Project title:	The Augmented Agronomist: Synthesis of Privacy-Preserving Neural Networks and Robotics to Assist Decision Support
Project number:	
Project leader:	George Onoufriou
Report:	Annual report, October 2019
Previous report:	N/A
Key staff:	Marc Hanheide Georgios Leontidis
Location of project:	Lincoln
Industry Representative:	Richard Harnden, Berry Gardens,
Date project commenced:	30 November 2018

DISCLAIMER

While the Agriculture and Horticulture Development Board seeks to ensure that the information contained within this document is accurate at the time of printing, no warranty is given in respect thereof and, to the maximum extent permitted by law the Agriculture and Horticulture Development Board accepts no liability for loss, damage or injury howsoever caused (including that caused by negligence) or suffered directly or indirectly in relation to information and opinions contained in or omitted from this document.

The results and conclusions in this report are based on an investigation conducted over a one-year period. The conditions under which the experiments were carried out and the results have been reported in detail and with accuracy. However, because of the biological nature of the work it must be borne in mind that different circumstances and conditions could produce different results. Therefore, care must be taken with interpretation of the results, especially if they are used as the basis for commercial product recommendations.

AUTHENTICATION

We declare that this work was done under our supervision according to the procedures described herein and that the report represents a true and accurate record of the results obtained.

George Onoufriou

PhD student

University of Lincoln

Signature Date 2020-02-26

Report authorised by:

Marc Hanheide

Professor of Intelligent Robotics & Interactive Systems

University of Lincoln, Lincoln Centre for Autonomous Systems Research

Signature Date 2020-02-26

Georgios Leontidis

Senior Lecturer in Computer Science (Machine Learning)

University of Lincoln, Machine Learning Group

Signature Date 2020-02-26

CONTENTS

GROWER SUMMARY

Headline

To provide automated agronomy support for agronomists at scale using machine/ deep learning techniques for yield prediction, from high dimensional spatio-temporal data.

This approach will reduce costs whilst maximizing specialist human time in areas that require the most attention.

Background

This work on the augmented agronomist has been undertaken to help focus human time to the most vital areas, and act as an arm for agronomists to help locate problem areas in the crop earlier than possible and improve yield prediction. This system is also being created to improve trust, and security around the usually enigmatised deep learning models, and ensure data owners privacy.

Summary

Over the course of this project we intend to complete the following key objectives:

- **Provide agronomists and growers with yield predictions.** This is the primary advantage provided by the augmented agronomist system which will provide alerts to the operator of deviations from forecasts, and highlight areas where predicted yield potential is not on target. This information will enable the operator to focus efforts in areas which require most attention in order to maximise yield potential.
- **Create an autonomous data collection system**. Hand collecting data at scale would be infeasible due to both time and cost investments being too high while also providing inconsistent results. We will develop a repeatable and autonomous data collection platform so that we can collect spacio-temporal data for yield consistently and at scale.
- **Create a data aggregation and utilization pipeline.** This pipeline will be designed to be able to collect and aggregate distributed (i.e. from multiple growing sites) and autonomous data.
- **Deploy an agronomy assistive neural network to predict plant yield ahead of harvest.** This includes neural network assessment of its own certainty in its predictions, allowing certainty metrics to be used to determine areas of interest.

Financial Benefits

At this point in time it is unclear how significant the financial benefits will be as it is still early in research. However, if we compare this work to similar deep learning studies we would expect around an error of less than 15% (RMSE%) in yield prediction (Konstantinos et al 2018; Maimaitijiang et al 2020). Adding to this the uncertainty metrics to flag the uncertain cases to reduce the model's error over time, and we aim to eventually achieve <5% error.

Action Points

- There are no action points at this early stage of the project.

SCIENCE SECTION

Introduction

Machine/ Deep learning is becoming a bigger and more important part of our daily lives through the rise of an ever-increasing quantity of available data. The use of machine learning in combination with user data is becoming increasingly widespread and impactful in everyday society performing tasks ranging from, natural language processing (Do et al. 2019), image recognition, diagnosis (Biswas et al. 2019), detection, classification (Fawaz et al. 2019), medical diagnosis (Anderson et al. 2019), self-driving cars (Huval et al. 2015), facial recognition (G¨uera and Delp 2018) among many other examples. However one area with which deep learning has remained relatively underutilised is in agriculture, where the data is scarce. The existing research has relied on classical techniques, and remote sensing datasets and to date has not taken advantage of the recent advances such as generative adversarial networks (GANs; Alvarez 2009; Chlingaryan, Sukkarieh, and Whelan 2018; Prasad et al. 2006). The primary reason why agriculture has not innovated in this area for so long is likely to be the lack of consistent data, but also the lack of willingness and trust of the growers/ agriculturalists to release data which could compromise their competitive advantage. Thus if there is little to no data there can be little advancement with deep learning techniques, meaning we will likely have to collect our own data to find any meaningful relations between the features and targets with which to predict yield accurately and far enough ahead to facilitate timely and effective actions. We have access to a plethora of data collection possibilities including; through the RASBERRY project, a collaboration between University of Lincoln (UoL), Saga Robotics, and Berry Gardens (BG), funding autonomous strawberry data collection; and through members of the consortium which fund the Collaborative Training Partnership Fruit Crop Research (CTP-FCR) studentship programme. The involvement of Saga Robotics (SR) allows us to work on a common generic expandable robotic platform called Thorvald. Thorvald is an autonomous robot ready for use in many terrains and an ideal candidate platform to use for our own data collection and usage system thanks to its autonomy, funding, and available resources. The seasonality of strawberry crops mean that fruit is only available for harvest between late June to early October in our experimental system. We intend to use and publish a labelled dataset of strawberries to make a large open dataset for yield prediction. Becoming the only multi-modal plant yield dataset openly available, consisting of both environmental and image features over time.

Materials and methods

Figure 2: Thorvald robot adjacent to strawberry tabletop plantation at the University of Lincoln Riseholme campus. This Thorvald is equipped with 3 realsense cameras; RGB, depth, one Raspberry Pi Zero, one raspberry pi camera v2, one BME680; temperature, humidity, and air-quality sensor

Initially we had to create a system to be able to attain our data before we could do anything with it. This should also serve as the basis of the augmented agronomist. In collaboration with the Lincoln Centre for Autonomous Systems (LCAS), and Saga Robotics who produce the Thorvald an autonomous data collection platform was developed.

The creation of the data collection pipeline is a collaborative process in which this project developed/ and integrated autonomous data capture, secure communication/ encryption, databases and deep learning. In developing the data collection pipeline this study worked closely with a specialist (Raymond Kirk, 3rd year CTP-FCR PhD student) in autonomous robotic control using the robot operating system (ROS), Thorvalds, and deep learning. Thus, the work reported herein is a result of this collaborative effort due to the cross disciplinary skill sets required. The work has been completed in two work packages:

- Data Collection; robotic control, pathing, orchestration, and data capture.
- Data Management; data aggregation/ pipelines, Deployment (dockerisation), MongoDB distributed replica sets and networking.

Data Collection

Owing to the volume of work here, and to keep things concise this section will outline first the objective and then our findings and methods. There will primarily be a focus only in areas where this PhD is principally concerned:

- Data collection planning; Explore the production system (strawberries grown on tabletop), and determine what intricacies the system may have before data collection. An interesting aspect was the effect that insects have on the outcome/ yield such as the burrowing of wasps in the strawberries. Although it is recognised that wasp damage is not a major commercial issue in strawberry production it provides a visual example to develop and demonstrate the capabilities of the augmented agronomist. There initially appeared little way to be able to predict or account for sudden surges in yield loss due to wasp damage, but we found that wasps would only burrow into slightly or completely overripe strawberries (Figure 3). This is a direct consequence of poor picking, as any unpicked strawberries over ripen and attract wasps, which is something we could potentially see in the dataset and could reasonably predict.
- Automatically traverse the strawberry tabletop. This is primarily thanks to the work of Kirk, that the robot traverses the strawberry tabletop safely and consistently.
- Automatically and repeatedly collect data at set intervals down the row. As above.
- Collect data suitable for as many of our associated project needs as possible. For this project the key requirements were environmental factors such as temperature, humidity, etc.
- Create sensors appropriate for data collection. Now that we know what data we require we could decide upon what sensors and other constituents the system required to gather these. In the case of this project a need for a sensor to collect temperature, and humidity resulted in use of a BME680 (INSERT SUPPLIER INFO) which is a small, cheap, i2c gpio board that integrates well into small ARM boards such as the Raspberry Pi. In the absence of accessible GPIO pins on the Thorvald system a compartmentalized system for data collection was developed. The many iterations of this can be seen in figure 4 with each iteration becoming more efficient with regards to space, and power.
- Store the data locally on the robot platform due to difficulties transmitting large amounts of data wirelessly from inside the polytunnels at any given point in time. In previous projects data has been stored as files on the filesystem making them difficult to access, maintain, make consistent between multiple machines, and are unindexable/ unsearchable, resulting in a lot of wasted time. This project used MongoDB, which provides a common database to store all the data captures. This now meant that data could become live when connectivity was possible/ stable.

As a result of quite some monumental effort to create a system suitable for both the requirements of the funding body, and the near-real time data provisioning of this PhD, we created an automated Thorvald data collection platform, gathering many different varieties of spatio-temporal data, while maintaining their consistency, availability, and fault tolerance securely. The data collected key features such as humidity, temperature, light intensity, RGB images of the strawberry plants and fruit from multiple angles, as well as depth, and positional information, to list a few. We have employed 3 different data collection methods, including an intensive stationary sensor array to capture images and sensor data every 30 minutes of a few select plants to model them more precisely. We have also collected continuous capture of images, and sensor data to create sensor-array-video. Lastly the main data collection uses 20 cm step sampling of the strawberry tabletop to model the complex intra-crop environmental changes, which the vast majority of other datasets are not granular enough to do. We can pinpoint exactly at which position the robot was during any given data point, what its orientation was, and all the associated sensor array readings.

Figure 3: Wasp burrowing behavior, eating unpicked overripe strawberries, which is not immediately apparent as it could easily be mistaken for a ripe strawberry.

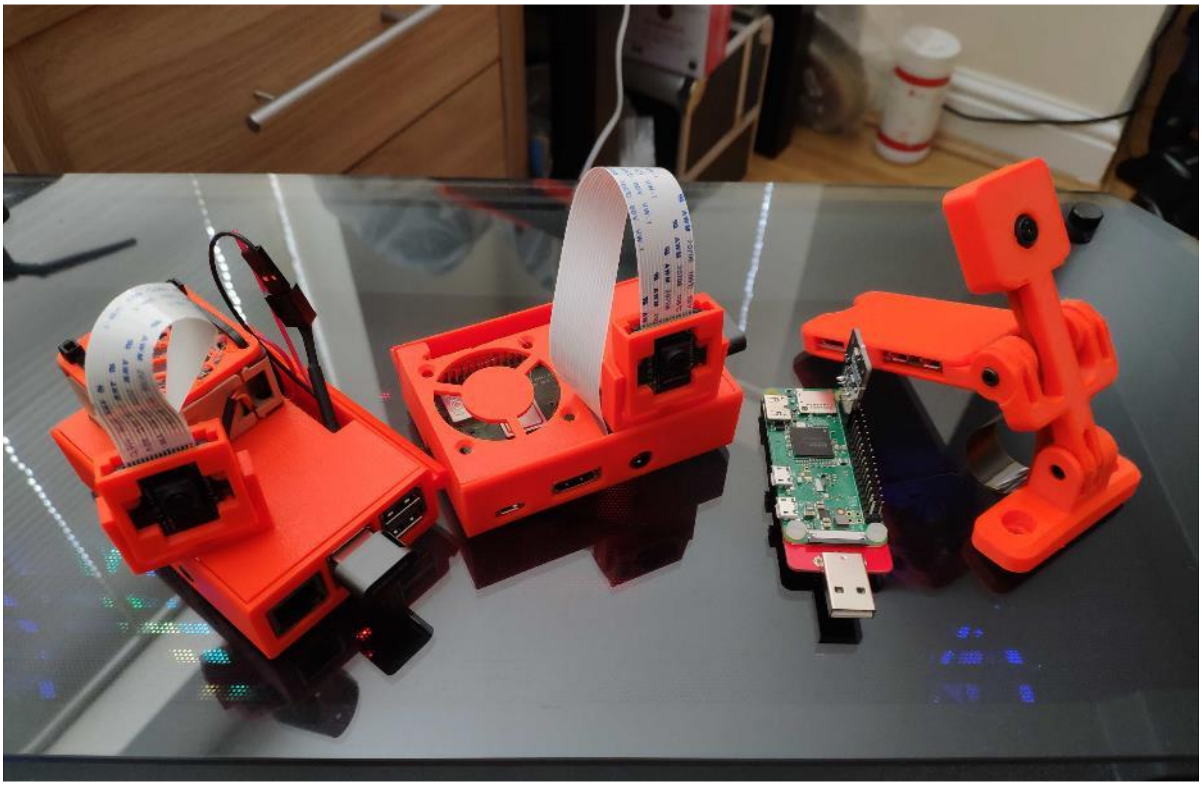

Figure 4: Plethora of cameras, and sensors created over the course of data acquisition stage. The majority of which had specially designed, and 3D printed protective polyethylene terephthalate glycol (PETG) casings which doubled as mounts to be able to attach to the Thorvald robots. PETG is also highly UV and weather resistant.

Data Management

Data management requires the development of a consistent, reliable, and available method to gather data from the pipeline developed above that is potentially distributed between not only different Thorvald robots, but also multiple data collection sites. This work package is summarised as below:

- Aggregate the data. One of the primary reasons for using MongoDB during data collection was an awareness of the distributed, sharding/ replica set functionality ingrained in MongoDB along with its ease of use and security when properly configured. MongoDB was found to be capable of automated data distribution such that each robots database would synchronise its contents with a larger more powerful network of MongoDB shards that was distributed using our Nemesyst framework (Onoufriou 2019a).
- Back up the data/ add redundancy. Simultaneously to aggregating the data from each of the robotic platforms these shards, and replica sets provided guaranteed redundancy should any individual or even multiple datasets fail, with an automatic voting system to adjust which remaining replica set becomes the leader/ master. Using a logged database

such as MongoDB on a simultaneously journaled file-system such as BTRFS or EXT4 meant that almost any data can be recovered that has been removed, deleted, corrupted or otherwise.

- Make the data accessible to others. Since many other projects are interdependent on this data we implemented a key based user authentication mechanism, along with in place encryption, transport layer security, server authentication, replica set cross authentication and encryption.

Figure 5: MongoDB, and monitoring. This allows us to continuously be aware of the state of the replica sets, their load, who is the master/ primary, network traffic, etc. This is an invaluable resource to the management of such a system and will provide us with many useful tools to analyze the effects of our research on this distributed implementation.

Results and Discussion

Thus far through the PhD we have collected, and aggregated in our new pipelines:

- 35033 rows of weather data, or every 15 minutes for the year following 2019-01-01
- 23 records for number of punnets produced over the season
- 2200 intensive records of strawberry growth in the span of a week, including RGB, depth, infrared, location data and environmental data adjacent to the plant.
- 1503 (in progress annotation of each ripe and unripe strawberry) images every 20cm down each row, with the same features as above.
- 188 top down images of strawberries, with the same features as above.

- 188 bottom up images of strawberries with the same features as above.
- 5200 units? of camera footage going down each row.

We have begun cleaning and analysis of the data to prepare it for further deep learning efforts. However at this time we have no further results to disclose as our first year necessarily entailed a large data collection effort to be automated for subsequent years and data collections.

Lastly we have also begun testing of our privacy preserving methods, and found them to be negligibly different in error compared to non-private models and data, although we are not prepared to disclose this in great detail publicly yet since it is the future of the PhD and critical to its success.

Conclusions

In summary we have created a completely new, distributed, and near-real time autonomous data collection platform so that we can evaluate yield prediction at arbitrary scale. We have iterated on our existing database enabled deep learning framework (Nemesyst) with the generalised functionality missing that was required for this application, and published a paper on its application to a similar problem. We have already begun our first tests into integrating pipelines for TensorFlow v2 to accelerate tensors on the GPU while preserving privacy.

Knowledge and Technology Transfer

Our work primarily in data aggregation pipelines are primarily Nemesyst, our open source distributed deep learning framework, for which we have detailed documentation (https://nemesyst.readthedocs.io/), and openly licenced MIT source code (https://github.com/DreamingRaven/nemesyst) which seeks to allow the technology to be more widely adopted.

References

Alvarez, R (2009). “Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach”. In: European Journal of Agronomy 30.2, pp. 70–77.

Anderson, Rachel et al. (2019). “Evaluating deep learning techniques for dynamic contrast-enhanced MRI in the diagnosis of breast cancer”. In: Medical Imaging 2019: Computer-Aided Diagnosis. Vol. 10950. International Society for Optics and Photonics, p. 1095006.

Biswas, M et al. (2019). “State-of-the-art review on deep learning in medical imaging.” In: Frontiers in bioscience (Landmark edition) 24, pp. 392–426.

Chlingaryan, Anna, Salah Sukkarieh, and Brett Whelan (2018). “Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review”. In: Computers and electronics in agriculture 151, pp. 61–69.

Do, Hai Ha et al. (2019). “Deep learning for aspect-based sentiment analysis: a comparative review”. In: Expert Systems with Applications 118, pp. 272–299. Ershov, Mikhail (2015). Survey of Algebra. url: http://people.virginia. edu/~mve2x/3354_Spring2015/ (visited on 11/10/2019).

Fawaz, Hassan Ismail et al. (2019). “Deep learning for time series classification: a review”. In: Data Mining and Knowledge Discovery 33.4, pp. 917–963.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). Deep learning. MIT press. Goodfellow, Ian, Jean Pouget-Abadie, et al. (2014). “Generative adversarial nets”. In: Advances in neural information processing systems, pp. 2672–2680.

G¨uera, David and Edward J Delp (2018). “Deepfake video detection using recurrent neural networks”. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, pp. 1–6. Hayes,

Jamie et al. (2017). “LOGAN: evaluating privacy leakage of generative models using generative adversarial networks”. In: arXiv preprint arXiv:1705.07663. Huval, Brody et al. (2015). “An empirical evaluation of deep learning on highway driving”. In: arXiv preprint arXiv:1504.01716.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). “Deep learning”. In: nature 521.7553, pp. 436–444. 15 Liakos, Konstantinos G et al. (2018). “Machine learning in agriculture: A review”. In: Sensors 18.8, p. 2674.

Niedba la, Gniewko (2019). “Application of Artificial Neural Networks for MultiCriteria Yield Prediction of Winter Rapeseed”. In: Sustainability 11.2, p. 533.

Onoufriou, George (2019a). Nemesyst documentation. url: https://nemesyst. readthedocs.io (visited on 11/10/2019). — (2019b).

Nemesyst; Generalised and highly customisable, hybrid-parallelism, database based, deep learning framework. url: https : / / github . com / DreamingRaven/nemesyst (visited on 11/10/2019).

Onoufriou, George et al. (2019). “Nemesyst: A Hybrid Parallelism Deep LearningBased Framework Applied for Internet of Things Enabled Food Retailing Refrigeration Systems”. In: arXiv preprint arXiv:1906.01600.

Prasad, Anup K et al. (2006). “Crop yield estimation model for Iowa using remote sensing and surface parameters”. In: International Journal of Applied Earth Observation and Geoinformation 8.1, pp. 26–33.

Rahnemoonfar, Maryam and Clay Sheppard (2017). “Real-time yield estimation based on deep learning”. In: Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping II. Vol. 10218. International Society for Optics and Photonics, p. 1021809.

Wang, Anna X et al. (2018). “Deep transfer learning for crop yield prediction with remote sensing data”. In: Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies. ACM, p. 50. You, Jiaxuan et al. (2017). “Deep gaussian process for crop yield prediction based on remote sensing data”. In: Thirty-First AAAI Conference on Artificial Intelligence.

Liakos, Konstantinos G., Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. "Machine learning in agriculture: A review." *Sensors* 18, no. 8 (2018): 2674.

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F. and Fritschi, F.B., 2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sensing of Environment*, *237*, p.111599.

