Image-Based Date Fruit Classification

Abdulhamid Haidar
Massachusetts Institute of Technology
Cambridge, MA
ahaidar@mit.edu

Haiwei Dong and Nikolaos Mavridis
Department of Computer Engineering
New York University Abu Dhabi
Abu Dhabi, UAE
{haiwei.dong; nikolaos.mavridis}@nyu.edu

***Abstract*— Date fruits are small fruits that are abundant and popular in the Middle East, and have growing international presence. There are many different types of dates, each with different features. Sorting of dates is a key process in the date industry, and can be a tedious job. In this paper, we present a method for automatic classification of date fruits based on computer vision and pattern recognition. The method was implemented, and empirically tested on an image data spanning seven different categories of dates. In our method, an appropriately crafted mixture of fifteen different visual features was extracted, and then, multiple methods of classification were tried out, until satisfactory performance was achieved. Top accuracies ranged between 89% and 99%.**

***Index Terms*-date feature extraction, automated sorting, computer vision.**

I. Introduction

Dates – fruits of date palm trees – are very popular in the Middle East. Dates have very important historical, cultural, and religious significance, and there is evidence that date palm trees have been cultivated in Arabia as far back as 6000 BCE (Alvarez-Mon 2006). Dates are very often served with traditional food, and can be served as either appetizers or sweets. Dates also carry religious importance, and are associated with Muslim events such as Ramadan. Dates are also important economically. In 2010, global production exceeded 7.75 Million Tons, valued at over 3.8 Billion US Dollars [1]. The bulk of production, as well as consumption, came from the Middle East (Fig.1-2). In 2010, the top five producers – Egypt, Iran, Saudi Arabia, Pakistan and the UAE – alone accounted for well over 60% of global production.

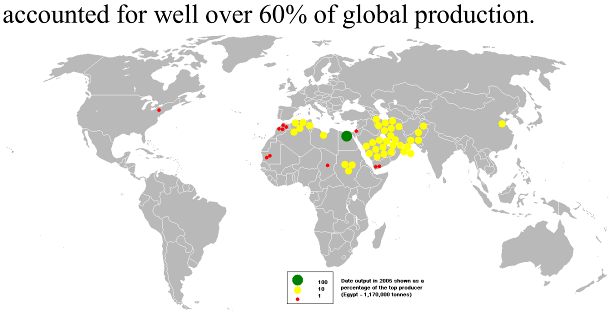


Figure 1. Date output distribution. This map shows the global distribution of date output in 2005 as a percentage of the top producer (from Wikipedia) [2].

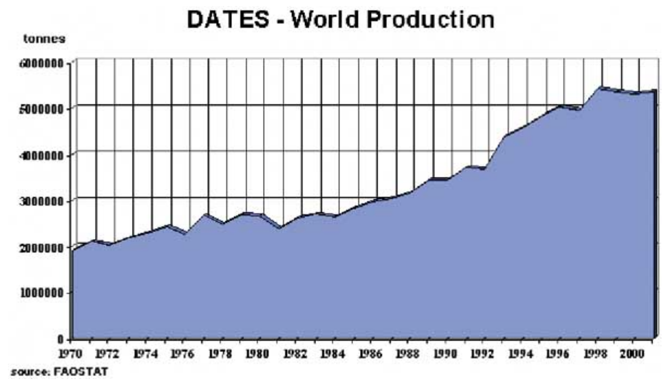


Figure 2. World date production. It increased through the late 1900s [3].

The popularity of dates is on the rise – dates from the UAE were the best selling product at SIAL China 2012, one of Asia's top food exhibitions [4]. With new international attention, there is large economic potential in dates, both regionally and internationally.

There are over 40 unique types of dates, and over 400 varieties, which cover a wide range of taste, shape, and color, as well as price and importance. The process of classifying dates is thus quite important, particularly because a large percentage of consumers can not differentiate between many different types – and thus, one could even envision a cell phone camera-based application, which could be used by consumers in the marketplace. It is also particularly important to be able to classify dates visually for automated factory classification. Ohali reported in 2010 that the process of sorting dates is completed manually by humans, is time intensive, and is a key source of delay in the date production cycle [5]. Thus, implementation of automated computer vision-based classification system has the potential of radically improving the entire date industry. Lee for example, successfully built a sorting system incorporating computer vision-based grading, which was installed and operated in a packaging facility [6]. He reports that his system, relative to manual grading and sorting, provided greater accuracy and a reduction in operational costs.

In this research, we attempt to classify different types of dates using computer vision, without the need for complex physical measurements or composition-based techniques. Thus, our methods can be directly applied to the industry. Although different date types can be distinguished by visual characteristics, one technical problem in visual-based recognition is that there sometimes exists a wide range of certain features within samples of a specific date type, i.e. large intra-class variation (Fig. 3). Such variations could result from natural differences in exposure to water, heat, etc. On the other hand, several date types are not distinguishable merely by basic visual measurements, such as color and size, and thus further visual detail is required to separate such date types.

Figure 3. Safawi Al Madina dates. Two dates of the same type (Safawi Al Madina) with different sizes and color distributions.

A good amount of research has gone towards image-based techniques and pattern recognition regarding fruits and vegetables. D. Unaya et al. have done much successful work related to machine-vision based apple-grading [7] [8] [9]. However, a large portion of such research focuses on identifying faults, deformations and defects. Besides apples, potatoes have also been focused on [10]. There has been little research dedicated to palm date fruits, and most literature on the subject revolves on grading and quality classification. Schmilovitch et. al used near-infrared spectrometry to measure 'maturity' of dates [11]. Lee used a similar computer vision-based method (digital reflective near-infrared imaging) to grade dates [6]. Al-Janobi explored visual-based classification based on color [12] and texture [13] features, also for the purpose of grading.

Literature concerning the classification of different types of dates, on the other hand, is very scarce. Hobani et Al. developed a system for classifying seven different varieties of dates, with accuracies reaching 99.6% [14]. However, their system was based only partially on computer vision and image analysis, and made heavy use of physical properties that were measured carefully, such as weight, volume, and moisture level. In fact, only three of their eleven features were based on visual data, and those related only to the color of the dates. Such requirements make their system a poor candidate for any automated industrial classification system, as measurements for any given date would require an immense amount of time, and probably manual labor as well. Fadel, on the other hand, presented a classification program based entirely on visual features, but relied only on color data – RGB means and standard deviations [15]. His method classifies five different types of dates, with accuracies averaging 80%. The most comprehensive computer-vision system that deals with dates is probably the one developed by Ohali, which accounts for size, shape, and flabbiness, among other things [5]. However, his method – like most – classifies dates based on grade, and achieves a maximum success rate of 80%. Regarding the pattern recognition methods that were utilized, there has been a lot of concentration on the usage of artificial neural networks for classification of fruits [16] [17]. This is also true for projects concerning dates. Ohali, Fadel, and Hobani et. al, as well as others, used neural networks for classification [14] [15] [5].

As we shall see, although our method utilizes solely image-based features, which are inexpensive and very fast to acquire, it achieves accuracies comparable to those that rely heavily on physical features, thus supporting the design and implementation of computer vision-based industrial sorting mechanisms. Moreover, we explore the uses of alternative methods to neural networks and compare the results achieved.

In this paper, in order to classify the dates, fifteen visual features were extracted from training images, including: means and standard deviations of colors, size, shape descriptors, as well as texture descriptors (entropy and energy). In order to extract features, images went through several stages of image processing, including color threshold masking, color filtering, and region identification. Then, multiple classification methods were used, including nearest neighbor, artificial neural networks and linear discriminant analysis. We will start by presenting our method in more detail, describing image preprocessing, the customized features that were extracted, as well as the classification methods that were tried out, and then we will present detailed results, followed by discussion section, and our concluding remarks.

II. Feature Extraction

For our project, a total of 140 images of dates were available, divided evenly between seven common market date types – Sukkery, Mabroom, Khedri, Safawi Al Madina, Madina Ajwa, Amber Al Madina, and Safree. Images had a size of 480 x 640 and were in JPEG format, and were taken from a set position, so that all images had the same dimension scales. Dates were placed in different positions, but were always fully contained in the frame. Although they were placed on a white background, lighting varied considerably between different images. Shadows almost always existed around the dates. In addition, in some cases the white background did not cover the entire frame, leading to dark triangles on edges and corners. In order to perform segmentation of the date regions, an adaptive thresholding method was implemented to separate the date from the rest of the image. No uniform threshold worked well, due to significant differences amongst the date colors, as well as lighting and shadow issues.

Thresholds that worked well on lighter images often captured shadows when used with darker images, and those that worked well for dark images failed to capture some lighter-colored date parts in lighter images.

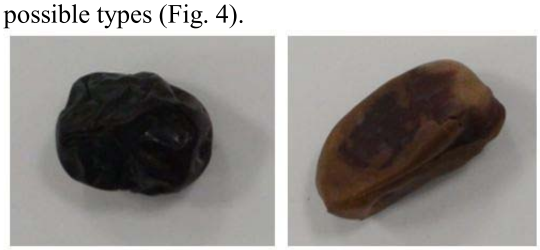
Thus, a customized threshold was used for every image, which was calculated on the basis of the color distribution. Table 2 shows a sample date image from each type, as well as the image after background-removal and the color histograms from which the thresholds were extracted. The histograms contain two high-frequency regions (bimodal distribution) – the first being the date colors, and the second being the light-grey/white background.

The histograms were filtered in order to make the graph smooth and continuous, as not all intensity values with observed, leaving zero readings throughout the histogram. The filter used was a one-dimensional finite-impulse response filter. The target threshold was the color value at the end of the first intense region (the beginning of the 'plateau'), thus effectively separating the pixels belonging to the date vs. the background.

A total of 15 features, summarized in Table 1, were chosen to represent each date image, and were calculated after the date region was extracted from the image using the adaptive threshold-based segmentation-derived mask.

A. Color-Related Features

Colors are perhaps the most important features, as some date types vary considerably in color. Thus, color alone can often be used to restrict the number of possible types (Fig. 4).



(a) (b)

Figure 4. Color feature. (a) A typical Madina Ajwa date has an uniform and dark color. (b) Safawi Al Madina dates contain lighter and darker regions.

Distributions of colors are also quite important. While some date types have uniform colors, others contain both light and dark regions, and such differences could possibly identify certain types. Once the date region is extracted from the image, several resultant images are produced. First, red, green and blue component regions are extracted, and from each the average corresponding color value is calculated. The standard deviation of each color component region is also measured and used as a feature, in order to represent the distribution of the three colors.

TABLE I. FEATURES AND CORRESPONDING MEASUREMENT FORMATS

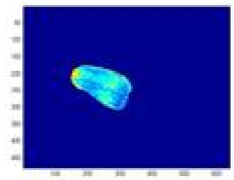
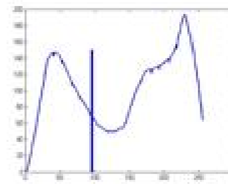
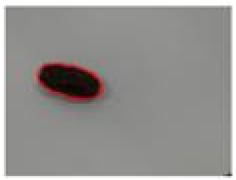
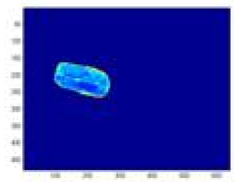
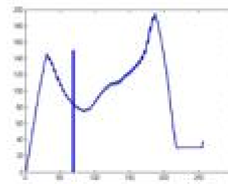
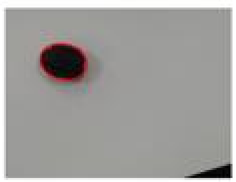
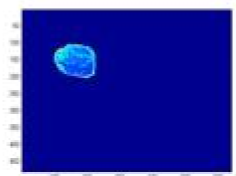
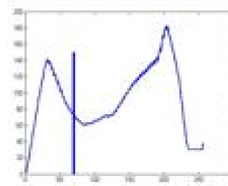
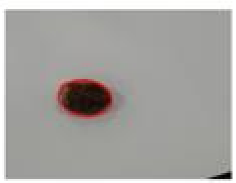
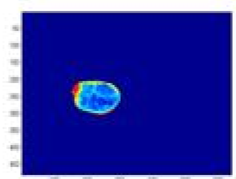
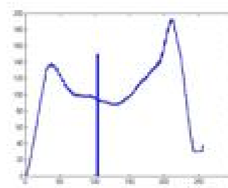
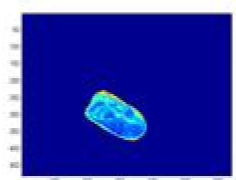
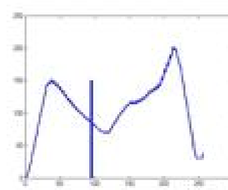
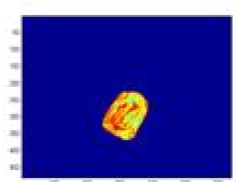
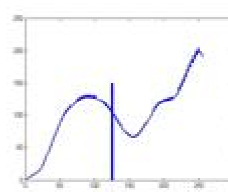
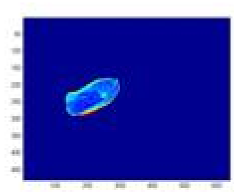
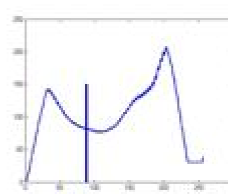
Feature Group	Feature	Measurement Format
Color-Related Features	Red Filter Mean	Float from 8-bit Int (0 - 255)
	Green Filter Mean	Float from 8-bit Int (0 - 255)
	Blue Filter Mean	Float from 8-bit Int (0 - 255)
	Red Filter Standard Deviation	Float from 8-bit Int (0 - 255)
	Green Filter Standard Deviation	Float from 8-bit Int (0 - 255)
	Blue Filter Standard Deviation	Float from 8-bit Int (0 - 255)
Size and Shape Features	Area	Number of Pixels
	Perimeter	Number of Pixels
	Ellipse Eccentricity	Ratio between foci separation and major axis length
	Major Axis Length	Distance in terms of pixels
	Minor Axis Length	Distance in terms of pixels
Texture-Related Features	Red Filter Entropy	Bit
	Green Filter Entropy	Bit
	Blue Filter Entropy	Bit
	Energy	Sum of squares of Gray Level Co-Occurance Matrix entries

B. Shape and Size Features

For each date, the ellipse with the same normalized second central moments was used as a best-fitting ellipse. Ellipses were chosen as a modeling shape due to the natural shapes of the dates. From the best-fitting ellipse, the lengths of the major and minor axis were calculated and used as features. In addition, the eccentricity – defined as the ratio between the major axis length and the distance separating the two foci – was also used.

The size of each date is a key feature, as it has great separating power amongst different date types. Although dates are not perfectly symmetrical, the area of each date in the images was used to represent size. Areas were obtained simply by counting the number of pixels in the mask, determined by the adaptive threshold. In addition, the perimeter of the date image segments was also used, which was simply the number of date-pixels bordering non-date-pixels.

TABLE II. COLOR DISTRIBUTION ANALYSIS

Date Type	Original Image, With Best-Fitting Ellipse	Colorscaled Image After Background Separation Via Masking	Color Histogram – Vertical Line Represents Adaptive Threshhold
Amber Al Madina			
Mabroom			
Madina Ajwa			
Safawi Al Madina			
Safree			
Sukkery			
Khedri			

C. *Texture-Related Features*

It was also necessary to take texture and randomness measures into consideration, as some date types can be visually separated from others by their smoothness – or lack thereof. In order to represent texture, two tactics were used.

First, the gray level co-occurrence matrix was calculated for each image, using gray-scale versions of the images. Gray-level co-occurrence matrices indicate how often pixels of certain intensities adjoin each other. For its calculation, the gray-scale range was divided into eight segments, and then each pixel was replaced with its gray level, ranging between one and eight. A master scale of 20-150 was used, as opposed to the full scale of 0 – 255, for added accuracy, as very few date pixel values existed outside the chosen range. Then, for every pixel, the gray level of the pixel and the one directly to its right were observed. Finally, the co-occurrence matrix was constructed using the numbers of instances of all possible neighborhoods. Each entry (i, j) represents the number of instances of a j-pixel being to the right of an i-pixel (Fig. 5).

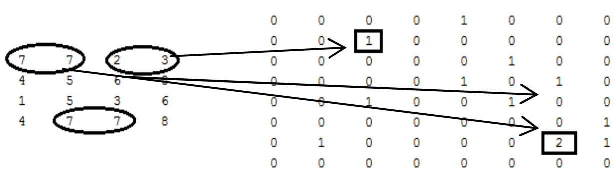


Figure 5. Formation of a gray-level co-occurance matrix from segmented matrix. The (2, 3) entry represents the number of occurrences of (2 3) in the matrix of regions as horizontal neighbors. The entry (7, 7) is a 2, corresponding to two instances of neighboring 7s.

The gray-level co-occurrence matrix, thus, gives us an indication of the smoothness of an image. In an image with completely uniform rows, for example, all neighborhoods will be between pixels of the same level, leading to a diagonal matrix. On the other hand, an image where no neighboring pixels are similar in color will produce a matrix with zeros along the diagonal and positive values elsewhere. From the gray-level co-occurrence matrix, several single values can be extracted to represent randomness or texture. Energy – the sum of the squares of the elements in the matrix – was chosen. Besides energy, entropies – statistical measures of randomness - were determined for each of the red, green, and blue filters, and were used as features.

I. CLASSIFICATION

Three classification methods were used – nearest neighbor, linear discriminant analysis, and neural networks. In more detail:

A. *Nearest Neighbor*

K-Nearest Neighbor was tested, in which each date of a testing set was assigned to the class of the date with the closest feature vector in the training set. In order to measure accuracy, 10-fold validation was implemented: the set of all images was divided into 10 subsets, each of equal size. Each subset was chosen randomly, but in a way that guaranteed balance between date types. Thus, each subset contained two images of each date subset, with a size of 14 images. In turn, every subset was tested using the other nine subsets (126 images) as the training test.

Two types of distance calculation were implemented. First, the standard Euclidean distance between any two points (L2), computed using Pythagorean trigonometry, was used. As an alternative measurement, City Block distances (L1) were also used. City Block distances are defined as the sum of distances along each dimension. In other words, the differences in each of the features are measured independently, and then all differences are summed. Results differed only slightly between the two distance calculation methods.

For each method and k-value, the test was repeated 100 times, each with different subsets.

B. *Discriminant Analysis*

Classification was also performed with Linear Discriminant Analysis. In this process, new features, constructed from linear combinations of the original features, are found and used to better separate the set of features into their respective categories. This is done based on modeling each training subset of the same data type as a multivariate normal density distribution. As in K-Nearest Neighbor, 10-fold validation was implemented to measure accuracy when generalizing. Once again, the date subsets were randomly chosen in a way that guaranteed equal distribution of date types.

C. *Neural Networks*

Classification using artificial neural networks, most likely the most popular method in the field of vision-based automated recognition of fruits, was explored as well. The networks were trained using scaled conjugate gradient back propagation, with 70% of the data reserved for training (98 images), 15% for validation, and 15% for testing (21 each). The validation subset was used to measure how well the network performs with new data, after being trained with the training subset. The neural network was trained repeatedly (with the same training data) until results on the validation subset ceased to improve (measured by the mean squared error on the validation samples), in order to avoid over-fitting. Then, with training complete, the network was put to the test with the testing subset. The networks used were two-layer feed-forward networks, with sigmoid hidden and output neurons. Training spanned neural networks with a range of number of hidden layers, and the optimal net found contained 10 hidden layers.

The entire process of multiple training, validating, and testing at the end was repeated many times with different numbers of hidden neuron layers until an adequate total accuracy was achieved. It must be noted that this method – particularly with repetition until satisfaction – was computationally expensive in comparison to the first two methods.

II. Results

A. Method Comparison

The results are summarized in Table 3. After testing 1- through 20-nearest neighbor, 3-Nearest Neighbor with Euclidean distances emerged as the most accurate test, with an accuracy rate of 90.3 %. City-Block distances were only slightly less informative, as with it and 4-nearest neighbor an accuracy of 89.4% was achieved.

TABLE III. Accuracies Using The Four Methods

Method	Nearest Neighbor (CityBlock)	Nearest Neighbor (Euclidean)	Linear Discriminant Analysis	Artificial Neural Network
Top Accuracy	89%	90%	96%	99%

Linear Discriminant Analysis proved to be a quite successful way to classify different types of dates. This can be attributed to the nature of the different types and the features chosen to represent them – many dates can be quite definitively identified using a small number of features, and with 15 features, linear separation proved to be productive.

The strong separability of some date categories can be visualized directly, by viewing inside the feature space. For example, Fig. 6 plots the major axis length, mean red value and mean blue value of each date, which were carefully chosen to maximize visual separation. It can be seen that several groups form clusters that can easily be separated from others using these features. As shown in Fig. 6, Safree dates (top right rectangle), Madina Ajwa dates (bottom left rectangle), Sukkery dates (bottom right polygon) and Safawi Al Madina dates (central polygon) all form exclusive clusters. For each, the entire population of that type – and none of any other type – lies in the cluster. It must also be noted that a fifth date type – Amber Al Madina (top left rectangle) – would also be exclusive to a certain cluster, had there not been two Khedri dates with very similar features. Thus, using only three features one can cluster and separate many different groups with a large degree of accuracy. More important is the fact that many clusters are linearly separable – a positive indication for a linear discriminant analysis classifier. Also, while Fig. 6 shows only three distinct features, the classifiers have access to fifteen. With several separating planes visible in three dimensions, certainly many can exist with combinations of all fifteen dimensions. It is therefore no surprise that linear discriminant analysis achieved an accuracy of 96% with 10-fold validation.

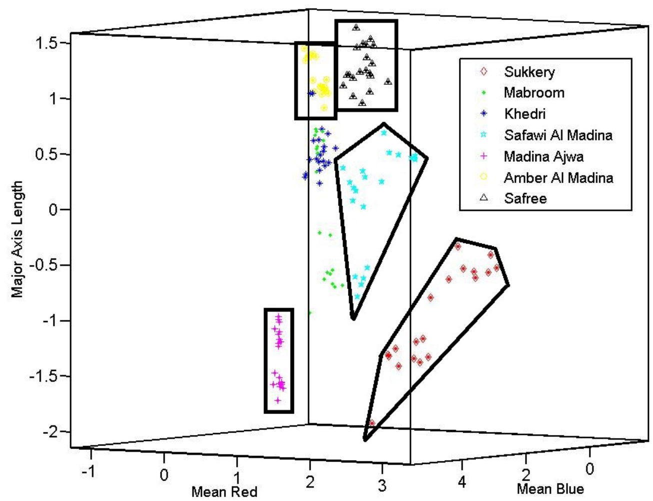


Figure 6. Visualization of date features. The features include major axis lengths, mean red values and mean blue values.

Classification via an artificial neural networks proved to be the most accurate method, as the best neural network found achieved a total accuracy of 99.29 % - the largest success rate achieved during this study. It must be noted that it took a great deal of time and training to arrive at this network – the majority of trained networks had significantly lower success rates, but after approximately 20 minutes, we arrived at our selected classifier, which has performance and generalizability superior to all other classification methods that we tested, as can easily be witnessed in Table 3.

B. Discussion

Artificial Neural Networks emerged with the best final accuracy, reaching 98.6%. However, linear discriminant analysis performed nearly as well, with an accuracy of 96%. The success of LDA can be in part attributed to the existence of easily computable features which can achieve high seperability across a number of the date categories that we are classifying. Nevertheless, our experiment confirmed that in terms of absolute accuracy, neural networks are unmatched. However, one should not underestimate the time and effort required to find a network with a minimum set error rate. Results using nearest neighbor methods are not as accurate as either of the other two methods, but fall within acceptable range.

III. Conclusion

Date palms are not only an ancient fruit with important historical as well as symbolic significance, but are also a noticeable dietary component of several nations in the modern world, a material that is transformed to many others, and a product with a huge worldwide market and strong economic significance. In this paper, we have presented a system for classifying date palms, based on images of the dates, without the need for time-consuming and intricate physical measurements. Our system consists of a customized feature extraction stage, followed by classification. An extensive empirical trial is also presented, exposing a number of results regarding classifier choice and tuning, while guaranteeing generalizability, and most importantly, illustrating the real-world effectiveness and direct applicability of our system. Also, many extensions of the proposed system are possible, thus enabling technology to increase the utilization as well as the quality of one of the world's most ancient fruits, the date palm. Other research involving date classification has provided far less accurate classification methods [15], or has required the measurement of physical attributes [14]. This paper thus shows that it is possible to classify dates very accurately using computer vision, and supports further research in the topic, as well as implementation of vision-based systems in date production plants.

References

[1] Food and Agriculture Organization of the United Nations, "Food and Agricultural commodities production 2010".

[2] Anwar_saadat, "Date Output in 2005", Wikipedia, 2007. http://en.wikipedia.org/wiki/File:2005dattes.png.

[3] A. Zaid, "Date Palm Cultivation," Rome, 2002.

[4] Abu Dhabi Food Control Authority, *UAE Dates Named Best Selling Product in SIAL China 2012,* 2012.

[5] Y. A. Ohali, "Computer vision based date fruit grading system: Design," *Journal of King Saud University,* vol. 23, pp. 29 - 36, 2011.

[6] D. J. Lee, "Development of a machine vision system for automatic date grading using digital reflective near-infrared imaging.," *Journal of Food Engineering,* vol. 86, pp. 388-398, 2008.

[7] D. Unay and B. Gosselin, "Thresholding-based segmentation and apple grading by machine vision," in *Proc. of European Signal Processing Conference*, 2005.

[8] D. Unay, B. Gosselin, O. Kleynen, V. Leemans, M. Destain and O. Debeir, "Automatic grading of Bi-colored apples by multispectral machine vision," *Computers and Electronics in Agriculture,* vol. 75, no. 1, pp. 204-212, 2011.

[9] D. UNAY, "Artificial neural network-based segmentation and apple grading by machine vision," in *Proc. of IEEE Internation Conference on Image Processing*, 2005.

[10] F. Pedreschi, D. Mery, F. Mendoza and J. Aguilera, "Classification of Potato Chips Using Pattern Recognition," *Journal of Food Science,* vol. 69, no. 6, pp. E264 - E270, 2004.

[11] Z. Schmilovitch1, A. Hoffman, H. Egozi, R. Ben-Zvi, Z. Bernstein and V. Alchanatis, "Maturity determination of fresh dates by near infrared spectrometry," *Journal of the Science of Food and Agriculture,* vol. 79, pp. 86-90, 1999.

[12] A. Al-Janobi, "Date inspection by color machine vision.," *Journal of King Saud University,* vol. 12, no. 1, pp. 69-97, 2000.

[13] A. A. Al-Janobi, "Application of co-occurrence matrix method in grading date fruits," in *ASAE Paper, no. 98-3024*, 1998.

[14] A. I. Hobani, A. M. Thottam and K. A. Ahmed, "Development of a neural network classifier for date fruit varieties using some physical attributes," King Saud University - Agricultural Research Center, 2003.

[15] M. Fadel, "Date fruits classification using probabilistic neural networks," *Agricultural Engineering International: the CIGR Ejournal,* vol. 9, 2007.

[16] D. Unay, B. Gosselin, "Apple defect detection and quality classification with MLP-neural networks," in *13th Annual Workshop on Circuits, Systems and Signal Processing*, 2002.

[17] D. Unay and B. Gosselin, "Apple defect segmentation by artificial neural networks," in *Proceedings of BeNeLux Conference on Artificial Intelligence*, 2006.

[18] N. Alavi, "Date grading using rule-based fuzzy inference system," *Journal of Agricultural Technology,* vol. 8, no. 4, pp. 1243-1254, 2012.

[19] S. M. Mazloumzadeh, M. Shamsi and H. Nezamabadi-pour, "Evaluation of generalpurpose lifters for the date harvest industry based on a fuzzy inference system," *Computers and Electronics in Agriculture,* vol. 60, pp. 60-66, 2008.