I.J. Image, Graphics and Signal Processing, **2020, 5, 37-43**
Published Online October 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2020.05.04

Variance Analysis Based Mango Recognition Using Correlation Distance

Farhana Tazmim Pinki, S.M. Mohidul Islam
Computer Science and Engineering Discipline, Khulna University, Khulna, Bangladesh
Email: farhana.kucse@gmail.com, mohid@cse.ku.ac.bd

Received: 20 December 2019; Accepted: 02 February 2020; Published: 08 October 2020

Abstract: Mango plays a major role in the Agro industry and it is a very popular fruit to most of the people due to its flavor and taste. There are many varieties of mangoes that are differentiable based on their various characteristics. Sometimes it is difficult and time consuming for general people or farmers to categorize the mango into different types due to intra-class variation among various types of mangoes. This paper has proposed an automatic system to recognize mangoes thus it becomes convenient to identify various types of mangoes. In this method, mangoes are recognized into different categories based on variance analysis or data dispersion measures. Measures include five number summary, variance, mean deviation, skewness, coefficient of variation which are used as features. From both training and query images, feature vectors are created. Correlation is used to recognize mangoes into various categories. The proposed method shows better result than some existing methods.

Index Terms: Image Processing, Variance, Correlation, Feature Vector, Mango Recognition.

1. Introduction

Mango is one of the most common and flavorsome fruit in most continents, particularly in Asia. It is a tropical fruit that matures on very large trees. Mangoes have a variation of colors including red, green, and yellow [1]. This fruit is native to India, Bangladesh, and Pakistan [1]. There are different varieties of mangoes in the world. These varieties are categorized by various specialized agronomists. But it is hard to identify mango types by common people. They usually consult with specialists, query with mango guidebooks or browse relevant web pages through keywords searching to know the names or characterizations of mango [2].

Nowadays it is a very difficult task of computer vision to recognize object from image. An automatic classification system can solve the problem of object classification. Mango classification is not a very easy task because different varieties of mangoes contain intra-class similarity and inter-class variation. Some mangoes are well formed and some are deformed. So it is a challenging task to recognize various types of mangoes automatically.

The main objective of the proposed system is to develop an automatic mango recognition system using image processing and data mining techniques. The pospoded method highlights the classification issue because different types of mangoes has different market demand. Variance analysis or data dispersion measures are used for feature extraction and correlation is applied for mango recognition. The proposed system is able to correctly classify various types of mangoes which helps the farmer of rural areas without the help of agriculturalists. Some papers, such as [9] performed mango grading only based on quality, maturity, and size but did not classify mango. Some papers [3][7] performed the classification works but their accuracy is lower than the proposed method.

The remainder of this paper is organized as follows. In section II, some existing works are discussed. The proposed method is described in detail in section III, experimental results are presented in section IV, and finally, the paper ends with the conclusion in section V.

2. Related Works

Several types of researches are related to mango classification. Behera et al. [3] classified different types of mangoes like Langra, Amprapali, Himsagar, Kesar. For this, at first, the images were preprocessed by median filter. Then K-means clustering was used for image segmentation and features were extracted from the segmented images. Contrast, correlation, homogeneity, energy, mean, standard, deviation, entropy, RMS, variance, smoothness, kurtosis, skewness, and IDM were used as features. For classification, they used multiclass SVM. The methodology acquired around 90% accuracy.

Dameshwari et al. [4] developed a methodology for identifying defect and maturity of mangoes. They performed several preprocessing steps including background subtraction and RGB to Gray conversion. They implemented two different algorithms for defect detection and maturity specification. For defect detection, the ratio of the defected area was calculated from binary image and a certain threshold was set for making decision. For maturity identification, the contour line was extracted and the decision was made about the maturity of mangoes from matrix difference.

Anurekha et al. [5] used GANFIS (Genetic Adaptive Neuro Fuzzy Inference System) technique for grading of mangoes and identifying them efficiently. At first, adaptive median filter was applied to eliminate the noise from the images. The proposed GANFIS based algorithm used genetic algorithm to read the input images and extracted the shape, size, and texture features. These extracted features were used to make covariance matrix and LBP (Local Binary Pattern). These features were used to develop neural network for classification and the fuzzy rules are applied on each level of neurons.

Roomi et al. [6] proposed an algorithm for classifying several types of mangoes such as Totapuri, Alfonza, and Rumani. They converted the RGB images into HSV color spaces and Region of Interest was extracted from the converted image using Otsu's method. Different features like translation, rotation and scale invariant shape which are time invariant features were extracted from image. The research mainly focused on object contour modeling which was used for automatic selection of apriori probability. Some region based descriptors such as major axis length, minor axis length were also extracted and from them, eccentricity and area to square perimeter were calculated. Bayes classifier is used for classification.

Abbas et al. [7] developed a mango classification system using shape and texture feature. They also used MaZda package along with the B11 program [8] which can be used for texture analysis and visualization. The set of different mango types with different lengths and widths were trained by the B11 model. Then Region Of Interest (ROI) technique was used to extract various texture features and then, data processing was performed through Lobe Component Analysis (LCA), Linear Discriminant Analysis (LDA) and Nonlinear Discriminant Analysis (NDA) to extract texture feature. These features were stored in the B11 database. Finally the classification was performed using size, texture and shape features along with the B11 program. They used total 22 mangos of 7 types for training and 6 images for testing. The accuracy of their system was 83%.

Rashmi Pandey et al. [9] proposed a system which is divided in two halves: First part discussed selecting healthy mangoes and then classifying it into ripe and unripe category. Second part talked about grading mangoes based on its size. CIELAB color model with Dominant density range method was used for color feature extraction which easily discriminate color and classify healthy and diseased mangoes. Then Healthy mangoes were classified into ripe and unripe category using the same method. Then size measure was evaluated using fuzzy expert system for grading of mango. Size feature was calculated using ellipse properties in order to classify in different grades. At final stage, size feature was fed to fuzzy expert system for grading. The whole system achieved 97.47% average accuracy.

3. Proposed Method

The proposed system focuses on recognizing various types of mango. The steps that are followed to execute the proposed method is shown in the following:

Step 1: Load the image dataset.

Step 2: Find the data dispersion measure by extracting five number summary, variance, mean deviation, skewness, coefficient of variation of the image and create feature vector.

Step 3: Store the features in a database.

Step 4: Load stored feature values which are stored in database of Step 3.

Step 5: Load the query image folder.

Step 6: Extract feature from query image and create feature vector as described in Step 2.

Step 7: Find the correlation between images in the dataset and query images using their extracted features.

Step 8: Select the class of the image of the dataset as the predicted class of the query image which is mostly correlated with the query image.

The system architecture of the proposed method is shown in Fig. 1 and we describe in the following how the steps of our proposed method is performed for automatic recognition of various mango types.

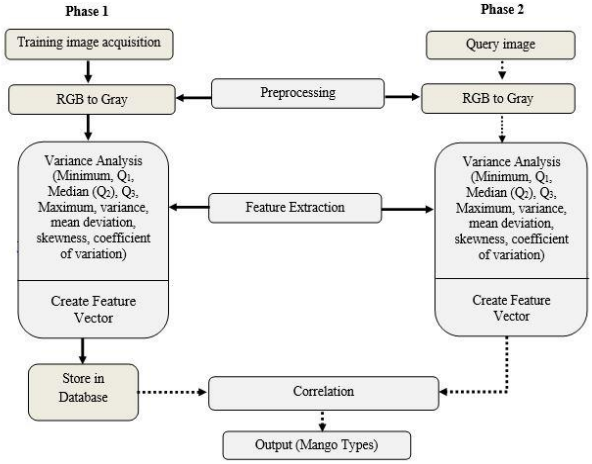

Fig.1. System architecture of of the proposed system

The proposed system inovolves two phases. In first phase, features of the mangoes in the dataset are extracted and stored in database which is used in the second phase to find the correlation with the feature of the query image.

A. Phase 1

Some common steps such as image acquisition, preprocessing, and feature extraction are the parts of both phases. These steps are described in the following.

1) Image Acquisition

The mango images are collected from field and some images are collected from Internet. The collected images are taken as input image in RGB (Red, Green, and Blue) form. Different types of mango samples are shown in Fig. 2.

a b c

Fig.2. Sample of (a) Badami (b) Kesar (c) Totapuri mango types

2) Image Preprocessing

The images which are obtained from image acquisition may not be suited for next step. So preprocessing is necessary. The captured images are in RGB color space which is one of the models of color images [10]. Since RGB color space depends on the individual devices, so it is mapped to grayscale image.

3) Feature Extraction

Feature plays an important role in describing a large set of data. It reduces the amount of resources required to describe a large set of data. For this proposed mango recognition system, data dispersion measures are used as features. These measures are used to assess the variance or spread of numeric data [11]. It includes five number summary, variance, mean deviation, skewness, coefficient of variation.

Five Number Summary: The five number summary of a distribution consists of the median (Q2), the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order of Minimum, Q1, Median, Q3, and Maximum. Quantiles are points taken at regular intervals of a data distribution and divide it into essentially equal-size consecutive sets. The 2-quantile data point divides the lower and upper halves of the data distribution. It refers to the median. The 4-quantiles are the three data points and they split the data distribution into four equal parts, each part represents one-fourth of the data distribution. They are more commonly referred to as quartiles. The 100-quantiles are known to as percentiles and they divide the data distribution into 100 equal-sized consecutive sets. The median, quartiles, and percentiles are the most commonly used forms of quantiles. The quartiles represent an indication of a distribution's center, spread, and shape.

The 1st quartile (Q1) is the 25th percentile. It cuts off the 25% of the data. The 3rd quartile (Q3) is the 75th percentile. The 2nd quartile (Q2) is the 50th percentile, also known as median.

Variance: The variance of N observations, $x_1, x_2, \ldots x_n$ for a numeric attribute X is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \quad (1)$$

Mean deviation: The mean deviation is defined is the arithmetic mean of the absolute deviation from the means and calculated as

$$mean\ deviation = \frac{\sum_{i=1}^{N}|x_i - \mu|}{N} \quad (2)$$

where μ is the arithmetic mean of the values and N is the total number of values. This value will be greater for distributions with a larger spread.

Skewness: A common measure of skewness

$$skewness = \frac{x - mode}{s} \quad (3)$$

This indicates how far (in standard deviation, s) the mean (x) is from the mode and whether it is greater or less than the mode.

Coefficient of variation: The coefficient of variation is the standard deviation expressed as a percentage of arithmetic mean and is calculated as

$$Coefficient\ of\ variation = \frac{s}{x} \times 100 \quad (4)$$

The variability in groups of observations with widely differing means can be compared using this measure.

4) Creating Feature vector and store in database

After extracting features using data dispersion measures, feature vector is created for each image and store these in a database.

B. Phase 2

The preprocessing and feature extraction of query images are same as phase 1. A feature vector is created using data dispersion measure and it is used to identify the class of the query image by comparing it with image in the database.

A correlation is a statistical measure of the association between two variables. The measure is top used in variables that prove a linear relationship between each other [12]. The correlation coefficient that indicates the strength of the relationship between two images is found using the following formula.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (5)$$

where r_{xy} is the correlation coefficient of the linear relationship between the known image in the daatset, x, and the query image, y. x_i is the values of the image x in a sample, $\bar{x}$ is the mean of the values of x, y_i is the values of the image y in a sample, $\bar{y}$ is the mean of the values of y.

4. Experimental Results

The experiment is performed to recognize the mango into three types: Badami, Kesar, and Totapuri. We have used total 580 samples of mango images from the dataset [13] where 151 samples belong to Badami, 136 samples belong to Kesar, and 193 samples belong to Totapuri. So, the training set contains total 480 samples of mango images. In preprocessing steps, the images are resized into 640×480 pixels and converted from RGB to Gray. Then features are extracted using variance analysis. Here we extract 9 features and they are five-number summary (minimum, Q1, median (Q2), Q3, and maximum), variance, mean deviation, skewness, coefficient of variation. A feature vector is created using these features and they are stored in a database. To evaluate the performance of the proposed method, 100 samples of mango images are used for testing. The query images also consist of three categories where 40 images are Badami, 25 images are Kesar, and 35 images are Totapuri. According to the phase 1, the preprocessing and feature extraction phases are performed for the query images. Finally, the correlation measure is used to recognize mango. This measure is computed for feature vectors of all images of the database with feature vector of query images. The class of the most

correlated image from the training images is considered as the predicted class of the query image. The User Interface (UI) of whole process of experiment is shown in Fig. 3.

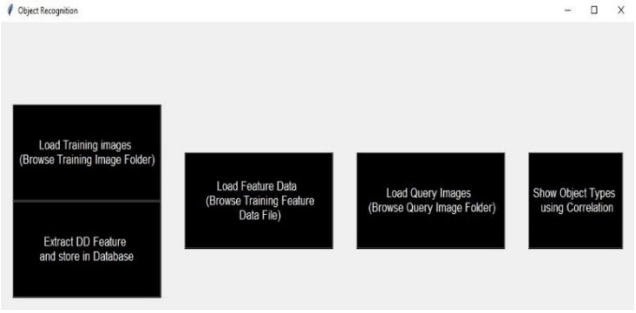

Fig.3. User interface of the proposed method

Some samples of training image features which are extracted using variance analysis and saved in excel file are shown in Fig. 4. The mango types for all query images are also saved in excel file. The output for some sample query images is shown in Fig. 5.

	A	B	C	D	E	F	G	H	I	J	K
1	Label	Minimum	Q1	Median	Q3	Maximum	Variance	MeanDev	Skewness	CoefficientOfVariation	
2	Badami100.jpg	47	164	194	206	230	704.4628	22.0451	-0.76131	0.142325	
3	Badami101.jpg	14	169	215	228	254	1669.277	33.44802	-1.06608	0.203138	
4	Badami102.jpg	23	125	157	199	235	1373.557	34.89334	-0.08556	0.228473	
5	Badami103.jpg	10	188	219	231	254	1757.718	33.81547	-1.19779	0.205887	
6	Badami104.jpg	40	201	210	218	247	921.1901	24.14187	-1.24312	0.152405	
7	Badami105.jpg	50	197	207	215	234	987.0257	25.02469	-1.25493	0.160177	
8	Badami106.jpg	19	150	197	208	245	1076.776	29.03633	-0.67785	0.180193	
9	Badami107.jpg	32	165	208	218	249	1125.134	29.16413	-0.87219	0.174025	
10	Badami108.jpg	14	149	194	207	247	1358.347	31.65984	-0.95526	0.206321	
11	Badami109.jpg	43	147	195	211	238	1219.521	30.84252	-0.54876	0.191607	
12	Badami110.jpg	23	182	203	211	252	795.1679	22.97775	-1.09032	0.146649	
13	Badami111.jpg	32	176	208	215	232	959.1942	25.38442	-1.16479	0.158807	
14	Badami112.jpg	30	185	202	211	253	913.7104	24.12882	-1.19003	0.159149	
15	Badami113.jpg	27	129	175	198	226	1319.193	33.45118	-0.36346	0.222185	
16	Badami114.jpg	26	192	213	221	255	925.7302	24.53754	-1.23524	0.15172	
17	Badami115.jpg	25	181	209	218	254	843.8217	23.82326	-1.07019	0.146634	
18	Badami116.jpg	22	182	214	223	250	942.9436	25.14831	-1.19678	0.152236	
19	Badami117.jpg	26	193	211	220	250	863.1499	23.83618	-1.17574	0.146601	
20	Badami118.jpg	24	194	215	224	254	944.2107	24.83037	-1.27058	0.151131	

Fig.4. Sample features of training images

	A	B	D
1	TestImageName	ObjectType	
2	Badami1.jpg	Badami83.jpg	
3	Badami10.jpg	Badami79.jpg	
4	Badami11.jpg	Badami106.jpg	
5	Badami147.jpg	Badami145.jpg	
6	Badami148.jpg	totapuri176.jpg	
7	Badami149.jpg	totapuri110.jpg	
8	Badami150.jpg	Badami135.jpg	
9	Badami151.jpg	Badami191.jpg	
10	Badami152.jpg	Badami135.jpg	
11	Badami153.jpg	Badami22.jpg	
12	Badami158.jpg	Badami125.jpg	
13	Badami159.jpg	Badami172.jpg	
14	Badami160.jpg	Badami173.jpg	
15	Badami161.jpg	Badami125.jpg	
16	Badami162.jpg	Badami165.jpg	
17	Badami163.jpg	Badami127.jpg	
18	Badami164.jpg	Badami134.jpg	
19	Badami175.jpg	Badami125.jpg	
20	Badami2.jpg	Badami89.jpg	

Fig.5. Some sample output of query images

The performance of the proposed system is measured from confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known [14]. From a confusion matrix TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) values can be found. The diagonal values of the confusion matrix represent the number of images that are correctly recognized and non-diagonal values represent the images that are misclassified [15]. The values of confusion matrix are calculed from the output file which is shown partly in Fig. 5. Table 1 shows the confusion matrix of our mango dataset.

Table 1. Confusion Matrix

Type of Mango	Badami	Kesar	Totapuri
Badami	38	0	2
Kesar	0	25	0
Totapuri	7	0	28

The individual success rate of various types of mango is shown in Fig. 6.

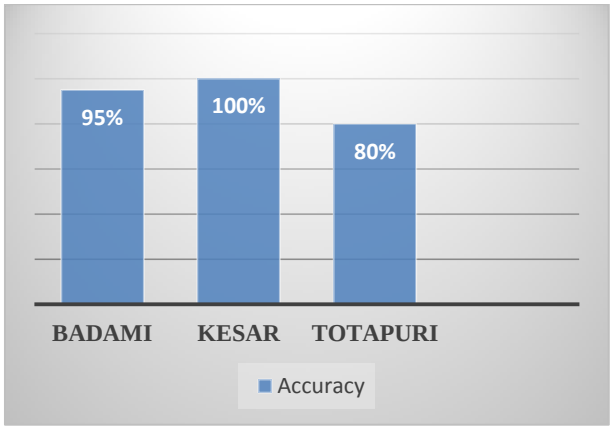

Fig.6. Performance measure of various types of mango recognition

The success rate and misclassification rate of the proposed method is evaluated from the following.

$$success\ rate = \frac{TP+TN}{TP+FP+TN+FN} \tag{6}$$

$$\text{misclassification rate} = 100 - success\ rate \tag{7}$$

Using (6) we get the following success rate of our system for recognizing various types of mango:

$$Accuracy = \frac{38+25+28}{40+25+35} \times 100\% = 91\%$$

So, we can say that the overall accuracy of our proposed method is 91% and misclassification rate is 9%, which is better than some exsiting methods such as [3][7].

5. Conclusion

Image processing is a vast area which is growing day by day. It is really helpful for automatic object detection and classification of different objects. Recognition of three types of mangoes viz Badami, Kesar, and Totapuri is presented in the experiments of the proposed method. This recognition process is done based on variance analysis and correlation, and the result shows better results than some existing mthods in case of mango classification. It will be very helpful for farmers to identify different types of mangoes efficiently and accurately. We have experimented on three categories of mangoes, in future, the number of mango samples and more categories of mangoes can be added to expand the area of identification of mangoes.

Acknowledgment

This research work is funded by Information and Communication Technology (ICT) Division, Ministry of Post, Telecommunication, and Information Technology, Government of the People's Republic of Bangladesh through ICT fellowship.

References

[1] The Top Mango Producing Countries in the World, https://www.worldatlas.com/articles/the-top-mango-producing-countries-in-the-world.html, Accessed on 9 April 2019.

[2] Rohan Sriram, Amar Tejas M, Prof. J. Girija, "Mango Classification using Convolutional Neural Networks", *International Research Journal of Engineering and Technology (IRJET)*, Vol. 5, Issue. 11, 2018.

[3] Behera, Santi Kumari, et al., "Automatic Classification of Mango Using Statistical Feature and SVM." *Advances in Computer, Communication and Control*. Springer, Singapore, pp. 469-475, 2019.

[4] Sahu, Dameshwari, and Ravindra Manohar Potdar, "Defect identification and maturity detection of mango fruits using image analysis.", *American Journal of Artificial Intelligence*. Vol. 1, No. 1, pp. 5-14, 2017.

[5] R Anurekha, D., and R. A. Sankaran, "Efficient classification and grading of MANGOES with GANFIS for improved performance.", *Multimedia Tools and Applications*, pp. 1-16, 2019.

[6] Roomi, S. Mohamed Mansoor, et al., "Classification of mangoes by object features and contour modeling." *2012 International Conference on Machine Vision and Image Processing (MVIP)*. IEEE, 2012.
[7] Abbas, Q., Iqbal, M. M., Niazi, S., Noureen, M., Ahmad, M. S., Nisa, M., & Malik, M. K.: Mango Classification Using Texture & Shape Features. *International Journal of Computer Science and Network Security*, 18(8), pp. 132-138, 2018.
[8] MaZda Web Site. In: Eletel.p.lodz.pl, http://www.eletel.p.lodz.pl/programy/mazda/index.php?action=mazda, Accessed on 29 June 2019.
[9] Pandey, Rashmi, Nikunj Gamit, and Sapan Naik, "A novel non-destructive grading method for Mango (Mangifera Indica L.) using fuzzy expert system.", *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE*, 2014.
[10] Mishra, Alok, Pallavi Asthana, and Pooja Khanna., "The quality identification of fruits in image processing using Matlab.", *International Journal of Research in Engineering and Technology*, Vol. 3, No. 10, pp. 92-95, 2014.
[11] Han, J., Pei, J., & Kamber, M.: Data Mining Concepts and Techniques. 3rd edn. *Elsevier*, 2011.
[12] Correlation - Overview, Formula, and Practical Example, https://corporatefinanceinstitute.com/resources/knowledge/finance/correlation/, Accessed on 5 July 2019.
[13] Mango Dataset - Studio Setup, https://data.mendeley.com/datasets/fmfncxjz3v/1, Ac-cessed on 27 March 2019.
[14] Confusion Matrix in Machine Learning - GeeksforGeeks. In: GeeksforGeeks. https://www.geeksforgeeks.org/confusion-matrix-machine-learning/, Accessed on 9 June 2019.
[15] Singla, A., & Garg, M.: CBIR approach based on combined HSV auto correlogram color moments and Gabor wavelet. *International Journal of Engineering and Computer Science*, 3(10), pp. 9007-9012, 2014.

Authors' Profiles

Farhana Tazmim Pinki is a MSc student at the Computer Science and Engineering Discipline, Khulna University, Bangladesh. She is now working as a Faculty member of Norther University of Business and Technology Khulna, Bangladesh. Her research interests include Machine learning, Data Mining, and Digital Image Processing.

S.M. Mohidul Islam is an Associate Professor at the Computer Science and Engineering Discipline, Khulna University, Bangladesh. He received his B.Sc. Engg. and M.Sc. Engg. degree from Khulna University. His research interests include Community ICT, Machine learning, Data Mining, Pattern Recognition, and Digital Image Processing.

How to cite this paper: Farhana Tazmim Pinki, S.M. Mohidul Islam, " Variance Analysis Based Mango Recognition Using Correlation Distance", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.12, No.5, pp. 37-43, 2020.DOI: 10.5815/ijigsp.2020.05.04