Abstract— In a supermarket checkout, employees deal with a great number of different fruits and vegetables. Sometimes, the correct pricing of those items is a difficult task, due to the short available time or even inexperience. In this paper we present a complete method for the automatic classification of fruits, extendable for vegetables. We use patches extracted from the images and represented by means of MPEG-7 color and texture descriptors. To obtain the patches, a simple and robust technique is used. This is an important characteristic of the method, since dispenses the use of strong segmentation, which is usually difficult to be achieved, especially in real-world conditions. Two classification methods are evaluated: a multi-layer perceptron network and a support vector machine. An image database with 320 images of eight different fruits was constructed and used in the assessments. The proposed method presents excellent results, with accuracy values higher than 90%.

Keywords— Fruit classification, fruits image dataset, MPEG-7 descriptors, neural network, support vector machine.

I. INTRODUCTION

Due to the great variety of fruits available in Brazilian supermarkets, cashiers usually do not know them all. In order to classify and price them accordingly, they need to look up in a booklet containing pictures and numerical codes of each item to be priced. This procedure is susceptible to human errors both during the code typing and, mainly, at the classification task [1]. In this context, computer vision approaches can be used to assist this process, providing valuable information to the cashiers about the variety of fruit or vegetable.

Traditionally, the architectures for the solution of this kind of problem involve the object segmentation for further feature extraction and classification [1,2,3]. However, the segmentation usually consists in a difficult process and its quality strongly impacts on the overall performance of the system. In order to decrease the dependence on the segmentation process, we propose the use of image patches that can be easily obtained using a particular acquisition method associated with a simple algorithm.

In this paper, we present an image database of fruit images, as well as details about the acquisition process, in order to expand the database to include new varieties of fruits and vegetables. Also, we propose classification methods using MPEG-7 visual descriptors, multi-layer perceptron (MLP) artificial neural network (ANN) and support vector machine (SVM) with radial basis function (RBF) kernel.

This paper is organized as follows: section II presents a short review of the related work and the used machine learning methods; sections III and IV present an overview and detailed descriptions of the system, respectively; section V the results and we conclude in section VI.

II. BACKGROUND AND CONTEXT

Fruits and vegetables classification is a problem not extensively studied yet, although it can be employed in many applications such as automatic harvesting of fresh fruits [4], quality and maturation inspection [5], subgroup sorting of a specific fruit [8] and identification at supermarkets checkout [1].

In [1], an image database was build and some features (color, texture and shape) and classification methods (supervised and unsupervised) were used to show that computer vision methods based on segmentation and machine learning can be used to tackle this problem. In [2], it was used the same image database as in [1], this time proposing the use of just color and texture features, including a combination of both. In [3] the author focuses on the use of shape descriptors but does not detail the segmentation step. It is worth mention that all those methods rely on an accurate segmentation of the image to achieve good results.

ANNs are well-known methods used in classification and pattern recognition in computer vision tasks. In most current available implementations, it is easy to configure important parameters of the ANN, such as changes in the activation function and amounts of inputs or neurons in the hidden layer. Some of the available ANN implementations are the Matlab ANN toolbox [6] and the application on Neural Networks for Mathematica [7]. Moreover, good overall results can be achieved whether the problem is considered linear or not.

In [8], MLP and RBF networks are used to sort different subgroups of date fruits. Images are acquired with a Sony color camera, segmented and then processed. The authors use color and shape features such as RGB values, perimeter, length, width and length-to-width ratio.
In [5], just MLP network is used to classify the different ripening of oranges according to the definitions of the Brazilian Center of Quality in Horticulture. The method consists in a system for image capture using a white background under carefully controlled light condition for easy segmentation, color RGB feature extraction and the MLP network.

Support Vector Machines (SVMs) are a supervised learning technique widely used for many different kinds of classification tasks. They were initially conceived to solve classification problems between only two classes, but they can be employed in multi-class problems by using one-against-all or one-against-one techniques [9].

III. OVERVIEW OF THE ARCHITECTURE

A. Image database

The database is composed by 8 groups, each one for a single variety of fruit: bahia orange, pear orange, sicilian lemon, thaiti lime, fuji apple, gala apple, tangerine and tomato, as shown in Fig. 1. In each group, 5 fruits were photographed in 8 different poses, giving a total of 40 images per group. Therefore, the complete database has 320 images.

In the algorithm, a region of interest (patch) containing part of the vegetable being inspected is extracted and used as input to the feature extraction step.

Fig. 1. Sample images of each variety of fruit, from the 320 images dataset. From left to right, top-down: bahia orange, pear orange, sicilian lemon, thaiti lime, fuji apple, gala apple, tangerine and tomato.

B. Features

According to [10], “the main objective of the MPEG-7 visual standard is to provide standardized descriptions of streamed or stored images (...) that help users or applications to identify, categorize or filter images or video”. Since we need robust and representative descriptors and the MPEG-7 visual standard provides them [11], we decided to follow this standard. Amongst the available implementations of the MPEG-7 descriptors extraction we choose [11] due to its easy adaptability to our development environment and clear documentation.

C. Classification

Preliminary tests gave us evidences that linear classifiers would not be suited for this task. Also, the broad range of characteristics that fruits and vegetables can assume suggests that non-linear approaches are best suited to this task. Thus, we adopted in this work artificial neural networks (ANN) and support vector machines (SVM).

The ANN used is a MLP, trained by back-propagation, due to its proven ability to handle non-linear pattern recognition and classification problems [12].

Although SVMs are designed to do only two-class classification, they have shown consistent performance when applied to multi-class classification problems, using one-against-one or one-against-all approaches [9].

IV. THE PROPOSED METHOD

A. Image acquisition

In order to present a lower dependence on accurate segmentation processes, our method uses small image patches of the object instead of the entire object within its accurate contour. Image acquisition plays an important role in the process of the patches detection (due to the technique used to detect the patches). The image acquisition system needs to ensure that a significant part of the image area is covered by the target fruit and that the target is brighter than the rest.

Therefore, the image acquisition process employed here consists of a black background on a workbench, where the target is positioned. The illumination consists of a 25 W dichroic lamp positioned right over the camera in order to avoid shadows. The photograph equipment is a Samsung PL51 digital still camera, operating at a 1024 x 768 pixels resolution. In order to provide stability to the equipment, a tripod is used at a fixed location and constant viewpoint, as depicts the diagram in Fig. 2. Every target was manually moved to assume 8 different poses.

Fig. 2. The main components of the image acquisition system for the database construction.

B. Patch detection

The method for patch detection consists of locating the brightest image pixel and cropping a 192 by 192 window around it. This size was empirically selected to be larger enough so it has sufficient information under the considered scale, and smaller enough so it is still fast to compute.

The brightest pixel is computed by converting the image to grayscale according to (1) and then finding its global maximum.

\[
g = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B \tag{1}
\]

where R,G and B are the red, green and blue components respectively.
Fig. 3 shows one example of each class after applying the patch detection method. Note that a shiny spot is visible in the center of the patches as a result of the illumination system and the patch detection method.

![Fig. 3. Results of the patch detection. From left to right, top-down: bahia orange, pear orange, sicilian lime, thati lemon, fuji apple, gala apple, tangerine and tomato.](image)

C. Feature extraction

For the feature extraction step, it uses the code from [11], since it is based on the standard reference code adding minor bug corrections. The following MPEG-7 descriptors are computed:

- CSD - Color Structure Descriptor (64 integer values): captures both color content and the position where they appear, it is basically a color histogram computed in a window sliding over the image;
- SCD - Scalable Color Descriptor (128 integer values): color histogram in HSV encoded by a Haar transform;
- CLD - Color Layout Descriptor (120 integer values): small scale representation of the image (lowered resolution) encoded by a DCT and quantized;
- HTD - Homogeneous Texture Descriptor (62 integer values): Assumes that texture is homogeneous and characterizes its properties;
- EHD - Edge Histogram Descriptor (80 integer values): local histogram of four directional and one non-directional edge types; can be aggregated into global histograms.

The adopted feature vector is the concatenation of all the above descriptors, resulting in a total of 454 bins.

D. Classification methods

A general feed-forward MLP network capable of classifying inputs into classes is illustrated in Fig. 4 (assuming that it follows Stone-Weierstrass theorem [12]). The input data is a matrix containing feature vectors of the samples. Those inputs are computed in a group of artificial neurons (AN). An AN is a structure formed by a scalar value from the input multiplied by a weight and added to a bias. The result is then used as argument to an activation function. In this case, this function is a sigmoid. Fig. 5 and (2) show a representation of an AN. In order to compute the output values (a in Fig. 5) when the ANN has matricial inputs, (3) can be used. This set of AN is called hidden layer. Its size (number of neurons) can vary.

![Fig. 4. A general representation of a MLP network used for classification and pattern recognition problems. Redrawn from [13].](image)

![Fig. 5. A representation of an AN. Redrawn from [13].](image)

\[ a = f(p^*w+b) \]  \hspace{1cm} (2)

where \( a \) is the output, \( f \) is the activation function, \( p \) is the input, \( w \) is the weight and \( b \) is the bias.

\[ a_j = f_j^{\text{hidden}}(\sum_i x_i w_{ji} + b_j) \]  \hspace{1cm} (3)

where \( a_j \) is the \( j \)th output of the neuron, \( x_i \) is the \( i \)th network input, \( w_{ji} \) is weight value in the \( j \)th neuron of the \( i \)th input, \( b_j \) the biasing value of the \( j \)th neuron and \( f_j^{\text{hidden}} \) is the activation function in the \( j \)th neuron.

Finally, the outputs of the hidden layer are used as inputs to the output layer. Thus, the output layer is responsible for sorting the data. Necessarily, the number of neurons in the output layer is equal to the amount of classes. The resulting output is a matrix identifying in which class the sample belongs.

The back-propagation algorithm is used to train the network. This algorithm consists in recalculating weights and biases from the output to the input layer. The updated weights and bias can be computed by (4) and (5).

\[ w_j(k+1) = w_j(k) - \alpha e_j^* x_i \]  \hspace{1cm} (4)

\[ b_j(k+1) = b_j(k) - \alpha e_j \]  \hspace{1cm} (5)

where \( w_j(k+1) \) is the new weight of the \( j \)th neuron, \( w_j(k) \) is the actual weight value of the \( j \)th neuron, \( \alpha \) is the learning rate, \( e_j \) is the error value for the \( j \)th neuron, \( b_j(k+1) \) is new bias for the \( j \)th neuron and \( b_j(k) \) is the actual bias of the \( j \)th neuron.

A SVM maps the data into a higher dimensional space, where it can then be separable by a hyperplane. Fig. 6 shows
an example of a hyperplane and its respective support vectors (samples crossed by dashed lines) defining the maximum margins (distance from dashed to the continuous line).

Fig. 6. Example of support vectors in a 2 dimensional space. In this example, just the 3 samples crossed by dashed lines (support vectors) are needed to further classification. From [14].

Training a SVM is done by solving a quadratic optimization problem described in details in section 2.1 of [9]. In the specific case where the RBF kernel is used, two parameters need to be specified: the standard deviation (sigma) of the kernel and the regularization parameter C (see [9] for more details).

In order to achieve best performance, those parameters can be adjusted to minimize the error on a cross-validation set, in a process called grid-search. Grid-search was conducted in our dataset, using 224 images for the training set and 48 for the cross-validation, leaving more 48 out for the test set, and the parameters space shown in Fig. 7 leads us to \( \gamma = 1 \times 10^{-5} \) and \( C = 100 \) to be the best.

Fig. 7. Grid-search result. Red circle shows the maximum accuracy in the cross-validation set.

Since the fruits classification problem is multiclass, and LIBSVM [9] was used to train and test the SVM, their approach was the one-against-one [9]. Thus, 28 SVMs were generated and trained, and the results from each one were then combined, based on the probability output from each one, to generate the final output.

V. RESULTS

All results presented here consider the following numbering for the classes: 1 – bahia orange, 2 – pear orange, 3 – sicilian lemon, 4 – thai lime, 5 – fuji apple, 6 – gala apple, 7 – tangerine and 8 – tomato.

The MLP was trained to a cross-validation error of 12.5%. It was reached after 6 epochs, achieving an accuracy of 93.75% in the test set. The confusion matrices for the cross-validation and for the test set can be seen in the Fig. 8 and Fig. 9 respectively.

![Cross-validation confusion matrix for MLP ANN.](image)

![Test confusion matrix for ANN MLP.](image)

The SVM parameters were chosen after a grid search performed by training and cross-validating it. Finally it was retrained with the best parameters found, achieving an accuracy of 93.75% in the test set. The confusion matrices for the SVM are presented in Fig. 10 and Fig. 11.
As can be seen in the confusion matrices, both systems are able to handle the classification task very well, although some errors still happen. Fig. 12 and Fig. 13 show the 3 patches misclassified by the MLP and SVM, respectively, side by side. Note that they are indeed quite similar, what explains the occurrence of classification errors for these patches.

VI. CONCLUSIONS

The method presented in this paper for the classification of fruits images uses MPEG-7 visual descriptors and non-linear supervised learning techniques. Specifically, feature vectors originated from the concatenation of CSD, SCD, CLD, HTD and EHD descriptors are extracted from image patches and feeds an ANN and a SVM. Both ANN and SVM presented similar results in terms of accuracy: 93.75%. This is an expressive performance, considering that some varieties of fruits are quite similar (such as bahia and pear orange) and the method dispenses strong segmentation operations. Moreover, the “general-use” MPEG-7 color and texture descriptors presented very good performance in association with the tested classification methods. This indicates that eventual efforts in the design of application-specific descriptors may be carefully considered. In other words, we suggest that performance improvements are more likely to be obtained by means of a fine tuning of the available descriptors (to build the feature vectors), than by means of the design of application-specific descriptors.

It is important to stress that the current available approaches for fruits and vegetables automatic classification are dependent on a segmentation step. Such automatic segmentation has to be extremely reliable in order to allow good performance of the system’s further steps – feature extraction and classification based on some machine learning technique. The proposed method does not rely on strong segmentation, but on a simple technique to detect a squared patch of the image.

Future work includes the expansion of the database, considering common vegetables found in the supermarkets, such as different kinds of lettuce, broccoli and potatoes, just to cite a few. Since the adopted MPEG-7 descriptors consider color and texture, we believe that the method will present good performance with these categories as well. In addition, future work includes the construction of an apparatus
embedding the illumination and image acquisition components, to be evaluated in real-world scenarios.

Finally, the authors would like to mention that the present work was strongly motivated and inspired by a presentation available in the web [15], showing a fully operational prototype of a system for fruits and vegetable classification. Due to the non-scientific nature of the presentation and the lack of official information from Toshiba, the supposed manufacturer, we consider that it is not possible to confirm the correctness of the information presented in the video. However, since the video [15] is an incontestable resource for the contextualization of the present work, we assume that it is worth to mention.

ACKNOWLEDGMENT

To the Brazilian Coordination for Enhancement of Higher Education Personnel (CAPES) for Fáveros’ Scientific Initiation studentship.

REFERENCES


