Robot-supervised Learning of Crop Row Segmentation*

Marianne Bakken[1,2], Vignesh Raja Ponnambalam[1], Richard J. D. Moore[2],
Jon Glenn Omholt Gjevestad[1] and Pål Johan From[1]

***Abstract*— We propose an approach for robot-supervised learning that automates label generation for semantic segmentation with Convolutional Neural Networks (CNNs) for crop row detection in a field. Using a training robot equipped with RTK GNSS and RGB camera, we train a neural network that can later be used for pure vision-based navigation. We test our approach on an agri-robot in a strawberry field and successfully train crop row segmentation without any hand-drawn image labels. Our main finding is that the resulting segmentation output of the CNN shows better performance than the noisy labels it was trained on. Finally, we conduct open-loop field trials with our agri-robot and show that row-following based on the segmentation result is likely accurate enough for closed-loop guidance. We conclude that automatically generating noisy segmentation labels is a promising approach for vision-based row following that can be quickly and easily adapted to new scenes.**

I. INTRODUCTION

Automating agricultural practices through the use of robots (i.e. agri-robots, Fig. 1) is a key strategy for improving farm productivity and achieving sustainable food production to meet the needs of future generations. One of the basic requirements for such robots is to be able to navigate autonomously to and from their base station and along the crop rows. Finding robust, fast, and cost-efficient navigation solutions that can generalise across different field types is an active research topic that can facilitate more wide-spread use of agri-robots.

There is a wide range of sensing options for agri-robot navigation, all with different strengths for different types of fields. In open fields, real-time kinematic (RTK) GNSS provides an accurate position for the robot but does not inherently describe the location nor extent of the crops, thus requiring additional setup effort and cost. Onboard sensors such as scanning lidar or machine vision cameras enable direct sensing of the the crops and structures surrounding the robot. Lidar-based navigation has been shown to work well in structured environments such as strawberry polytunnels [1]. Vision-based crop row following using RGB images is a well-established strategy, typically employing colour (e.g. greenness) to segment crops from soil, followed by line extraction to locate crop rows [2]. This has been demonstrated to work well for several crop types, particularly where the crop can be imaged from overhead and/or crop rows are well delineated.

*This work was partly funded by The Norwegian Research Council, grant no. 259869

[1] Norwegian University of Life Sciences, 1430 Ås, Norway
[2] SINTEF Digital, Forskningsveien 1, 0373 Oslo, Norway. marianne.bakken@sintef.no

Fig. 1. The Thorvald II agri-robot platform operating in a strawberry field during data collection. We perform CNN-based segmentation to detect crop rows for visual guidance and propose an automatic labelling strategy for generating training data. During training, a mask representing crop locations is projected onto the camera image using the pose of the robot, measured with a dual-antenna RTK GNSS system. After training, CNN-based image segmentation can be used to guide the robot along the crop rows, without hand-drawn training labels.

Fig. 2. Example appearance variation in strawberry fields on a Norwegian farm. From top left: Thin plants, lanes partly covered with hay, strong shadows, crops with red leaves in autumn, clean lanes without offshoots and lanes covered completely with green offshoots.

With large seasonal variations, as illustrated in Fig. 2 for strawberry crops, greenness index is not always sufficient to get a good separation of plants and lanes. By utilising the recent advances in deep learning, it should be possible to learn a wider variety of features from labelled data. Recent work [3], [4] has shown promising results using Convolutional Neural Networks (CNNs) for semantic segmentation in agricultural scenes. However, hand-labelled training data covering all possible seasonal variations and crop types does not scale very well, and a neural network trained with insufficient data does not necessarily produce features that are more generalisable than traditional methods.

To overcome this limitation we have developed a robot-supervised learning approach that enables us to successfully

train a CNN for semantic segmentation of crop rows without hand-labelled training images, as illustrated in Fig. 1. To achieve this, we utilise knowledge of the sensor setup, structure of the field, and robot pose during the training phase to learn robust features with a CNN that can later be reused on cheaper robot platforms without RTK GNSS, or in sections of the field that do not have GNSS labelled rows. We develop and test our method on data from a strawberry field, but the approach can be applied to any type of field with row-based geometry.

Our hypothesis is that the CNN will be able to learn good features for crop row segmentation despite the reduced accuracy at the borders of the automatically generated labels. We test this hypothesis against hand-labelled real world data and with open-loop field trials. Our test field had relatively limited variation and distinct crop and lane appearances (Fig. 1), which allowed us to isolate and analyse the effect of noisy labels. The ability to generate and label training data on-the-fly will be critical for adapting our system to more complex scenes (Fig. 2).

The main contributions of this paper are: 1) We present an approach for automated generation of training labels for crop row segmentation with a robot platform. 2) Evaluation on real field data show that automatic labelling gives comparable (or better) network performance to manual labelling. 3) Field trials indicate that the segmentation result should be accurate enough for autonomous robot guidance.

II. Related Work

1) Vision-based crop row following: Vision-based crop row following in agriculture has been a research topic for decades, and several works have shown accurate and robust row detection for various crop types. To get a good segmentation separating plants from soil, these methods typically involve some variation of greenness identification (e.g. Excess Green Index (ExG) [5]), combined with thresholding and morphological operations. Then, lines are typically estimated in the segmented image with e.g. Hough Transform as in [2], [6] or least squares fit as in [7], [8], [9] to extract paths that can be used for guidance of autonomous robots.

While methods using greenness index as the main feature can do a great job in many types of fields, there are several situations where this approach may fail. The plants can be covered by dirt after a rainfall, seasons may change the spectral signature of the leaves, or the ground can be covered in vegetation due to weed or offshoots (as in the strawberry field in Fig. 2), to name just a few. There are a few examples of classical methods that use other features, like [10] who propose a learning-based method with Support Vector Machines (SVM) to tackle plants covered in dirt after a rainfall, or [11], who uses stereo cameras to create an elevated crop row map. In any case, tailoring new features for each new field/crop type or appearance does not scale well.

2) Supervised learning: Supervised deep learning approaches, particularly CNNs for semantic segmentation, have been successfully applied for vision-based guidance of autonomous vehicles, and more recently also for off-road and agricultural environments. Maturana et al. [12] build their own off-road dataset with semantic labels and elevation maps, and demonstrate autonomous driving on off-road paths. Valada et al. [13] collect data with RGB, NIR and depth from forest roads, and fuse these modalities in a CNN for segmentation that shows good results in challenging light conditions and appearance variations. Recently, learning-based semantic segmentation has also been applied for row following in agri-cultural environments, like tea plantations [3], and our earlier work in strawberry fields [14]. These works all relied on large quantities of manually-labelled training images to learn the different semantic classes.

3) Self-supervised learning: One way to overcome the need for manually labelled data is using a self-supervised learning strategy, where labels are automatically generated from the input data. There have been many different approaches to label generation for semantic segmentation, for instance using knowledge of the scene and camera viewpoint [15], other sensor modalities [16], [17], [18], [19], or correspondences [20]. Zeng et al. [15] automatically generate a big dataset with segmentation labels for robot grasping, using knowledge of the setup and camera viewpoint. They showed that features learned in a such a simplified setup perform well in cluttered scenes as well. In mobile robotics, it is more common to use other sensor modalities to guide the training. [16] use a hyperspectral scanner to automatically extract training data for weed classification with RGB camera. For autonomous offroad driving applications, 3D sensors (e.g. stereo cameras or scanning lidars) have been used to initially identify and label ground and non-ground regions in matching imagery [17]. Similar approaches have also been applied to the guidance of tractors in agricultural settings for the classification of driving surfaces [18] and localisation of cut plant material for automatic baling [19]. These approaches all require an initial classification of 3D sensor data in order to generate training labels for the visual classifier.

4) End-to-end learning: Another option for avoiding manual labelling is to perform training in an end-to-end mannner, i.e. learn some form of control policy directly from input images. Recently, both reinforcement learning [21] and CNN-based approaches [22], [23], have been used for vision-based guidance of mobile platforms. This eliminates the need for detailed per-pixel image labels for training the underlying networks, and simplifies the labelling process. In our previous work, we have shown that this approach can be applied to the guidance of agri-robots [24]. However, end-to-end learning approaches do not separate the process of learning visual features from classification or policy-learning, and their black-box nature can make it hard to adapt the system to new settings or perform troubleshooting. They also require orders of magnitude more training images than supervised semantic segmentation.

III. METHOD

The pipeline for visual crop row following in this paper consists of three steps: 1) automatic generation of training

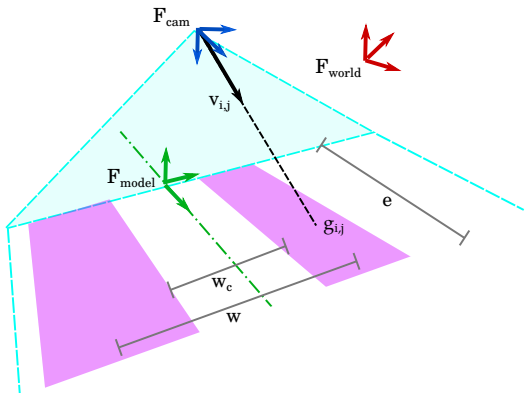

Fig. 3. Illustration of the label projection principle. The (local) virtual field model represent the crop rows as adjacent rectangles, specified by the lane spacing w and the crop width w_c with extent e. The model is projected to the camera image to create semantic labels.

data through label projection, 2) training a CNN for crop row segmentation, and 3) using the result for visual guidance by estimating robot pose from the segmentation. In this section, we focus on label projection, and briefly report the setup of the CNN for segmentation and the procedure for row following for completeness.

A. Automatic label generation

We use camera projection and position of the robot relative to the crop rows to project approximate segmentation labels to camera images, as illustrated in Fig. 3. To perform this projection, we make a few assumptions about the geometry of the field: 1) the crop rows are locally straight and parallel with a fixed width, and 2) the ground is flat and parallel with the robot coordinate frame.

We create a virtual field model using a set of rectangles that are locally aligned with the crop rows. The position and alignment of the virtual model is done in the following way: We measure consecutive points on the crop row centreline with a GNSS receiver, and position reference frame of the virtual model $F_{\text{model,i}}$ on each point. The orientation of the rectangles is then aligned with the crop row using a local linear fit of the centreline points.

Using position and heading information from GNSS on the robot, we compute the lateral offset and the yaw angle deviation compared to the local crop row centreline, which is used to transform F_{cam} to the nearest model frame $F_{\text{model,i}}$. Then the projection from pixel to ground point is computed using the camera intrinsics and extrinsics.

The geometry of the virtual field model can be adjusted as indicated in Fig. 3. Typically, the lane spacing is fixed, as it corresponds to the wheel spacing of the tractor. The crop width will vary and has to be measured separately for different fields and seasons.

B. CNN for crop row segmentation

Based on the automatically generated labels, we train a CNN for semantic segmentation that gives a per-pixel classification of the input image with the labels of interest, i.e., crops, lanes, and background in this case. The training procedure and network architecture are straightforward and well-tested, but are listed here for completeness. We use the SegNet [25] implementation from the Keras Image Segmentation Library [26] with ResNet50 [27] as the base model, input size *360×640*, output size *320×176* and 3 output classes. Training is performed with the following setup: categorical cross-entropy loss ignoring the zero class, adadelta as an optimiser, and regularisation through early stopping (choosing the epoch with the lowest validation loss.) The setup is identical for training with automatic and manual labels, but the epoch for early stopping will vary.

For the experiments in this paper, we trained the models on our field dataset as described in IV-A.3.

C. Crop row following

In order to compute steering commands for crop row following, we must estimate the instantaneous heading angle deviation and lateral offset of the robot from the centreline of the crop row. To compute these parameters during open-loop field trials, we used the following approach: 1) The image region corresponding to the active crop row was isolated from the predicted segmentation mask image (CNN output) using a heuristic algorithm. 2) The set of pixels corresponding to the midline of the extracted crop row blob was computed. 3) The set of midline pixels was projected onto the ground plane using the intrinsic and extrinsic calibration parameters for the camera. 4) A robust linear fit was applied to the projected midline points to compute the relative heading deviation and lateral offset of the robot.

IV. EXPERIMENTAL EVALUATION

Our approach is tested with a robot in a real field to evaluate how our approximations and possible inaccuracies in positioning affect the label quality. We then train a CNN based on automatically generated labels and compare the segmentation results to a CNN trained on manual labels. Finally, we evaluate whether our training and segmentation approach is sufficiently accurate for row following with our agri-robot.

A. Experimental setup

1) Robot and positioning system: We use the *Thorvald* [28] agri-robot platform from Saga Robotics to collect images and robot pose data in the field. The sensor setup with dual GNSS antennas and camera is shown in Fig. 4.

The dual-antenna GNSS receiver AsteRx4 from Septentrio is used to record accurate robot pose whilst driving in the field. In the current setup, the GNSS receiver is equipped with two AntCom G8Ant-3A4TB1-M1 antennas with approximately $0.5\,\text{m}$ separation, which provides high accuracy positions and attitude information (i.e. true heading and roll) at $10\,\text{Hz}$. With RTK GNSS, the position accuracy is estimated to be $1.5\,\text{cm}$ horizontally and $3\,\text{cm}$ vertically, and the heading accuracy $0.3°$ with this setup.

Fig. 4. Our data collection setup showing RealSense camera (for this study we use only RGB images) and dual GNSS antennas mounted on Saga Robotics' Thorvald platform.

Measurements of the static GNSS position of the crop row centreline were obtained manually using a Topcon HIPER SR geodetic GNSS RTK receiver.

Both GNSS receivers utilise corrections from the virtual reference network CPOS from the Norwegian Mapping Authority (NMA) to obtain integer fixed carrier phase RTK GNSS solutions.

The robot has an Intel Realsense D345 camera mounted at centre front, with a tilt of $22.5°$ downwards. The colour images from the camera have a resolution of 640×480 (we do not use the depth data in this paper), and the framerate was set to 6 fps. Images and position data are synchronised through ROS [29] and data is recorded using rosbags.

2) Field data collection: Data collection was performed in a strawberry field with an uneven and hilly terrain with slightly curved rows. The lane spacing is at a fixed $1.25\,\mathrm{m}$, but the width of the crops varies and was measured individually for each row of the recordings.

During data capture, the robot was driven manually along the rows in both directions, in two different patterns 1) straight and centred (approximately) and 2) turning from side to side in a slalom pattern. The driving speed was approximately $0.5\,\mathrm{m\,s^{-1}}$.

3) Dataset: In order to assess the quality of the automatically generated labels and validate the final segmentation result, we required some reference manual image annotations. The manual labelling was performed with the open-source annotation tool Labelme [30] where labels are hand-drawn with piecewise linear boundaries. Background, Crops, and Lanes were assigned labels 0, 1, and 2. The pixels that fall outside the 3 middle crop rows or 2 middle lanes in the image were labelled as background.

For training and testing the CNN, we used images recorded in a slalom pattern to get variation in angular and lateral offset. The training set consists of 195 images from one row, recorded in both directions and sampled at a 20 frames interval to avoid too much overlap between frames. After annotation, 20% of the data was reserved for validation during training of the CNN, to choose hyper-parameters. The test set was recorded in a different row from the same field that was not seen during training, by driving in a similar pattern as for the training set. There are 46 images in the test set. This dataset does not cover all the variation shown in Fig. 2, but since we focus on the performance of the labelling approach and not the overall generalisation of the segmentation, we believe it is sufficient for this purpose.

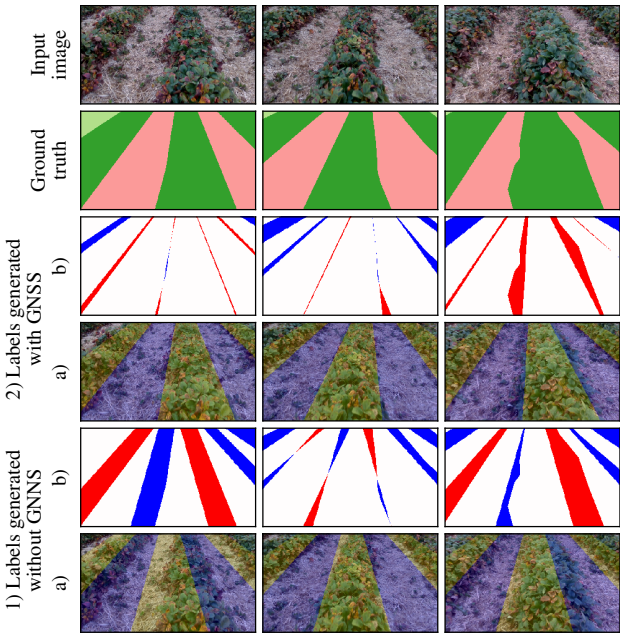

Fig. 5. Examples of automatically generated labels for crop rows in a strawberry field 1) without and 2) with GNSS positions. The label visualisations show a) mask overlayed on image and b) false positives and false negatives (blue) for the lane class.

TABLE I

MASK QUALITY OF AUTOMATIC LABELS, MEASURED IN MEAN IOU COMPARED TO MANUAL LABELS.

	IoU	
Driving pattern	**With GNSS**	**Without GNSS**
Straight and centred	0.78	0.80
Slalom	0.79	0.51

4) Evaluation metrics: For quantitative comparisons of label masks and segmentation results, we use frequency-weighted Intersection over Union (IoU) ignoring the background class.

B. Automatic labels

We compare our automatically generated labels with manual hand-drawn labels for the two different driving patterns described above. In addition to the standard setup with GNSS data, we also generated a set of labels without accounting for robot motion with GNSS data, i.e. assuming perfect alignment with the crop row. We report the mean IoU between manual and automatic masks in Table I, and visualisation of a few examples is shown in Fig. 5. When using GNSS data to project the labels, the same performance

TABLE II

SEGMENTATION RESULTS ON TEST SET, MEASURED IN MEAN IOU COMPARED TO MANUAL LABELS.

Labelling strategy	IoU
Manual	0.93
Automatic with GNSS	0.88
Automatic without GNSS	0.53

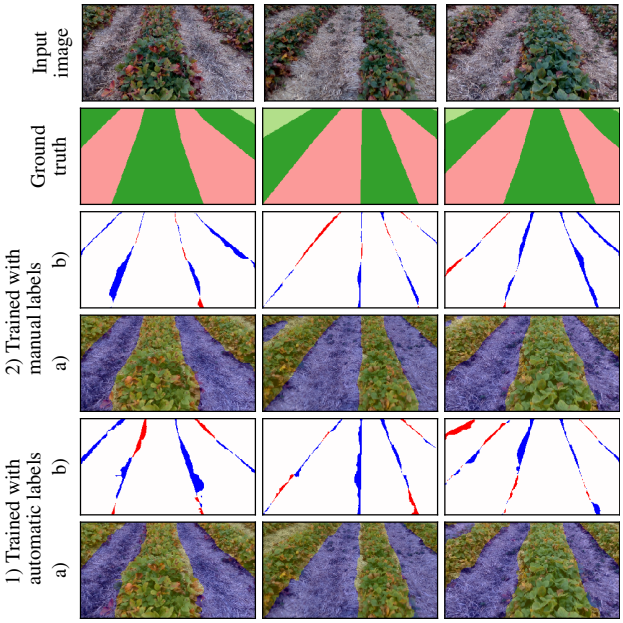

Fig. 6. Example segmentation results for models trained on 1) automatically generated masks with GNSS positions and 2) manual (hand-drawn) labels and. Visualisations show a) segmentation overlayed on image and b) false positives (red) and false negatives (blue) for the lane class.

is achieved for the slalom driving pattern as the straight and centred one.

C. Segmentation

The model for crop row segmentation was trained as described in Section III-B on labels automatically generated with GNSS data, from the same row as shown in Section IV-B. For comparison, we also trained a model on the same images, with manual labels. The model used for testing was picked based on minimum validation error, which was after 9 epochs for the automatic labels and 4 epochs for the manual labels. The models were tested on the separate test set, using manually labelled data as ground truth. The mean IoU of the segmentation masks are reported in Table II, and some example segmentation masks and their pixel-wise errors displayed in Fig. 6.

From the numbers in Table II and the examples in Fig. 6, we see that the model trained on automatic labels performs quite well. The mean IoU of 0.88, is actually slightly higher than the IoU of the masks it was trained on. When larger patches are misplaced, as is the case for labels generated without GNSS in sharp turns, the CNN is not able to learn any general features, as expected.

TABLE III

OPEN-LOOP ROBOT TRIALS. MEAN ABSOLUTE ERROR (MAE) OF ESTIMATED YAW AND POSITION COMPARED TO GNSS GROUND TRUTH.

	Yaw angle MAE	Lateral offset MAE
Manual labels	0.6°	4.8 cm
Automatic labels	0.1°	0.6 cm
Predicted masks	1.6°	4.6 cm

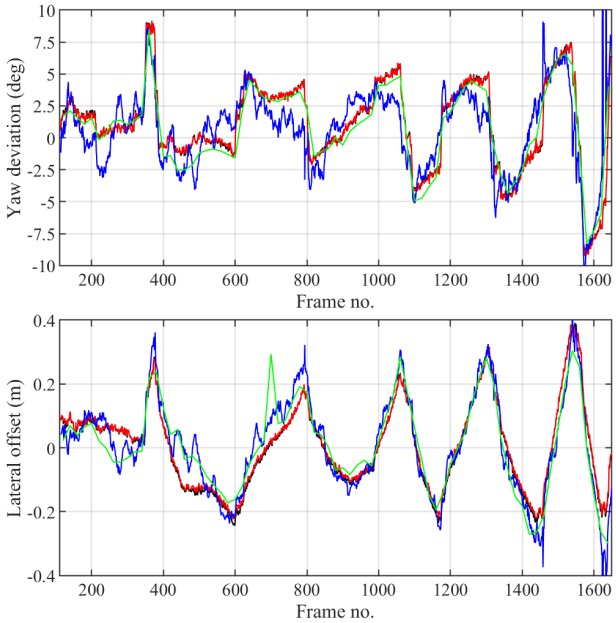

Fig. 7. Yaw angle deviation (top) and lateral offset (bottom) between robot and crop row centreline for a slalom drive. Traces show the yaw/lateral offset estimated from predicted masks (blue) as well as directly from manual labels (green) and automatically generated labels (red), and also the GNSS-based ground truth (black). Best viewed in colour.

D. Open-loop robot trials

We performed a series of open-loop field tests to validate that our CNN trained with automatic labels produced predicted crop masks with accuracy sufficient for closed-loop robot crop following. We compared the yaw angle deviation and lateral offset estimated from the predicted masks (see Section III-C) against the same values estimated from GNSS ground truth.

Fig. 7 shows that both the yaw angle deviation and lateral offset estimated from the segmentation results closely followed the ground truth for the entire dataset. The mean errors between estimated and ground truth values are summarised in Table III. For comparison, yaw deviation and lateral offset were also computed from the training labels, for both manual and automatic labels. Estimating yaw and lateral offset from the automatic labels was predictably closest to GNSS-derived ground truth because the automatic labels were generated from GNSS data. This result at least confirms that our approach for extracting yaw angle and lateral offset from the mask images is valid.

V. DISCUSSION

1) Automatic labelling: Our results indicate that when accounting for robot motions with GNSS, the overall alignment of the automatically generated masks is equally good for any driving pattern. However, the assumptions made when generating the mask introduce some errors, and the three most common are summarised in Fig. 8. The first column shows a lateral bias in predicted masks, possibly due to uncorrected roll angle of robot w.r.t. ground plane due to uneven track depths. The second shows a section where the height and width of the crop row is larger than the value estimated at the beginning of the row. In the third, there is a dent in the crop row boundary, that is not captured by straight boundaries of the projected mask. The first issue could be reduced by computing a full 6-DOF pose of the robot, while the other two are expected due to the limitation of the rectangular fixed-size field mask.

2) Segmentation: The reported IoU values for the final segmentation actually showed better performance than the automatically generated labels it was trained on, indicating that the neural network was able to learn the general appearance of the classes despite noisy labels along the boundaries. This is probably because of the large amount of good pixel labels per image, which dominate the total loss during training. Closer inspection of the segmentation masks reveal some issues, as shown in Fig. 9. The first two cases show under- and over-estimation of the crop row width, which may arise from the label errors discussed above. As long as this is consistent within each image, it does not introduce any shift for the row following. For the third case, the errors are mostly due to the fact that the segmentation is more fine-grained than the simplified ground truth, which does not capture the detailed curves around the plants.

Finally, it should be noted that the segmentation results reported in this article pertain to a limited test set that does not encompass a large variety of crops or seasonal changes in appearance. However, we believe that this is sufficient for exploring the effect of training with inaccurate segmentation boundaries. The overall performance of the final segmentation and row following system needs to be further evaluated on more data and different crop types, which will be addressed in future work.

Overall, these results indicate that our proposed auto-labelling approach produces guidance information that should be of sufficient accuracy for future closed-loop crop following trials with our agri-robot platform.

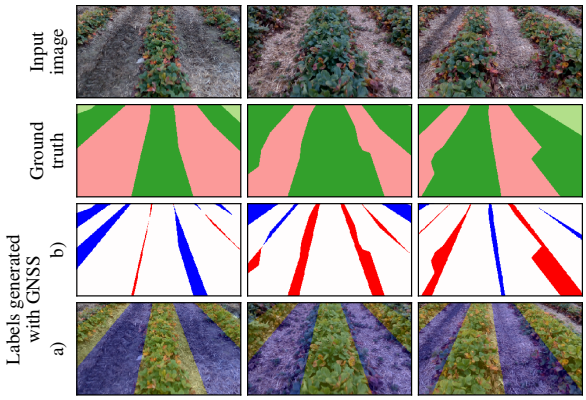

Fig. 8. Three example failure cases for automatically generated masks with GNSS positions. From left: 1) lateral bias in predicted masks; and 2) variation in crop row height and width unaccounted for by the rectangular crop labels; and 3) detail of rough crop boundaries not captured by straight label boundaries. Best viewed in colour.

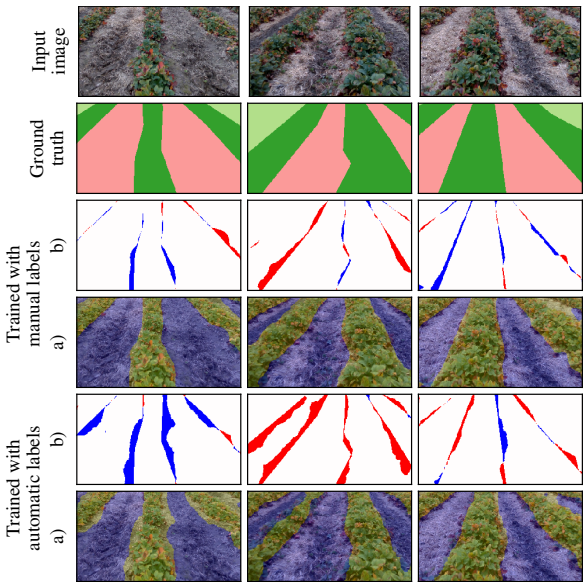

Fig. 9. Three example failure cases for segmentation with model trained on automatically generated masks. From left: 1) and 2) under- and over-estimation of crop row width; and 3) the simplified ground truth does not follow detailed curves around the plants. For reference, segmentation result for model trained with manual labels is also shown. The different visualisations are described in Fig. 6. Best viewed in colour.

VI. CONCLUSIONS

In this paper, we have proposed a new approach for automated labelling for crop row segmentation, using GNSS data from an agri-robot in the field. As expected, the simplified mask introduces some labelling errors near the class boundaries, however the resulting segmentation output of the CNN showed slightly better performance than the noisy labels it was trained on. This indicates that the neural network was able to learn general features despite inaccuracies in the labelled crop regions. Our open-loop field trials indicate that the segmentation accuracy is sufficient for row-based guidance of an agri-robot.

We conclude that training with labels that are generated automatically but noisily is a promising approach for quickly and easily adapting a vision-based row following robot to seasonal variations and new crops or fields on-the-go. In future work we will test our approach on a broader dataset to investigate its capabilities for generalisation.

ACKNOWLEDGEMENT

The authors would like to thank Per Fredrik Saxebøl for access to his strawberry farm, as well as Lars Grimstad at NMBU for help with the robot setup, and Sigvald Marholm for proofreading.

REFERENCES

[1] T. D. Le, V. R. Ponnambalam, J. G. O. Gjevestad, and P. J. From, "A low-cost and efficient autonomous row-following robot for food production in polytunnels," *Journal of Field Robotics*, no. April, pp. 1–13, 2019. [Online]. Available: http://doi.wiley.com/10.1002/rob.21878

[2] J. A. Marchant and R. Brivot, "Real-Time Tracking of Plant Rows Using a Hough Transform," *Real-Time Imaging*, vol. 1, no. 5, pp. 363–371, 11 1995.

[3] Y.-K. Lin and S.-F. Chen, "Development of navigation system for tea field machine using semantic segmentation," *IFAC-PapersOnLine*, vol. 52, no. 30, pp. 108–113, 2019.

[4] P. Lottes, M. Hörferlin, S. Sander, and C. Stachniss, "Effective vision-based classification for separating sugar beets and weeds for precision farming," *Journal of Field Robotics*, vol. 34, no. 6, pp. 1160–1178, 2017.

[5] D. M. Woebbecke, G. E. Meyer, K. V. Bargen, and D. A. Mortensen, "Color indices for weed identification under various soil, residue, and lighting conditions," *Transactions of the American Society of Agricultural Engineers*, vol. 38, no. 1, pp. 259–269, 1 1995. [Online]. Available: https://pennstate.pure.elsevier.com/en/publications/color-indices-for-weed-identification-under-various-soil-residue-

[6] B. Åstrand and A. J. Baerveldt, "A vision based row-following system for agricultural field machinery," *Mechatronics*, vol. 15, no. 2, pp. 251–269, 3 2005.

[7] X. Zhang, X. Li, B. Zhang, J. Zhou, G. Tian, Y. Xiong, and B. Gu, "Automated robust crop-row detection in maize fields based on position clustering algorithm and shortest path method," *Computers and Electronics in Agriculture*, vol. 154, pp. 165–175, 11 2018. [Online]. Available: www.elsevier.com/locate/compag

[8] I. García-Santillán, J. Miguel Guerrero, M. Montalvo, G. Pajares, and G. Pajares pajares, "Curved and straight crop row detection by accumulation of green pixels from images in maize fields," *Precision Agriculture*, vol. 19, pp. 18–41, 2018. [Online]. Available: https://doi.org/10.1007/s11119-016-9494-1

[9] A. Ahmadi, L. Nardi, N. Chebrolu, and C. Stachniss, "Visual Servoing-based Navigation for Monitoring Row-Crop Fields," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2020.

[10] J. M. Guerrero, G. Pajares, M. Montalvo, J. Romeo, and M. Guijarro, "Support Vector Machines for crop/weeds identification in maize fields," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11 149–11 155, 9 2012.

[11] M. Kise, Q. Zhang, and F. Rovira Más, "A stereovision-based crop row detection method for tractor-automated guidance," *Biosystems Engineering*, vol. 90, no. 4, pp. 357–367, 4 2005.

[12] D. Maturana, P.-W. Chou, M. Uenoyama, and S. Scherer, "Real-time semantic mapping for autonomous off-road navigation," in *Field and Service Robotics*, M. Hutter and R. Siegwart, Eds. Cham: Springer International Publishing, 2018, pp. 335–350.

[13] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, "Deep multispectral semantic scene understanding of forested environments using multimodal fusion," in *2016 International Symposium on Experimental Robotics*, D. Kulić, Y. Nakamura, O. Khatib, and G. Venture, Eds. Cham: Springer International Publishing, 2017, pp. 465–477.

[14] V. R. Ponnambalam, M. Bakken, R. J. Moore, J. Glenn Omholt Gjevestad, and P. Johan From, "Autonomous crop row guidance using adaptive multi-roi in strawberry fields," *Sensors*, vol. 20, no. 18, p. 5249, 2020.

[15] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1386–1383.

[16] A. Wendel and J. Underwood, "Self-supervised weed detection in vegetable crops using ground based hyperspectral imaging," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2016-June, pp. 5128–5135, 2016.

[17] S. Zhou, J. Xi, M. W. McDaniel, T. Nishihata, P. Salesses, and K. Iagnemma, "Self-supervised learning to visually detect terrain surfaces for autonomous robots operating in forested terrain," *Journal of Field Robotics*, vol. 29, no. 2, pp. 277–297, 2012.

[18] G. Reina and A. Milella, "Towards autonomous agriculture: Automatic ground detection using trinocular stereovision," *Sensors*, vol. 12, no. 9, pp. 12 405–12 423, 2012.

[19] M. R. Blas and M. Blanke, "Stereo vision with texture learning for fault-tolerant automatic baling," *Computers and electronics in agriculture*, vol. 75, no. 1, pp. 159–168, 2011.

[20] M. Larsson, E. Stenborg, C. Toft, L. Hammarstrand, T. Sattler, and F. Kahl, "Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 31–41.

[21] G. Ryou, Y. Sim, S. H. Yeon, and S. Seok, "Applying asynchronous deep classification networks and gaming reinforcement learning-based motion planners to mobile robots," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6268–6275.

[22] A. Loquercio, A. I. Maqueda, C. R. Del-Blanco, and D. Scaramuzza, "DroNet: Learning to Fly by Driving," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 1088–1095, 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8264734/

[23] A. Giusti, J. Guzzi, D. C. Cireşan, H. Fang-Lin, J. Rodriguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. Di Caro, D. Scaramuzza, and L. M. Gambardella, "A Machine Learning Approach to Visual Perception of Forest Trails for Mobile Robots," *Robotics and Automation Letters*, vol. 1, no. 2, pp. 661–667, 2016.

[24] M. Bakken, R. J. Moore, and P. From, "End-to-end learning for autonomous crop row-following," *IFAC-PapersOnLine*, vol. 52, no. 30, pp. 102–107, 2019.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[26] D. Gupta, "Implementation of various Deep Image Segmentation models in keras." https://github.com/divamgupta/image-segmentation-keras, 2017.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] L. Grimstad and P. J. From, "The Thorvald II Agricultural Robotic System," *Robotics*, vol. 6, no. 4, 2017.

[29] Stanford Artificial Intelligence Laboratory et al., "Robotic operating system." [Online]. Available: https://www.ros.org

[30] K. Wada, "labelme: Image Polygonal Annotation with Python," https://github.com/wkentaro/labelme, 2016.