Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 4 (2011) 1296–1305

Procedia Computer Science

International Conference on Computational Science, ICCS 2011

Nearest Neighbor For Histogram-based Feature Extraction

F.S. Mohamad[a,*], A.A. Manaf[b], S. Chuprat[a,b]

[a]*University of Technology Malaysia, Faculty of Computer Science and Information Science, UTM International Campus, 54100, Kuala Lumpur, Malaysia*

[b]*University of Technology Malaysia, Avanced Informatic School (AIS), UTM International Campus, 54100, Kuala Lumpur, Malaysia*

[a,b,*]*University of Technology Malaysia, Avanced Informatic School (AIS), UTM International Campus, , 54100, Kuala Lumpur, Malaysia*

Abstract

Manual grading process of Fresh Fruit Bunch (FFB) leads to misconduct and human error while inspecting the right category of fruits for the purpose of oil palm production at the mill. It is extremely important to identify the degree of ripeness of FFB are at 95% level of confidence as mentioned by Malaysian Palm Oil Board (MPOB). Therefore, wrong evaluation of graded fruits will result wrong report regarding the oil content. However, the most critical part of oil palm grading is the fruit classification. Error in classifying the right category of FFB will cause error in estimating the oil content. Research done by Federal Land Development Authority (FELDA) at mills show the estimated oil content for ripe fruit is 60%, while underripe is 40% and unripe is only 20% minus water and dirt. This indicates the importance of the right classification of FFB during grading process is essential to prevent from mistakenly claim low quality fruits as the good ones. Problem will occur while receiving the grading report claiming the high percentage of Basic Extraction Rate (BER) by the appointed graders while they have been proven to be poor quality fruits during oil production process. Fruit ripeness identification based on color is hard to measure especially when it involves the color intensity. The most suitable color space must be carefully selected to determine the right color especially when the color intensity is involved. HSV color space has proven to be a good choice because it has all the colors in the channel. Besides, it also offers color intensity which can be in variety level of intensity degrees. This paper explores the use of Nearest Neighbor Distance for histogram-based fruit ripeness identification. Promising results are obtained when value elements of HSV gives the highest recognition rate towards both ripe and unripe category.

Keywords: Fresh Fruit Bunch, Fruit Ripeness Identification; HSV; Nearest Neighbor Distance

1. Introduction

An image histogram is type of histogram which acts as a graphical representation of the tonal distribution in a digital image [1]. It plots the number of pixels for each tonal value. By looking at the histogram for a specific image

F.S.Mohamad. Tel.: +6-019-906-6074; fax: +6-03-2693-0933.
E-mail address: fatma@unisza.edu.my.

Selection and/or peer-review under responsibility of Prof. Mitsuhisa Sato and Prof. Satoshi Matsuoka
doi:10.1016/j.procs.2011.04.140

a viewer will be able to judge the entire tonal distribution at a glance. In “Statistics in Psychology”, [2] explains “a histogram is a graphical display of tabulated frequencies which shown as bars. It also shows what proportion of cases fall into each of several categories. The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent. The intervals (or bands, or bins) are generally of the same size”.[3] in their paper explain histograms are basically summary of useful data that convey the general shape of the frequency distribution (normal, chi-square, etc.), the symmetry of the distribution and whether it is skewed and modality, unimodal, bimodal, or multimodal.

The histogram of the frequency distribution can be converted to a probability distribution by dividing the tally in each group by the total number of data points to give the relative frequency. Moreover [3] added that the shape of the distribution conveys important information such as the probability distribution of the data. In cases in which the distribution is known, a histogram that does not fit the distribution may provide clues about a process and measurement problem. For example, a histogram that shows a higher than normal frequency in bins near one end and then a sharp drop-off may indicate that the observer is "helping" the results by classifying extreme data in the less extreme group.

2. Related Work

According to [4], direct tissue profiling by MALDI-TOF-MS is used to identify biomarkers in clinical tissue samples. Tomato ripening is used as a model system to study the plant protein. Three different stages of tomato viz, unripened (UR), medium ripened (MR) and fully ripened (FR) were used for direct tissue profiling MALDI-MS analysis. It was observed that 34-kDa and 44-kDa proteins were up regulated during fruit ripening. The result suggest that pectinesterase and heterotrimetric GTP-binding protein fragment, are the ripening specific markers in tomato, since their levels were upregulated during tomato ripening. Meanwhile Random Amplification of Polymorphic DNA (RAPD) markers are used by [5] to differentiate between two types of jackfruit. Results of the RAPD analysis revealed that the two fruit types may be distinguished. However, an easy and effective identification for Sequence Characterized Amplified Region SCAR markers need to be addressed in the future. The mRNA Differential Display Technique which coupled with silver-staining is used by [6] to identify and isolate cDNAs for apple fruit ripening. The mRNA Differential Display technique coupled with silver-staining was used with success as a method to identify and isolate cDNAs representing transcripts differentially expressed during this developmental process. The results provide a contribution to better characterise the changes in gene expression that accompanies the ripening process in apple; a specie which lacks information about the molecular regulation of ripening.

MALDI-QIT-TOF MS is explored by [7] to identify a great number of betacyanins in crude extracts from Amaranthus tricolor seedlings, Gomphrena globosa flowers, and Hylocereus polyrhizus fruits. However, the related isomers should be differentiated with the aid of HPLC. In [11], a computer assisted photogrammetric methodology is developed to correlates color of oil palm fruits with their ripeness by calculating the color Digital Numbers (DN). The result of developing a complete automation grading system is achieved. However, the images taken a day after the fruits were delivered will cause the fruits color change, less freshness, and this will affect the percentage of oil content in the fruits. This is in parallel with [10] which stated that fruits must be graded within 24 hours after harvested. Another automated grading system is also developed by [12] based on RGB color model. The color elements of Red, Green and Blue were analyzed using this grading system. The mean color intensity is used to differentiate between different color and its ripeness. However, results are only limited to Ripe, Under Ripe and Over Ripe categories of fruits which are insufficient to detect for the ripeness without take into account for Unripe fruits. This is important as unripe is major category for fruit ripeness indicator.

3. Methodology

An ongoing study is conducted to use similarity measures for oil palm fruit application. The approach is intended to classify the ripeness of oil palm fruit based on histogram. We are going to explore the potential features from the color histogram and incorporate these features into the distance measurement. For this purpose of study, Nearest Neighbor Distance is chosen for its suitability for Histogram Distance. Histogram is explored for the ability of increasing the global contrast of many images, especially when the usable data of the image is represented by close

contrast values. Through this adjustment, the intensities can be better distributed on the histogram. This allows for areas of lower local contrast to gain a higher contrast without affecting the global contrast [8].

Histogram equalization accomplishes this by effectively spreading out the most frequent intensity values. As mentioned by [4,5,6,7,11,12] who studied in various applications, we are currently working on estimating the ranges of features extracted from the identified categories of oil palm fruits. Since there are not many studies done on this particular topic, especially in identifying the oil palm fruit ripeness by using histogram-based distance metric, we are going to explore the possibility of using histogram and make use of its features and tested by using Nearest Neighbor Distance. The ripeness bunches of fruits are identified based on prior knowledge. A data of approximately 30 bunches of fruits are collected for each category. The ranges are then computed for each fruit category. Sample of two category of oil palm fruits are shown below:

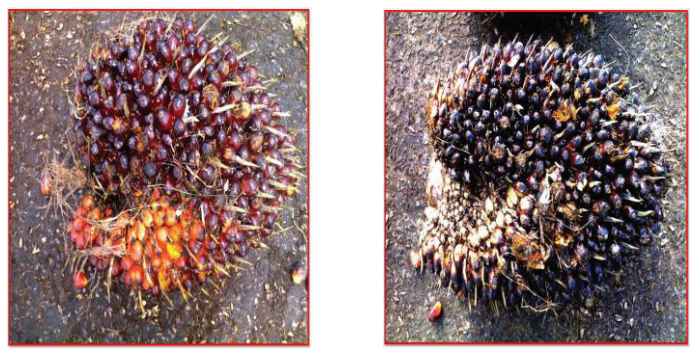

Fig. 1 (b) unripe. (a) ripe FFB

HSV color space is chosen because it has information about color in one channel [3]. HSV stands for Hue, Saturation and Value. Hue is the color itself. Saturation is the "quantity" of colors, meanwhile Value is a kind of brightness. Typical users perceive colors normally in the form of HSV color space. Besides, HSV color space is also chosen when you want to match for the right colors or to choose colors which looks similar.

As mentioned by [10], manual grading process by appointed grader will select about 50-100 bunches of FFB at random as sample from each consignment to be graded. This should represent top, middle and bottom portion of consignment. Problem will occur during grading process whenever human emotions involved, this may lead to mistakenly grade the low quality fruits as the good ones. Moreover, there might be some interference from the third party to influence the report from the grader for their own purpose.

Table 1.Estimated oil content

Fruit category	Estimated oil content
Ripe	60%
Underripe	40%
Unripe	20%z

Tab. 1 shows the estimated oil palm content for ripe, underripe and unripe category of FFB by research group of Federal Land Development Authority (FELDA) [12] at oil palm mills. The estimated of oil content is made after the consideration of the presence of water and dirt for every FFB. From the table, we can see that the ripe category of fruit provide the highest percentage of estimated of oil content and this indicate the importance of the right identification of FFB before it was taken for the sterilization process for oil production.

.

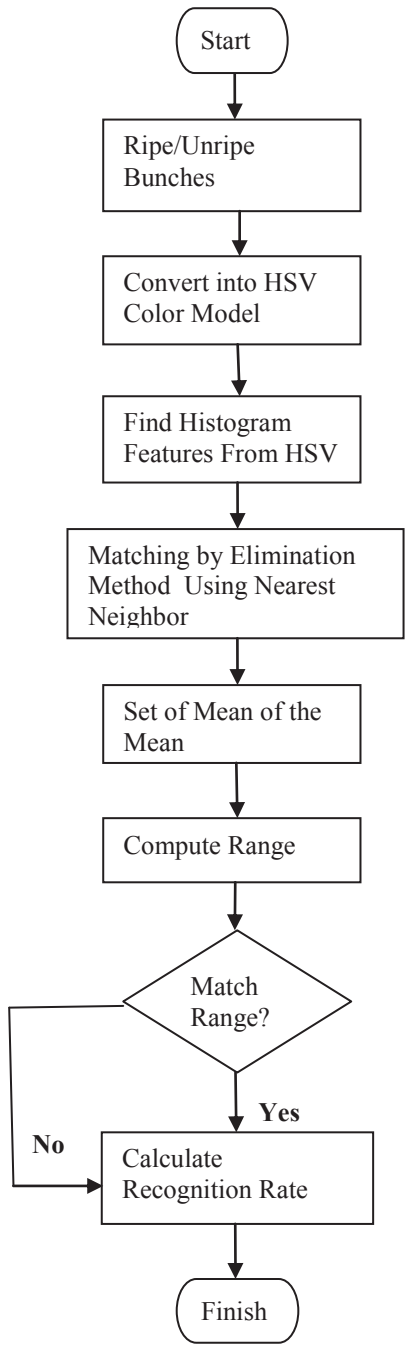

Fig.2. process flow

Referring to Fig.2., images of oil palm fruits in the form of JPEG format are captured at oil palm mill. Images are then converted to HSV color model from RGB color model. Then the histogram features are extracted for every element of Hue, Saturation and Value of HSV color model. Next, calculations are performed for the defined features extracted (mean of the mean value). Then, H, S, and V values are calculated for every bunch using Nearest Neighbor Distance. After that, a range of min and max values are also calculated for each H, S and V for every bunch of oil palm fruits for the ripe and unripe category. The same processes are then repeated for an unknown

sample of ripe fruits (30 bunches) and unripe fruits (30 bunches). Lastly, comparisons are made to match the unknown fruits with known category of fruits and the correct matches are then calculated

4. HSV Color Space

In image processing applications, there are various color spaces in use nowadays. Among the color spaces are RGB, HSV, HCI, HCL, CIE XYZ and many more. However, the two most popular color spaces are RGB and HSV. In RGB, there are primarily three color components: Red, Green and Blue. In addition, colors are not simply formed from these three fundamental colors. Each components of RGB are represented by pixel values and it is widely used in computer hardware like television screen or computer monitor as explained by [14]. In this study, we chose HSV color space because it offers color brightness that will be a good contribution to our research. This is in line with study conducted by [13] and [15] who experimented the features extracted from HSV color space, which decouples brightness from chromatic components, have demonstrated better performance than that from RGB color model. As mentioned by [15], HSV color space is used to describe perceptually relationship between colors more accurately compare to other color space. Basically, HSV color space has three components: Hue, Saturation and Value. Hue is the color attributes [13] which composed of variety of colors in the channel while saturation is the degree of color purity and value is the color intensity as described by [9]. In our previous work, we use RGB color space as an initial study. However, results obtained were not met with the objective because RGB color space cannot detect the color brightness in all the training images.

The HSV color model is defined as follows [16]:

$$H = \begin{cases} 60\frac{(G-B)}{R} & \text{if } MAX = R \\ 60\frac{(B-R)}{G} & \text{if } MAX = G \\ 60\frac{(R-G)}{B} & \text{if } MAX = B \\ \text{Not defined} \end{cases} \quad (1)$$

$$S \begin{cases} \frac{\delta}{MAX} & \text{if } MAX \neq 0 \\ 0 & \text{if } MAX = 0 \end{cases}$$

$$V = MAX$$

where δ = (MAX - MIN), MAX = max(R, G, B), and MIN = min(R, G, B). Note that the R, G, B values in Equation (1) are scaled to [0, 1]. In order to confine H within the range of [0, 360],

$H=H+360, \text{ if } H<0.$

5. Nearest-Neighbor Distance

We use nearest neighbor (NN) distance to compare the similarity between HSV elements of the histograms. In this study, histogram is used as a feature vector for feature extraction purpose. At this stage, mean value are extracted from the histogram for each H,S and V element. Then, Nearest Neighbor is used to compare between the histograms features using an elimination method for all the FFBs until a set of range value is obtained. The formula used is shown as below:

$$(h_1,h_2) = \sum_{i=n}^{n} \min(h_1,h_2) \quad (1)$$

h_1 is the known category of FFB while h_2 is the individual FFB which falls under unknown category. In this case, we have 4 different set of FFB bunches (30 bunches each) and the dataset obtained are from prior knowledge by appointed mill grader. The processes are divided into two stages. Details of the processes are explained below:

5.1 Known category

First, mean value is extracted from HSV histogram for ripe and unripe category of FFB. Total number of FFB used in this study is 30 for ripe and another 30 for unripe category. After that, we compare every FFB within the same category of fruits one by one until all 30 bunches using Nearest Neighbor Distance. Then, we calculate mean value for each matrix table for FFB 1 until FFB30. Since the comparison is made between FFB1 and FFB2, FFB1 and FFB3 until 30 bunches of fruits, the total number of FFB after the comparison is just 29. After that, for every mean value of 29 FFB, we compute the min and max value for the range. Then, mean of the mean value for every min and max for the total bunches are calculated for that particular category of each HSV elements.

5.2 Unknown category

For unknown category, we compare unknown set of 30 fruits for every FFB bunch using Nearest Neighbor Distance for every HSV element until mean is computed. Then, for every bunch of FFB, we calculate min and max to compute range value. After that, we match every bunch of FFB one by one with known range of particular category of fruits whether ripe or unripe. Finally, recognition rate is obtained for every HSV element.

6. Experimental Design

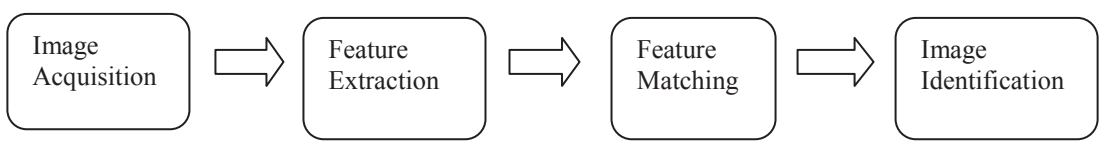

Fig.3. General Block Diagram

Fig.3 shows general block diagram for oil palm fruit ripeness identification. The processes start from Image Acquisition. At this stage, FFB images are taken using a digital camera by prior knowledge (appointed grader). Next, Feature Extraction processes from the histograms are obtained. After that, histogram features are compared using Nearest Neighbor and elimination technique for every FFB for Feature Matching process. Finally, Image Identification process is done and Recognition Rate is obtained.

The algorithm of the processes as below:

Step 1: Take 30 sample of ripe and another 30 sample of unripe oil palm fruits.

Step 2: Find mean value for each H, S, and V for every bunch of fruits

Step 3: Use Nearest Neighbor Distance and calculate min and max value for each H, S, and V for every bunch of fruits

Step 4: Calculate mean of min and mean of max value for each H, S, and V for every bunch of fruits

Step 5: Calculate mean of min and mean of max for each ripe and unripe category of oil palm fruits

Step 6: Take another sample of 30 bunches of unknown ripe fruits and another 30 bunches of unknown unripe fruits and repeat step 2 – step 5

Step 7: Match the result of unknown bunches of oil palm fruits with known ones. i.e. unknown ripe with ripe and unknown unripe with unripe

Step 8: Find the Recognition Rate or "match" rate as shown in Table 2 – 7.

7. Result and Analysis

Elimination method is used to compare between FFB for both category of fruits. Therefore, we are able to finalize the mean of the mean by using Nearest Neighbor Distance during elimination process. Based on the algorithm above, results are depicted in table 2 until 7 and divided into 2 categories which is Ripe and Unripe. Recognition rate is obtained and shown for Hue, Saturation and Value as below.

7.1. Ripe FFB

For ripe FFB, recognition rate for every element of Hue, Saturation and Value is shown in table 2-4. The process starts from comparing every unknown bunch of FFB with the range value obtained from Known category of ripe FFB.

Table 2. Hue ripe

	Hue
Total no. of fruits	29
Matched	11
Recognition rate	38%

Table 2 shows the recognition rate for Hue element of ripe category of oil palm fruits. From the table, 11 out of 29 bunches of FFB are matched within the range value of known ripe category and the recognition rate is 38%

Table 3. Saturation ripe

	Saturation
Total no. of fruits	29
Matched	15
Recognition rate	52%

Table 3 shows the Saturation element of the ripe category. Out of 29, 15 bunches of fruits are matched within the category and the recognition rate is slightly more than half, which is 52%.

Table 4. Value ripe

	Value
Total no. of fruits	29
Matched	26
Recognition rate	90%

In table 4, Value proved to be the best ripeness indicator which provide 90% recognition rate. Out of 29 bunches of FFB, 26 are matched. This provide good indicator for ripeness identification.

7.2. Unripe FFB

For unripe FFB, recognition rate is also obtained for every Hue, Saturation and Value element of FFB. In this process, individual unknown fruits are compared with range value obtained from known FFB.

Table 5. Hue unripe

	Hue
Total no. of fruits	29
Matched	7
Recognition rate	24%

Table 5 shows Hue result for the unripe category. From 29 bunches of FFB, only 7 bunches are matched and the recognition rate is 24%.

Table 6. Saturation unripe

	Saturation
Total no. of fruits	29
Matched	14
Recognition rate	48%

In table 6, from the total number of 29, 14 bunches are matched and this provide 48% recognition rate for the Saturation value.

Table 7. Value unripe

	Value
Total no. of fruits	29
Matched	24
Recognition rate	83%

Table 7 shows the highest recognition rate for Value (unripe category) which is 83%. Out of 29 bunches, 24 are

Table 7 shows the highest recognition rate for Value (unripe category) which is 83%. Out of 29 bunches of compared FFB, 24 are matched. This proved to be a significant indicator for ripeness identification.

Table 8. Unknown ripe vs. unknown unripe

	Hue	Saturation	Value
Ripe	38%	52%	90%
Unripe	24%	48%	83%

Table 8 shows the comparison result between unknown bunches which we assume ripe and unknown bunches which we assume unripe (based on prior knowledge). Overall result shows the recognition rate for Nearest Neighbor Distance falls under Value element of HSV color model. Value is proved to be a good indicator for ripeness identification. 90% of oil palm fruits are correctly identified as ripe and 83% for unripe. In manual grading process, colors are one of the main indicators to correctly identify whether the fruits are ripe or unripe. This is in line with the definition set by Malaysian Palm Oil Board (MPOB) which "ripe bunch is a fresh bunch which has reddish orange color" [10] which is hard to define since it involves color intensity. This is the main reason why HSV color space is chosen compare to RGB color space which is describing only on 3 primary colors without detailing on color intensity. Furthermore as described by [9], "value is the color lightness", this research proves to be very promising and encouraging research work for fruit ripeness identification

8. Conclusion and Future Work

Promising result in identifying ripe and unripe category of oil palm fruits is successfully obtained. The Value element of HSV color space defined by [9] as color lightness, has been proven to give the most distinctive difference between ripe and unripe category. Further investigation will look deep into color lightness as a major indicator. Nearest Neighbor proved as a good distance measurement for histogram-based features. Currently, we also experiment the applicability of Furthest Neighbor and Mean Distance and make comparison to see the best result. We will also try to combine the extracted features using PCA, K-SOM and few other techniques to be incorporated with existing Similarity Measurement Distance and find the best matching technique for fruit ripeness identification.

Acknowledgements

Special thanks to UTM International Campus and University of Sultan Zainal Abidin for the help and support of this research.

References

1. E. Sutton. "Histograms and the Zone System". Illustrated Photography. Accessed at http://www.illustratedphotography.com/photography-tips/basic/contrast
2. D.Howitt, and D.Cramer, "Statistics in Psychology". Prentice Hall.2008.
3. NetMBA. Internet Center for Management and Business Administration, Inc. The Histogram. Accessed at http://www.netmba.com/statistics/histogram/
4. K.Arvind.M, B Santhakumari and K.Mahesh.J., "Identification of specific proteins in tomato by intact tissue MALDI-TOF-MS,". Electronic Journal of Food and Plants Chemistry 3(1) 2008 10-13
5. D.K.N.G. Pushpakumara and S.A.Harris, "Potential of RAPD markers for identification of fruit types of *Artocarpus Heteropyllus Lam*. (jackfruit)," J. Nat.Sci.Foundation Sri Lanka 2007 35(3):175-179
6. G. Luis. F. and O. Cristina M., "Molecular identification of novel differentially expressed mRNAs up-regulated during ripening of apples," Plant Science Volume 172, Issue 2, February 2007, Pages 306-318

7. C.Yi-Zhang, X. Jie, S. Mei and C. Harold. “Rapid Identification of Betacyanins from Amaranthus tricolor, Gomphrena globosa, and Hylocereus polyrhizus by Matrix-Assisted Laser Desorption/Ionization Quadrupole Ion Trap Time-of-Flight Mass Spectrometry (MALDI-QIT-TOF MS),” j. Agric. Food Chem.2006.54, 6520-6526
8. “Histogram Equalization”. Wikipedia The Free Encyclopedia.2010.
9. SciVisHome. Color Principles - Hue, Saturation, and Value. 2000. Accessed at http://www.ncsu.edu/scivis/lessons/colormodels/color_models2.html#secondary
10. MPOB. Oil Palm Fruit Grading Manual. Second Edition.
11. A. Jaffar, R. Jaafar, N. Jamil, C. Y. Low and B. Abdullah. “Photogrammetric Grading of Oil Palm Fresh Fruit Bunches,” International Journal of Mechanical & Mechatronics Engineering IJMME Vol: 9 No: 10. 2009.
12. Felda Agricultural Services Sdn Bhd. Analisa Oil MPD Kilang-Kilang FPISB 2010.
13. Wen Chen, Yun Q. Shi and Guorong Xuan. Identifying Computer Graphics Using Hsv Color Model And Statistical Moments Of Characteristic Functions. ICME 2007. 1-4244-1017-7/07/$25.00 C2007 IEEE
14. Wikipedia. RGB Color Space. Accessed at http://www.fact-index.com/r/rg/rgb_color_space.html
15. Saed Mirghasemi and Ehsan Banihashem. Sea Target Detection Based on SVM Method Using HSV Color Space. IEEE Xplore 2009
16. A. R. Smith, "Color gamut transform pairs," Comput. Graph.12(3) (1978) 12-19.