|  |
| --- |
|  |
| NUTRITION ANALYSER – FRIUTS |
| Proffesional Readiness For Innovation ,  Employablity And Enterpreneurship. |
|  |
| **TEAM ID,**  **IBMTMID48053.**  **TEAM LEADER,**  **S.Thanigaivel,**  **TEAM MEMBERS,**  **R.Alageshwari,**  **T.Anbumozhi,**  **R.pavithra,** |
|  |

|  |
| --- |
|  |

INTRPODUCTION

According to [[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B1-sensors-16-01222)], sourcing skilled farm labour in the agriculture industry (especially horticulture) is one of the most cost-demanding factors in that industry. This is due to the rising values of supplies, such as power, water irrigation, agrochemicals, and so on. This is driving farm enterprises and horticultural industry to be under pressure with small profit margins. Under these challenges, food production still needs to meet the growing demands of an ever-growing world population, and this casts a critical problem to come.

Robotic harvesting can provide a potential solution to this problem by reducing the costs of labour (longer endurance and high repeatability) and increasing fruit quality. For these reasons, there has been growing interest in the use of agricultural robots for harvesting fruit and vegetables over the past three decades [[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B2-sensors-16-01222),[3](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B3-sensors-16-01222)]. The development of such platforms includes numerous challenging tasks, such as manipulation and picking. However, the development of an accurate fruit detection system is a crucial step toward fully-automated harvesting robots, as this is the front-end perception system before subsequent manipulation and grasping systems; if fruit is not detected or seen, it cannot be picked. This step is challenging due to various factors, among which are illumination variation, occlusions, as well as the cases when the fruit exhibits a similar visual appearance to the background, as shown in [Figure 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/figure/sensors-16-01222-f001/). To overcome these, a well-generalised model that is invariant and robust to brightness and viewpoint changes and highly discriminative feature representations are required.

### Fruit Detection Using Faster R-CNN

Despite the recent progress being made using deep convolutional neural networks on large-scale image classification and detection [[5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B5-sensors-16-01222)], accurate object detection still remains a challenging problem in the computer vision and machine learning fields. This task requires not only detecting which objects are in a scene, but also where they are located. Accurate region proposal algorithms thus play significant roles in the object detection task.

There are recent works, such selective search [[19](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B19-sensors-16-01222)], which merges super pixels based on low-level features, and EdgeBoxes [[20](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B20-sensors-16-01222)], making use of edge information to generate region proposals. However, these methods require as much running time as the detection to hypothesise object locations. Faster R-CNN [[21](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B21-sensors-16-01222)] was proposed to overcome this challenge by introducing the Region Proposal Network (RPN), which shares convolutional features with the classification network, and two networks are concatenated as one network that can be trained and tested through an end-to-end process. By doing that, the running time for region proposal generation takes around 10 ms, and this framework can maintain a 5 fps detection rate and outperform the state-of-the-art object detection accuracy using very deep models [[30](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B30-sensors-16-01222)

[[An external file that holds a picture, illustration, etc.
Object name is sensors-16-01222-g003.jpg](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g003.jpg)](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g003.jpg" \t "tileshopwindow)

[[An external file that holds a picture, illustration, etc.
Object name is sensors-16-01222-g004.jpg](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g004.jpg)](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g004.jpg" \t "tileshopwindow)

[Figure 4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/figure/sensors-16-01222-f004/)

(**a**) The 3 × 3 (pixels) Conv164 filters of the RGB network from VGG, (**b**) The input data and (**c**) One of the feature activations from the conv5 layer. The cyan boxes in (b) are manually labelled in the data input layer to highlight the corresponding fruits of the feature map.

Four thousands ninety six dimensions of feature vectors are extracted from the fully-connected 7 (fc7) layer and are fed into t-Distributed Stochastic Neighbour Embedding algorithm (t-SNE) [[32](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B32-sensors-16-01222)] with the corresponding labels. t-SNE is one of the popular dimensionality reduction methods that measures pairwise neighbouring similarities using the L2 norm distance in both high and low dimensions. The pairwise similarities are calculated around the sample points, and the Kullback–Leibler divergence is used to gauge the distance between two probability distributions (i.e., the similarities of high and low dimensions). Stochastic Gradient Decent (SGD) minimises the distance to keep the local structure in a low dimension space. [Figure 5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/figure/sensors-16-01222-f005/) shows low dimension (2D) feature visualisation using t-SNE. Each point represents a feature, and its colour is the corresponding label. It is obvious that sweet peppers (green) and rock melons (blue) are highly distinguishable from each other and the background (in red). This figure also shows that good detection results are expected given a reasonable classifier.

[[An external file that holds a picture, illustration, etc.
Object name is sensors-16-01222-g005.jpg](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g005.jpg)](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g005.jpg" \t "tileshopwindow)

**DeepFruits Training and its Deployment**

The data that we have are multi-modal, colour (RGB) and NIR in nature, and so, we fine-tune (adapt) the Faster R-CNN for each modality independently. Fine-tuning consists of updating, or adapting, the model parameters using the new data. In practice, this involves initialising a new classification layer and updating all of the layers, for both the region proposal and classification network. The classification network uses the same architecture as VGG [[30](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B30-sensors-16-01222)], as this provided the best performance.

The VGG network configuration used (Configuration D) consists of 13 convolutional layers followed by two fully-connected layers, referred to as VGG-D. The original implementation of Faster R-CNN was fine-tuned using the PASCAL VOC dataset (20 objects, 11 k images and 27 k annotated objects), and the network was initialised by the pre-trained ImageNet dataset, which consists of 1000 object categories, 1.2 million images and their bounding box annotations [[5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B5-sensors-16-01222)]. This implies that we are required to fine-tune again the network using our custom data; otherwise, Faster R-CNN can only detect the 20 ordinary objects on which the network was trained, such as aeroplane, bicycle, bird, cat, dog, and so on. By doing this, we can make use of features learned from a large-scale dataset which are well generalised to various visual recognition tasks.

Given the VGG-16 network, we define three classes named ‘background’, ‘sweet pepper’ and ‘rock melon’ and fine tune the network. Regarding this fine-tuning topic [[35](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B35-sensors-16-01222)], abundant resources are available from online [[36](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B36-sensors-16-01222)], and we also have made publicly available our implementation and tutorial document [[6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B6-sensors-16-01222)].

[Table 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/table/sensors-16-01222-t001/) shows the number of training images used by CRF and Faster R-CNN only for the performance evaluation. We can only use a relatively small number of images due to the limited pixel-wise image annotation datasets from [[4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B4-sensors-16-01222)]. For a fair comparison, the same training and testing images are utilised, and the experimental results are presented in [Section 4.2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#sec4dot2-sensors-16-01222). We also conduct further experiments by increasing the number of classes and training images to detect another fruit and to demonstrate its generalisation.

### Table 1

Number of images used for training and testing for CRF and Faster R-CNN.

|  | **Train** | **Test** | **Total** |
| --- | --- | --- | --- |
| **(RGB + NIR)** | **(RGB + NIR)** |
| CRF and Faster R-CNN | 100 (82%) | 22 (18%) | 122 |

[Open in a separate window](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/table/sensors-16-01222-t001/?report=objectonly)

After the training, we deploy the trained fruit detector on a laptop that has Intel i7, 64-bit 2.90 GHz quad-core CPUs, a GeForce GTX 980M 8 GB GPU (1536 CUDA cores) and 16 GB of memory space running on an Ubuntu 14.04 Linux system. Input images are obtained from a multi-spectral camera, the JAI AD-130GE, and a Microsoft Kinect 2. Each camera has a resolution of 1296 × 964 and 1920 × 1080, respectively. Processing for the detection takes an average of 341 ms with a 4 ms standard deviation for JAI and 393 ms with 3 ms for the Kinect 2 image. The processing time gap is caused by an external library for reading different resolution images.

### 3.4. Multi-Modal Fusion

In the previous section, we introduced the proposed fruit detection approach using the Faster R-CNN framework; here, we present the two methods, late and early fusion, that we use to combine the multi-modal (RGB and NIR) imagery that we have. Late fusion combines the classification decisions from the two modalities. Early fusion alters the structure of the input layer of the VGG network so that 4 channels, rather than 3, are provided.

#### 3.4.1. Late Fusion

Late fusion combines the classification information from the two modalities, colour and NIR imagery. Using the independently-trained models for each modality (see [Section 3.2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#sec3dot2-sensors-16-01222)), we combine the classification information in the following manner.

Each modality m produces Nm,P region proposals. To combine the two modalities, these region proposals are combined to form a single set of NP\* = m × Nm,P region proposals. A score **s**m,p is then proposed for the p-th proposed region of the m-th modality. A single score for the p-th region is produced by averaging the response across the modalities,

sp=∑m=1Msm,p

(1)

The score is a C-dimensional variable, where C is the number of classes to be classified.

#### 3.4.2. Early Fusion

Early fusion alters the structure of the input layer of the VGG network so that the input data layer has Nc = 4 channels (3 channels from RGB and 1 channel from NIR), rather than Nc = 3. The VGG network is modified and adapted to receive RGB and NIR information simultaneously. An overview of this is provided in [Figure 6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/figure/sensors-16-01222-f006/). To achieve this, we duplicate the R response from the VGG-D network and initialise the extra, NIR channel using this; the R channel (620–750 nm) is chosen, as it is closest to the NIR channel’s wavelength (750–1400 nm). This early fusion network is then fine-tuned as previously described.

[[An external file that holds a picture, illustration, etc.
Object name is sensors-16-01222-g006.jpg](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g006.jpg)](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g006.jpg" \t "tileshopwindow)

## 4. Experimental Results

In this section, we qualitatively and quantitatively evaluate our proposed method on five experimental settings: (1) we compare the early and late fusion performance; (2) we evaluate the performance between the baseline algorithm (CRF) and the proposed method; (3) we inspect the performance of RPN; (4) we exam the generalisation of the proposed method by performing spatial-temporal independent condition experiments; (5) we evaluate the extensibility of the proposed approach by applying it to several other fruits.

Prior to presenting the experimental results, we mention the creation of the ground truth of the dataset. [Figure 7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/figure/sensors-16-01222-f007/)a depicts hand-labelled bounding boxes (yellow) based on the colour image and the NIR image. In [Figure 7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/figure/sensors-16-01222-f007/)b, the cyan colour box missing from [Figure 7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/figure/sensors-16-01222-f007/)a highlights the missing annotation of a sweet pepper in the NIR image due to its poor visibility; whereas it is more obvious to see the sweet pepper in the RGB image because of the reflection from the sweet pepper. This also happens the other way around. A fruit in the dark is difficult to see in the RGB-based image, but can be identified easily in an NIR image. We thus merge these two ground truth sources using both RGB and NIR images by computing the pairwise Intersection of Union (IoU) of bounding boxes shown in [Figure 7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/figure/sensors-16-01222-f007/)c. The remainder of this article refers the merged ground truth as merged GT, and the other two ground truths are referred to as RGB GT and NIR GT based on the image sources used for making the ground truth.

[[An external file that holds a picture, illustration, etc.
Object name is sensors-16-01222-g007.jpg](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g007.jpg)](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g007.jpg" \t "tileshopwindow)

Given this threshold, the precision (P), recall (R) and F1 score are computed as:

P=TPTP+FP,R=TPTP+FN,F1=2⋅P⋅RP+R

(2)

where TP is the number of true positives (correct detections), FP is the number of false positives (false detection), FN is the number of false negatives (miss) and TN is the number of true negatives (correct rejection).

### 4.1. Early and Late Fusion Performance Comparison

Multi-modal visual sensing techniques are widely used in the agricultural field because they can often capture necessary signatures utilised for detection. We present the results of our proposed early and late fusion methods introduced in [Section 3.4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#sec3dot4-sensors-16-01222). The specifications of the training and testing dataset are shown in [Table 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/table/sensors-16-01222-t001/).

[Open in a separate window](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/table/sensors-16-01222-t002/?report=objectonly)

### 4.2. Fruit Detection Performance Comparison with CRF and Faster R-CNN

As previously mentioned in [Section 3.1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#sec3dot1-sensors-16-01222), fruit detection performance evaluation is conducted between CRF and the fine-tuned Faster R-CNN. We use the same training and test settings as described in [Table 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/table/sensors-16-01222-t001/). The only difference is that the pixel-annotated training set is utilised as the ground truth for CRF training, while bounding box annotations are used for Faster R-CNN (see [Figure 2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/figure/sensors-16-01222-f002/)). The ground truth for test images remains identical. We should note that the output from CRF is a pixel-level likelihood map representing how much the pixel belongs to a specific label. In order to have a fair comparison with the bounding box outputs of Faster R-CNN, we use a Laplacian of Gaussian (LoG) multi-scale blob detector [[37](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/#B37-sensors-16-01222)] for the CRF-based method to produce detected fruit regions (i.e., bounding boxes).

### Table 3

F1 scores of CRF and fused networks.

| **CRF** | **Early Fusion** | **Late Fusion** |
| --- | --- | --- |
| 0.807 | 0.799 | 0.838 |

[[An external file that holds a picture, illustration, etc.
Object name is sensors-16-01222-g010.jpg](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g010.jpg)](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g010.jpg" \t "tileshopwindow)

Performance evaluation of the region proposals of four different networks.

This study also explains the incomplete curve. The maximum performance all networks can achieve is less than one due to the limited performance of RPN. For CRF, it can only achieve 0.75 of the maximum detection rate.

The proposals’ generation time demonstrates the efficiency of this method; the results are presented in [Table 4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/table/sensors-16-01222-t004/). It can be seen that the computation time increases almost proportionally as the number of proposals increases.

### Table 4

Region proposal generating time including detection.

|  | **# Prop.** | **10** | **50** | **100** | **300** | **500** | **1000** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Net.** |  |
| RGB network (in s) | | 0.305 | 0.315 | 0.325 | 0.347 | 0.367 | 0.425 |
| Early fusion (in s) | | 0.263 | 0.268 | 0.291 | 0.309 | 0.317 | 0.374 |

[Open in a separate window](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017387/table/sensors-16-01222-t004/?report=objectonly)

### 4.4. Impact of the Number of Training Images for Fruit Detection

We presented the impact of the number of proposals versus the fruit detection rate in the previous section. In this section, we address the study of the impact of the number of images being used for retraining the network. The aim of this study is to demonstrate the performance of the fruit detector by varying the number of training image sets.

It is known that more training images leads to better performance under the framework of the deep convolutional neural network. The same trend can be seen in  It is, however, interesting that the impact of fine-tuning shows impressive results, as only 10 training images are utilised for retraining, but produce a 0.7 F1 score. Using 50 images (green) yields a slightly lower F1 score than using 25 images (blue), but it coversa wider recall area with higher precision. In addition, this study also implies that it is feasible to achieve better performance with more fruit images.

[[An external file that holds a picture, illustration, etc.
Object name is sensors-16-01222-g011.jpg](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g011.jpg)](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g011.jpg" \t "tileshopwindow)

[[An external file that holds a picture, illustration, etc.
Object name is sensors-16-01222-g012.jpg](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g012.jpg)](https://www.ncbi.nlm.nih.gov/core/lw/2.0/html/tileshop_pmc/tileshop_pmc_inline.html?title=Click%20on%20image%20to%20zoom&p=PMC3&id=5017387_sensors-16-01222-g012.jpg" \t "tileshopwindow)

Number of images used for training and testing for different fruits.

| **Name of Fruits** | **Train (# Images), 80%** | **Test (# Images), 20%** | **Total, 100%** |
| --- | --- | --- | --- |
| Sweet pepper | 100 | 22 | 122 |
| Rock melon | 109 | 26 | 135 |
| Apple | 51 | 13 | 64 |
| Avocado | 43 | 11 | 54 |
| Mango | 136 | 34 | 170 |
| Orange | 45 | 12 | 57 |

**CONCLUSION**

We present approaches for a vision-based fruit detection system that can perform up to a 0.83 F1 score with a field farm dataset, maintaining fast detection and a low burden for ground truth annotation. This is a competitive result compared to our previous pixel-based detector of 0.80. We also demonstrated qualitative results to show how well the trained model using a small dataset generalises to entirely independent (unseen) environments.

In developing this system, we performed fine-tuning of the VGG16 network based on the pre-trained ImageNet model. The novel use of RGB and NIR multi-modal information within early and late fusion networks provides improvements over a single DCNN. Furthermore, we investigated the performance of region proposal networks to narrow down a possible bottleneck of performance degradation. Our findings are returned to the relevant communities through an open dataset and tutorial documentation.

Future work involves the integration of the proposed algorithm with our custom-built harvesting robot and the collection of an enormous amount of ground truth annotations for a variety of fruits by utilising Amazon Mechanical Turk or other out-sourcing supplies to achieve more accurate performance.