

Apple detection in natural tree canopies from multimodal images

J. P. Wachs^{1,2}, H. I. Stern², T. Burks³ and V. Alchanatis¹

¹ Institute of Agricultural Engineering, Agricultural Research Organization, the Volcani Center, Bet-Dagan, Israel. victor@volcani.agri.gov.il

² Dept. of Industrial Engineering, Ben-Gurion University of the Negev, Israel

³ Agricultural and Biological Engineering, University of Florida, Gainesville, FL, 110570.

Abstract.

In this work we develop a real time system that recognizes occluded green apples within a tree canopy using infra-red and color images in order to achieve automated harvesting. Infra-red provides clues regarding the physical structure and location of the apples based on their temperature (leaves accumulate less heat and radiate faster than apples), while color images provide evidence of circular shape. Initially the optimal registration parameters are obtained using maximization of mutual information. Haar features are then applied separately to color and infra-red images through a process called Boosting, to detect apples from the background. A contribution reported in this work, is the voting scheme added to the output of the RGB Haar detector which reduces false alarms without affecting the recognition rate. The resulting classifiers alone can partially recognize the on-trees apples however when combined together the recognition accuracy is increased.

Keywords: Mutual information, multi-modal registration, sensor fusion, Haar detector, apple detection.

Introduction

In the last few years, object recognition algorithms are focusing on the efficient detection of objects in natural scenes. A system is developed to recognize in real-time partially occluded apples regardless of position, scale, shadow pattern and illumination within a tree canopy.

The work is motivated by the fact that labor for orchard tasks constitutes the largest expense (Jiménez *et al.*, 2000), and hence there is a need to develop autonomous robotic fruit picking systems. Here we address the first step in such a system by tackling the problem of on tree green apple detection using real-time machine vision algorithms. The complexity of the task involves the successful discrimination of “green” apples within scenes of “green leaves”, shadow patterns, branches and other objects found in natural tree canopies. Color and edges are features highly dependent on illumination while texture is highly sensitive to the proximity (scale) of the object. An excellent review regarding apple recognition systems was presented in (Jiménez *et al.*, 2000b). The concept of background modeling using Gaussian mixture color distributions in RGB images was used in Tabb *et al.*, (2006). This algorithm detected 85 to 96 percent of both red and yellow apples assuming a uniform background in an artificial environment. Color distribution models for fruit, leaf and background classes were used in Annamalai *et al.*, (2003) in a citrus fruit counting algorithm. In Stanjko *et al.*, (2004) pixel thermal values were mapped to RGB values and detected using the normalized difference index. However the efficiency of the algorithm was affected by the apple’s position on the tree and degree of sunlight. In Sapina (2001), textural features extracted from the gray level co-occurrence matrix were used to discriminate between warm objects and their background in thermal images. On the same vein, a

threshold selection approach was proposed by Fernandez *et al.* (1993) based on apple’s texture features in grayscale images. The authors assume that all the apples have a bright spot (due to their exposure to sunlight) and the apple region is practically homogenous and spherical. These assumptions have limited validity in natural un-controlled scenarios. Texture based edge detection combined with a measure of redness are used in Zhao *et al.*, (2005) for the detection of green and red apples in trees. The authors claim that their method can deal with occluded apples, clustered apples and cluttered environments. However no recognition rates are reported. A robust system using an infrared laser is presented in Jiménez *et al.*, (2000) which considers illumination, shadows and background objects. The authors report a rate of 80-90% of detection when used with an artificial orange tree.

Our paper proposes the use of two modalities; infra-red and color. Infra-red provides clues regarding the physical structure and location of the apples based on their temperature (leaves accumulate less heat and radiate it faster than apples), while color images provide evidence of circular shape.

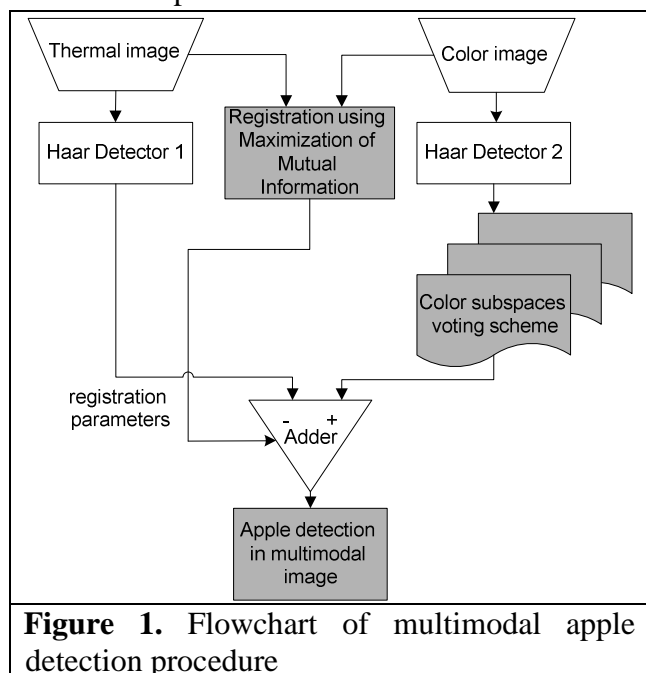


Figure 1. Flowchart of multimodal apple detection procedure

Our approach consists of a pipeline of registration, detection, color space voting and combining stages as shown in Fig. 1. In registration correspondence matching between a color and a thermal image is achieved using the maximization of mutual information technique and the registration parameters are obtained. At the same time, apples are detected using a Viola –Jones classifier (Viola and Jones, 2004) based on Haar-like features in the detection phase. The color detections are converted to hypotheses that are tested each by a voting scheme. The resulting detections are combined with the thermal results and transformed using the registration parameters.

This paper is structured as follows. The registration algorithm based on maximization of mutual information is described, then the

classification fusion scheme is presented. The results of each modality independently, their combination and the resulting enhancement are given in the last section.

Materials & Methods

Multimodal image registration using mutual information

Multi-modal image registration is a fundamental step preceding detection and recognition in image processing pipelines used by the pattern recognition community. This preprocessing stage concerns the comparison of two images –the base and sensed images- acquired from the same scenario at different times or with different sensors in such a way that every point in one image has a corresponding point on the other images, in order to align the images. In our problem, the transformation between two images of different modalities is affine which means; rotations,

translations and scaling are allowed. Transformation of the coordinates P_A and P_B from the sensed image A to the base image B is given by Equation 1.

$$(P_B - C_B) = sR(\theta)(P_A - C_A) + t$$

$$R(\theta) = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \quad t = \begin{pmatrix} t_x \\ t_y \end{pmatrix} \quad (1)$$

Where C_A and C_B are the coordinates of the centers of the images, s is a scaling factor, $R(\theta)$ is the rotation matrix, and t is the translation vector.

We shall compare five different registration methods using the similarity indices: cross correlation normalized (CC_1), correlation coefficient (CC_2), correlation coefficient normalized (CC_3), the Bhattacharyya coefficient (BC) and the Mutual Information index (MI).

We first introduce the mutual information (MI) method (Viola and Wells, 1995) as this will be compared to other methods for registration. Let A, B be two random variables, $p_A(a)$ and $p_B(b)$ with marginal probability distributions and $p_{AB}(a,b)$ a joint probability distribution. The degree of dependence between A and B can be obtained by the MI, according to Equation 2.

$$I(A, B) = \sum_{a,b} p_{AB}(a,b) \log \frac{p_{AB}(a,b)}{p_A(a)p_B(b)} \quad (2)$$

A data set including 125 color and thermal images of apple trees were acquired from a digital RGB camera and an IR FLIR camera. These images were registered by the five indices mentioned earlier. Table 1 shows the root mean squared errors (RMS) of the five indices for each registration parameter.

Table 1. Registration parameters RMS error using the five 5 similarity indices.

Measure	RMS			
	Δs	$\Delta \theta$	Δt_x (%)	Δt_y (%)
bc	0.226	2.205	3.958	4.328
mi	0.175	1.701	3.547	3.912
cc_1	0.196	1.929	3.985	3.868
cc_2	0.196	1.715	6.030	6.875
cc_3	0.196	1.713	6.067	6.848



Figure 2. Color and thermal registered image

By observing the results in Table 1, the mutual information technique performed better than the other four methods for three parameters (Δs , $\Delta \theta$, Δt_x), and comparable to cc_1 for the last parameter (Δt_y). Therefore MI was selected as the preferred method for registering the whole set of images. Fig. 2 shows an example of a pair of images registered from the dataset.

Apple detection using Haar classifiers

Apple detection using Haar classifiers are applied separately in color and thermal images. We also use a boosted cascade of simple classifiers inspired by Viola and Jones (2004). This classifier relies on features called Haar-like, since they follow the same arrangement as the Haar basis. The eleven basis features, i.e. edge, line, diagonal and center surround features, are presented in Fig. 3.

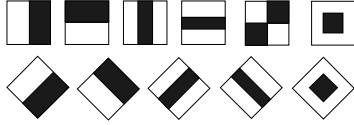


Figure 3. Eleven Haar features: edge, line, diagonal, center surround and rotated features

Since the number of features to be computed is quite large, integral images are adopted for fast computation. Let I be a temporary image, representing the sub-window to be classified, which includes the sum of gray scale pixel values of the sub-window N with height y and width x , such that:

$$I(x, y) = \sum_{x'=0}^x \sum_{y'=0}^y N(x', y') \quad (3)$$

The integral image is calculated recursively: $I(x, y) = I(x, y-1) + I(x-1, y) + N(x, y) - I(x-1, y-1)$ where $I(-1, y) = I(x, -1) = I(-1, -1) = 0$. This requires one scan over the input sub-window. Rotated features can be computed effectively in a similar way (Lienhart and Maydt, 2002).

A feature is detected when the computation of the weighted differences between the white and black areas of the rectangles (see Fig .3) are higher than a threshold. This threshold is determined during the training process in such a way that the minimum number of samples is misclassified. The set of selected features is learned through a Classification and Regression Tree (CART) technique, which is a form of binary recursive tree. To achieve a given detection and error rate, a set of simple CARTS is selected through the Gentle Adaboost algorithm (Freund and Shapire, 1996).

In order to improve the overall performance of the classifiers, they are arranged in a cascade structure, where in every stage of the cascade, a decision is made whether the sub-window includes the object to be detected. At every stage, at least a high hit rate is assured, e.g., 0.995 and at least half of the false alarms are discarded. In spite of the hit rate and the false alarms are reduced, the hit rate decreases slower than the false alarms rate (FA). For example for 20 stages, since every stage keeps the hit rate to 0.995 at least, after 20 stages, the hit rate is $0.995 \times 10^{20} = 0.904$. The false alarms rate (FA) is decreased in every stage so half of the FA detections is rejected every stage. For every stage the classification function is learned until the maximum number of stages is reached or the minimum acceptable FA rate is obtained.

Learning color subspaces using A voting scheme

In this section separate artificial neural network classifiers are trained and tested for each of the three color spaces; L^*a^*b , hsv and rgb. Since, as we will show, the accuracies obtained for all the color spaces are identical, it was decided to see if a fusion method would provide any advantage. We will show that combining the output of the three classifiers as an ensemble by “majority voting” will decrease the false alarms without affecting the recognition rate. Thermal images are not considered here since their intensity information can lead to ambiguity between classes.

Training the classifiers

For each window obtained from the Haar detector in the RGB images the hypothesis of whether the window is or is not an apple was subsequently tested. For this purpose three classifiers of the type MLP (feed forward multi-layer perceptrons) were used. Each was trained and tested by splitting a sample set of 751100 vectors of dimension three. The dataset was constructed using the following procedure: 1) a user selected and labeled manually rectangular regions of interest (sub-windows) from the color image dataset according to 5 classes: apples, leaves, branches, sky and ground, and 2) each selected window was resized to 10x10 pixels and the values of each of

the three channels of all pixel was stored as a set of 3D vectors. This process was repeated for three color models: L*a*b, HSV and RGB; and hence three datasets were obtained. Each classifier was trained and tested with a different dataset; therefore each classifier is used for one color space. The details of the datasets are given in Table 2. There are 3 such data sets , one for each of the color models.

Table 2. Dataset used to train the classifiers

Class	Sub-windows	Pixels
1 – apples	1416	141600
2 – leaves	2263	226300
3 - branches	1535	153500
4 – sky	1583	158300
5 – ground	714	71400
All	7511	751100

Each classifier had the same topology: 3-layer perceptron with 3 inputs, 5 outputs and two hidden layers including 100 neurons each. A symmetrical sigmoid activation function was used $f(x)=\beta*(1-e^{-\alpha x})/(1+e^{-\alpha x})$ with $\alpha=0.66$ and $\beta=1.71$. The training consisted of maximally 300 iterations resulting in the accuracies of 0.784, 0.78 and 0.782 for training and 0.782, 0.78 and 0.78 for testing, for the L*a*b, hsv and rgb classifiers respectively.

Since the accuracy values obtained using different classifiers are the same, in the next section a fusion approach is tested to see if an improved solution can be obtained.

Majority voting in classifier combination

One possible way of combining the output of the three classifiers is in an ensemble that is called “majority voting”. For a given triplet of values z , let define a classifier B_i that responds with an output vector y_i such that the entry $y_{ij}=1$ if z is classified as class j , otherwise 0. In our case $i=1,..,3$ and $j=1,..,5$. Lets define another type of classifier D_i that produces an output vector $[d_{i,1},..,d_{i,c}]$ where the value $d_{i,j}$ represents the base to the hypothesis that the sub-window w being tested on classifier i belong to class j . Each measurement level $d_{i,j}$ can be obtained by Equation 4.

$$d_{ij} = \frac{1}{|w|} \sum_{z \in w} B_{ij}(z) \quad (4)$$

For example, for window w_1 , the response vector $D_1=[0.2 \ 0.2 \ 0.1 \ 0.4 \ 0.1]$ means that 20% ,20% 10%, 40% and 10% of the pixels in the window belong to classes “apples”, ‘leaves”, “branches”, “ground” and “sky” respectively. However, to discriminate between true hits and false alarms, it is enough to classify the sub-window in two classes “apple” and “not apple”. Therefore vector $[d_{i,1},..,d_{i,c}]$ can be converted to a binary two dimensional vector $[e_{i,1},e_{i,2}]$ such that:

$$e_{i,l} = \begin{cases} 1, & \text{if } \sum_{j=1}^k d_{ij} > \sum_{j=k+1}^c d_{ij} \\ 0, & \text{otherwise} \end{cases}, \quad e_{i,2} = \bar{e}_{i,1} \quad (5)$$

where k is the partition index between classes, and c is the number of classes. For example, to consider the first two classes in one group (likely to be an apple), and all the rest in a different group (not likely to be an apple), $k=2$, $n=5$ and $i=3$.

Then, the majority vote scheme (Equation 6) determines the label L of the sub window detected by the 3 classifiers. The scheme is presented in Fig. 4.

$$L = \arg \max_{j=1}^c \sum_{i=1}^{\ell} d_{i,j} \quad (6)$$

The majority voting scheme was used to accept or reject the hypothesis about whether the detected sub-windows were or were not apples. In addition, two rules were implemented to accept a hypothesis: a) the detected window does not include sub-windows, b) the detected sub-window size is smaller than $k \cdot \text{median}(W, H)$, where we used $k=1.5$.

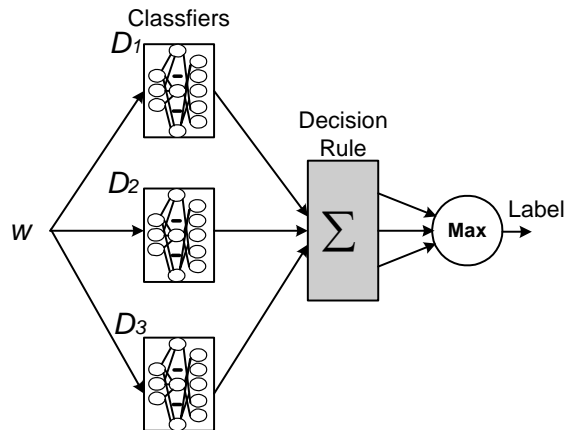


Figure 4. Classification combination scheme

Results & Discussion

The following subsections describe the performance of the multimodal apple detection system using first, the RGB and IR Haar detectors independently.

Results of the RGB Haar detector

To train the RGB detector, a set of 146 color images of apple trees was used which included a total of 9420 green apples under natural conditions. The classifier was tested on 34 images including 1972 apples. There were 30 stages in the detector's cascade, where each stage reached a hit rate of 0.995 with two splits, and its base resolution was 20x20 pixels. Fig 5 shows the detections found in a sub-region of a testing image.

The figure shows the classifier's ability to generalize apples (e.g. partially occluded with leaves, non-occluded, pits showing or not). False alarms were reduced using the voting scheme in classifier combination presented in Section 4. Table 3 presents the hits over the total number of apples, the missed apples over the total number of apples, and the false alarms when using the RGB Haar detector alone (single color space) and after adding the voting scheme (multiple color spaces). The voting scheme affected the correct detections only by less than 0.8% while dropping the FA rate by 7%.



Figure 5. Six apples detected by the RGB Haar detector

Results of the IR Haar detector

The apple detector classifier with IR images was trained with a training set of 286 images including 2330 apples from the same trees used to train the RGB Haar detector. Due to the lower resolution of the thermal camera, the area captured by the image is much smaller, and hence contained less apples. This classifier was trained with a cascade of 20 stages, with a minimum hit rate of 0.995 in each stage, with two splits and a base resolution of 24x24 pixels detection window. Fig. 6 shows apples detected in an IR sub-image.

The performance of this detector is given in Table 4 for stages 17-20. For each stage, the total number of apples, the total hits and the false alarms are presented. These results show the dependency between hit rate and false alarms. The cascade with 18 stages was used for the experiments, more stages decrease significantly the hit rate, while an increase yielded a drastic increase in the FA.

Table 3. Detection rate using the color Haar detector with and without the voting

	Hits	Missed	FA
RGB Haar	1326/197 2	646/1972	53 6
RGB Haar+Voting	1307/197 2	646/1972	53 6

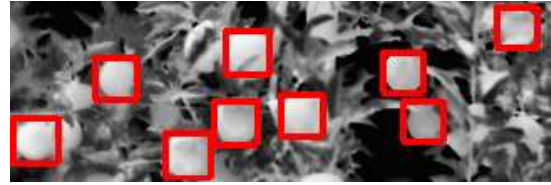


Figure 6. Nine apples detected by the IR Haar detector

Results of combined Color-IR scheme

The detected hits resulted from the voting scheme are added to the output of the IR Haar detector after applying the transformation parameters. First, the registration parameters for each pair of images (color and IR) are found using mutual information. Then, the RGB and IR Haar detectors are applied to the color and infra-red images respectively. Later, the affine transformation is applied to the set of detections obtained using the IR Haar detector. Finally, the total number of detections is the sum of both sets, RGB and IR. The apples considered for the detection in this step are those found in the common area between the color and IR images.

The results are presented in Table 5 when applied to 34 pairs of testing images. The combination approach shows that the recognition accuracy was increased (74%) compared to the conventional approach of detection using either the color (66%) or the IR (52%) modalities alone. One interesting feature of the methodology is that the three main processes: registration, Haar feature detection in RGB and IR are independent and hence can be easily parallelized by assigning each process to a different CPU.

Table 4. Hit rate and false alarms per stage of the Haar detector

Stage #	Hits	Missed	FA
17	274/504	0.456	80
18	263/504	241/504	61
19	245/504	259/504	51
20	231/504	273/504	47

Table 5. Performance when using single and combined modalities

Modality	Hits	Missed	FA
Color+Voting	1307/1972	665/1972	498
IR	263/504	241/504	61
Combined	679/913	234/913	344

Conclusions

We presented an algorithm for apple detection in natural scenes using a multimodal approach. Initially the optimal registration parameters are obtained using maximization of mutual information and are stored for later use. Then, Haar features in color and infra-red images are

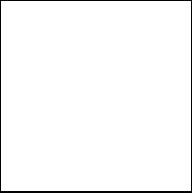
obtained through an Adaboost algorithm. Later, a voting scheme is used to improve the detection results. Finally, the detection results are fused after applying the best transformation found in the first step. A contribution reported in this work, is the voting scheme added to the output of the RGB Haar detector which drops the false alarms with little effect on the recognition rate. The resulting classifiers alone can partially recognize the on-trees apples however when combined together the recognition accuracy is increased. Although the algorithm did not detect all apples and contains false alarms, the main concern is its implementation in a robotic fruit picking scenario. In this case, the performance of the algorithm seems to be a sufficient for prepositioning a robot picking arm. Since images will be acquired from cameras mounted on the robotic arm which can be oriented to take close up pictures, gradually all the apples in the tree can be found and false alarms can be identified as the robot arm explores the canopy. In spite of the relatively low recognition accuracy, this is the first system, to our knowledge, that can deal with “green” apple detection, that are partially occluded with shadow patterns, from a tree canopy of “green” leaves, branches, and sky background. Future work will include increasing the robustness of the Haar classifiers by increasing the sample set, and incorporate morphologic information to the voting scheme.

Acknowledgments

This research was supported by Research Grant No US-3715-05 from BARD, The United States - Israel Binational Agricultural Research and Development Fund, and by the Paul Ivanier Center for Robotics Research and Production Management, Ben-Gurion University of the Negev.

References

- Annamalai, P. and Lee, W.S., 2003. Citrus Yield Mapping System Using Machine Vision. ASAE Annual International Meeting. Paper number 031002
- Fernandez-Maloigne, C., Laugier, D. and Boscolo, C., 1993. Detection of apples with texture analyses for an apple picker robot. Intelligent Vehicles '93 Symposium. 1993: 323-328.
- Freund Y. and Shapire, R.E., 1996. Experiments with a new boosting algorithm. In Machine Learning: Proc. of the 13th Int. Conf., 148-156.
- Jiménez, A.R., Ceres, R. and Pons, J.L., 2000a. A vision system based on a laser range-finder applied to robotic fruit harvesting. Machine Vision and Applications. Springer Berlin / Heidelberg. Volume 11, Number 6 / May, 2000. 321-329.
- Jiménez, A.R., Ceres, R., and Pons, J.L., 2000b. A survey of computer vision methods for locating fruit on trees. Trans. ASAE 43(6): 1911-1920.
- Lienhart, R. and Maydt, J., 2002. An Extended Set of Haar-like features for Rapid Object Detection. In Proc. of the IEEE Conf. on Image Processing (ICIP '02), pp. 155-162.
- Sapina, R., 2001. Computing textural features based on co-occurrence matrix for infrared images. Proceedings of 2nd International Symposium on Image and Signal Processing and Analysis, 2001. ISPA 2001. 373-376.
- Stanjnko, D., Lakota, M. and Hocevar, M., 2004. Estimation of number and diameter of apple fruits in an orchard during the growing season by thermal imaging. Computers and Electronics in Agr. 42: 31-42.
- Tabb, A.L. Peterson, D.L., and Park, J., 2006. Segmentation of Apple Fruit from Video via Background Modeling. ASABE Annual International Meeting, 2006. Paper number 063060
- Viola, P. and Jones, M.J., 2004. Robust real-time face detection. International Journal of Computer Vision, 57(2):137-154.
- Viola, P. and Wells, W.M. III, 1995. Alignment by maximization of mutual information. In Proc. 5th Int. Conf. Computer Vision, June 1995. 16-23



Zhao, J., Tow, J. and Katupitiya, J., 2005. On-tree fruit recognition using texture properties and color data. IEEE/RSJ Int. Conf. Intell. Robots and Systems, 263- 268