Utilising Metadata to Aid Image Classification in Orchards

Suchet Bargoti and James Underwood[1]

***Abstract*— Accurate image scene parsing is a crucial component underlying high-level robotic perception tasks in agriculture. This is a difficult task when operating with orchard image data in outdoor scenes due to the undesirable intra-class variations caused by changes in illumination, pose, tree types etc. However, given the presence of structure, in both the orchard and how exactly the data was obtained, some factors of variation may be modelled via readily available metadata, including extrinsic experimental data such as the sun incidence angle, position within farm, etc. Using image classification based on a multi-scale Multi-Layered Perceptron, such metadata can be incorporated with the image data to aid scene parsing. Experimental results are shown for fruit segmentation and yield estimation over data collected at an apple orchard. The results show a 6% improvement in the classification f1-score, subsequently increasing the yield accuracy from 82% to 87%.**

I. INTRODUCTION

With recent advances in robotics and automation, it is possible to obtain low cost image data at high spatial and temporal resolutions over large farms. Reliable and accurate image processing techniques are required to obtain high level information of the crop status, such as its health and distribution. This can allow farmers to optimise precision agricultural operations according to variations in the field, which ultimately leads to maximising yield and quality.

Orchard image data, which is generally captured in the day time, is subject to large intra-class variation due to illumination conditions and crop type. Discriminative tasks such as fruit or tree segmentation [1], [2] need to be invariant to properties such as incident lighting, fruit maturity, tree types etc. Given enough training data and model complexity, a parametric model can be trained to be invariant to such factors. However, due to the costs associated with obtaining labelled data, this invariance is never completely met.

This paper provides a means to utilise prior information, which correlates with some of the observed intra-class variation, to aid image classification. This information, referred to here as metadata, is orthogonal to the image data and helps overcome some of the limitations mentioned above. In particular, the contributions of this paper are:

- A novel extension to a multi-scale feature learning algorithm previously presented in [3] by incorporating metadata relating to appearance variations in image data in order to aid discriminative classification.
- A study of different metadata at an apple orchard and its effects on image classification and yield estimation.

[1]The authors are with the Australian Centre for Field Robotics, The University of Sydney, 2006, Australia. s.bargoti, j.underwood @acfr.usyd.edu.au

The remainder of the paper is organised as follows. Section II presents the related work for scene parsing in outdoor conditions with a focus on agriculture data. Section III describes the standard classification framework and builds upon it with the inclusion of metadata. Sections IV and V contain the experiment setup and classification results, followed by apple yield estimation. We conclude in Section VI, with a discussion of future directions.

II. RELATED WORK

General purpose supervised feature learning algorithms learn an encoding of input image data into a discriminative feature space, and have been shown to outperform threshold based algorithms typically used in agriculture [3]. However, as mentioned before, in natural scene data, it is difficult to model the inter-class variations (i.e. differentiation between trees, leaves and fruits), while being invariant to intra-class (within-class) variability due to the naturally occurring extrinsic factors such as illumination, pose and tree type.

As a workaround, the intra-class variability in data can be restricted manually. For example, in [1] and [4], locally constrained classifiers are trained to learn specialised representations of orchard and urban data respectively. This can allow for parallel training operations but prevents underlying similarities in data splits to be shared between classifiers. Another approach can involve minimising the extrinsic variations during data gathering. For example, pepper detection is performed in [5] at a greenhouse plantation with controlled illumination conditions. In [6], [7], the data gathering is done at night using strobes to restrict the illumination variance. However, in orchards it is more practical to operate experimental systems under natural day-light conditions. Additionally, sensing hardware can be easily incorporated onto tractors, which typically operate during the day. This leaves image classification under natural conditions an open and important problem.

In this paper we propose that knowledge about the structure of orchards and the way in which data is obtained, can be used to explicitly model some of the underlying factors of variations. For example, when working with standard natural scene data such as PASCAL VOC [8], access to the generally unavailable prior knowledge such as illumination conditions and object and camera pose could assist image classification frameworks. The inherent structure of orchards provides some predictability in the illumination variation over the images. Additionally, the sun position can be used to evaluate the illumination incidence angle. Such Metadata, available at no extra cost, can allow the classifier to explicitly capture some aspects of the intra-class variation.

III. Classification Method

In this section we first present a binary (fruit/non-fruit) pixel level image classification algorithm, based on previous work in [3]. We then introduce the classification architecture used to incorporate orchard metadata.

A. Multi-scale Scene Parsing

The classification framework used in this paper is based on a multi-scale Multi-Layered Perceptron (MLP) consisting of three hidden layers. The classifier input represents a contextual window around each pixel in RGB space captured over multiple image scales. This provides scale invariance for classification and allows us to capture local variations at different scales such as the edges between fruits and leaves and between the trees and the skyline.

The MLP architecture is illustrated in Figure 1. Individual scale hidden representations are encoded via non-linear sigmoid transformation and then concatenated as per:

$$\psi^{(i)} = \bigcup_{s=1}^{S} \sigma(W_s x_s^{(i)} + b_s) \tag{1}$$

where, $x_s^{(i)}$ is the raw (or processed) RGB input for scale $s \in S$, $\sigma(z)$ is the sigmoid activation function and W_s are the set of filters/weights to be learnt. Patches from each scale are treated independently during the encoding phase. The weights are initialised with unsupervised pre-training, using a sparse De-noising Auto Encoder, which has been shown in literature (and through experimentation) to boost classification performance [9]. The concatenated output ψ is then propagated through a softmax layer to obtain class labels[1]. The MLP is trained via back-propagation while minimising a cross-entropy loss function, with an L_2 penalty term to minimise over fitting.

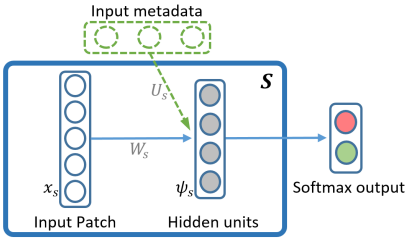

Fig. 1. The multi-scale Multi-Layered Perceptron architecture. The default setup is in blue, with x_s representing RGB patches captured at a given scale $s \in S$. Metadata configuration shown in green acts as a weighting to each of the hidden nodes.

B. Adding Metadata

Metadata corresponding to individual pixels can be incorporated to the multi-scale MLP architecture by appending the information at the input layer (shown in green in Figure 1). We can define each set of meta-data for a given input instance $x^{(i)}$, by $d_k^{(i)}$ $\forall k \in K$, where K is the set of different meta-data types, e.g. sun position, tree type etc. The different metadata can then be concatenated together as $D^{(i)} = \bigcup_{k=1}^{K} d_k^{(i)}$. The propagation to the hidden layer is then given by

$$\psi^{(i)} = \bigcup_{s=1}^{S} \sigma(W_s x_s^{(i)} + U_s D^{(i)} + b_s) \tag{2}$$

where, U_s are the weights learnt over each scale for the scale independent metadata input D. Qualitatively, U_s can be thought of as a weighting function for the different filters W_s, implicitly describing a relationship between each metadata component and the corresponding response from individual colour and edge filters.

The network is then trained using the same back-propagation algorithm as before. The computational expense of an MLP is linear to the number of input units. As the size of the input metadata is significantly smaller than the colour data, the additional computational expense is negligible.

IV. Experimental Setup

The image data was acquired over different rows of an apple orchard in Victoria, Australia, by Shrimp, which is a general purpose research ground vehicle, built at the Australian Centre for Field Robotics. Each tree image was 1616 by 1232 pixels. The training data (Figure 2) was collected by randomly sampling 1100 sub-images (308×202), corresponding to 1% of the entire data, and manually generating pixel-level labels for the fruit and non-fruit classes. Sub-sectioning the raw image allows for easier labelling and results in greater spatial variance within the training set.

Fig. 2. Sample images randomly extracted from over the orchard dataset for training. Images are ordered according to height in the original data.

Training patches of size $[8 \times 8 \times 3]$ were extracted over scales $[1, 1/2, 1/4, 1/8]$ with balanced random sampling over the labelled dataset. The MLP was configured with 50 hidden units per scale and trained using Stochastic Gradient Descent (SGD) with a learning rate and an L_2 penalty of 0.05 and 1×10^{-5} respectively (evaluated via a grid search and cross-fold validation). The algorithm was implemented using the open-source machine learning library, Pylearn2 [10].

The metadata used included a combination of pixel positions, orchard row numbers and the sun's azimuth angle relative to the vehicle body frame (calculated by using the time of day and vehicle's pose and geographical position). These properties were hypothesised to correlate with some of the intra-class variations, such as image height and illumination changes (Figure 2) and row numbering and changes in fruit/tree variety. A one-hot encoding was used to discretise continuous data into a number of discrete channels. A few channel sizes were tested and ultimately 8 channels were used for pixel i, j positions (p_i, p_j) and the sun azimuth angle (s_ψ). The row numbers (r_n) could be encoded directly as one-hot vectors.

[1]The process can therefore easily be extended to the multi-class scenario.

V. Classification Results

For evaluation, the labelled dataset was randomly split into an $80:10:10$ split of training, validation and testing images. The classification was repeated over multiple iterations, the three sets shuffled randomly each time. Training was done in the default configuration and with different combinations of metadata information. The results are reported in Table I.

TABLE I

Fruit classification and yield estimation results using the default method with combination of metadata such as pixel position (p_i, p_j), row number (r_n) and sun azimuth (s_ψ).

Config	F1-score	r-squared	Yield Est Acc (%)
None	0.683 ± 0.008	0.68 ± 0.01	$81.6\% \pm 0.3$
p_i	$+0.032 \pm 0.005$	–	–
p_j	$+0.000 \pm 0.002$	–	–
r_n	$+0.011 \pm 0.004$	–	–
s_ψ	$+0.001 \pm 0.003$	–	–
p_i, r_n	$+0.038 \pm 0.005$	–	–
p_i, p_j, r_n, s_ψ	$+0.042 \pm 0.005$	0.78 ± 0.02	$86.8\% \pm 0.8$

There is a clear improvement in classification results with the inclusion of all of the metadata increasing the f1-score by 6.1% (F1: $0.683 \rightarrow 0.725$). On it's own, p_i (pixel height) was the most important metadata, which also qualitatively portrayed the largest appearance changes in the data. Additionally, if some metadata was hypothesised to have no direct correlations with the extrinsic variations in data (such as p_j), the classification results were no worse. The lack of information from the sun position (s_ψ) is possibly due to the fact that in this dataset the sun azimuth was either $+90^\circ$ or -90° and therefore did not cause much variation on its own to the image data. The magnitude of learnt weights U_s was weaker for irrelevant metadata, guiding us towards the important factors of variations, which in turn can be used to improve subsequent data gathering tasks.

A. Yield Estimation

A desire for improving classification results was to perform accurate yield estimation. Due to occlusion in the data, even a perfect classifier cannot directly observe the true number of apples. Instead, we assume that on average there is a fixed ratio of occluded to visible fruit allowing us to map variations in yield over the farm and perform yield estimation given calibration data. For this data, the grower counted and weighed the post-harvest produce of 15 rows individually, which provided ground truth (this is too labour intensive for a commercial orchard to routinely perform).

Images were captured at 5 Hz and down-selected with a 0.5 m spacing along the rows to avoid double counting in subsequent frames. Yield estimation was done by linearly regressing true counts with pixel counts[2] accumulated along each whole row, as done in [1]. The yield estimate results were evaluated over multiple training iterations and are shown in Table I. The default classification method resulted in an r-squared value of 0.69, comparable to 0.656 reported in [1][3]. With the inclusion of metadata, the r-squared value increased to 0.78. The regression model was then used to estimate the yield per row, resulting in a yield estimate accuracy, which increased from 82% to 87%. The addition of metadata comes at no cost, and allows a greater number of fruits to be detected with fewer false positives.

[2]Erosion and dilation were used to filter out classification noise.

VI. Conclusion and Future Work

We have presented an image classification approach, which utilises extrinsic information corresponding to intra-class variations to produce more accurate classification results. The pixel-wise classification algorithm was based on a multi-scale Multi-Layered Perceptron within which, metadata relating to observed variations in data was incorporated. We evaluated the system over fruit classification at an apple orchard, using freely available information such as pixel position, row numbering and sun position as metadata. As a result, image classification performance improved by 6.1% to $F1 = 0.73$, and yield estimation accuracy over multiple rows improved from 81.6% to 86.8%.

Future work will involve incorporating additional metadata such as weather conditions, fruit types and seasons, covering over a larger orchard dataset. Further verification of the approach will also be conducted by extending to other domains where similar forms of metadata may be available. Additionally, image classification and the effects of metadata will be explored over different supervised learning architectures such as convolution or deeper networks.

References

[1] C. Hung, J. P. Underwood, J. I. Nieto, and S. Sukkarieh, "A Feature Learning Based Approach for Automated Fruit Yield Estimation," in *Fsr*. Springer, 2013, pp. 1–14.

[2] S. Bargoti, J. P. Underwood, J. I. Nieto, and S. Sukkarieh, "A Pipeline for Trunk Detection in Trellis Structured Apple Orchards," *Journal of Field Robotics*, May 2015.

[3] C. Hung, J. Nieto, Z. Taylor, J. Underwood, and S. Sukkarieh, "Orchard fruit segmentation using multi-spectral feature learning," in *IEEE International Conference on Intelligent Robots and Systems*, 2013, pp. 5314–5320.

[4] K. Dang and J. Yuan, "Location Constrained Pixel Classifiers for Image Parsing with Regular Spatial Layout," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[5] Y. Song, C. a. Glasbey, G. W. Horgan, G. Polder, J. a. Dieleman, and G. W. a. M. van der Heijden, "Automatic fruit recognition and counting from multiple images," *Biosystems Engineering*, vol. 118, no. 1, pp. 203–215, Feb. 2014.

[6] S. Nuske, K. Wilshusen, S. Achar, L. Yoder, and S. Singh, "Automated visual yield estimation in vineyards," *Journal of Field Robotics*, vol. 31, no. 5, pp. 837–860, Sept. 2014.

[7] A. Payne, K. Walsh, P. Subedi, and D. Jarvis, "Estimating mango crop yield using image analysis using fruit at 'stone hardening' stage and night time imaging," *Computers and Electronics in Agriculture*, vol. 100, pp. 160–167, Jan. 2014.

[8] M. Everingham and Others, "The PASCAL Visual Object Classes Challenge 2010 Results," 2010.

[9] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why Does Unsupervised Pre-training Help Deep Learning?" *J. Mach. Learn. Res.*, vol. 11, pp. 625–660, Mar. 2010.

[10] I. Goodfellow and D. Warde-Farley, "Pylearn2: a machine learning research library," *arXiv*, pp. 1–9, 2013.

[3]The previous publication mistakenly reported an r-squared value of 0.81, which was in fact the r-value.