Orchard Fruit Segmentation using Multi-spectral Feature Learning

Calvin Hung, Juan Nieto, Zachary Taylor, James Underwood and Salah Sukkarieh

***Abstract*— This paper presents a multi-class image segmentation approach to automate fruit segmentation. A feature learning algorithm combined with a *conditional random field* is applied to multi-spectral image data. Current classification methods used in agriculture scenarios tend to use hand crafted application-based features. In contrast, our approach uses unsupervised feature learning to automatically capture most relevant features from the data. This property makes our approach robust against variance in canopy trees and therefore has the potential to be applied to different domains. The proposed algorithm is applied to a fruit segmentation problem for a robotic agricultural surveillance mission, aiming to provide yield estimation with high accuracy and robustness against fruit variance. Experimental results with data collected in an almond farm are shown. The segmentation is performed with features extracted from multi-spectral (colour and infrared) data. We achieve a global classification accuracy of 88%.**

I. Introduction

The world's growing population demands an increase in food production. According to a number of studies, production must double by 2050 to meet these demands [1], [2]. There are several problems to achieve this target with current farming practices. First of all, there is a shortage in labour in rural areas, mainly due to migration and urbanization. Secondly, an increase in production combined with climate change evokes a need for innovation in current farming methods, in order to make agriculture a sustainable practice. Specialty crops (fruits and vegetables, tree nuts, dried fruits, horticulture, and nursery crops) are particularly labour demanding. Automation is a technology that is expected to have huge impact in farming, by increasing efficiency in production and reducing labour cost.

This paper presents an approach for automatic fruit segmentation. Reliable and accurate multi-class segmentation is a crucial component underlying higher-level robotic perception tasks. In the context of agricultural robotics, the multi-class image segmentation algorithm allows the robot to, for example, understand the environment, estimate yield health and calculate production [3].

Traditionally, in the context of vegetation classification in remote sensing, indices such as Normalised Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) are commonly used. These indices highlight the target object and allow simple classifier such as thresholding to extract and segment vegetations. Following a similar idea, the work presented in [3] presents a pipeline for yield estimation for apple orchards based on a set of carefully designed rules. Although these approaches present promising results, they require the redesign of the rule sets for every new application, since they cannot handle either, within or across class variabilities.

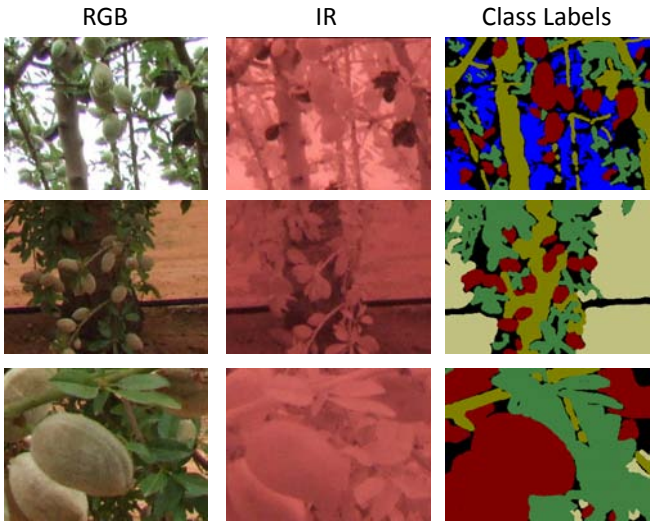

Fig. 1. The RGB+IR Dataset: Example of multi-spectral image and multi-class image segmentation: In multi-class image segmentation each pixel in the image is assigned to a class label. The right column shows the RGB images, the middle column is the corresponding observation in the IR band, the left column is the hand labelled multi-class ground truth. The multi-class segmentation results can be used for change detection, planning, fruit detection and ultimately yield estimation.

This paper takes a different approach by using feature learning to automatically obtain the rule sets from the data itself, instead of using fixed pre-defined features descriptors. This makes our approach suitable to a variety of crops since it can inherently handle variance. The algorithm presented applies feature learning within a *conditional random field* (CRF) framework. This approach simplifies the training process by removing the need for learning separately the feature models and the weights required to represent the feature importance. Feature learning has been applied successfully in vision and robotic applications and achieves state-of-the-art performance in object detection, image classification and object recognition tasks [5]. Existing approaches typically utilize spatial pooling of image statistics and are therefore only suitable for per-image classification rather than per-pixel multi-class image segmentation. To perform per-pixel classification, this work collects high-level spatial information via multi-scale features that are learnt from generic unlabeled

This work is supported in part by the Australian Centre for Field Robotics at the University of Sydney and Horticulture Australia Limited through project AH11009 Autonomous Perception Systems for Horticulture Tree Crops.

Calvin Hung, Juan Nieto, Zachary Taylor, James Underwood and Salah Sukkarieh are with Australian Centre for Field Robotics, School of Aerospace, Mechanical and Mechatronic Engineering The University of Sydney, NSW 2006 Australia c.hung, j.nieto, z.taylor, j.underwood, s.sukkarieh@acfr.usyd.edu.au

image datasets at different resolutions.

The specific contributions of our work include:

- A general semi-supervised approach for segmentation. The approach couples feature learning with Markov fields.
- Multi-scale feature learning for RGB-IR data. The learning approach is flexible, allowing the proposed algorithm to incorporate different data modality.
- A comprehensive evaluation using real data. The algorithm was evaluated with data collected in an almond farm.

The rest of the paper is organised as follows. Section II presents a summary of related work. Section III describes our approach including an overview of image segmentation using undirected graphical models (such as the CRF) and unsupervised feature learning. Section IV provides an overview of the experimental setup to validate the approach. Results are presented in Section V and in Section VI a discussion on the results and comparison to the related work in fruit segmentation is provided. We finally present our conclusions in Section VII.

II. Related Work

Existing fruit segmentation work focuses on grapes [6] [7], mangoes [8], oranges [9] and apples [3], [10]. These applications have the characteristic that the fruit can in general be well distinguished using a basic colour model or shape constrains. The work on this paper presents a study on segmenting almond fruits. Almonds are particularly difficult because they have similar colour to the branches, similar size and shape to the leaves, and the fruit size varies widely in the image frame depending on the fruit position within the tree.

The majority of the fruit segmentation studies focus on a binary approach, i.e. fruit vs. others segmentation problem. While the fruit class is the most important information for yield estimation, other classes such as leaves and branches can offer additional information about the general health of the plantation. The work presented in [6] tackles the multi-class (fruit, leaves and branches) problem using vision and depth data. In contrast, our work aims to evaluate classification accuracy using vision-only systems. We present comparative studies using colour (RGB) and infrared (IR) data.

A common trend in orchard tree classification is to use colour and shape features [11]. The classifiers used vary widely according to the application. For example, the work presented in [10] uses a simple intensity threshold while [8] uses colour threshold. Rule set approaches using fuzzy logic are presented in [9] and hybrid approaches using colour threshold followed by shape check are used in [3]. Colour-based apple segmentation followed by smoothing using erosion and dilation is presented in [12]. Finally, the authors in [7] propose the use of shape based detection followed by a colour and texture classifier for grape detection.

To summarise, most of the existing work is domain specific, exploiting particular properties (colour, texture, shape) of the target fruit to be classified. The framework proposed here has been designed with the aim of automatically adapting the feature sets for different fruits; the feature extraction and classification rules are all obtained via learning. Therefore our approach does not require domain specific assumptions and can be applied to different types of trees.

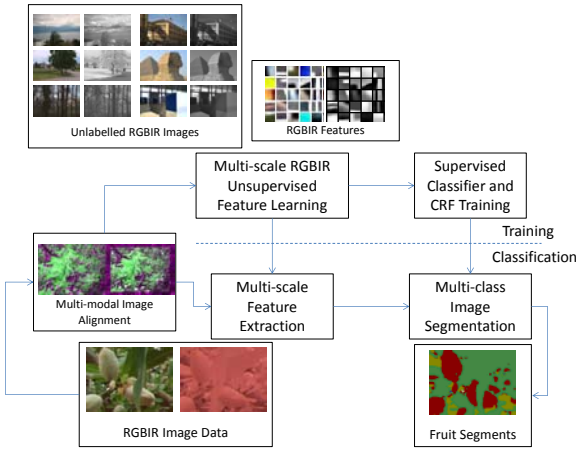

Fig. 2. Algorithm Overview: The multi-scale features are learned using unsupervised feature learning from a public RGB-IR dataset. The logistic regression classifier and the CRF are trained on our labelled dataset. The learnt models are used to perform image segmentation.

III. Algorithm

It has been shown that in classification problems, the algorithms perform better on meaningful feature descriptors instead of classifying the raw (noisy) data. For example, in RGB image segmentation problems it is standard to perform classification using colour, texture and shape features. There exists a large set of RGB feature descriptors from the computer vision community, however for the RGB-IR dataset there is currently no known feature set. Therefore, in order to learn informative RGB-IR features we apply unsupervised feature learning on a publicly available dataset [13].

An option for object classification is to perform image segmentation using pixel classification. The disadvantage is that the per-pixel classification approach does not take into account the correlation between neighbouring pixels in the image. This paper takes a similar approach to [6] by modelling correlations in the image using a CRF framework.

An overview of our approach is shown in Fig. 2. The different processing blocks are explained in the following section.

A. Image Modelling using Conditional Random Fields

Our approach models the image data using graphical models, in particular we use a CRF framework [14]. The graphical model for an image consists of a two dimensional lattice $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$ where $\mathcal{V}$ is a set of pixels representing the vertices of the graph and $\mathcal{E}$ is a set of edges modeling the relationships between the neighbouring pixels.

Image segmentation is performed by assigning to every pixel $x_i \in \mathcal{V}$ in the image a meaningful label $l_i \in \mathcal{L}$. For

multi-class image segmentation, the label set $\mathcal{L}$ may contain multiple labels up to k classes $\mathcal{L} \in \{1, ..., k\}$. The optimal labeling $l^\star$ is obtained via energy minimisation on the graph structure $\mathcal{G}$ with the energy function defined as in Eq. 1

$$E(\mathbf{l}) = \sum_{i \in \mathcal{V}} \psi_i(l_i, x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(l_i, l_j, x_i, x_j) - log(Z(\mathbf{x})) \tag{1}$$

where $\psi_i(l_i, x_i)$ is the unary potential which models the likelihood of a pixel taking a certain label, $\psi_{ij}(l_i, l_j)$ is the pairwise potential which models the assumption that the neighbouring pixels should take the same label, and $log(Z(\mathbf{x}))$ is the partition function.

Conventionally the unary potential is computed using the features in the image, for example grey level intensity [15], colour [16], or texture [17]. The unary potential $E_d(l)$ can be written as

$$E_d(l) = \sum_{i \in \mathcal{V}} \psi_i(l_i, x_i) = \sum_{i \in \mathcal{V}} \sum_{feat} w_{feat} \psi_{feat}(x_i | \theta_{feat}) \tag{2}$$

where the subscript $feat$ corresponds to features and w_{feat} models the relative importance of the individual feature functions. Our approach uses semi-supervised learning to learn the feature descriptors, as described in Section III-C.2.

The smoothness energy $E_s(l_i, l_j)$ measures the coherence of the neighbouring pixel labels, where $\mathcal{N}$ is the set of unordered pairs i, j of neighbouring pixels in $\mathcal{P}$. The smoothness term $E_s(l_i, l_j)$ penalises the neighbouring pixels with similar features from taking different labels as shown in Eq. 3.

$$\begin{aligned} E_s(l_i, l_j, x_i, x_j) &= \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(l_i, l_j, x_i, x_j) \\ &= \sum_{(i,j) \in \mathcal{E}} w_{i,j}(x_i, x_j) \cdot (l_i - l_j) \end{aligned} \tag{3}$$

The weight $w_{i,j}(x_i, x_j)$ measures the similarity between neighbouring pixels. In this paper, $w_{i,j}(x_i, x_j)$ is calculated using the Euclidean distance of the pixels x_i and x_j in colour space [16]. Weights $w_{i,j}$ are higher when the two pixels are similar and are used to penalise the neighbouring pixels from taking different labels. The $(l_i - l_j)$ term ensures that the smoothness energy is zero when the neighbouring pixels have the same label.

B. Feature Learning

The state-of-the-art approach to obtain good representations of the data is via feature learning. Feature learning methods originate from deep learning [4], [18] in which a set of trainable modules implementing complex non-linear functions are stacked together to first capture the underlying structure in unlabelled data, then a classifier layer is added to learn the label association for classification. Semi-supervised feature learning has been successfully applied to various computer vision problems and currently achieves state-of-the-art performance in digit classification [19], object detection and RGBD object recognition [5].

This section describes our feature learning approach. We first apply a sparse autoencoder [4] at different scales to obtain the multi-scale RGBIR dictionaries. A logistic regression classifier is then used to learn the label association to the multiscale responses. The classifier output is then passed into the CRF as the unary term described in Eq. 1 for multi-class image segmentation.

1) Unsupervised Feature Learning: Unsupervised feature learning captures the features from an unlabelled dataset. In this paper a sparse autoencoder is used to learn the dictionaries from randomly sampled RGBIR image patches. A sparse autoencoder minimises squared reconstruction error with an extra sparsity constraint [4] as shown in Eq. 4,

$$\begin{aligned} J(W,b)_{sparse} &= [\frac{1}{2m} \sum_{i=1}^{m} (\| h^k_{W,b}(x^{(i)}) - x^{(i)} \|^2)] + \dots \\ & \frac{\lambda}{2} \sum_{l=1}^{n_l - 1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_l + 1} (W^{(l)}_{ji})^2 + \beta \sum_{j=1}^{s2} KL(\rho \| \hat{\rho}_j) \end{aligned} \tag{4}$$

where the first term is the reconstruction error, which is the difference between the input $x^{(i)}$ and output $h^k_{W,b}$. By minimising the reconstruction error a mapping function which reconstructs the input at the output is learnt. The second term is the regulariser and the third term $KL(\rho \| \hat{\rho}_j)$ is the Kullback-Leibler (KL) divergence between ρ and $\hat{\rho}_j$ to enforce sparsity ρ. Sparsity is the proportion of activated nodes to the total number of nodes within a layer. The sparsity constraint limits the amount of activations in the hidden nodes. With the sparsity constraint, the number of hidden nodes can be higher than the number of input and output nodes and only a small number of hidden nodes are activated at a given time. For further discussion of sparse autoencoders, the reader is referred to [20].

2) Semi-supervised Feature Learning: With the low dimensional dictionary code obtained using unsupervised feature learning, an additional supervised label assignment step can be used to train a classifier. In this paper a softmax regression classifier is used. Softmax regression was chosen because it can be formulated as a single layer perceptron and trained using back-propagation.

Instead of using the raw image data x, the hypothesis from the hidden layer $h^{(2)}$ is used as the input of the softmax regression classifier. The hypothesis from the hidden layer contains sparse features extracted during learning. The output hypothesis of the softmax regression is shown in Eq. 5

$$h_\theta(x_i) = p(l_i = k | x_i; \theta) = \frac{exp(x_i \theta_k)}{1 + \sum_{j=1}^{J} exp(x_i \theta_j)} \tag{5}$$

where x_i are the input features. In this case the multi-scale augmented features $l_i \in \{1, 2, 3...J\}$ are the labels. The classification is done by selecting the class with the most probable hypothesis.

The cost and its gradient are required to obtain the optimal classifier parameters θ. The cost function is shown in Eq. 6

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{l_i = j\} \log \frac{exp(\theta_j^T x_i)}{\sum_{l=1}^{k} exp(\theta_l^T x_i)} + \ldots \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2 \quad (6)$$

where $1\{\cdot\}$ is an indicator function such that when the argument is true the function outputs 1 and when the argument is false the function outputs zero. λ is a weight decay parameter used to avoid overfitting by regularising θ.

C. Multi-scale Feature Learning

Our application requires an algorithm that can handle differences in feature scale. One approach to incorporate the multiscale property is to use pooling methods. The pooling approach down-samples the feature response from the original image dimension, and has proven very useful for image/object classification problems where a label is given to an image [21]. Despite its utility in per image classification, pooling cannot be used for the image segmentation problem described in this paper where pixel-wise labelling is required.

To incorporate the multi-scale information without using pooling, the feature learning approach described in Section III-B.2 is explicitly performed at multiple scales. The output of the feature learning algorithm is incorporated into the CRF framework for multi-class image segmentation.

1) Multi-scale Unsupervised Learning: The first step of multi-scale feature learning is to use unsupervised feature learning on different scales. The multi-scale patches are extracted by down-sampling the same image to extract the training data at different resolutions.

The unsupervised training patches are collected randomly over the images across the entire training dataset. The sparse autoencoder is used to learn the multi-scale representation from the extracted multi-scale patches.

2) Multi-scale Semi-supervised Learning: The second step of the multi-scale feature learning is to use the ground truth training label to train the multi-scale softmax regression classifier. The training patches are collected randomly over the images across the entire training dataset, but this time the multi-scale patches are aligned by the centroids and the labels at the patch centroids are used as training labels.

D. Feature Learning and CRF

The energy function of the CRF (see Eq. 1) is the equivalent of the cost function in feature learning. The function that is obtained via semi-supervised learning models the likelihood of x given l, and is equivalent to the unary function in the CRF framework. Consequently, the two frameworks can be integrated and optimised together. The unary potentials in Eq. 2 are generated using the learnt feature descriptors. The overall framework of the CRF with likelihood defined by the learnt features is shown in Fig. 3.

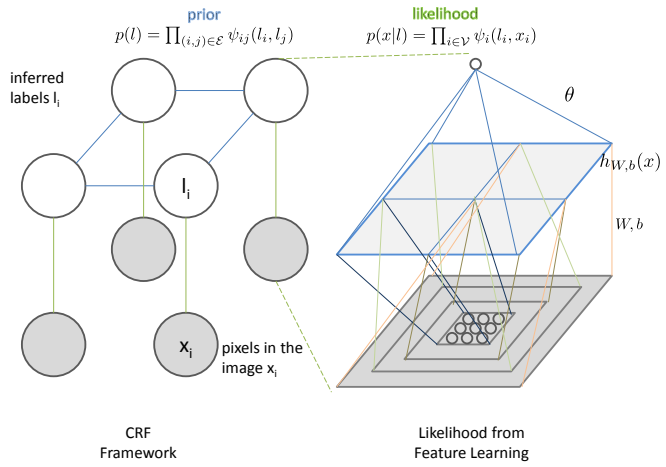

Fig. 3. Incorporating Feature Learning within a CRF Framework: x are the pixels from the image, $h_{W,b}(x)$ are the extracted multi-scale features and the classifier outputs the likelihood $p(x|l)$ to the CRF framework.

IV. Experimental Setup

The RGB-NIR Scene Dataset [13] was used for unsupervised feature learning. This dataset consists of 477 images captured in RGB and NIR. Random patches were sampled at 4 different scales (1, 1/2, 1/4, 1/8) to learn the multi-scale features.

Fig. 4 shows the platforms utilised to collect the experimental data in the almond orchard. The RGBIR images collected were used for supervised training and evaluation. This dataset consists of 1600 images at 320 $\times$ 240 resolution, of which 80 are labelled to pixel accuracy with five class labels, almond, trunk, leaves, ground and sky.

For the proposed algorithm two sets of results were computed, the first is the segmentation results based solely on feature learning and the second is the results with CRF. The CRF was trained with the Graphical Models/ Conditional Random Fields toolbox [22]. The unary potentials are generated using the fine-tuned multi-scale filters, and the pairwise potentials have the form of a contrast sensitive Potts model [15].

The existing fruit segmentation algorithms are application specific, since this paper presents the first study on almond fruit segmentation, there are no existing methods to compare. To present a quantitative comparison we implemented the feature set described in [23]. Our implementation uses the first 52 feature descriptors. The reason behind selecting this feature set was that it contains the most commonly used features including colour, texture, shape and super-pixel properties and is designed to be general and flexible to different segmentation problems. In comparison most existing work on fruit segmentation uses features ranging from RGB colour only to 34 dimensional colour and texture filters [7]. We expect the 52 dimensional benchmark implementation to be comparable to the state-of-the-art implementation used in fruit segmentation.

The second experiments use the entire RGB-IR dataset and the aim was to test if the extra IR channel improves the

Fig. 4. Robotic platforms used in the agricultural surveillance trial.

image segmentation accuracy.

Apart from the individual classification accuracy, three evaluation metrics were applied, the global, average accuracy and the F measure. The global accuracy measures the number of correctly classified pixels of the entire dataset whereas the average accuracy measures the average performance over all classes. The F measure was computed by averaging the F measure of the individual classes.

A. Multi-Spectral Image Registration

The RGB and IR images were taken with the same camera using an IR filter to change the modality captured. The original images were slightly misaligned due to wind affecting the trees and small disturbances affecting the cameras position and orientation between images. This misalignment had to be corrected before the images could be used together.

To correct for this misalignment we applied an affine transform to the IR image. Finding the optimal parameters for this affine transform is a challenging problem due to the difference in modalities between the images. This difference meant that standard mono-modal alignment techniques such as SIFT are unreliable, giving few matches and a significant number of outliers. Because of these issues we used normalized mutual information (NMI), a technique we had previously used to register multi-modal sensors [24]. Fig. 5 shows an example of the results obtained.

V. Results

This section shows the feature dictionaries obtained via feature learning, followed by the RGB based multi-class fruit segmentation results and the RGB-IR based segmentation results.

The feature dictionaries learnt using the sparse autoencoder in the unsupervised feature learning stage are shown in Fig. 6. The algorithm managed to capture the colour, colour gradient and edge filters from the RGB dataset, in addition because the training was done on all channels (RGB-IR) the correlations between the RGB and IR channels were also captured. These dictionaries were then used to extract feature descriptors for multi-class segmentation.

A. Using RGB channels

The RGB based multi-class segmentation results are shown in Fig. 7 and Table IV. The dataset consists of leaves,

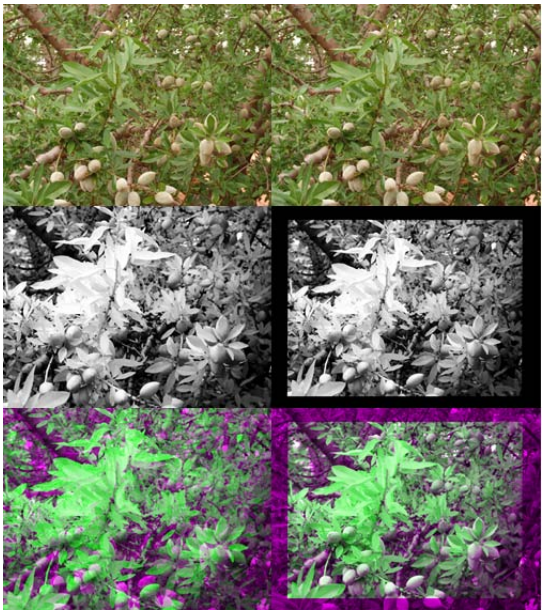

Fig. 5. One of the images before (left) and after alignment (right). RGB is shown at the top, IR shown in the middle and alignment (two image superimposed) shown at the bottom.

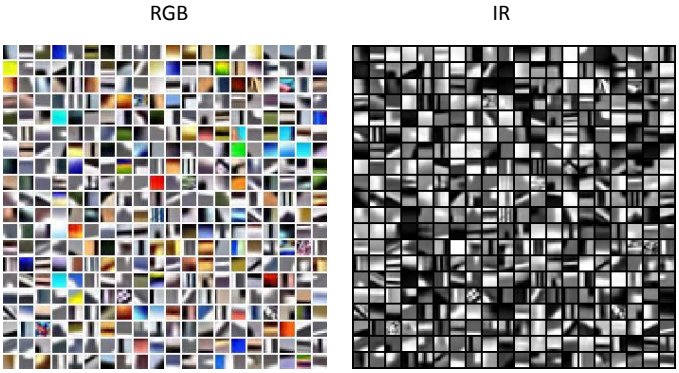

Fig. 6. Learnt Feature Dictionaries: The features on the left are learnt from the RGB channels, whereas the features from the right are learnt from the IR channel.

almonds and tree trunks at different lighting conditions and scales. The algorithm was able to segment various objects and also the background scenes (sky and ground). The proposed feature learning algorithm outperformed the benchmark [23] with and without the CRF. We did not apply CRF smoothing to the benchmark because it was already super-pixel based. The proposed algorithm was able to segment the fruits at different scales while the benchmark algorithm sometimes was unable to do so due to a few feature descriptors being super-pixel dependent. The global and average accuracy and the F measure of the proposed approach improved from 80.2% to 86.9%, 80.8% to 84.3% and 78.9% to 84.8% correspondingly by using CRF.

To investigate the sources of error the confusion matrices are also shown in Table II and Table III. It showed that the majority of error occurred between the almond and trunk classes. This was due to the spectral and textural similarity between the two classes.

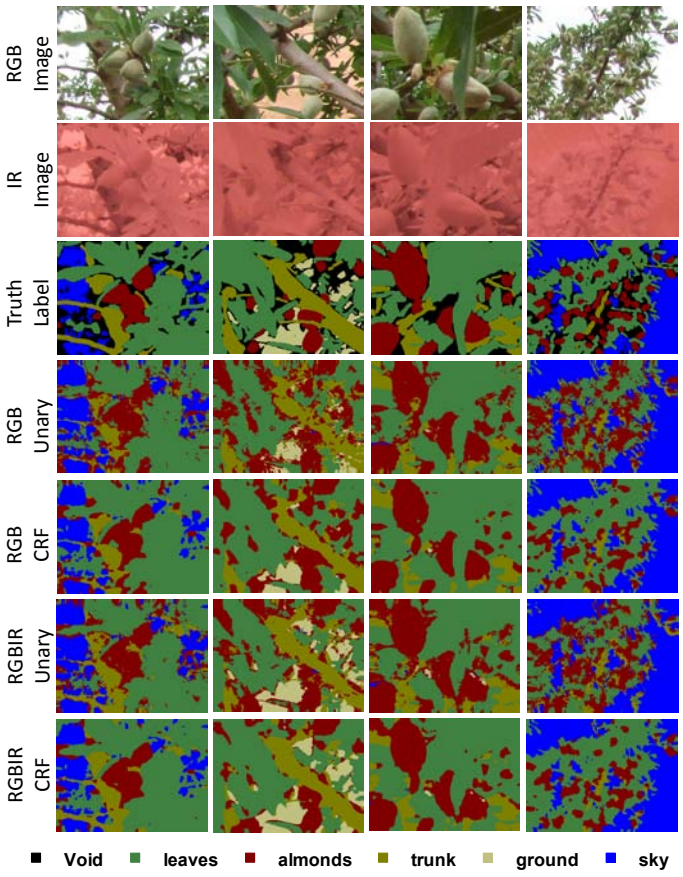

Fig. 7. Orchard qualitative results using RGB and RGB-IR channels: This dataset was collected during an orchard surveying mission aiming to automate the yield estimation and harvesting process. The dataset contains scenes of the almond trees at different zoom settings and lighting conditions.

	Leaves	Almonds	Trunk	Ground	Sky	Global	Average	F Measure
Hoiem [23]	91.9	54.6	55.8	81.0	93.0	79.9	75.3	77.7
This Work (Unary)	81.7	65.9	74.2	86.6	95.5	80.2	80.8	78.9
This Work (CRF)	94.0	69.8	72.8	89.1	95.8	86.9	84.3	84.8

TABLE I

ORCHARD RGB QUANTITATIVE RESULTS: COMPARISON WITH BENCHMARK ALGORITHM.

	leaves	**almonds**	**trunk**	**ground**	**sky**
leaves	81.7	11.1	7.6	0.6	1.1
almonds	8.8	65.9	14.4	3.4	1.5
trunk	7.9	18.3	74.2	9.4	1.9
ground	0.5	3.4	3.1	86.6	0.0
sky	1.1	1.3	0.7	0.0	95.5

TABLE II

RGB UNARY CONFUSION MATRIX

	leaves	**almonds**	**trunk**	**ground**	**sky**
leaves	94.0	16.1	12.7	2.4	2.9
almonds	2.7	69.8	10.5	1.8	0.6
trunk	2.5	11.7	72.8	6.6	0.7
ground	0.2	1.8	3.4	89.1	0.0
sky	0.5	0.6	0.6	0.0	95.8

TABLE III

RGB CRF CONFUSION MATRIX

	Leaves	Almonds	Trunk	Ground	Sky	Global	Average	F Measure
Unary	78.9	67.9	76.9	92.5	95.1	79.9	82.3	79.4
CRF	93.8	71.3	76.4	93.5	95.5	88.0	86.1	86.1

TABLE IV

ORCHARD RGB-IR QUANTITATIVE RESULTS.

B. Using RGB + IR

The segmentation results using all RGB-IR channels are shown in Fig. 7 and Table IV. The global, average accuracy and the F measure with CRF improved from 79.9% to 88.0%, 82.3% to 86.1% and 79.4% to 86.1% correspondingly. There are no known RGB-IR image segmentation algorithm so the results are compared to the RGB based segmentation.

To investigate the sources of error the confusion matrices are also shown in Table VI and Table VI. Compared to the RGB based confusion matrix, the misclassifcation between the almond and trunk class has reduced significantly by using the extra IR channel.

VI. DISCUSSION

The results showed that the proposed segmentation algorithm could be effectively used for fruit segmentation tasks. With the RGB channels it achieved a global accuracy of 87% and average accuracy of 84% (in comparison a random guessing algorithm on a 5 classes classification problem will have an accuracy of 20%). It outperformed the existing

	leaves	**almonds**	**trunk**	**ground**	**sky**
leaves	78.9	8.1	6.6	0.6	1.1
almonds	10.4	67.9	12.7	1.6	1.8
trunk	9.0	18.2	76.9	5.2	1.9
ground	0.5	4.4	2.8	92.5	0.0
sky	1.2	1.4	1.0	0.0	95.1

TABLE V

RGBIR UNARY CONFUSION MATRIX

	leaves	**almonds**	**trunk**	**ground**	**sky**
leaves	93.8	13.2	12.2	1.9	3.1
almonds	2.6	71.3	7.7	1.6	0.6
trunk	2.8	11.7	76.3	3.0	0.7
ground	0.2	3.2	3.2	93.5	0.0
sky	0.6	0.6	0.6	0.0	95.5

TABLE VI

RGBIR CRF CONFUSION MATRIX

benchmark segmentation algorithm [23] without the use of the CRF.

As expected the CRF increases the overall accuracy. By introducing the extra IR channel the segmentation accuracy improves further due to the increase in the discrimination power. The best performance achieved by the propose algorithm was using the RGB-IR data with the CRF, with 88% global accuracy, 86% average accuracy and the F measure of 86.1% .

A. Algorithm Strengths

The proposed algorithm is flexible in incorporating new data modalities. In contrast to the approach of designing new features for new data, applications or modalities, the feature learning approach learnt the new representation from the dataset itself, allowing the incorporation of the extra IR channel without the need of redesigning new feature sets.

This work also showed that by using multi-scale feature learning we were able to extend the feature learning based application from the standard image classification tasks to pixel-wise image segmentation tasks.

The multi-class segmentation approach extends the standard fruit/non-fruit binary segmentation to include other useful classes such as branches and leaves.

B. Limitations and Current Work

While the proposed algorithm provides reliable pixel labels, it does not incorporate the concept of objects and so does not provide the actual fruit count. The next stage of the project consists in estimating the fruit counts using the approach presented. The large amount of occlusion found may require extra sensing modalities. We are currently investigating the use of geometric features (e.g. using a stereo camera) to improve segmentation.

Currently the algorithm does not run in real time (5 seconds per image) because it was not the main requirement for this application. The aim was to use the autonomous platform to perform site surveillance and to provide offline yield estimation. However to extend this segmentation algorithm to other time-critical applications (such as autonomous fruit picking tasks) further work is required to speed up the process.

VII. Conclusions

This paper presented an approach for multi-class image segmentation applied to an almond fruit segmentation application. The proposed algorithm was able to perform multi-class fruit segmentation, with no artificial lighting, no prior assumptions on target fruit properties, and therefore is able to segment fruit with great variation in size. The multi-class approach was also trained to segment branches and leaves. The learning approach makes the algorithm more robust against variations such as changes in illumination. The proposed algorithm achieves state-of-the-art segmentation accuracy. While we demonstrated the performance on the almond dataset, the proposed approach is general and can be applied to other applications.

References

[1] U. Nations, "Food production must double by 2050," http://www.un.org/News/Press/docs/2009/gaef3242.doc.htm.

[2] GRID-Arendal, "The environment food crisis," http://www.grida.no/publications/rr/food-crisis/page/3457.aspx.

[3] Q. Wang, S. Nuske, M. Bergerman, and S. Singh, "Automated crop yield estimation for apple orchards," in *13th International Symposium on Experimental Robotics*, 2012.

[4] H. Lee, C. Ekanadham, and A. Ng, "Sparse deep belief net model for visual area v2," *Advances in neural information processing systems*, vol. 19, 2007.

[5] L. Bo, X. Ren, and D. Fox, "Unsupervised feature learning for rgb-d based object recognition," *ISER, June*, 2012.

[6] D. Dey, L. Mummert, and R. Sukthankar, "Classification of plant structures from uncalibrated image sequences," in *Applications of Computer Vision (WACV), 2012 IEEE Workshop on*. IEEE, 2012, pp. 329–336.

[7] S. Nuske, S. Achar, T. Bates, S. Narasimhan, and S. Singh, "Yield estimation in vineyards by visual grape detection," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*. IEEE, 2011, pp. 2352–2358.

[8] A. Payne, K. Walsh, P. Subedi, and D. Jarvis, "Estimation of mango crop yield using image analysis–segmentation method," *Computers and Electronics in Agriculture*, vol. 91, pp. 57–64, 2013.

[9] D. Bulanon, T. Burks, and V. Alchanatis, "Image fusion of visible and thermal images for fruit detection," *Biosystems Engineering*, vol. 103, no. 1, pp. 12–22, 2009.

[10] A. Aggelopoulou, D. Bochtis, S. Fountas, K. C. Swain, T. Gemtos, and G. Nanos, "Yield prediction in apple orchards based on image processing," *Precision Agriculture*, vol. 12, no. 3, pp. 448–456, 2011.

[11] A. Jimenez, R. Ceres, J. Pons *et al.*, "A survey of computer vision methods for locating fruit on trees," *Transactions of the ASAE-American Society of Agricultural Engineers*, vol. 43, no. 6, pp. 1911–1920, 2000.

[12] S. Singh, M. Bergerman, J. Cannons, B. Grocholsky, B. Hamner, G. Holguin, L. Hull, V. Jones, G. Kantor, H. Koselka *et al.*, "Comprehensive automation for specialty crops: Year 1 results and lessons learned," *Intelligent Service Robotics*, vol. 3, no. 4, pp. 245–262, 2010.

[13] M. Brown and S. Susstrunk, "Multi-spectral sift for scene category recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 177–184.

[14] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1068–1080, 2007.

[15] Y. Boykov and M. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 105–112.

[16] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3. ACM, 2004, pp. 309–314.

[17] L. Ladicky, C. Russell, P. Kohli, and P. Torr, "Associative hierarchical crfs for object class image segmentation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 739–746.

[18] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[19] D. Ciresan, U. Meier, J. Masci, L. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *International Joint Conference on Artificial Intelligence*, 2011.

[20] A. Ng, "Cs294a lecture notes: Sparse autoencoder," 2010.

[21] H. Lee, R. Grosse, R. Ranganath, and A. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 609–616.

[22] J. Domke, "Beating the likelihood: Marginalization-based parameter learning in graphical models."

[23] D. Hoiem, A. Efros, and M. Hebert, "Recovering surface layout from an image," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 151–172, 2007.

[24] Z. Taylor and J. Nieto, in *Australasian Conference on Robotics and Automation*.