LEHMANN + PARTNER GmbH, Kesselsdorf; Technische Universität Ilmenau

AUTOMATISCHE DETEKTION UND OBJEKTSCHARFE GEOREFERENZIERUNG VON FAHRBAHNSCHÄDEN AUS BILDDATEN EINES MOBILE-MAPPING-SYSTEMS MITHILFE VON DEEP LEARNING

Maximilian Sesselmann, Ronny Stricker, Thorsten Naber, Steffen Schelle

Zusammenfassung: Im Kontext des Straßeninfrastrukturmonitorings werden mit schnellfahrenden Mobile-Mapping-Systemen Kamera- und Laserscannerdaten des Straßenraums aufgenommen. Die georeferenzierende Zustands- und Objekterfassung aus solchen Bilddaten ist noch immer ein mehrheitlich manueller und damit ein zeit- und kostenintensiver Prozess. Dabei haben Deep-Learning-basierte Ansätze unter Einsatz künstlicher neuronaler Faltungsnetzwerke ihre Leistungsfähigkeit hinsichtlich automatisierter Bildanalyse bereits in verschiedensten Anwendungsdomänen bewiesen. In diesem Beitrag wird einerseits gezeigt, wie Fahrbahnschäden mithilfe tiefer neuronaler Netze aus Bildern automatisiert detektiert und klassifiziert werden können und andererseits, wie die erkannten Schadstellen objektscharf und mit einer absoluten Genauigkeit besser 10 cm georeferenziert werden können.

Schlüsselwörter: Mobile Mapping, Straßenzustand, Deep Learning, automatische Schadstellenkartierung

AUTOMATIC DETECTION AND GEOREFERENCING OF ROAD DAMAGE FROM A MOBILE MAPPING SYSTEMS IMAGERY WITH THE HELP OF DEEP LEARNING

Abstract: In the context of road infrastructure monitoring, asset inventory and condition survey from camera image data acquired with fastdriving mobile mapping systems is still a manual and therefore timeconsuming and costintensive process. Deep learning based approaches using convolutional neural networks have already proven their efficiency in automated image analysis in various application domains. The present contribution shows how road damage can be detected and classified automatically from images using deep neural networks. Furthermore, the detected damage can be georeferenced with an absolute accuracy better than 10 cm.

Keywords: Mobile mapping, road condition, deep learning, automated damage mapping

Autoren

Dipl.-Geogr. Maximilian Sesselmann
Dipl.-Ing. Thorsten Naber
Dipl.-Ing. Steffen Scheller
LEHMANN+PARTNER GmbH
Fachbereich Forschung & Entwicklung
Sachsenallee 24
D-01723 Kesselsdorf
E: sesselmann@lehmann-partner.de
naber@lehmann-partner.de
scheller@lehmann-partner.de

Dipl.-Inf. Ronny Stricker
Technische Universität Ilmenau
Institut für Technische Informatik und Ingenieurinformatik
Fachgebiet Neuroinformatik und Kognitive Robotik
D-98684 Ilmenau
E: ronny.stricker@tu-ilmenau.de

1 EINLEITUNG

Die Erfassung von Geoobjekten aus Kamerabilddaten, die mit schnellfahrenden Mobile-Mapping-Systemen aufgenommen werden, ist heute Alltag für Dienstleister im Bereich des Straßeninfrastrukturmonitorings. Dabei werden in der Regel Fahrbahnschäden, Straßeninventar oder Flächennutzungen in den Bilddaten digitalisiert und mithilfe verschiedenster Verfahren georeferenziert. Die bislang überwiegend manuelle Bildauswertung ist dabei vor allem ein zeit- und damit kostenintensiver Prozess. Im Kontext automatisierter Datenauswertung sind Deep-Learning-Ansätze unter Einsatz künstlicher neuronaler Faltungsnetzwerke (Convolutional Neural Networks, kurz: CNN) mittlerweile Stand der Technik. Für das Anwendungsfeld der Straßenzustandserfassung wurde im Rahmen des ASINVOS-Projekts systematisch untersucht, wie tiefe neuronale Faltungsnetze für eine Automatisierung der Schadenserkennung eingesetzt werden können (BMBF 2016). Die Untersuchungen beschränkten sich hierbei jedoch auf hochauflösende Oberflächenbilder, die mithilfe schnellfahrender Messfahrzeuge aus der Vogelperspektive aufgenommen wurden. In diesem Beitrag wird nun untersucht, wie die dort entwickelte Methodik für Bilddaten des Straßenraums adaptiert werden kann, die von in Fahrtrichtung ausgerichteten Kameras erfasst werden. In den Bildern dieser sogenannten Umfeldkameras können aufgrund der Aufnahmegeometrie Schäden und Objekte gleicher Größe in sehr unterschiedlichen Skalierungs- und Verzerrungsgraden abgebildet sein – ein Umstand, für den das bildausschnittbasierte CNN des ASINVOS-Systems eigentlich nicht konzipiert wurde und der das zu lösende Klassifikationsproblem potenziell erschwert. Das Ziel dieses Beitrags ist es daher zu überprüfen, ob und in welchen Grenzen straßenzustandsrelevante Oberflächenschäden mit einem bildausschnittbasierten CNN automatisiert aus Bildern detektiert werden können, die nicht aus der Draufsicht, sondern aus einer High-Angle-Perspektive aufgenommen wurden. Zu diesem Zweck wird ein umfangreicher Trainingsdatensatz mit Bilddaten aus dem deutschen Straßennetz aufbereitet und Klassifikatoren auf dieser Basis trainiert und evaluiert. Weiterhin wird ein 3D-Bildverarbeitungsansatz für monokulare Bildaufnahmen vorgestellt und getestet, mit dem die im Bild erkannten Schäden als Geoobjekte kartiert werden können. Hierzu werden verschiedene Sensordaten eines Mobile-Mapping-Systems fusioniert.

2 MOBILE-MAPPING-SYSTEM I.R.I.S

Im Rahmen der Straßenzustand- und Inventarerfassung sind schnellfahrende Multisensorsysteme Stand der Technik. Derartige Systeme, die im DACH-Raum eingesetzt werden, sind beispielsweise RoadSTAR (Austrian Institute of Technology GmbH), ARGUS (TÜV Rheinland Schniering GmbH), infra3D (iNovitas AG) sowie S.T.I.E.R und I.R.I.S (Lehmann + Partner GmbH). Die Datengrundlage für den vorliegenden Beitrag wurde mit dem kinematischen Erfassungssystem I.R.I.S (Integrated Road Information System) aufgenommen. I.R.I.S ist ein schnellfahrendes Mobile-Mapping-System zur bildhaften und dreidimensionalen Erfassung des Straßenraums. Das Messfahrzeug ist so konzipiert und dimensioniert, dass eine Befahrung sowohl des Fernstraßennetzes als auch des kommunalen Straßennetzes im fließenden Verkehr möglich ist. Die auf dem Fahrzeugdach montierte Rahmenkonstruktion trägt die Sensorik, wie zum Beispiel mehrere Umfeldkameras, Laserscanner und ein Positionierungssystem (Abbildung 1). Nachstehend werden die für den Beitrag relevanten Komponenten des I.R.I.S-Messsystems kurz vorgestellt.

Zu den Kernkomponenten zählt das Positionierungssystem vom Typ Applanix POS LV 420. Dabei handelt es sich um ein integriertes System aus globalem Navigationssatellitensystem (GNSS), inertialer Messeinheit (IMU) und einem Wegstreckenmesser. Durch die Kombination dieser Bestandteile ist sowohl die Bestimmung der absoluten Position als auch der relativen

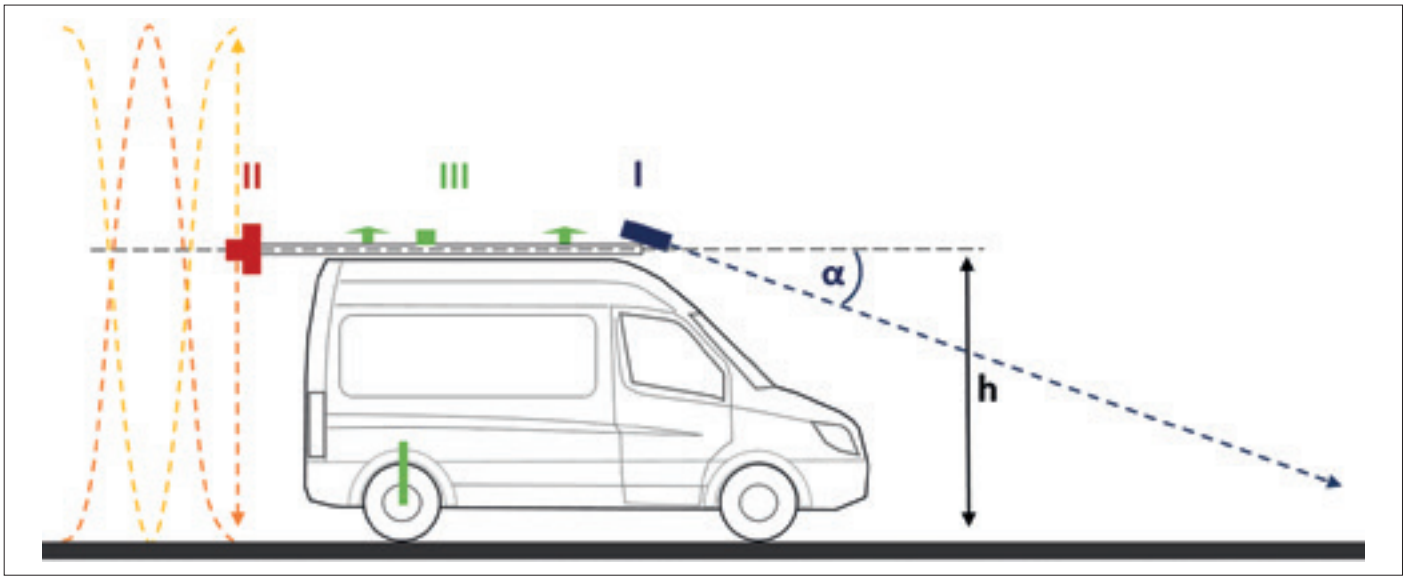


Abbildung 1: Prinzip der Sensorkonfiguration des Mobile-Mapping-Systems I.R.I.S. Im Beitrag werden die Daten der in Fahrtrichtung schräg (α) auf die Fahrbahnoberfläche ausgerichteten Umfeldkamera (I) und des am Heck montierten Fraunhofer Clearance-Profile-Scanners (II) verwendet. Letzterer ist ein Rotationslaserscanner zur Erfassung von Lichtraumprofilen. Mithilfe eines Positionierungssystems (III) bestehend aus globalem Navigationssatellitensystem, inertialer Messeinheit und Wegstreckenmesser wird die Trajektorie des Messfahrzeugs aufgezeichnet. Die Messsensoren sind auf einem Trägerrahmen in einer Höhe h von 3 m verbaut. Der Anstellwinkel α beträgt 14°.

Positionsänderung inklusive aller Raumwinkel und Beschleunigungen möglich. Das GNSS dient dabei der Positionierung des Systems im globalen Raumbezug. Die IMU nutzt zur Bestimmung der Winkeländerungen und der relativen Orientierung einen faseroptischen Kreisel mit drei Beschleunigungssensoren für jede Bewegungsrichtung. Mit einer Frequenz von 200 Hz werden die Beschleunigungen und Winkeländerungen für alle drei Raumachsen aufgezeichnet. Die eingesetzte IMU besitzt einen Winkelfehler von weniger als 0.6 Grad je Stunde, sodass auch bei schlechten GNSS-Bedingungen eine Positionsbestimmung mit einer geringen sensorspezifischen Drift fortgeführt werden kann. Zusätzlich ist zur Erfassung des zurückgelegten Wegs ein Wegstreckenmesser an einem Rad montiert. Je Radumdrehung werden etwa 1 800 Pulse aufgezeichnet. Die Auflösung beträgt somit, je nach Radumfang, circa 1 mm je Puls. Aus den Rohdaten aller im Fahrzeug verbauten Sensoren wird während der Messung mithilfe eines Kalman-Filters eine Realtime-Trajektorie berechnet, welche die Position und alle Raumwinkel auf Basis der Rohdaten enthält. Die Ermittlung der finalen Positionen und Winkel erfolgt mithilfe der proprietären Software POSPac MMS unter Nutzung externer Korrekturdaten (Landau et al. 2002). Im Ergebnis der Berechnung liegt die Fahrzeugtrajektorie als Punktfolge mit allen relevanten Informationen, wie beispielsweise Zeitstempel, Beschleunigungen, Raumwinkel und Koordinaten, vor. Das eingesetzte Positionierungssystem ist bei GNSS-Abdeckung mit 2 cm und bei 60-sekündigem GNSS-Ausfall mit 12 cm Lagegenauigkeit (RMS) spezifiziert (Applanix 2019).

Zur bildhaften Dokumentation des Straßenraums und messtechnischen Objekterfassung ist das I.R.I.S-System mit fünf Messkameras ausgerüstet, die den Straßenraum aus einer High-Angle-Perspektive in verschiedene Richtungen aufnehmen. Alle Kameras nutzen CCD-Sensoren, nehmen RGB-Bilder mit einer Auflösung von 2 560 x 1 920 Pixeln auf und sind photogrammetrisch kalibriert. Die Umfeldkameras werden weggesteuert ausgelöst und jeder Aufnahme wird dabei ein eindeutiger Zeitstempel zugewiesen. Hiermit kann jedem aufgenommenen Bild eine Position und Orientierung auf der Fahrzeugtrajektorie zugeordnet werden. Dies ist die Grundlage, um abgebildete Objekte präzise mithilfe der Kamerabilder zu georeferenzieren. Im Beitrag wird ausschließlich Bildmaterial einer Kamera verwendet, die den Straßenraum vor dem Messfahrzeug abbildet (Abbildung 1).

Zur dreidimensionalen Erfassung eines circa 30 bis 60 m breiten Korridors entlang der Fahrzeugtrajektorie ist das I.R.I.S-Messfahrzeug mit einem LiDAR-System ausgerüstet. Der integrierte Fraunhofer Clearance-Profile-Scanner (CPS) ist ein Rotationslaserscanner zur Erfassung von Lichtraumprofilen, der nach dem Phasenvergleichsverfahren misst. Durch die Rotation eines Spiegels erfasst der Laserscanner ein 2D-Profil in einem Aufnahmebereich von 350°. Der CPS befindet sich im Heckbereich und ist so ausgerichtet, dass die einzelnen Profile quer zur Fahrtrichtung aufgenommen werden. Durch die Bewegung des Messfahrzeugs beschreibt der Laserstrahl eine Helix, wodurch der Straßenraum entlang der Trajektorie sukzessive abgetastet wird (Abbildung 1). Der CPS erreicht eine Datenrate von 1 MHz und erfasst circa 200 Profile je Sekunde. Die Genauigkeit der Längenmessung beträgt je nach Reflexionseigenschaft einer Oberfläche in 5 m Entfernung 3 bis 7 mm. Messungen im öffentlichen Raum sind aufgrund der verwendeten Laserklasse 1 ohne Einschränkung möglich. Weitere technische Details des CPS sind beim Fraunhofer Institute for Physical Measurement Techniques IPM (2019) beschrieben. Wie auch die Bildaufnahmen werden die Rohmessungen des CPS (Winkel, Entfernung und Reflexionsintensität) mit Zeitstempeln versehen. Somit können auch die Lasermessungen mit Positionen und Orientierungen auf der Fahrzeugtrajektorie synchronisiert und somit 3D-Punktwolken als Datenprodukt erzeugt werden.

3 METHODIK

Das dem Beitrag zugrunde liegende Konzept zur automatischen Kartierung von

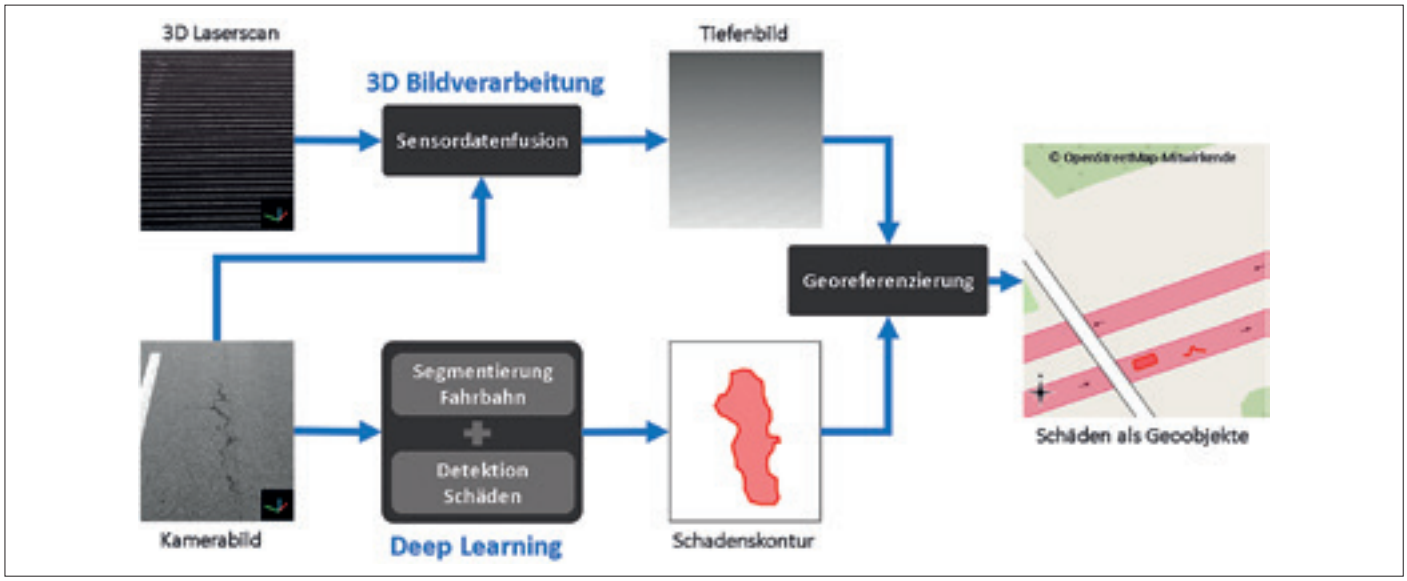


Abbildung 2: Schematische Darstellung des Ablaufs zur automatischen Kartierung von Fahrbahnschäden aus Kamerabildern und 3D-Laserscans des Mobile-Mapping-Systems I.R.I.S. Die Bildanalyse mithilfe tiefer neuronaler Netze bildet dabei den methodischen Kern. Die Georeferenzierung der Detektionsergebnisse wird über die Tiefeninformation der 3D-Laserscans realisiert.

Fahrbahnschäden mithilfe von Deep Learning und 3D-Bildverarbeitung auf Basis von Messdaten eines Mobile-Mapping-Systems ist in Abbildung 2 schematisch dargestellt. Tiefe neuronale Netze zur Analyse von Bilddaten bilden dabei den methodischen Kern der Verarbeitungskette. In vielen Anwendungsdomänen der Objekterkennung und Bildsegmentierung zeigten sich Deep-Learning-Ansätze konventionellen maschinellen Lernverfahren überlegen (Liu et al. 2016, Karaca et al. 2017). Für das Anlernen dieser Netzwerke kommen häufig überwachte Lernverfahren zum Einsatz, bei denen Wissen in Form von annotierten Daten zu Verfügung gestellt werden muss. Das Modell lernt, wie sich die Daten mit den vorgegebenen Annotationen (Labels) verknüpfen lassen. Im Gegensatz zu konventionellen maschinellen Lernverfahren, bei denen ein explizites Design der Merkmalsextraktion nötig ist, ist die Merkmalsextraktion bei tiefen neuronalen Faltungsnetzen impliziter Teil der Netzwerkarchitektur: Welche Merkmale in welcher Ausprägung relevant sind, wird während eines Trainingsprozesses datengetrieben gelernt. Gute Trainingsdaten und einen geeigneten Trainingsprozess vorausgesetzt, ist das trainierte Modell in der Lage, zu generalisieren und auch auf unbekannten Daten korrekte Schätzungen zu erzeugen.

Für die Erkennung und Klassifikation von Fahrbahnschäden wird die Architektur eines Schadstellen- und Objektdetektionsnetzwerks genutzt, das bei Stricker et al. (2019) vorgestellt wurde. Dieses tiefe bildausschnittbasierte Faltungsnetzwerk wurde entworfen, um Fahrbahnschäden aus Bilddaten zu erkennen und zu klassifizieren, die ausschließlich die Fahrbahnoberfläche fokussieren. Die Daten, für die dieses CNN ursprünglich entwickelt wurde, weisen aufgrund ihrer speziellen Aufnahmegeometrie (Draufsicht aus circa 3 m Höhe) kaum eine Variation der Schadens- und Objektausprägung infolge perspektivischer Effekte auf. Die für den vorliegenden Beitrag relevanten Umfeldkameras bilden dagegen aufgrund ihrer Perspektive Schäden und Objekte nicht nur in unterschiedlichsten Skalierungs- und Verzerrungsgraden ab, sondern sie nehmen neben der Fahrbahnoberfläche auch zahlreiche andere Objekte des Straßenraums, wie beispielsweise Fahrzeuge, Vegetation oder Gebäude auf. Daher wird zunächst in einem Vorverarbeitungsschritt die Fahrbahn segmentiert, bevor die eigentliche Detektion der Fahrbahnschäden durchgeführt wird. Für die Segmentierung der Fahrbahnoberfläche kann ein etabliertes Verfahren mit einem vortrainierten Modell verwendet werden. Für die Schadenserkennung wird hingegen ein spezielles neuronales Netz basierend auf der bei Stricker et al. (2019) erläuterten Architektur konzipiert und ein entsprechendes Modell trainiert. Grundlage dieses Trainings ist der in Kapitel 4 beschriebene Datensatz.

Da das Ziel dieses Beitrags nicht nur eine Lokalisierung und Klassifizierung von Fahrbahnschäden im Bildkoordinatensystem, sondern darüber hinaus die Kartierung von Schäden als Geoobjekte ist, werden im letzten Methodenteil die hierfür nötigen Aspekte der 3D-Bildverarbeitung thematisiert. Die dort präsentierte Sensordatenfusion ermöglicht sowohl ein manuelles Messen von 3D-Koordinaten aus den monokularen Bildaufnahmen des I.R.I.S-Systems als auch die automatische Georeferenzierung von im Bild erkannten Schadenskonturen.

3.1 VOLLBILDBASIERTE SEGMENTIERUNG DER FAHRBAHNOBERFLÄCHE

Die Kamerabilder der Umfeldkameras bilden, wie vorstehend beschrieben, weit mehr als nur die Straßenoberfläche ab. Daher ist es sinnvoll, eine Straßenszene zunächst semantisch zu segmentieren, um nur jene Pixel zu extrahieren, die tatsächlich Fahrbahnbereiche zeigen. Wird anschließend das neuronale Netz, welches die eigentliche Schadstellendetektion durchführt, nur auf die als Fahrbahn erkannten Bildbereiche angewendet, kann das Auftreten potenzieller falschpositiver Detektionen, beispielsweise auf Gebäuden, deutlich reduziert werden.

Im Bereich der automatisierten Objekterkennung aus Bilddaten ist der von Google-Entwicklern im Jahr 2018 vorgestellte, vollbildbasierte DeepLabv3+-Ansatz aufgrund seiner Leistungsfähigkeit weit verbreitet (Chen et al. 2018). Nicht nur auf dem PASCAL Visual-Object-Classes-Referenzdatensatz für Objekterkennung (Everingham et al. 2015), sondern auch auf Benchmark-Datensätzen, bei denen es primär um die Segmentierung von urbanen Straßenszenen geht, erzielt DeepLabv3+ sehr gute Ergebnisse (Cordts et al. 2016). DeepLabv3+ macht Gebrauch vom Konzept des sogenannten Atrous Spatial Pyramid Pooling bzw. der dilatierten Fal-

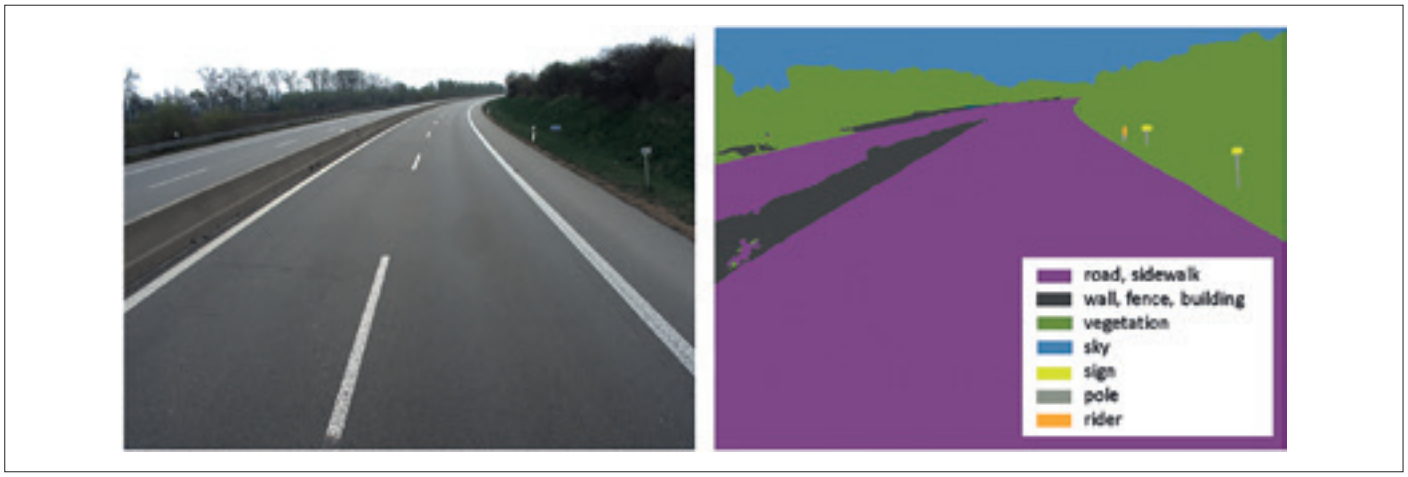


Abbildung 3: Vergleich von Kamerabild (links) und Segmentierungsergebnis (rechts), welches mit DeepLabv3+ erzeugt wurde. Die Klassen „road" und „sidewalk" (violett) dienen als begrenzende Maske für die Anwendung der Schadstellendetektion im anschließenden Verarbeitungsschritt.

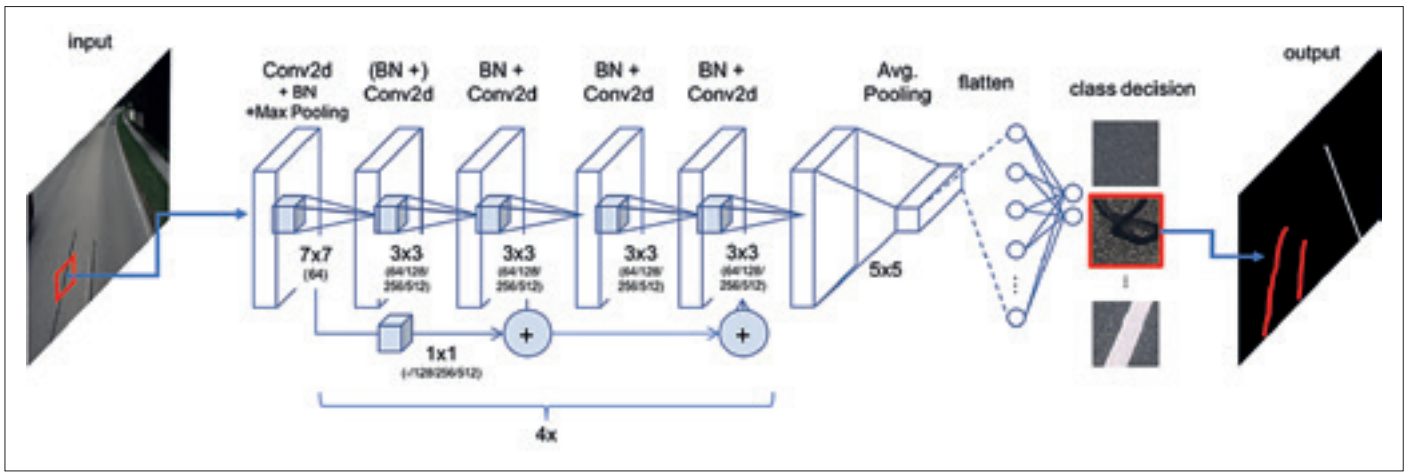


Abbildung 4: Schematische Darstellung eines Residual Networks mit 18 Faltungsschichten (Conv2d). Ein Block besteht jeweils aus zwei Faltungsschichten und die Ausgaben des Vorgängerblocks sind immer mit der Ausgabe des aktuellen Blocks durch eine Additionsschicht verbunden. Nach jeder Faltungsschicht folgt jeweils eine Batch-Normalization-Schicht (BN). Jeder Bildausschnitt (Patch) des Bilds wird jeweils einer Klasse zugeordnet, sodass sich als Ausgabebild eine Schadstellen- oder Objektkarte für das Eingabebild ergibt.

tung, um Kontextinformation auf verschiedenen Skalenniveaus zu codieren und nutzt eine Encoder-Decoder-Architektur, mit deren Hilfe räumliche Information wiederhergestellt werden kann. Details zur Netzwerkarchitektur sind bei Chen et al. (2018) beschrieben. Vortrainierte Modelle, die auf dem Cityscapes-Datensatz basieren (Cordts et al. 2016) und über Github (2019) zugänglich sind, können neben einer Vielzahl anderer Objekte auch Straßen und Gehwege erkennen. Da es sich bei DeepLabv3+ um einen semantischen Segmentierungsansatz handelt, wird jedem Pixel des Eingabebilds eine Objektklasse zugeordnet und darauf aufbauend Masken erzeugt, die das gesamte Eingabebild lückenlos abdecken. Wie in Abbildung 3 dargestellt ist, werden Objekte konturscharf erkannt und präzise im Bild lokalisiert. In diesem Beitrag werden nur die Masken der Klassen „road" und „sidewalk" weiterverwendet, um in einem nachfolgenden Verarbeitungsschritt die Detektion von Fahrbahnschäden nur auf den Bildbereichen anzuwenden, die zur Fahrbahn gehören. Auf dem Cityscapes-Benchmark-Datensatz erreicht das DeepLabv3+-Modell Intersection-Over-Union (IoU) Werte von 98.7 % für die Klasse „road" und 87 % für die Klasse „sidewalk" (Cityscapes 2019).

3.2 BILDAUSSCHNITTBASIERTE SCHADSTELLENDETEKTION MITTELS TIEFEM FALTUNGSNETZWERK

Für die Detektion und Klassifikation von Fahrbahnschäden auf Basis hochauflösender, aus der Draufsicht aufgenommener Oberflächenbilder wurde im Rahmen eines vom BMBF geförderten Forschungsprojekts der sogenannte ASINVOS-Ansatz entwickelt (Eisenbach et al. 2017, Seichter et al. 2018, Stricker et al. 2019). Die genutzten Oberflächenbilder waren mit 11 505 x 5 115 Pixel und einer Pixelauflösung von circa 1,2 mm in Relation zu den zu erkennenden Schäden sehr groß. Da auch feinste Risse erkannt werden sollten, wäre eine Skalierung der Bilddaten, welche den Einsatz klassischer Segmentierungsnetzwerke ermöglichen würde, nicht sinnvoll gewesen. In dem Projekt kam daher ein bildausschnittbasiertes CNN zum Einsatz. Um hierbei Trainingsdaten für tiefe Faltungsnetzwerke zu generieren, werden verschiedene Teilbilder gleicher Größe aus Kamerabildern ausgeschnitten. Diesen Bildausschnitten (Patches) wird jeweils eine Zielklasse (zum Beispiel „Riss" oder „Flickstelle") zugewiesen und das CNN damit trainiert. Anschließend kann das Netzwerk durch Ausschneiden von Patches auf einem unbekannten Bild angewendet werden und erzeugt als Ergebnis eine Schadstellenmaske. Weitere Details zu diesem Trainingsvorgang und wie das Netzwerk transformiert werden kann, sodass die Detektion auf einem ganzen Bild erfolgt, ohne dass dazu quadratische Teilbilder ausgeschnitten werden müssen, können in Eisenbach et al. (2017) nachgeschlagen werden.

Im Lauf der letzten Jahre haben sich bei Faltungsnetzwerken unterschiedliche Netzwerkarchitekturen durchgesetzt, welche sich im Aufbau der Schichten, aber auch hinsichtlich der genutzten Techniken unterscheiden. VGG-basierte Netzwerke (Simonyan & Zisserman 2015) stellen zum Beispiel eine klassische Netzwerkarchitektur für Faltungsnetzwerke dar. Aufgrund der großen Anzahl an Gewichten müssen bei dieser Netzwerkarchitektur jedoch viele Regularisierungstechniken eingesetzt werden, damit sich das Netzwerk nicht zu stark auf die Trainingsdaten spezialisiert und auch auf Testdaten eine gute Klassifikation ermöglicht. Für die Experimente in diesem Beitrag kommt eine modernere Netzwerkarchitektur in Form des Residual Neural Networks (He et al. 2016) zum Einsatz, die für die gegebene Problemstellung deutlich schneller als VGG-basierte Netzwerke trainiert werden können. Auch für Fragestellungen aus dem Bereich Fernerkundung, bei denen es beispielsweise um die Detektion und Klassifikation von Schiffen aus Satellitenbildern geht, werden Residual Networks (ResNets) erfolgreich eingesetzt (Voinov 2020). Die grundlegende Idee von ResNets ist die Einführung von Shortcut-Connections. Das Netzwerk wird dabei in Blöcken von Faltungsschichten organisiert, wobei jeweils zwischen zwei aufeinanderfolgenden Blöcken eine zusätzliche Shortcut-Connection existiert (Abbildung 4). Dadurch wird der Gradientenfluss während des Trainings verbessert und ein Block kann sich jeweils auf die Lösung eines Teilproblems konzentrieren. Da ResNets außerdem auf vollverschaltete Schichten weitestgehend verzichten, kommen sie, trotz größerer Tiefe, mit deutlich weniger Gewichten als VGG-basierte Netzwerke aus. Die gewählte Netzwerkarchitektur liefert im Gegensatz zu vollbildbasier-

ten Segmentierungsverfahren nur für kleine Bildausschnitte eine Klassenentscheidung. Für die Anwendung des trainierten Netzes auf einem vollständigen Eingabekamerabild werden deshalb schrittweise mehrere überlappende Bildausschnitte klassifiziert. Die Klasse mit der maximalen Netzwerkausgabe wird, entsprechend der Position des präsentierten Bildausschnitts, in eine Ergebnismaske eingetragen, wodurch einzelnen Bildbereiche bestimmten Klassen zugeordnet werden können (Abbildung 4).

3.3 3D-BILDVERARBEITUNG: SENSORDATENFUSION ZU TIEFENBILDERN

In der digitalen Bildverarbeitung werden Tiefenbilder in der Regel genutzt, um die Entfernung eines Szenenobjekts von einem Standpunkt aus zu quantifizieren oder um relative Aussagen über die Tiefenstaffelung von Objekten im Bild zu treffen. Es existieren verschiedene Möglichkeiten, solche Tiefeninformation für Kamerabilddaten zu generieren. Klassischerweise werden hierzu photogrammetrische Verfahren der Mehrbildauswertung genutzt (Luhmann 2018). Andere Ansätze nutzen Deep Learning zur Tiefenschätzung aus monokularen Bildaufnahmen (Bhoi 2019). Da im I.R.I.S-System jedoch ein Laserscanner integriert ist, kann die Tiefeninformation zu jedem Kamerabild über eine Sensordatenfusion von Kamerabild- und Laserscannerdaten erzeugt werden. Ziel dieser 3D-Bildverarbeitung im Rahmen des Beitrags ist es, mithilfe der Tiefeninformation und den Kalibrierparametern der beteiligten Sensoren von den Pixelkoordinaten eines Objekts im Kamerabild dessen 3D Koordinaten im globalen Koordinatensystem abzuleiten.

Maßgeblich für die Georeferenzierung von Bildinhalten ist, neben der in Kapitel 2 beschriebenen Messsensorik, vor allem deren präzise Kalibrierung sowie zeitliche Synchronisation. Der Kalibrierprozess umfasst sowohl die innere und äußere Sensorkalibrierung als auch die Orientierung der Messplattform im übergeordneten Koordinatensystem. Über die Parameter Sensorgröße, Pixelgröße, Kamerakonstante, Hauptpunktverschiebungen sowie verschiedenen Verzeichnungsparametern der Optik wird der Strahlengang jeder einzelnen Kamera beschrieben (Luhmann 2018). Diese innere Orientierung verknüpft jedes Element im Pixelkoordinatensystem mit dem jeweiligen Bild- und Sensorkoordinatensystem. Über die Kollinearitätsgleichung und die Verzeichnungsparameter können subpixelgenau dreidimensionale Objektstrahlen berechnet werden. Diese beschreiben mathematisch den Verlauf eines Lichtstrahls durch das Objektiv in den Objektraum. Die äußere Kalibrierung stellt die eindeutige Beschreibung der Rotation und Translation vom jeweiligen Sensorkoordinatensystem zum Fahrzeugkoordinatensystem dar. Diese Kalibrierungen und Transformationen sind für die Abbildung eines Messobjekts durch verschiedene Sensoren an ein und derselben Position im globalen Koordinatensystem von grundlegender Bedeutung.

In einem ersten Schritt der Tiefenbilderzeugung werden die Rohmessdaten des Laserscanners durch die Messungen des Positionierungssystems und die Kalibrierung des Sensors zu einer dreidimensionalen Punktwolke im globalen Koordinatensystem verortet. Im zweiten Schritt werden die Messkameras eindeutig mit dem Positionierungssystem synchronisiert, sodass eine mathematische Beschreibung des Sensors im übergeordneten Koordinatensystem erfolgen kann. Nun wird in einem dritten Schritt pro Bild jeder einzelne Objektstrahl jedes Pixels der Kamera mit der 3D-Punktwolke verschnitten. Trifft ein Objektstrahl dabei auf eine 3D-Punktmessung, wird die Entfernung von der Bildebene zum Objektpunkt als Tiefeninformation an der entsprechenden Bildposition codiert. Im Ergebnis wird parallel zum Kamerabild eine Bildmatrix mit realen Abständen zu den 3D-Messungen des Laserscanners erzeugt. In umgekehrter Richtung kann mithilfe dieses Tiefenbilds zu jedem beliebigen Pixel im zugehörigen Kamerabild ein Objektstrahl erzeugt und unter Berücksichtigung der Entfernungsinformation des Tiefenbilds sowie der Kalibrierparameter eine 3D-Geokoordinate berechnet werden.

4 DATENSATZ

Die Datenbasis für die im Beitrag avisierte Schadstellendetektion sind 1013 Kamerabilder des I.R.I.S-Systems. Um eine möglichst diverse Datenbasis zu erhalten, wurden Bilddaten aus 30 Messkampagnen ausgewählt, die in verschiedenen Regionen Deutschlands durchgeführt wurden. Im zweiten Schritt wurden Bilder selektiert, die möglichst diverse Ausprägungen der relevanten Fahrbahnschäden aufweisen. Im dritten Schritt wurde jedes Bild manuell mithilfe von Polygonmasken nach dem Vorbild des Cityscapes-Datensatzes gelabelt (Cordts et al. 2016). Abbildung 6 zeigt beispielhaft links ein Kamerabild und rechts die gelabelten Objekte und Schäden in Form farbcodierter Masken, die das Bild semantisch segmentieren. Der Klassenkatalog umfasst dabei nicht nur an das für

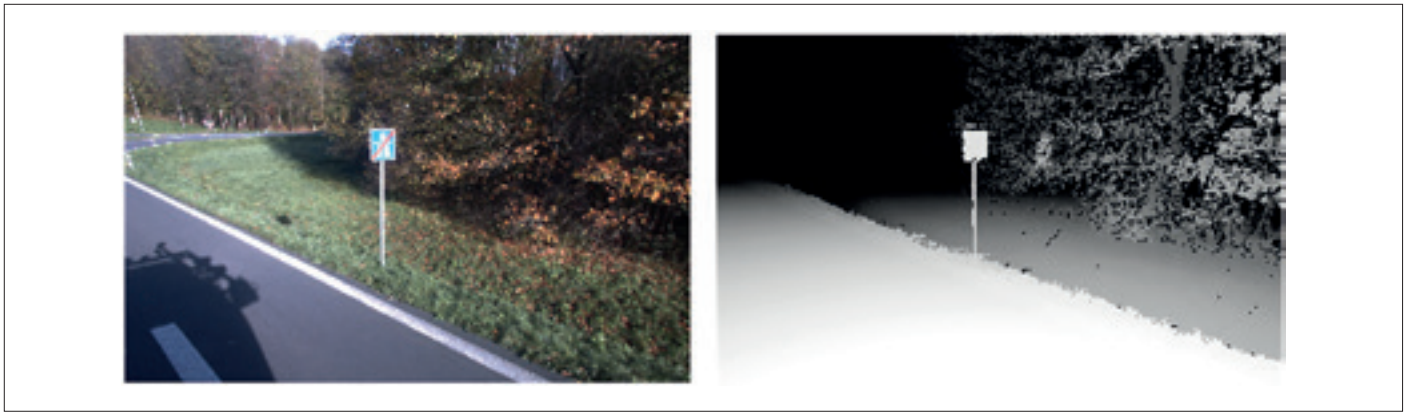


Abbildung 5: Beispiel für ein Bild der rechten Seitenkamera (links) und ein aus der Sensorfusion resultierendes Tiefenbild (rechts). Die Helligkeit der Pixel im Tiefenbild codiert die Entfernung der Sensorebene zum 3D-Messpunkt im Objektraum.

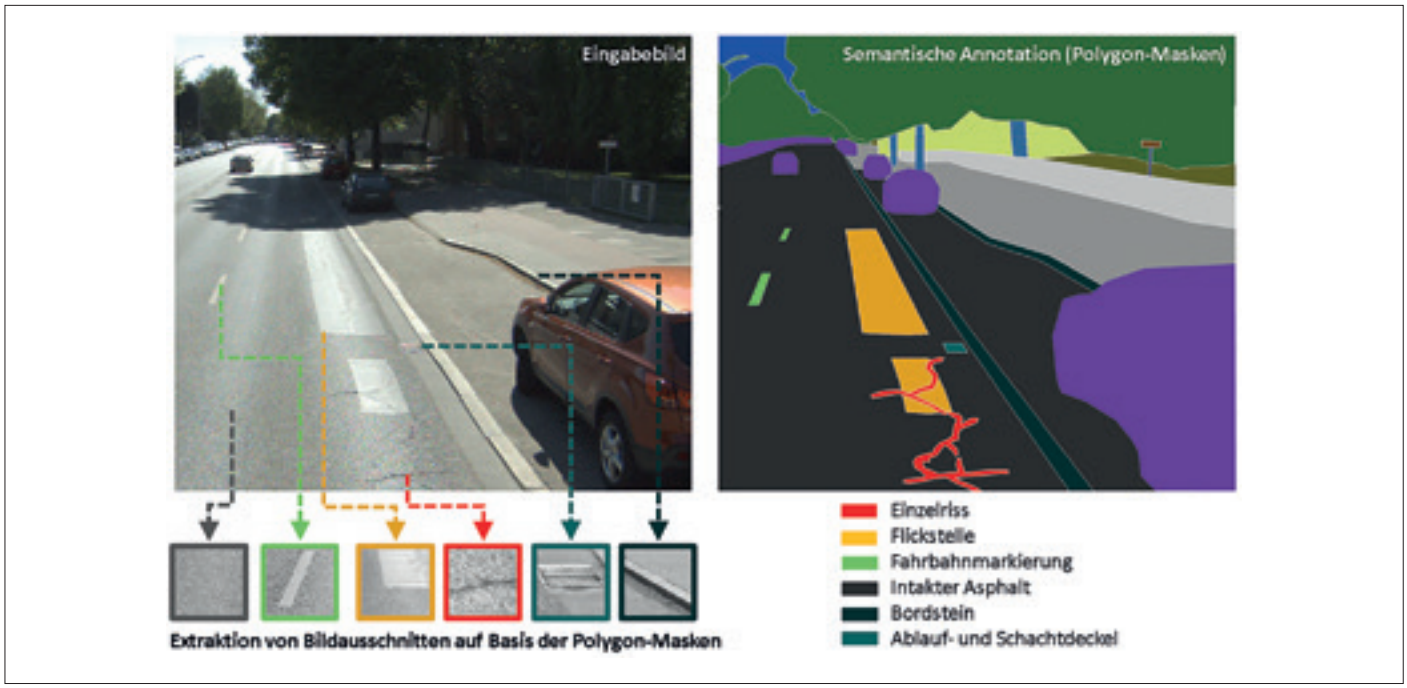


Abbildung 6: Prinzipskizze zur Aufbereitung des Datensatzes. Kamerabilder der Umfeldkameras wurden manuell mithilfe von Polygon-Masken semantisch annotiert. Auf Basis dieser Masken wurden anschließend quadratische Bildausschnitte (Patches) extrahiert, die den eigentlichen Datensatz zum Training und zur Evaluation des neuronalen Netzes darstellen.

Deutschland maßgebliche Regelwerk zur Zustandserfassung und -bewertung von Straßen angelehnte Schadklassen (FGSV 2018), sondern auch Objektklassen, die für die automatische Klassifikation als potenziell hilfreich angenommen wurden. Zu letzterem zählen unter anderem Bordsteine oder Fahrbahnmarkierungen. Tabelle 1 gibt einen Überblick über den Klassenkatalog.

Da das im Beitrag verwendete neuronale Netzwerk auf einer festen und relativ kleinen Bildgröße arbeitet (Stricker et al. 2019), kann es nicht direkt auf den vollen Originalbildern trainiert werden, sondern es müssen vor dem eigentlichen Training auf Basis der Label-Polygone kleine Teilbilder gleicher Größe – sogenannte Bildausschnitte oder Patches – aus den Bildern extrahiert werden. Während diese Extraktion für die Klasse „Flickstelle" nur entlang der Maskenkontur stattfand, wurden die Bildausschnitte für alle anderen Klassen zentral aus der Fläche der entsprechend gelabelten Polygonmaske gesampelt. Die resultierenden circa 260000 quadratischen Patches bilden den eigentlichen Datensatz, auf dem die Klassifikatoren trainiert und evaluiert wurden (siehe Kapitel 5.1).

Der Gesamtdatensatz wurde in mehrere Teildatensätze aufgeteilt: Der Trainingsdatensatz wurde für das tatsächliche Training der Netzwerke genutzt. Um eine Optimierung des neuronalen Netzes ausschließlich auf diesen Trainingsdaten zu verhindern, wurde zusätzlich ein Validierungs- und ein Testdatensatz erstellt, welche ausschließlich Bildausschnitte von Bildern enthalten, die nicht im Training genutzt werden. Der Validierungsdatensatz dient der Hyperparameteroptimierung und der Modellauswahl. Der Testdatensatz dagegen wird ausschließlich für die abschließende Bewertung der Leistungsfähigkeit herangezogen. Tabelle 1 bietet einen Überblick über die Klassenbesetzung der Teildatensätze und zeigt Beispiel-Patches je Klasse. Die von dem Detektionssystem zu erkennenden Schadens- und Objektklassen besitzen im Straßennetz jeweils unterschiedliche Auftrittswahrscheinlichkeiten und führen zu einer ungleichen Repräsentation der Klassen im Datensatz. Um diesem Umstand entgegenzuwirken, wurden die Klassen, welche eine geringe optische Varianz aufweisen

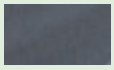
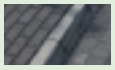
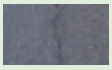
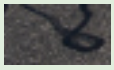
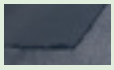
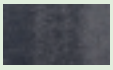
Code	OKAY	CURB	COV	MARK	SI_CR	AL_CR	SE_CR	POTHO	PATCH	BLEED
Klasse	Intakter Asphalt	Bordstein	Ablauf	Markierung	Einzelriss	Netzriss	Riss saniert	Ausbruch	Flickstelle	Bindemittelaustritt
Patch	81	111	0	0	0	0	0	0	0	192
Beispiel										
Train.	94660	15131	10044	9768	19913	9843	14139	4794	13869	7022
Valid.	9993	1535	1030	1001	2004	1024	1495	72	1521	729
Test	20029	2954	1946	2005	3999	1993	2289	399	2889	1284

Tabelle 1: Anzahl der Patches für die Teildatensätze für Training, Validierung und Test. Objektklassen sind blau und Schadklassen sind orange markiert.

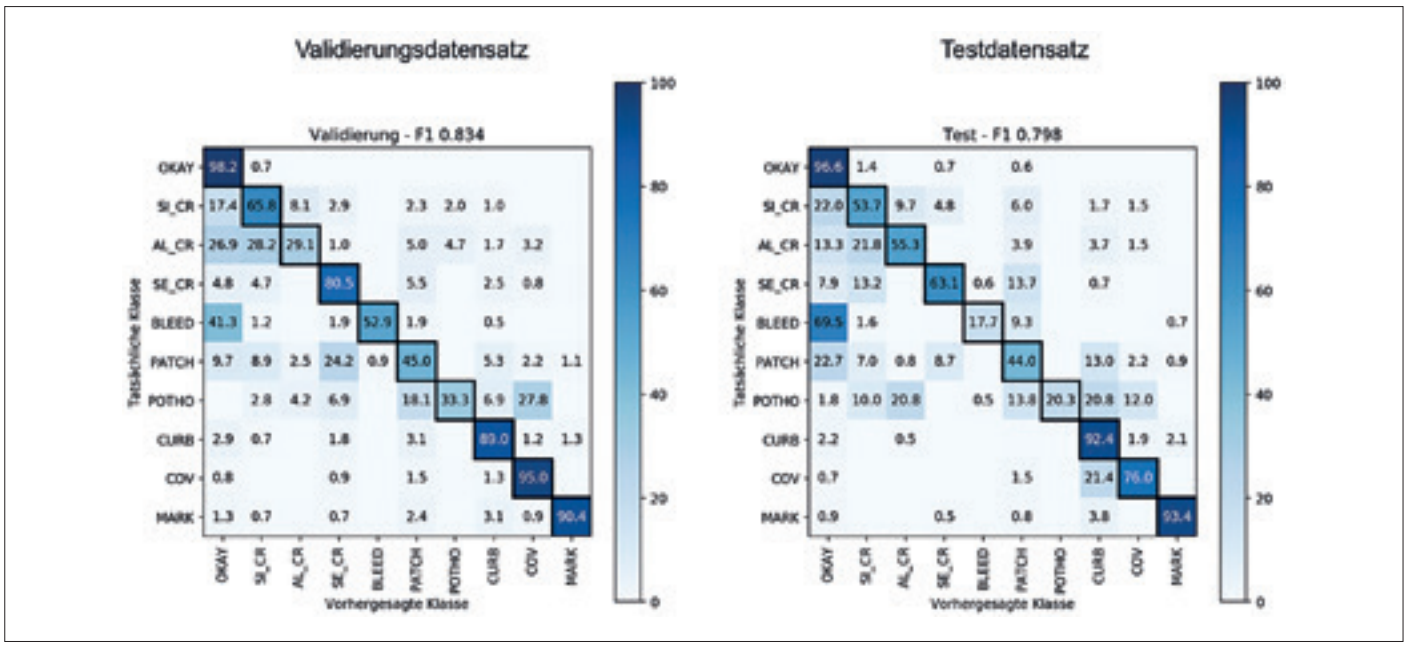


Abbildung 7: Konfusionsmatrizen für das beste grauwertbasierte Residual Network mit Angabe der Detektionsgenauigkeit in Prozent. Werte unter 0.5 Prozent sind aus Gründen der besseren Übersichtlichkeit ausgeblendet.

und sehr häufig in dem Ausgangsdatensatz enthalten waren, nur subgesampelt in die Teildatensätze übernommen (siehe Tabelle 1). Die dennoch vorhandene Ungleichheit der Klassenbesetzung wird durch die Verwendung des F1-Bewertungsmaßes in Verbindung mit Makro-Averaging bei der Auswertung berücksichtigt.

5 ERGEBNISSE

Für die Evaluation der Leistungsfähigkeit des vorgestellten Systems wird im Folgenden zuerst die Qualität der bildausschnittbasierten Schadstellen- und Objektdetektion und anschließend die Qualität der automatischen Schadstellenkartierung untersucht. Während sich Erstere auf den im vorherigen Kapitel beschriebenen Datensatz stützt, wurde für Letzteres eine circa 200 m lange Referenzstrecke aufgenommen.

5.1 QUALITÄT DER SCHADSTELLEN- UND OBJEKTDETEKTION

Das Training und die Anwendung der trainierten Modelle wurden mit dem Keras Framework (Chollet 2015) und Theano Backend (Theano Development Team 2016) realisiert. Für die Hyperparameteroptimierung wurden alle Trainingsprozesse mit unterschiedlicher Zufallsinitialisierung dreimal wiederholt durchgeführt, sodass auch Varianzen für die Parameterkonfigurationen bestimmt werden konnten (Tabelle 2). Neben unterschiedlichen Lernraten wurden auch ResNets mit unterschiedlichen Netzwerktiefen und unterschiedlich große Patchgrößen als Input getestet. Zur synthetischen Datenanreicherung der Trainingsdaten wurden die Bildausschnitte in dem Netzwerk gespiegelt, rotiert und leicht verschoben präsentiert (data augmentation). Jedes Netzwerk wurde für 200 Epochen trainiert, wobei die beste Epoche auf den Trainingsdaten nach spätestens 160 Epochen erreicht wurde. Auf einer NVIDIA GeForce GTX 1080Ti Grafikkarte benötigte das Training einer Epoche circa 90 Sekunden. Die Anwendung des trainierten Netzwerks auf den Befahrungsbildern des I.R.I.S-Systems (2560 x 1920 Pixel) benötigt hingegen auf der Grafikkarte lediglich 2.3 Sekunden.

In den Experimenten hat sich ein Residual Netzwerk mit 18 Faltungsschichten und eine Bildausschnittgröße von 128 x 128 Pixeln als guter Kompromiss zwischen Laufzeit und Genauigkeit herausgestellt (Abbildung 4). Kleinere Patchgrößen stellen für das Anwendungsproblem oft nicht genügend Kontextinformationen für das Netz zur Verfügung, während bei größeren Patchgrößen teilweise keine eindeutigen Klassenentscheidungen mehr getroffen werden können, da der Bildausschnitt dann mehrere Klassen beinhalten kann. Darüber hinaus ergab ein Vergleich von farb- und grauwertbildbasierten Netzwerken, dass die Verwendung der Farbinformation keinen wesentlichen Vorteil im Vergleich zur Verwendung von Graustufenbildern erbringt.

Insgesamt erreicht der auf Grauwertbildern trainierte Klassifikator auf dem ca. 40000 Patches umfassenden Testdatensatz eine Accuracy von 0.96 bzw. einen F1-Score von 0.80 (siehe Tabelle 2). Aufgrund des unbalancierten Datensatzes ist das F1-Maß dabei die adäquate Metrik, um die Leistungsfähigkeit des Klassifikators einzuschätzen. Die Konfusionsmatrizen in Abbildung 7 visualisieren die numerische Leistungsevaluierung auf den Bildausschnitten des Validierungs- und Testdatensatzes und schlüsseln die Erkennungsgenauigkeiten für die einzelnen Klassen im Detail auf.

Bewertungsmaß	Teildatensatz: Validierung	Teildatensatz: Test
Accuracy	0.968 (0.0004)	0.961 (0.0014)
F1	0.834 (0.0016)	0.798 (0.0075)

Tabelle 2: Darstellung der Leistungs- und Generalisierungsfähigkeit des mit Graustufenbildern trainierten Faltungsnetzwerks. Die Standardabweichung über die Trainingsläufe mit unterschiedlichen Initialisierungen ist in Klammern angegeben

Abbildung 8: Visualisierung von Eingabebildern (links) und den überlagerten Ergebnissen der automatischen Schadstellen- und Objektdetektion (rechts) für die unteren zwei Drittel des Kamerabilds. Das obere Bilddrittel wurde bei der automatischen Analyse ausgespart und ist deshalb hier nicht dargestellt. Der Klassifikator wurde nur in Bildbereichen angewendet, der vorher durch DeepLabv3+ als Fahrbahnoberfläche segmentiert wurde. Die Konturpolygone für offene Einzel- und Netzrisse sind rot, für versiegelte Risse orange, für Flickstellen gelb und für Markierungslinien grün eingefärbt. Während oben ein Beispiel dargestellt ist, bei dem sowohl die Detektion als auch die Unterscheidung der Schadensklasse sehr gut funktioniert, zeigt das untere Bildbeispiel einen Fall, bei dem falschpositive Flickstellen im Schatten unter einer Brücke erkannt werden.

Gute bis sehr gute Erkennungsraten sind für alle Objektklassen, also Bordsteine, Ablauf- und Schachtdeckel, Fahrbahnmarkierungen und intakte Fahrbahnbereiche, festzuhalten. Diese werden in 76 % bis knapp 97 % der Fälle korrekt klassifiziert. Für die Schadklassen ergibt sich ein anderes Bild: Die beste Detektionsleistung kann mit 63 % korrekten Entscheidungen für die Klasse der versiegelten Risse erzielt werden. Offene Einzelrisse werden in knapp 54 % der Fälle korrekt klassifiziert. Knapp 15 % der Fehlklassifizierungen entfallen auf Netzrisse und versiegelte Risse und bleiben somit zumindest innerhalb der Schadenskategorie Riss. Etwa 25 % wurden jedoch als intakter Asphalt oder eine andere Objektklasse fehlklassifiziert. Für die Klassen Ausbruch und Bindemittelaustritt ist die Erkennungsleistung auf einem sehr niedrigen Niveau. Dabei ist für die Klasse Ausbruch positiv zu bewerten, dass Verwechslungen zu mehr als 30 % innerhalb der Schadklassen auftreten, mit denen Ausbrüche häufig in Vergesellschaftung auftreten, wie z. B. Netzrissen. Die Klasse Bindemittelaustritt wird in knapp 70 % der Fälle als intakter Asphalt fehlklassifiziert und somit bei der Schadensanalyse potenziell übersehen.

Um einen Eindruck zu bekommen, welche Ergebnisse der trainierte Klassifikator bei der Anwendung auf einem gesamten Bild liefert, wurden Bildsequenzen automatisch analysiert (Abbildung 8). Zunächst wurde mithilfe von DeepLabv3+ die Fahrbahnoberfläche segmentiert und als Maske für die Anwendung des trainierten Klassifikators genutzt. Die visuell qualitative Inspektion der Klassifikationsergebnisse kommt zu folgender Einschätzung: Für Einzel- und Netzrisse lässt sich im Allgemeinen eine gute und wenig lückenhafte automatische Detektion feststellen. Es fällt jedoch auf, dass aufgrund des bildausschnittbasierten Klassifikationsansatzes die eigentlich linearen Rissstrukturen eher flächenhaft als betroffener Bildbereich detektiert werden. Für Flickstellen, Bindemittelaustritt und Ausbrüche überzeugen die Detektionsergebnisse nicht. Moderate bis gute Ergebnisse werden im Allgemeinen erzielt, solange Oberflächenbelag und Beleuchtungssituation ausreichend durch die Trainingsdaten abgedeckt sind. Bei stark abweichenden Situationen, wie beispielsweise tief stehendem Sonnenlicht, kommt es vermehrt zu Fehlklassifikationen. Die Generalisierungseigenschaften des trainierten Klassifikators sind demnach noch nicht optimal.

5.2 QUALITÄT DER AUTOMATISCHEN SCHADSTELLENKARTIERUNG

Für die nachstehend beschriebenen Untersuchungen wurde eine circa 200 m lange Teststrecke aufgenommen, die weder im Trainings-, Validierungs- noch Testdatensatz des Schadstellendetektors enthalten war. Im Vorfeld der Befahrung wurde eine Passpunktmarkierung auf dem Standstreifen aufgebracht und diese mittels RTK terrestrisch vermessen. Die absolute Lagequalität der Passpunktmessung betrug circa 2 cm. Im Anschluss erfolgte die Messung mit dem I.R.I.S-System bei einer Geschwindigkeit von circa 80 km/h, wobei Trajektorie, Kamerabilder und Laserscans erfasst wurden. Das Kamerabildmaterial wurde mit einem Auslöseintervall von 5 m aufgenommen, sodass eine Bildsequenz ein Objekt bzw. einen Schaden möglichst mehrfach aus verschiedenen Entfernungen und somit Perspektiven abbildet. Analog der in Kapitel 3.3 beschriebenen Vorgehensweise wurde je Kamerabild ein entsprechendes Tiefenbild erzeugt. Somit kann zu jedem beliebigen Kamerabild-Pixel ein Objektstrahl erzeugt und über die Entfernungsinformation eine 3D-Geokoordinate berechnet werden.

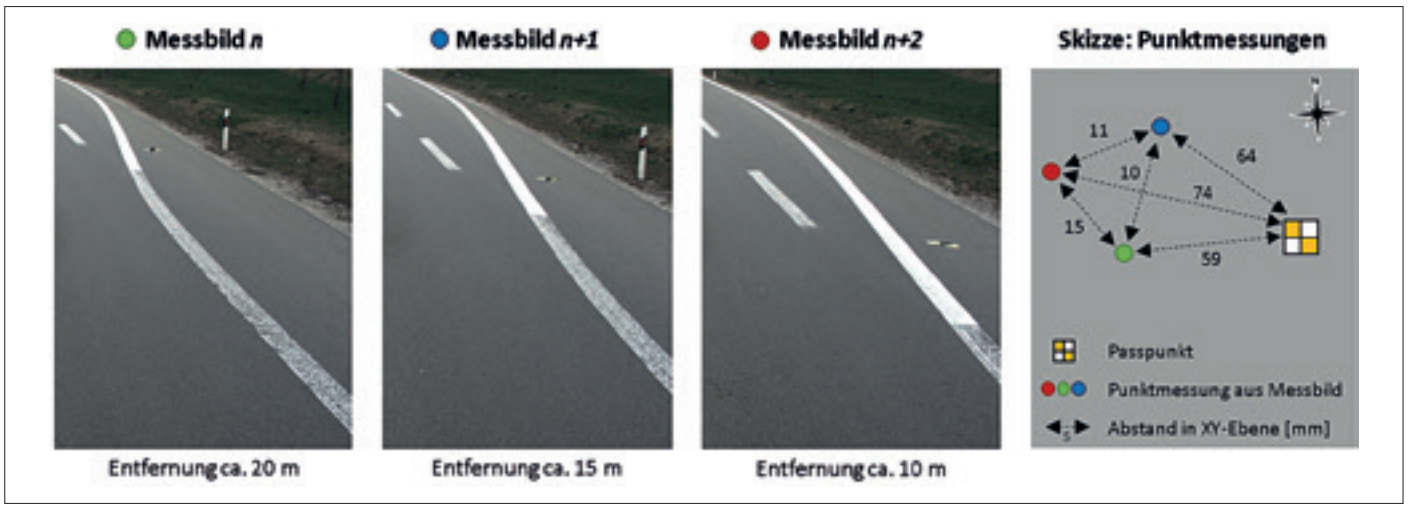


Abbildung 9: Evaluierung der Qualität der tiefenbildbasierten Georeferenzierung mittels Passpunkt. Die Abstände der drei Punktmessungen aus den einzelnen Messbildern zueinander ergeben eine relative Genauigkeit von besser 2 cm. Die absolute Genauigkeit (Abstände der Punktmessungen zu Passpunkt) ist besser als 10 cm. Die Skizze rechts entspricht hinsichtlich der Punktanordnung der tatsächlichen Situation, ist aber aus Darstellungsgründen nicht maßstabsgetreu.

Im ersten Schritt wurde die absolute Genauigkeit der mit dem I.R.I.S-System aufgenommenen Bilddaten in Bezug auf die Passpunktmessung evaluiert. Die Passpunktmarkierung war in drei aufeinanderfolgenden Bildern der in Fahrtrichtung ausgerichteten Umfeldkamera sichtbar und wurde aus jedem Bild manuell punktuell georeferenziert. Die Ergebnisse sind in Abbildung 9 dargestellt und zeigen, dass für die Punktmessung in einem Kamerabild eine absolute Genauigkeit im Bereich 6 cm bis 8 cm erreicht werden kann. Hierbei ist der maßgebliche Faktor die Qualität der Positionierungslösung. Die wiederholte Messung ein und desselben Objekts aus drei unterschiedlichen Entfernungen erbringt eine relative Genauigkeit von 1 cm bis 2 cm.

Der zweite Teil der Evaluation widmete sich der Frage, wie gut die vom neuronalen Netz erkannten Schäden als Geoobjekte kartiert werden können. Hierzu wurde für den Fahrbahnabschnitt eine automatische mit einer manuellen Schadstellenkartierung verglichen und statistisch ausgewertet. Dabei wurden nur Einzel- und Netzrisse als kombinierte Schadklasse Riss betrachtet, da sie erstens im Zuge der quantitativen Analyse hohe Erkennungsraten aufwiesen (Abbildung 7) und zweitens

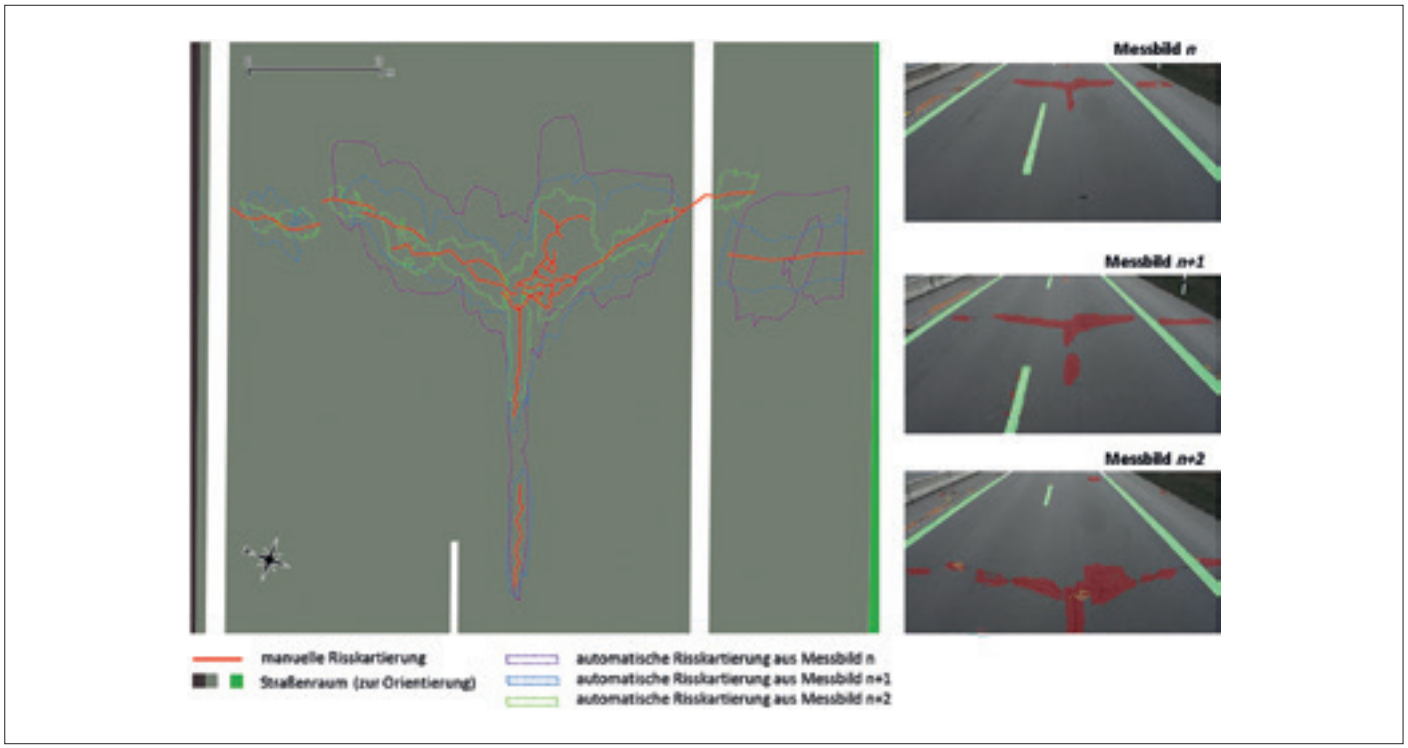


Abbildung 10: Ergebnis der automatischen Schadstellendetektion am Beispiel einer Bildsequenz, die im 5-m-Intervall aufgenommen wurde (rechts) sowie der Vergleich von manueller und automatischer Schadenskartierung, die jeweils mithilfe des Tiefenbilds realisiert wurde (links). Die manuelle Risskartierung wurde dabei linienhaft ausgeführt (rot). Die Ergebnisse der automatischen Risserkennung liegen dagegen als Polygon vor (violett, blau, hellgrün). Da der im Beispiel rechts rot dargestellte Riss in drei aufeinanderfolgenden Messbildern sichtbar ist, ergeben sich die links unterschiedlich eingefärbten Ergebnispolygone.

	Anzahl	Länge [m] Mittelwert	Länge [m] Standardabweichung	Länge [m] Minimum	Länge [m] Maximum
Richtig-Positiv	82	0.86	0.59	0.10	2.47
Falsch-Negativ	9	0.28	0.12	0.09	0.45
Falsch-Positiv	6	1.09	1.00	0.35	3.01

Tabelle 3: Statistische Auswertung der automatischen Risskartierung. Im Bereich der 200 m langen Teststrecke wurden 91 Rissobjekte manuell und 26 automatisch kartiert. Häufig beinhaltet ein automatisch kartiertes Risspolygon mehrere manuell kartierte Risslinien. Ergänzend sind statistische Kennwerte zur Geometrielänge angegeben. Da im Fall der Falsch-Positiv-Detektionen keine Riss-Linien als Referenz vorlagen, deren Länge hätte statistisch ausgewertet werden können, wurde die Maximaldistanz der Stützpunkte innerhalb des automatisch kartierten Risspolygons als Länge eingesetzt.

der deutsche Auswertestandard diese Rissausprägungen nicht getrennt bewertet (FGSV 2018). Zunächst wurde eine konventionelle, vollmanuelle Referenzkartierung unter Nutzung der Kamera- und Tiefenbilder vorgenommen. Die Bildfolgen wurden visuell inspiziert und beim Auftreten von Rissen die entsprechenden dominanten Tiefenlinien des Schadens nachvollzogen. Im Ergebnis lagen georeferenzierte Liniengeometrien vor. Anschließend wurde die automatische Fahrbahnextraktion und Schadstellendetektion auf die Kamerabilder angewendet sowie die Detektionsergebnisse (Ergebnismasken) in Riss-Konturen überführt. In beiden Fällen wurde dann die Entfernungsinformation des Tiefenbilds genutzt, um entweder für jeden manuell geklickten Stützpunkt der Riss-Linien oder für die Stützpunkte der automatisch erkannten Riss-Kontur globale 3D-Koordinaten zu berechnen.

Da die Kameraausrichtung sowie das Bildaufnahmeintervall so gewählt sind, dass sich Bildbereiche einer Aufnahmesequenz überlappen, wird auch ein und derselbe Schaden typischerweise in mehreren aufeinanderfolgenden Bildern erkannt. Für einen Schaden können demnach mehrere Geoobjekte in Form von Schadstellenpolygonen erzeugt werden (Abbildung 10). Die Überbestimmung, die aus der Mehrfachmessung von Schäden aus unterschiedlichen Entfernungen und Perspektiven resultiert, ermöglicht es, die Detektionssicherheit im Vergleich zur Einzelmessung noch zu erhöhen. In Abbildung 10 ist exemplarisch dargestellt, welche Ergebnisse die automatisierte Kartierung von Rissen aus einer Bildsequenz im Vergleich zu einer manuellen Risskartierung liefert. Im rechten Bereich der Abbildung ist ersichtlich, dass der von Rissen betroffene Fahrbahnbereich in den aufeinanderfolgenden Bildern aufgrund des unterschiedlichen Abstands zum Messfahrzeug und aufgrund der Aufnahmeperspektive an verschiedenen Bildpositionen unterschiedlich stark verzerrt abgebildet ist. Jedoch kann durch den im Beitrag beschriebenen tiefenbildbasierten Ansatz zur Georeferenzierung von Bildinhalten der Schaden immer an derselben Geoposition verortet werden – wenn auch mit unterschiedlich ausgeprägter Kontur. Die Detektionsergebnisse aus den verschiedenen Bildern der Sequenz ergänzen und bestätigen sich, sodass die kombinierte Gesamtgeometrie der automatisch erzeugten Polygone im Beispiel circa 85 % der manuell erfassten Risslänge abdeckt.

Für die statistische Auswertung der circa 200 m langen Referenzstrecke wurden 66 sich teilweise oder vollständig überlappende Riss-Polygone geometrisch miteinander vereinigt. Dadurch entstanden 26 Riss-Polygone, die mit den 91 manuell kartierten Riss-Linien abgeglichen wurden. Die statistische Auswertung der automatischen Risskartierung auf Geoobjekt-Ebene kommt zu folgendem Ergebnis (siehe Tabelle 3): Als richtigpositive Detektion wurden nur Fälle gezählt, bei denen mindestens

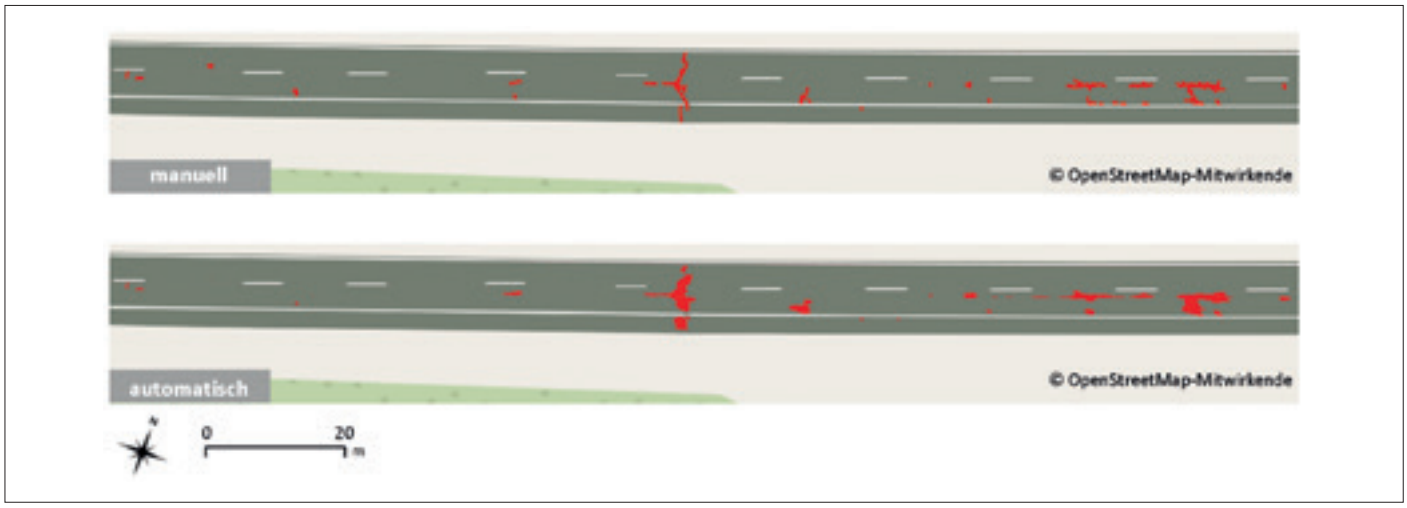


Abbildung 11: Vergleich von manueller (oben) und automatischer Risskartierung (unten) für die circa 200 m lange Teststrecke. Mit den Kamerabildern einer Messfahrt (Hauptfahrstreifen) konnten Risse auch auf dem angrenzenden Fahr- und Standstreifen erfasst werden. Beide Teilabbildungen sind mit OpenStreetMap-Basiskarten hinterlegt.

25 % der Risslänge von den automatisch kartierten Riss-Polygonen abgedeckt wurde. Demnach wurden 85 % der Risse korrekt kartiert, bei 6 % der Risse handelt es sich um Fehler erster Art (Falsch-Positiv) und bei 9 % um Fehler zweiter Art (Falsch-Negativ). Abbildung 11 stellt die manuellen und automatischen Kartierungsergebnisse grafisch gegenüber. Der T-förmig quer über die Fahrbahn verlaufende Riss aus Abbildung 10 befindet sich etwa in der Mitte der Referenzstrecke.

6 FAZIT UND AUSBLICK

Ziel des Beitrags war es zu untersuchen, ob Kamerabilder eines Mobile-Mapping-Systems, welche die Fahrbahnoberfläche aus einer Schrägansicht abbilden, mit einem bildausschnittbasierten CNN-Ansatz hinsichtlich Fahrbahnschäden automatisch klassifiziert werden können. Hierzu wurde ein umfangreicher Datensatz aufbereitet und ein tiefes neuronales Netz darauf trainiert und getestet. Es muss festgehalten werden, dass vor allem die High-Angle-Perspektive der Kamerabilder und die damit einhergehende Verzerrung und Skalierung von Schäden und Objekten im Bild die Limitierung des mit fixen Bildausschnittgrößen arbeitenden tiefen Faltungsnetzes im Anwendungskontext offenlegen. Gerade für großflächige Schadklassen, wie zum Beispiel Flickstellen, konnte kein befriedigendes Detektionsniveau erzielt werden. Für eher linear ausgeprägte Klassen, wie beispielsweise Risse oder Fahrbahnmarkierungen, hingegen liefert das trainierte Modell moderate bis gute Ergebnisse. Vollbildbasierte semantische Segmentierungsansätze könnten für die gegebene Problemstellung potenziell besser geeignet sein, da sie Kontextinformationen für die Klassifikation besser nutzen, als dies mit der festen Patchgröße bildausschnittbasierter Netze möglich ist.

Außerdem lag der Fokus des Beitrags auf der Verortung der im Bild erkannten Schäden als Geoobjekte. Hierzu lässt sich festhalten, dass die vorgestellte Sensordatenfusion ein präzises Messen von 3D-Koordinaten aus monokularen Bildaufnahmen ermöglicht. Mithilfe der aus der 3D-Bildverarbeitung resultierenden Tiefenbilder können die im Kamerabild erkannten Konturen von Fahrbahnschäden in georeferenzierte Polygon-Geometrien überführt werden. Der Abgleich mit einer terrestrischen Passpunktaufnahme ergibt eine relative Genauigkeit besser als 2 cm sowie eine absolute Lagegenauigkeit besser als 10 cm. Am Beispiel von Rissen wurde gezeigt, dass für Schad- und Objektklassen, die robust detektiert werden, mit dem vorgestellten Ansatz eine konturscharfe Kartierung als Geoobjekte automatisiert auf einem bereits hohen Niveau möglich ist, wobei eine bessere Segmentierung der Schäden im Bild auch eine genauere Kartierung zur Folge hat.

Weiterführende Arbeiten werden sich daher zum einen darauf konzentrieren, moderne Segmentierungsnetzwerke im Anwendungskontext zu testen und zum anderen die Datenbasis sukzessive zu erweitern. Letzteres betrifft vor allem die Schadklassen, die im aktuellen Datensatz noch unterrepräsentiert sind und daher auch keine stabilen Detektionsraten aufweisen.

Literatur

Applanix (2019): POS LV. https://www.applanix.com/downloads/products/specs/POS-LV-Datasheet.pdf, Zugriff 11/2019.

Bhoi, A. (2019): Monocular Depth Estimation: A Survey. https://arxiv.org/pdf/1901.09402.pdf, Zugriff 11/2019.

BMBF–Bundesministerium für Bildung und Forschung (2016): ASINVOS – Assistierendes und interaktiv lernfähiges Vi-deoinspektionssystem für Oberflächenstrukturen am Beispiel von Straßenbelägen und Rohrleitungen. Förderkennzeichen: 01IS15036.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. (2018): Encoder-Decoder with Atrous Seperable Convolution for Semantic Image Segmentation. https://arxiv.org/pdf/1802.02611.pdf, Zugriff 11/2019.

Chollet, F. (2015): Keras. https://keras.io, Zugriff 11/2019.

Cityscapes (2019): Detailed Results for 'DeepLabv3+' – Pixel-Level Semantic Labeling Task. https://www.cityscapes-dataset.com/detailed results/, Zugriff 11/2019.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. (2016): The Cityscapes Dataset for Semantic Urban Scene Understanding. https://www.cityscapes-dataset.com/wordpress/wp-content/papercite-data/pdf/cordts2016cityscapes.pdf, Zugriff 11/2019.

Eisenbach, M.; Stricker, R.; Seichter, D.; Amende K.; Debes, K.; Sesselmann, M.; Ebersbach, D.; Stöckert, U.; Gross, H.-M. (2017): How to Get Pavement Distress Detection Ready for Deep Learning? A Systematic Approach.In: Int. Joint Conf. on Neural Networks (IJCNN), Anchorage, USA, S. 2039-2047.

Everingham, M.; Eslami, S. M.; Gool, L.; Williams, C. K.; Winn, J.; Zisserman, A. (2015): The Pascal Visual Object Classes Challenge: A Retrospective. In: International Journal of Computer Vision, 111 (1), S. 98-136.

FGSV–Forschungsgesellschaft für Straßen- und Verkehrswesen (2018): ZTV ZEB-StB–Zusätzliche Technische Vertrags-bedingungen und Richtlinien zur Zustandserfassung und -bewertung von Straßen, Ausgabe 2006, geänderter und korrigierter Nachdruck 2018 unter Berücksichtigung des BMV ARS 6/2018. Köln.

Fraunhofer Institute for Physical Measurement Techniques IPM (2019): Clearance Profile Scanner CPS. https://www.ipm.fraunhofer.de/content/dam/ipm/en/PDFs/product-information/OF/MTS/Clearance-Profile-Scanner-CPS.pdf, Zugriff 11/2019.

Github (2019): DeepLab – Deep Labelling for Semantic Image Segmentation. https://github.com/tensorflow/models/tree/master/research/deeplab, Zugriff 11/2019.

He, K.; Zhang, X.; Ren, S.; Sun, J. (2016): Identity mappings in deep residual networks. In: European conference on computer vision. Springer, S. 630-645.

Karaca, Y.; Cattani, C.; Moonis, M. (2017): Comparison of Deep Learning and Support Vector Machine Learning for Sub-groups of Multiple Sclerosis. In: Computational Science and Its Applications – ICCSA 2017, July 3–6, 2017, Trieste, Italy, S. 142-153.

Landau, H.; Vollath, U.; Chen, X. (2002): Virtual Reference Station Systems. In: Journal of Global Positioning Systems, 1 (2), S. 137-143.

Liu, P.; Choo, K.-K. R.; Wang, L.; Huan, F. (2016): SVM or deep learning? A comparative study on remote sensing image classification. In: Soft Computing, 21(23), S. 7053-7065.

Luhmann, T. (2018): Nahbereichsphotogrammetrie – Grundlagen, Methoden, Beispiele. Wichmann, Berlin/Offenbach.

Seichter, D.; Eisenbach, M.; Stricker, R.; Gross, H.-M. (2018): How to Improve Deep Learning based Pavement Distress Detection while Minimizing Human Effort. In: IEEE Int. Conf. on Automation Science and Engineering (CASE), Munich, S. 63-68.

Simonyan, K.; Zisserman, A. (2015): Very deep convolutional networks for large-scale image recognition. https://arxiv.org/pdf/1409.1556.pdf, Zugriff 11/2019.

Stricker, R.; Eisenbach, M.; Sesselmann, M.; Debes, K.; Gross, H.-M. (2019): Improving Visual Road Condition Assessment by Extensive Experiments on the Extended GAPs Dataset. In: Int. Joint Conf. on Neural Networks (IJCNN), Budapest, Hungary, Paper N-20496.

Theano Development Team (2016): Theano: A Python framework for fast computation of mathematical expressions. http://arxiv.org/abs/1605.02688, Zugriff 11/2019.

Voinov, S. (2020): Deep Learning-based multiclass vessel detection from very high resolution optical satellite images. In: gis.Science, 1/2020, S. 10-17.