Abteilung Signalverarbeitung und Machine Learning (Fingscheidt)

1. Forschungsfelder der Abteilung

Die Abteilung Signalverarbeitung und Machine Learning arbeitet mit Methoden des Deep Learning in den Forschungsfeldern Speech, Vision und Predictive Maintenance.

Im Forschungsfeld „Speech“ erforschen wir Verfahren zur Störgeräuschreduktion, zur akustischen Echokompensation, In-Car-Kommunikation sowie Strategien zum Training neuronaler Netze zu allen vorgenannten Verfahren [FIN4], [FIN5]. Weitere Themen sind Beamforming sowie (neuronale) Nachfilter für höherqualitative, aber standardkonforme Sprach- und Audio*de*coder. Darüber hinaus befassen wir uns mit Emotionserkennung, automatischer Spracherkennung (auch audiovisuell) und Informationsfusion.

Im Forschungsfeld „Vision“ forschen wir im Wesentlichen an Perzeptionsmethoden für das autonome Fahren. Dazu gehören semantische Segmentierung, Tiefenschätzung, gelernte Bildkompression, Corner-Case-Detektion sowie Fragestellungen des Domain Transfers und adversarialer Angriffe auf neuronale Netze.

Im Forschungsfeld „Predictive Maintenance“ arbeiten wir an Methoden der akustischen Fahrzeugüberwachung (akustische Event-Erkennung) und an der Detektion von fehlerhaften Zuständen in der Bahntechnik.

Unsere Forschungsergebnisse können Anwendung finden im Kontext von Fahrzeug-, Office- und Consumer-Umgebungen, der Verbesserung von Smartphones, Hörgeräten/Cochlea-Implantaten, Freisprechsystemen, Überwachungs- bzw. Produktionstechnologien bis hin zum autonomen Fahren.

2. Projekte

Neu im Berichtszeitraum gestartet wurde das vom Bundesministerium für Wirtschaft und Technologie (BMWi) geförderte dreijährige Verbundprojekt SPEAKER unter der Federführung des Fraunhofer-Instituts für Intelligente Analyse- und Informationssysteme IAIS sowie des Fraunhofer-Instituts für Integrierte Schaltungen IIS, in dem eine nationale Plattform für Spracherkennungslösungen entwickelt werden soll. In einem Verbund von 17 Konsortialpartnern ist die TU Braunschweig einer von drei universitären Partnern. Das IfN steuert ein visuelles Frontend für eine audiovisuelle Spracherkennung bei.

Ebenfalls neu gestartet und ebenfalls vom BMWi gefördert wird das dreijährige Projekt KI DataTooling mit 17 Partnern. Das IfN erforscht in diesem Rahmen gelernte Codierverfahren zur effizienten Speicherung und Übertragung von Bild-Daten, sowie die Detektion von Corner Cases zur Selektion von Datenmaterial für das Training und die Absicherung von maschinell gelernten Perzeptionsmethoden im hochautomatisierten Fahren.

Schließlich wurden wir per Unterauftrag der Volkswagen Group Innovation in das dreijährige, ebenfalls vom BMWi geförderte Verbundprojekt KI DeltaLearning eingebunden. Vorrangiges Ziel des IfN-Beitrags wird die Erforschung von Methoden des Continual Learning für die Domänenanpassung sein, konkreter, des fortwährenden Lernens neuronaler Netze in fahrenden Fahrzeugen im Betrieb, ohne dass die bestehenden Fähigkeiten der Netze abhanden kommen. Weiterhin werden Methoden untersucht, die es erlauben auf synthetischen Daten Perzeptionsverfahren zu trainieren, welche dann auch auf realen Daten eine hohe Performanz erreichen.

Es gelang uns, das dritte Forschungsfeld der Predictive Maintenance mit einem kleinen Kooperationsvorhaben zusammen mit der Siemens AG, dem Deutschen Forschungszentrum für künstliche Intelligenz (DFKI) und der DB Systel GmbH aus dem mFund-Programm des Bundesministeriums für Verkehr und digitale Infrastruktur (BMVI) zu stärken. Ziel ist eine Machbarkeitsstudie für ein größeres Verbundvorhaben im Bereich der Predictive Maintenance in der Bahntechnik.

Die Forschungsarbeiten im Bereich der iterativen Turbo-Informationsfusion in der automatischen Spracherkennung im Rahmen einer DFG-Einzelförderung wurden in diesem Berichtsjahr fortgesetzt (DFG-Kennzeichen FI 1494/6-1). Im Mittelpunkt standen unser neuartiges tiefes rekurrentes Netzwerk zum Probability-in-Probability-out Sequence Enhancement sowie Verfahren im Bereich der Large-Vocabulary Continuous Speech Recognition (LVCSR).

Ebenfalls fortgesetzt wurde das gut zweijährige Kooperationsprojekt MindMarker zusammen mit der Firma Mind Intelligence UG, Berlin, gefördert vom Bundesministerium für Bildung und Forschung (BMBF). Ziel des Projekts ist es, ein System zur Detektion von Depressionen anhand der Stimme eines Anrufers zu konzipieren. Im Berichtszeitraum entwickelte das IfN ein Emotionserkennungssystem, das – als neuronales Netz – rekurrente Schichten mit Long Short-Term Memory (LSTM)-Zellen einschließt, um aus dem emotionsbehafteten Sprachsignal die gewünschten kontinuierlichen Dominanz-Valenz-Aktivitätswerte (DVA) abzuschätzen. Das IfN entwickelt unterstützend dazu derzeit ein Verfahren zur Störgeräuschunterdrückung, um das gesamte Emotionserkennungssystem in einer gestörten Umgebung robuster zu machen.

Das Projekt KI-Absicherung, gefördert vom BMWi, ging in diesem Berichtszeitraum in das zweite Jahr. Es wurden unter anderem maschinell gelernte Verfahren zur Detektion einer Domänen-Abweichung aus Kamerabildern entwickelt. Des Weiteren wurde ein Teacher-Student-(TS-)Framework entwickelt, welches mit dem Best Paper Award auf einem CVPR-Workshop ausgezeichnet wurde. Das IfN ist über einen Unterauftrag der Volkswagen Group Automation eingebunden.

Auch das mehrjährige Projekt zur Predictive Maintenance mittels akustischer Diagnose in Fahrzeugen ging in das zweite Jahr der Kooperation mit der IAV GmbH. Mit der Fokussierung auf die Erkennung akustischer Events sollen dabei fehlerhafte Zustände eines Fahrzeugs durch überwachte Lernmethoden basierend auf akustischen Sensoren (z.B. Freisprechmikrofonen) detektiert und klassifiziert werden. In Zukunft wird die Anomalie-Detektion im Mittelpunkt stehen.

Die vorjährige Studie zu Teacher-Student-(TS-)Frameworks als Maßnahme zur Erhöhung der Robustheit gegen adversariale Angriffe wird über das Projekt KI-Absicherung weiter fortgesetzt. Über eine Unterbeauftragung der Volkswagen Group Automation werden nun universelle adversariale Angriffe näher untersucht. Dabei wird ein Fokus auf semantische Segmentierung und Videodaten gesetzt.

Das mit der Volkswagen Group Innovation durchgeführte Kooperationsprojekt zu sog. generativen adversarialen Netzwerken (GANs) für gelernte Bildkompressionsverfahren im automatischen Fahren wurde erfolgreich abgeschlossen. Es konnten Methoden zur objektspezifisch fokussierten Codierung sowie der effizienten Kompression mittels Vektorquantisierung erforscht werden.

Ebenfalls zu einem Abschluss gebracht wurde in Kooperation mit der Volkswagen Group Innovation unser Projekt zu methodischen Ansätzen zur Corner-Case-Detektion. Dabei handelt es sich um Online- und Offline-Ansätze zum Auffinden von kurzen Bildsequenzen, bei denen kritische bzw. seltene Ereignisse in einem Verkehrsszenario zu sehen sind. Es entstanden Lösungen zu sehr unterschiedlichen Typen von Corner Cases.

Das zweijährige, durch die NBank des Landes Niedersachsen geförderte Gemeinschaftsprojekt mit der Firma Pan Acoustics GmbH aus Wolfenbüttel ist erfolgreich zu Ende gegangen. Nachdem im ersten Projektjahr maßgeblich adaptives Beamforming für das Mikrofonmodul im Fokus stand, lag der Fokus im zweiten Projektjahr auf niedriger Verzögerungszeit für das Gesamtsystem und einer Soft-Audio-Decodierung für die Funkstrecke.

Mit der R&D-Gruppe der Firma NXP Semiconductors, Product Line Voice and Audio Solutions, Belgien, mittlerweile Goodix Technology (Belgium) B.V., wurden im Berichtszeitraum Forschungsarbeiten zur Störgeräuschreduktion für Sprach-

signale weitergeführt. Der Fokus des Vorhabens im vergangenen Jahr lag auf der Erforschung einstufig aufgebauter Systeme basierend auf tiefen neuronalen Netzen mit Rekursion und durchgehenden Faltungen, sowie in einem zweiten Projekt auf der Reduktion der Parameterzahl und Rechenkomplexität der Netze.

3. Mitarbeiterinnen und Mitarbeiter der Abteilung

Das Forschungsfeld Vision ist aufgrund der Vielzahl KI-bezogener Projekte weiter auf Expansionskurs. Zu Frau Breitenstein und den Herren Bär, Termöhlen, Klingner und Löhdefink ist die iranische Gastwissenschaftlerin Atiye Sadat Hashemi im Bereich der Sicherheit von neuronalen Netzen neu dazu gestoßen (seit 01.01.2020). Im Forschungsfeld Speech arbeiten weiterhin die Herren Franzen, Lohrenz, Strake, Xu und Zhao. Neu an Bord ist hier Herr Renzheng Shi (seit 01.03.2020). Verlassen haben uns im Berichtszeitraum die Herren Elshamy (bis 31.12.2019) und Meyer (bis 31.03.2020). Im Forschungsfeld Predictive Maintenance sind weiterhin die Herren Baumann und Klingner aktiv, sowie neu seit 15.09.2020 assoziiert die Gastwissenschaftlerin Frau Rashmi Kaslikar, Mitarbeiterin der iTUBS mbH. Damit arbeiteten zum Ende des Berichtszeitraums in der Abteilung Signalverarbeitung und Machine Learning neben Prof. Fingscheidt und Frau Erichsen-Rua 12 Wissenschaftler*innen sowie zwei Gastwissenschaftlerinnen mit. Im weiteren Umfeld zählen noch vier externe Doktorand*innen bei der Volkswagen AG dazu, die die Forschung im Bereich Vision, Karten und Scene Understanding erweitern. Im Berichtszeitraum haben bei uns 17 Studierende eine Masterarbeit und 8 Studierende eine Bachelorarbeit bzw. Projektarbeit abgeschlossen. Weiterhin hat uns noch eine Vielzahl studentischer Hilfskräfte unterstützt.

4. Forschungsfeld ”Speech”

4.1 Störgeräuschreduktion

Im Bereich der Störgeräuschreduktion beschäftigen wir uns mit der Unterdrückung von unerwünschten Nebengeräuschen in Sprachsignalen, um eine verständliche und qualitativ hochwertige Telekommunikation zu ermöglichen oder beispielsweise auch eine Vorverarbeitung für eine robuste Spracherkennung zu liefern. Der Fokus der Arbeiten liegt hier inzwischen auf Methoden des maschinellen Lernens mit modernen neuronalen Netzen, wobei neue Netztopologien, Verlustfunktionen für das Training und Methoden zur Reduktion der Rechenkomplexität erforscht werden.

Herr Strake führte das Projekt zum Thema Störgeräuschreduktion für Sprachsignale in Kooperation mit der R&D-Gruppe der Firma Goodix Technology (Bel-

Team	Realtime?	MOS (Rang)
Amazon	NRT	3.52 (1.)
North Western Polytechnical University, China	RT	3.42 (1.)
Amazon	RT	3.39 (2.)
TU Braunschweig (IfN) und Goodix Technology	NRT	3.38 (2.)
North Western Polytechnical University, China	NRT	3.38 (2.)
TU Braunschweig (IfN) und Goodix Technology	RT	3.36 (3.)
Sony und Carnegie Mellon University	NRT	3.34 (3.)
Ohne Störgeräuschunterdrückung	-	2.85

Tabelle 1: Ergebnisse des finalen subjektiven Qualitätstests unter den sieben Finalisten der Interspeech 2020 Deep Noise Suppression Challenge, angegeben als Mean Opinion Score (MOS) für den Realtime-Track (RT) und den Non-Realtime-Track (NRT) der Challenge.

gium) BV, ehemals NXP Semiconductors, Product Line Voice and Audio Solutions, im nunmehr insgesamt sechsten Projektjahr fort. Es wurde eine Fully Convolutional Recurrent Neural Network (FCRN) genannte Netztopologie entwickelt, welche die Modellierung zeitlicher Zusammenhänge ermöglicht und zudem durchgehend Faltungen als gelernte Transformationen verwendet. In einem Konferenzpaper [STR/FIN1] konnte gezeigt werden, dass sich diese Netztopologie gerade für die Störgeräuschunterdrückung in Sprache als vorteilhaft gegenüber oft verwendeten vollverbundenen Netzwerkschichten erweisen. Ein solches FCRN wurde in einer auf gemeinsame Reduktion von Störgeräuschen und Nachhall angepassten Version [STR/FIN2] für eine Teilnahme an der von Microsoft ausgerichteten Interspeech 2020 Deep Noise Suppression (DNS) Challenge genutzt. Die DNS Challenge zielte auf eine möglichst vergleichbare Evaluation von netzbasierten Systemen unter realistischen Anwendungsbedingungen ab und nutzte dafür umfangreiche Sprachqualitätstests mit Testhörern, die weltweit über Crowd Sourcing gewonnen wurden. Die Ergebnisse der Evaluation unter den sieben Finalisten (bei insgesamt 28 Einreichungen) ist in **Tabelle 1** gezeigt, wobei zwei verschiedene Komplexitätsstufen, zum einen für unter festen Hardwarevorgaben echtzeitfähige (Realtime-Track, RT), zum anderen für nicht-echtzeitfähige Systeme (Non-Realtime-Track, NRT) unterschieden wurden. Das von uns eingereichte System konnte sich den dritten Rang im RT und den zweiten Rang im NRT sichern. Außerhalb des Projekts wurden im Bereich der Forschung an Trainingsmethoden für neuronale Netze in einer Masterarbeit [Ma 20/005] Ansätze zur Nutzung von Generative Adversarial Networks (GANs) im Bereich der Störgeräuschunterdrückung untersucht.

Mit der Anwendung tiefer neuronaler Netze mit Rekursion und durchgehenden Faltungen ist auch die Anzahl der Parameter und vor allem die Rechenkomplexität der Netze gestiegen. Die hohe Zahl an Faltungsschichten fordert eine ent-

sprechend hohe Zahl an Multiplikationen, was – je nach Leistungsfähigkeit des Endgeräts – die Inferenz in Echtzeit erschwert oder gar unmöglich macht. Vielversprechende Ergebnisse zur Problemlösung wurden in einem zweiten Projekt mit Goodix erzielt: Um der Komplexitätsproblematik entgegenzuwirken, wurden Strukturen und Techniken untersucht und entwickelt, um die Anzahl an Multiplikationen zu reduzieren und gleichzeitig die Performanz des Systems zu erhalten.

Für das maskenbasierte neuronale Netz zur Sprachverbesserung in der Zeit-Frequenz-Domäne wird im Training eine Komponenten-Kostenfunktion (sog. components loss, CL) eingeführt. Dieser neue CL bietet separat Kontrolle über die Unterdrückung der Störgeräuschkomponente und den Erhalt der Sprachkomponente. Das Potenzial des vorgestellten CL für die Sprachverbesserung wird in [XU/FIN1] veranschaulicht. Wir erreichen eine Verbesserung in (fast) allen verwendeten instrumentellen Qualitätsmetriken gegenüber einem Training mit konventionellen Kostenfunktionen, wie dem Mean Squared Error (MSE) und der Perceptual Evaluation of Speech Quality (PESQ)-Kostenfunktion. Inspiriert durch die Resultate mit dem CL haben wir die Kostenfunktion weiter angepasst, um ein Netzwerk zu trainieren, das gleichzeitig Störgeräuschunterdrückung und Enthallung in einer Echtzeit-Sprachverbesserung ausführen kann. Diese Arbeit wird bei der ICASSP 2021 Deep Noise Suppression (DNS) Challenge eingereicht.

4.2 Echokompensation / In-Car-Kommunikation

Akustische Echo- und Feedback-Kompensation (engl. acoustic echo/feedback cancellation, AEC/AFC) erforschen wir im Bereich von Freisprechsystemen für Telefonie und von In-Car-Kommunikationssystemen (ICC-Systeme) zur Verbesserung der Kommunikation zwischen Passagieren innerhalb einer Fahrzeugkabine. In beiden Anwendungsfällen wird die akustische Rückkopplung von Lautsprechern in das Mikrofon geschätzt und unterdrückt.

Ein Ansatz, den wir hierfür nutzen, ist das frequenzbasierte Kalman-Filter (engl. frequency domain adaptive Kalman filter, FDAKF). In diesem Kontext hat Herr Franzen den Einsatz eines ein-, zwei- und sogar vierkanaligen FDAKF für ICC-Systeme statt eines üblicherweise verwendeten einkanaligen Algorithmus untersucht und diese miteinander verglichen. Die Vorteile der Varianten haben sich in verschiedenen akustischen Szenarien, darunter auch das zusätzliche Abspielen von Musik über die Fahrzeug-Lautsprecher, gezeigt. Die Ergebnisse wurden als Buchkapitel veröffentlicht [FRA/FIN1].

Neben der Anwendung eines adaptiven Filters zur Unterdrückung der Echo- bzw. Feedbackkomponente untersuchen wir als neuen Ansatz auch die Anwendung tiefer neuronaler Netze zur Echokompensation. Hierbei werden rekursive Netze mit durchgehenden Faltungen genutzt, welche die Aufgabe der Echokom-

pensation erlernen sollen: Basierend auf ausreichend vielen Trainingsdaten findet das Netz selbstständig einen Weg, die Echokomponente im Mikrofonsignal wiederzuerkennen und zu entfernen. In der [Ma 20/008] wurden hierzu diverse (strukturelle) Ideen erarbeitet und untersucht, welche sich letztlich auch in einer Einreichung für die ICASSP 2021 wiederfinden werden.

4.3 Sprach(de)codierung

Das NBank-Förderprojekt mit der Firma Pan Acoustics GmbH wurde erfolgreich abgeschlossen. Im Berichtszeitraum arbeitete Herr Zhao an der Entwicklung einer intelligenten Soft-Audio-Decodierung für eine Funkstrecke. Die Hardware einschließlich des rahmenweise arbeitenden Senders und Empfängers wurde von Pan Acoustics geliefert. Unter realistischen Übertragungsbedingungen diente sie als Plattform für Experimente zur sogenannten Soft-Audio-Decodierung am Empfänger. Da Übertragungsstörungen bei einer 2,4 GHz ISM-Band Funkstrecke in praktischen Anwendungen unvermeidlich sind, treten regelmäßig fehlerbehaftete Rahmen auf und einzelne Rahmen gehen sogar komplett verloren. Die Soft-Audio-Decodierung erzielt hier mithilfe eines Algorithmus zur Schätzung der Log-Likelihood Ratios (LLRs) auf Grundlage der Empfangspegelwerte ein deutlich verbessertes decodiertes Signal am Empfänger. Neben den fehlerbehafteten Rahmen werden in der Praxis gänzlich fehlende Rahmen zu einem ernsten Problem. Diese wären als plötzlicher Abbruch des Signals hörbar, insbesondere wenn mehrere aufeinander folgende Rahmen fehlen. Um solch wahrnehmbare Diskontinuitäten zu reduzieren, wurden zusätzliche Glättungsalgorithmen entwickelt. Eine große Anzahl von Experimenten wurde durchgeführt, um die praktischen Übertragungsbedingungen genau zu analysieren und den Algorithmus unter verschiedenen Umständen zu untersuchen. Hierzu gehörten zum Beispiel auch der Betrieb bei verschiedenen Störquellen wie z.B. Bluetooth, WLAN oder Mikrowellen.

4.4 Automatische Spracherkennung

Das DFG-Forschungsvorhaben für die Turbo-Fusion in der automatischen Spracherkennung hat sich im Berichtszeitraum auf die Entwicklung sogenannter Posterior-In-Posterior-Out (PIPO-)Netze konzentriert, mit dem Ziel, den bisher genutzten Turbo-forward-backward-Algorithmus (Turbo-FBA) durch ein bidirektionales long short-term memory (BLSTM) zu ersetzen.

In ersten Untersuchungen zu dieser PIPO-BLSTM-Netzstruktur haben sich durch die klar definierte Posterior-Input-Schnittstelle besondere Eigenschaften herausgestellt. Die besten Erkennungsergebnisse werden erreicht, wenn PIPO-BLSTMs auf Posterior-Vektoren trainiert werden, die aus einem akustischen Modell *ohne* zeitlichen Eingangskontext stammen, im Test aber dann mit einem

separaten akustischen Modell *mit großem* zeitlichen Eingangskontext evaluiert werden. Unerwarteterweise führt hier der größte Mismatch zu den besten Ergebnissen, da die PIPO-BLSTMs von den unschärferen Posterior-Verteilungen im kontextfreien Training profitieren. Diese Experimente wurden zu Beginn des Berichtszeitraums als Konferenzbeitrag auf dem ASRU Workshop in Singapur vorgestellt [LOH/STR/FIN1]. Wie geplant, konnten im Anschluss die PIPO-BLSTMs anstelle des Turbo-FBA in die Turbo-Informationsfusion integriert werden. In einem Fusionsszenario, in dem Filterbank-Merkmale fusioniert werden, die auf unterschiedlichen Fensterlängen in der spektralen Analyse basieren, erreicht die Turbo-Informationsfusion die besten Erkennungsergebnisse unter allen untersuchten Fusionsverfahren auf der bekannten TIMIT-Datenbank. Ein neuartiges Verfahren, das die Posterior-Trainingsdaten für die PIPO-BLSTMs durch Akquise während des iterativen Turbo-Prozesses vervielfacht, hat sich dabei als besonders effektiv herausgestellt. Nach Vorarbeiten in der Masterarbeit [Ma 20/015] wurden diese Ergebnisse als Konferenzbeitrag auf der INTERSPEECH vorgestellt [LOH/FIN1]. Weitere Arbeiten konzentrieren sich auf die Portierung der Turbo-Informationsfusion auf Spracherkennungsanwendungen mit großem Vokabular.

Im Rahmen des BMWI-Verbundprojekts "SPEAKER" zur audiovisuellen Spracherkennung wurde zunächst ein akustischer Spracherkenner für deutsche Sprache und mit großem Vokabular in einer Masterarbeit [Ma 20/004] implementiert. Auf dieser Grundlage wird nun ein visuelles Frontend entwickelt, welches durch automatisches Lippenlesen ein solches Spracherkennungssystem in sehr schwierigen akustischen Umgebungen unterstützen kann. Ein erstes visuelles Frontend – basierend auf neuronalen Faltungsnetzen – ist in der Masterarbeit [Ma 20/016] entstanden. Durch die Bereitstellung der audiovisuellen Datenbank SmartKOM der LMU München und eine klare Schnittstellendefinition des Moduls zur automatischen Spracherkennung der o.g. Fraunhofer-Institute sind alle Voraussetzungen für die Entwicklung eines visuellen Frontends geschaffen.

4.5 Emotionserkennung und Sprecher-Interferenzreduktion

Das BMBF-Projekt "Mind Marker", welches durch die Herren Xu und Meyer betreut wird, befindet sich mittlerweile im 2. Projektjahr. In Kooperation mit der Firma Mind Intelligence aus Berlin wird ein System zur Depressions-Erkennung auf Basis der Stimme einer/s Telefonanrufenden entwickelt, die/der durch einen künstlichen Dialog mit einem Sprachdialogsystem psychologisch motivierte Fragen gestellt bekommt. Das IfN hat im ersten Projektjahr ein System zur Emotionserkennung entwickelt, um die kontinuierlichen Dominanz-Valenz-Aktivitäts-Werte (DVA) einer Testperson während eines Gesprächs zu schätzen. Anhand der erkannten Emotionen soll später Einfluss auf den künstlichen Dialog genommen werden, um die Auswahl der psychologischen Fragen an die Stimmung

der Testperson anzupassen. Für die Umsetzung wurde neben einem Modell zur Emotionserkennung auch ein Modul zur Schätzung der Sprachaktivität implementiert. Hiermit wird der kontinuierliche Audiodatenstrom in Sätze segmentiert, da die sprachbasierte Emotionserkennung typischerweise auf Satzbasis arbeitet.

Eingangssignale können je nach Umgebung der Testperson zum Teil mit starken Störgeräuschen behaftet sein. Daher entwickelt das IfN derzeit einen Algorithmus zur Rauschunterdrückung, um eine robuste Erkennung der Emotionen zu gewährleisten. Für Deep-Learning-basierte neuronale Netze zur Sprachverbesserung wird im Training eine Komponenten-Kostenfunktion (sog. components loss, CL) eingeführt. Das Potenzial des vorgestellten CL für die Sprachverbesserung wird in [XU/FIN1] veranschaulicht.

Herr Meyer, der das IfN Ende März 2020 verlassen hat, hat sich während seiner Zeit am Institut mit der automatischen Analyse von Meetings auf Basis von Sprachsignalen beschäftigt. Ziel dabei war es, aufgenommene Sprachsignale aus einem Mehrpersonen-Meeting zunächst von Störungen zu befreien, um anschließend eine automatische sprachbasierte Emotionserkennung zu ermöglichen.

Herrn Meyers Forschung wurde von einer Kooperation mit der Abteilung Arbeits-, Organisations- und Sozial-Psychologie (AOS) des Instituts für Psychologie der TU Braunschweig begleitet, in welcher psychologische Aspekte von Meetings erforscht werden. Gemeinsam wurde ein Analysetool namens „Group Interaction and Annotation Tool (GiANT)“ entwickelt, welches die aufwändige psychologische Analyse von Meetings vereinfacht. GiANT beinhaltet unter anderem einen am IfN entwickelten Algorithmus zur automatischen Sprecherdetektierung und ist für Forschungszwecke auf der Plattform GitHub frei verfügbar. In einem gemeinsamen Journal-Artikel [FIN1] in der Zeitschrift Gruppe.Interaktion.Organisation wurde das entwickelte Tool im Januar 2020 vorgestellt.

Für die automatische Analyse der Meetings werden die Sprachsignale der einzelnen Personen mit Headsets in einem separaten Kanal aufgenommen, um eine optimale Sprachqualität zu erhalten. Dennoch koppeln die akustischen Äußerungen eines Sprechers nicht nur in sein eigenes Mikrofon, sondern auch in die Mikrofone aller anderen Personen mit einem gewissen Pegel ein. Diese Überlagerung von Sprachsignalen in einem Kanal muss durch eine Sprecher-Interferenzreduktion eliminiert werden, um eine automatische Emotionserkennung zu ermöglichen. Hierzu wurde ein bestehendes Wiener-Filter-Verfahren durch die Integration eines Kalman-Filters verbessert, welches zur Schätzung der Raum-Impulsantworten zwischen den störenden Sprechern und einem betrachteten Mikrofon genutzt wird. Mit Hilfe dieses Verfahrens können die Interferenzen um bis zu 12,5 dB reduziert werden, während die gewünschte Sprach-

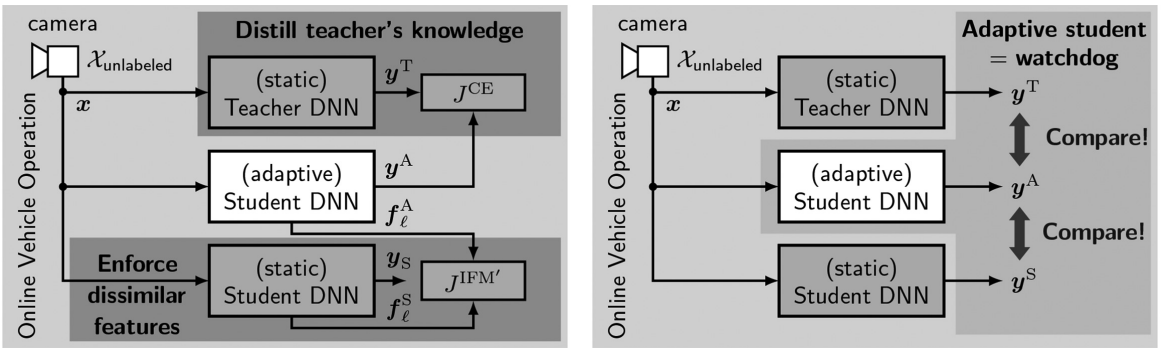
komponente des betrachteten Sprechers unberührt bleibt. Das Verfahren wurde in einem Journal-Artikel [FIN2] in der Zeitschrift EURASIP Journal on Audio, Speech, and Music Processing im September 2020 publiziert. Vertiefende Untersuchungen zu dem entwickelten Verfahren wurden außerdem im Rahmen von zwei internationalen Konferenzen, der diesjährigen ICASSP in Barcelona [FIN3] sowie der EUSIPCO in Amsterdam [FRA/FIN2], vorgestellt.

Neben der Forschung zur Sprecher-Interferenzreduktion in Meetings hat sich Herr Meyer während seiner Zeit am IfN vor allem mit der automatischen sprachbasierten Emotionserkennung befasst und dabei zwei Förderprojekte bearbeitet. Obwohl das Forschungsfeld der Emotionserkennung seit vielen Jahren existiert, gibt es noch immer grundlegende Probleme bezüglich einer einheitlichen Definition von Emotionen sowie der Akquise großer Datenbanken, die reale Emotionen enthalten. In diesem Kontext wurde während der Forschungsarbeiten entdeckt, dass die sehr bekannte und viel genutzte eNTERFACE-Datenbank zwar emotionale Sprache enthält, aufgrund ihrer Struktur jedoch nicht dazu geeignet ist, einem Klassifikationsmodell das Erkennen von Emotionen zu vermitteln. Im Gegenteil: Es stellte sich bei der Analyse heraus, dass verschiedene Klassifikationsmodelle keine Emotionen, sondern stattdessen den linguistischen Inhalt der Sprachäußerungen auswendig gelernt hatten, um so ein gutes Erkennungsergebnis zu erzielen. Darüber hinaus konnte auf Basis eines neuartigen Verfahrens ein neuronales Netzwerk entwickelt werden, dass auf einer renommierten Datenbank eines der besten Ergebnisse im internationalen Vergleich erzielt. Das Besondere an diesem Verfahren ist, dass die Erkennung der Emotionen auf einem zweidimensionalen Log-Mel-Spektrogramm basiert und somit im Grunde einem Bildklassifikator gleichkommt. Dies ermöglicht die Verwendung vieler neuer Strukturen im Bereich der Emotionserkennung, die bereits im Feld der Bilderkennung eingesetzt werden. Ein Tagungspaper mit den Ergebnissen des entwickelten Modells wurde bereits zu einer internationalen Konferenz eingereicht.

5. Forschungsfeld ”Vision”

5.1 Robuste semantische Segmentierung

Im Rahmen des Förderprojekts KI-Absicherung untersucht Herr Bär Methoden und Techniken, mittels derer man sich gegen adversariale Angriffe auf neuronale Netzwerke verteidigen kann. Adversariale Angriffe sind bösartige Manipulationen des Eingangssignals, die die Operation der Netze mit teils fataler Wirkung verändern. Hierzu wurde ein Verfahren [BAE/KLI/FIN1], basierend auf Teacher-Student-Netzwerken, entwickelt, welches auf dem diesjährigen Safe Artificial Intelligence for Autonomous Driving(SAIAD)-Workshop im Rahmen der IEEE Conference on Computer Vision and Pattern Recognition (CVPR) präsen-



(a) Adaptionsschritt (b) Inferenzschritt

Abbildung 5: Überblick über das Zusammenspiel der drei Netzwerke mit ihren Kostenfunktionen im Adaptionsschritt (linkes Bild) und im Inferenzschritt (rechtes Bild)

tiert und mit dem Best Paper Award ausgezeichnet wurde. Ein Artikel zum Best Paper Award, verfasst von Herrn Laurenz Kötter von der Pressestelle der TU Braunschweig, findet sich als Sonderbericht auf Seite 118.

Die **Abbildung 5** zeigt eine Übersicht über das entwickelte Verfahren. Im Fahrzeug wird der adaptive Student mithilfe des statischen Teachers und des statischen Studenten adaptiert (**Abbildung 5** (a)). So extrahiert der adaptive Student Wissen vom Teacher und wird gleichzeitig dazu gezwungen, seine Merkmale vom statischen Studenten zu entkoppeln. Anschließend wird der adaptive Student als eine Art Überwachungsnetzwerk verwendet, welches die Ausgänge der statischen Netzwerke überwacht, indem es diese auf Gleichheit prüft (**Abbildung 5** (b)). **Tabelle 2** fasst die wichtigsten Ergebnisse zusammen und zeigt, dass der adaptive Student fast immer in der Lage ist, unter den statischen Netzwerken dasjenige herauszupicken, welches gerade nicht angegriffen wird: Es handelt sich dabei um eine Korrektur durch Mehrheitsentscheid. Es ist geplant, das bisherige Setup auf zwei Netzwerke zu reduzieren, d.h., das Framework besteht künftig nur noch aus einem statischen Teacher und einem adaptiven Studenten.

Des Weiteren ist Herr Bär Projektleiter seitens des IfN im Projekt KI-Absicherung. Seit diesem Jahr betreut er das Ergebnis-Cluster “Adversarial Attacks and Teacher-Student Framework”, in dem ein Austausch mit den vielen Projektpartnern zum Thema adversariale Angriffe und Teacher-Student-Netzwerke erfolgt. Außerdem stehen seit diesem Jahr Projektdaten zur Verfügung, die synthetisch generiert wurden. Diese Daten wurden dazu verwendet, um u.a. die entwickelten Verfahren von Herrn Bär und Herrn Löhdefink in verschiedenen Szenarien zu testen. Der entwickelte Code im Projekt KI-Absicherung wird unter allen Partnern geteilt.

	Anzahl der Fälle / mIoU				
	$\epsilon = 0$	$\epsilon = 1$		$\epsilon = 10$	
model	clean	T-AE	S-AE	T-AE	S-AE
T	**440** / **75.77**	0 / 16.89	**441** / **72.12**	0 / 1.84	**441** / **54.02**
S	1 / 64.55	**441** / **58.39**	0 / 9.14	**441** / **41.21**	0 / 0.87

Tabelle 2: Anzahl der Fälle, in denen die semantische Segmentierung des adaptiven Studenten ähnlicher zum statischen Teacher (T) oder zum statischen Studenten (S) ist. Zusätzlich ist die mean intersection-over-union (mIoU) angegeben. Es wurden unveränderte Bilder ($\epsilon = 0$), sowie leichte ($\epsilon = 1$) und starke ($\epsilon = 10$) Angriffe untersucht. Die Angriffe sind jeweils für den statischen Studenten (S-AE) und den statischen Teacher (T-AE) berechnet worden. In nahezu allen Fällen erkennt der adaptive Student das derzeit besser funktionierende Netzwerk (fettgedruckt) ohne Kenntnis der Grundwahrheit (ground truth).

Neben der Projektarbeit in KI-Absicherung gibt es eine weitere Unterbeauftragung der Volkswagen Group Automation, die von Herrn Bär bearbeitet wird. Hierbei geht es speziell um universelle adversariale Angriffe (UAA). U.a. wurde über eine studentische Arbeit [Ma 20/023] untersucht, inwieweit die Integration von ConvLSTMs (LSTM-Netze durchgehend als Faltungsnetze ausgelegt, engl. convolutional neural network) in einen UAA-Generator einen Mehrwert für Angriffe auf Videodaten erzeugt. Des Weiteren entstand über die Finanzierung der Volkswagen Group Automation ein Übersichts-Artikel zu adversarialen Angriffen, welcher beim Signal Processing Magazine bereits letztes Jahr eingereicht wurde und im März diesen Jahres zur Veröffentlichung (Anfang 2021) akzeptiert wurde.

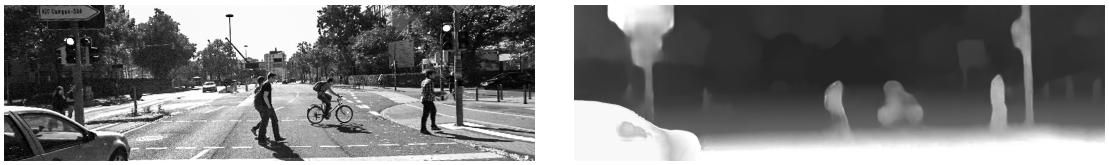
Aus der Unterbeauftragung seitens der Volkswagen Group Automation kam eine Zusammenarbeit von Herrn Bär mit dem externen Doktoranden Herrn Nikhil Kapoor zustande, wobei Herr Bär vorwiegend eine beratende Funktion einnimmt. Herr Kapoor erforscht die Eigenschaften von adversarialen Angriffen, die sich im Frequenzbereich zeigen. Dabei lässt sich beobachten, dass aus einem adversarialen Angriff entstandene manipulierte Eingangs-Bilder in ein neuronales Netz klar messbare Unterschiede zu ihrem unverrauschten Gegenpart im Frequenzbereich aufweisen. Basierend auf dieser Beobachtung wurde ein Wiener-Filter in seiner Funktion dahingehend weiterentwickelt, dass es nicht nur erfolgreich als Verteidigung gegen adversariale Angriffe verwendet werden kann, sondern auch bestehende verwandte Methoden des Stands der Technik in der Effektivität übertrifft. Beide Erkenntnisse tragen zum allgemeinen Verständnis von adversarialen Angriffen und der Natur von tiefen neuronalen Netzwerken bei. Eine entsprechende wissenschaftliche Veröffentlichung wurde auf der Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence eingereicht.

Anfang 2020 kam die Gastwissenschaftlerin Frau Atiye Hashemi aus dem Iran an das Institut für Nachrichtentechnik und betreibt bis Ende Januar 2021 mit Herrn Bär Grundlagenforschung im Bereich adversarialer Angriffe. In einer ersten wissenschaftlichen Zusammenarbeit wird dabei untersucht, inwieweit die Ähnlichkeit der Merkmalskarten neuronaler Netzwerke in den ersten Schichten dazu verwendet werden kann, die Transferierbarkeit von universellen Angriffen auf andere Netzwerktopologien zu erhöhen. Eine entsprechende wissenschaftliche Veröffentlichung wurde zu einer internationalen Tagung eingereicht. Eine weiterführende Arbeit, die sich auf die Erweiterung der Experimente mit semantischer Segmentierung fokussiert, wurde als Journalartikel in den IEEE Transactions on Intelligent Transportation Systems (T-ITS) eingereicht. Zusätzlich finden Untersuchungen zur Detektion und Korrektur von besonders stark von einem adversarialen Angriff betroffenen Pixeln statt.

Typischerweise wird eine semantische Segmentierung auf Einzelbildern trainiert und inferiert. Geht man nun von einer Abfolge von Einzelbildern mit zeitlichem Kontext aus (Videosequenz), so ist es erwünscht, dass die semantische Segmentierung dieser Abfolge trotz des Inferierens einzelner Bilder in sich zeitlich konsistent bleibt. Eine entsprechende Metrik entwickelte Herr Serin Varghese, externer Doktorand bei der Volkswagen Group Automation, u.a. in Zusammenarbeit mit Herrn Bär. Dabei werden unter Verwendung des optischen Flusses Segmentierungen zum nächsten Zeitschritt transformiert und anschließend mit der Prädiktion des Netzwerks verglichen. Ein Paper [BAE/KLI/FIN1] zu diesem Verfahren wurde auf dem SAIAD Workshop der IEEE Conference on Computer Vision and Pattern Recognition (CVPR) veröffentlicht.

Neben den Projekttätigkeiten fanden Abschlussarbeiten im Bereich performante semantische Segmentierung [Ba 20/02] und effiziente semantische Segmentierung [Ma 20/10] statt, um das Portfolio von Netzwerken zur semantischen Segmentierung des IfN aufzufrischen und weiterzuentwickeln.

Im Rahmen des SAIAD Workshops auf der IEEE Conference on Computer Vision and Pattern Recognition (CVPR) entstand darüber hinaus eine Veröffentlichung von Herrn Klingner zur Robustheit der semantischen Segmentierung gegenüber verrauschten oder attackierten Eingangsbildern [KLI/BAE/FIN1]. Es konnte gezeigt werden, dass durch einen Multitask-Lernansatz mit monokularer Tiefenschätzung auf mehreren Datensätzen die Robustheit des Netzwerks der semantischen Segmentierung gesteigert werden konnte – ohne zusätzliche Rechenkomplexität während der Inferenz. Gleichzeitig wird auch noch die absolute Performanz des Netzwerks gesteigert. Da dieser Multitask-Lernansatz viele positive Effekte hat, wird er weiter aktiv erforscht, um weitere Anwendungsgebiete zu erschließen.



(a) Eingangsbild (b) Tiefenschätzung

Abbildung 6: Qualitative Ergebnisse des im IfN entwickelten Netzwerks zur monokularen Tiefenschätzung. Jedem Pixel aus dem Eingangsbild (linkes Bild) wird eine Distanz zugeordnet (rechtes Bild).

Die semantische Segmentierung wird üblicherweise für eine fixe Anzahl an Klassen trainiert. Für manche Anwendungsfälle kann es im Nachhinein aber sinnvoll sein, zusätzliche Klassen in ein bereits vortrainiertes Netzwerk einzufügen. Im Rahmen einer Veröffentlichung auf der IEEE International Conference on Intelligent Transportation Systems (ITSC) wurde ein Verfahren hierfür entwickelt, welches insbesondere nicht den zum Vortraining benutzten Datensatz benötigt, sondern nur Labels für die neuen Klassen auf neuen Daten [BAE/FIN1]. Vertiefend wurde dies auch in einer Masterarbeit weiterentwickelt [Ma 20/009]. Ebenfalls abgeschlossen wurde eine Bachelorarbeit zur echtzeitfähigen Objekterkennung [Ba 20/006] und eine bei der BMW AG verfasste Masterarbeit zur Erhöhung der Fußgänger-Sicherheit beim Fahrzeugdesign durch Machine Learning [Ma 20/007].

5.2 Robuste Monokulare Tiefenschätzung

Ein erfolgreich neu erschlossenes Forschungsfeld der Arbeitsgruppe Signalverarbeitung und Machine Learning ist das Themengebiet der monokularen Tiefenschätzung aus Kamerabildern. Hierbei wird für jeden Inputpixel dessen Distanz von der Kamera vorhergesagt, so dass eine sogenannte Tiefenkarte entsteht (vgl. **Abbildung 6**). Besonders interessant wird dieses Verfahren dadurch, dass es eigenüberwacht aus sequentiellen Bilddaten trainiert werden kann und somit auf beliebige Videos anwendbar ist.

Die Kombination der Methode mit der bereits länger am Institut erforschten semantischen Segmentierung hat zu einer ersten Veröffentlichung der Arbeitsgruppe auf der European Conference on Computer Vision (ECCV) geführt [KLI/TER/FIN1]. Zusätzlich wurde eine neue Lösung für das Einbeziehen von dynamischen Objekten, wie Autos und Fußgängern, in den Trainingsprozess vorgestellt. Beide Neuerungen haben zu einem neuen Benchmark auf dem Datensatz KITTI Eigen Split geführt. Die Ergebnisse der Publikation beruhen auf den Untersuchungen einer am Institut durchgeführten Masterarbeit [Ma 20/006]. Eine Anwendung der Methode auf Fischaugen-Kameras wurde weiterhin in eine

Veröffentlichung auf dem Workshop on Applications of Computer Vision (WACV) eingebaut, die in Kooperation mit der Valeo Schalter und Sensoren GmbH entstand und zur Publikation im Januar 2021 angenommen wurde.

Das Thema ist darüber hinaus auch in mehreren Studien- und Abschlussarbeiten erforscht und erweitert worden. Themengebiete waren hier die halbüberwachte Tiefenschätzung [Ma 20/013], die eigenüberwachte Domänenanpassung durch Training der verschiedenen Aufgaben auf mehreren Datensätzen [Ma 20/014], [Ma 20/022] und die Übertragung des eigenüberwachten Lernkonzepts auf die Prädiktion des optischen Flusses [Ba 20/001]. Da die monokulare Tiefenschätzung so vielfältig einsetzbar ist, wird sie auch zukünftig intensiv in der Arbeitsgruppe erforscht werden.

5.3 Domain-Mismatch-Detektion und -Adaption

Die Performanz von Perzeptionsmethoden (z.B. neuronale Netze zur semantischen Segmentierung) ist stark abhängig von den verwendeten Trainingsdaten und sinkt typischerweise, wenn eine Änderung der Domäne vorliegt. Unter Domäne versteht man dabei die Verwendung gewisser (Kamera-)Sensoren, Tag- oder Nacht-Aufnahme oder ein gewisses Aussehen von Klassen in verschiedenen geographischen Regionen. Diese Änderung der Domäne gilt es einerseits zu detektieren und bestenfalls zu überbrücken.

Im Projekt KI-Absicherung werden u.a. Methoden entwickelt, um gelernte Perzeptionsfunktionen im autonomen Fahren gegen wechselnde Eingangsdomänen robuster zu machen. Ein vom IfN entwickeltes Verfahren leitet die Eingangsbilder einer zu überwachenden semantischen Segmentierung parallel in einen Autoencoder, dessen Aufgabe in der effizienten Codierung und Rekonstruktion der Bilder besteht. Der Trainingsprozess eines Autoencoders ist selbstüberwacht und hat damit den Vorteil, dass keine expliziten Annotationen zur Berechnung der Kostenfunktion und des Gütemaßes erforderlich sind. Durch das Auswerten der Autoencoder-Gütemaße lässt sich pro Datensatz ein Histogramm erzeugen, wodurch verschiedene Datensätze unter Zuhilfenahme der sogenannten Wasserstein-Distanz miteinander verglichen werden können. In Experimenten wurde gezeigt, dass der Unterschied in der Segmentierungsperformanz zwischen zwei Datensätzen mit der neu entwickelten Metrik korreliert ist und sich das Verfahren daher gut als Überwachungsfunktion für einen potentiellen Domain Mismatch im Fahrzeug eignet. Die Ergebnisse dieser Untersuchung wurden auf dem diesjährigen CVPR-Workshop SAIAD veröffentlicht und präsentiert [LOE/FIN1].

Zur Überbrückung der Abweichung zwischen Trainings- und Zieldomänen ist insbesondere die unüberwachte Domänenadaption von großer Bedeutung, da ein teures Labeling der Daten aus der Zieldomäne entfällt. Typischerweise erfordern

diese Methoden die Verfügbarkeit von Zieldomänen-Daten bereits während des Trainings auf der gelabelten Trainingsdomäne, z.B. im Labor, was allerdings in realen Anwendungen nicht immer gewährleistet werden kann. Herr Termöhlen untersuchte hierfür im Unterauftrag für die Volkswagen Group Innovation im Rahmen des Förderprojektes KI-DeltaLearning einen Ansatz, bei dem ein semantisches Segmentierungsnetz zunächst mit gelabelten Daten der Trainingsdomäne (im Labor) trainiert wird und anschließend (z.B. im Fahrzeug) eine unbeaufsichtigte Domänenanpassung durchgeführt wird. Damit die gelernten Informationen der Trainingsdomäne nicht vergessen werden, ist es hierbei nicht notwendig, den gesamten Datensatz zu speichern, sondern lediglich die Segmentierungslabels, die einen deutlich geringeren Speicherbedarf haben, und ein Bildgeneratormodell, welches aus diesen Segmentierungslabels nahezu fotorealistische Bilder erzeugen kann. Das Bildgeneratormodell wurde in einer Bachelorarbeit implementiert, in der auch erste erfolgreiche Experimente zum (überwachten) kontinuierlichen Training gemacht wurden [Ba 20/004]. Der entwickelte Ansatz für die kontinuierliche Domänenanpassung wurde auf einer großen Auswahl an Datensätzen und Netzwerkarchitekturen getestet.

Im Rahmen der Forschung am gleichzeitigen Training auf mehreren Datensätzen ist in der Abteilung Signalverarbeitung und Machine Learning auch ein weiteres Verfahren zur unüberwachten Domänenanpassung der semantischen Segmentierung entstanden. Das Besondere an diesem Verfahren ist, dass zunächst offline auf einem Datensatz trainiert werden kann und anschließend der Algorithmus online an einen neuen Datensatz angepasst werden kann, ohne dass Informationen über den ursprünglichen Trainingsdatensatz benötigt werden. Ein Paper zu diesem Verfahren wurde auf der IEEE Conference on Computer Vision and Pattern Recognition (CVPR) eingereicht.

5.4 Gelernte Bildkompression

Bis Ende des Jahres 2019 wurde die gelernte Bildkompression in dem von der Volkswagen Group Innovation finanzierten Industrieprojekt “GANs im autonomen Fahren” erforscht. Das Bildkompressions-Framework besteht aus einer Encoder-Quantisierer-Decoder-Kombination, welche durch neuronale Netzwerke realisiert ist und darauf trainiert wird, die Eingangsdaten nach einer starken Dimensions- und Auflösungsreduktion in möglichst hoher Qualität zu rekonstruieren. Die Quantisierung wird dabei entweder durch einen simplen Skalarquantisierer oder einen komplexeren, aber effizienteren Vektorquantisierer umgesetzt. Hierbei wurde der Einfluss verschiedener Kostenfunktionen, die während des Trainings neben der Verwendung von Skalar- oder Vektorquantisierung zum Einsatz kommen, auf die Rekonstruktionsqualität untersucht. Wie erwartet, hat sich in dieser Untersuchung gezeigt, dass die Vektorquantisierung eine deutlich effizientere Darstellung des Merkmalsraumes bewirkt und dass das übli-

cherweise verwendete Gütemaß der Bildkompression (peak signal-to-noise ratio, PSNR) zwar von einer MSE-Kostenfunktion profitiert, das subjektive Ergebnis allerdings mit der GAN-basierten Kostenfunktion deutlich besser ausfällt. Das Paper, welches aus diesen Erkenntnissen entstand, wurde auf der IEEE International Conference on Intelligent Transportation Systems (ITSC) eingereicht und angenommen [LOE/FIN2].

Nicht alle Regionen eines Bildes sind gleich relevant für Anwendungen des autonomen Fahrens. Um eine möglichst effiziente Verwendung der Bitrate, besonders im Low-Bitrate-Bereich, zu gewährleisten, ist das Ziel einer weiteren Untersuchung, die in Bildern eher irrelevanten Bereiche mit nachrangiger Priorität zu repräsentieren und mehr Bits für die wichtigen Bildbereiche zu verwenden. Dies wird im Encoder durch die Fusion des Bildes mit einer binären Maske erreicht, welche zuvor mittels einer semantischen Segmentierung aus dem Eingangsbild extrahiert wurde. Betrachtet man die als relevant gekennzeichneten Bereiche des Bildes, so konnte in Experimenten gezeigt werden, dass die Bildqualität im Vergleich zur Kompression ohne Fusionsverfahren auf Kosten der Gesamtqualität des Bildes gesteigert werden kann. Die Ergebnisse dieses Verfahrens wurden auf dem IEEE Intelligent Vehicles Symposium (IV) publiziert [LOE/BAE/FIN1].

Im Projekt KI DataTooling, welches im April 2020 begonnen hat, werden erstmals gesamtheitlich Tools und Methoden für die Bereitstellung von Daten aller Modalitäten (Kamera, Lidar, Radar, IMU, HD Karte, etc.) für KI-Anwendungen entwickelt und untersucht. Durch die integrierte Betrachtung von Realdaten und synthetischen Daten sowie der Nutzung von Effizienzpotentialen bei deren Kombination wird in diesem Vorhaben erstmals eine „Daten-Factory“ als Komplettlösung für das Training und die Validierung von KI-basierten automatisierten Fahrfunktionen entwickelt. Außerdem wird hier die gelernte Bildkompression fortgeführt. Erste Untersuchungen im Projekt befassen sich im Rahmen einer Masterarbeit [Ma 20/019] mit prädiktiver Quantisierung, bei der das zu komprimierende Signal zunächst prädiziert wird, so dass anschließend nur noch ein Restsignal zu übertragen ist. Nachdem das Codebuch der Quantisierung bisher durch den sogenannten (Linde-Buzo-Gray, LBG)-Algorithmus getrennt von dem eigentlichen Netzwerktraining erzeugt wurde, soll im weiteren Projektverlauf ein lernbares Codebuch entwickelt werden, wodurch die neuronalen Netze und die Quantisierung noch besser aneinander angepasst werden. Außerdem soll das Framework durch verschiedene Verfahren der Entropiecodierung erweitert werden, so dass eine weitere Senkung der Bitrate ohne Verringerung der Bildqualität möglich wird.

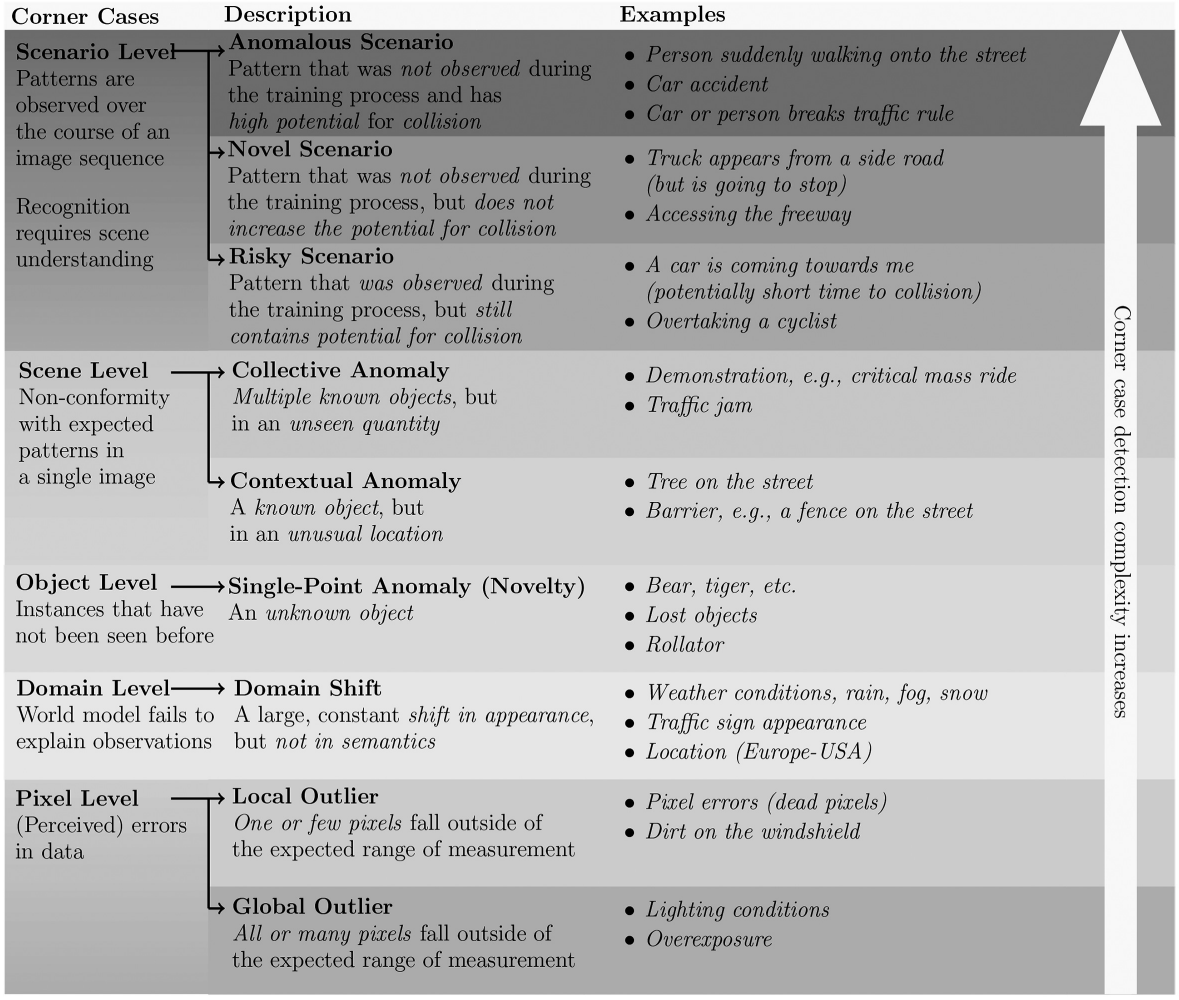


Abbildung 7: Systematisierung von Corner Cases auf den verschiedenen Leveln. Die zu erwartende technische Komplexität einer jeweiligen Detektion nimmt von unten nach oben zu.

5.5 Detektion von Corner Cases

Frau Breitenstein beschäftigt sich mit dem Themenfeld der sog. Corner-Case-Detektion. Unter Corner Cases versteht man zuerst einmal sämtliche Situationen oder Ereignisse, die im Straßenverkehr von der Normalität abweichen, ungewöhnlich oder sogar gefährlich sind. Um diese Definition zu schärfen, wurde zu Beginn dieses Jahres gemeinsam mit Herrn Termöhlen eine Systematisierung für Corner Cases für die visuelle Wahrnehmung im autonomen Fahren entwickelt und auf dem IEEE Intelligent Vehicles Symposium (IV) veröffentlicht [BRE/TER/FIN1]. **Abbildung 7** zeigt die entwickelte Systematisierung, die Corner Cases in sechs verschiedene Level gliedert: Pixel-, Domain-, Objekt-, Szenen- und Szenario-Level. Die zu erwartende technische Komplexität einer jeweili-

gen Detektion steigt dabei in dieser Reihenfolge, in der Abbildung 7 visualisiert durch den Pfeil. Die einzelnen Level werden dabei noch weiter in Unterkategorien unterteilt. Ziel dieser Systematisierung ist eine gezielte Entwicklung von Corner-Case-Detektoren und eine bessere Orientierung in der unübersichtlichen Menge möglicher kritischer Situationen.

Basierend auf der Systematisierung von Corner Cases wurden im Rahmen des Projekts mit der VW Group Innovation bestimmte Corner-Case-Fälle und ihre Detektion genauer untersucht und prototypische Detektoren dafür entwickelt. Frau Breitenstein hat sich dabei auf den Objekt- und Szenen-Level der Corner Cases fokussiert und drei prototypische Detektionsmethoden für Single-Point, Kontext- und kollektive Anomalien entwickelt. Unter Corner Cases auf Objektlevel, den Single-Point-Anomalien, verstehen wir unbekannte Objekte, die im Bild auftauchen. Für die Detektion dieser unbekannten Objekte wurde eine Entropie-basierte Detektions-Pipeline entwickelt. Im Rahmen einer Bachelorarbeit wurde diese Pipeline noch erweitert [Ba 20/012]. Bei Kontextanomalien handelt es sich um bekannte Objekte, die im Bild an ungewöhnlichen Orten auftauchen. Für ihre Detektion wurde ein sog. Orts-Prior auf Referenzdaten berechnet. Eine starke Orts-Abweichung der detektierten Objekte zu diesem Orts-Prior in der Anwendung führt zu einer Klassifikation als Kontextanomalie. In einer Studienarbeit wurden noch weitere Ansätze für die Berechnung von Orts-Prior-Wahrscheinlichkeiten zur Detektion von Kontextanomalien angewandt [Ba 20/013]. Kollektive Anomalien umfassen das Auftauchen bekannter Objekte in einer ungewöhnlichen Menge, z.B. Menschen auf einer Demonstration. Der Detektionsansatz basiert auf der Verteilung der Anzahl von Objektinstanzen pro Bild und berechnet so Abweichungen von neuen ungesehenen Daten zu Referenzdaten. Hierfür wird an einer Publikationseinreichung für 2021 gearbeitet. Zusätzlich ist im Rahmen des Projektes in einem umfangreichen Bericht [BRE/TER/FIN2] eine Erweiterung der Systematisierung um Ansätze zu Detektionsmethoden vorgelegt worden, welche für die verschiedenen Corner-Case-Fälle geeignete Detektionsansätze berücksichtigt.

Zum Thema Corner Cases wurde zudem eine Abschlussarbeit durchgeführt, die sich mit klassischen Methoden beschäftigt hat, um ungewöhnliche Nachbarschaftsverhältnisse in semantisch segmentierten Bildern zu detektieren [Ma 20/012]. Weiterhin wurde im Rahmen einer Bachelorarbeit untersucht, inwieweit sich die zeitliche Prädiktion von Segmentierungsmasken besser für die Corner-Case-Detektion eignet als die Prädiktion der eigentlichen Bilder [Ba 20/009]. In einer weiteren Studienarbeit wurde ein Datensatz zur Detektion von Überbelichtung in Bildern erstellt und eine erste Baseline-Methode implementiert [Ba 20/008].

Im Rahmen des Projekts KI DataTooling arbeitet Frau Breitenstein ab Oktober 2020 zunächst an einer Weiterentwicklung der Corner-Case-Definition basierend auf der bereits veröffentlichten Systematisierung [BRE/TER/FIN1].

5.6 Datensätze und Karten

In Zusammenarbeit mit Doktorand*innen der Volkswagen Group Innovation wurden in diesem Jahr zwei Publikationen zu Datensätzen und Karten veröffentlicht.

Gemeinsam mit Frau Breuer hat Herr Termöhlen ein Paper zum von Volkswagen Group Innovation veröffentlichten Datensatz openDD [TER/FIN1] verfasst, das auf der diesjährigen ITSC publiziert wurde. Der Datensatz beinhaltet Aufnahmen von sieben Kreisverkehren in Wolfsburg inklusive der Trajektorien und Begrenzungsboxen von über 80.000 verschiedenen Vekehrsbeteiligten. Die Aufnahmen wurden mit Hilfe von Drohnen angefertigt und haben eine Dauer von insgesamt 62 Stunden. Damit ist er derzeit einer der größten öffentlich verfügbaren Datensätze für Trajektorien aus der Drohnenperspektive. Der Datensatz bietet für jeden der erfassten Kreisverkehre eine hochaufgelöste Karte (inkl. befahrbarer Bereiche, Fahrspuren und Begrenzungslinien) und ein georeferenziertes Bild (inkl. befahrbarer Bereiche) an. Er ist für die Simulation von verschiedenen Verkehrsbeteiligten und für das Training und die Evaluation von Modellen zur Trajektorienprädiktion gedacht. Zusätzlich zu den Daten wurden im Paper sowohl Datensplits für wissenschaftliche Challenges als auch die zu verwendenden Metriken für die Trajektorienprädiktion beschrieben. Der Datensatz steht auf der Webseite des L3Pilot-Projektes zum Download verfügbar (http://www.l3pilot.eu/openDD) und darf mit einer Lizenz nach CC BY-ND 4.0 sowohl nicht-kommerziell als auch kommerziell genutzt werden.

Gemeinsam mit Herrn Plachetka hat Herr Termöhlen ein Paper zur Terminologie und Analyse von Abweichungen in hochaufgelösten Kartendaten für das autonome Fahren [TER/FIN2] auf dem Workshop on Online Map Validation and Road Model Creation veröffentlicht. Der Workshop fand in Verbindung mit dem diesjährigen IV-Symposium statt. Der Betrieb autonomer Fahrfunktionen beruht häufig auf Kartendaten und ist anfällig für Abweichungen zwischen der echten Welt, die durch die Fahrzeugsensorik erfasst wird, und den Kartendaten. In der Publikation wird das Konzept der Zuverlässigkeit aus dem Systems Engineering auf Kartendaten übertragen, um Zuverlässigkeit, Verfügbarkeit und eine sichere Nutzung der Karten gewährleisten zu können. Es wurden die notwendigen Begriffe zur Beschreibung und Messung von Kartenabweichungen definiert. Zusätzlich wurde aufbauend auf den Daten einer Messkampagne aus Hamburg im Dezember 2019 der vorgestellte Ansatz analysiert. Hierfür wurden Streckenabschnitte mit einer Länge von 127 km mit einem Messfahrzeug aufgenommen

und mit Kartendaten von 2017 verglichen. Die Ergebnisse zeigten, dass die analysierten Streckenabschnitte der Karte relativ wenige Fehler (0,07 / km) und keine offensichtlichen dauerhaften Veränderungen aufweisen. Es zeigte sich jedoch, dass temporäre Veränderungen (z.B. Baustellen) die Hauptursache für Kartenabweichungen in städtischen Gebieten sind, da fast 9 % der befahrenen Straßenabschnitte Abweichungen zur Karte zeigten.

6. Forschungsfeld „Predictive Maintenance“

6.1 Akustische Diagnose in Fahrzeugen

Herr Baumann erarbeitet in einem Kooperationsprojekt mit IAV GmbH Verfahren der KI-basierten, automatisierten Fehlergeräuscherkennung in der Fahrzeugdiagnose. Hier soll der momentane Zustand von Fahrzeug und Fahrzeugkomponenten mittels neuronalen Netzen aus einem Audiosignal geschätzt werden, welches z.B. durch die Freisprechmikrofone aufgenommen wird. Es wird dafür zunächst eine akustische Event-Erkennung mit neuronalen Netzen umgesetzt. So können unerwünschte oder Fehler-anzeigende Geräusche eines Fahrzeugs erkannt und durch Klassifikation teilweise auch die Geräuschquelle identifiziert werden. Des Weiteren wird aktuell eine Anomaliedetektion in akustischen Signalen umgesetzt. Diese kann unterstützend oder alleinstehend anomale Geräusche erkennen, ohne sie vorher im Training gesehen zu haben. Ein solches System zur Anomaliedetektion wurde in einer Abschlussarbeit [Ba 20/007] umgesetzt. Zur Erzeugung eines simulierten, realistischen Szenarios der akustischen Eventerkennung im Fahrzeug wurde seitens IAV ein Fahrzeuggeräusch-Datensatz erstellt, für den standardisiert mehrere Stunden Audiodaten aus verschiedenen, fehlerfreien Fahrzeugen aufgenommen wurden. Im Rahmen der DAGA 2020 wurde mit diesem Datensatz eine Erkennung von Babygeschrei im Fahrzeug [BAU/LOH/FRA/FIN1] vorgestellt. So soll z.B. eine potenzielle Ablenkung des Fahrers durch akustische Event-Erkennung vorausgesehen werden können. Im Zuge der Forschung an der akustischen Event-Erkennung wurde ein Paper [BAU/LOH/FIN1] auf der ICASSP 2020 veröffentlicht. Aufbauend auf einer Aufgabe der DCASE Challenge (Detection and Classification of Acoustic Scenes and Events) zur Erkennung von selten auftretenden akustischen Events wurde eine neue Aufgabe definiert, da in der ursprünglichen Aufgabenstellung leider eine unrealistische Situation vorlag, bei der Vorwissen zu der jeweils auftretenden Eventklasse genutzt werden durfte. Um ein realitätsnäheres Setup zu erzeugen, wurde eine neue Aufgabe definiert, bei der ohne besagtes Vorwissen alle Event-Klassen erkannt werden müssen. Unsere Experimente dazu zeigen, dass ein so trainierter Klassifizierer deutlich an Robustheit gewinnen kann.

6.2 Fehlerdiagnose aus betrieblichen Daten

Im Rahmen eines Förderprojekts aus dem mFUND-Programm des Bundesministeriums für Verkehr und digitale Infrastruktur (BMVI) arbeitet Herr Klingner in Kooperation mit der Siemens AG, dem Deutschen Forschungszentrum für künstliche Intelligenz (DFKI) und der DB Systel GmbH daran, neuronale Netzwerke auf die hochdimensionalen betrieblichen Daten eines Zuges, z.B. eines ICE, anzuwenden. Dabei ist insbesondere von Interesse, möglicherweise auftretende Fehler im Voraus zu diagnostizieren und so einen reibungsfreien Betriebsablauf zu unterstützen. Hierbei sollen verschiedene Methoden getestet werden, beispielsweise überwachte Lernverfahren, bei denen aus bereits aufgetretenen Fehlerfällen gelernt wird, diese vorherzusagen. Alternativ werden aber auch neuere unüberwachte Verfahren zur Anomaliedetektion verwendet, die aus normalen Betriebsdaten lernen und im Praxiseinsatz dann Abweichungen von diesem Zustand detektieren können. Die besondere Herausforderung des Projekts liegt allerdings darin, dass die KI nicht nur Fehlerdetektionen liefern soll, sondern Fehlerursachen auch noch sprachlich erklärt. Statt einer einfachen Wahrscheinlichkeit könnte somit ein Satz ausgegeben werden wie: “Bauteil xy weist einen außergewöhnlich hohen Verschleiß auf und sollte zeitnah ausgewechselt werden“. Das Projekt ist als Machbarkeitsstudie angelegt und soll somit vor allem die generelle Anwendbarkeit von neuronalen Netzwerken auf den konkreten Anwendungsfall bestätigen und eine Vorauswahl für erfolgversprechende Verfahren liefern.