Random forests based classification for crops ripeness stages

Esraa Elhariri[1,6], Nashwa El-Bendary[2,6], Aboul Ella Hassanien[3,6], Amr Badr[3], Ahmed M. M. Hussein[4], Vaclav Snasel[5]

[1] Faculty of Computers and Information, Fayoum University, Fayoum - Egypt
[2] Arab Academy for Science,Technology, and Maritime Transport, Cairo - Egypt
[3] Faculty of Computers and Information, Cairo University, Cairo - Egypt
[4] Faculty of Agriculture, Minia University, Minya - Egypt
[5] Electrical Engineering & Computer Science,VSB-TU of Ostrava, Ostrava - Czech Republic
[6] Scientific Research Group in Egypt (SRGE), http://www.egyptscience.net

Abstract. This article presents a classification approach based on random forests algorithm for estimating and classifying the different maturity/ripeness stages of two types of crops; namely tomato and bell pepper (sweet pepper). The proposed approach consists of three phases that are pre-processing, feature extraction, and classification phases. Surface color of tomato and bell pepper is the most important characteristic to observe ripeness. So, the proposed classification system uses color features for classifying ripeness stages. It implements principal components analysis (PCA) along with support vector machine (SVM) algorithms and random forests (RF) classifier for features extraction and classification of ripeness stages, respectively. The datasets used for experiments were constructed based on real sample images for both tomatoes and bell pepper at different stages, which were collected from farms in Minya city, Upper Egypt. Datasets of total 250 and 175 images for tomato and bell pepper, respectively were used for both training and testing datasets. Training dataset is divided into five classes representing the different stages of tomato and bell pepper ripeness. Experimental results showed that SVM with Linear Kernel function achieved accuracy better than RF.

Keywords: image classification; features extraction; ripeness; principal component analysis (PCA); tomato; bell pepper; support vector machine (SVM); random forests (RF)

1 Introduction

Agriculture crops play a crucial role in the life of an economy. It is the backbone of the country's economic system because, Agriculture is one of the prime sources of the countries' national income. One of the prime factors in ensuring consistent marketing of crops is the product quality. For many crops, the main indicator of product quality from the customer's perspective is crop ripeness. Also, one of the most worrying issues for producers is the product's appearance as it has a high influence on product's quality and consumer preferences [1, 3]. This is why, produce (fruits and vegetables) ripeness monitoring and controlling has become a very important issue in the crops' industry. However, up to this day, optimal harvest dates and prediction of storage life

are still mainly based on subjective interpretation and practical experience [1]. Hence, automation of the ripeness assessment process is a big gain for agriculture and industry fields. For agriculture, an automatic harvest system may be developed. Also, it may be used for the purpose of saving crops from damages caused by environmental changes. For industry, it may be used for many purposes such as developing an automatic sorting system or checking the quality of fruits system to increase customer satisfaction level [2, 1, 3]. So, to ensure an optimum yield of high quality products, an objective and accurate ripeness assessment of agricultural crops is required. Moreover, identifying physiological and harvest maturity of agricultural crops correctly, will ensure timely harvest to avoid cutting of either under or over-ripe agricultural crops [1, 4]. There are one or more apparent sign, which every fruit shows when it reaches different ripeness levels. For Tomato and Bell Pepper, Surface color is the most common characteristic for ripeness determination. So color is used as a major method in determining maturity/ ripeness of tomato and bell pepper [1].

Recently, utilizing computer vision in food products has become very widespread, especially for products where measuring color or other spectral features enable estimating the ripeness stage [6]. This article presents a classification approach based on random forests algorithm for estimating and classifying the different maturity/ripeness stages of two types of crops; namely tomato and bell pepper (sweet pepper). Also, this article presents a comparative analysis of experimental results of applying Support Vector Machine (SVM) and Random Forests classifiers for estimating the ripeness level of tomato and bell pepper (sweet pepper) via investigating and classifying the different maturity/ripeness stages based on the color features.

The rest of this article is organized as follows. Section 2 introduces related research work. Section 3 presents the core concepts of SVM and RF algorithms. Section 4 describes the different phases of the proposed content-based classification system; namely pre-processing, feature extraction, and classification phases. Section 5 discusses the tested image dataset and presented the obtained experimental results. Finally, Section 6 presents conclusions.

2 Related Work

This section reviews current approaches tackling the problem of fruits/vegetables ripeness monitoring and classification.

In [1], authors present a content-based image classification system for the problem of estimation the ripeness levels of tomato via investigating and classifying the different maturity/ripeness stages.They use color features for classifying ripeness stages because color of tomato surface is the most important indicator for tomato ripeness. Principal Component Analysis (PCA) algorithm is applied feature extraction in order to generate a feature vector for each image in the dataset. Then, Support Vector Machine (SVM) algorithm were applied for classification of ripeness stages. While in [3] they applied their system in order to estimate the ripeness level of bell pepper (sweet pepper). they applied the one-against-one multi-class SVM approach for classification of ripeness stages. This approach achieved an efficiency of 93.89%.

Also, in [7], authors proposed an approach for classifying the ripeness stages of Apple fruit, this approach based on color image segmentation and fuzzy logic technique. the proposed approach depends on the mean values of RGB color components. They use fuzzy logic system for ripeness classification purpose.

Furthermore, in [8] an image processing based approach is designed for cherry sorting and grading system. The proposed approach depends on RGB color components of the captured images of cherry. The proposed, sorting system of cherries used color criteria and the TTS (Total Soluble Solids) in fruit to classify it to the right ripeness stage. This system achieved 92% accuracy in sorting cherries according to their ripeness.

Also, in [9] authors proposed an approach for lime maturity and ripeness identification. This approach based on image processing and Artificial Neural Networks . The used characteristics for ripeness identification in this approach are area, shape factor , RGB color index and texture features of lime. This approach achieved 100% accuracy in classifying the maturity and ripeness of lime

3 Preliminaries

3.1 Color features

A widely used feature in image retrieval and image classification problems is the color, which is as well an important feature for image representation [10]. Moreover, the first three color moments, which are mean, standard deviation, and skewness [10, 11], have been proved to be efficient and effective way for representing color distribution in any image.

Mean, standard deviation, and skewness for a colored image of size $N \times M$ pixels are defined by the following equations (1)-(3).

$$\bar{x}_i = \frac{\sum_{j=1}^{M \cdot N} x_{ij}}{M \cdot N} \quad (1) \qquad \partial_i = \sqrt{\frac{1}{M \cdot N} \sum_{j=1}^{M.N} (x_{ij} - \bar{x}_i)^2} \quad (2)$$

$$S_i = \sqrt[3]{\frac{1}{M \cdot N} \sum_{j=1}^{M.N} (x_{ij} - \bar{x}_i)^3} \quad (3)$$

where x_{ij} is the value of image pixel j of color channel i (e.g RGB, HSV and etc..), $\bar{x}_i$ is the mean for each channel i=(H,S and V), ∂_i is the standard deviation, and S_i is the skewness for each channel [10, 11]. On the other hand, colored histogram is a color descriptor that shows representation of the distribution of colors in an image. It represents the number of pixels that have colors in each range of colors [10].

3.2 Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a Machine Learning (ML) algorithm that is used for classification and regression of high dimensional datasets with great results [12–14]. SVM solves the classification problem via trying to find an optimal separating hyperplane between classes. It depends on the training cases which are placed on the

edge of class descriptor this is called support vectors, any other cases are discarded [13–15].

SVM algorithm seeks to maximize the margin around a hyperplane that separates a positive class from a negative class [12–14]. Given a training dataset with n samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where x_i is a feature vector in a v-dimensional feature space and with labels $y_i \in -1, 1$ belonging to either of two linearly separable classes C_1 and C_2. Geometrically, the SVM modeling algorithm finds an optimal hyperplane with the maximal margin to separate two classes, which requires to solve the optimization problem, as shown in equations (4) and (5).

$$maximize \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j . K(x_i, x_j) \tag{4}$$

$$Subject - to : \sum_{i=1}^{n} \alpha_i y_i, 0 \leq \alpha_i \leq C \tag{5}$$

where, α_i is the weight assigned to the training sample x_i. If $\alpha_i > 0$, x_i is called a support vector. C is a regulation parameter used to trade-off the training accuracy and the model complexity so that a superior generalization capability can be achieved. K is a kernel function, which is used to measure the similarity between two samples. .

3.3 Random Forests (RF)

The Random Forests (RF) is one of the best known classification and regression techniques, which has the ability to classify large dataset with excellent accuracy. Random Forests algorithm generates an ensemble of decision trees. Ensemble methods main principle is to group weak learners together to build a strong learner [20, 21]. The input is entered at the top and as it traverses down the tree, the original data is sampled in random, but with replacement into smaller and smaller sets. The class of sample is determined using random forests trees, which are of an arbitrary number. RF algorithm can be performed by applying the following steps [20]:

- Step 1: Draw N_{tree} bootstrap samples from the original data.
- Step 2: For each of the bootstrap samples, grow an un-pruned classification or regression tree.
- Step 3: At each internal node, rather than choosing the best split among all predictors, randomly select m_{try} of the M predictors and determine the best split using only those predictors.
- Step 4: Save tree as is, alongside those built thus far(Do not perform cost complexity pruning).
- Step 5: Predict new data by aggregating the predictions of the N_{tree} trees.

The predictions of the Random Forests are taken to be the majority votes of the predictions of all trees for classification and for regression are taken to be the average of the predictions of the all trees as shown in equation (6) [20, 21]:

$$S = \frac{1}{K} \sum_{K=1}^{K} K^{th} \tag{6}$$

Where S is a random forests prediction, K^{th} is a tree response, and k is the index runs over the individual trees in the forest.

4 The Proposed Classification System

In particular, the proposed framework is capable of recognizing the different ripeness stages of tomato and bell pepper, as shown in figure 1. This paper presents a ripeness classification system for tomato and bell pepper.

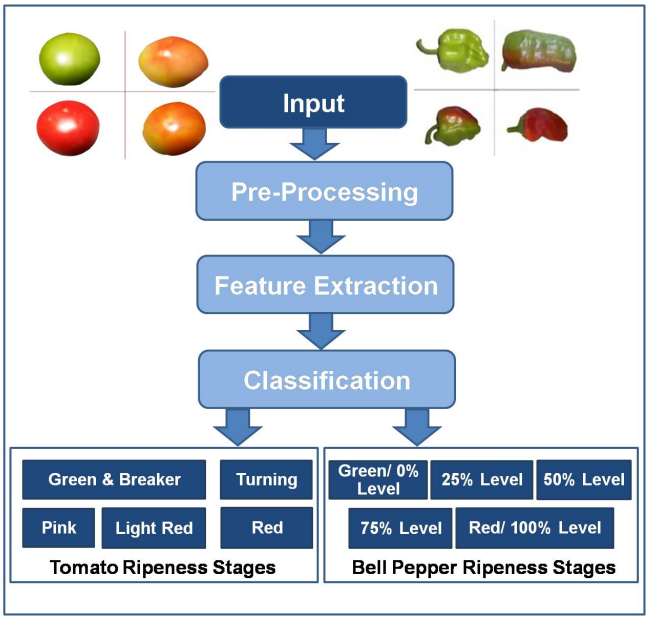


Fig. 1. Architecture of the proposed ripeness classification approach

The proposed system is composed of three main phases; namely ***Pre-processing***, ***Feature Extraction***, and ***Classification*** phases.

The datasets used for experiments were constructed based on real sample images for tomato and bell pepper at different ripeness stages, which were collected from farms in Minya city. Collected datasets contain colored JPEG images of resolution 3664×2748 pixels that were captured using Kodak C1013 digital camera of 10.3 mega pixels resolution.

4.1 Pre-processing phase

During pre-processing phase, the proposed approach resizes images to 250×250 pixels, in order to reduce their color index, and the background of each image will be removed

using background subtraction technique. Figure 2 shows an example of the background removal algorithm. Also, each image is converted from RGB to HSV color space as it is widely used in the field of color vision and close to the categories of human color perception [15].

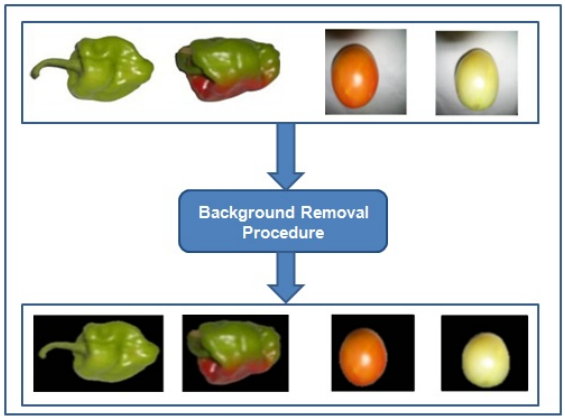


Fig. 2. Sample images before and after applying background subtraction algorithm

4.2 Feature extraction phase

As previously stated, since tomato and bell pepper surface color is the most important characteristic to asset the ripeness, this system uses HSV colored histogram and color moments for ripeness stages classification. For this phase, Principal Component Analysis (PCA) algorithm is applied as features extraction technique in order to generate a feature vector for each image in the dataset.

PCA is known as a statistical common technique, which is widely used in image recognition and compression for a dimensionality reduction, data representation and feature extraction tool as it ensures better classification [16–19]. It transforms the input space into sub-spaces for dimensionality reduction by discarding all ineffective minor components), after completing the previous 1D $16 \times 4 \times 4$ HSV histogram, 16 level for hue and 4 level for each of saturation and value. In addition, nine color moments, three for each channel (H, S and V channels) (mean, standard deviation, and skewness), will be computed. Then, a feature vector will be formed as a combination of HSV 1D histogram and the nine color moments.

4.3 Classification phase

Finally, for classification phase, the proposed approach applied two different algorithms for classification of ripeness stages SVM and RF. For SVM, The inputs are training

dataset feature vectors and their corresponding classes, whereas the outputs are the ripeness stage of each image in the testing dataset. For RF, The inputs are number of trees, training dataset feature vectors and their corresponding classes, whereas the outputs are the ripeness stage of each image in the testing dataset.

Since SVM is a binary class classification method and our problem is an N-class classification problem, so in this research the SVM algorithm is applied to Multi-class problem [22, 23]. We used one-against-one approach to do that.

SVM was trained and tested using (Linear and Polynomial with $order = 3$ kernel functions) and cross-validation.

5 Experimental Results

Simulation experiments in this article used a dataset of total 250 and 175 images for tomato and bell pepper, respectively for both training and testing datasets with 10-fold cross-validation. Bell pepper ripeness stages are shown in figure 3, while Tomato ripeness stages are shown in figure 4 [5, 24, 25].

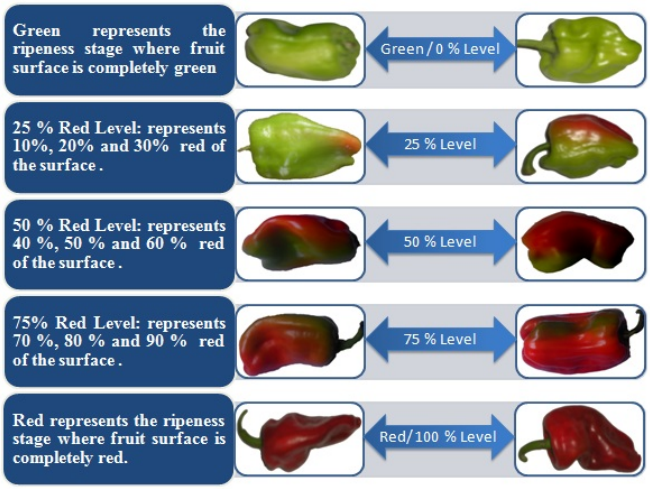


Fig. 3. Different ripeness stages of bell pepper

Training dataset is divided into five classes representing the different stages of tomato and bell pepper ripeness. The classes for bell pepper are *Green*, *25 % Level*, *50 % Level*, *75 % Level*, and *Red* stages, and for tomato are *Green & Breaker*, *Turning*, *Pink*, *Light Red*, and *Red* stages . Some samples of both training and testing datasets for both tomatoes and bell peppers are shown in figure 5.

The proposed approach has been implemented considering the One-against-One multi-class SVM system and RF using 10-fold cross validation and a total of 175 and

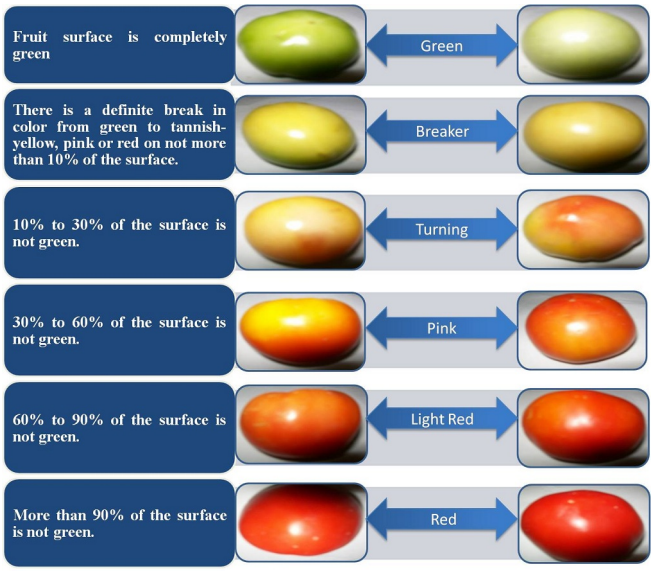


Fig. 4. Different ripeness stages of tomato

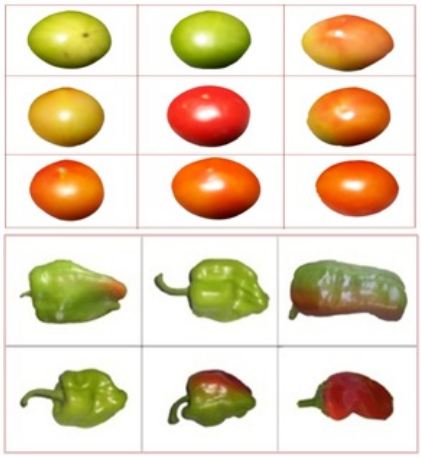
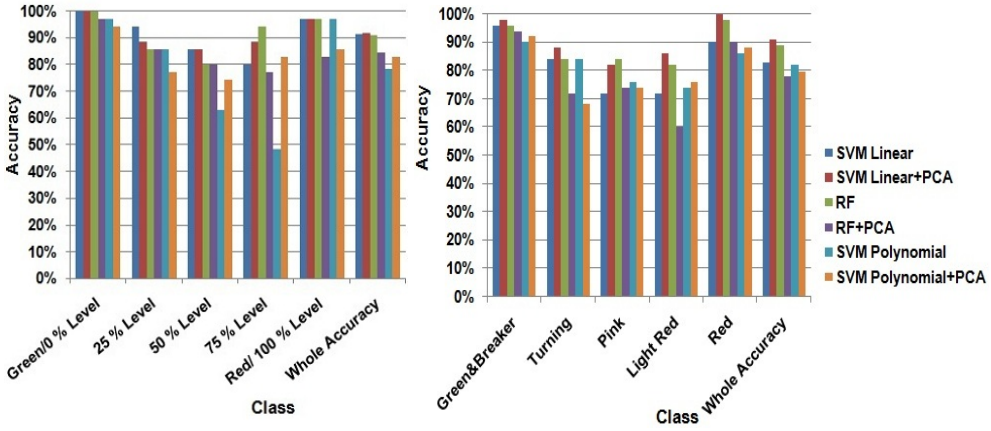


Fig. 5. Examples of training and testing samples

250 images of bell pepper and tomato, respectively for both of training and testing datasets. The used features for classification are a combination of colored HSV histogram and color moments. Moreover, SVM algorithm was employed with different kernel functions that are: *Linear* kernel and *Polynomial* kernel [26, 27] for ripeness stage classification.

Figure 6 shows classification accuracy for bell pepper and tomato ripeness obtained via applying each kernel function of SVM and RF.



(a) Bell Pepper Ripeness Stages (b) Tomato Ripeness Stages

Fig. 6. Comparative Study Results between SVM and RF Classifier for The Classification Accuracy

The accuracy is computed using equation (7):

$$Accuracy = \frac{Numberofcorrectlyclassifiedimages}{Totalnumberoftestingimages} * 100 \tag{7}$$

In this comparative study, we tested SVM and RF classifiers considering two scenarios: Firstly, the two classifiers were applied directly to the color features without applying PCA as features extraction technique. The result showed that both SVM and RF achieved good accuracy, SVM with linear kernel function, then RF and finally SVM with Polynomial kernel function, respectively. Secondly, the two classifiers were applied to features extracted by PCA as features extraction technique. The result showed that SVM achieved the best accuracy, while the accuracy of RF isn't good. This is because of the working method of RF described above. RF achieves high accuracy with huge amount of dataset which has high number of variables.

6 Conclusions

In this article, a comparative study between Support Vector Machine (SVM) and Random Forests (RF) Classifiers for the problem of estimating the ripeness level of tomato and bell pepper (sweet pepper) has been presented. The proposed system has three main stages; pre-processing, feature extraction and ripeness classification. The proposed classification approach was implemented by applying resizing, background removal, and extracting color components for each image. Then, feature extraction was applied to

each pre-processed image, HSV histogram and color moments are obtained as a feature vector, and used as a PCA inputs for transformation. Finally, SVM and RF models are developed for ripeness stage classification. The proposed approach has been implemented via applying One-against-One multi-class SVM system and RF system using10-fold cross-validation. Based on the obtained the experimental results, SVM with Linear Kernel function achieved accuracy better than RF.

References

1. Esraa Elhariri, Nashwa El-Bendary, Mohamed Mostafa M. Fouad, Jan Plato, Aboul Ella Hassanien, Ahmed M. M. Hussein, "Multi-class SVM Based Classification Approach for Tomato Ripeness", *Innovations in Bio-inspired Computing and Applications, Advances in Intelligent Systems and Computing* (Springer) Volume 237, pp. 175-186, (2013).
2. J.Brezmes, E.Llobet, X.Vilanova, G.Saiz, and X.Correig,"Fruit ripeness monitoring using an electronic nose", *The Journal of Sensors and Actuators B-Chem*, vol. 69,no.3, PP. 223-229, (2000).
3. Esraa Elhariri, Nashwa El-Bendary, Ahmed M. M. Hussein, Aboul Ella Hassanien and Amr Badr, "Bell Pepper Ripeness Classification based on Support Vector Machine, " in *The 2nd International Conference on Engineering and Technology (ICET 2014)*, Cairo, Egypt, (2014).(Accepted)
4. Z. May and M. H. Amaran, "Automated ripeness assessment of oil palm fruit using RGB and fuzzy logic technique, "in *Proc. the 13th WSEAS international conference on Mathematical and computational methods in science and engineering (MACMESE'11)*, Metin Demiralp, Zoran Bojkovic, and Angela Repanovici (Eds.), Wisconsin, USA, (2011), pp. 52–59.
5. Andrs F. Lpez Camelo,"Manual for the preparation and sale of fruits and vegetables From field to market", *Food ans Agriculture Organization (FAO) of the United Nations (UN)*, Agricultural Services Bulletin, Version 151, Rome (2004).
6. Francisco J. Rodrguez-Pulido, Beln Gordillo, M. Lourdes Gonzlez-Miret, Francisco J. Heredia, "Analysis of food appearance properties by computer vision applying ellipsoids to colour data", *Computers and Electronics in Agriculture* 99 (2013) 108115.
7. Meenu Dadwal and V.K.Banga, "Estimate Ripeness Level of fruits Using RGB Color Space and Fuzzy Logic Technique". In*International Journal of Engineering and Advanced Technology (IJEAT)*,2012 , 02 (01), 225-229.
8. Asghar Mousavi balestani ,Parviz Ahmadi Moghaddam, Asaad Modares motlaq and Hamed Dolaty , "Sorting and Grading of Cherries on the Basis of Ripeness, Size and Defects by Using Image Processing Techniques". In *International Journal of Agriculture and Crop Sciences(IJACS)*.2012 , 4 (16), 1144-1149.
9. hami Johar Damiri and Cepy Slamet,"Application of Image Processing and Artificial Neural Networks to Identify Ripeness and Maturity of the Lime(citrus medica)".In *INTERNATIONAL JOURNAL OF BASIC AND APPLIED SCIENCE* ,2012, 01 (02), 171-179.
10. A. Shahbahrami, D. Borodin, and B. Juurlink, "Comparison between color and texture features for image retrieval, "in *Proc. 19th Annual Workshop on Circuits, Systems and Signal Processing (ProRisc 2008)*, Veldhoven, The Netherlands, (2008).
11. S. Soman, M. Ghorpade, V. Sonone, and S, Chavan, "Content Based Image Retrieval using Advanced Color and Texture Features, "in *Proc. International Conference in Computational Intelligence (ICCIA2012)*, New York, USA, (2012).
12. Q. Wu and D.-X. Zhou, "Analysis of support vector machine classification, "*J. Comput. Anal. Appl.*, vol. 8, pp. 99–119, (2006).

13. H. M. Zawbaa, N. El-Bendary, A. E. Hassanien, and A. Abraham, "SVM-based Soccer Video Summarization System, "in *Proc. The Third IEEE World Congress on Nature and Biologically Inspired Computing (NaBIC2011)*, Salamanca, Spain, (2011), pp. 7–11.
14. H. M. Zawbaa, N. El-Bendary, A. E. Hassanien, and T. H. Kim, "Machine Learning-Based Soccer Video Summarization System, "in *Proc. Multimedia, Computer Graphics and Broadcasting FGIT-MulGraB (2)*, Jeju Island, Korea, (2011), Springer, vol. 263, pp. 19–28.
15. S.R. Suralkar and A. H. Karode, and P. W. Pawade, "Texture Image Classification Using Support Vector Machine, "*International Journal of Computer Applications in Technology*, vol. 3, no. 1, pp. 71–75, (2012).
16. M. Suganthy and P. Ramamoorthy, "Principal Component Analysis Based Feature Extraction, Morphological Edge Detection and Localization for Fast Iris Recognition, "*Journal of Computer Science*, vol. 8, no. 9, pp. 1428–1433, (2012).
17. Ada and RajneetKaur, "Feature Extraction and Principal Component Analysis for Lung Cancer Detection in CT scan Images, "*International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 3, (2013).
18. N. El-Bendary, H. M. Zawbaa, A. E. Hassanien, and V. Snasel, "PCA-based Home Videos Annotation System, "*The International Journal of Reasoning-based Intelligent Systems (IJRIS)*, vol. 3, no. 2, pp. 71–79, (2011).
19. B. Xiao, "Principal component analysis for feature extraction of image sequence, "in *Proc. International Conference On Computer and Communication Technologies in Agriculture Engineering (CCTAE)*, Chengdu, China, (2010), vol. 1, pp. 250–253.
20. Vrushali Y Kulkarni and Pradeep K Sinha, "Efficient Learning of Random Forest Classifier using Disjoint Partitioning Approach", in*Proceedings of the World Congress on Engineering*. Vol. 2. 2013.
21. Anna Bosch, Andrew Zisserman and Xavier Munoz, "Image Classification using Random Forests and Ferns,"in *IEEE 11th International Conference on Computer Vision*, 2007.
22. Y. Liu and Y. F. Zheng, "One-against-all multi-class SVM classification using reliability measures, "in *Proc. IEEE International Joint Conference on Neural Networks (IJCNN'05)*, Montreal, Quebec, Canada, (2005), vol. 2, pp. 849–854.
23. G. Anthony, H. Gregg, and M. Tshilidzi, "Image Classification Using SVMs: One-against-One Vs One-against-All, "in *Proc. of the 28th Asian Conference on Remote Sensing*, 2007.
24. Fox, A. J., Del Pozo-Insfran, D., Lee, J. H., Sargent, S. A., and Talcott, S. T., "Ripening-induced Chemical and Antioxidant Changes in Bell Peppers as Affected by Harvest Maturity and Postharvest Ethylene Exposure, "*AMERICAN SOCIETY FOR HORTICULTURAL SCIENCE, HORTSCIENCE*, vol. 40, no. 3, pp. 732-736, (2005).
25. Antoniali, Silvia, Leal, Paulo Ademar Martins, Magalhes, Ana Maria de, Fuziki, Rogrio Tsuyoshi, and Sanches, Juliana, "Physico-chemical characterization of 'Zarco HS' yellow bell pepper for different ripeness stages, "*Sci. agric. (Piracicaba, Braz.)*, vol. 64, no. 1, pp. 19-22, (2007) .
26. B. Vanschoenwinkel and B. Manderick. "Appropriate kernel functions for support vector machine learning with sequences of symbolic data, "*Deterministic and Statistical Methods in Machine Learning, Lecture Notes in Computer Science*, vol. 3635, pp. 256–280, 2005.
27. D. Boolchandani, and Vineet Sahula, "Exploring Efficient Kernel Functions for Support Vector Machine Based Feasibility Models for Analog Circuits, " *Int. Journal of design , analysis, and tools for circuits and systems*, vol. 1(1), pp. 1–8, JUNE 2011.