpress.com/2010/02/ 13/buku-ta-k- nearestneighborknn/* Retrieved August 10 2016 8 p.m
[12] Kadir A and Susanto A 2012 *Teori dan Aplikasi Pengolahan Citra* Yogyakarta: CV Andi Offset