2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)
April 09-11. 2021, ISTTS Surabaya, Indonesia

Strawberry Ripeness Identification Using Feature Extraction of RGB and K-Nearest Neighbor

1st Siska Anraeni
Faculty of Computer Science
Universitas Muslim Indonesia
Makassar, Indonesia
siska.anraeni@umi.ac.id

2nd Dolly Indra
Faculty of Computer Science
Universitas Muslim Indonesia
Makassar, Indonesia
dolly.indra@umi.ac.id

3rd Desrial Adirahmadi
Faculty of Computer Science
Universitas Muslim Indonesia
Makassar, Indonesia
desrialadi7@gmail.com

4th Suwito Pomalingo
Faculty of Computer Science
Universitas Muslim Indonesia
Makassar, Indonesia
suwitopoms@gmail.com

5th Sugiarti
Faculty of Computer Science
Universitas Muslim Indonesia
Makassar, Indonesia
sugiarti@umi.ac.id

6th St. Hajrah Mansyur
Faculty of Computer Science
Universitas Muslim Indonesia
Makassar, Indonesia
shazwal12@gmail.com

2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT) | 978-1-6654-0514-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/EIConCIT50028.2021.9431854

***Abstract*— Nowadays, Indonesia has not been playing an active role in fulfilling the demand for strawberries in foreign markets yet. One of the reasons is the low quality of fruit selection that still uses conventional methods. Therefore, a proper method to group strawberries automatically is considered necessary. This research aims to identify the ripeness of strawberries using RGB feature extraction and the K-Nearest Neighbor (k-NN) algorithm. The strawberry image data used in this study is divided into two, namely training data consisting of 30 images, and test data with 20 images which is classified into four categories, i.e., ripe, unripe, raw, and not strawberry. Based on the test results obtained, incorrect classification is discovered happened on the unripe strawberry images due to the tendency of the red or green dominantly but not uniformly distributed. However, the accuracy of the ripeness classification is 85% for value of k used is 7. Therefore, it can be concluded that the system is able to detect the image of the strawberry category as well as the non-strawberry category.**

Keywords*— ripeness identification, RGB, k-nearest neighbor

I. Introduction

Among agricultural commodities, Strawberries demand is one of the greatest demands in supermarkets, hotels, restaurants, factories, and households globally. However, as a matter of fact, Indonesia as a major producer of agricultural products does not take part in meeting foreign market demands for strawberries. This is due to the low quality of fruit selection which is still using conventional methods depending on the human bare eyes. In addition, the length of the selection process decreases the freshness of the fruit to be sold. Therefore, an appropriate method to classify strawberries automatically with a high degree of accuracy is necessary. Several digital image processing applications implemented in agriculture include detecting and estimating the number of yields that may be obtained at harvest time [1], observing and detecting diseases that infect the fruit and leaves [2], classifying plant varieties [3], and harvesting using robots that are able to detect and sort ready-to-harvest fruits automatically [4]. Such wide application of digital image processing in agriculture will lead to a smart farming model of agricultural land processing [5].

Numerous research on ripeness identification of strawberry have been performed such as classification based on skin colour using Multi-Class Support Vector Machine obtaining 85,64% of accuracy [6]. Convolutional Neural Network (CNN) was used to classify the early ripe and ripe strawberry samples gaining the accuracy of 98.6% for testing dataset [7]. Multivariate nonlinear model had the highest identification accuracy (which was over 94%) in the greenhouse for rapid recognition of strawberry maturity [8]. In addition, hyperspectral imaging technology has been applied for automatic identification of strawberry ripeness [9]. Considering the importance of developing agricultural technology in Indonesia particularly in the processing of strawberry plantations, this research presented the application of digital image processing to identify the ripeness of strawberries which are categorized into four categories, namely ripe, unripe, raw, and not strawberry.

There have been many studies using the K-Nearest Neighbour (k-NN) method, particularly in classification issues. The k-NN algorithm is a supervised algorithm that uses the distance vector feature to classify data [10]. Research entitled identification of the quality of maize-based on colour and texture features using the K-Nearest Neighbour (K-NN) method has been performed with an accuracy of 90% at the value of k = 5 [11]. In addition, another study, namely the classification of the level of maturity of tomatoes based on the colour features of Hue, Saturation, and Value (HSV) using K-NN resulted in the highest accuracy of 92.5% in images measuring 1000x1000 pixels with the parameter k= 3 [12]. Thus, this study performed by extracting RGB values and computing the similarity of RGB values to define the ripeness of the strawberries. The expected result in this research is to obtain the best accuracy in the recognition.

II. Experimental Detail

A. Training and Testing Data

Fifty strawberry images are divided into two parts; 30 images for training data and 20 images of testing data which are all classified into 4 classes, namely the class of the ripe, unripe, raw, and not strawberry as shown in Fig. 1. The image size used is 128x128 pixels in JPG format.

(a) (b) (c)

Fig. 1. Strawberry image classification: (a) ripe, (b) unripe, and (c) raw

B. Research Method

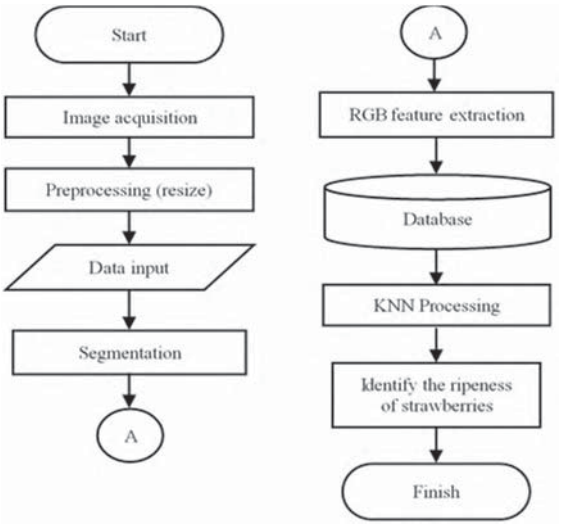

Fig. 2. Flowchart of System

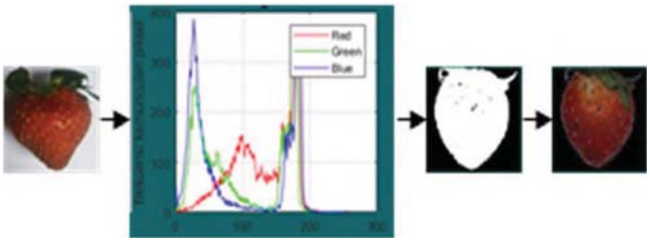

Fig. 3. Segmentation Process of RGB Images

Ripeness identification of strawberry using RGB feature extraction and k-NN is performed as presented in Fig. 2. This research consists of 8 stages: (1) Image acquisition; the reading process is carried out on the image stored in file form; (2) Pre-processing; resizing the image so that it is uniform with the overall image, from a size of 1024x768 pixels to 128x128 pixels; (3) Input data; reading training data and image test data; (4) Segmentation; RGB component separation process using the Thresholding method which aims to make the round colour value zero (black) and displayed in binary and RGB images as shown in Fig. 3; (5) RGB feature extraction; calculating the RGB component values, area features, roundness, and slenderness of the image; (6) Save to database; the final centroid value of each RGB component in the training data is saved into the database; (7) k-NN Processing; Carrying out the calculation of the Euclidean distance between the centroid values in the test data and training data. The result of the calculation of the closest or smallest distance will be the result of identifying the ripeness of the strawberry.

C. K-Nearest Neighbor

In classifying an object, a feature is required to distinguish each class. The k-NN algorithm is a method in classification analysis learning data based on the closest distance to the object [13]. Classification using the K-NN algorithm is performed by grouping the testing data based on the distance to the nearest k neighbours of the training data. The value of k used in this research is k = 3 and k = 7. The principle of k-NN is to calculate distance using the Euclidean distance. This Euclidean distance functions to measure the amount of distance between the evaluated data and the training data [14] which is represented in (1).

$$d_{(x,y)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

As the initial stage of k-NN algorithm is to determine the k value and continued to the process of calculating the distance between each data to be evaluated with all training. The next stage is sorting the distanced resulted earlier before determining the closest distance. Next, the class pairs will be corresponded. Lastly, the number of classes from the closest neighbors and the class of data will be defined.

III. Result and Discussion

A. Image Data

Strawberry image data used is in the form of color images consisted of 30 images of training data divided into 10 images of ripe strawberries, 10 images of unripe ones, and 10 images of raw strawberries. Meanwhile, the test data of 20 images are divided into 5 images each for ripe, unripe, raw, and non-strawberries consecutively. The training image data is presented in Table I.

B. Image Reading and Initial Centroid Generation

Image data processing is carried out by extracting the value of R, G, and B components. In the process, the three matrices are all in 128x128 pixels size for each component i.e., R, G, and B consecutively. The next step following the image reading process is to generate random initial centroid data. The centroid matrix size is $k \times$ the number of parameters, where k is the number of classes, and the number of input parameter is 3, namely parameters of R, G, and B. The value of k is 2 for the case, so the initial centroid matrix will be 2×3 in size. Thus, the initial centroid used for classification can be seen in Table II. The generation of initial centroid values is carried out with a range between the minimum and maximum values of the pixel components R, G, and B. Based on Table III, it is known that the ranges of each component are as 52 to 233, 47 to 227, and 25 to 229 for R, G, and B component, respectively.

TABLE I. Training Data Categorization

No.	Training data	Class		No.	Training data	Class
1	CL01.jpg	Ripe		16	CL16.jpg	Unripe
2	CL02.jpg	Ripe		17	CL17.jpg	Unripe
3	CL03.jpg	Ripe		18	CL18.jpg	Unripe
4	CL04.jpg	Ripe		19	CL19.jpg	Unripe
5	CL05.jpg	Ripe		20	CL20.jpg	Unripe
6	CL06.jpg	Ripe		21	CL21.jpg	Raw
7	CL07.jpg	Ripe		22	CL22.jpg	Raw
8	CL08.jpg	Ripe		23	CL23.jpg	Raw
9	CL09.jpg	Ripe		24	CL24.jpg	Raw
10	CL10.jpg	Ripe		25	CL25.jpg	Raw
11	CL11.jpg	Unripe		26	CL26.jpg	Raw
12	CL12.jpg	Unripe		27	CL27.jpg	Raw
13	CL13.jpg	Unripe		28	CL28.jpg	Raw
14	CL14.jpg	Unripe		29	CL29.jpg	Raw
15	CL15.jpg	Unripe		30	CL30.jpg	Raw

R			G			B		
52	62	88	47	80	57	25	30	28
133	180	225	71	110	150	32	61	93
125	166	233	120	115	227	124	60	229

Fig. 4. Illustration of R,G, B values of a color image.

TABLE II. INITIAL CENTROID

Class	Component Centroid		
	R	*G*	*B*
1	52	47	25
2	233	227	229

TABLE III. EUCLIDEAN DISTANCE TO THE INITIAL CENTROID

Pixel's order	Component			Euclidean Distance	
	R	*G*	*B*	*k1*	*k2*
1	52	47	25	0	326.77
2	133	71	32	84.77	270.45
3	125	120	124	143.03	184.76
4	62	80	30	34.84	300.75
5	180	110	61	147.13	211.47
6	166	115	60	137.27	213.52
7	88	57	28	37.48	300.54
8	225	150	93	212.51	156.48
9	233	227	229	326.77	0

C. Distance Calculation of Pixel to Centroid

The calculation of the Euclidean distance of the image pixel intensity given in Fig. 3 to the center point of each class represented in Table II is performed using (1) for each RGB component. The following (2) shows the Euclidean distance of the first pixel to the centroid of each class. Table III obtains the Euclidean distance of all image pixel data to the initial centroid of each class.

$$d_k = \sqrt{\sum_{i=1}^{n} pixel_i - centroid_i} \qquad (2)$$

D. Centroid Updating

The following step is to classify the class of each pixel data based on the shortest distance. Based on Table III, the first to 7th pixel data has the closest distance to the class 1 centroid. So that the data is grouped into class 1 (k1). The 8th and 9th pixel data have the closest distance to the 2nd class. Thus, the data is grouped in the 2nd class (k2). The process is continued to centroid updating for each class. Here (3) is the latest centroid calculation for each class. Hence, the final centroid values are illustrated in Table IV.

E. Identification Process

In the identification process using the k-NN algorithm, Firstly, $k = 7$ is used as the number of the nearest neighbors considered. Table V presents the result of Euclidean distance result between tested data and training data which have been evaluated. In addition, the comparison of the results of classification and the factual data are given in Table VI.

$$c_i = average \sum_{i=1}^{n} pixel_i \qquad (3)$$

TABLE IV. THE FINAL CENTROID

Class	Component Centroid		
	R	*G*	*B*
1	115.14	85.71	51.43
2	229	188.5	161

TABLE V. EUCLIDEAN DISTANCE RESULT

Training Data	Final Centroid Feature			Euclidean Distance
	R	*G*	*B*	
CL01	0,45073	0,16291	0,13865	0,068468586
CL02	0,40561	0,21997	0,18665	0,061708318
CL03	0,36036	0,15116	0,11631	0,049418176
CL04	0,3766	0,18972	0,13256	0,01180072
CL05	0,4176	0,1788	0,17012	0,045260678
CL06	0,38171	0,19025	0,17604	0,038351712
CL07	0,38062	0,17594	0,14328	0,013489989
CL08	0,45002	0,21249	0,14032	0,068426349
CL09	0,44864	0,16968	0,13806	0,064382869
CL10	0,44298	0,23693	0,16672	0,080616071
CL11	0,36739	0,30097	0,19156	0,127420166
CL12	0,46416	0,36036	0,20052	0,199993528
CL13	0,40634	0,3626	0,23772	0,202923278
CL14	0,41411	0,36764	0,24403	0,21127692
CL15	0,38877	0,30764	0,1797	0,127717164
CL16	0,39078	0,31358	0,19672	0,139636218
CL17	0,421	0,29487	0,2001	0,129154338
CL18	0,43695	0,28779	0,19688	0,127142225
CL19	0,40634	0,3626	0,23772	0,202923278
CL20	0,4611	0,29335	0,29335	0,142855141
CL21	0,32197	0,281	0,17787	0,120877655
CL22	0,3041	0,24958	0,14041	0,103640695
CL23	0,33774	0,29145	0,18605	0,124958006
CL24	0,28487	0,26355	0,17318	0,132109233
CL25	0,32077	0,30825	0,22568	0,163472748
CL26	0,33398	0,30382	0,19133	0,138814991
CL27	0,32666	0,28795	0,18336	0,125896272
CL28	0,34428	0,28632	0,16482	0,111304071
CL29	0,30041	27152	0,18859	0,130901286
CL30	0,30298	0,22445	0,12655	0,092399232

TABLE VI. COMPARISON OF OBSERVED DATA AND CLASSIFICATION RESULT

Testing Data	Observed Data	System	Identification
CU01.jpg	Ripe	Ripe	True
CU02.jpg	Ripe	Ripe	True
CU03.jpg	Ripe	Ripe	True
CU04.jpg	Ripe	Ripe	True
CU05.jpg	Ripe	Ripe	True
CU06.jpg	Unripe	Raw	False
CU07.jpg	Unripe	Unripe	True
CU08.jpg	Unripe	Unripe	True
CU09.jpg	Unripe	Raw	False
CU10.jpg	Unripe	Ripe	False
CU11.jpg	Raw	Raw	True
CU12.jpg	Raw	Raw	True
CU13.jpg	Raw	Raw	True
CU14.jpg	Raw	Raw	True
CU15.jpg	Raw	Raw	True
CU16.jpg	Other	Other	True
CU17.jpg	Other	Other	True
CU18.jpg	Other	Other	True
CU19.jpg	Other	Other	True
CU20.jpg	Other	Other	True

Based on Table VI, it can be computed that the accuracy of the system in recognizing the image of strawberries is:

$$Accuracy = \frac{the\ number\ of\ true\ classification}{total\ of\ testing\ data} \times 100\%$$

$$= \frac{17}{20} \times 100\% = 85\,\% \quad (4)$$

It was discovered that the incorrect identification results were in the classification of the unripe image where 3 out of 5 confirmed images are identified as 2 raw images and 1 ripe image. This is caused by the colours of unripe strawberries image which is not equally distributed or uniform. There is a tendency to be nearly red and the other is dominated of green colour. Therefore, for further system development is needed to increase the accuracy level.

IV. Conclusion

Based on the testing process performed on strawberry images using k-NN algorithm, it obtains 85% of accuracy in classifying the strawberry images. In conclusion, the system is able to identify those that are not strawberries (other fruits and objects) well. As a future work, It is recommended to use other algorithms that have better performance with more data, and the system development can be conducted on android-based.

References

[1] S. Nuske, S. Achar, T. Bates, S. Narasimhan, and S. Singh, "Yield estimation in vineyards by visual grape detection," pp. 2352–2358, 2011, doi: 10.1109/iros.2011.6095069.

[2] G. S. Gill, "Detection Of Diseased Section In Leaves Using Image Processing," vol. 11, no. 7, pp. 296–305, 2020, doi: 10.34218/IJARET.11.7.2020.030.

[3] K. P. Lewis and J. D. Espineli, "Classification and detection of nutritional deficiencies in coffee plants using image processing and convolutional neural network (Cnn)," *Int. J. Sci. Technol. Res.*, vol. 9, no. 4, pp. 2076–2081, 2020.

[4] L. Zhang, G. Gui, A. M. Khattak, M. Wang, W. Gao, and J. Jia, "Multi-task cascaded convolutional networks based intelligent fruit detection for designing automated robot," *IEEE Access*, vol. 7, pp. 56028–56038, 2019, doi: 10.1109/ACCESS.2019.2899940.

[5] J. Pandya, P. A. B, and M. E. Student, "Image Processing for Pomegranate Disease Detection : A Survey," vol. 1, no. 4, pp. 111–117, 2017.

[6] I. Indrabayu, N. Arifin, and I. S. Areni, "Strawberry Ripeness Classification System Based On Skin Tone Color using Multi-Class Support Vector Machine," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, Jul. 2019, pp. 191–195, doi: 10.1109/ICOIACT46704.2019.8938457.

[7] Z. Gao, Y. Shao, G. Xuan, Y. Wang, Y. Liu, and X. Han, "Arti fi cial Intelligence in Agriculture Real-time hyperspectral imaging for the in-fi eld estimation of strawberry ripeness with deep learning," *Artif. Intell. Agric.*, vol. 4, pp. 31–38, 2020, doi: 10.1016/j.aiia.2020.04.003.

[8] X. Yue, Z. Shang, J. Yang, L. Huang, and Y. Wang, "A smart data-driven rapid method to recognize the strawberry maturity," *Inf. Process. Agric.*, vol. 7, no. 4, pp. 575–584, 2020, doi: 10.1016/j.inpa.2019.10.005.

[9] H. Jiang, C. Zhang, F. Liu, H. Zhu, and Y. He, "[Identification of Strawberry Ripeness Based on Multispectral Indexes Extracted from Hyperspectral Images].," *Guang Pu Xue Yu Guang Pu Fen Xi*, vol. 36, no. 5, pp. 1423–1427, May 2016.

[10] S. Khan, H. Ali, Z. Ullah, N. Minallah, S. Maqsood, and A. Hafeez, "KNN and ANN-based recognition of handwritten pashto letters using zoning features," *arXiv*, vol. 9, no. 10, 2019, doi: 10.14569/IJACSA.2018.091070.

[11] M. Effendi, M. Jannah, and U. Effendi, "Corn quality identification using image processing with k-nearest neighbor classifier based on color and texture features," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 230, no. 1, 2019, doi: 10.1088/1755-1315/230/1/012066.

[12] S. Sanjaya, M. L. Pura, S. K. Gusti, F. Yanto, and F. Syafria, "K-Nearest Neighbor for Classification of Tomato Maturity Level Based on Hue, Saturation, and Value Colors," *Indones. J. Artif. Intell. Data Min.*, vol. 2, no. 2, p. 101, 2019, doi: 10.24014/ijaidm.v2i2.7975.

[13] T. Dharani and I. L. Aroquiaraj, "Content Based Image Retrieval System using Feature Classification with Modified KNN Algorithm," 2013, [Online]. Available: http://arxiv.org/abs/1307.4717.

[14] S. Anraeni and Herman, "Hybrid lacunarity and euclidean distance algorithms for kidney health classification through iris image," *Int. J. Sci. Technol. Res.*, vol. 8, no. 11, pp. 486–488, 2019.