Expert Systems with Applications xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Using machine learning techniques for evaluating tomato ripeness

Nashwa El-Bendary [a,d,*], Esraa El Hariri [b,d], Aboul Ella Hassanien [c,d], Amr Badr [c]

[a] Arab Academy for Science, Technology, and Maritime Transport, Cairo, Egypt
[b] Faculty of Computers and Information, Fayoum University, Egypt
[c] Faculty of Computers and Information, Cairo University, Egypt
[d] Scientific Research Group in Egypt, Egypt

A R T I C L E I N F O

Article history:
Available online xxxx

Keywords:
Image classification
Features extraction
Ripeness
Principal component analysis (PCA)
Support Vector Machines (SVMs)
Linear Discriminant Analysis (LDA)

A B S T R A C T

Tomato quality is one of the most important factors that helps ensuring a consistent marketing of tomato fruit. As ripeness is the main indicator for tomato quality from customers perspective, the determination of tomato ripeness stages is a basic industrial concern regarding tomato production in order to get high quality product. Automatic ripeness evaluation of tomato is an essential research topic as it may prove benefits in ensuring optimum yield of high quality product, this will increase the income because tomato is one of the most important crops in the world. This article presents an automated multi-class classification approach for tomato ripeness measurement and evaluation via investigating and classifying the different maturity/ripeness stages. The proposed approach uses color features for classifying tomato ripeness stages. The approach proposed in this article uses Principal Components Analysis (PCA) in addition to Support Vector Machines (SVMs) and Linear Discriminant Analysis (LDA) algorithms for feature extraction and classification, respectively. Experiments have been conducted on a dataset of total 250 images that has been used for both training and testing datasets with 10-fold cross validation. Experimental results showed that the proposed classification approach has obtained ripeness classification accuracy of 90.80%, using one-against-one (OAO) multi-class SVMs algorithm with linear kernel function, ripeness classification accuracy of 84.80% using one-against-all (OAA) multi-class SVMs algorithm with linear kernel function, and ripeness classification accuracy of 84% using LDA algorithm.

1. Introduction

Fruits and vegetables development is characterized by a short period of cell division followed by a longer period of cell elongation by water uptake. The final fruit size mainly depends on initial cell number, rather than cell size (Cowan, Cripps, Richings, & Taylor, 2001). Fruit ripening, on the other hand, is characterized by the development of color, flavor, texture and aroma. The actual time from anthesis until full maturity can vary tremendously among species/cultivars due to genetic and environmental differences. Even between fruit on the same plant, fruit development and ripening can take more or less time depending on local microclimate conditions and differences in sink/source relations within the plant. In addition, when a fruit is harvested, the time of anthesis of a particular fruit is generally unknown, as is its full history (El Hariri, El-Bendary, Hassanien, & Badr, 2014; Lang & Hübert, 2012; Wei, Liu, Qiu, Shao, & He, 2014).

According to (FAOSTAT Database) Food and of the United Nations (FAO-UN) (2012), tomatoes world production was about 162 million tons fresh fruits produced in the year 2012 with income about 592 trillion dollars. Tomatoes are taking an important place among the fruits and vegetables all over the world, due to their continuously prevailing daily nutrition, dietary value and production income. Moreover, tomato is the fourth most important crop next to soybeans at world production (Camelo, 2004). Tomato production has been reported for 144 countries and Egypt occupies the fifth place in these countries at tomato production in both income and weight of fruit produced. Furthermore, tomato is the first most important crop in Egypt. Based on the previously stated facts, in this article we selected to focus on monitoring the ripeness stages of tomato crop, especially in Egypt. Another matter of fact is, on the categorization of crops into the two types of *climacteric* that are able to continue ripening after picking from the mother plant and *non-climacteric* that can ripe only when it is

* Corresponding author at: Arab Academy for Science, Technology, and Maritime Transport, Cairo, Egypt.

E-mail addresses: nashwa.elbendary@ieee.org (N. El-Bendary), eng.esraa.elhariri@gmail.com (E. El Hariri), aboitcairo@gmail.com (A.E. Hassanien), a.badr.fci@gmail.com (A. Badr).

URLs: http://www.egyptscience.net (N. El-Bendary), http://www.egyptscience.net (E. El Hariri), http://www.egyptscience.net (A.E. Hassanien).

http://dx.doi.org/10.1016/j.eswa.2014.09.057
0957-4174/© 2014 Published by Elsevier Ltd.

attached to the mother plant (Camelo, 2004; Coates & Johnson, 1997), tomato belongs to climacteric category as it can reach over-ripening stage after being harvested. Also, tomatoes have many different ripeness stages, which are (1) Green, (2) Breaker, (3) Turning, (4) Pink, (5) Light red and (6) Red stages, so they reach full red color even when harvested green. Red stage is the most preferred ripeness stage commercially (Camelo, 2004; El Hariri et al., 2014).

As has been noted, monitoring and controlling produce (fruits and vegetables) ripeness has become a very important issue in the crops industry, since ripeness is perceived by customers as the main quality indicator. Also, the product's appearance is one of the most worrying issues for producers as it has a high influence on product's quality and consumer preferences. However, up to this day, optimal harvest dates and prediction of storage life are still mainly based on subjective interpretation and practical experience (El Hariri et al., 2014). Hence, automation of that process is of a great gain for agriculture and industry fields. For agriculture, it may be used to develop automatic harvest systems and saving crops from damages caused by environmental changes. On the other hand, for industry, it is used to develop automatic sorting system or checking the quality of fruits to increase customer satisfaction level (Brezmes, Llobet, Vilanova, Saiz, & Correig, 2000; Elhariri et al., 2014). Accordingly, an objective and accurate ripeness assessment of agricultural crops is important in ensuring optimum yield of high quality products. Moreover, identifying physiological and harvest maturity of agricultural crops correctly, will ensure timely harvest to avoid cutting of either under-ripe and over-ripe agricultural crops (El Hariri et al., 2014, 2014; May & Amaran, 2011).

To put it briefly, the main research motivation of the approach proposed in this article is providing an automated multi-class classification approach for tomato ripeness measurement and evaluation via investigating and classifying the different maturity/ripeness stages based on the color features. As previously stated, among the 144 countries reported for tomato production, Egypt occupied the fifth place in both income and weight of fruit produced as well as the fact that tomato is the first most important crop in Egypt. Thus, another research motivation in this article is the selection to focus on monitoring the ripeness stages of tomato crop, especially in Egypt. The dataset used for experiments were constructed based on real sample images for tomato at different stages, which were collected from different farms in Minya city, Upper Egypt. Dataset of total 250 images was used for both training and testing datasets with 10-fold cross-validation. Training dataset is divided into 5 classes representing the different stages of tomato ripeness. The proposed approach consists of three phases; namely *pre-processing*, *feature extraction*, and *classification* phases. 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 has been removed using background subtraction technique. Also, each image has been converted from RGB to HSV color space. For feature extraction phase, Principal component analysis (PCA) algorithm was applied in order to generate a feature vector for each image in the dataset. Finally, for classification phase, the proposed approach applied Support Vector Machines (SVMs) and Linear Discriminant Analysis (LDA) algorithms for classification of ripeness stages.

Another basic research motivation is that, to the best of our knowledge, none of the recent ripeness classification related research works have addressed the dependency of the classification approach performance on statistics of the experimented dataset(s).

So, another contribution of this article is that it highlights the most appropriate classification algorithm considering the dependency of the classification approach performance on statistics of the experimented dataset. That has been achieved via adopting the utilization of principal component analysis (PCA) in addition to Support Vector Machines (SVMs) and Linear Discriminant Analysis (LDA) algorithms for feature extraction and classification, respectively, for tomato ripeness stages evaluation and classification considering the color features. Also, both training and testing datasets have been generated via employing the 10-fold cross validation.

An essential finding is that the performance of LDA and SVMs was highly dependent on statistics of the dataset. That is, on datasets with fewer classes (ripeness categories), and many training examples per class, SVMs had an advantage over the LDA classification approach.

The selection of both SVMs classification algorithm depended on the facts that the application of SVMs classification algorithm has may advantages such as, it deliver a unique solution, it does not need any assumptions about the functional form of the transformation, because the kernel implicitly contains a non-linear transformation. Also, if an appropriate generalization grade was chosen, even when the training sample has some bias, SVMs can be robust. Moreover, by choosing an appropriate kernel, one can put more stress on the similarity between samples. However, there is as well some limitations for using SVMs algorithm, that is the lack of transparency of results and the need for very large training time when using large datasets (Auria & Moro, 2008). On the other hand, the selection of both LDA classification algorithm depended on its advantages that are LDA has some advantages such as, the employment of projection that solves the problem of illumination by maximizing between-class scatter and minimizing within-class scatter and it need less samples in order to obtain a reliable classifier. However, one common disadvantage of LDA is the singularity problem as well as it fails when all scatter matrix are singular (Kumar & Kaur, 2012).

In general, the limitations we faced in this research are the dataset size that's needed to be larger, as the accuracy of SVMs increases by increasing the number of images per training class, and accordingly a maximum accuracy of 90.2% has been achieved.

The rest of this article is organized as follows. Section 2 introduces some recent research work related to monitoring and classification of maturity stages for tomatoes and other fruits/vegetables. Section 3 presents the core concepts of SVMs, LDA and PCA 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 and addresses a number of future research suggestions.

2. Related work

This section reviews a number of current research approaches that tackle the problem of ripeness monitoring and classification for tomatoes and other fruits/vegetables.

First of all, for tomato ripeness classification, various research works have been proposed. In Zhang and McCarthy (2012), authors offered tomato maturity evaluation approach using magnetic resonance imaging (MRI). For the proposed approach, MR images were captured for tomatoes that were harvested from the field at different maturity stages. Then, for each of the MR images, the mean and histogram features of the voxel intensities in the region of interest (RoI) were calculated. Finally, partial least square discriminant analysis (PLS-DA) algorithm was applied using both the calculated features and maturity classes variables in order to deduce a maturity classification model showing that different maturity stages are embedded in MR images signal intensity.