A Hidden Markov Model Development for Star Fruit Sugariness Identification

Agus Buono[1], Nursidik Heru Praptono[1], Irmansyah[2], and Aziz Kustiyo[1]
[1]*Computer Sciences Department, Faculty of Mathematics and Natural Sciences, Bogor Agriculture University, Bogor-West Java, Indonesia*
[2]*Physics Department, Faculty of Mathematics and Natural Sciences, Bogor Agriculture University, Bogor-West Java, Indonesia*
Email: pudesha@yahoo.co.id

***Abstract*—This research is addressed the implementation of Hidden Markov Model (HMM) dedicated to identify the star fruit sugariness. This pixel based identification transforms the RGB (Red-Green-Blue) image into the total dissolved solids (TDS) for each pixel by using a linear regression equation, and categorized them into one of the six available Class. Then the sequence of pixels with one out of six available symbols becomes the input sequence. Based on this sequence, we identify the star fruit sugariness into one of the three categories by using HMM. Number of hidden state is varying from 3 until 8. This research shown that is a positive correlation between TDS with the Red component, and negative correlation with Green component. There is non significant correlation between TDS with the blue component. The best regression model that relating TDS with the RGB component is TDS=f(R,G) with 0.85 as determination coefficients (R^2). The accuracy of the HMM is varies from 69% to 75 %, with 73.5% as average. It seen that the classification is tend to the low class. It means, the misclassifications occur from class 2 classified into class 1, and class 3 classified into class 2. All objects in class 1 are classified into appropriate class.**

I. Introduction

Sorting fruit based on the condition that the consumer interest is part of post-harvest process, which classifies fruits that are relatively uniform in one group and the fruits of which differ greatly in other groups. The process helps in determining prices and distribution in accordance with the type of consumers who become the target sales. In general, consumers prefer the taste of star fruit sweet than sour. Fruit sweetness level is determined by the dissolved solid content, the higher content of dissolved solids, then the fruit is more sweet. This substance can be detected through laboratory testing. Normally, star fruit with a dominant red color will be much sweeter than the less red color. Therefore, the image color of the fruit is often used as a basis for sorting fruit.

In the automatic sorting fruit process, the RGB image of a certain fruit is read pixel by pixel and for each pixel is given a score of R, G or B according to the most dominant color. Furthermore, fruit classification determined by the number of pixels with a dominant red color, [1]. By this way, the position of pixels with a dominant red color ignored in the process of classification. Therefore, in this research the classification of fruits is done by considering the value of pixels as well as its position. The suitable technique for this approach is Hidden Markov Model. In this case each pixel is labeled based on RGB values. By using the sequence of the label, we construct HMM model for each class, which will then be used as a reference in identifying the level of sweetness for the new object.

The remainder of this paper is organized as follows: Section 2 presents the data preprocessing and experimental setup. Section 3 the principle of the HMM. Result and discussion is presented in Section 4, and finally, Section 5 is dedicated to the conclusions of this research.

II. Data Preprocessing and Experimental Setup

A. Data Preprocessing

TABEL I
DISTRIBUTION OF TDS CONTENT FOR EACH CLASS

Class of sweetness	TDS Content
Class 1 (less)	TDS<8.18
Class 2 (moderate)	8.18≤TDS≤9.70
Class 3 (high)	TDS>9.7

This research uses the RGB image data of star fruit (Averrhoa Carambola L) that consists of 3 classes, namely class with the sweetness of less (class 1), the class with the sweetness of moderate (class 2), and class with high sweetness (class 3). Each class consists of 45 images of size 192x256 pixels, and has known the value of total dissolved solids (TDS) through laboratory tests.

The higher content of TDS, the star fruit is more sweet and vice versa, the lower the TDS content, the fruit is more or less sweet (sour). Table 1 presents the range of total dissolved solids value of each class. While Figure 1 presents the boxplot of TDS content.

In this research, 75% (33 image for each class) of the data are used as training data and the remaining (25%, 12 image for each class) as testing data. The first stage of preprocess is resizing the image size with a value of 1 / 25 of the original by using the technique described in [2], such that the image size of 11x8 pixels. The next stage is to convert RGB values from 0-255 into 0-1 at each pixel with these formulas:

$$(1)\ r = \frac{R}{R+G+B},\ (2)\ g = \frac{G}{R+G+B},\ \text{and}\ (3)\ b = \frac{B}{R+G+B}$$

For each star fruit calculated the average r, g and b from a number of 11x8 = 88 pixels. Then, we develop the best regression model that connects the TDS as response variable with the average r, g and b as the predictors, TDS = f (r, g, b), using the existing image of star fruit in the training set.

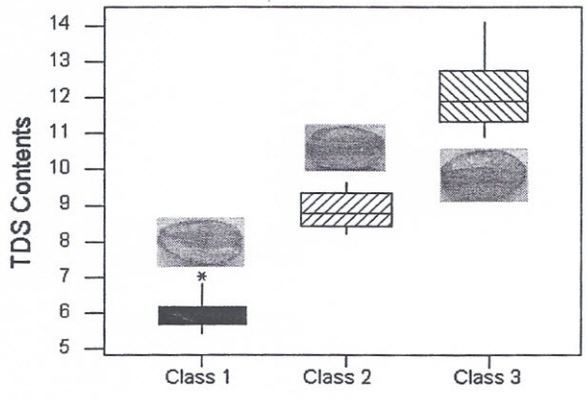
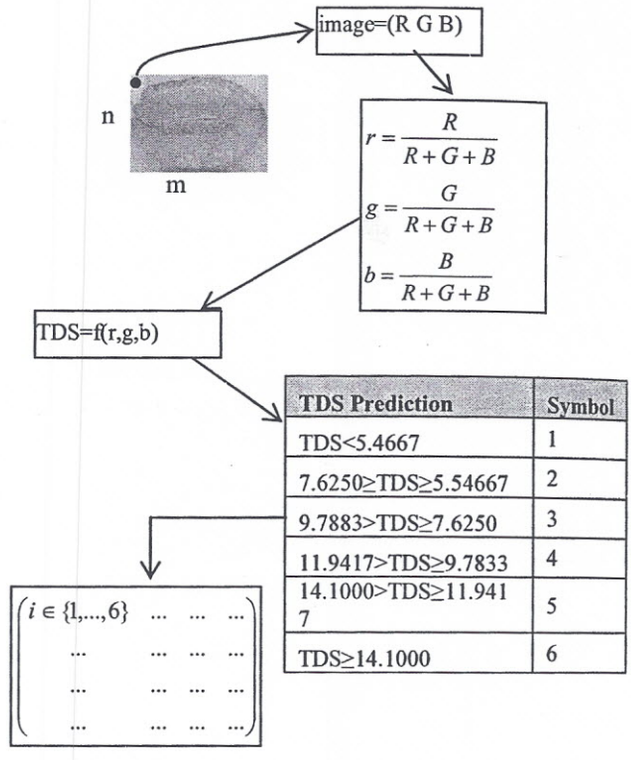


Fig. 1. Boxplot of the TDS content of the 3 classes of star fruit

Regression equation obtained from the above steps will then be used to map each pixel to the TDS domain, which then categorized into a symbol. Figure 2 presents the categorization of a pixel into one symbol of the 6 existing symbol.

B. Data Preprocessing

Figure 2 shows the flow diagram of the research. There are 4 stages, i.e. data preprocessing, regression modeling, HMM training parameters, and class identification. Detail process of the data preprocessing has been described in 2.A above. In this stage, we resize the image size and converting the RGB image into labeling image. It means, we give a label (1, 2, 3, …, or 6) to each pixel based on the TDS contents that predicted using regression equation with r, g and b as predictors. Then, we have a sequence of symbol 1, 2, 3, …, or 6 for an image. This process is done for all images in the training set (33 images for each class) and in the testing set (12 images for each class).



TDS Prediction	Symbol
TDS<5.4667	1
7.6250≥TDS≥5.54667	2
9.7883>TDS≥7.6250	3
11.9417>TDS≥9.7833	4
14.1000>TDS≥11.9417	5
TDS≥14.1000	6

Fig. 2. The process of pixels labeling

The 33 sequences in the training set for each class are combine together into one vector of 33x8x11=2904 symbols. Then, we use the vector to predict the HMM's parameters for each class, i.e. HMM1, HMM2, and HMM3 for class1, class2, and class 3, respectively, by using Baum-Welch Algorithm, [3]. In this research, we use the left-right HMM with number of hidden state are 3, 4, 5, …, 8. Finally we have 3x6=18 HMM. Due to the type of HMM, the training algorithm will predict the transition probability matrix (A) and the observation probability matrix (HMM) of the model. The stopping criteria is the absolute difference between elements of the matrix parameters for iteration t by iteration t +1. The algorithm will stop if the value smaller than 0.00001. The next stage is to validate the model using the testing data with the forward algorithm, [3].

III. THE LEFT-RIGHT HIDDEN MARKOV MODEL

Hidden Markov Model (HMM) is a temporal probability model that describes the relation between

state variables (hidden state) in the successive time index, and between state variables with observable variables. Visually, this model can be described using a finite state automata. There are two kinds of model, i.e. ergodic HMM and left-right HMM.

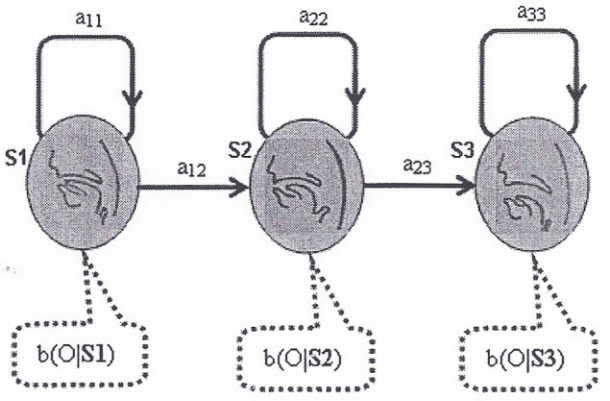


Fig. 3. Left-Right HMM with 3 hidden state

Figure 3 presents the left-right HMM with 3 hidden state. The parameters for the model, i.e. A, and B are:

$$A=\begin{pmatrix} a_{11} & a_{12} & 0 \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix} \quad B=\begin{pmatrix} b_1(1) & b_1(2) & \dots & b_1(K) \\ b_2(1) & b_2(2) & \dots & b_2(K) \\ b_3(1) & b_3(2) & \dots & b_3(K) \end{pmatrix}$$

The initial parameter, Л, for the model is [1 0 0]. In order to simplify the mathematical formulation, following are the notations on the HMM, [4]:

- A: matrix with elements $a_{ij}=P(q_{t+1}=j\ q_t=S_i)$, the probability in state S_j at time t+1, given at time t was in state S_i.
- B: matrix with elements $b_j(k)=P(V_K$ at time $t\ q_t=S_j)$, the probability of appearance of a symbol V_K, given the hidden state is S_j. K is the number of symbol in the observable set.
- Л: vector with element $л_i=P(q1=i)$, the probability in the early stage is in state i. In this case $\sum_{i=1}^{N}\pi_i = 1$, N is the number of hidden state.

There are three problems with the HMM, i.e. evaluation, decoding and learning. In this research, we use forward algorithm to evaluate the probability of a given observable sequence, and the Baum-Welch algorithm to train the HMM. Detail explanation of the HMM could be found in [3] and [4].

IV. Result and Discussion

Figure 4 presents the pattern of the relationship between TDS with color intensity (r, g and b). From the pictures can be seen that the intensity of green and red colors have a coefficient of determination 0.8584 (or correlation above 0.9) and 0505 (or correlation above 0.7), respectively. This shows that the intensity of the two colors can be used to predict the TDS content, which shows the level of star fruit sweetness. In this case the higher the intensity of the green, then the lower the TDS fruit, with a decrease of about 230 units for each increase of 1 unit of green intensity. The opposite occurs for red color. With the higher intensity of red color, the higher the TDS content, which increased by 125 times for every increase of 1 unit of red intensity. From the picture is also seen that the range of red and green intensity (34 - 39) were more dominant than blue intensity (26 - 33). The figure also shows the fact that the relationship between the intensity of blue color with the TDS is not as strong with two other colors (red and green).

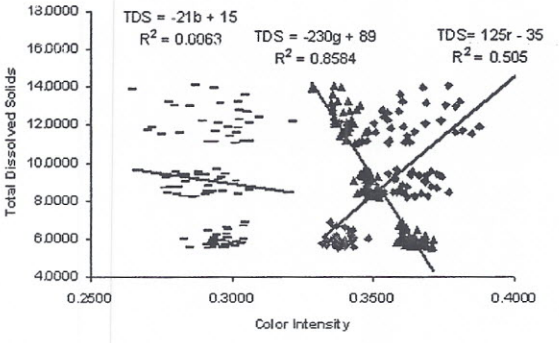


Fig. 4. TDS relationship patterns with color intensity

By considering the three patterns between TDS content and the intensity of color presented by Figure 4, then the predictors in linear regression models to estimate the value of TDS are the intensity of red color and intensity of green color. The regression equation is as follows:

$$TDS = 99.0 + 9.7r - 267g$$

The coefficients determination of the regression model is 85.9 . Based on the regression equation, the TDS content prediction is then performed for each pixel. Then the pixels are labeled by using the boundaries as given in Table 1. After that, we identify the class of each star fruit on the test data by using the HMM. Figure 5 presents the accuracy of the identification of star fruit with a number of hidden state 4, 5, 6, 7, and 8.

Figure 5 presents the accuracy of the system at different number of hidden state of the HMM. The average accuracy is ranging from 69 to 75 . From the picture we can see the fact that more and more number of hidden state, does not guarantee the accuracy higher.

The optimum number of hidden state is 4 or 7. This is around 9 over the accuracy obtained in [5], which only reached 66.9 .

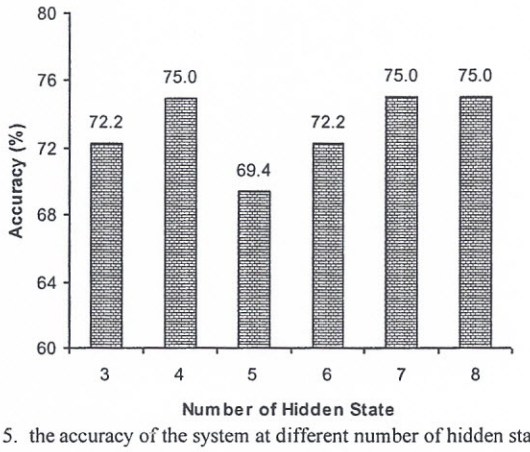


Fig. 5. the accuracy of the system at different number of hidden state

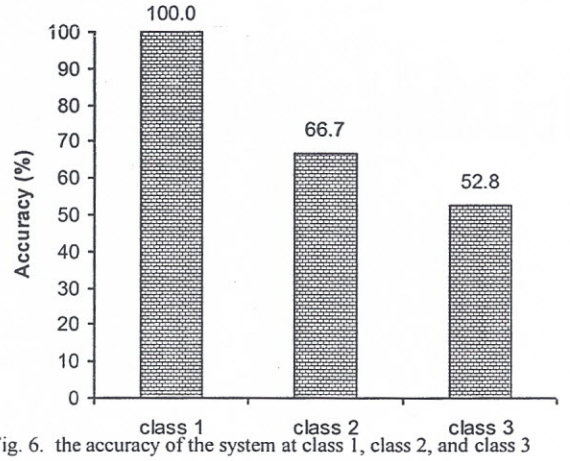


Fig. 6. the accuracy of the system at class 1, class 2, and class 3

Based on the accuracy for each class, we see that class 1 (class who have low TDS content) can be identified 100 . While class 2 and class 3 the average accuracy for each was 66.7 and 52.8 , as shown in Figure 6. From the picture is also seen that the system is less able to identify the fruit with moderate to high TDS content. It is also shown in Table 2. From the table shows that the wrong classification is to class with TDS content of one level below it, ie class 2 are classified into class 1 (33), class 3 was detected as a class 2 (42).

The facts can be explained by using Figure 7 which presents a scatter plot between the predicted TDS content (using the regression function) with the observation from the laboratory. The picture shows that the prediction is overestimate for the class 2. In this case the predicted value is greater than the true value. As for grade 3, in general, the prediction value is smaller than the true value (underestimate). These facts indicates that the error classification was not due to HMM model as a classifier, but in the feature extraction. There are two parts in feature extraction process, the first is a regression function that converts RGB values into the TDS content. The second is the boundaries in categorization the TDS content into the labels of 1, 2, 3, ..., or 6.

TABEL II
CONFUSSION MATRIX FOR HMM WITH 8 HIDDEN STATE (NUMBER OF OBJECTS FOR EACH CLASS IS 12)

Class		*Classifi ation*		
		1	*2*	
ser ation	*1*	12	0	0
	2	4	8	0
		0	5	7

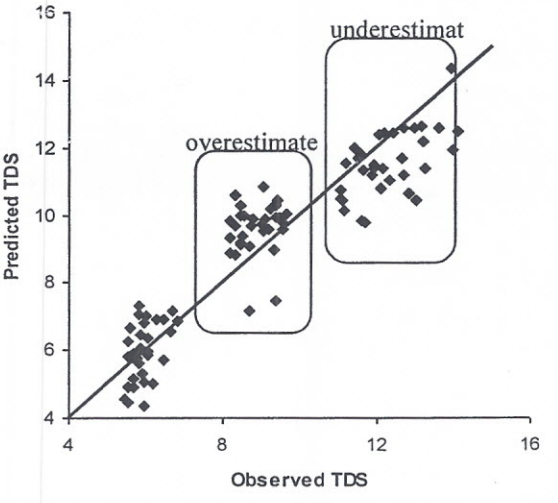


Fig. 7. Scatter plot of Predicted TDS with Observation TDS Content

Figure 8 presents the dotplot of TDS observations and predictions as well as class boundaries and TDS boundaries for the state categorization. It appears that the boundaries of TDS for state catogorization did not follow the pattern of distribution of the TDS observation. From the pictures can be seen that there is overlapping between the TDS results of prediction for the class 2 and class 3. This has become one of the reasons why errors occur between the two classes.

V. CONCLUSIONS

Based on the experiment, we can conclude several things:

a. Correlation between the r (red) and g (green) intensity with the content of TDS are 0.71 and 0.93 each for the red and green intensity, respectively.

Then, we use r and g intensity to predict the TDS content of the star fruit.

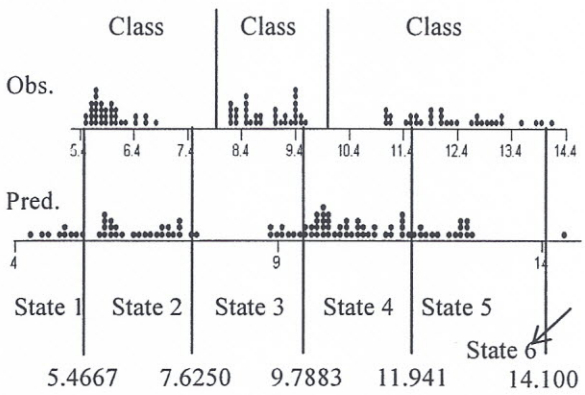


Fig. 8. Dotplot of the TDS observation and prediction combined with the boundaries

b. The best regression model with r and g as the predictors, TDS content as the response variable has a high coefficients of determination, i.e. 0.859. But, the model is not good enough. Because, the prediction is underestimate for the moderate TDS and overestimate for the high TDS.

c. The system accuracy is 75 in the average with number of hidden state of the HMM are 4, 7, or 8. It is caused by at least two problems in the feature extraction process, i.e. a regression function that predicting the TDS content based on the r and g intensity, and categorization boundaries that convert the TDS level to the one of the existing 6 symbols.

In order to increase the system accuracy, for future word, we will focused on the feature extraction process.

REFERENCES

[1] Irmansyah. *Star Fruit Classification Based on Color and Taste By Using Image Processing and Fuzzy Logic*. [PhD. Dissertation]. Bogor: Agriculture Engineering Study Program, Graduate School, Bogor Agricultural University, 2009.

[2] Gonzales, RC and Woods, RE. *Digital Image Processing 2nd Edition*. New Jersey: Prentice-Hall, 2002.

[3] Rabiner, RL, "A Tutotial on Hidden Markov Models and Selected Applications in Speech Recognition", In *Proceeding of the IEEE*. Vol.77, No. 2, 1989, pp. 257-283

[4] Dugad R. and Desai UB. *A Tutorial on Hidden Markov Models*. Technical Report, Department of Electrical Engineering, Indian Institute of Technology – Bombay, India, 1996.

[5] Buono, A., and Irmansyah., "Total Dissolved Solids of Star Fruit Identification Based on RGB Image by Using Principle Component Analysis as Feature Extraction and Euclidean Distance as Classifier", *Jurnal Ilmu Komputer dan Informasi*, Faculty of Computer Science, University of Indonesia, Vol. 2., No. 1., February 2009, pp. 35-41.