POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Strawberries picking scheduling: challenges in robotic harvesting

Laurea Magistrale in Automation and Control Engineering - Ingegneria dell'Automazione

Author: Chiara Bigi

Advisor: Prof. Paolo Rocco

Co-advisor: Prof. Amir Ghalamzan Esfahani

Academic year: 2021-2022

1. Introduction

The field of Agricultural Robotics has been increasing over the past years. The demand for new technologies has been imposed by different social, political, and economic factors that are an indication of how relying on human labour is not safe for the agricultural chain [1].
The global population keeps growing, meaning agricultural production is asked to increase as well. Nevertheless, the labour shortage has been a rising problem for the past decade. Being an agricultural worker comes with many risks, that have been aggravated by climate change, covid restrictions, and the recent consequences of the Ukranian conflict.
The objective of agricultural robotics is to improve both productivity and working conditions. These new technologies are already increasing production yields for farmers in various ways automating slow, repetitive, and dull tasks.
Collaborative robots are now commonly used for the collection of fruits. Labour represents the largest cost and a vast operational uncertainty for berry farmers. The major concern over robots picking fruit comes from harvesting soft crops such as strawberries which can easily be damaged or missed entirely. The state-of-the-art still lacks AI techniques for the scheduling of the collection, notwithstanding how teaching a machine to output an optimized order for the picking of fruit would complete the harvesting automation.
This research was conducted at the University of Lincoln. It first addresses the problem of identification of a strawberry in an image, and classification based on ripeness and occlusion properties. Furthermore, it proposes some improvements on a deep model for trajectory generation for the reach-to-pick task. The novelty of this work stands in the introduction of a Graph Attention Network for the prediction of strawberry harvesting scheduling order with human-like reasoning, to reduce failure in the picking. This scheduling decision and trajectory generation process was tested on a Franka Emika manipulator provided with a RealSense camera.

2. Deep-ProMP for Motion Planning

Deep Probabilistic Motion Planning (d-PMP) [2] is a model for the generation of trajectories for a reach-to-pick task starting from images of a cluster of strawberries. d-PMP is divided into two parts. The first consists of the encoding

of an RGB input into a low-dimensional latent space. Different baselines were tested for this operation, such as auto-encoder (AE), Variational AE (VAE), and Conditional VAE (cVAE). AE is an unsupervised artificial neural network composed of an encoder and a decoder. The encoder learns how to efficiently compress and encode data, instead, the decoder learns how to reconstruct the data back from the reduced encoded representation to a representation that is as close to the original input as possible. By design, the model reduces data dimensions by learning how to ignore the noise in the data. This has been exploited for multiple tasks, such as dimensionality reduction or recovery of information. A next step is done in VAEs, in which the latent space is represented by the mean and covariance of a distribution so that multiple samples can be generated from it. Further evolution is found in cVAE, which inserts label information in the latent space to force a deterministic constrained representation of the learned data.
In the second part of d-PMP, the latent embedded vector is mapped to the mean and covariance of the trajectory's weights' distributions using a multilayer perceptron (MLP) per joint. MPL is a neural network connecting multiple layers in a directed graph. The concept of perceiving a trajectory as a sample from a Gaussian-like distribution comes from the Probabilistic Movement Primitives (ProMP) paper [3], taking, in turn, the idea of trajectory parameterization of Movement Primitives (MP). A visualization of MP and ProMP functioning is in Figure 1.

The d-PMP model lacks joint correlation, which can be achieved by predicting the weights' distributions for the trajectories of all seven degrees of freedom of the manipulator as the output of the same MLP. For the training of such a model, it was collected a dataset containing the images of strawberry configurations and a set of ten trajectories to reach all the singular targets. The trajectories of each joint were converted into weights to extract their mean and covariance of every set. The weights are extrapolated with the hypothesis that the trajectories follow the trend of some Gaussian distribution. The information of the covariance matrix was compressed with principal component analysis

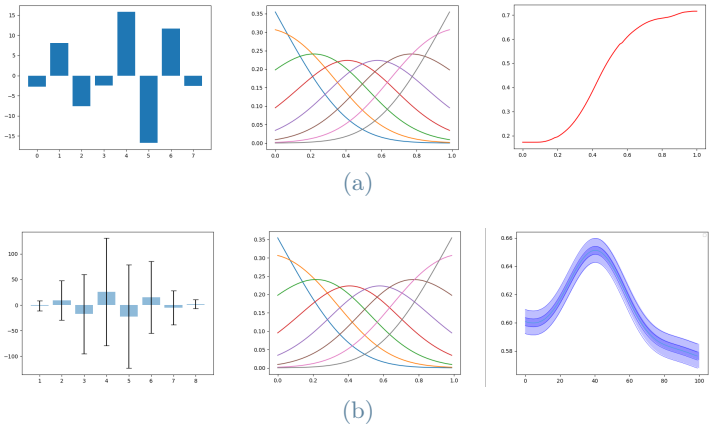


Figure 1: Starting from the left: weights (a) or weight distribution (b) of a trajectory for each basis function; Gaussian basis functions; trajectory (a) or trajectory distribution (b) of a single DoF

obtaining three vectors u, s and v. This allows retrieving the distribution (μ_T, Σ_T) of a trajectory of a robotic arm with seven joints j with only 189 elements, with the following equations:

$$\Sigma_{w_j} = u_j * s_j * v_j^T$$
$$\Sigma_w = diag(\Sigma_{w_1}, \dots, \Sigma_{w_7})$$
$$\mu_w = [\mu_1, \dots, \mu_7]^T$$
$$\Phi = diag(\phi)$$
$$\phi \quad \text{gaussian basis function}$$
$$\mu_T = \Phi^T \mu_w, \quad \Sigma_T = \Phi^T \Sigma_w \Phi$$

This improvement was only prepared by the author, but it was then implemented and tested on a thesis developed in parallel to the present work by another MSc student.

3. Scheduling Decision Problem

Visual detection of fruit is necessary for autonomous harvesting. The University of Lincoln made available for this work a dataset consisting of 2000 images of clusters of strawberries. For each "raw" image there is a corresponding JSON file containing the annotations about the ripe fruits. The structure of the annotation is composed by name and size (width×height) of the image file, and this information for each bounding box: pixel coordinates of the left down corner, width and height; an occlusion property chosen between "occluded", "occluding", "occluded/occluding", and "neither"; a scheduling label chosen by taking into consideration the po-

sition of the strawberry in the cluster and the occlusion properties.
End-to-end Object Detection with Transformer (DETR) [4] was shown to significantly outperform competitive baselines of state-of-the-art object detectors by using an innovative encoder-decoder architecture based on transformers - a deep learning tool that adopts the mechanism of self-attention. A DETR benchmark model was fine-tuned on the provided dataset to identify and classify the occlusion properties of strawberries. A pre-trained Detectron2 model [5] was instead exploited for the distinction between ripe and unripe berries.
To retrieve a harvesting order from an image, it is not enough to train a classifier by comparing images of strawberries with the same picking scheduling number. The detected berries are instead represented as the nodes of a connected graph having as features: an occlusion weight; a ripeness indicator; information of the patch of the image inside the bounding box, compressed with EfficientNet [6] (a competitive CNN classifier that takes images as input; to retrieve just the compressed information of the inputted image, the last layers are not used, since those are for the classification task). The edges between the nodes are represented by the pixel Euclidean distance between the bounding boxes of the fruits. These graphs are fed to a Graph Attention Network (GAT) [7], which exchanges and aggregates information about the node's features through the edges' connections. Furthermore, in these message-passing layers, each node has a self-attention mechanism that allows computing how much to attend to each neighbour. In this way, every strawberry learns its representation in the contest of the cluster.
The layers of this model are:

1. **Graph Convolutional Attention Layer**. It is a convolution, adapted to work on graph structures, that uses the self-attention mechanism to compute how every node attends to each neighbour.
2. **ReLu** activation function. The activation functions decide if a neuron in the neural network needs to be activated or not. They introduce non-linearity for the learning of more complex tasks. The Rectified Linear Unit (ReLu) is expressed as $f(x) = max(0, x)$.
3. **Dropout layer**. It randomly drops out neurons during training allowing for generalization. It is used to help prevent overfitting, which happens when a model adapts to the observations it was trained on, and it is unable to perform correctly on unseen data.
4. second **Graph Convolutional Attention Layer**.
5. **Linear Layer**.
6. **Sigmoid** activation function $f(x) = \frac{1}{1+e^{-x}}$, which allows all the output values to be between 0 and 1.

The reason for avoiding having too many convolutional layers is the risk of oversmoothing. If the message passing and aggregation are done too many times, every node will end up with the same information making it impossible to distinguish and classify them. The output of this GAT scheduling model is a node binary classification that assigns to every strawberry a probability of being the first target to be picked. The training for this model was done over a dataset with human-decided scheduling. This was confronted with a couple of heuristic scheduling computations. The first one favours isolated non-occluded berries. The second one aims to assign an absolute easiness score by multiplying the minimum isolation distance with an occlusion weight. This weight takes into consideration if the occlusion is by a leaf or by another berry, and the percentage of non-occluded area in the bounding box of the considered strawberry. A GAT score prediction model was trained on this latter heuristic computation, similar to the GAT classification one but without needing a Sigmoid activation function.

4. Results

This work provides a DETR model for the visual detection of the strawberries and the classification of their occlusion properties. The gain of using such a model with a self-attention mechanism will be underlined even more in the next chapter, but Table 1 shows its cons: without a huge amount of data and computational power state-of-the-art performances can never be achieved.
The trends between the average precision values are similar to the state-of-the-art. For example, the average precision values are higher for the

DETR-DC5 model	Benchmark	Customized
number of GPUs	8	1
dataset size	123k	2k
AP	43.3	21.2
AP_{50}	63.1	29.8
AP_{75}	45.9	24.1
AP_S	22.5	30.0
AP_M	47.3	35.0
AP_L	61.1	46.5

Table 1: DETR comparison. Benchmark data are taken from [4]. AP is the abbreviation of "average precision". The subscript number is the threshold of bounding box overlap. S, M and L stand respectively for small, medium and large.

bigger identified object. This is a known issue with DETR, already addressed in the original paper [4]. This matter is here somewhat discouraged. Since the given dataset has all the annotated bounding boxes "small", meaning smaller than a quarter of the image size, the model is trained more to detect little targets. Consequently, the AP_S value of the customized model outperforms the benchmark one.
For the specific application of this thesis research, these limited results were not a restriction since the training of the GAT scheduling model was done with the information in the annotations and not with the outputs of the perception models as it happens in the test exposed in Figure 6. Moreover, many other models for a perception task can be adapted to the provided strawberry dataset without the need for more data or GPUs. For example, the Detectron2 model used to add information on unripe berries is easily adaptable to detect occlusion properties as well.
For the scheduling problem, the first models were trained on data graphs without information on unripe berries or image patches. This allowed faster training to tune hyperparameters and try different configurations of layers and number of nodes.
The simpler GAT scheduling binary classification model can identify the first target to be picked from an image of a cluster of strawberries with an accuracy of 70%. Figure 2 shows the loss and accuracy trend during training and the value obtained for the best model with the

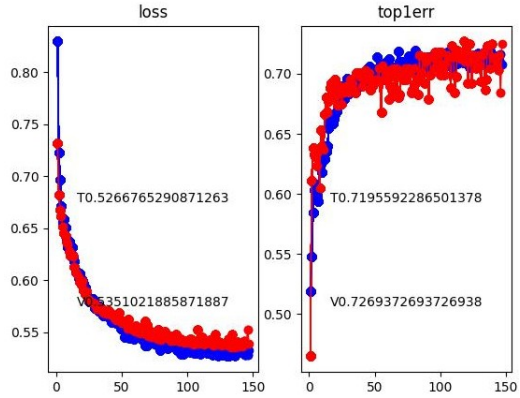
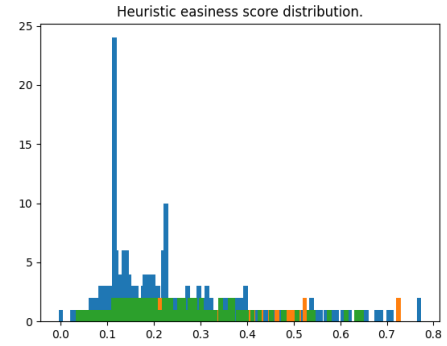
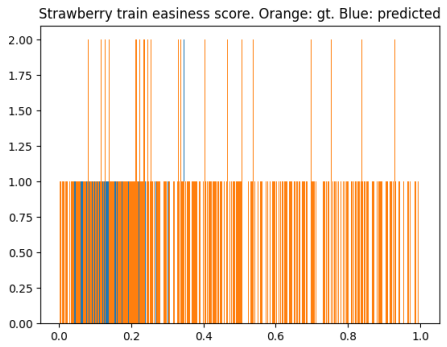


Figure 2: Simpler GAT scheduling model loss and accuracy training plot

T train and V validation sets.
The models for the prediction of an absolute easiness score for each fruit had some complications. Figure 3 (a) shows the distribution of the scores in the dataset. The predictions of the model in Figure 3 (b) are all values close to zero. An explanation behind this could be that the MSE loss favours small values. To discourage this, a custom LeakyReLu activation function was introduced. Another cause is the imbalance in the dataset, which led to reasoning on a new easiness score computation with less weight on the pixel-wise Euclidean distance between the strawberries.
To predict the whole scheduling of picking, the simpler GAT scheduling binary classification model was enriched with information about the unripe berries of the cluster and the patches of the image inside the bounding boxes. This led to a more accurate correspondence between the harvesting order output of the model and the human-decided of the dataset, but it also led to less accuracy in the prediction of the first target to pick. This could be caused by the imbalance introduced by the unripe fruits, which are now nodes in the graph with a label and will be classified by the model. Figure 4 shows a heatmap where every element of the cell $c_{i,j}$ indicates the number of strawberries with human-decided scheduling label i and predicted scheduling label j; if the two schedulings coincided, the heatmap would be a diagonal matrix. Here the percentage of total correspondence is 38.7%.
Figure 5 displays the occlusion properties of the scheduled strawberries. Each cell $c_{i,j}$ indicates the number of berries with scheduling label j



(a)



(b)

Figure 3: This data bar shows the number of strawberries (y-axis) for each rounded easiness score (x-axis). Figure (a): in blue is represented the train, in orange the validation, and in green the test set. Figure (b): in blue are represented the predictions and in orange the train set.

which have occlusion property i, chosen between "non-occluded", "occluded by a leaf", and "occluded by a berry". This shows that the majority of non-occluded strawberries are usually picked first by the model, as expected. A surprisingly high amount of strawberries occluded by a berry are picked early in the scheduling. This suggests that a more aggressive weight should be given as occlusion in the node features.

It is fair to comment that during the testing of the in-development work it was noted that after the picking of a single berry, the configuration of the cluster might change because the manipulator can collide with leaves or other fruits. Although being able to output directly the whole sequence of order of picking is an interesting challenge, it could be fairly important to evaluate just which is the easiest element to be picked.

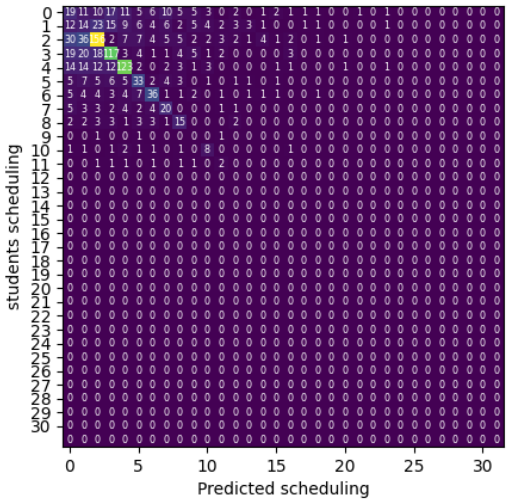


Figure 4: Comparison between students scheduling choice and GAT scheduling model prediction

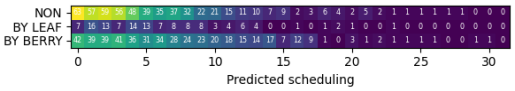


Figure 5: Occlusion properties of the scheduled strawberries

5. Conclusions

This thesis work proposes solutions to some of the main problems of autonomous robotic strawberry harvesting, suggesting another way how automation can help in the agricultural field.
For the reach-to-pick task from visual input, some improvements were proposed starting from the work of d-PMP for the generation of more precise trajectories with joint correlation.
To reduce failures in autonomous picking, the choice of the target berry in a cluster is crucial. The scheduling decision prediction from perception is a novelty in the harvesting of small fruit. Since picking scheduling prediction with GNNs is a new topic, many future improvements can grow from this work. Starting from the perceptron problem, a new dataset would be needed to train a model that recognises both the ripeness and the occlusion of the strawberries. An interesting development in the graph representation would be exploiting more DETR for the node features. In fact, the decoder part of the transformer outputs an attention map for every detected object, that measures how pixels attend to each other. Another progress could surely be to give graphs depth perception by adding it as another node feature: analyzing

the results of the easiness score, it turns out that pixel distance and bounding box size are not enough to capture the separation between berries. As already introduced, a new scheduling score could benefit from this information. Moreover, a custom loss could be studied for the GAT scheduling model that assigns an absolute easiness score to each strawberry, since the MSE method favours small values and is slowing down the learning of correct prediction.

References

[1] Tom Duckett, Simon Pearson, Simon Blackmore, Bruce Grieve, Wen-Hua Chen, Grzegorz Cielniak, Jason Cleaversmith, Jian Dai, Steve Davis, Charles Fox, et al. Agricultural robotics: the future of robotic agriculture. *arXiv preprint arXiv:1806.06762*, 2018.

[2] Alessandra Tafuro, Bappaditya Debnath, Andrea M Zanchettin, and E Amir Ghalamzan. dpmp-deep probabilistic motion planning: A use case in strawberry picking robot. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8675–8681. IEEE, 2022.

[3] Alexandros Paraschos, Christian Daniel, Jan R Peters, and Gerhard Neumann. Probabilistic movement primitives. *Advances in neural information processing systems*, 26, 2013.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[5] Alessandra Tafuro, Adeayo Adewumi, Soran Parsa, Ghalamzan E Amir, and Bappaditya Debnath. Strawberry picking point localization ripeness and weight estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2295–2302. IEEE, 2022.

[6] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[7] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

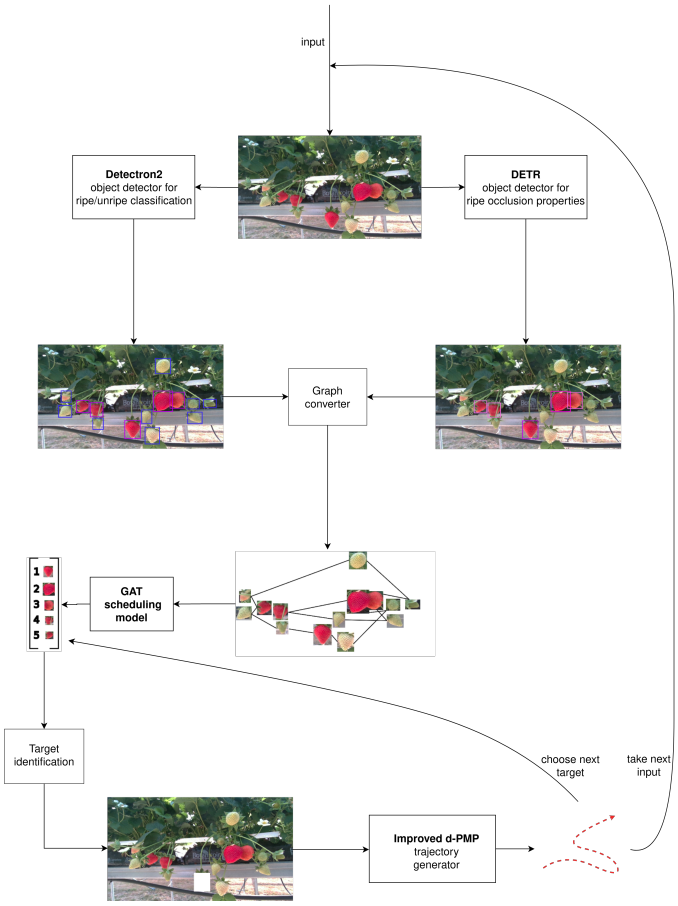


Figure 6: Logical scheme for a real-life application. Starting from an input image of strawberries taken from the robot's home position, this is fed to two object detectors to retrieve information on the ripeness and occlusion properties of the detected fruits. The information is converted into graph representation and fed to the GAT scheduling model. Its output allows for identifying the first berry to be picked. An image with a white patch on the target is given to an improved d-PMP model so that the manipulator can execute the generated trajectory. The human user can choose to collect the sequence of targets indicated by the scheduling model or take a new image to input into the whole sequence above described after the reaching of every strawberry.