Detektion eines Grünlandschwades mit Stereo-RGB Kamera

Peter Riegler-Nurscher[1], Johann Prankl[1], Markus Hofinger[2] und Markus Vincze[3]

Abstract: Robustes Detektieren von Grünlandschwaden ist die Grundlage für die Automatisierung bei der Heu- und Silage-Ernte. Vor allem bei kleinem Schwadvolumen ist die Detektion basierend auf Daten von 3D-Sensoren fehleranfällig. Es wird eine neue Methode zur Segmentierung einer Schwad in einem RGB-Bild basierend auf einem Convolutional Neural Network (CNN) vorgestellt. Die Methode wird mit der Segmentierung von 3D-Tiefendaten einer Stereo-Kamera mittels Ebenen-Detektion verglichen. Zur Validierung beider Methoden wurden Aufnahmen bei der Silage- und bei der Heuernte manuell annotiert. Es kann gezeigt werden, dass die CNN-basierte Schwaderkennung bei kleinem Volumen eine höhere Genauigkeit erreicht.

Keywords: Grünland, Schwaderkennung, Convolutional Neural Network

1 Einleitung

Automatische Schwadführung, Ertragserfassung, Maschinensteuerung und Logistikoptimierung bei der Heu- und Silage-Ernte setzen genaues Lokalisieren und Vermessen einer Grünlandschwad voraus [Sc08]. Aktuelle Methoden nutzen Ultraschall-, Radar- oder Laser-Sensoren bzw. Kameras für diese Aufgabe. Kamerabasierte Schwad-Erkennung ermöglicht es, zusätzliche Informationen, wie Textur und Farbe, zu erfassen [BB11]. Vor allem bei geringen Volumina haben Methoden basierend auf 3D-Daten von Laser-Ranger oder reinen 3D-Kameras Probleme bei der Segmentierung der Schwad. Die in [BB11] vorgestellte Methode zeigt das Potenzial zur Segmentierung basierend auf Stereo-Tiefendaten und Texturinformation.

In diesem Beitrag wird eine neue Methode zur Segmentierung einer Schwad in einem RGB-Bild basierend auf einem Convolutional Neural Network (CNN) vorgestellt. Durch Transfer Learning kann die Methode mit einer geringen Anzahl an Trainingsbildern trainiert werden. Die Kamera ist ca. 45° zur Schwad in Fahrtrichtung geneigt, sodass ein höherer Schwad-Längenausschnitt erfasst werden kann. Durch den Aufbau als Stereo-System können zusätzlich das Volumen und absolute Abstände erfasst werden. Ebenso wurde eine Methode basierend auf den Tiefendaten der Stereokamera zur Schwadsegmentierung implementiert. Ein Vergleich der beiden Methoden zeigt deutlich

[1] Josephinum Research, 3250 Wieselburg, Austria, p.riegler-nurscher@josephinum.at
[2] Pöttinger Landtechnik GmbH, 4710 Grieskirchen, Austria, markus.hofinger@poettinger.at
[3] Technische Universität Wien, Institut für Automatisierungs- und Regelungstechnik, 1040 Wien, Austria, vincze@acin.tuwien.ac.at

bessere Erkennungsraten bei der auf CNN basierenden Methode gegenüber der Segmentierung der Tiefendaten.

2 Material und Methoden

2.1 Kamerasystem

Zur Aufnahme des Schwades wurde eine Stereokamera bestehend aus zwei RGB-Kameras mit Trigger zur Synchronisierung der Bildaufnahme verwendet. Die Kamera wurde dabei aus einer schrägen Perspektive auf die Schwad gerichtet. Zur Segmentierung basierend auf den Tiefendaten wird zuerst eine Stereo-Rekonstruktion durchgeführt. Bei der CNN-Segmentierung wird nur das RGB-Bild der linken Kamera verwendet.

2.2 Stereo-Rekonstruktion

Um eine 3D-Punktwolke aus den Stereo-Aufnahmen der Kameras zu generieren, müssen die Kameras kalibriert werden. Dabei werden die intrinsischen und extrinsischen Kameraparameter bestimmt. Nach dieser Kalibrierung müssen die Kameraparameter sowie Kamerapositionen relativ zueinander fix gehalten werden.

Nach der Aufnahme der Stereobilder wird eine Entzerrung und Rektifizierung durchgeführt, damit Epipolarlinien in den beiden Bildern horizontal und co-linear ausgerichtet sind. Beim anschließenden Matching wird versucht, Paare von Bildpunkten in der rechten und linken Aufnahme zu finden. Dazu wird das Blockmatching in der opencv Library [Op19] verwendet. Die Punkt-Paare der texturreichen Schwad können damit effizient bestimmt werden. Aus jedem Bildpunkt-Paar kann danach ein Disparity-Wert, der den Versatz in x-Richtung angibt, bestimmt werden. Die Disparity-Map dient als Basis zur Berechnung der 3D-Punktwolke basierend auf den extrinsischen Kameraparametern. Abschließend wird die Punktwolke gefiltert und Outlier entfernt.

2.3 Segmentierung des Schwades basierend auf den Tiefendaten

Bei der Segmentierung des Schwades basierend auf den Tiefendaten wird die generierte 3D-Punktwolke der Stereokamera herangezogen. Bei dieser Methode wird angenommen, dass sich der Schwad im mittleren Bildbereich befindet und der Boden geometrisch eben ist. Die Schritte von Bildaufnahme bis Schwadsegmentierung sind in Abbildung 1 dargestellt.

Im ersten Schritt wird dazu der mittlere Bereich aus der Punktwolke, in dem der Schwad erwartet wird, entfernt. Die verbleibenden Punkte werden zum Bestimmen einer Ebene

herangezogen. Diese Ebene approximiert die Bodenoberfläche und wird mit der SAC-Segmentation der Point Cloud Library (PCL) [Pc19] bestimmt. Danach wird die Punktwolke rotiert, sodass die Bodenoberfläche in der XY-Ebene liegt. In einem abschließenden Schritt werden alle Punkte oberhalb der Ebene in der ursprünglichen Punktwolke markiert.

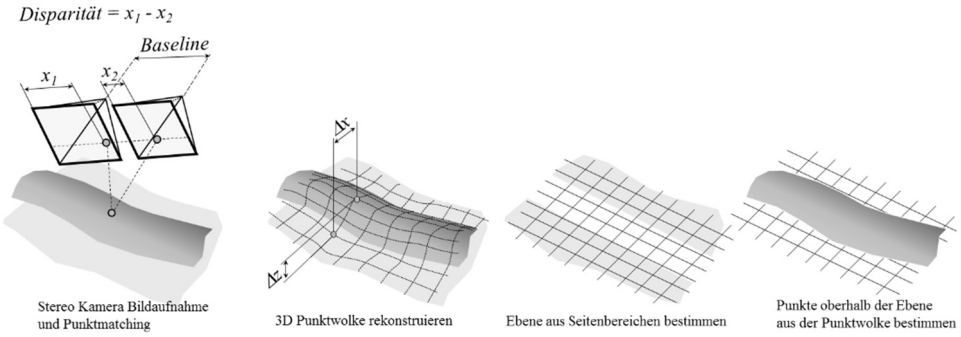

Abb. 1: Schritte von der Bildaufnahme bis zur Segmentierung basierend auf den Tiefendaten der Stereokamera

2.4 Segmentierung des Schwades basierend auf CNN

Bei der zweiten Methode wird eine pixelweise Segmentierung der Schwad-Bereiche mit einem Machine Learning Modell durgeführt.

Als Ausgangsbasis für das Training des Modells dienen pixelweise manuell annotierte Trainingsbilder in die Klassen Schwad oder keine Schwad. Es wurden 154 Trainingsbilder auf ein vortrainiertes CNN, ein ERFNet [Ro18], trainiert. Das Modell wurde auf ein Zuckerrübendatenset (sugar beet dataset [Ch17]) vortrainiert. Zur einfachen Anwendung wurde das Framework aus [MS19] verwendet.

CNNs für die semantische Segmentierung verwenden einen Codierer- und Decodiererblock nacheinander. Durch Downsampling oder Codierung können tiefere Schichten mehr Kontext erfassen und somit die Klassifizierung verbessern. Dies hat jedoch den Nachteil, dass die Pixelgenauigkeit für die semantische Segmentierung verringert wird. ERFNet verwendet sogenannte Factorized Convolutions mit Residual Connections und führt Non-Bottleneck-1D (non-bt-1D) Layer ein. Diese Kombination von 1D-Filtern ist schneller und hat weniger Parameter als Bottleneck Layer. Die Genauigkeit bleibt dadurch gleich wie bei Non-Bottleneck Layer. ERFNet verwendet einfache Entfaltungsschichten mit Stride 2, um die Speicher- und Rechenanforderungen für die Dekodierung zu verringern. Die Architektur eines CNNs für die semantische Segmentierung ist in Abbildung 2 dargestellt.

Nach dem Training mit dem Trainingsdatenset können neue 2D-RGB-Testbilder klassifiziert werden. Nach der Klassifizierung der Bildpunkte im RGB-Bild können alle Punkte in die 3D-Punktwolke projiziert werden, um die entsprechenden 3D-Punkte zu markieren.

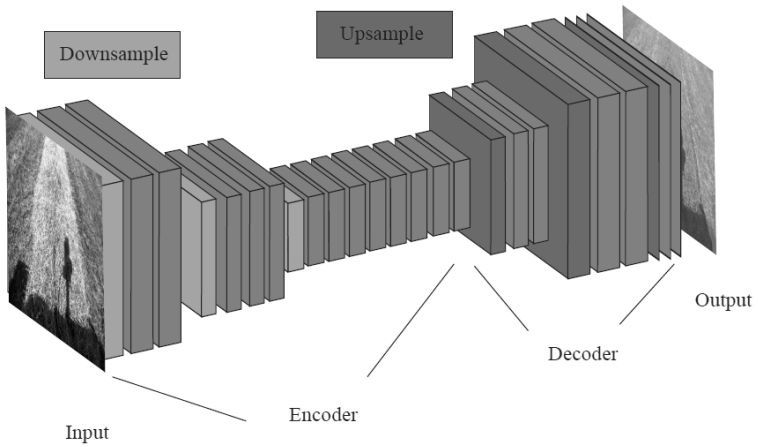

Abb. 2: Architektur des CNN für Semantic Segmentation

3 Ergebnisse

Die Segmentierung inklusive 3D-Punktwolke können in nachfolgenden Prozessen, wie in der Einleitung beispielhaft beschrieben, verwendet werden. Abbildung 3 zeigt ein Beispielbild mit den Stereo-Aufnahmen des Schwades, der 3D-Rekonstruktion und dem markierten Schwad aus der CNN-Segmentierung.

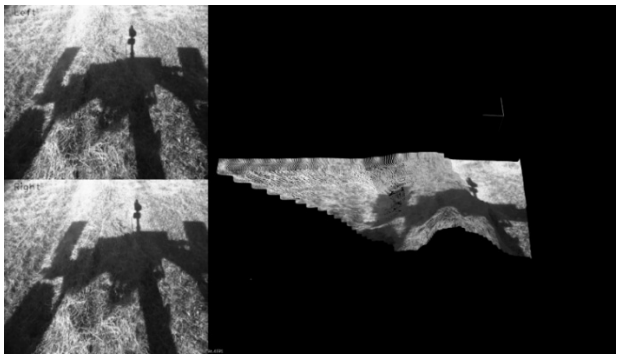

Abb. 3: RGB-Stereo-Bilder eines Schwades (links), Segmentierungsergebnis eines Schwades basierend auf einem CNN (rechts)

Zur Bewertung der Erkennungsrate der beiden Methoden wurden 39 Testsamples aus verschiedenen Szenarien ausgewertet und die in der Bildverarbeitung übliche Metrik der Korrektklassifikationsrate (Accuracy) berechnet. Dabei wird die Anzahl aller richtig klassifizierten Pixel der Gesamtpixelzahl gegenübergestellt. Die Testsamples sind nicht in den Trainingsdaten der CNN-Segmentierung enthalten.

Die Auswertung ergab eine Accuracy von 0,934 für die Segmentierung basierend auf CNN bzw. von 0,820 für die Segmentierung basierend auf der geschätzten Ebene der Tiefendaten.

Abbildung 4 zeigt einige Beispielbilder aus den Testdaten. Bei Schwaden mit geringem Volumen bzw. geringer Höhe, wie in Beispielbild 4 und 5, resultiert bei der Segmentierung basierend auf Tiefendaten ein größerer Fehler. Dies kann durch die geringere Tiefenauflösung des Stereo-Setups gegenüber der Bildauflösung erklärt werden.

Generell können die Genauigkeitseinbußen der Methode basierend auf den Tiefendaten durch geringere Tiefenauflösung, lokale Unebenheiten der Böden und Fehler bei der Ebenschätzung erklärt werden.

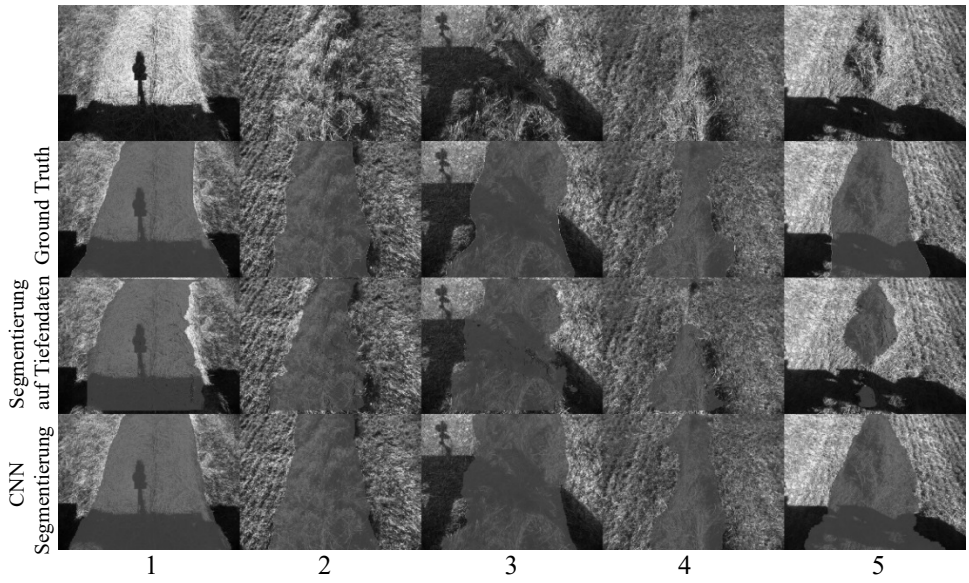

Abb. 4: Beispielbilder von Schwaden mit der manuellen Annotierung (Ground Truth), Segmentierung auf den Tiefendaten und der CNN-basierten Segmentierung

4 Fazit

Die Detektion und Vermessung von Schwaden sind Grundlage für die Automatisierung bei der Heu- und Silage-Ernte und geben Aufschluss über Teilflächenerträge im Grünland. In diesem Beitrag wurde eine neue Methode zur Segmentierung von Schwaden in monokularen Farbbildern basierend auf einem Convolutional Neural Network (CNN) vorgestellt. Die Methode wurde mit dem Stand der Technik, einer Segmentierung in 3D-Punktwolken durch RANSAC und einem Ebenenmodell verglichen. Es wurde gezeigt, dass die CNN-basierte Methode eine höhere Genauigkeit erreicht und auch Schwaden bei geringem Bestand, die sich nicht mehr signifikant von der Bodenoberfläche unterscheiden, segmentieren kann.

In Folgeprojekten soll der Trainingsdatensatz um Bilder von weiteren Kulturen und Stroh, die derzeit nicht abgedeckt sind, erweitert werden. Zusätzlich sollen in die CNN-basierte Methode Tiefendaten integriert werden. Es ist zu erwarten, dass dadurch eine höhere Robustheit bei unterschiedlichen Lichtbedingungen und Schwadgeometrien erreicht werden kann.

Literaturverzeichnis

[BB11] Blas, M. R.; Blanke, M.: Stereo vision with texture learning for fault-tolerant automatic baling, Computers and Electronics in Agriculture, Volume 75, Issue 1, Pages 159-168, ISSN 0168-1699, 2011.

[Ch17] Chebrolu, N.: Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. The International Journal of Robotics Research, 36(10), 1045-1052, 2017.

[MS19] Milioto, A.; Stachniss, C.: Bonnet: An Open-Source Training and Deployment Framework for Semantic Segmentation in Robotics using CNNs, In Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA), 2019.

[Op19] OpenCV Library, https://opencv.org/, Stand: 28.10.2019

[Pc19] PCL Library, http://pointclouds.org/, Stand: 28.10.2019

[Ro18] Romera, E.: ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation, in IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 1, pp. 263-272, 2018.

[Sc08] Schellberg, J. et.al.: Precision agriculture on grassland: Applications, perspectives and constraints. European Journal of Agronomy, 29(2-3), 59-71, 2008.

[SV14] Schellberg, J.; Verbruggen, E.: Frontiers and perspectives on research strategies in grassland technology, Crop and Pasture Science, 65(6), 508-523, 2014