Straw Row Position And Orientation Reconstruction Through Image Segmentation

Aleksandr Yurievich Shkanaev¹, Dmitry Valerevich Polevoy², Aleksei Vladimirovich Panchenko³
Yuri Yurievich Borisov⁴ and Maxim Igorevich Kolomeychenko⁵

¹²Cognitive Technologies LTD, 60-letya Oktiabrya Avenue, 9, Moscow, 117312, Russia
³¹ISA FRC CSC RAS, 60-letya Oktiabrya Avenue, 9, Moscow, 117312, Russia
⁵⁴DeepSystems LTD, 3-ya Yamskogo Polya ul., 2 building 13, Moscow, 125040, Russia

Abstract
This paper investigates a new approach in straw row position and orientation reconstruction in an open field, based on image segmentation with Fully Convolutional Networks (FCN). The model architecture consists of an encoder (for feature extraction) and decoder (produces segmentation map from encoded features) modules and similar to [1] except for two fully connected layers. The heatmaps produced by the FCN are used to determine orientations and spatial arrangements of the straw rows relatively to harvester via transforming the bird's eye view and Fast Hough Transform (FHT). This leads to real-time harvester trajectory optimization over treated area of the field by correction conditions calculation through the row’s directions family.

Keywords: grain harvester, semantic segmentation, convolutional neural networks,

INTRODUCTION
In recent years information technologies development lead to a significant increase in R&D projects in the area of self-driving vehicles and robotic systems; advanced driver assistance systems (ADAS), self-driving cars, manufacturing robots [2], etc. No wonder, that major part of these projects is dedicated to the agricultural area: with a predicted world population increase up to 10 billion by the 2050 year, it is obvious, that population problem can’t be solved without corresponding increases of agricultural efficiency and automatization. During the last 20 years, these studies made a significant progress: in 90’s of XX-century, Japan, USA and Europe first introduced the usage of GPS-integrated agricultural vehicles within the "Precision Agriculture" concept. Now the complexity and accuracy of such solutions has significantly increased because of various types of new sensors, such as optical encoders, LIDAR/LEDDAR, as well as mathematical algorithms and steering controller systems, etc. All these technologies are already implemented in manufacturing solutions in various fields of agriculture: spraying [3], soil preparation and seedling [4], rootstock planting [5], harvesting [6] etc.

RELATED WORKS
Nowadays in the field of Precision Agriculture and Livestock Farming many types of sensors are used in the autonomous and semi autonomous driving [7-9] to solve cut-edge detection [10-12] and row detection [13-15] tasks. Machine vision-based systems mimic the operator’s perceptive process [13]. Images can be acquired without depth information with monochrome (near infra-red) or color cameras and stereo vision technology [16-18] and are widely used to restore depth information. RGB-D and flash light detection (ToF) cameras give direct 3D vision information and points clouds that are captured from laser (LADAR) or light detection and ranging (LIDAR) devices [8, 9, 19]. In vision-based systems camera location and algorithm development are related. For example, the algorithms discussed in [11] were used with a camera mounted directly on the head of the combine harvester. In spite of a more complicated scene geometry vehicle cabin mounted camera systems [15, 16, 20] are more universal with respect to different types of tasks and have become a de-facto standard. Monocular vision-based crop rows detectors generally separate plants images from soil, and then find rows by fitting straight lines to the resulting binary image. Authors [14, 21] use near-infrared spectrum monochrome cameras and image binarization (intensity segmentation). More sophisticated color-based crop row detectors use color segmentation focused on the separation of green plants (crops and weeds) from the rest. Many “vegetation indices” ([22-24] etc.), complementary techniques like fuzzy clustering [25, 26] and binarization method selection are used to become more robust to illumination changes and shadows. Color segmentation approach is ineffective when segmented objects are no longer green and using color-based only method becomes difficult. Crop rows have regular structure and several approaches avoid the segmentation step by instead searching for periodic variations in image intensity ([21, 27-29], etc.) under the assumption that the crop rows lay at the peaks in these periodic variations. Crop rows detection is predominantly carried out by fitting lines to the row fragments (segments) with standard image analysis techniques such as: fixed template matching [30]; linear regression [14, 28]; the Hough transform [22, 31, 32]; random sample consensus [16]. Due to perspective distortion parallel
rows became non-parallel on images and vanishing point detection technics can be used. Another popular approach is bird’s eye view [12, 31]. Due to the complexity of the image segmentation task under controlled conditions recent works [15, 33, 59] try to combine intensity, color and texture information to obtain more robust algorithms.

Over the past several years Convolutional Neural Networks [34] have shown remarkable results in computer vision and related tasks, such as image classification [35-37], object detection [38-40], transfer learning [41, 42]. Recent works have also shown that CNNs can equally well apply to pixel-level labelling tasks, such as semantic segmentation [1, 43]. Most earlier works on semantic segmentation rely on a pre-segmentation into superpixels or other segment candidates, and extract features and categories from individual segments and from various combinations of neighboring segments [44-50]. Then an inference procedure based on a graphical model finds the most consistent set of segments which covers the image.

There is a number of problems with the approach described above. First of all, final semantic segmentation results heavily depend on heuristic segment proposal algorithm. Secondly, it is not clear how to integrate large available context from the whole image into decision making procedure about the region being classified. And, at last, pre-segmentation approaches are very computationally expensive and can not be applied in real time settings.

Another line of works uses convolutional networks in a patch based fashion [51, 52] where the model classifies the central pixel of an image crop which is relatively small compared to the whole image. The segmentation map is obtained by cropping an image and making a class prediction at every possible location. Although the properties of convolutional network architecture allow to calculate the segmentation map efficiently [52], the problem of integrating global features from the whole image (the problem of small receptive field) into decision making still remains.

The most successful recent results for semantic image segmentation are based on Fully Convolutional Networks (FCN) and their modifications [1, 43, 53, 54]. FCN-based models take the image as an input and produce another image as an output which usually is interpreted as segmentation map. In this formulation, one can construct FCN-based models with arbitrary large receptive fields making use of global contextual information from an entire image. Natural end to end training abilities, fast and straightforward inference procedure are another key properties of FCN models. The direct prediction with FCN-based models usually results in low-resolution segmentation map. To address this problem, a number of recent methods focus on refining the low-resolution prediction to obtain high resolution segmentation map. The following works [53, 54] use graphical models on top of FCN, [1] introduces a promising approach of using learnable deconvolutional layers to upsample the low-resolution predictions. In this work we chose to adopt the deconvolutional network architecture from [1] for the task of straw row position and orientation reconstruction through image segmentation.

**NETWORK ARCHITECTURE**

The model architecture consists of encoder and decoder modules as shown in Fig. 1. Generally speaking, the encoder module is responsible for feature extraction, whereas decoder module generates segmentation map given the encoder features. The encoder and decoder modules have symmetrical configuration in a sense that for every convolutional layer in encoder there is a deconvolutional layer in decoder and each pooling layer in encoders corresponds to a specific unpooling layer in decoder. Our architecture is similar to the one described in [1], except that we have removed two fully connected layers.

**DATASET**

The dataset consists of images that were obtained from a camera mounted on a performing agricultural work harvester as schematically depicted in Fig. 2. The images were manually annotated to form the Ground Truth segmentation maps. The segmentation maps include following classes of objects:

- Background: none of the following
- Straw Row
- Harvested Crop
- Uncut Crop
- Combine Head

All training segmentation maps were made manually. All in all there are 925 marked up images in Full HD image format. Segments are represented by a set of simple polygons.
The minibatch size was set to 64. The network constant. L2 regularization with weight decay value of 0.0005. During the training process the learning rate value was momentum had the values of 0.01 and 0.9 respectively.

The proposed network implementation was based on Caffe initialized with zero-mean Gaussians. The straw row class is the only one to be allocated from the deconv+ respectively. The quantitative results of the models are presented in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>Background</th>
<th>Uncut Crop</th>
<th>Straw Row</th>
<th>Combine Head</th>
<th>Harvested Crop</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>IoU</td>
<td>0.997</td>
<td>0.966</td>
<td>0.914</td>
<td>0.747</td>
<td>0.672</td>
<td>0.966</td>
</tr>
<tr>
<td>Acc</td>
<td>0.987</td>
<td>0.956</td>
<td>0.914</td>
<td>0.747</td>
<td>0.672</td>
<td>0.966</td>
</tr>
</tbody>
</table>

Table 4: Per class evaluation metrics on the test set

From Table 4 we can see that training data augmentation was helpful in a sense that the corresponding model works better in terms of intersection over union (IoU) and per pixel accuracy (Acc). The model that makes use of data augmentation correctly classifies 95.8 percent of all pixels on the test set. Besides, for the well represented classes such as background, the accuracy is extremely high. On the other hand, the accuracy for underrepresented class of straw row is below 85 percent.

STRAW ROW GEOMETRY RECONSTRUCTION

The goal of active rear-wheel steering is to keep straw row between forward wheels of the harvester during moving over treated area. Trajectory above the middle line of straw row additionally preserve harvester from dangerous situations, since straw rows are automatically formed in a such way that it is guaranteed that harvester will meet no stationary obstacles on its way when moving along straw rows. After receiving heatmap from CNN depicted on Fig. 5 it is possible to solve task of straw rows geometry reconstruction, i.e. to find the direction and the offset of the straw row relative to harvester reference frame. Orientation and spatial arrangement of the rows relatively to harvester are required to generate corrective action in a steering subsystem of a harvester.

RESULTS EVALUATION

Two models were trained: with and without data augmentations. Let’s denote these models as deconv+ and deconv respectively.
Figure 6: Segmentation examples produced by our model. Original image (left column), predicted segmentation (right column), segmentation shown on top of the original image (middle column)
On Fig. 7 mask for straw row segments is depicted. The bird’s eye view transform applied to that mask. The bird’s-eye view is perspective transformation technique which generates a top view perspective of an image. Parallel structures on the ground become truly parallel. To perform the bird’s-eye perspective warping accurate camera calibration technique is essential. As result of bird’s-eye view transformation straw rows transformed into segments with persistent form and similar orientation as depicted on Fig. 8.

To detect the relative directions and offset of the rows on the bird’s-eye image the Fast Hough Transform (FHT) [57] is applied. After normalization to the range $[0, 255]$ there are several bright blobs appeared as depicted on Fig. 9. Each blob corresponds to a separate straw row on the bird’s-eye image.

In purpose of direction measurement noise reduction the additional FHT is applied to the result of first FHT. FHT of order 2 (FHT2) sharpens the maximum essentially as depicted on Fig. 10. The physical meaning of the FHT2 is that it detects the mean direction of all observable straw row segments.

Number of bright lobs on Fig. 9 represents the amount of straw rows. To find the best directions for each row independently it is required to perform the FHT for each row independently and choose the local direction with minimal deviation from global direction of all straw rows.

The last stage is to select the unique line with optimum criteria — select the closest straw row. On previous stage there was the family of directions. For the best choice of unique direction line denote cost function as a perpendicular distance from the center of harvester head and to a line (direction) in Euclidean metrics.

The resulting line corresponds to the middle line of the closest segment. The resulting line can be further processed with
some control subsystem to generate steering commands to follow the straw row.

EXPERIMENTAL RESULTS

The proposed CNN was tested on the Advantech UNO UNO-3083G-D64E embedded computer. Brief computer configuration is given below in Table 13.

<table>
<thead>
<tr>
<th>Processor</th>
<th>Core i7-3555LE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU memory</td>
<td>4gb</td>
</tr>
<tr>
<td>GPU card</td>
<td>PNY Quadro k1200 4GB PCIe 4xmDP DVI 1046/1253 128-bit DDR5 512 Cores 4xmDP to DVI-D adapter</td>
</tr>
<tr>
<td>Operational System</td>
<td>Ubuntu 16.04 LTS</td>
</tr>
</tbody>
</table>

Table 13: Hardware details

Processing time of each frame was 165 ms: processing frame by CNN – 150 ms, computing straw row position, orientation – 15 ms. This processing time resulted in a final performance of 6 frames per second. For this purpose of the harvester real-time automatic steering processing frequency of 6 FPS is sufficient.

CONCLUSION

This work demonstrates that the adopted deconvolution network architecture is suitable for straw row segmentation. A series of experiments was conducted to validate high quality segmentation.

Proposed neural network architecture is efficient both in terms of memory and computing time. The network is trained end-to-end since the proposed approach yields accurate straw row segmentation in roughly 6 frames per second.

Accurate camera calibration and Bird’s-eye view transform is essential for straw row geometry reconstruction. The straw row geometry reconstruction can be done with sufficient processing speed to meet the requirements of the harvester automatic steering along the rows.

ACKNOWLEDGEMENTS

This work was financially supported by the Ministry of Education and Science of Russian Federation (Agreement № 14.579.21.0122, Applied Research and Experimental Development individual identifier RFMEFI57915X0122).

REFERENCES


