Project title:	Safe Human Robot Interaction
Project number:	SF/TF 170
Project leader:	Dr. Paul Baxter
Report:	Annual report, 2019
Previous report:	N/A
Key staff:	Alexander Gabriel
Location of project:	University of Lincoln, United Kingdom
Industry Representative:	Richard Harnden, Berry Garden Growers Ltd
Date project commenced:	January 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.

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.

[Name] Nicola Bellotto

[Position] Associate Professor

[Organisation] University of Lincoln

Signature Nicola Bellotto .... Date ....28/04/20..............................

[Name]

[Position]

[Organisation]

Signature .... Date ....

Report authorised by:

[Name]

[Position]

[Organisation]

Signature .... Date ....

[Name]

[Position]

[Organisation]

Signature .... Date ....

CONTENTS

GROWER SUMMARY

Headline

Working towards labour cost savings of up to 20% by reducing the time between picking and processing by letting your human workforce concentrate on picking fruit while your robotic workforce transports the produce.

Background

Fruit production is labour intensive and relies heavily on seasonal migrant labour. Socio-economic changes (e.g. Brexit) which have already led to labour shortages, together with existing pressures associated with decreasing margins from multiples retailers pose challenges for profitability and sustainability in the horticultural sector making a strong case to reduce reliance on manual work. Automation can help, but automation solutions aren't yet commercially available. The agricultural environment poses a number of challenges to both Robotics as well as Human-Robot-Interaction that must be overcome before this technology can be considered mature enough to be applied productively in agricultural settings. This work contributes to this effort by developing solutions that enable comfortable, safe and efficient Human-Robot-Interaction.

Summary

This project is part of the RASBerry research programme. This project aims to develop an autonomous fleet of robots for in-field transportation. Specifically, the robots are expected to aid human fruit pickers by transporting crates from the picker's point of work to locations outside the field or polytunnel. Introduction of robots into this workspace will significantly reduce the costs of producing berries and is the first step towards fully autonomous agricultural systems.

Within the RASBerry research programme, this project is concerned with the safe interaction of humans and robots, specifically the recognition and estimation of human behaviour and its interpretation as commands given to the robot. The results of this project will enable the robot to better prioritize its tasks and allow for a comfortable interaction between human and robot co-workers.

Financial Benefits

A robotic fruit transport system could save 20% labour costs by reducing the time human workers spend with secondary tasks like transportation (From et al. 2018).

Action Points

- Support research and development in robotics and artificial intelligence by engaging with researchers in this area to help shape the applications of this new and innovative technology.

SCIENCE SECTION

Introduction

Introducing robots into a human working space can increase efficiency but should not come at the cost of comfort or safety. To achieve this balance in a challenging setting like agriculture, a robot needs to understand the intentions behind their co-workers' behaviour and basic communication. Gestures form an ideal medium to maintain reliability in adverse circumstances but are limited to situations where the human has their hands free. Additional clues from the environment as well as behaviour analysis can be used to estimate their state. Our interpretation of intentions (Gabriel et al. 2019b) sees them as the meaning of, explanation for, or idea behind an action, plan or utterance (Fleischman et al. 2005, Tahboub et al. 2006 and Youn et al. 2007 respectively). In the agricultural setting, on which this project focuses, workers pick berries into crates in a polytunnel environment. The robot is acting in a supporting role, supplying the human with empty crates, taking away full crates and staying out of the way the rest of the time. To facilitate the robot's autonomy, we created an integrated sensor data processing pipeline and Belief-Desire-Intention (BDI) (Bratman et al. 1987) agent system. The general motivation for this system is that in order to 'understand' the intentions of their human interaction partner (from observable behaviour) and to generate appropriate responses, the robot should consider both the environmental context but also its own goals (or 'desires'): this supports our use of a BDI architecture.

Methodology

Data Processing

The robot perceives its environment through a stereo RGB-D camera, a thermal camera, 2D and 3D LIDAR (Light Detection and Ranging), as well as differential GPS and odometry. It can also receive its co-workers' location either provided by GPS or ultrasonic localization and is supplied with predefined topological and laser scan maps.

During simulation, the robot determines its position using simulated laser scans and odometry. The location of the human is supplied by a picker-simulation engine and abstracted using Qualitative Trajectory Calculus (QTC) [Van de Weghe2006].

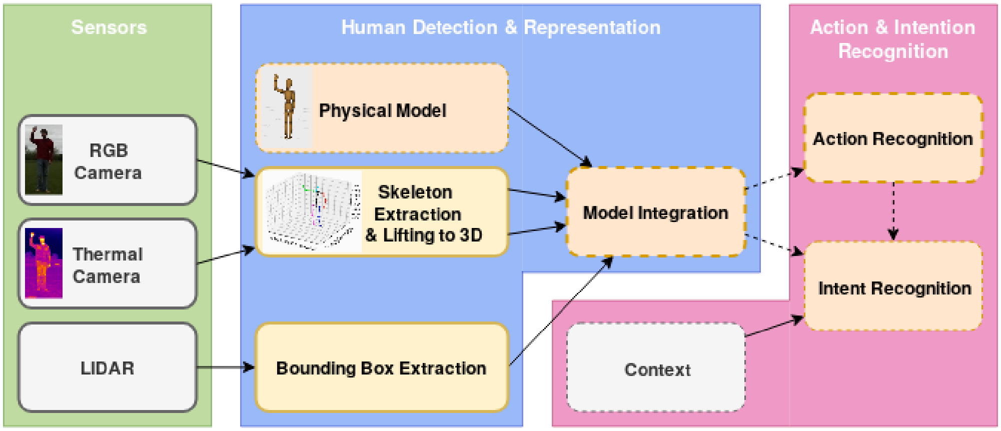


Figure 1. Overview of Data Processing System highlighting different processing stages.

As shown in Figure 1, the video is first pre-processed using OpenPose (Cao et al. 2017) to extract joint positions. Those are further processed to extract joint angles before both are fed into a naïve classifier that produces pose labels for each frame of the video individually, based on predefined prototype poses. Series of frames are classified using voting rounds where each frame contributes a single vote towards a pose. Whichever pose first wins 10 votes, labels the round.

The movement samples used in our first evaluations (used as training or testing/validating?) are part of a new dataset for Action Recognition in agri-robotics, that we collected last summer (2019). The dataset contains examples or 'samples' of behaviours, such as picking of berries and carrying of crates, and 'samples' of gestures to communicate with the robot. The samples were recorded from 10 different subjects in different lighting scenarios (between morning and early afternoon) in a polytunnel environment growing strawberry plants on a tabletop system.

Belief-Desire-Intention Agent

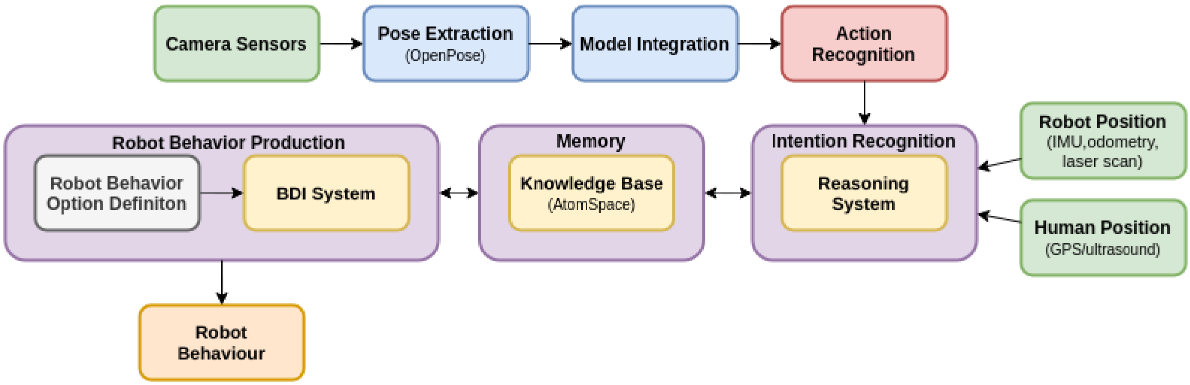

Camera Sensors
Pose Extraction (OpenPose)
Model Integration
Action Recognition
Robot Behavior Production
Robot Behavior Option Definiton
BDI System
Memory
Knowledge Base (AtomSpace)
Intention Recognition
Reasoning System
Robot Position (IMU,odometry, laser scan)
Human Position (GPS/ultrasound)
Robot Behaviour

Figure 2. Overview of the agent system.

The Belief-Desire-Intention (BDI) system (Figure 2) chooses 'intentions' (plans to reach a goal) from its 'desires' (a list of abstract goals) based on its beliefs as captured in the

Knowledge Base (KB). OpenCog's AtomSpace (Goertzel et al. 2014) was used as the Knowledge Base in this system.

The use of a BDI system separates reasoning about which goals to achieve from managing the execution of said goals. This allows us to consider more contextual information when deciding which goal to follow and leads to an aesthetic analogue to our idea of human motivation, intention and action on the robot. In this system, plans are represented as ordered tree structures with executable actions as 'leaves'.

Actions have a set of preconditions and expected consequences, which combine to form the preconditions and expected consequences of a plan. When the agent decides which of its desires are applicable in a given situation, it searches the KB for patterns of beliefs that match its desires' preconditions. If successful, it produces a corresponding intention. When idle, the robot chooses from the possible next actions defined by its current intentions, based on utility and expected time requirements.

Evaluation Strategy

The visual conditions in polytunnels vary a lot depending on the crop stage. Bigger plants obstruct the field of view and obscure the human fruit picker. In order to simulate these conditions for robust evaluation our in-field testing period is limited to the summer months. This period is further limited due to experiments involving robots and humans being time consuming and expensive to setup.

Experiments in simulation on the other hand are relatively cheaper, easier to set up, and can often be automated to run unattended at night. Unfortunately, they can't reproduce the natural environment and human reactions nearly close enough to produce dependable results.

Our evaluation strategy is thus threefold. First, we evaluate the performance of the perception part of the system. Our initial evaluation of the whole system takes place in simulation, keeping as close as possible to the real world, while our later evaluation will take place in the real world.

Results

Evaluation of the Perception System

We tested extracting 2D skeletons (joint positions) from RGB images as well as thermal images, at up to 25m distance from the robot (Figure 3). The skeletons were extracted using a deep-learning-based multi-person skeleton extractor called OpenPose (Cao et al. 2017).

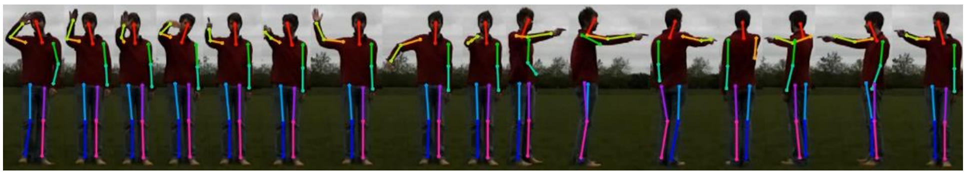


Figure 3. A sample of the gestures we collected for the dataset. From left to right: wave, come, stop, go away, thumb up, thumb down, lower arm up, lower arm down, pointing anti-clockwise at 45° intervals. The skeletons shown are 2D skeletons back-projected from 3D skeletons generated by the 'Lifting from the Deep algorithm' (Tome et al. 2017) run with OpenPose (Cao et al. 2017) 2D skeletons as input.

The average confidence score for skeleton extraction on RGB images shown in Figure 4 are averages over the confidence scores produced by OpenPose (Cao et al. 2017) for each skeleton over the duration of an action at various distances.

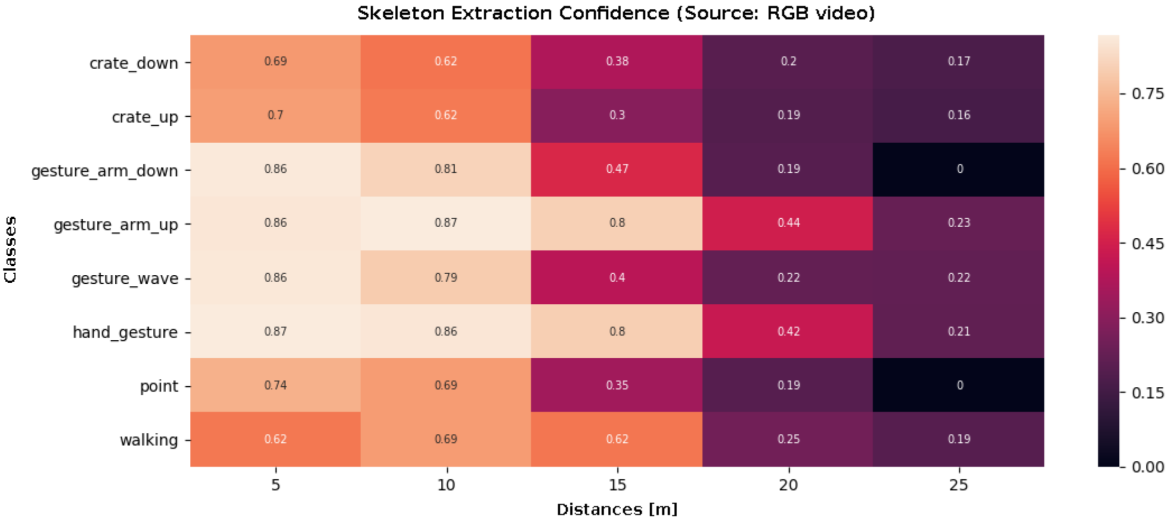


Figure 4. Average Skeleton detection confidence from ZED RGB+D camera sensor (single RGB video) source. Distances on the X-axis from 5m to 25m, confidence values ranging from 0 to 1 with larger (brighter) values indicating better performance. Notable here is the expected loss of extraction confidence with increasing distance. For a more detailed analysis and comparison to skeletons extracted from thermal false-colour video, refer to (Gabriel et al. 2019a).

The data shows significantly better skeleton extraction for action classes where the actor is facing the camera (wave, hand and arm gestures) compared to classes where the actor is facing to the side or away (crate actions, pointing) for a part of the sample set. This results from self-occlusion of the further body side occurring in the side views and self-occlusion of the arms when the actor is performing some action while facing away from the camera.

As expected, the extraction confidence progressively diminishes with distance, as the number of pixels covering the subject grows smaller (Figure 5).

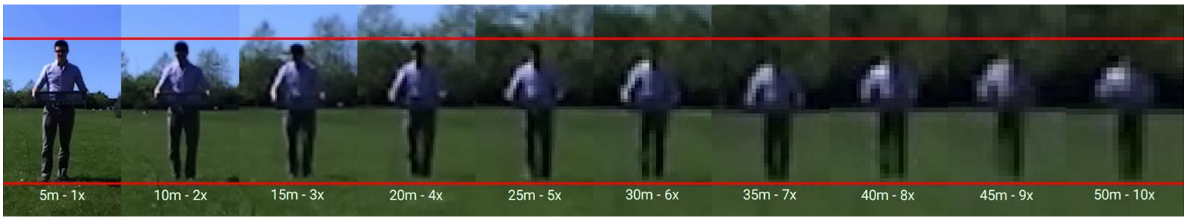


Figure 5. An indication of resolution of a human at increasing distance from the camera, highlighting the human detection problem over longer distances.

We have also tested extracting 2D bounding boxes from RGB images to identify humans. We used a deep-learning-based single-shot object detector called YOLOv3 (Redmon et al. 2018). The detector is run using the pre-trained model for the COCO dataset (Lin et al. 2014). Figure 6 shows some examples of person detection using YOLOv3.

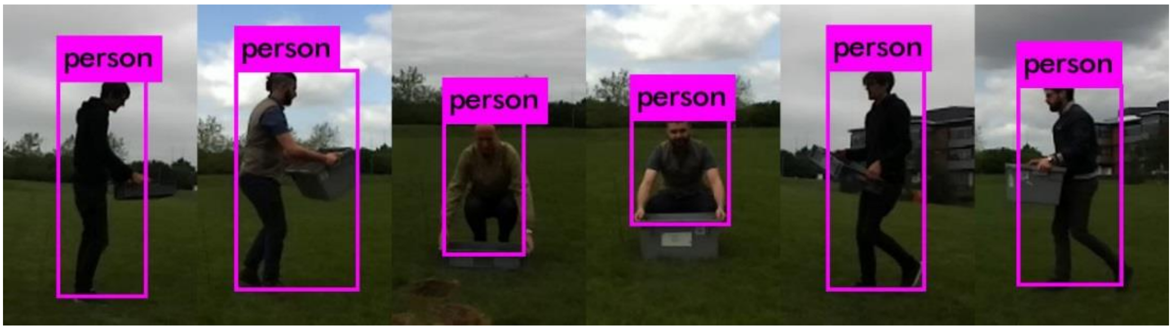
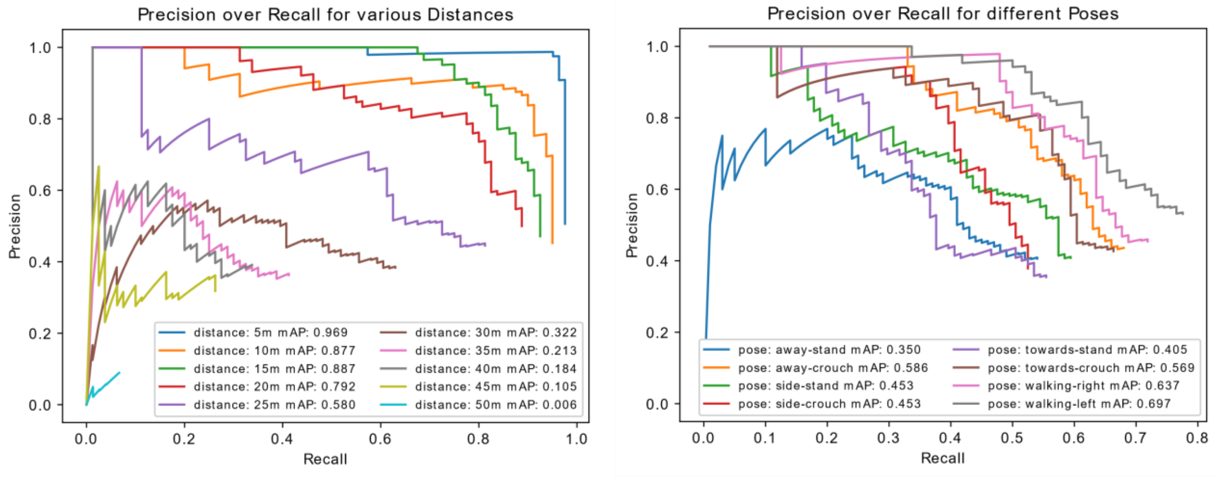


Figure 6. Examples of person detection whilst performing actions with crates, using the YOLOv3 algorithm.

We evaluated the performance of the person detector by running on 800 annotated images from the dataset. Following the PASCAL Visual Object Classes Challenge (Everingham et al. 2015), the precision and recall rates are calculated by assuming a correct detection, if the area of overlap between the predicted bounding box and ground truth bounding box exceeds 50%.

Precision-Recall curves show; Precision, which is the percentage of correctly classified samples out of all samples classified as a class over; Recall, which is the percentage of correctly classified samples out of all samples in that class. Generally speaking, high Precision is positive but with increasing Recall (decreasing classification sensitivity) a loss of Precision is expected. The Precision-Recall curves for various poses and distances (Figure 7), show that person detection works best when people walk laterally. The worst performance is obtained when people face away from the robot. We can also see that after 25m, the algorithm fails to detect people, highlighting an area for prospective improvement in outdoor environments.


Precision over Recall for various Distances

Precision over Recall for different Poses

Figure 7. Precision-Recall curves of person detector for various distances (left) and poses (right).

Initial BDI System Evaluation

In the context of interactions between a robot and a human in an agricultural context, we explore three scenarios (Figure 8). They cover different situations in the work environment: starting to work (crate delivery, left), moving around (evading the human, middle), and resupply (crate exchange, right). Both the delivery and exchange scenarios are initiated by the human gesturing to the robot but require a different response. The robot can make the distinction based on prior observed human behaviour (whether the person has been picking berries). The evasion scenario is triggered by human behaviour (approaching) without any conscious interaction.

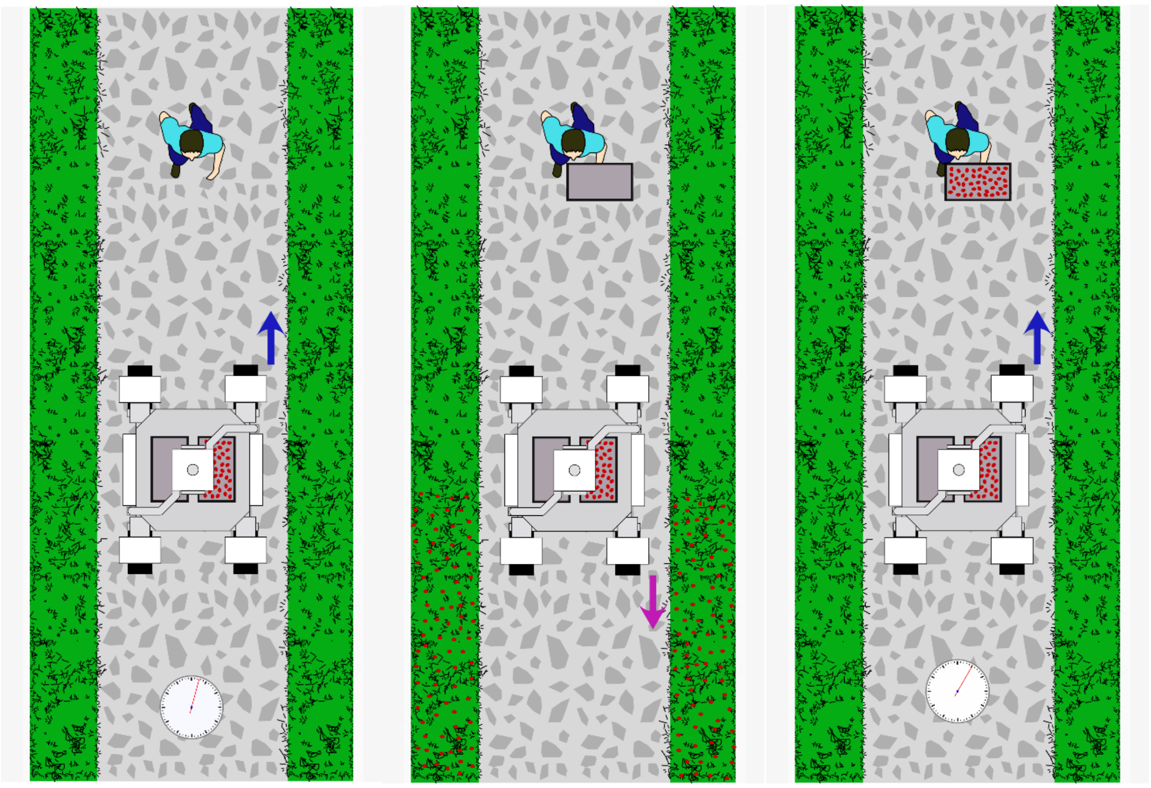


Figure 8. Illustrations of the three interactions between robot and human in an agricultural context explored in this project. Left) Delivery Scenario; Meet the human, wait for 2.5 seconds, leave. Middle) Evasion Scenario; On human approach, move to the next waypoint. Right) Exchange Scenario; Meet the human, wait for 5 seconds, leave.

The recorded human behaviours in polytunnels, are passed as input to the proposed system (Figure 2), acting on a simulated environment. Given 10 recorded subjects, 20 simulations per subject are performed (given a stochastic simulation), resulting in 200 simulations per scenario.

Table 1. Experimental Results Initial System Evaluation

Scenario	Success Rate	Meeting Distance [m]	Time to Service [s]
Delivery	0.99	µ: 0.37 σ: 0.006	µ: 12.29 σ: 0.227
Evasion	1.00	N/A	µ: 9.86 σ: 3.684
Exchange	0.99	µ: 0.37 σ: 0.006	µ: 12.28 σ: 0.574

Table 1 shows for each scenario the three metrics of evaluation: success rate, meeting distance and time to service. The mean (µ) and standard deviation (σ) is shown for the meeting distance and time to service metrics. Success rate is defined as the share of experiment runs that ended with the robot successfully interpreting the situation and performing the expected actions. Meeting distance is the distance between human and robot at which the robot decided to halt to facilitate the delivery or exchange of a crate. The variance for this metric can be interpreted as an indicator of how much reasoning affects the agent's reaction time (the robot and human don't meet in the Evasion case). Time to service is the time between the human displaying behaviour that should trigger a change in robot behaviour, and the time at which the robot performed the expected action (delivered or exchanged a crate, or moved away from the human to the next waypoint). This time consists mainly of the time it takes to detect the behaviour, the time it takes to meet the human (~4s), and the time for the delivery (2.5s) or exchange (5s) of crates.

Discussion

Our sensor data processing evaluation experiments show good results within a range of about 10 meters and dropping performance thereafter. This led us to limit the collection of the second dataset to 10 meters and design our simulated scenarios within that range. Our initial evaluation of the overall system is limited in its real-world significance due to only taking place in simulation, but the high percentage of correctly performed behaviours is promising. The short mean time to service provides an indication of the human work time that could potentially be saved compared to existing practice without robots.

Conclusions

In the last year, we have recorded a second dataset, this time within a polytunnel environment and we have implemented the data processing pipeline as well as a first version of the reasoning stage of the system. Our initial evaluation shows a system that is working for the most part, but still needs some tweaks before it can meet a real-world scenario. It also needs to be tested in the real-world to make sure the promising results of the simulation experiments translate into the real-world.

Glossary

RGB-D: Red, Green, Blue, Depth. Image format storing colour and distance values.

LIDAR: Light Detection and Ranging, like RADAR, but using laser light instead of radio waves.

Odometry: Estimation of change in position over time based on motion data

Topological map: Map consisting of interconnected waypoints.

Laser scan map: Map consisting of 3 dimensional points.

Qualitative Trajectory Calculus (QTC): Calculus that enables the abstraction of time-series of positional data into a symbolic representation (e.g. Entity X is moving away, coming closer)

Knowledge Base: A repository of information facilitating a selection of interaction scenarios with the stored information.

Deep Learning: A class of machine learning algorithms utilizing a layered approach to extract progressively complex features from raw data.

Bounding Box: A rectangle (2D) or cuboid (3D) bounding the borders of an area of interest.

References

Bratman, M. et al (1987). Intention, plans, and practical reason, volume 10. Harvard University Press Cambridge, MA.

Cao, Z. et al (2017). Realtime multi-person 2D pose estimation using part affinity fields. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, 2017-Januar:1302–1310.

Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A. (2015) The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision 111(1), 98–136.

Fleischman, M. and Roy, D. (2005). Why Verbs are Harder to Learn than Nouns: Initial Insights from a Computational Model of Intention Recognition in Situated Word Learning. In Proc. of the Annual Meeting of the Cognitive Science Society.

From, P., Grimstad, L., Hanheide, M., Pearson, S. and Cielniak, G. (2018). RASberry - Robotic and Autonomous Systems for Berry Production. Mechanical Engineering Magazine Select Articles, 140 (6). ISSN 0025-6501.

Gabriel, A., Bellotto, N., Baxter, P. (2019a). Towards a Dataset of Activities for Action Recognition on Open Fields. In: to be published in the Proceedings of the UKRAS 2019 Conference on Embedded Intelligence.

Gabriel, A., Coşar, S., Bellotto, N., and Baxter, P. (2019b). A dataset for action recognition in the wild. In Annual Conference Towards Autonomous Robotic Systems, pages 362–374. Springer.

Goertzel, B., Pennachin, C. and Geisweiller, N. (2014). The opencog framework. In Engineering General Intelligence, Part 2, pages 3–29. Springer.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L. (2014). Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham.

Redmon, J., Farhadi, A. (2018). Yolov3: An incremental improvement. ArXiv.

Tahboub, K.A. (2006). Intelligent human-machine interaction based on Dynamic Bayesian Networks probabilistic intention recognition. Journal of Intelli-gent and Robotic Systems: Theory and Applications, 45(1):31–52.

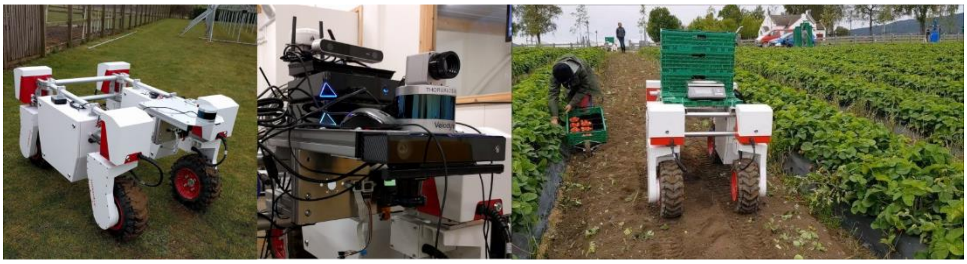
Tome D., Russell C., and Agapito L. (2017). Lifting from the deep: Convolutional 3D pose estimation from a single image, CVPR, pp. 5689–5698.

Van de Weghe, N. et al (2006). A qualitative trajectory calculus as a basis for representing moving objects in geographical information systems. Control and cybernetics, 35(1):97–119.

Youn, S. and Oh, K. (2007). Intention recognition using a graph representation. International Journal of Applied Science, Engineering and Technology, 4(1):13–18.

Appendices

Example photos of the robot and sensor setup for the initial data collection:

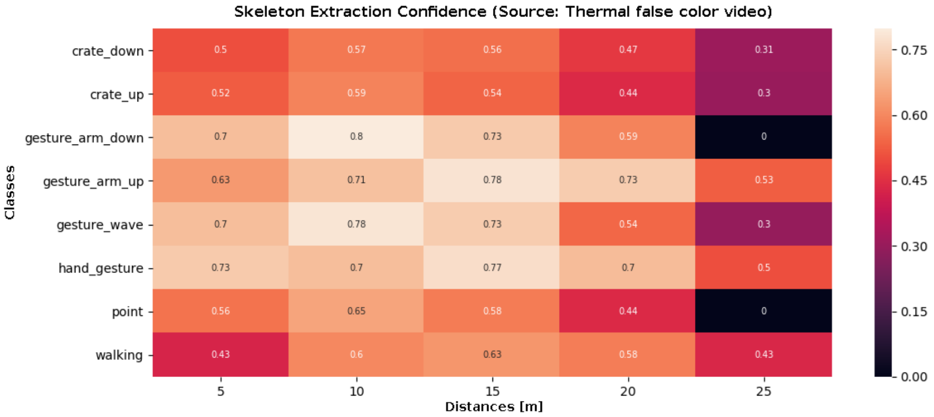


Example photo of robot showing the sensor setup for the polytunnel data collection. The human is directing the robot to move to the side:

Example picture of polytunnel data collection. Subject picking berries approximately 7 meters from the robot:

Examples of varying light and weather conditions during the recording of the first dataset:

Skeleton extraction confidence values from thermal false colour video. Larger/brighter values indicate better performance:


Skeleton Extraction Confidence (Source: Thermal false color video)
Classes
crate_down 0.5 0.57 0.56 0.47 0.31
crate_up 0.52 0.59 0.54 0.44 0.3
gesture_arm_down 0.7 0.8 0.73 0.59 0
gesture_arm_up 0.63 0.71 0.78 0.73 0.53
gesture_wave 0.7 0.78 0.73 0.54 0.3
hand_gesture 0.73 0.7 0.77 0.7 0.5
point 0.56 0.65 0.58 0.44 0
walking 0.43 0.6 0.63 0.58 0.43
Distances [m]