Engineering Applications of Artificial Intelligence 77 (2019) 59–69

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

MangoNet: A deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard

Ramesh Kestur[a,*,1], Avadesh Meduri[b,1], Omkar Narasipura[a]

[a] Department of Aerospace Engineering, Indian Institute of Science, Bangalore, India
[b] Department of Mechanical Engineering, Birla Institute of Technology and Science, Pilani, India

ARTICLE INFO

Keywords:
Deep learning
Semantic segmentation
Object detection
Mango detection
Convolutional neural networks
Fully convolutional networks

ABSTRACT

This work presents a method for detection and counting of mangoes in RGB images for further yield estimation. The RGB images are acquired in open field conditions from a mango orchard in the pre-harvest stage. The proposed method uses MangoNet, a deep convolutional neural network based architecture for mango detection using semantic segmentation. Further, mango objects are detected in the semantic segmented output using contour based connected object detection. The MangoNet is trained using 11,096 image patches of size 200 × 200 obtained from 40 images. Testing was carried out on 1500 image patches generated from 4 test images. The results are analyzed for performance of segmentation and detection of mangoes. Results are analyzed using the precision, recall, F1 parameters derived from contingency matrix. Results demonstrate the robustness of detection for a multitude of factors such as scale, occlusion, distance and illumination conditions, characteristic to open field conditions. The performance of the MangoNet is compared with FCN variant architectures trained on the same data. MangoNet outperforms its variant architectures.

1. Introduction

India is the worlds largest producer of Mango (Magnifera Indica L). India contributes over 52% of global production of approximately 23 million metric tonnes. 30% of the farmers sell the produce to pre-harvest contractors (Srikanth et al., 2015). The value of the crop is determined by estimating the projected crop yield. Hence the crop yield estimate is critical in assessing the monetary value of the produce and therefore has a direct bearing on the monetary value that the farmer can realize from the crop. Current yield estimates of mango crop are manual. The manual process typically involves manual sampling of blocks in conditions of heat and humidity typical of tropical weathers in which mangoes are grown. Such a manual process limits the accuracy of assessment of the whole orchard making the entire manual process tedious, labor intensive, time consuming and error prone with a high degree of variability.

A machine vision based system could replace the manual system. Several sensor technologies such as LIDAR, thermal infrared, multi-spectral and vision based (RGB) imaging. Lidar based techniques have higher accuracy and hence better localization accuracy, however they are bulky and costly. The most favored solution is the use of RGB imagery to match human abilities to perceive fruit within a canopy on the basis of color, geometry and texture (Qureshi et al., 2017).

Traditional machine vision approaches to fruit detection are based on pixel features, shape, texture features or integrated approaches. These approaches have been used to classify a variety of fruits. Schillaci et al. (2012), Roy et al. (2011), Sengupta and Lee (2014), Diago et al. (2012), Hung et al. (2013) and Kurtulmus et al. (2014) used shape and color features to identify fruit pixels in images of tomato, pomegranate, citrus, grape, apples, green almonds and peach fruits on plants respectively. Chaivivatrakul et al. (2010), Chaivivatrakul and Dailey (2014) and Moonrinta et al. (2010) used texture based features to identify fruit regions in images of pineapple plants. Further, integrated approaches (Gongal et al., 2015) that combine both pixel features and shape features provide better performance as compared to single feature based classification

Image processing approaches to fruit detection are based on color (Qiang et al., 2014), geometry (Senthilnath et al., 2016) and texture features (Zhao et al., 2005; Chaivivatrakul and Dailey, 2014). Image classification algorithms use these features to create hand engineered features to encode visual attributes that discriminate fruit from non fruit regions. Payne et al. (2014, 2013) analyzed night time

[*] Corresponding author.
E-mail addresses: rkestur@gmail.com (Ramesh K.), avadesh02@gmail.com (Avadesh M.).
[1] Both Authors contributed equally to the work.

https://doi.org/10.1016/j.engappai.2018.09.011
Received 16 April 2018; Received in revised form 7 August 2018; Accepted 16 September 2018
Available online xxxx

acquired imagery to detect apple fruits. Color based segmentation using RGB an YCb Cr color space was used in image segmentation. Wang et al. (2013) used Hue saturation and value in the HSV color space to detect red and green apples. Zhao et al. (2005) used RGB/HIS color space to detect both green and red apples on a tree. Thresholded values of red–blue and red green–red was used to develop a color model to detect the fruits (Dorj et al., 2017). Detected and counted citrus fruits using orange color was detected by thresholding of RGB to HSV converted images. Further watershed algorithm was used to count the fruits. These works utilize hand engineered features to encode fruit and non fruit regions. Wachs et al. (2010) used unsupervised K-means clustering based on a and b channels of from L*A*B color space to detect green apples from background. Qureshi et al. (2017) used supervised learning based K Nearest Neighbor (KNN) to count fruits in images of mango tree canopies. KNN was used to classify pixels based on color and smoothness followed by contour segmentation and detection using an elliptic shape model. Supervised learning based fruit detection include soft computing methods such as ANN and SVM. Nanaa et al. (2014) used a back propagation neural network based method to recognize mango fruits in images of mango fruit trees. Qiang et al. (2014) used multi-class SVM method for detection of citrus fruits.

From the authors review of literature it is observed that image processing methods use color and color threshold based methods. However, variable lighting conditions affect the intensity of reflected light and hence the performance of the color and threshold based models reduce. One of the options to address variable illumination is the use of controlled lighting in orchards such as night time imaging. However, this limits the operation time in commercial operations and makes the acquisition operations complex and costly. The occlusion of fruits by leaves, branches and fruit clustering affect the geometry based models. Further, the features used in image segmentation algorithms are hand crafted features. Hand crafted features are image and data specific and hence are not robust.

In recent years, there is considerable research interest in the application of deep convolutional neural networks for fruit detection. Deep learning based vision systems identify objects by recognizing their unique features implicitly unlike the hand crafted features in traditional image processing methods. Bargoti and Underwood (2017) used object detection based deep framework for mango counting. Deep learning methods such as Object detection (Bargoti and Underwood, 2016) and Semantic segmentation (Bargoti and Underwood, 2017) are applied in mango detection. The object detection based systems comprise of two steps. In the first step region proposal are carried out. Region proposal involves identification of regions in the image that have a high probability of containing objects. In the second step, the proposed regions are input to an R-CNN (Ren et al., 2015) for predicting objects in the region. Positive predicted object regions are resized to enclose only the objects. The region proposal step is achieved through non-learning algorithms such as selective search (Uijlings et al., 2013) and edge boxes (Zitnick and Dollár, 2014). These non-learning algorithms propose regions by measuring the number of super pixels and edges respectively. Hence these non-learning algorithms would propose irrelevant regions on our image dataset as each image contains many leaves, stones and trees apart from mangoes, resulting in false positives. Further, R-CNNs are computationally intensive and detection time depends on the number of objects leading to longer runtimes for images with more number of mangoes.

In semantic segmentation methods, classification is carried out at a pixel level (Long et al., 2015). Semantic segmentation is achieved, either by labeling single pixel at a time or by labeling all pixels simultaneously. In a single pixel labeling semantic segmentation approach, classifiers are trained using a neighborhood region of the pixel, also called a patch. Bargoti and Underwood (2017) used patch based semantic segmentation. Multi layer Perceptron(MLP) and Convolutional neural networks (CNN) were used for classification of pixels. The patch based segmentation achieved better mango detection accuracy as compared to traditional approaches. However they are sensitive to occlusion and varied illumination. Further, patch based segmentation label one pixel at a time resulting in increased runtime.

In semantic segmentation methods involving simultaneous classification of pixels, typically Fully Convolutional Neural network (FCN) (Long et al., 2015) are utilized. FCNs accept variable input image sizes and output the corresponding predicted segmented image. FCNs have achieved promising results in semantic segmentation challenges like the PASCAL VOC (Everingham et al., 2010).

In this work we propose a deep learning framework for detection and counting of mangoes in an open mango field. The proposed method uses a novel deep semantic segmentation architecture, The MangoNet, to segment mangoes in RGB images. The MangoNet is capable of segmenting images of different input sizes. The proposed method displays robustness to various conditions of illumination, scale mango density and occlusion. Further, the run time of detection using the proposed method is invariant to the number of mangoes in the image as the MangoNet classifies pixels simultaneously during semantic segmentation.

The Mangoes are detected in the image by identifying the locations of the connected objects in the semantic output from the MangoNet using contour detection, as each connected object in the semantic output represents a predicted mango in the original input image. Mango Detection is further improved by post processing methods. Candidate proposals that include noise are eliminated and candidate proposals which are enclosed by larger sized candidate proposals are eliminated. Counting results are realized by comparing the predicted regions with annotated bounding boxes. Performance of the proposed method is evaluated by calculating mango detection accuracy and F1 score. The performance of the MangoNet is measured by its semantic segmentation results. A pixel wise F1 score is calculated to evaluate the MangoNet's segmentation results. The MangoNet's performance is compared with FCN variant architectures trained on the same data. MangoNet performs better as compared to its variant architectures.

The rest of this document is organized as follows. Section 2 contains a review of Convolutional neural networks and Fully convolutional networks. The methodology is discussed in Section 3. The evaluation scheme used to measure the performance of the MangoNet and proposed method, is detailed in Section 4. The results and inference are discussed in Section 5. The conclusion and future scope are detailed in Section 6.

2. Convolution Neural Network (CNN)

This section includes an over review of a convolution, the CNN architecture and FCN which are the basis for the proposed MangoNet architecture.

An Artificial Neural Network (ANN) is a system of interconnected neurons used to model non linear relationships. The basic ANN model is composed of three types layers. An input layer, one or more hidden layers and an output layer.

A CNN is type of ANN in which neurons are locally connected. Further, CNN contains convolutional layers instead of input and hidden layers. A convolution block is a convolutional layer, followed by a non linear activation function and a pooling layer. A deep CNN comprises a stack of convolution blocks. Deep CNNs progressively train high level feature representations of input data (Masci et al., 2011; Simonyan et al., 2013).

A convolution layer is given by:

$$x_j^l = f(\sum_{k=1}^{m}(x_k^{l-1} * W_{kj}^l) + b_j^l) \quad (1)$$

where * is a convolution operation, x_k^{l-1} defines the kth feature map of the $l-1$th layer, x_j^l is the jth feature map of the lth layer and M is the number of input feature maps. W_{kj}^l are the trainable weights also called as kernel or filter and b_j^l is the bias which is also trainable. f is